Random forests to evaluate interspecific interactions in fish distribution models

By Vezza P.[1]*, Muñoz-Mas R.[1], Martinez-Capel F.[1], and Mouton A.[2]

[1] Institut d'Investigació per a la Gestió Integrada de Zones Costaneres (IGIC)
Universitat Politècnica de València, C/ Paranimf 1, 46730 Grau de Gandia. València. España.
[2] Research Institute for Nature and Forest (INBO), Kliniekstraat 25, B-1070 Brussels, Belgium

*Corresponding author: Paolo Vezza, e-mail: paovez@upv.es, voice: +34633856260

Abstract

Previous research indicated that high predictive performance in species distribution modelling can be obtained by combining both biotic and abiotic habitat variables. However, models developed for fish often only address physical habitat characteristics, thus omitting potentially important biotic factors. Therefore, we assessed the impact of biotic variables on fish habitat preferences in four selected stretches of the upper Cabriel River (E Spain).The occurrence of *Squalius pyrenaicus* and *Luciobarbus guiraonis* was related to environmental variables describing interspecific interactions (inferred by relationships among fish abundances) and channel hydro-morphological characteristics. Random Forests (RF) models were trained and then validated using independent datasets. In both training and validation phases, RF showed high performance. Water depth, channel width, fine substrate and water-surface gradient were selected as most important habitat variables for both fish. Results showed clear habitat overlapping between fish species and suggest that interspecific competition is not a strong factor in the study area.

Keywords: Interspecific interactions, Random Forests, *Squalius*, *Barbus*, species distribution modelling, mesohabitat

1. Introduction

According to the IUCN (International Union for Conservation of Nature), 56% of Mediterranean freshwater species are threatened (Smith and Darwall, 2006) and, given the high degree of endemicity of freshwater biota, native fish should be the target of actions for biodiversity conservation (Corbacho and Sánchez, 2001; Doadrio, 2002). Consequently in the last decade, efforts to understand the link between habitat attributes and fish habitat use have increased, and currently habitat modelling for freshwater fish is considered an important field of research (Guay et al., 2000; Lamouroux and Jowett, 2005; Olden et al., 2008; Strayer and Dudgeon, 2010; Mouton et al., 2011; Fukuda et al., 2012).

This study focused on *Squalius pyrenaicus* (Southern Iberian Chub) and *Luciobarbus guiraonis* (Eastern Iberian barbel), two threatened fish species (Baillie et al., 2004) characteristic for Mediterranean rivers of Eastern Spain (Crivelli, 1996). These two species may act as an indicator for other Mediterranean fish since they face similar threats and knowledge gaps (Doadrio, 2001). Specifically, these fish populations have been declining due to habitat modification and water abstraction, as well as due to the introduction of alien species (e.g., *Esox lucius*, Hermoso et al., 2010; Maceda-Veiga, 2012). Few studies have investigated the ecology of these fish (Crivelli, 1996) and, to our knowledge, no habitat or fish distribution models are currently available for either *S. pyrenaicus* or *L. guiraonis*, like for most endemic fish species of the Iberian Peninsula (Grossman and De Sostoa, 1994; Magalhães et al., 2002; Martínez-Capel et al., 2009; Costa et al., 2012).

S. pyrenaicus is distributed in most of the large river basins of the Eastern and Southern Iberian Peninsula (Doadrio and Carmona, 2006). However, the species has become rare due to habitat loss and it was classified as Near Threatened (NT) in the IUCN red list (Baillie et al., 2004). Pires et al. (2000) investigated the ecology and life history strategies of *S. pyrenaicus* in some reaches of the middle Guadiana basin (Portugal), focusing on its growth rates and behavioural adaptations to summer drought. Kottelat and Freyhof (2007) described *S. pyrenaicus* as an ubiquitous species that inhabits small to medium-sized streams with a Mediterranean flow regime. Ferreira et al. (2007) found that *S. pyrenaicus* occurrence in the streams of central and Northern Portugal depends on the availability of coarse substrate and shading by overhanging trees.

L. guiraonis is a native species of the middle and lower river courses of the Jucar River Basin District, dwelling also in lakes and reservoirs (Crivelli, 1996). In particular, its natural range is restricted to the region between the rivers Mijares and Serpis, but it has also been translocated in the upper part of the Guadiana river basin (Hermoso et al., 2011). The species is classified as a vulnerable species (Baillie et al., 2004) and local populations are heavily affected by habitat alteration and water abstraction. It is a large barbel (up to 50 cm in length) that migrates to upstream stretches during the spawning season (from April to June, Kottelat and Freyhof, 2007).

When studying fish distribution, researchers assume that the associations of fish species and habitat characteristics arise from either biotic or abiotic variables or some combination of the two (Guisan and Thuiller, 2005). However, very few habitat models explicitly include biotic factors, which can be used to infer or provide clues about inter-specific interactions (Elith and Leathwick, 2009). Indeed, habitat requirements for fish are often defined as abiotic features of the environment that are necessary for the survival and persistence of individuals or populations (Rosenfield, 2003, Ahmadi-Nedushan et al. 2006). The habitat suitability index (HSI, Bovee, 1982), the most commonly used index of habitat quality, is an analytical tool used to represent preferences of different aquatic species for physical instream variables (e.g., velocity, depth, substrate, cover). This approach has been criticized because such models almost exclusively address physical habitat characteristics, thus omitting potentially important biotic factors (Armstrong et al., 2003; Rosenfeld, 2003; Teichert et al., 2010) and because the relationships fit poorly when transferred across different river morphologies (Armstrong et al., 2003).

Wisz et al. (2013) reported that one solution to account for interspecific interactions is to use species distribution models in concert with biotic surrogate variables that reflect spatial turnover or gradients in the distribution of biotic interactions. To model species distribution, Random Forests (RF, Breiman, 2001), a statistical method based on an automatic combination of decision trees, is currently considered a promising technique in ecology (Cutler et al., 2007; Franklin, 2010; Drew et al., 2011; Cheng et al., 2012). RF has been applied in freshwater fish studies (Buisson et al., 2010; Grenouillet et al., 2011; Markovic et al., 2012) and several authors have shown that, compared to other methodologies, RF often reach top performance in building predictive models of species distribution (Svetnik et al., 2003; Siroky, 2009; He et al., 2010; Mouton et al., 2011). Moreover, RF has been recently included in mesohabitat simulation tools, i.e., MesoHABSIM (Parasiewicz et al., 2013; Vezza et al., 2014a) to model fish ecological response to hydro-morphological alterations. However, current applications at the mesohabitat scale (or mesoscale) focus on the evaluation of physical habitat for aquatic species and no studies are currently available to include both biotic and abiotic habitat variables in these analyses.

To develop a reliable and ecologically relevant species distribution model, we used RF to predict fish distribution at the mesohabitat scale, based on both biotic and abiotic habitat variables. The aims of the study were: (i) to investigate which are the most important variables predicting the presence of *S. pyrenaicus* and *L. Guiraonis*, (ii) evaluate how interspecific interactions affect habitat use and (iii) validate the developed models using an independent data set to test its values for potential users.

2. Methods

2.1 Study area

Data were collected on eight sampling sites of the Cabriel River (Fig. 1), which were selected based on their natural habitat conditions (i.e., absence of water abstractions, natural flow regime and river morphology) and the presence of age-structured populations of *S. pyrenaicus* and *L.*

guiraonis. The Cabriel River is part of the Júcar River Basin, which is one of the pilot basins for the implementation of the Water Framework Directive in Spain. In total, the river is 220 km long, the catchment elevation ranges from 490 to 1790 m a.s.l. and its drainage area covers 4750 km^2. The study area has a typical Mediterranean climate with a mean annual precipitation of ca. 500 mm, resulting in low flows and high evapotranspiration in summer and high flows in spring and autumn.

Study sites for both model training and validation (named, respectively, T1, T2, T3 and T4 and V1, V2, V3 and V4 in downstream order, Fig. 1) were all located in the upper part of the Cabriel catchment (province of Cuenca, Spain), upstream of the large Contreras Dam. In this part of the watershed, the average riverbed slope is 1.1% and land cover (from the Corine Land Cover classification; Bossard et al., 2000) mainly consists of forested areas (86%) and crops (12%). The study sites were selected as to differ in both their morphological characteristics (mean gradient, channel size and substrate composition), and flow duration curves, and, at the most downstream site (V4), the low (Q95), mean (Q50) and high (Q5) flow are, respectively, 0.94, 2.74 and 15.83 m^{3}s^{-1}. Salmonidae and cyprinidae are the predominant families. Besides *S. pyrenaicus* and *L. Guiraonis*, *Parachondrostomas arrigonis* (Júcar nase), *Pseudochondrostoma polylepis* (Tagus nase), *Gobio lozanoi* (Iberian gudgeon) and *Salmo trutta* (brown trout) are the species present in the upstream course of the Cabriel River (CHJ, 2007).

