Evaluating the performance of habitat models for predicting the environmental flow requirements of benthic macroinvertebrates

Christos Theodoropoulos¹,²,*, Aikaterini Vourka¹, Nikolaos Skoulikidis¹, Peter Rutschmann³, Anastasios Stamou²

¹ Hellenic Centre for Marine Research, Institute of Marine Biological Resources and Inland Waters, 46.7km Athens-Sounio ave., 19013, Anavyssos, Greece
² National Technical University of Athens, Department of Water Resources and Environmental Engineering, Iroon Polytechniou 5, 15780, Athens, Greece
³ Technical University of Munich, Chair of Hydraulic and Water Resources Engineering, Arcisstr. 21, D-80333 Munich, Germany

* Corresponding author:
Phone: +302291076335; E-mail: ctheodor@hcmr.gr

Abstract - Although various methods are currently available for modelling the habitat preferences of aquatic biota, studies comparing the performance of data-driven habitat models are limited. In this study, we assembled a benthic-macroinvertebrate microhabitat-preference dataset and used it to evaluate the predictive accuracy of regression-based univariate Habitat Suitability Curves (HSC), Boosted Regression Trees (BRT), Random Forests (RF), fuzzy-logic-based models using the weighted average (FLWA), maximum membership (FLMM), mean of maximum (FLM) and centroid (FLC) defuzzification algorithms and fuzzy rule-based Bayesian inference (FRB). The results show that the BRT model was the most accurate, closely followed by RF, FRB, FLM and FLMM while the FLC and FLWA algorithms had the lowest performance. However, due to the imbalanced nature of the dataset and in contrast to the fuzzy rule-based algorithms, the HSC, BRT and RF models failed to accurately predict the habitat suitability in low-scored microhabitats. We conclude that, given balanced datasets, BRT and RF can be effectively used in habitat suitability modelling. For imbalanced datasets, a properly adjusted RF model can be applied but when the input dataset is large enough to provide sufficient data-driven IF-THEN rules to train a FRB, FLMM or FLM algorithm, these models will produce the most accurate predictions.

1. Introduction

It was in 1924 that J. Grinnell first used the term ‘ecological niche’ to describe ‘the ultimate distributional unit, within which each species is held by its structural and instinctive limitations’ (Grinnell and Storer 1924). And it was G.E. Hutchinson (1957) who tried to quantify the niche, describing it as a multi-dimensional hypervolume,
where the different dimensions represent environmental conditions and resources, which define the requirements of an individual or a species to live and of its populations to persist and thrive. As the processing power and data storage capacity of modern computers increased over the last decades, this fundamental framework has been the springboard for operational hydroecological applications—termed as habitat suitability models—that are used to explore, interpret and predict the distribution of biota along gradients of hydrological or hydraulic conditions (Mouton et al. 2011; Papadaki et al. 2016; Leitner et al. 2017).

Habitat suitability models are usually incorporated as modules in hydrodynamic habitat models (HHMs), which (i) determine the water depth (D) and flow velocity (V) in a properly constructed computational grid representing the area under investigation (hydrodynamic module) and (ii) assess the habitat suitability of target biotic elements of the aquatic ecosystem (e.g. fish, benthic macroinvertebrates, macrophytes and riparian vegetation) for various discharge scenarios (Benjankar et al. 2014; Theodoropoulos et al. 2015), by integrating their habitat preferences with the predicted values of D and V (habitat module). Focusing either on the microhabitat—the unique combination of V, D, substrate type (S) and often temperature (T)—or the mesohabitat scale, which includes wider hydromorphologic units of species- and life-stage-dependent hydraulic configurations (Parasiewicz 2007; Schneider et al. 2014), their goal is often to derive ecosystem-based environmental flow scenarios downstream of anthropogenic pressures causing hydrological alteration of surface water bodies (Katopodis 2003; Bockelmann et al. 2004; Muñoz-Mas et al. 2016).

Despite the continuous efforts to increase the predictive accuracy of HHMs, there has been much controversy and criticism of the methods applied in their habitat modules to quantify the relationship between the environmental predictors and the biological response variables (Ahmadi-Nedushan et al. 2006; Hirzel and Le Lay 2008; Jowett et al. 2008). The traditional approach is the one described by Bovee (1986), which requires the development of three types of univariate habitat suitability curves (HSC) based either on expert knowledge (category I—professional judgment curves), or on actual field data (categories II and III—habitat utilization and habitat preference curves) to quantify the relationship between the abiotic variables (V, D, S and/or T) and the biotic response variables (usually the abundance of organisms). Based on this rather simplistic approach and in combination with the criticism received, complex mathematical methods evolved, including three basic types:

1. Multivariate statistical models, usually regression-oriented, which attempt to fit a curve between the abiotic predictors (environmental and hydraulic data) and the biotic response variables, based on specific assumptions about the distribution of data. Over time, these models have been enhanced to include various alternatives of the fundamental curve-fitting concept to avoid the assumptions of linearity between the predictor and response variables (Generalized Linear Models (GLMs) - McCullagh and Nelder 1989) or to account for the cumulative influence of the predictors on the response variables (Generalized Additive Models (GAMs) - Hastie and Tibshirani 1990).

