**Original Article**

**Modelling potential distribution of fluvial fish species for expanding conservation knowledge: Case study of the genus *Barbus* in Iran**

**Hossein Mostafavi**<sup>1</sup>, Jafar Kambouzia<sup>2</sup>

<sup>1</sup>Department of Biodiversity and Ecosystem Management, Environmental Sciences Research Institute, Shahid Beheshti University, GC, Tehran, Iran.  
<sup>2</sup>Department of Agroecology, Environmental Sciences Research Institute, Shahid Beheshti University, GC, Tehran, Iran.

**Abstract:** Species inhabiting fresh waters are severely influenced by anthropogenic factors. Effective management and conservation plans require high accurate and reliable species distribution forecasts. Here, we modelled potential distribution of the genus *Barbus* in Iran, based on environmental variables using Species Distribution Models (SDMs). Six environmental predictors (i.e. slope, bankfull width, elevation, mean air temperature, range of air temperature and annual precipitation) were applied for modelling. The models were selected among different technique (GLM, GAM, CTA, SRE, GBM, RF, MARS, and FDA) which their results were summarized through ensemble forecasting approaches. According to the TSS (True Skill Statistic), the accuracy of the implemented models was greater than 0.8. The results showed that the projected distributions not only were observed in the same recorded basins but also in the new basins. Presented results deepen the conservation knowledge in Iran and act as a guidance for management decisions aimed at legal identification of critical habitats for species as well as informing them for translocation of threatened or captive-bred populations.

**Introduction**

Nowadays, there is a growing concern about the long-term viability of fish species due to human pressures especially in freshwater ecosystems. Human activities such as excessive withdrawal, water pollution, alterations of river flow, habitat degradation, alien species, etc. can negatively affect the distribution of many species in their habitats (Dudgeon et al., 2006; Mostafavi et al., 2015). Freshwater ecosystems, especially rivers are one of the most important habitats of the world. Although they account for only 0.01 percent of global water resources and just cover small portion of the planet’s surface, but they support 10 percent of all known wildlife species and one-third of all vertebrate species e.g. about 40 percent of the world's fish species (McAllister et al., 1997; Pringle, 2003). Therefore, conservation biologists and natural resource managers have been investigating for ways to protect these ecosystems and their species.

For proper conservation and management of fish species in these ecosystems, we need to apply new methods for protecting them and provide practical and reliable information for managers and researchers. Nowadays, with the advancement of science, especially computer science, the use of modeling methods along with other methods has provided many opportunities for them to accomplish better and more effective environmental decision making. In this regard, modelling the distribution of species has become an important component of conservation programs (Thuiller et al., 2005; Morid et al., 2016; Amiri et al., 2017) and one of the most important environmental modelling methods is Species Distribution Models (SDMs). Forecasting species’ distributions has become a significant part of conservation planning in the last decade, and there is a wide range of modeling techniques for this purpose (Guisan and Thuiller, 2005). The spatial distribution of environments that are suitable for the species can then be estimated across a study region. This approach
is valuable for generating geospatial information that can be applied across a broad range of fields, including conservation biology, ecology, and evolutionary biology. SDMs usually use the relationship between environmental variables and species records to identify the environmental conditions in which populations can be present. It means the spatial distribution of the suitable environment for species is modelled and then the distribution of species is estimated for the whole region (Pearson et al., 2007). Many countries in Europe, the United States, Australia and some Asian countries have made extensive studies in this regard, however few studies have been done in Iran in aquatic ecosystems (Mostafavi et al., 2014).

Iran has high biological diversity, especially in freshwater fish species (Esmaeili et al., 2018) but freshwater ecosystems in Iran have been affected by a variety of human activities, such as damming, industrial development, urbanization, agriculture, etc., causing more destruction of these habitats that has resulted in serious threats to fish biodiversity (Mostafavi et al., 2014). Previous studies have been mostly carried out on identifying biodiversity of fishes in freshwater ecosystems (Tabatabaei et al., 2015; Zamani Faradonbe and Eagderi 2015; Ghasemi et al., 2015; Hoghoghi et al., 2016; Morid et al., 2016; Zamani Faradonbe et al., 2017), but due to various reasons such as lack of facilities, budget and lack of access to all rivers, only the main rivers have been investigated and in many rivers the presence and absence of many species has not been thoroughly studied.