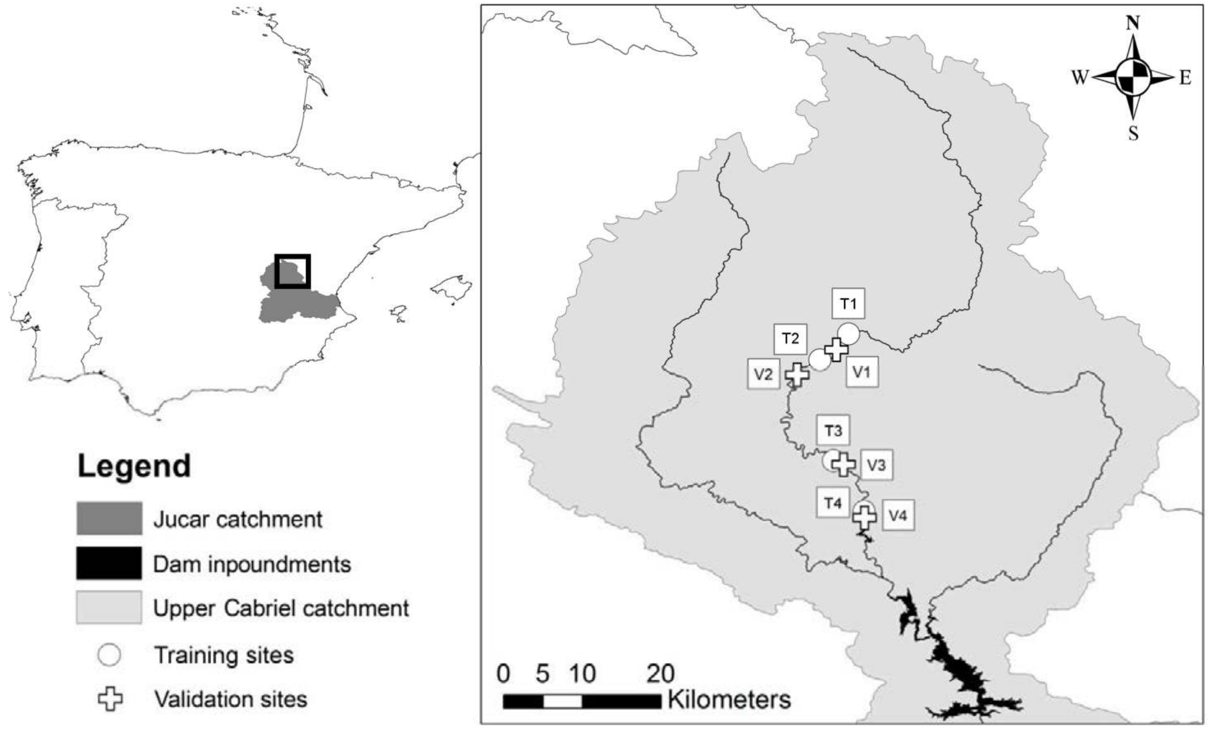

Figure 1. Location of the training (T$_i$) and validation (V$_i$) study sites in the upper Cabriel catchment (Júcar River basin, Spain). The main watercourses and the large reservoir of Contreras are also shown.

2.2 Habitat description and fish data

Data were collected at the mesohabitat scale and the hydromorphological unit – HMU (e.g., pools, riffles, rapids) was considered the sampling unit for this study. HMUs often correspond in size and location to mesohabitats (Bain and Knight, 1996; Parasiewicz, 2007; Hauer et al., 2010) and can be used to capture the confounded effects of biotic and abiotic environmental variables, focusing on how aquatic species interact with the spatial arrangement of different habitat characteristics (Addicott et al., 1987; Kemp et al., 1999).

Each site used for model training was at least 1 km long and was surveyed two to three times to record the distribution of HMUs and habitat variables. The total length of each sampling site was not constant, as the size of HMUs varied with flow (Costa et al., 2012). The four river stretches for model validation were shorter (ranging from 0.3 km to 0.6 km) and, due to the limited availability of access points to the river, V3 and V4 partially overlapped T3 and T4 respectively, but were surveyed at different moments in time. Specifically, habitat surveys and fish population assessment for model training were carried out between 2006 and 2009, whereas, data for model validation were collected between 2011 and 2012. Although a partial overlapping between training and validation sites occurred, this temporal distance and the variation in flow conditions between fish sampling campaigns ensured the independence of validation data from those used for model training. Surveys took place from June to October, i.e. after both species' spawning period (Doadrio, 2001), and during low to medium flows (i.e. ranging from Q98 to Q40) to represent the habitat availability in the upper Cabriel River.

Following previous research in Mediterranean rivers (Alcaraz-Hernández et al., 2011), five types of HMUs were considered: pool, glide, run, riffle and rapid. Pools were characterized by moderate to high water depth (> 0.5 m) generally associated with erosion phenomena, low flow velocity and a very low gradient. Glides were characterized by moderate to high water depth (> 0.5 m), low flow velocity and nearly symmetrical cross-sections. Riffles were characterised by the occurrence of surface ripples and moderate to high flow velocity (> 0.2 m/s) , whereas runs are similar to riffles but lack pronounced waves and ripples on the water surface. Finally, rapids were characterized by shallowness, a moderate to high gradient and abundant white-waters and macro-roughness elements. For each HMU, the following habitat variables were collected: longitudinal length, channel width, water-surface gradient, mean water depth, mean flow velocity, substrate composition and cover (Table 1). The first three variables, used to describe HMU size and longitudinal slope of the water surface, were measured through the CMII Hip Chain (CSP Forestry Ltd. Alford, Scotland), the laser distancemeter DISTO A5 (Leica Geosystems, Heerbrugg, Switzerland), and the Haglöf HEC Electronic Clinometer (Haglöf Sweden AB, Långsele, Sweden), respectively.

Mean water depth was calculated from point measurements uniformly distributed in four to eight cross-sections along the HMU, and each cross-section was entirely located in only one HMU type. The mean flow velocity of each HMU was calculated by dividing the value of the discharge during the survey (available at Pajaroncillo gauging station) by the mean HMU cross-section area. The substrate composition was assessed by eye and expressed as percentage of bedrock, coarse substrate (boulders and cobbles), fine substrate (gravel and sand), sludge (silt and clay) and submerged vegetation. To represent cover availability for fish, canopy shading (as the percentage of the overall HMU's area), undercut banks (as the percentage of the HMU's length) and the presence of large boulders and woody debris were included. Finally, both the reach mean width and gradient of each sampling site were included in the analysis as proxies of channel morphology to evaluate possible site-scale effects on fish distribution.

Table 1. Code, description, unit and range of the habitat variables included in RF models. Each fish species abundance was considered as a biotic habitat variable and was expressed by three classes: abs = absent, pres = present and abu = abundant.

Variable code	Description	Unit	Range
Len	Longitudinal length of the hydro-morphological unit	m	9 - 108
Wid	Mean channel width	m	2.7 - 20.0
Dmed	Mean water depth	m	0.29 – 3.52
Vmed	Mean flow velocity	m/s	0.04 – 1.05
Grad	HMU gradient (longitudinal slope of the water surface)	%	0.0 – 9.3
RK	Bedrock substrate	%	0-100
CS	Coarse substrate (boulders and cobbles)	%	0-100
FS	Fine substrate (gravel and sand)	%	0-100
SC	Silt and clay substrate	%	0-60
SV	Submerged vegetation	%	0-90
Sh	Canopy shading	%	0-100
UB	Undercut banks	%	0-100
WD	Woody debris	-	yes/no
B	Boulder cover	-	yes/no
RWid	Reach mean channel width	m	6.5-11.9
RGrad	Reach mean gradient	%	1.4-3.5
ASP	Abundance of *Squalius pyrenaicus* (Southern Iberian chub)	-	abs/pres/abu
ALG	Abundance of *Luciobarbus guiraonis* (Eastern Iberian barbel)	-	abs/pres/abu
APA	Abundance of *Parachondrostomas arrigonis* (Júcar nase)	-	abs/pres/abu
APP	Abundance of *Pseudochondrostoma polylepis* (Tagus nase)	-	abs/pres/abu
AGL	Abundance of *Gobio lozanoi* (Iberian gudgeon)	-	abs/pres/abu
AST	Abundance of *Salmo trutta* (brown trout)	-	abs/pres/abu

Fish were counted in each HMU by snorkelling, as to observe habitat use during their diurnal routine. Two divers conducted the underwater counts in three independent passes from downstream to upstream (Baillie et al., 2004) throughout each HMU of each sampling site (Costa et al., 2012). Three snorkelling passes were considered enough to ensure a reasonably uniform probability of detection (Schill and Griffith, 1984), and, for each HMU, the sampling effort (expressed in minutes per unit area) and the number of counted fish was consistent among passes (coefficient of determination between two independent passes, $R^2 > 0.95$). To ensure that each pass was independent, and not affected by previous passes, a time delay of about two hours was programmed between successive counts (*sensu*, Bain et al., 1985). The snorkelling technique was chosen for its effectiveness to assess fish population density at the mesoscale and to avoid any damage to the threatened target species. Moreover, we considered it the most appropriate methodology for this study due to the morphological characteristics of the river (i.e. clear water, presence of pools and low channel width). However, underwater counts may fail to observe and classify fish in the shortest length class (Joyce and Hubert, 2003) and only fish > 5 cm for *S. pyrenaicus* and >10 cm for *L. guiraonis* were considered in the analysis. This allowed us to focus on adult fish and develop habitat models for 2+ or older individuals (García de Jalón et al., 1999; Pires et al., 2000).

To produce species distribution models, which can be implemented in common mesohabitat simulation tools, the dependent variable was defined as a binary response (i.e., fish absence/presence) for both *S. pyrenaicus* and *L. guiraonis*. To investigate the influence of interspecific interactions, the abundance of each observed fish species was included as biotic independent variable (Table 1). Specifically, for each species we classified fish abundance in three classes (absent, present and abundant). The cutoff value (expressed in individuals/m^2) for low and high abundance was determined as the inflection point of the envelope curve of the fish density histograms (Parasiewicz, 2007).

Data from 240 HMUs were used for *S. pyrenaicus* model training, whereas an independent dataset of 48 HMUs (20% of the training data-set) was used for model validation. For *L. guiraonis*, due to the absence of adult specimens in T1 and V1 sampling sites, the data from these stretches were excluded from model development and only 110 and 22 HMUs were considered respectively for model training and validation. T1, showing the highest gradient and the narrowest and most constrained channel, is the most diverse and variable stretch based on flow conditions. Due to the exclusion of T1 from *L. guiraonis* model construction, the two databases mainly differed in terms of number of observations, minimum channel width and maximum gradient of riffles and rapids (Table 2). In terms of fish occurrence, the model prevalence for *S. pyrenaicus* was 0.54 in training and 0.38 in validation, whereas for *L. guiraonis* it was 0.64 and 0.59, respectively.