2. Machine learning models; a series of non-parametric algorithms, which combine regression or classification functions based on the characteristics of available data. These algorithms are able to deal with complex relationships and interactions between the environmental variables; they can handle large amounts of data with possible non-linear relationships between the predictors and are able to process complex and noisy data (Recknagel 2001). The most widely used examples include Classification and Regression Trees (Breiman et al. 1984; Dakou et al. 2007), Artificial Neural Networks (Broomhead and Lowe 1988; Tirelli and Pessani 2009), Random Forests (Breiman 2001; Vezza et al. 2015; Booker 2016) and Boosted Regression Trees (Elith et al. 2008; Theodoropoulos et al. 2017).

3. Fuzzy-rule based models (Zadeh 1965), often enhanced by Bayesian Belief Networks (BBNs - Pearl 1988); based on sets of IF-THEN rules, fuzzy models convert the actual values of the predictor and
response variables to membership functions ranging from 0 to 1 and calculate the habitat suitability by applying logical operators (AND/OR). These methods can handle the inherent vagueness of the input and output data, as well as the possible interaction between the predictor variables (Ahmadi-Nedushan et al. 2006) and have been applied in habitat modelling, in combination with BBNs to enhance predictive accuracy (van Broekhoven et al. 2006; Liu et al. 2013; Lange et al. 2015).

The available literature indicates that most of the research on HHMs has been carried out using fish fauna as the target biotic element of the aquatic ecosystem. In contrast, benthic-macroinvertebrate (BM) benchmark datasets and relevant comparative modelling studies are disproportionately limited (Poff and Zimmerman 2010; Shearer et al. 2015; Theodoropoulos et al. 2015).

The purpose of this article is to evaluate the predictive performance of widely applied habitat modelling algorithms based on a reference hydroecological dataset consisting of microhabitat-defining predictor variables (V, D, S and T) and BM-related response variables, assembled from an extended sampling network in Greek streams and rivers. We were particularly interested in answering the following question: ‘Do the specific distributional properties of this and similar hydroecological datasets - i.e. (i) imbalanced nature, (ii) complex interactions between the microhabitat-defining variables, (iii) high variation of the response BM variables- require the application of specific habitat modelling methods to avoid producing misleading results and enhance the predictive accuracy of HHMs?’ In addition to answering this question, working solutions are proposed that address the difficulty of working with highly variable hydroecological data in model-based environmental flow assessments.

2. Materials and methods

2.1. Study area and site selection

The study area was intentionally selected to include natural landscapes with no or very minor anthropogenic activities, with the purpose of eliminating bias originating from water quality degradation and focus on the influence of the hydraulic properties of the water on the communities. Data were collected from pluvial perennial streams and rivers in central and southern Greece (Fig. 1; Table 1), belonging to the RM-1 (10-100 km²; altitudes between 200-800 m.a.s.l.; mixed geology), RM-2 (100-1000 km²; altitude <600 m.a.s.l.; mixed geology) and RM-4 (10-1000 km²; altitudes between 400-1500 m.a.s.l.; mixed geology) European intercalibration types (van de Bund et al. 2004). The study area is characterized by semi-mountainous to mountainous relief with altitudes ranging from 116 to 972 m.a.s.l., large areas with evergreen forests, often mixed with deciduous forests of Quercus sp., Cornus sp., Pinus sp. and Fagus sp. The riparian vegetation is composed of thick forests including plane trees, willows, poplars, alders and ash trees.

The climatic conditions are typical of temperate Mediterranean areas, with hot, dry summers and cold, wet winters. Temperatures range between 0°C (or lower) and 30°C (or higher during extreme events). The greatest amount of precipitation falls between October and April, with the driest months being July, August and September. Maximum and minimum precipitation was recorded from nearby meteorological stations for the northern parts of the study area (sites 1, 2 and 3) in January (395.7 mm) and August (48.5 mm). For the sites of central Greece (4 and 5) the amount of precipitation was 158.2 mm in January and 24 mm in August, while in the southern parts of the area (sites 6, 7, 8 and 9), the relevant amount was 211.4 mm in March and 2.8 mm in July, respectively.

According to the REFCOND guidelines for reference site selection (WFD CIS Guidance Document No. 10 2003), nine sites were selected in the study area (Fig. 1; Table 1).
Fig. 1. Study area
Table 1: Geomorphological properties of the study sites. T: Tributary, M: Main stem

<table>
<thead>
<tr>
<th>Site No.</th>
<th>River</th>
<th>Site coordinates</th>
<th>Altitude (m)</th>
<th>Strahler order</th>
<th>Distance to source (km)</th>
<th>Catchment area (m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Acheloos (T)</td>
<td>39°31'24.16&quot;N 21°26'34.88&quot;E</td>
<td>972</td>
<td>2</td>
<td>6.68</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>Acheloos (T)</td>
<td>39°25'42.01&quot;N 21°14'51.68&quot;E</td>
<td>563</td>
<td>1</td>
<td>2.81</td>
<td>56</td>
</tr>
<tr>
<td>3</td>
<td>Acheloos (T)</td>
<td>39°16'23.11&quot;N 21°31'39.99&quot;E</td>
<td>659</td>
<td>1</td>
<td>2.64</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>Evinos (M)</td>
<td>38°43'20.39&quot;N 22°01'08.20&quot;E</td>
<td>774</td>
<td>2</td>
<td>5.16</td>
<td>86</td>
</tr>
<tr>
<td>5</td>
<td>Mornos (M)</td>
<td>38°35'53.43&quot;N 22°11'18.01&quot;E</td>
<td>407</td>
<td>3</td>
<td>15.46</td>
<td>222</td>
</tr>
<tr>
<td>6</td>
<td>Krathis (M)</td>
<td>38°00'55.29&quot;N 22°15'44.10&quot;E</td>
<td>911</td>
<td>1</td>
<td>7.32</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>Alfeios (T)</td>
<td>37°41'36.95&quot;N 21°57'57.87&quot;E</td>
<td>403</td>
<td>1</td>
<td>7.86</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>Parapeiros (M)</td>
<td>37°59'47.93&quot;N 21°46'00.27&quot;E</td>
<td>291</td>
<td>2</td>
<td>6.71</td>
<td>39</td>
</tr>
<tr>
<td>9</td>
<td>Parapeiros (M)</td>
<td>38°03'07.38&quot;N 21°42'55.96&quot;E</td>
<td>116</td>
<td>2</td>
<td>15.09</td>
<td>108</td>
</tr>
</tbody>
</table>

