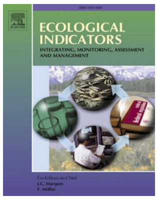
Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Atlantic salmon habitat-abundance modeling using machine learning methods

Bähar Jelovica [a,*], Jaakko Erkinaro [b], Panu Orell [b], Bjørn Kløve [a], Ali Torabi Haghighi [a], Hannu Marttila [a]

[a] *Water, Energy and Environmental Engineering Research Unit, University of Oulu, Finland*
[b] *Natural Resource Institute Finland (LUKE), Finland*

ARTICLE INFO

Keywords:
Atlantic salmon Abundance
Machine Learning Modelling
Habitat-Abundance Relationship
Arctic

ABSTRACT

Climate change and anthropogenic activities have impacts on fish habitat suitability, demanding more accurate modeling of species abundance for effective conservation and management. In this study, we applied Machine Learning techniques to model the habitat-abundance relationship of juvenile Atlantic salmon (*Salmo salar*) in the Teno catchment in Finland and Norway. To capture the complexity and nonlinearity of the habitat-abundance relationship, we employed Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting (GB), and Support Vector Classification (SVC) and compared their performances. Among the regression models considered, those incorporating input variables such as substrate, shade, and vegetation demonstrate higher performance. Support Vector Regression yields the highest mean cross-validation score ($R^2 = 0.58$), and Gradient Boosting produces the highest test score ($R^2 = 0.6$) among the regression techniques. The mean cross-validation and test scores obtained for the classification models are notably higher compared to the regression models across all scenarios. A comparison between regression and classification results highlights the challenges of accurately modeling the habitat-abundance relationship. This study provides insights into the challenges and potential of machine learning techniques for juvenile Atlantic salmon habitat-abundance modeling in complex riverine habitat environments. The findings emphasize the importance of considering the limitations of machine learning models, particularly in ecological contexts, and the need for further research to address temporal variations and improve the precision of habitat-abundance modeling.

1. Introduction

Climate change and anthropogenic activities influence fish habitat suitability by changing the river temperature (Isaak et al., 2012), vegetation distribution (Ackerly et al., 2015), food availability (Cameron et al., 2019), water quality (Ritson et al., 2014), and hydrological regimes (Jelovica et al., 2022). Since distribution and abundance of fluvial fish are strongly impacted by the habitat (Armstrong et al., 2003), it is essential to measure the indicators which are reflecting the habitat quality and could support river conservation and improvements (Giorgio et al., 2016). For example, stream depth, substrate, flow, shelter, temperature, and oxygen availability are some of the essential factors that influence the salmon abundance in their various life cycles. Abiotic factors such as riverbed geomorphology, hydrology, water quality, and aquatic environment are complex and have intricate independence. The habitat dataset is complex and the relationships among variables are not necessarily linear, making the modeling of the freshwater communities challenging (Armstrong et al., 2003, Mondal and Bhat, 2021). These challenges promote application of data-driven tools such as Machine Learning (ML) techniques (Lee et al., 2003, De'ath et al., 2000), which support nonlinear relations.

Machine learning (ML) is a strong statistical tool to identify nonlinear relationships in natural phenomena (Naghibi et al., 2016). In ecological studies, ML has been applied to model complex species community composition and abundance (Matsuzawa et al., 2023). For example, Mondal and Bhat (2021) modeled the species richness and diversity in eastern and central India using various ML approaches, leveraging abundance and ecological data. Wellman et al. (2020) used machine learning for modeling the ecology of urban birds and their habitats, whereas Xu et al., (2024) used support vector regression, RF, and extreme gradient boosting to predict the phytoplankton biomass using environment variables. In general, Support Vector Machine (SVM)

* Corresponding author.
E-mail address: bahar.jelovica@oulu.fi (B. Jelovica).

https://doi.org/10.1016/j.ecolind.2024.111832
Received 31 October 2023; Received in revised form 19 February 2024; Accepted 29 February 2024
Available online 7 March 2024

(Kang et al., 2022, Ahmadi et al., 2021, Fan et al., 2017, Park et al., 2015), Random Forest (RF) (Yang et al., 2020, Guo et al., 2019, Woo et al., 2019), and Gradient Boosting Regression (GBR) (Ficsór and Csabai, 2023, Garcia et al., 2018, Welchowski et al., 2022) have been widely used to solve ecohydrological and environmental problems. SVM proposed by Vapnik (1998) is one of the most popular ML techniques to describe nonlinear and complex data and has been used for uncertainty analysis (Liu et al., 2013, Singh et al., 2011). SVM has an excellent generalization performance and produces competitive results with the smallest amount of model tuning (Granata et al., 2017, Hoang et al., 2010). Fan et al. (2017) used SVM to predict the bio-indicators of an aquatic ecosystem in the Taizi River in China, and Kang et al. (2022) applied SVM to estimate the fish assessment index in South Korean rivers. On the other hand, RF method has been used in many research and studies due to its high accuracy and superiority (Ho, 1995, Amit and Geman, 1997, Breiman, 1996). Martínez-Santos et al. (2021) used SVM and RF to predict aquatic ecosystem mapping. Olaya-Marín et al. (2013) applied RF for fish richness in Mediterranean region. RF was an effective approach to assess the stream habitat conditions and demonstrate the seasonal, longitudinal, and local co-occurrence pattern of fish species in Yagawa River (Matsuzawa et al., 2023). Yang et al. (2020) developed a RF model to show the composition of fish species between two reservoirs in Yangtze River China. GBR model has successfully modelled problems with many variables and nonlinear relationships and shown high prediction accuracy (De'ath and Fabricius, 2000). As an example, Leathwick et al. (2006) analyzed the relationships between demersal fish species richness, environment, and trawl characteristics using GBR. Ficsór and Csabai (2023) applied various machine learning models such as RF and GB to predict the distribution of *Hydropsyche* and explained the impact of environmental factors on the dominant presence of the species.

Although, ML models exhibit significant capabilities in addressing challenges such as small dataset sizes, high dimensionality, and nonlinear problem domains (Ding et al., 2011), their performance and accuracy significantly depend on the size and quality of the training dataset. The habitat datasets are often small, scarce, and imbalanced due to costly and labor measurements. Therefore, it is challenging to precisely solve the problems related to these datasets (Danandeh Mehr et al., 2022, Crisci et al., 2012). To address this issue, we applied ML techniques that are less sensitive to the sample size. Since each model has advantages, we used multiple models and compared their performance to estimate habitat-abundance relationships in juvenile Atlantic salmon (*Salmo salar*). To enhance the performance of models, we applied a grid search algorithm to select the hyperparameters which control the learning process and model results. The grid search algorithm coupled with K-fold cross-validation (Fayed and Atiya, 2019) is employed to optimize the hyperparameters in the models. This utilization of cross-validation not only enhances the model's reliability and mitigates overfitting, but also provides a more accurate estimation of the model's generalization performance on the test set (Abobakr Yahya et al., 2019, Danandeh Mehr et al., 2022).

This study aims to model the relationships between abundance of juvenile Atlantic salmon and their fluvial habitat using datasets from the subarctic Teno catchment in the northernmost Scandinavia. Abundance refers to the density of juvenile Atlantic salmon per 100 m^2 in the studied area. Given the relatively small size of the habitat data and the intricate relationship between habitat characteristics and juvenile salmon abundance, we employed a range of ML techniques to model the habitat-abundance relationship of juvenile Atlantic salmon in two distinct age categories: fry (age- 0 +) and parr (age- 1 + and older).

2. Methods

2.1. Study area and data

The subarctic Teno River (Tana in Norwegian, Deatnu in Sami) forms the border between northernmost Finland and Norway at 70°N. With a catchment area of 16,386 km^2, it is one of the largest Atlantic salmon rivers running to the North-East Atlantic Ocean. The mean annual discharge of the river is 177 m^{3}s^{-1}, with spring flood peaking up to 2000–3000 m^{3}s^{-1}. The mean annual temperature is between ca. 0 to −3 °C and annual precipitation ranges from ca. 300–500 mm (Koster et al., 2005). Atlantic salmon in the Teno River are distributed over more than 1100 km of the main branch and tributaries. The population complex shows extraordinary diversity by numerous genetically distinct subpopulations (Vähä et al., 2017) and by vast variation in life-history strategies (Erkinaro et al., 2019).

Estimating juvenile Atlantic salmon abundance has been part of the long-term monitoring program in the Teno (Niemelä et al., 2005). Permanent monitoring sites (Fig. 1), have been distributed along three main branches of the Teno system including the Teno main stem and two of its large tributaries: Inarijoki and Utsjoki. Most of the Inarijoki follows the Finnish-Norwegian border. The Teno mainstream starts from the confluence of Inarijoki and another large headwater tributary, Karasjohka. Inarijoki has a length of 153 km with a drainage area of 3,152 km^2 and average monthly discharge of 36.4 m^{3}s^{-1}. Utsjoki is the largest tributary on the Finnish side of Teno catchment with a drainage area of 1,652 km^2 and mean discharge of 18 m^{3}s^{-1}.