Table 2. Description of the five HMU types in the study area. Proportion of samples HMUs, range of mean water depth, mean flow velocity and channel width, dominant substrates and proportional occurrence of fish are reported for each category. See Table 1 for substrate codes.

Squalius pyrenaicus

HMU (N. Tot = 240)	% over sampled HMUs	Water depth	Flow velocity	Channel width	Dominant substrate	Fish occurrence
Units	(%)	(m)	(m/s)	(m)	(-)	(%)
Pool	34	0.54-3.52	0.04-0.33	4.5-15.2	FS-SV	77
Glide	4	0.50-1.73	0.08-0.28	4.4-14.7	FS-SV	72
Riffle	45	0.29-2.38	0.18-0.84	3.2-20.0	CS-FS-SV	47
Run	3	0.92-1.39	0.27-0.41	8.2-12.3	CS-FS	80
Rapid	14	0.30-0.88	0.13-1.05	2.7-13.6	CS	9

Luciobarbus guiraonis

HMU (N. Tot = 110)	% over sampled HMUs	Water depth	Flow velocity	Channel width	Dominant substrate	Fish occurrence
Units	(%)	(m)	(m/s)	(m)	(-)	(%)
Pool	32	0.62-3.52	0.08-0.33	6.05-15.2	FS	86
Glide	6	0.80-1.70	0.12-0.28	11.4-14.7	FS-SV	83
Riffle	39	0.32-2.38	0.14-0.84	4.25-20.0	CS-FS-SV	54
Run	4	0.92-1.39	0.27-0.41	9.3-12.3	CS-FS	40
Rapid	19	0.30-0.75	0.21-1.05	4.7-13.0	CS	28

2.3 Data analysis

Since many sampling units were contiguous, we firstly measured and tested spatial autocorrelation by means of Moran’s *I* with associated z-values (R package “spdep”, Bivand, 2012). For this analysis, the fish data collected in each HMU and the Euclidean distance between HMU centroids were used to calculate Moran’s *I* and z-values in each surveyed river reach (Elith and Leathwick, 2009; Planque et al., 2011).

To find effective habitat suitability criteria, the relationship between habitat variables and fish presence was explored by Random Forests (Breiman, 2001; Cutler et al., 2007), as implemented in R (R Development Core Team 2009; Liaw and Wiener, 2002). RF is an ensemble learning technique based on the combination of a large set of decision trees (i.e., Classification and Regression Trees - CART, Breiman et al., 1984). The CART technique splits a learning sample using an algorithm known as binary recursive partitioning, by which the data set is divided into two parts by maximizing the homogeneity in the two child nodes. This splitting or partitioning starts from the most important variable to the less important ones and it is applied to each of the new branches of the tree (Vezza et al., 2010).

In RF, each tree of the forest is grown by selecting a random bootstrap subset X_i (where i = the index of the bootstrap iteration, ranging from 1 to the maximum number of trees t) of the original dataset X and a random set of predictive variables (Liaw and Wiener, 2002). This represents the main difference compared to standard decision trees, where each node is split using the best split among all predictive variables. Moreover, RF corrects many of the known issues in CART, such as over-fitting (Breiman, 2001), and provides very well-supported predictions with large numbers of independent variables (Cutler et al., 2007). As the response variable was categorical (fish presence/absence), we confined our attention to classification RF models. The algorithm for growing a RF of t classification trees performs as follows (for full details see Breiman, 2001):

i) t bootstrap subsets X_i (the training dataset) are randomly drawn with replacement from the original dataset, each containing approximately two third of the elements of the original dataset X. The elements not included in the training dataset are referred to as out-of-bag (OOB) data for that bootstrap sample. On average, each element of X is an OOB element in one-third of the t iterations.

ii) For each bootstrap sample X_i, an unpruned classification tree is grown. At each node m variables are randomly selected and the best split is chosen between them.

iii) The trees are fully grown and each tree is used to predict OOB observations. New predictions (for the OOB elements) are calculated by means of the majority vote of OOB predictions of the t generated trees. In particular, the predictions from all the trees are combined to predict an observation class (as well as a probabilistic prediction output) for that observation. Note that, as OOB observations are not used in the fitting of RF trees, the out-of-bag estimates are essentially cross-validated accuracy estimates.

iv) Global RF accuracies and error rates (i.e. the OOB error, E_{OOB}, and within-class errors, $E_{Class(j)}$) are finally computed using OOB predictions.

The E_{OOB} is also used to choose an optimal value of t. In our analysis E_{OOB} stabilization occurred between t = 1500 and t = 2500 replicates. However, a heuristic estimation of t taking into account for E_{OOB} stabilization and variable interaction with a large set of independent variables is defined as [2*(t for E_{OOB} stabilization) = 5000] (Evans and Cushman, 2009). The m parameter (indicating the number of variables permutated at each node) is defined as the square root of the total number of predictor variables included in each model, with a minimum of m = 2 (Breiman, 2001).

To assess the importance of a specific predictor, in RF the values of each variable are randomly permuted for the OOB observations, and then the modified OOB data are passed down the tree to get new predictions. The difference between the prediction accuracy before and after the permutation gives the importance of a variable for one tree, and the importance of the variable for the forest is computed as an average over all trees. However, the permutation importance embedded in the RF algorithm overestimates the variable importance of highly correlated variables. Thus, a conditional variable importance, proposed by Strobl et al. (2008), was used in this study to avoid bias towards correlated predictor variables.

As model parsimony is important for future model applications (i.e., less variables to be surveyed), the most parsimonious model was identified by the Model Improvement Ratio (MIR, Murphy et al., 2010) technique. The improvement ratio was calculated as [*In/Imax*], where *In* is the importance of a given variable and *Imax* is the maximum model improvement score. Starting from *MIR* = 0, we then iterated through MIR thresholds (i.e. 0.02 increments), with all variables above the threshold retained for each model (Evans and Cushman, 2009). The models corresponding to different subsets were then compared and the model exhibiting the minimum E_{OOB} and the lowest maximum $E_{Class(j)}$ was selected (Fig. 2). Lastly, to avoid collinearity effects on the model performance, the correlation among the selected variables was tested using a correlation matrix. For models including both numerical and categorical variables, an heterogeneous correlation matrix was computed using the polycor package in R (Fox, 2007).

The performance of the predictive models was evaluated using five performance metrics, i.e., accuracy, sensitivity, specificity, Cohen’s kappa (*k)* area under Receiver Operating Characteristic (ROC) curve (AUC), and true skill statistic (TSS), which are commonly used in ecological modeling (Mouton et al., 2010). Accuracy represents the proportion of overall correctly classified observations, while sensitivity and specificity, respectively, refer to the proportion of actual positives and negatives correctly identified as such. The *k* coefficient, which takes into account the agreement occurring by chance, is a statistical measure of inter-rater agreement for categorical items. However, the chance percentage can provide misleading results as a low kappa (i.e. 0) could result for a model with good agreement if one category dominates the data (i.e., low or high prevalence, Bennett et al., 2013). To address this issue, AUC, measured from ROC plots, and TSS (Allouche et al., 2006) are used as performance metrics that are independent of prevalence (Mouton et al. 2010) and represents useful measures of how well a model is parameterized and calibrated. Furthermore, a confusion matrix, expressed as a bar chart, allowed visualization of model performance.