In addition, in Iran for the introduction of species into new environments, the potential of the regions has not been properly studied. According to Guisan et al. (2013), SDMs are a tool for differentiating habitat quality at a range-wide scale, therefore, based on our results, critical habitats were typically defined as habitats necessary for the persistence, or long-term recovery, of sensitive/threatened species. Moreover, according to these results, SDMs can potentially inform the translocation decision process. Indeed, if translocation is deemed necessary, SDMs can identify potential recipient sites, which may be climate refugia within the current range, or sites that are projected to become newly suitable. Or SDMs can be used to identify which local species may be at risk of impact from the introduction of a translocated species through predicted overlapping distributions, in the same way as they are used to identify conflict areas between native and invasive species. In addition, SDMs are able to identify sites, where species may have been absent due to habitat degradation. Consequently, based on more detailed future studies effective conservation and restoration measures can be undertaken to maintain and (re)establish species populations. This can be useful for conservation decision because the habitats of some species were already degraded due to human pressures (Mostafavi et al., 2015).

The widespread Barbus populations from the southern tributaries of the Caspian Sea, Lake Namak and Urmia Lake basins, and the Euphrates and Tigris drainages previously identified as Barbus lacerta (Nikmehr et al., 2016; Khaefi et al., 2017a; Esmaeili et al., 2018). Naseka and Bogutskaya (2009) recognize B. cyri as a valid species from the Caspian Sea basin. Khaefi et al. (2017b) revalidated B. miliaris, a nominal species found in the Lake Namak basin. Furthermore, Khaefi et al. (2017a) described B. karunensis from the Karun River drainage in Iran, increasing the number of Barbus species in Iran to four. The objective of this study are (1) identification of potential distribution member of the genus Barbus over the extent of Iran, (2) determining the importance of environmental variables on the distribution of members of this genus and (3) developing recommendations that can aid in the conservation of riverine species in Iran.

**Materials and Methods**

**Study area, fish data and environmental predictors:**
This study was conducted in Iran which encompasses 19 basins (Coad, 2019). Member of the genus Barbus widely distributed in Iran and adjacent countries. They are found in the Caspian Sea, Urmia Lake, Namak Lake and Tigris basins in Iran (Mostafavi et al., 2014; Khaefi et al., 2017a). Data used in this study covers
several time periods (1970-2000) obtained from databases originating from the previous field samplings, several museums and literature (Berg, 1949; Saadati, 1977). Sites where fish were occurred/observed are called actual data. Moreover, sites which had an unclear position to the river network or were outside the temporal period between 1970 and 2000, as well stocked with species population and located in any lakes and wetlands were excluded. The primary database contained approximately 1700 sites which were reduced to 1090 sites after a detailed quality check concerning the reliability of the biological and spatial information.

After collecting the fish data, we considered three environmental variables at local scale (elevation (ELE), stream slope (SLO), and bank-full width (B_WID)), and five variables at the regional scale (maximum air temperature (Max_TEM), minimum air temperature (Min_TEM), mean air temperature (A_TEM), the range of air temperature (R_TEM) and annual precipitation (PRE)). We extracted ELE and B_WID from Google Earth (Google Inc. 2009, Version 5). B_WID was maximum width the stream attains, typically marked by changes in vegetation, topography, or texture of sediment. SLO was calculated for a 1 km stretch extending upstream of each site. Climate variables were extracted from WorldClim data (Hijmans et al., 2005, 2007) to characterize annual climate trends based on records for 30 years of monthly means (1970 to 2000), and interpolated at 30 arc-seconds grid extent (approximately 1 km at the Equator). Climate variables were extracted in a circular buffer (5 km) around each sampling site (Mostafavi et al., 2014) as a catchment layer similar to CCM2 (Catchment Characterization and Modelling database) (Vogt et al., 2003, 2007; De Jager and Vogt, 2010) is not available for Iran. Variable redundancy within environmental variables was tested by Spearman's rank correlation (r). If two variables were highly correlated (r>|0.75|) (Filipe et al., 2013; Mostafavi et al., 2014); one of them was excluded to avoid co-linearity according to our expert judgement and literatures.

**Modelling process (techniques, calibration, evaluation and ensemble forecasting):** In this study, the BIOMOD2 (BIOdiversity MODelling) package (Thuiller, 2003) was used within the R software (R Development Core Team, 2011) for modelling. As above mentioned, our fish data is based on a heterogeneous data set containing information from several sources, therefore the “presence-background modelling” approach was used according to Chefaoui and Lobo (2008) and Barbet-Massin et al. (2012). Then, the following nine modelling techniques were applied: Generalised Linear Models (GLM), Generalized Additive Model (GAM), Classification Tree Analysis (CTA), Artificial Neural Network (ANN), Surface Range Envelops (SRE), Generalized Boosting Method (GBM), Random forest (RF), Multivariate Adaptive Regression Splines (MARS), and Flexible Discriminant Analysis (FDA). We applied a cross-validation procedure by randomly splitting the data into calibration (80% of the data) and validating (20%) data sets with 10 repetition runs to assess model performance stability (Mostafavi et al., 2014). Model evaluation was based on the True Skill Statistic (TSS) which corresponds to the sum of sensitivity and specificity minus 1, and is independent of prevalence (Lobo et al., 2008; Thuiller et al., 2009a, b). Finally, all nine modelling techniques were combined in an ensemble-forecasting framework as recommended by Araújo and New (2007). Variable importance was calculated by a permutation procedure used in BIOMOD2, which is independent of the modelling technique (Thuiller et al., 2009a, b). We used the software ArcGIS Desktop 10.3 (ESRI© 1999-2008) to map the spatial pattern of the predicted distributions of the studied fish species within the 19 drainage basins of Iran.