Habitat data was collected from permanent electrofishing sites, (Fig. 1), in Teno, Inarijoki, and Utsjoki Rivers in two consecutive years from July to October. At each electrofishing site, three points on three transects, two points close to the edges and one in the middle of the site, were selected. Each point measured 0.25 m^2 (0.5 x 0.5 m). The habitat variables i.e., habitat characteristics of the electrofishing sites, include water temperature, average depth (cm), average velocity (cms^{-1}), substrate types, shade types, vegetation, and shelter index. The substrate was categorized into four groups including organic-silt-sand, gravel (2–16 mm), cobble (17–130 mm), stone (131–500 mm), and boulder (greater than 500 mm). The observed shades were mainly boulders and sometimes other structures such as large wooden debris. Vegetation includes moss, algae, and other plants. Shelter was estimated by visually identifying all potential interstitial spaces in the substratum. The depth was measured with a flexible PVC tube (13 mm diameter) where distances of 3, 5 and 10 cm were marked off (cf. Finstad et al., 2007). Spaces deeper than 3 cm (25–100 % of body length of the fish) were counted as a shelter, and three shelter size (depth) groups were identified: 3–5 cm; 5–10 cm; > 10 cm. Juvenile salmon abundance was defined as the number of fish per 100 m^2 (single-pass electrofishing, no removal estimates used) in two age groups: fry (age- 0 +) and parr (age- 1 + and older).

2.2. Regression and classification models

We employed SVR, RF, and GB as a set of regression techniques along with SVC as a classification tool to model the habitat-abundance relationship of juvenile Atlantic salmon. For the SVC model, fry and parr abundance are classified into distinct classes using their abundance histograms (Fig. 2). The classification of abundance offers a reduction in the complexity of the model's results as opposed to using regression models. We identified two and four distinctive classes for fry and parr abundance, respectively. Given the limited size of the habitat dataset, which comprises 14 habitat variables and only 114 records of data, we developed multiple models to assess and compare their performance.

2.2.1. Support vector Machine (SVM)

The Support Vector Machine technique is based on the dimension theory by Vapnik-Chervonenkis (1971) and provides a robust solution for both regression and classification problems based on a maximal margin hyperplane. SVM finds a dividing hyperplane with maximum margin. For a simple two-dimensional plane, the hyperplane is defined as $f(x) = \omega^T x_i + b$ where ω is the support vector and $b/\|\omega\|$ determines

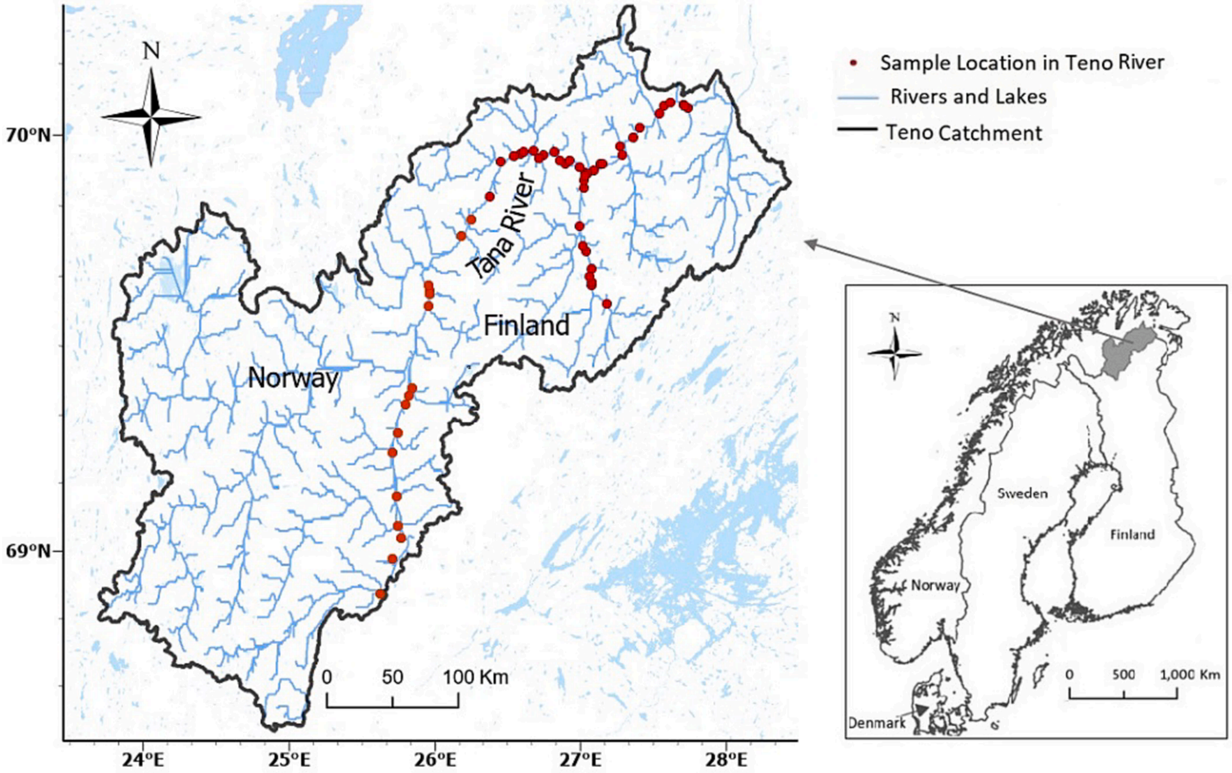


Fig. 1. Teno catchment and sampling sites in the Teno River located on the border between Finland and Norway.

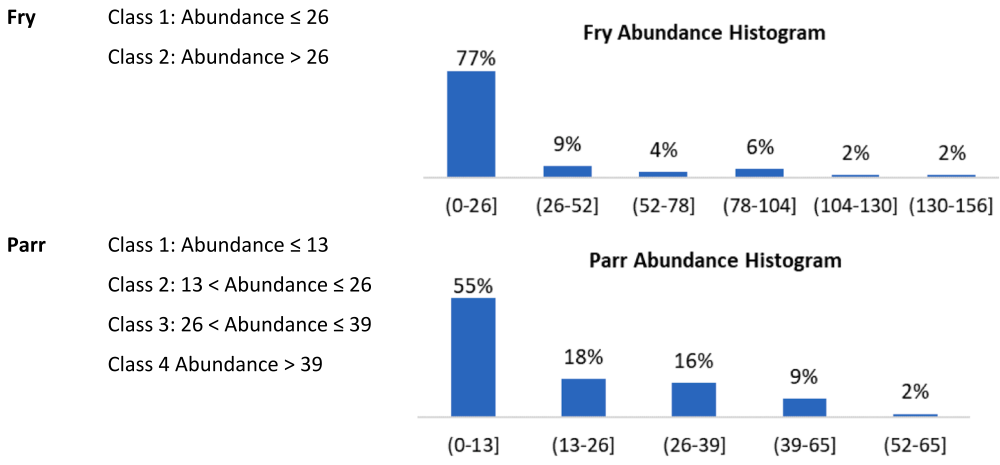


Fig. 2. The classification of fry and parr abundance based on their abundance histogram per 100 m^2 for the SVC model. This approach discerns two distinct classes for fry abundance and four distinct classes for parr abundance.

the offset of the hyperplane from the origin. In the case of non-linear relationships, SVM uses a technique called kernel trick which can project the data into high dimensional space. A detailed explanation of SVM technique can be found in the works of Abobakr Yahya et al. (2019), Rahimian Boogar et al. (2019), and Auerbach and Fremie, (2022).

The performance of SVM relies on the selection of kernel functions and hyperparameters such as gamma and Capacity (C). The common kernel functions are radial basis function (RBF), sigmoid, and polynomial. Notably, RBF exhibits higher efficiency compared to other kernel functions, as highlighted by Yang et al. (2021), making it the preferred choice in this study for constructing the SVR and SVC models.

The hyperparameters Gamma and C play essential roles in the performance of SVM. The accuracy of prediction is influenced by the appropriate selection of these hyperparameters. The Gamma parameter describes the width or slope of the kernel function which controls the complexity of the model whereas C affects the fundamental tradeoff. It is crucial to note that choosing a smaller value of C may result in underfitting (Abobakr Yahya et al., 2019).

2.2.2. *Random forest (RF)*

The Random Forest algorithm, introduced by Breiman (2001), is based on the concept of model aggregation to produce accurate predictions for both regression and classification problems. It is one of the most popular machine learning techniques widely employed in environmental studies (Vorpahl et al., 2012, Prasad et al., 2006). The technique has a fast-learning rate and can handle multidimensional datasets (Li et al., 2021).