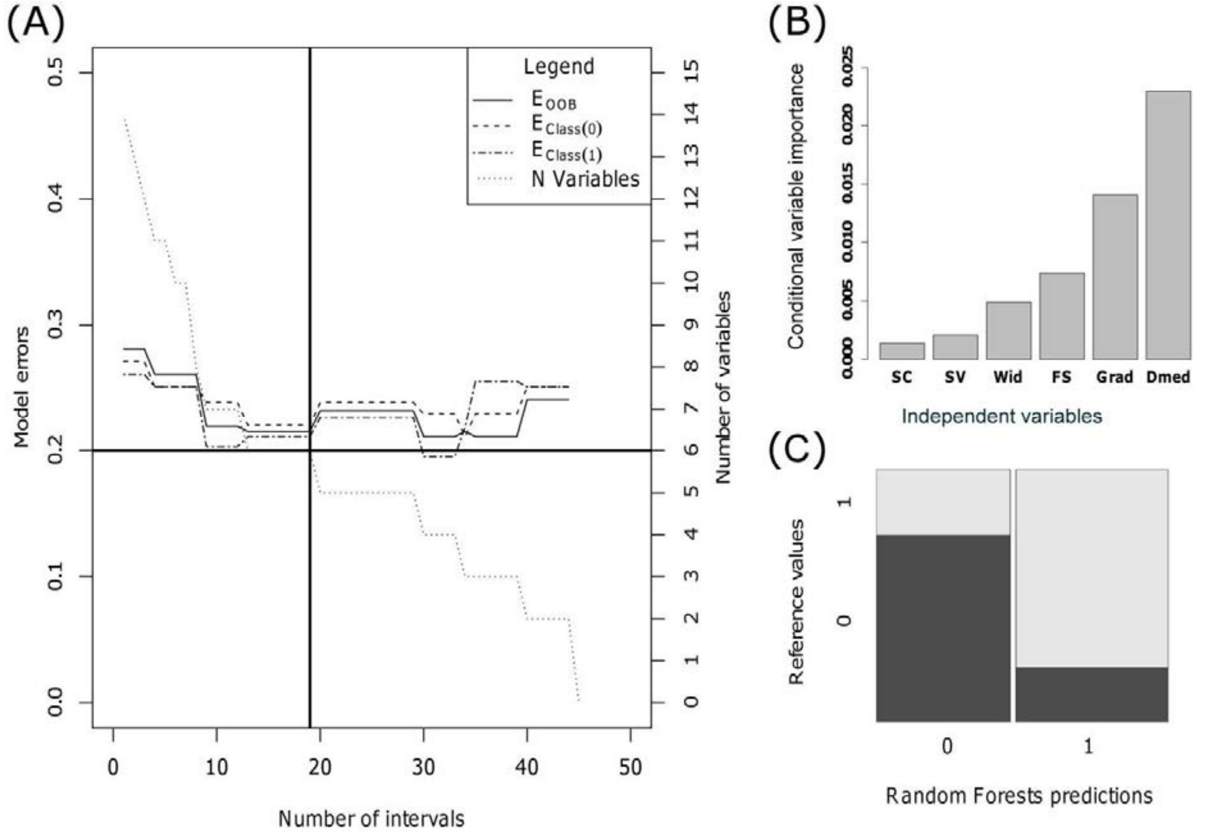

Figure 2. Habitat model for *S. pyrenaicus*. (A) Model Improvement Ratio technique (Murphy et al., 2010) showing the out-of-bag error (E_{OOB}) and within class errors ($E_{Class(j)}$) related to increment intervals which defined the number of selected variables for each subset model. The model that exhibits the minimum E_{OOB} and lowest maximum $E_{Class(j)}$ was selected (i.e. 6 input variables). (B) Relevant habitat variables for *S. pyrenaicus* presence and their relative importance by conditional variable importance (Strobl et al., 2008). Mean water depth (Dmed), HMU gradient (Grad), channel width (Wid), proportion of fine substrate (FS), submerged vegetation (SV) and sludge (silt and clay, SC) were selected as the most important habitat variables. (C) Confusion matrix of the selected RF model expressed as bar charts.

The partial dependence plots provided a way to visualize the marginal effect of the selected independent variables on the probability of fish presence (Cutler et al., 2007). Specifically, these plots can be used to graphically characterize the relationship between habitat variables and the predicted probabilities of fish presence obtained by RF. Finally, to test for the influence of interspecific interactions, three binary models were constructed: (i) one using only abiotic habitat variables, (ii) one using both biotic and abiotic habitat variables, and (iii) one using only biotic habitat variables.

3. Results

Pools, riffles and rapids (occurrence = 34%, 45%, 14%, respectively) were the most common hydromorphological units (HMUs) in the upper Cabriel River, whereas glide and run (occurrence = 4% and 3%, respectively) could be considered as rare. *S. pyrenaicus* occurred most frequently in pools, glides and runs, whilst it was less frequent in riffles and almost absent in rapids. *L. guiraonis* showed a similar distribution pattern, its frequency of occurrence decreasing as the flow velocity was increasing; most barbels were found in HMUs classified as pools, whereas their presence was the lowest in rapids (Table 2). Spatial dependency in fish distribution was tested by Moran's *I* with associated z-values, that suggested a random spatial pattern (z-values <|1.96|) and showed no evidence of spatial autocorrelation.

The models including only abiotic variables showed 76% and 84% accuracy for *S. pyrenaicus* and *L. guiraonis*, respectively, whereas Cohen's kappa, AUC and TSS were respectively 0.52, 0.80 and 0.54 for *S. pyrenaicus*, and 0.66, 0.85 and 0.68 for *L. guiraonis* (Fig. 3). Although these models performed well, considering biological interactions among species slightly increased the models performance. Specifically, the models for *S. pyrenaicus* and *L. guiraonis* reached 80% and 91% accuracy, Cohen's Kappa values of 0.59 and 0.80, AUC values of 0.85 and 0.95, and TSS values of 0.60 and 0.80, respectively. The model built using only biotic variables showed the lowest performance, i.e., 72% and 77% accuracy, 0.45 and 0.53 Cohen's Kappa, 0.72 and 0.76 AUC, and 0.44 and 0.55 TSS, for the *S. pyrenaicus* and *L. guiraonis* models, respectively (Fig. 3).

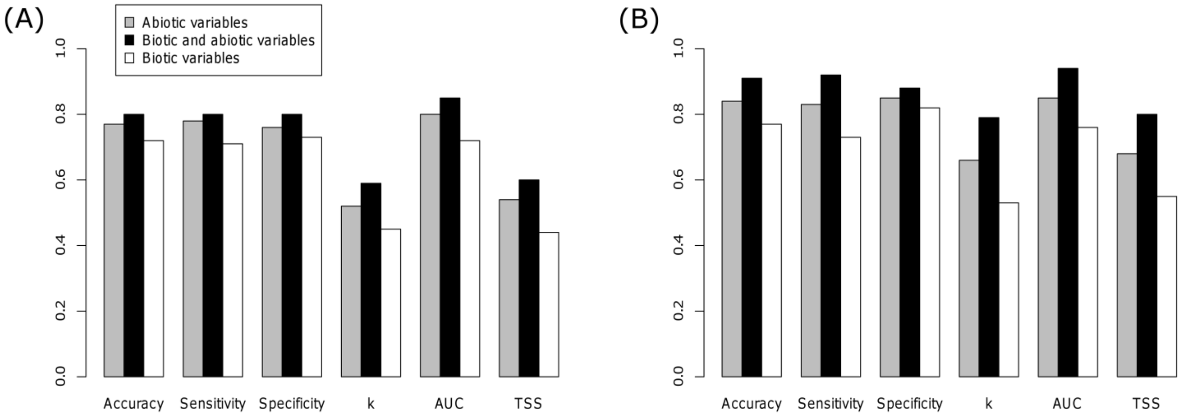

Figure 3. Random Forests model performance for (A) *S. pyrenaicus* and (B) *L. guiraonis* using (i) only abiotic, (ii) both biotic and abiotic, and (iii) only biotic habitat variables. Model accuracy (in terms of correctly classified observations), sensitivity, specificity, Cohen's kappa (k), area under the ROC curve (AUC) and true skill statistic (TSS) are shown for each model.

According to the partial dependence plots (Fig. 4), the models developed using only abiotic variables provided similar sets of selected inputs for the two species, although variable were ranked differently. Specifically, mean water depth (Dmed), channel width (Wid) and the proportion of fine substrate (FS) were positively correlated with the presence of both fish species, whilst HMU gradient (Grad) was negatively related to the presence of both species. The probability of presence of *S. pyrenaicus* also increased with the proportion of submerged vegetation (SV) and decreased with the percentage of sludge (silt and clay, SC) (Fig. 4).

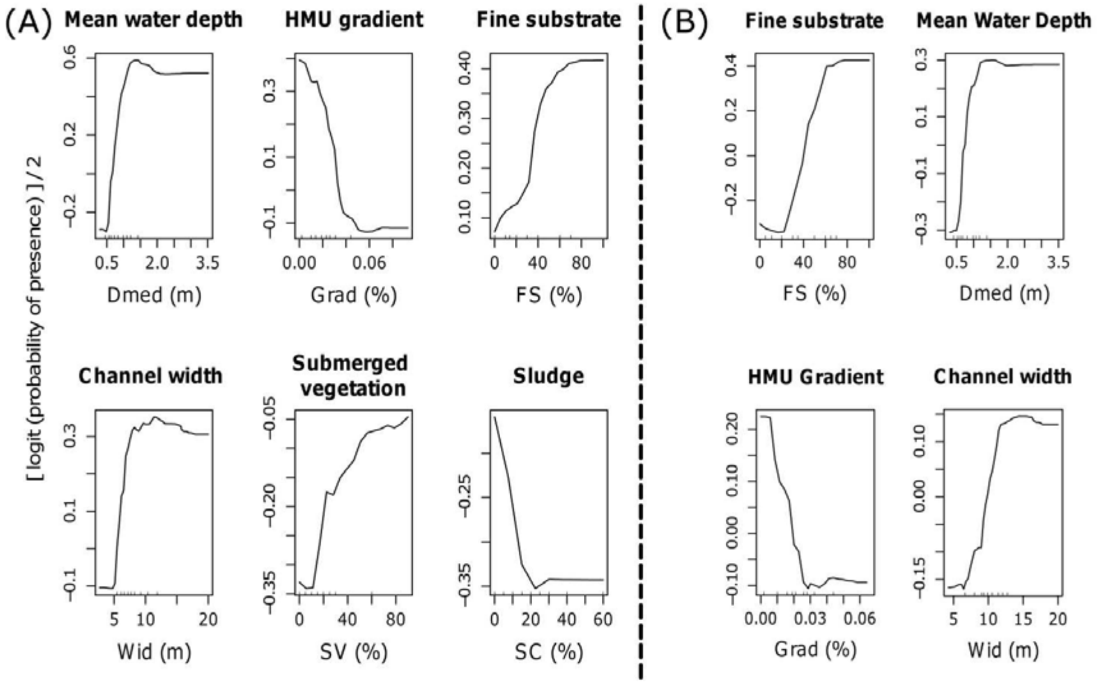

Figure 4. Partial dependence plots of the habitat models for (A) *S. pyrenaicus* and (B) *L. guiraonis*. Partial plots represent the marginal effect of a single variable included in the RF model on the probability of fish presence, while averaging out the effect of all the other parameters (Cutler et al., 2007). Selected variables are reported in order of importance.

In the models including both biotic and abiotic variables, the abundance of three cyprinid species was positively correlated to the probability of presence of both target fish species (Fig. 5). Specifically, the abundances of *L. guiraonis*, *P. arrigonis* and *G. lozanoi* were selected in the model for *S. pyrenaicus*, whereas the abundances of *P. Arrigonis*, *S. pyrenaicus* and *G. lozanoi* were selected in the model for *L. guiraonis*. However, when *L. Guiraonis* was abundant, the probability of presence of *S. pyrenaicus* slightly decreased. Mean water depth and the proportion of fine substrate were also selected as important abiotic variables in both fish models, whereas only HMU gradient was selected in the model for *S. pyrenaicus*.