2.2. Physicochemical data
At each site, measurements of water temperature, electrical conductivity, dissolved oxygen, and total dissolved solids were made using the Aquaread AP-2000 Multiparameter Meter. Water samples were collected, stored in plastic bottles (250 mL) at 4°C, transferred to the lab and analyzed for major ions (NO$_3^-$, NO$_2^-$, NH$_4^+$ and PO$_4^{3-}$) using the Merck Nova 60 Spectroquant Photometer, with the results (not shown) indicating that all sites conformed to the REFCOND guidelines (NO$_3^-$ < 0.5 mg/L; NO$_2^-$ < 0.3 mg/L; NH$_4^+$ < 0.01 mg/L; PO$_4^{3-}$ < 0.15 mg/L).

2.3. Hydraulic-habitat and biological data
At each site, immediately after the physicochemical measurements, following a modified AQEM approach (AQEM Consortium 2002), benthic macroinvertebrates were sampled from a maximum of 20 rectangular microhabitats delineated as combinations of V, D and S. A 0.25 x 0.25 m Surber sampler with a mesh size of 500 μm was used, resulting in a total sampling area of 0.0625 m² at each microhabitat. Each BM sample was separately preserved in plastic bottles containing 70% ethanol. At each microhabitat, flow velocity (V), water depth (D) and the type of substrate (S) were afterwards recorded. Flow velocity was measured using the Swoffer 2100 current velocity meter at 0.6 x D (measuring from the surface) when D ≤ 0.75 m and by averaging 0.2 x D and 0.8 x D when D > 0.75 m according to Nolan and Shields (2000). Water depth was measured using a water-depth measurement rod and substrate was visually assessed using the categories defined by Schneider et al. (2010).

After taking into account safety considerations and avoiding sampling during and after extreme rainfall events, a benchmark dataset of 380 microhabitats was collected over three sampling periods (spring, summer and autumn 2015). Detailed characteristics of each microhabitat are shown in Table S1.

2.4. Data analysis
In the lab, all specimens from each BM sample were sorted and identified at the family level using a stereo microscope (magnification 6.39 - 639), and macroinvertebrate identification guides for the Mediterranean region (Tachet et al. 2010). Abundance of macroinvertebrates and BM metrics, including taxonomic richness,
diversity (Shannon’s index) and EPT richness (Ephemeroptera, Plecoptera, Trichoptera) were calculated using the ASTERICS 3.1.1 software (Wageningen Software Labs 2004).

2.4.1. Habitat suitability
The (micro) habitat suitability was calculated based on BM metrics commonly applied to assess the quality-suitability in relevant studies (Englund and Malmqvist 1996; Monk et al. 2006; Waddle and Holmquist 2011; Holmquist et al. 2015). The following equation was used:

$$\kappa = 0.4 \times \frac{n_{ij}}{n_{jmax}} + 0.3 \times \frac{H_{ij}}{H_{jmax}} + 0.2 \times \frac{EPT_{ij}}{EPT_{jmax}} + 0.1 \times \frac{a_{ij}}{a_{jmax}}$$

where,

- $\kappa$ is the habitat suitability of the $i^{th}$ microhabitat of the $j^{th}$ site, ranging from 0 to 1
- $n_{ij}$ denotes the number of BM taxa (families) found at the $i^{th}$ microhabitat of the $j^{th}$ site
- $H_{ij}$ denotes the Shannon’s diversity index for the $i^{th}$ microhabitat of the $j^{th}$ site
- $EPT_{ij}$ is the number of EPT taxa found at the $i^{th}$ microhabitat of the $j^{th}$ site
- $a_{ij}$ is the abundance of benthic macroinvertebrates found at the $i^{th}$ microhabitat of the $j^{th}$ site
- $n_{jmax}$, $H_{jmax}$, $EPT_{jmax}$ and $a_{jmax}$ denote the maximum value of the relevant variables observed at the $j^{th}$ site.

It must be noted that the equation used, ‘relaxes’ any seasonal or site-specific variation in the BM communities due to the particular hydrological and climatological features of each site, since the values of each BM metric are normalized per site and per season to a range between 0 (unsuitable) and 1 (suitable) by dividing by the maximum observed value of the particular metric at each site for each sampling period.