Random forest is composed of numerous binary decision trees that use bootstrapping samples from the training dataset and a random selection of explanatory features at each node (Amit and Geman, 1997, Ho, 1998, Breiman, 2001). RF uses the bootstrap method to divide the original dataset into random subsets. A decision tree is trained independently for each subset. The result is obtained by averaging the predictions of all decision trees (Li et al., 2021). The randomness enforces the model's robustness and improves the learning process by changing from one random partitioned inventory subset to another to obtain the patterns of interest (Prasad et al., 2006). The Random Forest algorithm's performance is influenced by several hyperparameters, including the number of trees employed in the ensemble (Vorpahl et al., 2012).

2.2.3. Gradient boosting regression (GBR)

Gradient Boosting is a versatile technique applicable to both regression and classification problems. The technique relies on a set of weak learners or models such as decision trees. Since it is a boosting method, it builds the model by stages and achieves a single strong ensemble model optimizing a loss function. Friedman suggested the negative gradient of loss function $L(y, F(x))$ to approximate the loss in a Classification and Regression Tree (CART) $\widehat{F} = argminE_{x,y}[L(y, F(x))]$ where $\widehat{F}$ is an estimate of the function $F(x)$ and $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ is the training dataset (Friedman, 2002, Bühlmann and Hothorn, 2007).

The GBR method encompasses several hyperparameters that impact its performance such as number of estimators, learning rate, maximum depth, and minimum of sample leaf. A comprehensive description of GBR and its associated hyperparameters can be found in García Nieto et al., 2021.

2.2.4. Hyperparameter optimization

Although machine learning models are powerful in solving small dataset, high dimensional, and non-linear problems (Ding et al., 2011), the selection of hyperparameters affect their performance. Therefore, hyperparameter tuning is crucial for finding the optimal combination of hyperparameters that maximize the model's performance (Fayed and Atiya, 2019).

There are various approaches to tune the hyperparameters including the grid search algorithm (Fayed and Atiya, 2019), genetic algorithm (Sanz-Garcia et al., 2015), and swarm intelligence optimization algorithm (Adachi and Yoshida, 1995). The grid search involves a K-fold cross-validation widely used to assess the model's parameters.

During the K-fold cross-validation, the training dataset is partitioned into K-folds. A model is trained in sequence on K-1 folds and tested on the fold that is not used during training. This process is repeated for each fold, and the model's score is averaged over the folds. Cross-validation improves the model's reliability, mitigates overfitting, and provides a better estimate of generalization performance on the test set (Abobakr Yahya et al., 2019, Danandeh Mehr et al., 2022).

Table 1 presents the models and the values considered for the hyperparameter optimization. The values and interval boundaries are determined through trial and error, aiming to explore a comprehensive range of possibilities.

Table 1
Hyperparameter values employed in the grid search algorithm during cross-validation.

Model	Hyperparameter	Value
SVM	C	A set of evenly spaced numbers generated in the range of [0.01, 285]
	Gamma	A set of evenly spaced numbers generated in the range of [0.01, 380]
	Kernel	RBF
RF	Maximum depth	A set of evenly spaced numbers generated in the range of [1, 50]
GBR	Number of estimators	A set of evenly spaced numbers generated in the range of [100, 1000]
	Learning rate	{0.1, 0.05, 0.02}
	Maximum depth	{2, 4, 6}
	Minimum sample leaf	{3, 5, 9}

2.3. Selection of habitat variables and data preprocessing

We defined various scenarios for the selection of habitat variables incorporated in the models. These scenarios play a pivotal role in determining the importance of habitat variables when examined collectively (Scenario 1), individually (Scenario 2), or in various groups (Scenarios 3, 4, and 5) with regards to the models' results. One can choose alternative scenarios based on their specific modeling objectives to assess their impact on the results. Scenarios 3, 4 and 5 were specifically chosen following the model's results obtained for Scenarios 1 and 2. This selection enables a deeper exploration of the significance of the habitat variables on the juvenile salmon abundance. A conceptual map of the study is shown in Fig. 3.

- Scenario 1 includes all habitat variables in the models i.e., water temperature, mean depth, mean velocity, substrate types (organic-silt-sand, gravel, cobble, stone, boulder), shade types (boulder shade, other shade), and vegetation (algae, moss, plants), and shelter index.
- Scenario 2 considers each habitat variable individually. This scenario particularly investigates if a certain variable has a higher impact on the juvenile salmon abundance.
- Scenario 3 considers only substrate variables.
- Scenario 4 considers only shade and plants.
- Scenario 5 explores a combination of various substrates, shades, and vegetation (a combination of scenario 3 and 4).

Prior to model training, the dataset is divided into the training and test sets, with a split ratio of 20 %. The quality of the dataset would impact the model's output. Hence it is important to remove the outliers and missing values (Ro et al., 2015). Since the number of records is small in the dataset, we only excluded the outliers in the water temperature of less than 10° C.

A conventional strategy to deal with missing data is to remove the entire rows or columns containing missing values. However, this comes at the price of potentially losing valuable data. A more effective strategy is to infer the missing values from the known part of the data using the most suitable imputation technique. In this study, we deployed the mean strategy to impute the missing numeric values.

Since the dataset values exhibit different ranges, normalization is necessary to ensure optimal training speed and accurate results. We employ Min-max normalization method, which maps all the data to the range between 0 and 1 using equation (1) (Yang et al., 2021).

$$X_{nor} = (x - x_{min})/(x_{max} - x_{min}) \tag{1}$$

Where x is the original data, x_{nor} is the normalized data, x_{max} and x_{min} are the maximum and minimum values of the data, respectively.

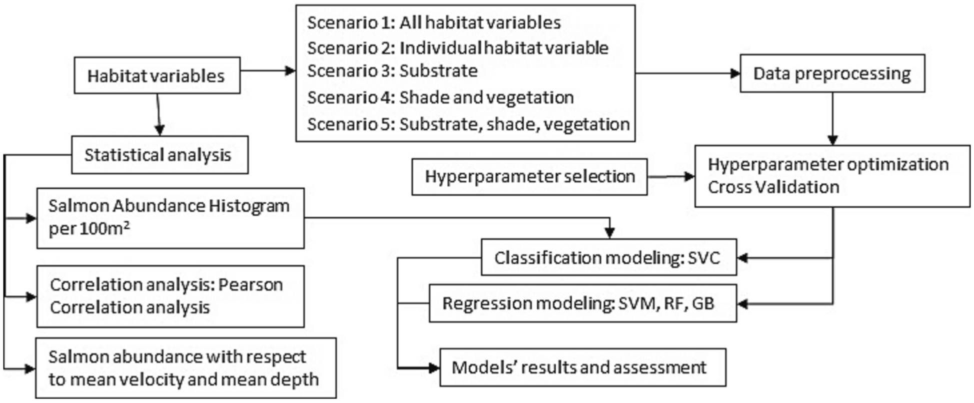
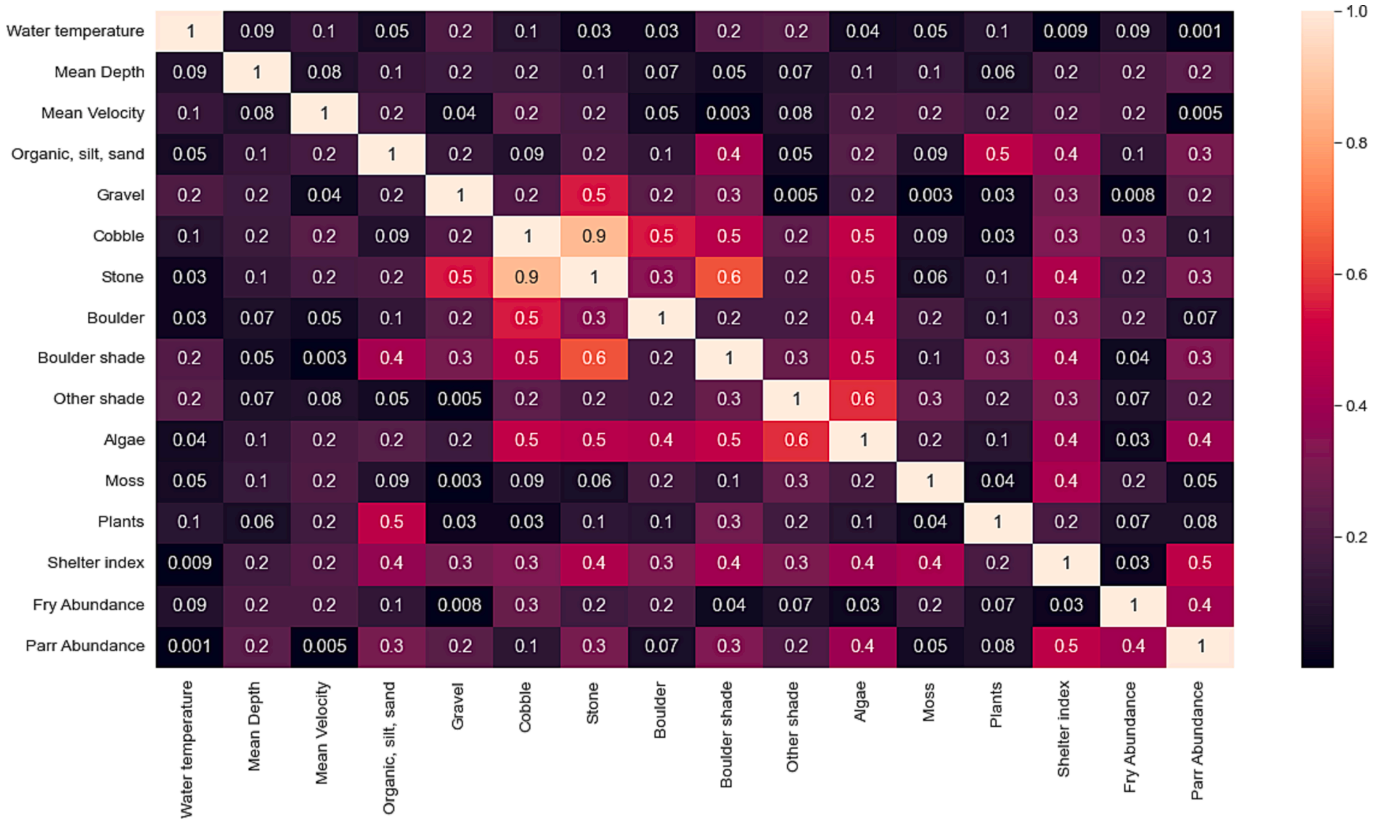