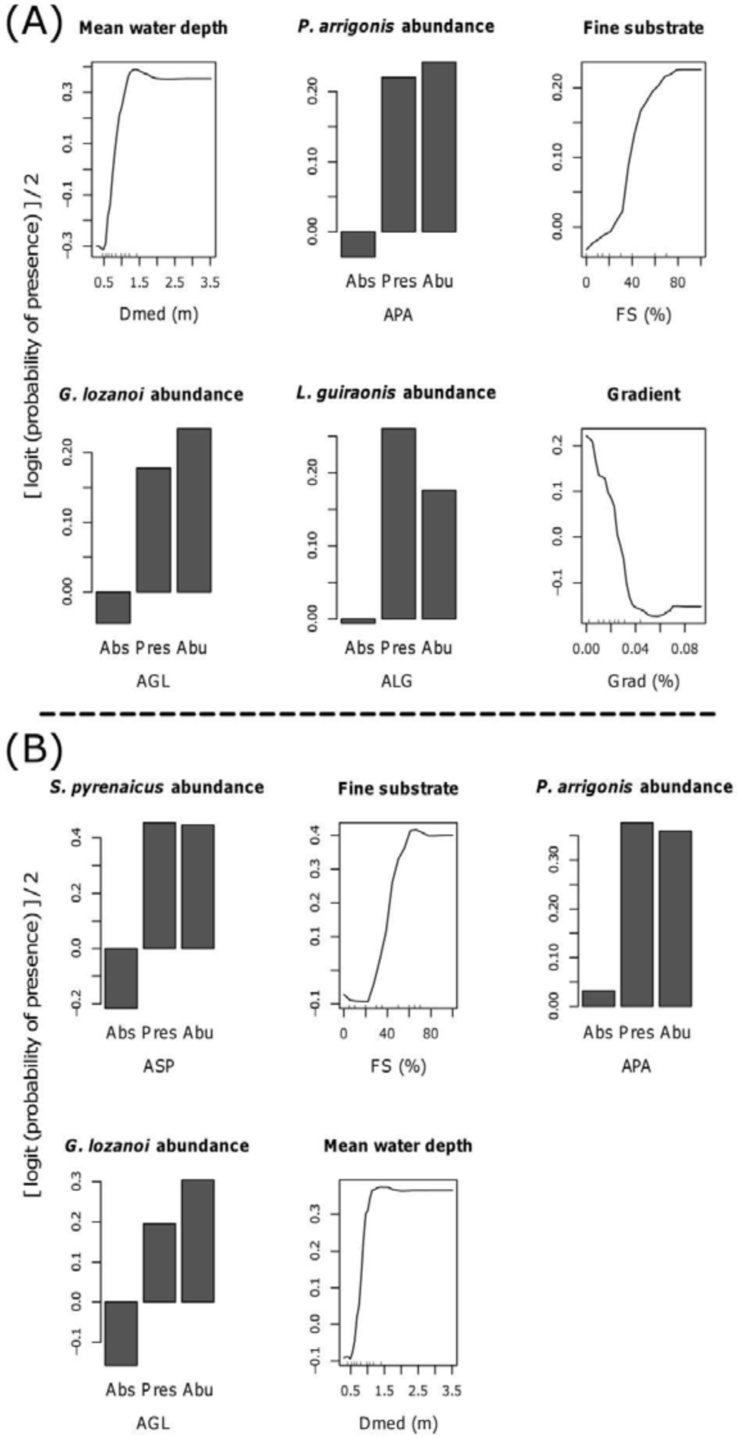

Figure 5. Partial plots of variable marginal effects in the RF models for (A) *S. pyrenaicus* and (B) *L. guiraonis*, considering both biotic and abiotic habitat variables. Fish abundance was expressed by three classes: Abs = absent, Pres = present and Abu = abundant. Selected variables are reported in order or importance.

As for the model built using both biotic and abiotic habitat variables, the same fish abundances were selected in the model built using only biotic variables, i.e., *P. Arrigonis*, *G. lozanoi* and *L. guiraonis* for *S. pyrenaicus*, and *P. Arrigonis*, *S. pyrenaicus* and *G. lozanoi* for *L.guiraonis* (Fig. 6).

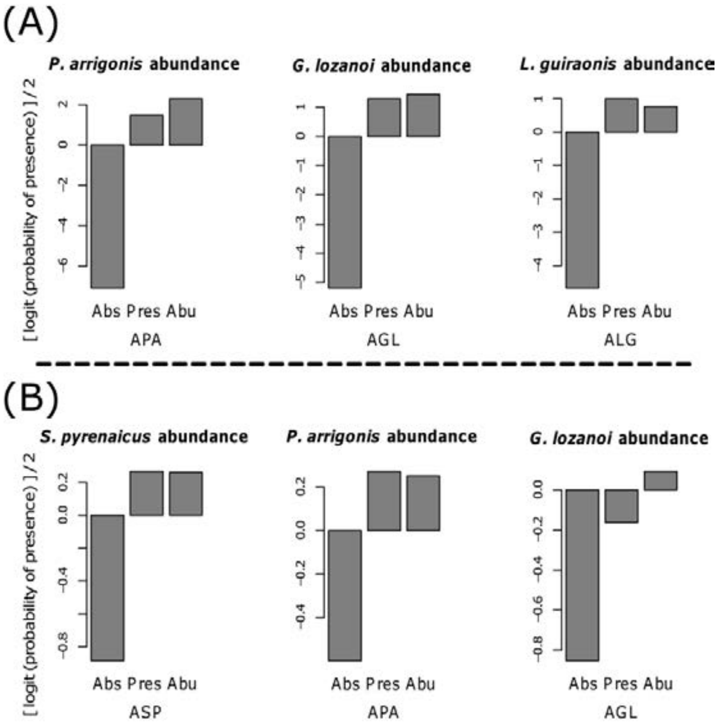

Figure 6. Partial plots of variable marginal effects in the RF models for (A) *S. pyrenaicus* and (B) *L. guiraonis*, considering only biotic habitat variables. Fish abundance was expressed by three classes: Abs = absent, Pres = present and Abu = abundant. Selected variables are reported in order or importance.

Due to the ecological relevance and the high model performance, model validation with an independent dataset was carried out only for the predictive models built by abiotic variables. For *S. pyrenaicus*, the model showed an accuracy of 75%, a Cohen's kappa of 0.51 and a TSS of 0.55, although being slightly over-predictive (sensitivity = 0.93 and specificity = 0.62). The *L. guiraonis* model performance was even higher, achieving an accuracy of 81%, whereas Cohen's kappa and TSS were equal to 0.60 (Fig. 6). Compared to model training, the area under ROC curve (AUC) decreased, showing a value of 0.75 for *S.pyrenaicus* model and 0.81 for *L. guiraonis* model.

Table 3: Validation of the habitat models for *S. pyrenaicus* and *L. Guiraonis* built using only abiotic variables. Model accuracy, specificity, sensitivity, Cohen's kappa, area under ROC curve and true skill statistic are reported in the table, together with the prevalence, number of observations and proportion of the training data for each validation dataset.

Model validation	***S. pyrenaicus***	***L. guiraonis***
Accuracy	0.73	0.81
Sensitivity	0.93	0.77
Specificity	0.62	0.83
Cohen's kappa	0.47	0.60
Area under ROC curve	0.75	0.81
True skill statistic	0.55	0.60
Prevalence	0.38	0.59
Number of obs.	48	22
Proportion of the calibration dataset (%)	20	20

4. Discussion

This study focused on the prediction of *S. pyrenaicus* and *L. guiraonis* distribution in the upper Cabriel River (Eastern Spain), taking into account the relative importance of both biotic and abiotic habitat variables. In particular, we evaluated the role of interspecific interactions to shape fish distribution, which constitute a valuable contribution for modelling and evaluating habitat for fish.

Random Forests (RF) was effective in predicting the probability of fish presence in response to habitat variables and the conditional variable importance (Strobl et al., 2008) provided a fair means of comparison that can help identify the truly relevant predictor variables. For the first time in species distribution modelling, the conditional variable importance was used together with the Model Improvement Ratio (MIR) technique (Murphy et al., 2010) and the procedure showed effectiveness in identifying a parsimonious set of not correlated variables, which minimize noise and improve model performance. Furthermore, the MIR procedure can be considered appropriate for parsimonious model construction as RF is noted to be robust to overfitting when the number of noise variables increases (Hastie et al., 2009). According to Freeman at al. (2012), we did not balance the species prevalence in model construction phases (e.g., re-sampling the data to have prevalence = 0.5), due to its negligible influence on RF results. All models showed high accuracy, sensitivity/specificity values and Cohen's kappa statistics indicating reliable predictions with low cross-classification errors. Moreover, the area under ROC curve (AUC) and the true skill statistic (TSS), which can also be considered independent of prevalence (Vaughan and Ormerod, 2005; Maggini et al., 2006), suggested good to excellent model performance (Pearce and Ferrier, 2000; Allouche et al., 2006).

The presence of *S. pyrenaicus* and *L. guiraonis* in pools and glides, but also in moderate to fast water habitats, such as riffles (Table 2), is in accordance with the classification of both fish as eurytopic species (Matono et al., 2006; Capela, 2007). *S. pyrenaicus* had been previously defined as lithophilic (Doadrio, 2001), as riffles with abundant gravel are important spawning sites for the fish (Granado-Lorencio, 1996; Ilhéu et al., 1999; Doadrio, 2001). This spawning behaviour is in accordance with the one described for *S. cephalus* (European chub), which selects shallow running waters as spawning sites (Fredrich et al., 2003). In our study, the preference shown by both fish species for pools, glides and riffles may depend on the selected survey period (June-October), in which the main drivers of the species distribution may be related to daily feeding and resting activities rather than spawning (Doadrio, 2001). Considering the diel and seasonal variation of habitat requirements (sensu Davey et al., 2011), one can state that the protection and enhancement of habitat diversity seems to be the best strategy to favour the conservation of these endemic Iberian species (Ilhéu et al., 1999; Magalhães et al., 2002).