2.4.2. Hydroecological relationships
To examine the possibility of using the whole dataset to train and cross validate the selected models/methods, microhabitat samples were grouped according to their typological (RM-1, RM-2 and RM-4) and geographical (altitude, geographical location) properties. The Non-metric Multidimensional Scaling (NMDS) procedure, implemented in the PRIMER 6 statistical software with the Bray-Curtis distance measure was used in order to identify structural similarities between the BM communities. BM data (abundance) were square root transformed prior to the analysis to approximate normality (Quinn and Keough 2002; Leps and Smilauer 2003).

Hydroecological relationships between the predicted habitat suitability ($\kappa$) scores and the V, D, S and T hydraulic and environmental predictors were examined using five selected methods: (i) univariate habitat suitability curves (Bovee 1986), (ii) Boosted Regression Trees (Elith et al. 2008) and (iii) Random Forests (Breiman 2001), implemented in R version 3.1.0. (R Core Team 2014), (iv) the Mamdani-Assilian fuzzy inference process with various defuzzification algorithms (Zadeh 1965; Mamdani and Assilian 1975; Ross 2010) and (v) a fuzzy Bayesian inference process based on Brookes et al. (2010), both implemented in FORTRAN, using the HABFUZZ software (Theodoropoulos et al. 2016).

2.4.3. Univariate Habitat Suitability Curves (HSC)
Based on the approach of Bovee (1986), habitat utilization curves (category II) were developed to relate the observed habitat suitability ($\kappa$) with the predictor variables. The range of values of the continuous predictors was divided into classes, with intervals of 0.1 m/s for V, 0.1 m for D and 1 °C for T. For each predictor, the average $\kappa$ of each class was calculated and visualized using histograms. A polynomial regression function (2$^{nd}$ to 4$^{th}$ order) was fitted to each ‘predictor-$\kappa$’ plot-, adjusting it to achieve $R^2$ values higher than 0.6 and $p$ values.
lower than 0.05. The categorical predictor (S) was already divided into classes and the average $\kappa$ for each S class was calculated. Respective polynomial models were used to predict $K$ for each value of the predictor variable (V, D, S and T), thus deriving four different $K$ values (one for each predictor, except for S since it was directly related with the average $K$), which were then integrated using the following equation:

$$K = \sqrt[4]{K_V \times K_D \times K_S \times K_T}$$

where,

$K$ is the predicted habitat suitability

$K_V$, $K_D$, $K_S$ and $K_T$ denote the habitat suitability values derived from the polynomial equation of each predictor variable.

2.4.4. Boosted Regression Trees (BRT)

BRT is a powerful machine-learning model, which combines two techniques; (i) regression trees, using recursive binary splits to adjust the response to predictor variables and (ii) gradient boosting, a method, which combines several, moderately accurate models to produce a very accurate predictive model. The BRT has been proven to efficiently explain and predict the response of aquatic communities to environmental variability (Leathwick et al. 2006; Leclere et al. 2011; Waite et al. 2014; Pilière et al. 2014). In this study we used it as one of the best available regression-based option because BRT produces the most reliable best-fit curve among the relevant methods (Elith et al. 2008; Theodoropoulos et al. 2017). The BRT model was fitted using the gbm.step function of the ‘dismo’ package v1.1-4 (Hijmans et al. 2016) in R statistical software version 3.1.0 (R Core Team 2014). Bag fraction, learning rate and tree complexity were adjusted to develop a model with the lowest cross-validation error. The values of the final model were 0.5, 0.005 and 2 respectively, achieving at least 1000 trees, following the rule of thumb suggested by Elith et al. (2008). The BRT model was developed using untransformed data, according to Elith et al. (2008).

2.4.5. Random Forests (RF)

RF is an ensemble learning model that generates a large number of decision trees, which are then aggregated to compute the final classification. RF applies the bootstrapping technique (resampling with replacement) and increases the tree diversity by randomly changing the sets of predictor variables over the different tree induction processes. Each classification tree is grown using another bootstrap subset $X_i$ of the original data set $X$ (where $i$ denotes the index of the bootstrap iteration, ranging from 1 to the maximum number of trees $k$) and the nodes are split using the best split predictor variable among a subset of $m$ randomly selected predictors (Liaw and Wiener 2002). In detail, the algorithm for growing a RF of $k$ classification trees goes as follows (Peters et al. 2007; Vezza et al. 2015):

1. A bootstrap subset $X_i$ containing approximately 2/3 of the elements of the original data set $X$ is selected. The elements not included in the training dataset are referred to as out-of-bag (OOB) data for that bootstrap sample.

2. $X_i$ is used to grow an unpruned classification tree to the maximum depth but, rather than choosing the best split among all predictive variables, $m$ predictor variables are randomly selected and the best split is chosen among these variables.

3. Each tree is fully grown and used to predict OOB observations. Prediction is applied according to the majority vote of the ensemble of $k$ trees (the predictions from all the trees are combined to predict an observation class as well as a probabilistic prediction output for that observation). As OOB
observations are not used in the fitting of the trees, this procedure is similar to the cross-validation process, which prevents overfitting.