Habitat variables
Scenario 1: All habitat variables
Scenario 2: Individual habitat variable
Scenario 3: Substrate
Scenario 4: Shade and vegetation
Scenario 5: Substrate, shade, vegetation
Data preprocessing
Statistical analysis
Hyperparameter selection
Hyperparameter optimization Cross Validation
Salmon Abundance Histogram per 100m²
Classification modeling: SVC
Correlation analysis: Pearson Correlation analysis
Regression modeling: SVM, RF, GB
Salmon abundance with respect to mean velocity and mean depth
Models' results and assessment

Fig. 3. Conceptual map of this study including statistical analysis of the habitat variables, various scenarios to select habitat variables for modeling, data preprocessing, hyperparameter optimization and cross validation imbedded in the regression and classification modeling, and model evaluation.



	Water temperature	Mean Depth	Mean Velocity	Organic, silt, sand	Gravel	Cobble	Stone	Boulder	Boulder shade	Other shade	Algae	Moss	Plants	Shelter index	Fry Abundance	Parr Abundance
Water temperature	1	0.09	0.1	0.05	0.2	0.1	0.03	0.03	0.2	0.2	0.04	0.05	0.1	0.009	0.09	0.001
Mean Depth	0.09	1	0.08	0.1	0.2	0.2	0.1	0.07	0.05	0.07	0.1	0.1	0.06	0.2	0.2	0.2
Mean Velocity	0.1	0.08	1	0.2	0.04	0.2	0.2	0.05	0.003	0.08	0.2	0.2	0.2	0.2	0.2	0.005
Organic, silt, sand	0.05	0.1	0.2	1	0.2	0.09	0.2	0.1	0.4	0.05	0.2	0.09	0.5	0.4	0.1	0.3
Gravel	0.2	0.2	0.04	0.2	1	0.2	0.5	0.2	0.3	0.005	0.2	0.003	0.03	0.3	0.008	0.2
Cobble	0.1	0.2	0.2	0.09	0.2	1	0.9	0.5	0.5	0.2	0.5	0.09	0.03	0.3	0.3	0.1
Stone	0.03	0.1	0.2	0.2	0.5	0.9	1	0.3	0.6	0.2	0.5	0.06	0.1	0.4	0.2	0.3
Boulder	0.03	0.07	0.05	0.1	0.2	0.5	0.3	1	0.2	0.2	0.4	0.2	0.1	0.3	0.2	0.07
Boulder shade	0.2	0.05	0.003	0.4	0.3	0.5	0.6	0.2	1	0.3	0.5	0.1	0.3	0.4	0.04	0.3
Other shade	0.2	0.07	0.08	0.05	0.005	0.2	0.2	0.2	0.3	1	0.6	0.3	0.2	0.3	0.07	0.2
Algae	0.04	0.1	0.2	0.2	0.2	0.5	0.5	0.4	0.5	0.6	1	0.2	0.1	0.4	0.03	0.4
Moss	0.05	0.1	0.2	0.09	0.003	0.09	0.06	0.2	0.1	0.3	0.2	1	0.04	0.4	0.2	0.05
Plants	0.1	0.06	0.2	0.5	0.03	0.03	0.1	0.1	0.3	0.2	0.1	0.04	1	0.2	0.07	0.08
Shelter index	0.009	0.2	0.2	0.4	0.3	0.3	0.4	0.3	0.4	0.3	0.4	0.4	0.2	1	0.03	0.5
Fry Abundance	0.09	0.2	0.2	0.1	0.008	0.3	0.2	0.2	0.04	0.07	0.03	0.2	0.07	0.03	1	0.4
Parr Abundance	0.001	0.2	0.005	0.3	0.2	0.1	0.3	0.07	0.3	0.2	0.4	0.05	0.08	0.5	0.4	1

Fig. 4. Pearson Correlation Coefficients among habitat data and juvenile salmon abundance, ranges from 0 to 1 with one indicates a strong positive correlation.

3. Results

3.1. Habitat data

The linear relationships among habitat variables and juvenile salmon abundance are evaluated using Pearson Correlation Coefficient (PCC) (Fig. 4) which ranges from 0 to 1 with one indicates a strong positive correlation. The PCC revealed no significant correlations between fry abundance and the habitat variables. The highest PCC value of 0.44 is observed between fry and parr abundance.

Among the habitat variables, the shelter index, algae, boulder shade, stone and organic substrate exhibit higher PCC values with parr abundance. Notably, the shelter index demonstrates the highest PCC value of 0.48 with parr abundance. However, the correlations between the habitat variables and parr abundance are not found to be statistically significant overall.

Among habitat variables, a notable high correlation of 0.9 is observed between stone and cobble. Boulder shade exhibits a relatively stronger correlation of 0.6 with stone, and both other shade and algae demonstrate a comparatively higher PCC of 0.6 when compared to the remaining habitat variables.

3.2. Juvenile salmon abundance with respect to the depth and water velocity

The distribution of mean water velocity and mean depth with respect to the juvenile salmon abundance graphs are shown in Fig. 5a and b respectively. The fry and parr abundance graphs present the total

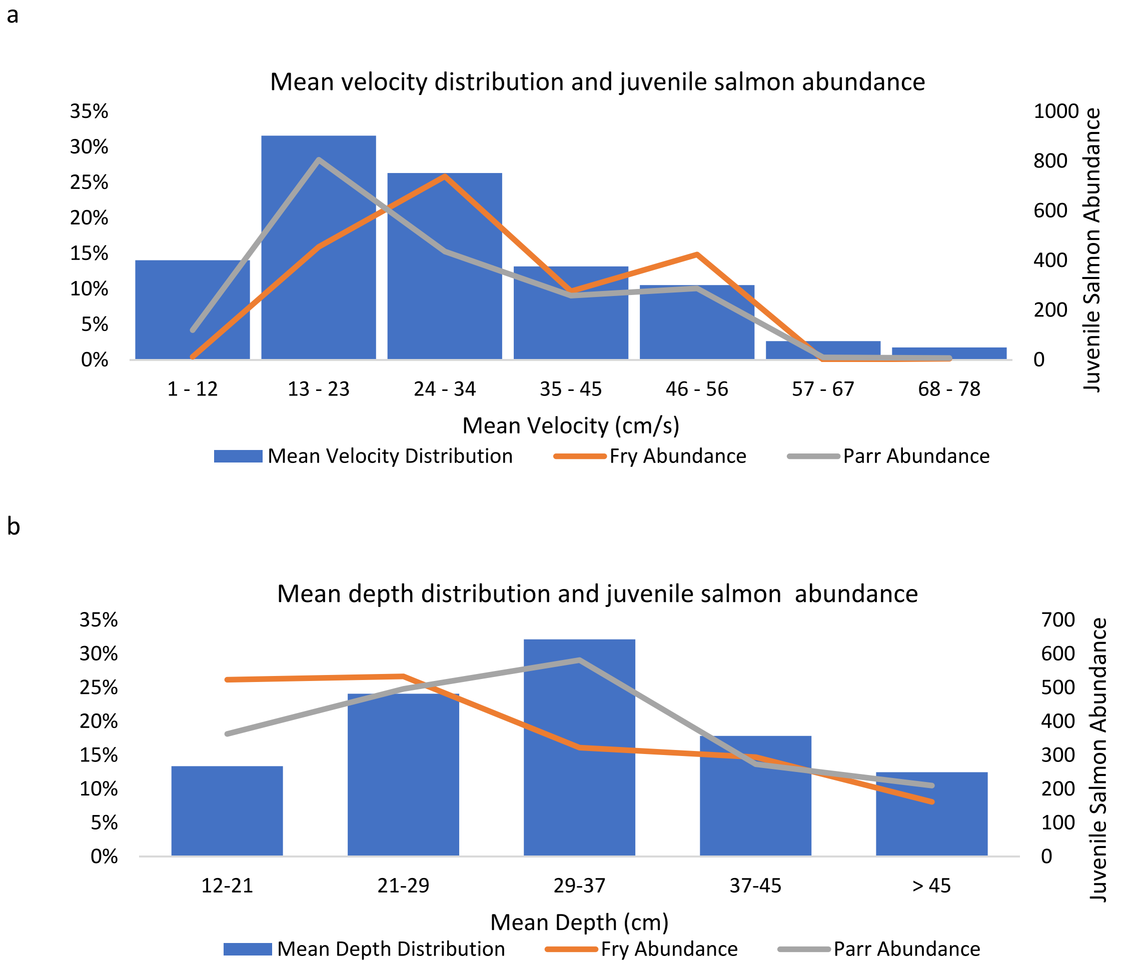


Fig. 5. A) mean velocity distribution, b) mean depth distribution with respect to fry and parr abundance which are estimated across all electrofishing site in two years.