Although the predictive models for the two target species were built using two different training datasets (Table 2), the selected biotic and abiotic inputs and their influence on the probability of presence were similar. This results may suggest that the fish distribution patterns are similar and the two species generally occupy similar habitats. Indeed, *S. pyrenaicus* and *L. guiraonis* were frequently observed in mixed species groups during the surveys. Therefore, the positive effect of cyprinid abundances on the probability of fish presence (Fig. 5 and 6) may not indicate positive interspecific interactions but only habitat overlapping. Only when *L. Guiraonis* was classified as abundant, the probability of presence of *S. pyrenaicus* slightly decreased, which can be indicative of possible competition between the two fish species in such a condition. The Iberian species of chub and barbel are considered generalist, mainly relying on invertebrates, detritus and plants in accordance to their relative availability (Granado-Lorencio, 1996; Valladolid and Przybylski, 1996; Carmona et al., 1999), although at the microhabitat scale, differences in the feeding habits can lead to the differential use of the water column (Grossman and De Sostoa, 1994). This resource partitioning can therefore explain the coexistence between species and the overlap in habitat use shown by the models (Martínez-Capel, 2000). Indeed, the analysis on the correlation between fish densities, and particularly the correlation between the two target species and other fish species (Fig. 7), revealed that cyprinid densities were positively correlated (Spearman's coefficient ranging from 0.28 to 0.77), hence emphasizing the habitat overlapping and the limited role of interspecific competition. However, it is important to state that competition can limit population size without completely excluding species from habitats (e.g. competitors do coexist), and further analysis of fish abundance may provide valuable additional information (Fukuda et al., 2012; Olaya-Marín et al., 2013).

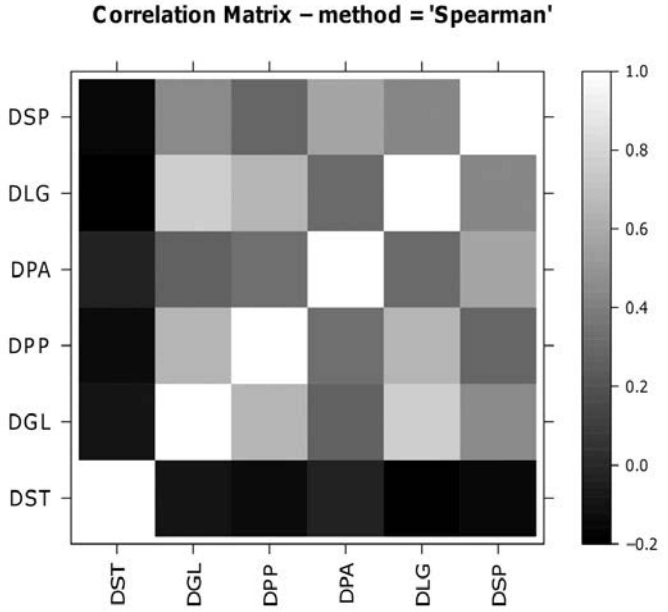

Figure 7. Spearman correlation coefficients among fish density values in the study area. Species codes are reported in Table 1.

It is important to note that all modelling approaches designed to account for biotic interactions have important limitations in inferring causation from spatial data. If the distribution of one species is shown to be highly dependent on the distribution of another species it can be difficult to differentiate if this is due to a real biotic interaction between the two species or is better explained by one or more overlooked environmental factors not accounted for in the model (Wisz et al., 2013). Building three different models, which account for (i) only abiotic (Fig. 4), (ii) both biotic and abiotic (Fig. 5), and (iii) only biotic variables (Fig. 6) can be seen a possible approach to gain insights on the role of the different drivers of species distribution. However, the proposed approach needs some prior knowledge on the ecology of the species under study to include the appropriate environmental predictors at the appropriate scale resolution, in order to avoid the risk of concluding that there is completion or mutualism when this is not the case (Wisz et al., 2013).

Looking at the selected abiotic variables (Fig. 4), the positive effect of water depth and channel width on cyprinids occurrence has been pointed out in Iberian rivers (Godinho et al., 1997; Carmona et al., 1999; Pires et al., 2000). Particularly, studies carried out at the micro-scale showed that both *Squalius* and *Barbus* prefer deep-water habitats (Grossman and De Sostoa, 1994; Martínez-Capel et al., 2009). However, contrary to Ferreira et al. (2007), the proportion of fine substrate (which is almost absent in the upper Cabriel River) was shown to be an important variable for *S. pyrenaicus* occurrence. The importance of submerged vegetation for fish has been demonstrated in a number of studies (Arlinghaus and Wolter, 2003; Oliva-Paterna et al., 2003; Santos et al., 2004; Gomes-Ferreira et al., 2005), and has been related to a combination of factors including physical stresses, food availability and predation risk (Ferreira et al., 2007). Clavero et al. (2004) also stressed its importance in well-conserved upper reaches of Iberian rivers as refuge for small-sized *S. pyrenaicus*. This fish can also respond negatively to pressure related to morphological alteration (CEMAGREF, 2008); and, according to our model, an increase of the proportion of silt and clay substrate may result in a decrease of *S. pyrenaicus* occurrence.

The performance in validation (Table 3) demonstrated the great efficacy and the ecological relevance of the selected abiotic variables in predicting fish distribution at the mesoscale, and this result is coherent with the previous ecological knowledge on habitat selection by Mediterranean cyprinids (Granado-Lorencio, 1996). It is important to highlight here that the use of independent data for validation is a not common procedure, often omitted in species distribution models (Elith and Leathwick, 2009). Current practice usually involves testing predictive performance using data resampling (e.g., split-sample or cross-validation procedures), and more experimental verification of modelled fish-habitat relationships is needed to provide valuable insights on model effectiveness and transferability (Bennett et al., 2013). Indeed, model generality should be tested on a spatially independent data-set since the use of accuracy estimates based on cross-validation procedures tend to differ (Edwards Jr et al., 2006). However, collecting new data is costly and needs to be optimized. Some work has attempted to identify the minimum sample requirements for deriving robust predictions at minimal costs, and have shown that different modelling methods might require different minimum sampling size (Stockwell and Peterson, 2002). Following Freeman et al. (2012) and Stockwell & Peterson (2002), we assumed that, for RF, 20% of the training data-set and more than 20 observations per species were suitable for model validation. Moreover, improving model parsimony was useful to identify the lowest number of variable to be surveyed, and this approach will help in the case of future model applications.

The mesoscale resolution and the potential of RF in considering categorical and continuous variables allowed us to gain an insight into the influence of both biotic and abiotic variables on fish habitat use and to test if fish habitat selection was mainly driven (or not) by instream physical characteristics. The presented approach substantially differs from the traditional, more common micro-scale analysis, which is less flexible in accounting for multiple species and biotic interactions (Parasiewicz et al. 2013). This study represents a step towards including interspecific interactions in mesohabitat simulation tools (e.g., MesoHABSIM, Parasiewicz, 2007, MesoCaSiMiR, Eisner et al., 2005) in order to clarify the role of biotic interactions more rigorously across different spatial scales (from HMU scale, to river segments, to entire catchment). As reported in Hirzel and Guisan (2002), collecting fine-grained observational data across large spatial extents, stratified to represent variation in environmental gradients, can be useful to better investigate the effect of biotic interactions on species distribution. Such cross-scale analyses could be performed for freshwater fish using mesoscale approaches by building regional fish distribution models that describe how biotic interactions influence species assemblages and the processes that shape them (Araújo and Rozenfeld, 2014). Regional habitat models built at the mesohabitat scale have been already proposed in Vezza et al. (2014a,b). Following the proposed modelling procedure, the incorporation of biotic variables as predictors of fish distribution could also be considered in future studies using habitat simulation models to design environmental flows and river restoration actions to allow a better understanding of complex impact sources on the habitat use by fish (Boavida et al., 2012).

Nevertheless, the proposed modeling procedure has the limitation of ignoring the importance of population dynamics which can generate time lags in the relationship between environmental conditions and species' abundances. Looking at the results, our findings may represent the “ecological snapshot” of the upper Cabriel River and more studies would be needed to clarify the structure of freshwater fish assemblages in the Mediterranean area. The Upper Cabriel River constitutes a natural (unimpacted) study area to develop reference habitat models, which can be useful for the management of local populations. A more regional approach would be needed to validate the obtained results across different catchments in the Jucar River Basin District to gain more insight on habitat requirements of the considered fish species. However, samples from different rivers in nearly natural conditions are difficult to collect given the high degree of hydro-morphological alteration of Mediterranean rivers (Belmar et al., 2013; Feio et al., 2013) and the sensitive state of the fish (Crivelli, 1996; Baillie et al., 2004).

Apart from their ecological relevance, the obtained predictive models are based on variables which can be objectively measured and can be very useful to support habitat simulation tools. RF can be seen as a promising tool for the ecological management of Mediterranean rivers and predictive models can be implemented in the context of hydraulic-habitat simulation systems (Vezza et al., 2014b). Species distribution models should include the effects of interspecific interactions (Elith and Leathwick, 2009) and many conservation actions could benefit from modelling approaches that include both abiotic and biotic habitat variables (Guisan and Thuiller, 2005). Perspectives for refining predictions of fish distribution by accounting for biotic interactions remain in the early stages of development (Wisz et al., 2013). This approach is considered as an interesting line of research and further studies in Mediterranean rivers have been already planned for the near future.