In our study, the RF model was built using the randomForestSRC 2.4.2 package (Ishwaran and Kogalur 2017) and was further explored using the ggRandomForests 2.0.1 (Ehrlinger 2016) and ggplot2 2.2.1 (Wickham and Chang 2016) R packages. The parameters in our model were optimized in order to minimize the overall error rate. Specifically, the number of trees was set to 3500 and the number of randomly selected variables to split the nodes was set to 2. Furthermore, since the $\kappa$ classes were not equally represented in the benchmark dataset, resulting in poor predictive accuracy for the minority class, we applied (i) a weighted RF (W-RF) to increase the probability of the rarely observed $\kappa$-class samples to be selected during bootstrapping and (ii) a RF where the majority classes were down-sampled (D-RF) so that all classes were represented equally in each tree. The class-weight and down-sampling balancing were implemented using the classwt and the sampsize arguments, respectively, in randomForest 4.6-12 (Liaw and Wiener 2002) R package.

2.4.6. Fuzzy logic and fuzzy rule-based Bayesian inference

Fuzzy rule-based models use linguistic descriptions, such as ‘low’, ‘moderate’ and ‘high’ to convert the fixed, numerical values of the predictor and response variables (called crisp sets) to overlapping membership functions (fuzzy sets), usually trapezoidal or triangular-shaped, characterized by four parameters ($a_1$, $a_2$, $a_3$, $a_4$) (van Broekhoven et al. 2006). By this procedure, called fuzzification, each numerical value of the predictor and the response variables is assigned to one or more membership functions with a specific membership degree ranging from zero to one. In trapezoidal-shaped membership functions, the membership degree linearly increases from 0 to 1 for values between $a_1$ and $a_2$, remains constant for values between $a_2$ and $a_3$ and linearly decreases from 1 to 0 for values between $a_3$ and $a_4$. In triangular-shaped membership functions, $a_2$ is equal to $a_3$. The membership functions of the predictor variables (in our case V, D, S and T) are afterwards related with the habitat suitability (note that $\kappa$ is calculated based on BM community metrics and not the abundance of particular BM taxa) by a set of IF-THEN rules, either data-driven or expert-knowledge-based, such as ‘IF V is low AND D is moderate AND S is gravel AND T is moderate THEN $\kappa$ is high’. The ‘IF’ part is called ‘the antecedent’ and the ‘THEN’ part is called ‘the consequent’.

In typical fuzzy logic applications, the degree of fulfilment of each IF-THEN rule is afterwards calculated as the minimum of the membership degrees in its antecedent. Finally, to each linguistic habitat suitability value ($\kappa$) a fulfillment degree is also assigned, equal to the maximum of the fulfilment degrees of all rules with the output value under consideration in their consequent (van Broekhoven et al. 2006) and the final predicted $\kappa$ is calculated within a defuzzification process, in which the $\kappa$ fuzzy values are converted into a single crisp numerical value. The algorithms used for defuzzification were (i) centroid (FLC), (ii) weighted average (FLWA), (iii) maximum membership (FLMM) and (iv) mean of maximum (FLM) (Ross 2010).

We additionally applied a fuzzy rule-based Bayesian inference algorithm (FRB), which deviates from the typical approach (Brookes et al. 2010); the fuzzy membership degree of each predictor variable is considered as the probability of occurrence of the particular fuzzy set. The IF-THEN rules are then combined using the Bayesian joint probability, so that (referring to the previous example) the probability of a microhabitat’s $\kappa$ being high is the joint probability that V is low AND D is moderate AND S is gravel AND T is moderate. Since $\kappa$ is also a linguistic output, a score is assigned to each habitat suitability class. Given for example two membership functions $\{X_1, X_2, \ldots X_i\}$ and $\{Y_1, Y_2, \ldots Y_j\}$ for the predictor variables $X$ and $Y$, respectively, there are $M_{ij} = X_i \ast Y_j$ cross memberships and the final predicted $\kappa$ is derived as:

$$
K = \sum M_{ij} S_{ij}
$$
where,
\(K\) is the predicted habitat suitability
\(M_{ij}\) denotes the joint probability of occurrence of each \(K\) class (cross memberships)
\(S_{ij}\) denotes the score of each \(K\) class

In our study we used five-class, trapezoidal-shaped membership functions for the fuzzification of \(V\), \(D\) and \(T\) (Table 2). Habitat suitability (both the observed \(\kappa\) and the predicted \(K\)) was treated as a crisp input/output and classified using a 3-class scheme, as the minimum acceptable output scheme, and a 5-class scheme following the status classification system of the Water Framework Directive 2000/60/EC (WFD - European Commission 2000). The substrate type was also treated as a crisp input, keeping the classes of Schneider et al. (2010) and numbering them appropriately (Table 2) to be included in the HABFUZZ software.