abundance estimated across all electrofishing sites for two years. Fry and parr have significantly higher abundance at the velocity ranges between 13 and 56 cms^{-1} in comparison to the other velocity ranges. Conversely, their abundance is the lowest at velocities below 13 and above 56 cms^{-1}. Fry abundance increases significantly at velocities above 13 cms^{-1}, reaches its peak within 24 and 34 cms^{-1} and drops at velocities exceeding 35 cms^{-1}. On the other hand, the parr abundance reaches its highest at velocities between 13 and 23 cms^{-1} and declines at velocities beyond 24 cms^{-1}. The mean velocity ranges between 2 and 78 cms^{-1}, and 96 % of the studied area has a mean velocity below 56 cms^{-1}.

From the mean depth distribution and abundance graphs (Fig. 5b), we observed that the fry abundance peaks at the mean depths below 29 cm which constitute about 40 % of the studied area and it decreases at the depth beyond 29 cm. Parr abundance gradually increases within various depth ranges below 37 cm and reaches its highest at depths between 29 and 37 cm, but significantly declines at depths exceeding 37 cm. Approximately 70 % of the studied area exhibits mean depth ranges between 12 and 37 cm. Both fry and parr have the lowest abundance at depths above 45 cm.

3.3. *Regression modeling*

The SVR, RF, and GB regression techniques are evaluated using the coefficient of determination (R^2). The optimized parameter values and corresponding R^2 scores for the validation and test sets are presented in Table 2 for fry and Table 3 for parr. It is important to note that all individual habitat variables (scenario 2), except the shelter index, resulted in negative R^2 values for both the validation and test sets and, thus, are not reported.

For fry abundance, the SVR achieves the highest mean cross-validation score of 0.58 in scenario 5 which involves substrates, shades, and vegetation. Regarding the test score, SVR performs best in scenario 3 ($R^2 = 0.28$), which includes solely the substrates. On the other hand, both RF and GB attained the highest mean cross-validation scores ($R^2 = 0.33$ and $= 0.44$, respectively) when considering the shelter index (scenario 2), while achieving the highest test scores ($R^2 = 0.46$ and $R^2 = 0.49$, respectively) for substrates (scenario 3). Comparing all models and scenarios, SVR demonstrates the highest mean cross-validation score ($R^2 = 0.58$), indicating a better fit to the validation data. However, GBR demonstrates the highest test score (R^2 = 0.49), suggesting superior generalization performance on unseen data.

The grid search heatmaps for Fry habitat-abundance modeling, (Fig. 6), illustrate the influence of various combinations of hyperparameters on mean cross-validation scores (R^2). Darker colors show higher scores. These heatmaps focus exclusively on the models that attained the highest mean cross-validation scores across all scenarios, shown in bold font in Table 2. A subset of hyperparameter values is selected for each heatmap to optimize the clarity and interpretability in presenting the impact of hyperparameter tuning on model performance.

The SVR models exhibit varying performance with different values of

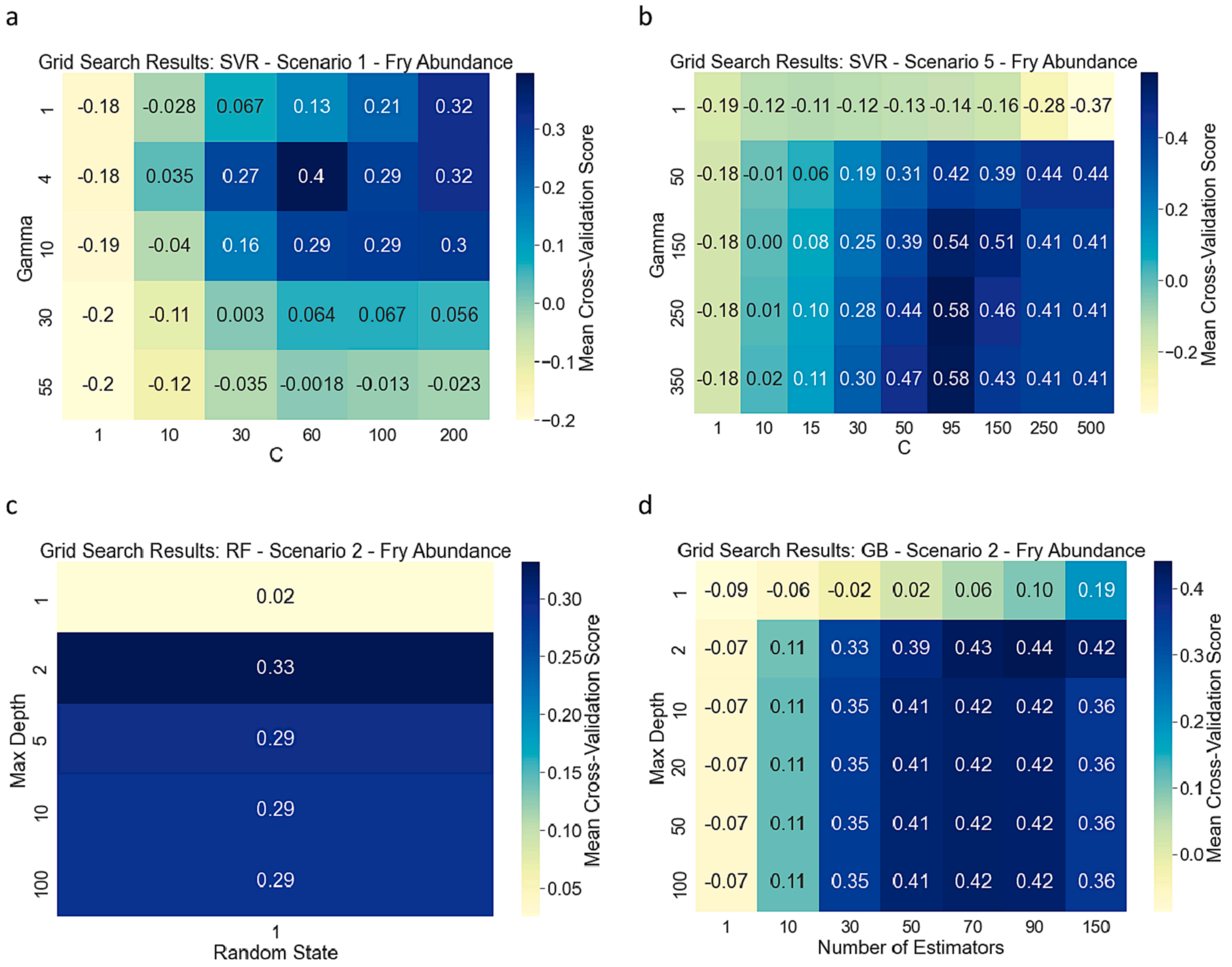


Fig. 6. Grid-Search heatmaps of hyperparameters and mean cross-validation scores (R^2) in Fry habitat-abundance models: a) SVR - Scenario 1, b) SVR - Scenario 5, c) RF - Scenario 2 (Shelter Index), d) GBR - Scenario 2 (Shelter Index). Darker colors show higher scores.

C and gamma hyperparameters in scenario 1 and 5 (Fig. 6a and b). In scenario 1, specifically for C and gamma values of 60 and 4, respectively, the model demonstrates optimal performance. Notably, lower values of C (1 and 10) across the suggested range of gamma values result in suboptimal performance. The model performance enhances gradually with larger values of C, particularly when gamma is set to 4 and 10.

In scenario 5 (Fig. 6b), the optimal C and gamma values are 95 and 250, respectively. The model performance is notably low for smaller values of C (1 and 10). It gradually increases with larger values of C until it reaches the optimum at 95, after which it begins to decline. Interestingly, the model's performance remains relatively consistent at higher C values. On the other hand, a smaller gamma value of 1 across the suggested range of C values results in poor model performance, while larger gamma values, especially in conjunction with larger C values (greater than 10), gradually enhance the model's performance.