**Results**

After correlation test, seven environmental variables (B_WID, SLO, ELE, A_TEM, R_TEM, and PRE) out of eight, remained as independent variables for the modelling. Table 1 describe their characteristics. Overall, TSS values ranged from 0.52 to 0.99 for testing validity (Fig. 1). In addition, the TSS was good in average (>|0.78). Moreover, whilst SRE showed the
lowest accuracy and RF was highest (Table 2).

The relative importance of environmental variables for each model is showed in Table 2. Among all variables, the most important ones are mean air temperature, annual precipitation and elevation, respectively (≥15%), whereas other variables show the low values (<15%). Moreover, the relative importance of predictors in each model was a little bit different to each other.

According to Figure 2, *Barbus* species were predicted not only in the same recorded basins (i.e. Caspian Sea, Urmia Lake, Namak Lake and Tigris basins) but also predicted in the new basins (i.e. Hari River basin and Upper Tigris River drainage). Moreover, in the same recorded basins (i.e. Caspian Sea, Urmia Lake, Namak Lake and Tigris basins) some new sites were identified which was not recorded in the literatures.

**Discussions**

The results of this study have important implications for conservation activities and management. The modelling framework has the ability to highlight the potential areas of species occurrence, and as well to identify sites, where species may have been absent due to habitat degradation.

**Model performance:** The results showed that the performance of the models was acceptable. RF was performed the best while SRE was performed the weakest. SRE or BIOCLIM (Bioclimatic Envelope

<table>
<thead>
<tr>
<th>Number of sites</th>
<th>B_WID (m)</th>
<th>SLO (%)</th>
<th>ELE (m)</th>
<th>A_TEM (°C)</th>
<th>R_TEM (°C)</th>
<th>PRE (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1090</td>
<td>Mean</td>
<td>112.5</td>
<td>1.6</td>
<td>731</td>
<td>19.1</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>1.0-3539.8</td>
<td>0.0-28.0</td>
<td>(-)27-2708</td>
<td>5.5-29.5</td>
<td>6.9-18.5</td>
</tr>
</tbody>
</table>

**Table 1.** Mean and range (minimum- maximum) of environmental variables.

**Table 2.** The order of relative importance of environmental predictors for genus *Barbus* fish.

<table>
<thead>
<tr>
<th>Environmental predictors</th>
<th>Average</th>
<th>SRE</th>
<th>CTA</th>
<th>RF</th>
<th>MARS</th>
<th>FDA</th>
<th>GLM</th>
<th>GAM</th>
<th>GBM</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_WID</td>
<td>0.08</td>
<td>0.15</td>
<td>0.10</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.16</td>
<td>0.01</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>SLO</td>
<td>0.04</td>
<td>0.10</td>
<td>0.14</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>ELE</td>
<td>0.15</td>
<td>0.12</td>
<td>0.12</td>
<td>0.05</td>
<td>0.20</td>
<td>0.34</td>
<td>0.18</td>
<td>0.04</td>
<td>0.03</td>
<td>0.24</td>
</tr>
<tr>
<td>R_TEM</td>
<td>0.07</td>
<td>0.13</td>
<td>0.06</td>
<td>0.13</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.07</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>A_TEM</td>
<td>0.44</td>
<td>0.45</td>
<td>0.53</td>
<td>0.29</td>
<td>0.47</td>
<td>0.55</td>
<td>0.63</td>
<td>0.35</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>PRE</td>
<td>0.38</td>
<td>0.30</td>
<td>0.43</td>
<td>0.24</td>
<td>0.51</td>
<td>0.36</td>
<td>0.27</td>
<td>0.31</td>
<td>0.32</td>
<td>0.67</td>
</tr>
</tbody>
</table>

**Abbreviations:** bankfull width (B_WID), stream slope (SLO), mean air temperature (A_TEM), annual precipitation (PRE), elevation (ELE), range of air temperature (R_TEM).