Table 2: Linguistic values of the abiotic (predictor) and biotic (response) variables. \(V\): flow velocity; \(D\): water depth, \(S\): substrate, \(T\): temperature, \(K\): habitat suitability. See Schneider et al. (2010) for substrate-type size. Fuzzy set parameters follow the \(\{a_1, a_2, a_3, a_4\}\) order as explained in the text.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Linguistic value</th>
<th>Fuzzy set parameters</th>
<th>Variable</th>
<th>Linguistic value</th>
<th>Crisp set parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(V) (m/s)</td>
<td>Very low</td>
<td>0, 0, 0.05, 0.1</td>
<td>(K) (3-class)</td>
<td>Low</td>
<td>0, 0.2</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>(0.05, 0.1, 0.15, 0.2)</td>
<td>Moderate</td>
<td>(0.2, 0.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>(0.15, 0.2, 0.4, 0.5)</td>
<td>High</td>
<td>(0.6, 1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>(0.4, 0.5, 0.7, 0.8)</td>
<td>Very high</td>
<td>(0.7, 0.8, 0.8, 0.8)</td>
<td></td>
</tr>
<tr>
<td>(D) (m)</td>
<td>Very shallow</td>
<td>0, 0, 0.1, 0.15</td>
<td>(K) (5-class)</td>
<td>Bad</td>
<td>0, 0.2</td>
</tr>
<tr>
<td></td>
<td>Shallow</td>
<td>(0.15, 0.2, 0.3, 0.35)</td>
<td>Moderate</td>
<td>(0.2, 0.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>(0.3, 0.35, 0.55, 0.6)</td>
<td>Good</td>
<td>(0.4, 0.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deep</td>
<td>(0.55, 0.6, 0.7, 0.75)</td>
<td>High</td>
<td>(0.6, 0.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Very deep</td>
<td>(0.75, 0.8, 0.8, 0.8)</td>
<td>(S)</td>
<td>Boulders (BO)</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Large stones (LS)</td>
<td>0.050</td>
</tr>
<tr>
<td>(T) (°C)</td>
<td>Very low</td>
<td>0, 0, 9, 10</td>
<td></td>
<td>Small stones (SS)</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>(9, 10, 13, 15)</td>
<td></td>
<td>Large gravel (LG)</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>(13, 15, 17, 19)</td>
<td></td>
<td>Medium gravel (MG)</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>(19, 20, 23, 25)</td>
<td></td>
<td>Fine gravel (FG)</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>Very high</td>
<td>(25, 27, 27, 27)</td>
<td></td>
<td>Sand (SA)</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Silt (SI)</td>
<td>0.020</td>
</tr>
</tbody>
</table>

2.5. Evaluation of the models’ performance and comparison
Ten times repeated ten-fold cross validation was applied to evaluate the predictive accuracy of each algorithm (Kohavi 1995). The initial dataset was randomly partitioned in ten equal-sized subsamples. Nine subsamples (342 microhabitats) were used as the training dataset and the remaining subsample (38 microhabitats) was used for model validation. This process was repeated ten times (folds), using a different subsample for validation at each iteration. The ten-fold cross validation was repeated ten times to acquire more robust results. The performance of each model was evaluated as the average percentage of the correctly classified instances (CCI) between each iteration of the ten-fold cross-validation process.
3. Results and Discussion

3.1. The benchmark dataset

A total of 26,758 individuals were isolated from the 380 sampled microhabitats and 70 benthic invertebrate families were identified (see Table S1 for details). The allocation of samples to the three-dimensional space according to the NMDS procedure showed structural community differences based on the typological and geographical grouping applied, mainly for sites 6, 8 and 9 (Figure 2). The microhabitats of these sites were concentrated at specific parts of the 3D plot and were partially isolated from the rest of the samples, indicating increased similarity between them and consequently, increased dissimilarity with the microhabitats of the other sites. This trend suggests that typological variation (site 8: RM-1, site 9: RM-2, site 6: RM-4) results in biological variation, as was initially expected.

Fig. 2. Allocation of microhabitats to the NMDS 3D space. (Top: Grouping based on river type, Bottom: Grouping based on geographical location). The number of each site is also depicted.
While in abundance-based habitat models, the aforementioned dissimilarities may suggest the exclusion of these microhabitats from model training, our working solution links the predictor variables (V, D, S and T) directly to the habitat suitability ($\kappa$) using BM metrics instead of the microhabitat’s macroinvertebrate abundance. This enables the application of the whole dataset to predict $K$ in sites with varying typological properties, as long as their specific microhabitat characteristics resemble the ones provided in the training dataset. In addition, since the response variable $\kappa$ is normalized per site and per season (and consequently per river type), typological, geographical, or even seasonal differences can be eliminated (Jowett et al. 2008), as each sample is compared with the samples of the same site.

With the inclusion of all microhabitat samples in the training dataset, the relationship between the predictor variables and the BM-metrics-based $\kappa$ is depicted in Figure 3. These are the typical distributions of such datasets (Fukuda et al. 2013), and the variation observed reflects the complex interactive relationships between the abiotic predictors (Jowett et al. 1991; Leclerc et al. 2003), which obviously cannot be effectively illustrated in the 2-dimensional space. The ideal visualization scheme to illustrate such complex interactions would require a 5D scatter plot (plotting V, D, S, T and $\kappa$ simultaneously). However, such a scheme, although scientifically meaningful, would be visually too complex to assimilate and susceptible to common pitfalls related with high dimensionality (Ronan et al. 2016). As a consequence, we used the 2D scatter plots only to enable an initial visualization of the raw dataset and facilitate further analysis.

![Fig. 3. Two-dimensional scatter plots between the predictor and response variables. V: flow velocity, D: water depth, T: temperature, S: substrate type, $\kappa$: observed habitat suitability. For S classes see Table 2.](image-url)
3.2. Application of the selected habitat models

According to the habitat suitability curves developed by both the HSC and BRT models (Figure 4), $\kappa$ values were highest in flow velocities between approximately 0.25 and 0.65 m/s, with the optimal $\kappa$ values being recorded near 0.6 m/s. Water depths less than 0.3 m showed increased $\kappa$, peaking near 0.2 m and decreasing rapidly to minimum in D values greater than 0.4 m. Finer substrates had (in most cases) lower habitat suitability, with the highest $\kappa$ values being recorded in boulders and large stones. Regarding $T$, the lowest $\kappa$ values were recorded in temperatures between 12 and 16°C, increasing in $T$ values less than 12°C and greater than 16°C. However, temperature had the weakest relationship with $\kappa$ among the four predictor variables.