The heatmaps for random forest and gradient boosting (Fig. 6c and d, respectively) are presented for shelter index in scenario 2, where the models illustrate the highest mean cross-validation scores (Table 2). The random forest performance is optimal at maximum depth hyperparameter of 2, with consistent performance for larger maximum depth values.

In the case of gradient boosting, performance gradually improves with an increase in the number of estimators, reaching optimal performance at 90 estimators and a maximum depth of 2. The model performance is relatively stable when number of estimators ranges between 50 and 90 and the maximum depth is greater than 2. The performance is notably low when the number of estimators is set to 1 across various suggested maximum depths.

In parr habitat-abundance modeling, SVR achieves the highest mean cross-validation scores for scenarios 3 (substrates, $R^2 = 0.55$), 4 (shade and vegetation, $R^2 = 0.53$) and 5 (substrates, shade, and vegetation, $R^2 = 0.53$). RF obtains the highest mean cross-validation score for scenario 4 (shades and vegetation, $R^2 = 0.4$). GBR exhibits the highest mean cross-validation scores when considering all variables (scenario 1, $R^2 = 0.44$) and the highest test score for scenario 4 (shade and vegetation, $R^2 = 0.6$).

The grid search heatmaps (Fig. 7) illustrate the influence of varying combinations of hyperparameters on mean cross-validation scores (R^2) in Parr habitat-abundance modeling. These heatmaps focus exclusively on the models that attained the highest mean cross-validation scores across all scenarios, shown in bold font in Table 3. A subset of hyperparameter values is selected for each heatmap to optimize the clarity and interpretability in presenting the impact of hyperparameter tuning on model performance.

Table 2
SVM, RF, GBR modeling results for fry with R^2 as a performance metric. (LR: Learning-Rate, MD: Max-Depth, MSL: Min-Samples-Leaf, NE: n-estimators). Bolded numbers indicate the highest mean cross-validation/test scores per model and scenario.

Model	Habitat variable	Optimized hyperparameter value	Mean Cross-Validation Score (R^2)	Test Score (R^2)
SVR	All variables (Scenario 1)	C: 60, gamma: 4	**0.4**	0.14
	Shelter index (Scenario 2)	C: 58.8 - gamma: 180	0.1	−0.24
	Substrate: organic-silt-sand, gravel, cobble, stone, boulder (Scenario 3)	C: 90 - gamma: 250	0.4	**0.28**
	Shade and vegetation: boulder shade, other shade, algae, moss, plants (Scenario 4)	C: 124 - gamma: 280	0.16	0.23
	Substrate, shade, and vegetation (Scenario 5)	C: 95 - gamma: 250	**0.58**	0.2
RF	All variables	MD: 11, RS: 9	−0.21	0.11
	Shelter index	MD: 2, RS: 1	**0.33**	−0.4
	Substrate: organic-silt-sand, gravel, cobble, stone, boulder	MD: 10, RS: 9	0.14	**0.46**
	Shade and vegetation: boulder shade, other shade, algae, moss, plants	MD: 1, RS: 1	−0.13	−0.05
	Substrate, shade, and vegetation	MD: 13, RS: 20	0.12	0.41
GB	All variables	LR: 0.02, MD: 2, MSL: 5, NE: 200	0.2	0.35
	Shelter index	LR: 0.02, MD: 2, MSL: 5, NE: 100	**0.44**	−0.18
	Substrate: organic-silt-sand, gravel, cobble, stone, boulder	LR: 0.02, MD: 6, MSL: 3, NE: 200	0.3	**0.49**
	Shade and vegetation: boulder shade, other shade, algae, moss, plants	LR: 0.02, MD: 2, MSL: 9, NE: 100	−0.3	−0.08
	Substrate, shade, and vegetation	LR: 0.05, MD: 4, MSL: 5, NE: 100	0.23	0.34

The SVR model performance in scenario 3 and 4 (Fig. 7a and b) are notably low for small C values (e.g., 1) across the suggested gamma values. In scenario 3 (Fig. 7a), the model performance improves as both C and gamma values increase, reaching optimum performance at C = 36 and gamma = 225. The model performance is significantly poor with smaller values of gamma (e.g., 1). In scenario 4 (Fig. 7b), smaller values of C and gamma (e.g., 1) result in low model performance. The model performance improves with values of C and gamma increasing and reaches the best performance at 257 and 226 respectively and stays consistent afterwards.

The random forest model performance in scenario 4, (Fig. 7c), is poor for smaller value of maximum depth (e.g., 1) and improves as maximum depth increase, reaching the optimum at 13. The model performance is not sensitive to larger values of maximum depth (e.g., greater than 13). In the case of gradient boosting in scenario 1, (Fig. 7d), performance is suboptimal when number of estimators set to 1, but it gradually improves with an increase in the number of estimators, reaching optimal performance at 900 estimators and a maximum depth of 2.

3.4. *Classification modeling*

The SVC models is applied to predict the fry and parr abundance based on the classes defined using their abundance histogram (Fig. 2). The mean cross-validation and test scores present the accuracy of the model. As explained in section 2.2, two and four classes are considered in the SVC modeling respectively. The mean cross-validation and test scores obtained for the classification models (Table 4 and Table 5) are notably higher compared to the regression models across all scenarios. Among the scenarios for fry (Table 4), the model achieves the highest accuracy when considering the shelter index, with mean cross-validation and test scores of 91 % and 96 %, respectively. Notably, the test scores remain consistent across all scenarios, except for the substrate types organic, silt, and sand, which exhibit slight variations in performance.

In the parr habitat-abundance modeling, scenario 5 yields the highest cross-validation accuracy which involves substrate, shade, and vegetation variables. Conversely, scenario 2 exhibits lower mean cross-validation scores compared to the other scenarios. Within scenario 2, most individual habitat variables display similar mean cross-validation accuracies, except for mean depth, stone, and shelter index, which demonstrate higher mean cross-validation scores. Across all scenarios, the test scores remain consistent, except for the algae and moss, which show lower accuracies.

4. Discussion

In our study at the subarctic Teno catchment, we employed several machine learning (ML) techniques to model the habitat-abundance relationships of juvenile Atlantic salmon. The performance and accuracy of ML techniques can be greatly influenced by the size of dataset. Since collection of long-term habitat data is often costly and laborious, these datasets tend to be scarce and imbalanced in nature (Niemelä et al., 2005, Danandeh Mehr et al., 2022, Matsuzawa et al., 2023). To address these issues, we used ML models which are less sensitive to the sample size and have been proven to overcome the complexity and uncertainties in the data (Davoudi Moghaddam et al., 2020, Liu et al., 2013, Singh et al., 2011, Guisan and Thuiller, 2005). Furthermore, we performed feature scaling (normalization), K-fold cross-validation, and hyperparameter optimization to improve the performance of the models. Despite our efforts, the R^2 scores obtained from the regression models are not notably high which could be due to the small sample size, uncertainties in habitat data and complex nature of species (Matsuzawa et al., 2023). The ML techniques rely on the data and would perform better with large datasets (Chapelle et al., 2002). However, most ecological datasets are small unless they are part of large-scale projects (Mondal and Bhat, 2021). Since data from long-term monitoring may help to improve the performance of the models, therefore, we suggest collecting more local habitat data or combining international datasets from similar river systems to ensure more data records in the modeling.

Grid-Search heatmaps visually depict how different combinations of hyperparameters affect the performance of regression models in our results (see Fig. 7). In SVR model, small C values indicate a wider range for decision boundary which may result in suboptimal performance. Increasing C typically leads to a narrower margin for the decision boundary, which can result in better performance. Gamma controls the curvature of the Gaussian Kernel function. Small values of gamma correspond to smoother decision boundary, leading to a broader acceptance of data points in the calculations. Larger values of gamma can lead to a more complex decision boundary, which might improve performance in cases where the relationship between input and output variables is highly non-linear. Very large gamma values lead to