5. Acknowledgements

This research has been developed in the frame of the HolRiverMed project (Marie Curie Actions, Intra-European Fellowships) and the SCARCE project (Assessing and predicting effects on water quantity and quality in Iberian rivers caused by global change, Consolider-Ingenio 2010 CSD2009-00065). Data collection has been funded partly by the Spanish Government (Ministerio de Medio Ambiente y Rural y Marino, Confederación Hidrográfica del Júcar), and partly by the Spanish Ministry of Education and Science (POTECOL, CGL2007-66412). We thank Juan Diego Alacaraz-Hernandez, Matias Peredo-Parada and Aina Hernandez-Mascarell for their help with field work and suggestions on data analysis.

6. References

Addicott, J., Aho, J., Antolin, M., Padilla, D., Richardson, J. and Soluk, D., 1987. Ecological neighborhoods: scaling environmental patterns. Oikos, 49, 340-346.

Alcaraz-Hernández, J.D., Martínez-Capel, F., Peredo-Parada, M. and Hernández-Mascarell, A.B., 2011. Mesohabitat heterogeneity in four mediterranean streams of the Jucar river basin (Eastern Spain). Limnetica, 30 (2), 363-378.

Allouche, O., Tsoar, A. and Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology, 43 (6), 1223-1232.

Araújo, M.B. and Rozenfeld, A., 2014. The geographic scaling of biotic interactions. Ecography, 37 (5), 406-415.

Arlinghaus, R. and Wolter, C., 2003. Amplitude of ecological potential: chub *Leuciscus cephalus* (L.) spawning in an artificial lowland canal. Journal of Applied Ichthyology, 19 (1), 52-54.

Armstrong, J.D., Kemp, P.S., Kennedy, G.J.A, Ladle, M. and Milner, N.J., 2003. Habitat requirements of Atlantic salmon and brown trout in rivers and streams, Fisheries Research, pp. 143-170.

Baillie, J.E.M., Hilton-Taylor, C. and Stuart, S.N., 2004. IUCN Red List of Threatened Species. A Global Species Assessment. IUCN, Gland, Switzerland.

Bain, M.B., Finn, J.T. and Booke, H.E., 1985. A quantitative method for sampling riverine microhabitats by electrofishing, North American Journal of Fisheries Management, pp. 489-493.

Bain, M.B. and Knight, J.G., 1996. Classifying stream habitat using fish community analysis. In: C.H. Leclerc M, Valentin S, Boudreau A, Cote Z (Editor), Ecohydraulics 2000, 2nd

International Symposium on Habitat Hydraulics,. INRS-Eau, Quebec City, Canada, pp. 107–117.

Belmar, O., Bruno, D., Martínez-Capel, F., Barquín, J. and Velasco, J., 2013. Effects of flow regime alteration on fluvial habitats and riparian quality in a semiarid Mediterranean basin. Ecological Indicators, 30 (0), 52-64.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D. and Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling & Software, 40 (0), 1-20.

Bivand, R., 2012. Package 'spdep'. Available: http://cran.r-project.org/web/packages/spdep/index.html. Accessed 2012 December 21.

Boavida, I., Santos, J.M., Cortes, R., Pinheiro, A. and Ferreira, M.T., 2012. Benchmarking river habitat improvement. River Research and Applications, 28 (10), 1768-1779.

Bossard, M., Feranec, J. and Otahel, J., 2000. CORINE land cover technical guide – Addendum 2000 Technical report No 40

Bovee, K.D., 1982. A guide to stream habitat analysis using the instream flow incremental methodology. Instream Flow Information Paper 12. U.S. Fish and Wildlife Service, Fort Collins, Colorado, USA.

Breiman, L., Friedman, J.H., Olshen, R. and Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Breiman, L., 2001. Random Forest. Machine Learning, 45, 5-32.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. and Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Global Change Biology, 16 (4), 1145-1157.

Capela, J.R., 2007. Methodology for establishing the ecological potential of reservoirs in Spain, according to the directive 2000/60/EC. Faculdade de Ciências - Departamento de biologia animal. Universidade de Lisboa, Lisbon, Portugal.

Carmona, J., Doadrio, I., Márquez, A., Real, R., Hugueny, B. and Vargas, J., 1999. Distribution Patterns of Indigenous Freshwater Fishes in the Tagus River Basin, Spain. Environmental Biology of Fishes, 54 (4), 371-387.

CEMAGREF, 2008. EFI+ report. Improvement and spatial extension of the European Fish Index, Aix en Provence, France.

Cheng, L., Lek, S., Lek-Ang, S. and Li, Z., 2012. Predicting fish assemblages and diversity in shallow lakes in the Yangtze River basin. Limnologica, 42 (2), 127-136.

CHJ, 2007. Confederación Hidrográfica del Júcar. Estudio general sobre la Demarcación Hidrográfica del Júcar. CHJ, Madrid, Spain.

Clavero, M., Blanco-Garrido, F. and Prenda, J., 2004. Fish fauna in Iberian Mediterranean river basins: biodiversity, introduced species and damming impacts. Aquatic Conservation: Marine and Freshwater Ecosystems, 14 (6), 575-585.

Corbacho, C. and Sánchez, J.M., 2001. Patterns of species richness and introduced species in native freshwater fish faunas of a Mediterranean-type basin: the Guadiana River (southwest Iberian Peninsula). Regulated Rivers: Research & Management, 17 (6), 699-707.

Costa, R.M.S., Martínez-Capel, F., Muñoz-Mas, R., Alcaraz-Hernández, J.D. and Garófano-Gómez, V., 2012. Habitat suitability modelling at mesohabitat scale and effects of dam operation on the endangered júcar nase, *Parachondrostoma arrigonis* (River Cabriel, Spain). River Research and Applications, 28 (6), 740-752.

Crivelli, A.J., 1996. The freshwater fish endemic to the Mediterranean region. An action plan for their conservation. Tour du Valat Publication, 171 pp.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. and Lawler, J.J., 2007. Random forests for classification in ecology. Ecology, 88 (11), 2783-2792.

Davey, A.J.H., Booker, D.J. and D.J., K., 2011. Diel variation in stream fish habitat suitability criteria: implications for instream flow assessment. Aquatic Conservation: Marine and Freshwater Ecosystems, 21, 132-145.

Doadrio, I., 2001. Atlas y Libro Rojo de los Peces Continentales de Espana.

Doadrio, I., 2002. Origen y Evolución de la Ictiofauna Continental Española. En: Atlas y Libro Rojo de los Peces Continentales de España. CSIC y Ministerio del Medio Ambiente, Madrid.

Doadrio, I. and Carmona, J.A., 2006. Phylogenetic overview of the genus Squalius (Actinopterygii, Cyprinidae) in the Iberian Peninsula, with description of two new species. Cybium, 30 (3), 199-214.

Drew, C.A., Wiersma, Y. and Huettmann, F., 2011. Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications. Springer, New York, 328 pp.

Edwards Jr, T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L. and Moisen, G.G., 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecological Modelling, 199 (2), 132-141.

Elith, J. and Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics, 40 (1), 677-697.

Evans, J. and Cushman, S., 2009. Gradient modeling of conifer species using random forests. Landscape Ecology, 24 (5), 673-683.

Feio, M.J., Aguiar, F.C., Almeida, S.F.P., Ferreira, J., Ferreira, M.T., Elias, C., Serra, S.R.Q., Buffagni, A., Cambra, J., Chauvin, C., Delmas, F., Dörflinger, G., Erba, S., Flor, N., Ferréol, M., Germ, M., Mancini, L., Manolaki, P., Marcheggiani, S., Minciardi, M.R., Munné, A., Papastergiadou, E., Prat, N., Puccinelli, C., Rosebery, J., Sabater, S., Ciadamidaro, S., Tornés, E., Tziortzis, I., Urbanič, G. and Vieira, C., 2013. Least Disturbed Condition for European Mediterranean rivers. Science of the Total Environment, 476-477, 745-756.

Ferreira, M.T., Sousa, L., Santos, J.M., Reino, L., Oliveira, J., Almeida, P.R. and Cortes, R.V., 2007. Regional and local environmental correlates of native Iberian fish fauna. Ecology of Freshwater Fish, 16 (4), 504-514.

Fox, J., 2007. polycor: Polychoric and Polyserial Correlations. R package version 0.7-5, http://CRAN.R-project.org/package=polycor. Accessed 3 April 2010.

Franklin, J., 2010. Mapping species distributions: spatial inference and prediction. Cambridge University Press, New York, 338 pp.

Fredrich, F., Ohmann, S., Curio, B. and Kirschbaum, F., 2003. Spawning migrations of the chub in the River Spree, Germany. Journal of Fish Biology, 63 (3), 710-723.

Freeman, E.A., Moisen, G.G. and Frescino, T.S., 2012. Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in Random Forest models of tree species distributions in Nevada. Ecological Modelling, 233 (0), 1-10.

Fukuda, S., Mouton, A. and De Baets, B., 2012. Abundance versus presence/absence data for modelling fish habitat preference with a genetic Takagi–Sugeno fuzzy system. Environmental Monitoring and Assessment, 184 (10), 6159-6171.

García de Jalón, D., Torralva, M.M., Lurueña, J., Andreu, A., Martínez-Capel, F., Oliva-Paterna, F.J. and Alonso, C., 1999. Plan de Gestión Piscícola de la Región de Murcia.

Godinho, F.N., Ferreira, M.T. and Cortes, R.V., 1997. Composition and spatial organization of fish assemblages in the lower Guadiana basin, southern Iberia. Ecology of Freshwater Fish, 6 (3), 134-143.

Gomes-Ferreira, A., Ribeiro, F., Moreira da Costa, L., Cowx, I.G. and Collares-Pereira, M.J., 2005. Variability in diet and foraging behaviour between sexes and ploidy forms of the hybridogenetic *Squalius alburnoides* complex (Cyprinidae) in the Guadiana River basin, Portugal. Journal of Fish Biology, 66 (2), 454-467.