Similar BM habitat preferences have been indicated in the majority of previous studies from various geographic locations with varying climatic and hydrodynamic conditions. With minor exceptions for certain macroinvertebrate taxa (Jowett et al. 1991), flow velocities higher than 0.7 m/s are not considered suitable for most macroinvertebrates (Gore et al. 2001; Li et al. 2009; Shearer et al. 2015) and, in agreement with our findings, the optimal $\kappa$ has been recorded in $V$ values between 0.1 m/s and 0.6 m/s. Regarding $D$, habitat suitability in low-order streams and rivers usually peaks at 0.25 m (Jowett et al. 1991; Gore et al. 2001; Li et al. 2009), ranging between 0.1 m and 1 m. Coarser substrates are generally considered more suitable for benthic macroinvertebrates but with high variation among studies. Our results also suggest that large stones and boulders were more suitable (except for a unique ‘silt’ sample with high $\kappa$ values).

![Fig. 4. BRT-fitted K in response to the values of the environmental/hydraulic predictors. Y axes are on the logit scale and centered to have zero mean over the data distribution. Marks at the inside bottom of the plots indicate data range and deciles of site distribution across each variable. The solid line is the BRT-modelled response curve. The dashed line (red) is the LOESS smoother (span 0.25) fitted to the response curve. S abbreviations are explained in Table 2.](image)
The probability of each $K$-class in response to the values of the environmental and hydraulic predictors, based on the RF model application, is illustrated in Figure 5. Probability ($p$) values between the high (H) and moderate (M) $K$ classes showed inverse trends for $V$, $D$ and $T$. The $p$ value of H increased above 0.5 in $V$ values between 0.1 and 0.7 m/s, in contrast with the $p$ value of the M class, which decreased below 0.5 in the same $V$ range. Regarding $D$, $p$ values higher than 0.5 were recorded for $D$ values less than 0.3 m, contrasting to the $p$ values of the M class. $T$ values higher than $15^\circ C$ resulted in increasing H class $p$ ($>0.5$) and decreasing M class $p$ ($<0.5$). With much deviation from the mean values, H-class $p$ increased in boulders, large and small stones and large gravel. The M-class $p$ increased in fine gravel, medium gravel and sand, while the L-class $p$ was higher in sand.

**Fig. 5.** Probability of $K$ in response to the values of the environmental/hydraulic predictors, based on the RF model. $V$: flow velocity, $D$: water depth, $S$: substrate type, $T$: temperature, H: high $K$ class, M: moderate $K$ class, L: low $K$ class. For S-classes see Table 2.

It has to be noted that due to the small representation of the low $k$ class in the benchmark dataset, its probability of occurrence was very low for all the predictor variables, increasing only to near 0.1 for $V$ values less than 0.2 m/s. This ‘weakness’ was observed not only in the RF algorithm but also in HSC and BRT and is a
common pitfall in imbalanced datasets as many algorithms are constructed to minimize the overall error rate, which often results in poor accuracy for the minority class (Sun et al. 2007; Sun et al. 2009; Ali et al. 2015). However, in our study adjusting the RF model in favor of the ‘rare’ class, either by class-weight or down-sampling, resulted in a decrease of the major classes’ classification accuracy and of the model’s overall predictive accuracy. Consequently, the selection of the ‘best’ model is highly subjective and mostly depends on the purpose of the model’s application. For example, in WFD-oriented habitat simulations, the user’s interest is mainly focused on accurately predicting and discriminating between the M and G classes (finding the M-G boundary), since in the WFD’s classification system only the ‘good’ and ‘high’ K classes are considered acceptable (European Commission 2000).

3.4. Testing the predictive performance of the models

Based on the overall performance metrics calculated during the 10-fold cross validation process (Table 3), and regarding the 3-class K scheme, the FLC and FLWA models had the lowest performance (50.13% and 49.14% respectively). The remaining fuzzy-logic-based models showed acceptable predictive accuracy (>50% CCI) but were outperformed by the HSC, RF and FRB models, which showed 61.38%, 61.94% and 61.2% accuracy, respectively, while the BRT model had the highest overall performance (67.92%), reaching a maximum of 81.58% during cross-validation. As previously mentioned, in contrast to the fuzzy rule-based algorithms, the HSC, BRT and RF models failed to accurately predict the low class K (all models had an accuracy lower than 10% for this class). Adjusting the RF model to account for the low k microhabitats increased the predictive accuracy of the low K class from 0% to 50% and 55% for the W-RF and the D-RF, respectively, but decreased the accuracy of the H class by 10% (W-RF) and 20% (D-RF), while the M class predictions showed varying shifts (+10% for W-RF and -10% for D-RF). The overall predictive accuracy decreased from 61.85% to 57% and 54% for the W-RF and D-RF, respectively.

Table 3: Model performance, evaluated as the average percentage of Correctly Classified Instances (%CCI) during the cross validation process.