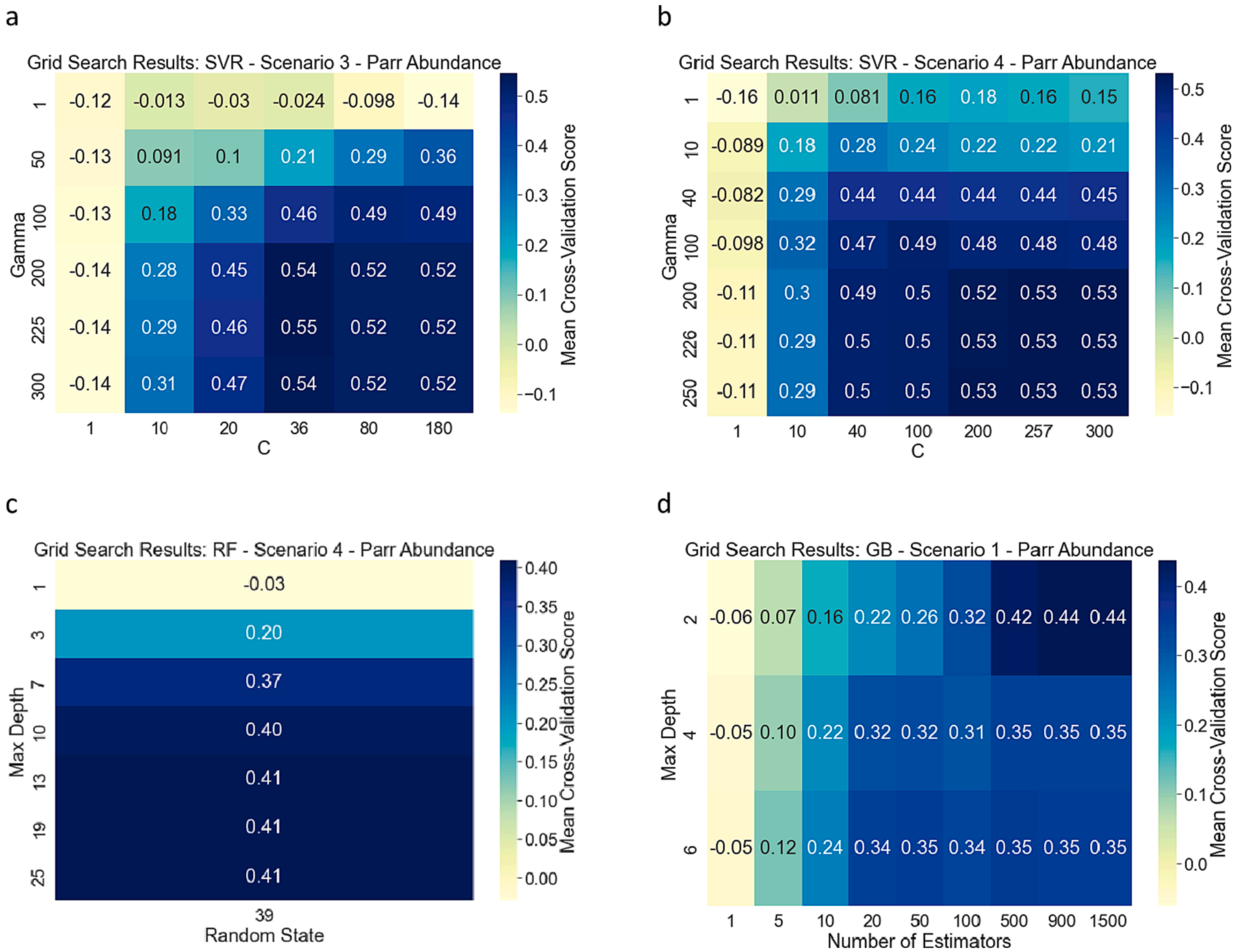


Fig. 7. Grid-Search heatmaps of hyperparameters and mean cross-validation scores (R^2) in Parr habitat-abundance regression models: a) SVR - Scenario 3, b) SVR - Scenario 4, c) RF - Scenario 4, d) GBR – Scenario 1. Darker colors show higher scores.

overfitting (Kalita et al., 2023). Therefore, it is necessary to find optimal hyperparameters using e.g., grid search method and cross validation to avoid overfitting (Fayed and Atiya, 2019). In all our SVR models small values of C and gamma (e.g., 1) resulted in poor performance which shows the complex non-linear relationship between habitat variables and salmon abundance. The SVR models' performance improves as the values of C and gamma increase as they reach the best performance. In RF and GB approaches, the depth of decision tree represents the length of each tree. The deeper decision tree means more split permitting the trees to describe more variation in the dataset (Breiman, 2001). In the presented heatmaps for RF and GB models, small value of maximum depth (i.e., 1) leads to low model performance whereas larger values do not necessarily improve the models. The R^2 for the RF models are the lowest among all models across all scenarios except in scenario 2 (shelter index) of fry modeling. In GB, each additional estimator contributes to model complexity and may improve the performance. Lower number of estimators result in suboptimal performance and under fitting (Sagi and Rokach, 2018) which we observed in GB heatmaps. Whereas larger number of estimators result in model improvement (Callens et al., 2020).

The feature selection is a helpful pre-processing strategy to build a simpler model and improve the performance (Li et al., 2017). In the regression models, we observed that the selection of specific habitat variables, such as substrates, shades, and vegetation, leads to improved performance. The R^2 values obtained for these variables in scenarios 3, 4, and 5 are significantly higher compared to scenario 2, where the models only consider the individual habitat variables and were unable to establish meaningful relationships between each habitat variables and juvenile salmon abundance. Our results indicated that a combination of habitat conditions, i.e. selection of certain features, such as substrates, shades, and vegetation is more effective than individual variables like water depth or water velocity, which have been reported to impact juvenile salmon in previous studies (e.g., Mäki-Petäys et al., 2004, Binns and Eiserman, 1979, Heggenes, 1990). The inconsistencies in the results of scenario 2 with the existing studies may be due to the robustness of the data. If the data is scarce or does not encompass the various aspects of the task, the learning could fall short, affecting the performance of ML (Mosavi et al., 2018).

To gain a better insight about the individual habitat variables such as water depth and water velocity where the models exhibited low performance, we specifically looked at the mean velocity and mean depth distributions with respect to the salmon abundance in all study sites (Fig. 5). A study conducted by Mäki-Petäys et al. (2002) in the Teno River reported that fry prefer near zero and below 20 cms^{-1} velocities, while parr exhibit a preference for velocities ranging between 35 and 80 cms^{-1}. Our results (Fig. 5a) reveal some deviations in the abundance

Table 3

SVM, RF, GBR modeling results for parr with R^2 as a performance metric. (LR: Learning-Rate, MD: Max-Depth, MSL: Min-Samples-Leaf, NE: n-estimators, RS: Random-State). Bolded numbers indicate the highest mean cross-validation/test scores per model and scenario.

Model	Habitat variable	Optimized hyperparameter value	Mean Cross-Validation Score	Test Score
SVR	All variables (Scenario 1)	C: 44.9 - gamma: 2	0.41	0.04
	Shelter index (Scenario 2)	C: 36.1 - gamma: 11	0.23	−0.04
	Substrate: organic-silt-sand, gravel, cobble, stone, boulder (Scenario 3)	C: 36 - gamma: 225	**0.55**	0.23
	Shade and vegetation: boulder shade, other shade, algae, moss, plants (Scenario 4)	C: 257.1 - gamma: 226.5	**0.53**	0.12
	Substrate, shade, and vegetation (Scenario 5)	C: 36.7 - gamma: 38.4	**0.53**	0.12
RF	All variables	MD: 5 - RS: 2	0.31	−0.29
	Shelter index	MD: 1 - RS: 12	0.12	−0.04
	Substrate: organic-silt-sand, gravel, cobble, stone, boulder	MD: 10 - RS: 35	0.31	−0.23
	Shade and vegetation: boulder shade, other shade, algae, moss, plants	MD: 13 - RS: 39	**0.4**	0.47
	Substrate, shade, and vegetation	MD: 9 - RS: 34	0.31	−0.09
GB	All variables	LR: 0.05, MD: 2, MSL: 5, NE: 900	**0.44**	−0.7
	Shelter index	LR: 0.02, MD: 4, MSL: 9, NE: 300	0.32	−0.3
	Substrate: organic-silt-sand, gravel, cobble, stone, boulder	LR: 0.05, MD: 2, MSL: 9, NE: 900	0.4	−1.07
	Shade and vegetation: boulder shade, other shade, algae, moss, plants	LR: 0.05, MD: 6, MSL: 5, NE: 100	0.34	**0.6**
	Substrate, shade, and vegetation	LR: 0.05, MD: 4, MSL: 3, NE: 200	0.4	0.16

Table 4

SVC results for modeling fry abundance. Bolded numbers indicate the highest mean cross-validation/test scores per model and scenario.