Granado-Lorencio, C., 1996. Ecología de Peces. Universidad de Sevilla, Sevilla.

Grenouillet, G., Buisson, L., Casajus, N. and Lek, S., 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography, 34 (1), 9-17.

Grossman, G.D. and De Sostoa, A., 1994. Microhabitat use by fish in the lower Rio Matarraña, Spain, 1984–1987. Ecology of Freshwater Fish, 3 (3), 123-136.

Guay, J.C., Boisclair, D., Rioux, D., Leclerc, M., Lapointe, M. and Legendre, P., 2000. Development and validation of numerical habitat models for juveniles of Atlantic salmon (*Salmo salar*), Canadian Journal of Fisheries and Aquatic Sciences, pp. 2065-2075.

Guisan, A. and Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8 (9), 993-1009.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. Springer Series in Statistics. Springer, New York, USA.

Hauer, C., Unfer, G., Tritthart, M., Formann, E. and Habersack, H., 2010. Variability of mesohabitat characteristics in riffle-pool reaches: testing an integrative evaluation concept (FGC) for MEM-application. River Research and Applications, doi: 10.1002/rra.1357.

He, Y., Wang, J., Lek-Ang, S. and Lek, S., 2010. Predicting assemblages and species richness of endemic fish in the upper Yangtze River. Science of the Total Environment, 408 (19), 4211-4220.

Hermoso, V., Clavero, M., Blanco-Garrido, F. and Prenda, J., 2010. Invasive species and habitat degradation in Iberian streams: an analysis of their role in freshwater fish diversity loss. Ecological Applications, 21 (1), 175-188.

Hermoso, V., Januchowski-Hartley, S., Linke, S. and Possingham, H.P., 2011. Reference vs. present-day condition: early planning decisions influence the achievement of conservation objectives. Aquatic Conservation: Marine and Freshwater Ecosystems, 21 (6), 500-509.

Hirzel, A. and Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling, 157 (2–3), 331-341.

Ilhéu, M., Costa, A.M. and Bernardo, J.M., 1999. Habitat use by fish species in a Mediterranean temporary river: the importance of riffles., Proceedings 3rd International Symposium on Ecohydraulics. Utah State University, Salt Lake City, pp. CD-ROM, 3 pp.

Joyce, M.P. and Hubert, W.A., 2003. Snorkeling as an alternative to depletion electrofishing for assessing cutthroat trout and brown trout in stream pools. Journal of Freshwater Ecology, 18 (2), 215-222.

Kemp, J.L., Harper, D.M. and Crosa, G.A., 1999. Use of 'functional habitats' to link ecology with morphology and hydrology in river rehabilitation. Aquatic Conservation: Marine and Freshwater Ecosystems, 9 (1), 159-178.

Kottelat, M. and Freyhof, J., 2007. Handbook of European freshwater fishes. Publications Kottelat, Cornol, Switzerland. 646 p.

Lamouroux, N. and Jowett, I.G., 2005. Generalized instream habitat models. Canadian Journal of Fisheries and Aquatic Sciences, 62 (1), 7-14.

Liaw, A. and Wiener, M., 2002. Classification and regression by Random Forest. R News, 2, 18-22.

Maceda-Veiga, A., 2012. Towards the conservation of freshwater fish: Iberian Rivers as an example of threats and management practices. Reviews in Fish Biology and Fisheries, 23 (1), 1-22.

Magalhães, M.F., Beja, P., Canas, C. and Collares-Pereira, M.J., 2002. Functional heterogeneity of dry-season fish refugia across a Mediterranean catchment: the role of habitat and predation. Freshwater Biology, 47 (10), 1919-1934.

Maggini, R., Lehmann, A., Zimmermann, N.E. and Guisan, A., 2006. Improving generalized regression analysis for the spatial prediction of forest communities. Journal of Biogeography, 33 (10), 1729-1749.

Markovic, D., Freyhof, J. and Wolter, C., 2012. Where Are All the Fish: Potential of Biogeographical Maps to Project Current and Future Distribution Patterns of Freshwater Species. PLoS ONE, 7 (7), e40530.

Martínez-Capel, F., García de Jalón, D., Werenitzky, D., Baeza, D. and Rodilla-Alamá, M., 2009. Microhabitat use by three endemic Iberian cyprinids in Mediterranean rivers (Tagus River Basin, Spain). Fisheries Management and Ecology, 16, 52–60.

Matono, P., Iihéu, M., Sousa, L., Bernardo, J.M., Formigo, N., Ferreira, M.T., de Almeida, P.R. and Cortes, R., 2006. Aplicação da directiva-quadro da água: tipos de rios portugueses com base na ictiofauna. In: A.P.d.R. Hídricos (Editor), VIII Congresso da Água. Associação Portuguesa dos Recursos Hídricos, Figueira da Foz, Portugal.

Mouton, A.M., De Baets, B. and Goethals, P.L.M., 2010. Ecological relevance of performance criteria for species distribution models. Ecological Modelling, 221 (16), 1995-2002.

Mouton, A.M., Alcaraz-Hernández, J.D., De Baets, B., Goethals, P.L.M. and Martínez-Capel, F., 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. Environmental Modelling & Software, 26 (5), 615-622.

Murphy, M.A., Evans, J.S. and Storfer, A., 2010. Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. Ecology, 91 (1), 252-261.

Olaya-Marín, E.J., Martinez-Capel, F. and Vezza, P., 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowledge and Management of Aquatic Ecosystems, 409 (07), 1-19.

Olden, J.D., Lawler, J.J. and Poff, N.L., 2008. Machine Learning Methods Without Tears: A Primer for Ecologists. The Quarterly Review of Biology, 83 (2), 171-193.

Oliva-Paterna, F., Miñnano, P. and Torralva, M., 2003. Habitat quality affects the condition of Barbus sclateri in Mediterranean semi-arid streams. Environmental Biology of Fishes, 67 (1), 13-22.

Parasiewicz, P., 2007. The MesoHABSIM model revisited. River Research and Applications, 23 (8), 893-903.

Parasiewicz, P., Rogers, J.N., Vezza, P., Gortazar, J., Seager, T., Pegg, M., Wiśniewolski, W. and Comoglio, C., 2013. Applications of the MesoHABSIM Simulation Model. In: H.A. Maddock I., Kemp P. and Wood P. (Editor), Ecohydraulics: an integrated approach. John Wiley & Sons Ltd, pp. 109-124

Pearce, J. and Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling, 133 (3), 225-245.

Pires, A.M., Cowx, I.G. and Coelho, M.M., 2000. Life history strategy of *Leuciscus pyrenaicus* (Cyprinidae) in intermittent streams of the Guadiana basin (Portugal). Cybium, 24 (3), 287-297.

Planque, B., Loots, C., Petitgas, P., LindstrøM, U.L.F. and Vaz, S., 2011. Understanding what controls the spatial distribution of fish populations using a multi-model approach. Fisheries Oceanography, 20 (1), 1-17.

Rosenfeld, J., 2003. Assessing the Habitat Requirements of Stream Fishes: An Overview and Evaluation of Different Approaches. Transactions of the American Fisheries Society, 132 (5), 953-968.

Santos, J.M., Godinho, F.N. and Ferreira, M.T., 2004. Microhabitat use by Iberian nase Chondrostoma polylepis and Iberian chub Squalius carolitertii in three small streams, north-west Portugal. Ecology of Freshwater Fish, 13 (3), 223-230.

Schill, D.J. and Griffith, J.S., 1984. Use of Underwater Observations to Estimate Cutthroat Trout Abundance in the Yellowstone River. North American Journal of Fisheries Management, 4 (4B), 479-487.

Siroky, D.S., 2009. Navigating Random Forests and related advances in algorithmic modeling. Stat Surveys, 3, 147-163.

Smith, K.G. and Darwall, W.R.T. (Editors), 2006. The status and distribution of freshwater fish endemic to the mediterranean basin. IUCN -The World Conservation Union, Gland, Switzerland/Cambridge, UK., 41 pp.

Stockwell, D.R.B. and Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling, 148 (1), 1-13.

Strayer, D.L. and Dudgeon, D., 2010. Freshwater biodiversity conservation: recent progress and future challenges. Journal of the North American Benthological Society, 29 (1), 344-358.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. and Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics, 9 (307).

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P., 2003. Random Forest:□ A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences, 43 (6), 1947-1958.

Teichert, M.A.K., Kvingedal, E., Forseth, T., Ugedal, O. and Finstad, A.G., 2010. Effects of discharge and local density on the growth of juvenile Atlantic salmon Salmo salar. Journal of Fish Biology, 76 (7), 1751-1769.

Valladolid, M. and Przybylski, M., 1996. Feeding relations among cyprinids in the Lozoya river (Madrid, Central Spain). Polskie Archiwum Hydrobiologii, 43, 213-223.

Vaughan, I.P. and Ormerod, S.J., 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology, 42 (4), 720-730.

Vezza, P., Comoglio, C., Rosso, M. and Viglione, A., 2010. Low Flows Regionalization in North-Western Italy. Water Resources Management, 24(14), 4049-4074.

Vezza, P., Parasiewicz, P., Calles, O., Spairani, M. and Comoglio, C., 2014a. Modelling habitat requirements of bullhead (*Cottus gobio*) in alpine streams. Aquatic Sciences, 76 (1), 1-15.

Vezza, P., Parasiewicz, P., Spairani, M. and Comoglio, C., 2014b. Habitat modelling in high gradient streams: the meso-scale approach and application. Ecological Applications, 24 (4), 844-861.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.-A., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. and Svenning, J.-C., 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological Reviews, 88 (1), 15-30.