<table>
<thead>
<tr>
<th>Model</th>
<th>3-class K %CCI</th>
<th>5-class K %CCI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>HSC</td>
<td>61.38</td>
<td>55.26</td>
</tr>
<tr>
<td>BRT</td>
<td>67.92</td>
<td>60.52</td>
</tr>
<tr>
<td>RF</td>
<td>61.85</td>
<td>-</td>
</tr>
<tr>
<td>W-RF</td>
<td>57.23</td>
<td>-</td>
</tr>
<tr>
<td>D-RF</td>
<td>54.38</td>
<td>-</td>
</tr>
<tr>
<td>FLC</td>
<td>50.13</td>
<td>46.84</td>
</tr>
<tr>
<td>FLWA</td>
<td>49.74</td>
<td>46.84</td>
</tr>
<tr>
<td>FLMM</td>
<td>61.16</td>
<td>58.16</td>
</tr>
<tr>
<td>FLM</td>
<td>60.68</td>
<td>55.79</td>
</tr>
<tr>
<td>FRB</td>
<td>61.20</td>
<td>58.95</td>
</tr>
</tbody>
</table>

These results suggest the use of the FRB model for reliable applications of habitat suitability modelling in imbalanced datasets, since it had the highest overall predictive accuracy and also performed well in predicting the low K-class samples (71% accurate), without any further adjustment (Figure 6). It must be noted however, that as the detail of the final prediction increased (by applying the 5-class K scheme) the predictive performance of all models, including the FRB, decreased dramatically to unacceptable prediction levels, ranging from 29.63% to 38.66% (again the BRT, RF and FRB models were the most accurate).

Previous studies show varying results regarding the predictive accuracy of habitat models, obviously due to the different methodological approaches applied and the variation of the input datasets. A 5-class K
scheme has rarely been previously adopted. As partially discussed in Lange et al. (2015) and our results confirm, the application of habitat suitability modelling within a 5-class $K$ system (according to the WFD classification) is prohibitive due to the low (unacceptable) predictive accuracy of the models. Moreover, existing habitat models relate habitat suitability directly to the abundance of particular aquatic taxa (either fish or benthic macroinvertebrates) or are based on presence-absence schemes. Within this context, 4-class $K$, BM abundance-based fuzzy models (van Broekhoven et al. 2006; Mouton et al. 2009) showed %CCI values between 50% and 66%. Fish-based presence-absence fuzzy models (Muñoz-Mas et al. 2016) calculated %CCI values between 45% and 48%. In addition, previous fish-based RF model applications (Mouton et al. 2011; Vezza et al. 2015) confirm the increased accuracy of the specific method, calculating %CCI values greater than 70%.

5. Conclusions
The results of the study suggest that BM-based data-driven habitat modelling requires the application of specific algorithms to avoid producing misleading results and enhance the predictive accuracy of the output.
habitat suitability, depending on the distributional properties of the input hydroecological dataset. To answer our initial question, it was observed that:

- The complex interactions between the microhabitat-defining predictors and the high variation of the response BM variables did not significantly influence overall model performances. All models including the univariate HSC algorithm (which does not account for interactions between predictors) showed almost equal overall performance.

- In contrast, the imbalanced nature of the dataset, reduced the predictive accuracy of the HSC, BRT and RF algorithms for the low-K class to unacceptable levels. This suggests that the imbalanced nature of a dataset should be the main factor to consider when selecting the most suitable algorithm.

- Random Forests and Boosted Regression Trees can be effectively applied to provide reliable habitat suitability predictions. The predictive accuracy of these models was 61.85% and 67.92% respectively, reflecting the increased predictive capacity of these algorithms. However, both models failed to efficiently predict the microhabitats with low habitat suitability, due to the disproportionately limited number of such microhabitats in the benchmark dataset. The use of W-RF and D-RF overcame this problem but at the cost of decreasing the overall predictive accuracy.

- The FRB, FLM and FLMM models also had high overall predictive accuracy. In addition, due to the intrinsic properties of these algorithms, which, using a set of data-driven IF-THEN rules, process and predict the habitat suitability of each microhabitat one-by-one, these models developed reliable predictions for all classes of habitat suitability.

- It must be noted however that all models failed to efficiently predict the microhabitat suitability when a 5-class scheme was applied, suggesting that as the detail of the final prediction increases, the application of the selected models becomes insufficient. As partially discussed in van Broekhoven et al. (2006) and in agreement with our ecological viewpoint, while a 3-class-based, accurate predictive model is considered sufficient for ecological applications (e.g. for model-based environmental flow assessments), other modelling options could further increase the predictive performance of habitat models.

We finally conclude that Boosted Regression Trees and Random Forests can be effectively used in habitat modelling, given balanced datasets. However, the fuzzy rule-based algorithms should be preferred when modelling imbalanced datasets. In addition, when the input dataset is large enough to provide sufficient data-driven IF-THEN rules to ‘feed’ a fuzzy-rule-based algorithm (FRB, FLM, FLMM), these models are likely to produce the most reliable habitat suitability predictions. The application of other machine-learning algorithms should be investigated because the predictive accuracy of Random Forests and Boosted Regression Trees indicates the high potential of such models for habitat modelling applications.

Acknowledgements
This research is part of a PhD project conducted in the Hellenic Centre for Marine Research (HCMR) under the supervision of the National Technical University of Athens (NTUA) and in close collaboration with the Technical University of Munich (TUM). The authors would like to thank the Institute of Marine Biological Resources and Inland Waters of the HCMR for providing the field sampling equipment and the NTUA and TUM for communicating the hydraulic experience and expertise to complete this project.
References