**Figure 1.** Evaluation of model accuracy according to TSS validation index.
Model) algorithm uses a rectangular envelope that enclose all (or a specified percentile) of the species records in the environmental space (Guisan et al., 2017). It is known as one the simplest modelling methods that only uses the presence records to model species environmental niche. SRE assumes no interaction between the environmental variables and has a simple response shape which is not capable of capturing all the complexity in species-environment relationship (Merow et al., 2014). Based on the number of selected environmental variables, SRE might lead to over or under fitting (Guisan et al., 2017). A comprehensive comparison study indicated the SRE to have a low predictive power (Elith et al., 2006). On the contrary, Random Forest (RF) has high transferability and predictive performance (Yates et al., 2018). RF is a machine learning algorithm built based on an ensemble of classification trees. It uses bagging or bootstrap aggregation technique to reduce the variance in classification trees and by selecting only a subset of variables on each node avoids overfitting (Hastie et al., 2009). Having the classification trees as its base learner; RF is capable of including high order interaction in model fitting that makes it powerful to draw more realistic species-environment relationship (Merow et al., 2014). These features make RF a robust predictive model.

Moreover, it is essential to be indicated that all applied models in this study come from different families ranging from statistical to machine learning and tree based models. Each of them have their own specific limitations such as ignoring interaction among the predictors (e.g, GAM and GLM) or assuming specific distributions, whereas the others can handle missing values, are insensitive to outliers and can handle large datasets. On the other hand, some models such as RF are quite flexible while a model like GLM is less flexible. The results of GAM and MARS are interpretable, nevertheless, models such as ANN belong to a black box category in which understanding the structure of the model is not straightforward (Valavi et al., 2018). Therefore, all nine modelling techniques were combined in an ensemble-forecasting framework as recommended by (Araújo and New, 2007) to reduce the uncertainty.

**Evaluation of selected variables:** The most important part of species distribution modelling is to select and use appropriate environmental variables. These variables can be used to describe geographical and biological conditions of the species and predict correct distribution range of the species. It is not possible to identify all variables and factors affecting species niche characteristics in modelling. This might be due to the deficiencies of ecological knowledge in studied species, the lack of knowledge for recognizing all dimensions of the species niche, availability limitations and the use of all variables in the form of information layers in modelling (due to the lack of information layers or their inaccuracies). Temperature appears to be one of the main determinant factors of spatial distribution for stream fishes (e.g. Buisson et al., 2008; Logez et al., 2012). Freshwater fishes are ectothermic (cold-blooded) animals and particularly sensitive to temperature. Fluctuation of temperature
effects the metabolism, breeding, development and growth of fishes (Mann, 1996). In addition, mean air temperature has been widely reported as an important variable affecting fish distribution (e.g. Pont et al., 2005; Buisson et al., 2008; Abdoli and Naderi, 2009; Mostafavi et al., 2014, 2015), which is in line with the results of this study. About the slope variable, it generally depends to the ecology of fishes which in this study it was not so much important for the species studied while in some studies (e.g. Filipe et al., 2013; Mostafavi et al., 2014) it was important for brown trout (*Salmo trutta*).

**Comparing the modelling results with available reports:** Comparing results of the studied species complex distribution modelling with existing reports (e.g. Mostafavi et al., 2014; Khaefi et al., 2017a) shows that the members of *Barbus* species was predicted not only in Caspian Sea, Urmia, Namak and Tigris basins, but also predicted in Hari River basin and Upper Tigris River drainage. Moreover, in the same recored basins, some new sites were identified. Two reasons may/can be stated for this difference. First, sampling methods conducted by researchers might be incomprehensive because almost every year new species are described from remote and mountainous regions of Iran (Khaefi et al., 2017a; Esmaeili et al., 2018). Second, these differences can also be attributed to the limiting factors, which influence distribution of freshwater fish species. Although we can consider the distribution of a species based on the relationship between the environmental conditions and the species inherent characteristics in terms of its potential distribution, but the various natural and artificial factors have constrained the realized distribution of species (Sexton et al., 2009). Since the stream network is the only corridor for distribution of fish species, the main limiting natural factor in freshwater ecosystems is the catchment border, by which movement of species between catchment would be restricted (Schmutz et al., 2000; Pont et al., 2005). Whilst, the habitat suitability for a species may be beyond its main catchment area, distribution of species within the catchment area may be confronted with various obstacles, especially artificial constructions like dams, resulting in limited distribution of a species within the catchment area. Species distribution can also be impacted by biological interactions in which the importance of some characteristics such as competition and predation is remarkable (Araújo and New, 2007). In addition, the absence of species in the new predictions for the same basin might be related to different sampling and monitoring methods conducted in diverse habitats and rivers in Iran as well as due to human pressures.

**Conclusion**

SDMs are important conservation and management tools because the models can reliably predict areas suitable for species occupation, and the species responses to particular environmental variables can suggest management alternatives.

**Acknowledgment**

This study was funded by Shahid Beheshti University, G.C. Special thanks and Gh. Amiri Ghadi for their support during our fish sampling.

**References**