Scenario	Habitat variables	Best Parameters	Mean cross-validation Score	Test Score
Scenario 1	All variables	C: 42.86 - gamma: 3.79	0.86	0.96
Scenario 2	Water temperature	C: 0.01 - gamma: 0.01	0.77	0.96
	Mean depth	C: 8.17 - gamma: 93.88	0.79	0.96
	Mean velocity	C: 0.01 - gamma: 0.01	0.77	0.96
	Organic, silt, sand	C: 36.74 - gamma: 44.9	0.78	0.74
	Gravel	C: 63.27 - gamma: 100	0.78	0.96
	Cobble	C: 0.01 - gamma: 0.01	0.77	0.96
	Stone	C: 0.01 - gamma: 0.01	0.77	0.96
	Boulder	C: 0.01 - gamma: 0.01	0.77	0.96
	Boulder shade	C: 4.09 - gamma: 77.55	0.78	0.96
	Other shade	C: 2.05 - gamma: 12.25	0.79	0.96
	Algae	C: 0.01 - gamma: 0.01	0.77	0.96
	Moss	C: 0.01 - gamma: 0.01	0.77	0.96
	Plants	C: 0.01 - gamma: 0.01	0.77	0.96
	Shelter index	C: 24.5 - gamma: 97.96	**0.91**	**0.96**
Scenario 3	Substrate: organic-silt-sand, gravel, cobble, stone, boulder	C: 32.66 - gamma: 53.07	0.82	0.96
Scenario 4	Shade and vegetation: boulder shade, other shade, algae, moss, plants	C: 48.98 - gamma: 97.96	0.80	0.96
Scenario 5	Substrate, shade, and vegetation	C: 4.091 - gamma: 89.8	0.85	0.96

patterns of fry and parr compared to the outcomes reported by Mäki-Petäys et al. (2002). For fry, our results indicated significantly lower abundance at near zero velocities and the highest at velocities between 24 and 34 cms^{-1}. For parr, the abundance peaked between 13 and 23 cms^{-1} and declined as the velocities increased beyond this range. On the other hand, we observed that the abundance of both fry and parr across various depth ranges align with the measurements reported by Mäki-Petäys et al. (2002) where fry and parr optimal depths are 5 to 25 cm and 5 to 35 cm respectively. Nevertheless, approximately 40 % of fry abundance were identified at depths greater than 29 cm, which is outside the optimum reported by Mäki Petäys et al. (2002). These deviations from the expected preferences in mean velocity and mean depth in our dataset may result from environmental factors such as local habitat, competition, or juvenile salmon behavior adaptation in the specific study area (Mäki-Petäys et al., (2002), Rosenfeld et al., 2005), or data quality (Mosavi et al., 2018). The data can be enriched through invariance assessment to obtain the group characteristic (Tsai and Yang, 2012) or by using casually dependent coefficients for handling missing values (Sivapalan et al., 2005). These approaches may enhance the robustness of the data which could impact the model development and performance. In addition, we recommend exploring other advanced ML techniques, such as deep learning, for modeling as in some cases they have proven to be useful in habitat abundance modeling (Ditria et al., 2020).

In this research, to decrease the complexity of regression models and improve the performance, we suggested a practical approach by employing support vector classification technique to classify the fry and parr abundance based on their abundance histograms. Transformation of a regression problem into a classification could improve the model performance (Salman and Kecman, 2012, Torgo and Gama, 1996). While the classification models do not predict the abundance values like regression models, they offer a useful approximation of juvenile salmon abundance within the specific defined classes. The fry abundance classification model demonstrates high accuracies on both the mean cross-validation and test sets across all scenarios. However, the accuracies

Table 5
SVC results for modeling parr abundance. Bolded numbers indicate the highest mean cross-validation/test scores per model and scenario.

Scenario	Habitat variables	Best Parameters	Mean cross-validation Score	Test Score
Scenario 1	All variables	C: 97.96 - gamma: 0.01	**0.55**	0.48
Scenario 2	Water temperature	C: 0.01 - gamma: 0.01	0.47	0.48
	Mean Depth	C: 4.09 - gamma: 20.41	**0.51**	0.48
	Mean Velocity	C: 0.01 - gamma: 0.01	0.47	0.48
	Organic, silt, sand	C: 0.01 - gamma: 0.01	0.47	0.48
	Gravel	C: 0.01 - gamma: 0.01	0.47	0.48
	Cobble	C: 16.33 - gamma: 30.62	0.48	0.48
	Stone	C: 97.96 - gamma: 79.6	**0.56**	0.48
	Boulder	C: 14.29 - gamma: 65.31	0.48	0.48
	Boulder shade	C: 0.01 - gamma: 0.01	0.47	0.48
	Other shade	C: 28.58 - gamma: 4.091	0.48	0.48
	Algae	C: 6.13 - gamma: 16.33	0.47	**0.43**
	Moss	C: 4.09 - gamma: 12.25	0.48	**0.35**
	Plants	C: 2.05 - gamma: 57.15	0.48	0.48
	Shelter index	C: 28.58 - gamma: 59.19	**0.53**	0.48
Scenario 3	Substrate: Organic-silt-sand, gravel, cobble, stone, boulder	C: 6.13 - gamma: 30.62	**0.55**	0.48
Scenario 4	Shade and vegetation: boulder shade, other shade, algae, moss, plants	C: 44.9 - gamma: 95.92	**0.52**	0.48
Scenario 5	Substrate, shade, and vegetation	C: 2.05 - gamma: 77.55	**0.59**	0.48

of SVC models for parr abundance are comparatively lower. This can be attributed to the increased complexity of the model, which considers four distinct classes instead of two in the SVC modeling for fry. While classification models may not capture the full complexity of habitat data, they can still offer valuable insights for understanding and managing juvenile salmon abundance.

Atlantic salmon is a critical natural resource and cultural element in Northern Europe (Landauer et al., 2023), supporting local societies, ecosystems, and ecosystem services. In addition, salmon has a special role in the culture of indigenous Sami people in the Teno River catchment (Hiedanpää et al., 2020). Currently, Atlantic salmon stocks in the North Atlantic area are declining (ICES, 2023) and more information is urgently needed to support environmental policy and management decisions for conservation and restoration (e.g. Lennox et al., 2021). This includes an improved understanding of habitat-abundance relationships for different life stages of Atlantic salmon in the riverine conditions. Our study demonstrated that small-size datasets and the presence of complex non-linear relationships between habitat and juvenile salmon abundance could impact the models' performance and reliability. These factors can pose challenges in capturing the relationship between species abundance and habitat conditions and the effectiveness of ML techniques (Danandeh Mehr et al., 2022; Guo et al., 2015; McPherson and Jetz, 2007). The availability of data and the characteristics of specific ecological communities could influence the selection and development of models (Mondal and Bhat, 2021). In future studies, it is crucial to consider factors like data quality and characteristics such as temporal variations, as they introduce uncertainties in both the dataset and the modeling process. It is recommended to consider uncertainty analysis as it is an effective approach to understand the degree of variability associated with models and level of confidence in predictions (Lin et al., 2015). Improved ML based models could help to identify critical locations in the riverine conditions and combined with process-based hydrodynamical modelling, the habitat-abundance estimation could be done even in real-time including varying conditions in the riverine habitat. However, before reaching that analytical level and getting ML results to support decision making, it is important to identify and quantify the sources of uncertainty as they contribute to the variance of ecological predictions (Buisson et al., 2010). Therefore, a comprehensive examination of these factors will enhance the accuracy and robustness of modeling. In addition, machine learning methods often struggle to precisely model the habitat data, as noted by Crisci et al., (2012). Thus, it is crucial to exercise caution when utilizing these models.

5. Conclusion

Recognizing the complex and non-linear nature of habitat-abundance modeling, we employed ML techniques to model the juvenile salmon abundance in Teno catchment, using a relatively small habitat dataset. A comparison between regression and classification models revealed that the relationship between the habitat dataset and juvenile salmon abundance is indeed intricate. Consequently, the regression techniques struggled to fit suitable models to the data, despite their inherent ability to handle complex non-linear relationships. Among the regression models, those incorporating substrates, shades, and vegetation demonstrated higher levels of accuracy. Notably, the support vector classification model outperformed the regression techniques in terms of modeling accuracy. Among the regression models, SVR demonstrates the highest performance.

This study provides insights into the challenges and potential of machine learning techniques for juvenile salmon habitat-abundance modeling in complex habitat environments. The findings emphasize the importance of considering the limitations of machine learning models, particularly in habitat contexts, and the need for further research to address temporal variations and improve the precision of habitat-abundance modeling. Such advancements will aid in the development of robust and reliable tools for fisheries management and conservation strategies, facilitating the sustainable management of Atlantic salmon populations and their habitats.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, Bähar Jelovica used chatGPT in order to improve the language and fluency of the text. After using this tool/service, Bähar Jelovica reviewed and edited the content as needed and takes full responsibility for the content of the publication.

CRediT authorship contribution statement

Bähar Jelovica: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jaakko Erkinaro:** Writing – review & editing, Validation, Data curation. **Panu Orell:** Writing – review & editing, Validation, Data curation. **Bjørn Kløve:** Writing – review & editing, Validation. **Ali Torabi Haghighi:** Writing – review & editing, Validation, Supervision, Funding acquisition, Conceptualization. **Hannu Marttila:** Writing – review & editing, Validation, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank OLVI Foundation (Grant No. 20230145). The study was supported by Academy of Finland funded EcoRiver (number: 323810) project and was part of Freshwater Competence Centre (FWCC) activities. We also thank the field crews of the former Finnish Game and Fisheries Research Institute (currently LUKE), especially Matti Kylmäaho, Jorma Ollila, Eero Niemelä, and Markku Julkunen.

Appendix

Appendix A

A1: Box plots of habitat variables and fish abundance shows the skewness and dispersions of all variables in the habitat dataset.

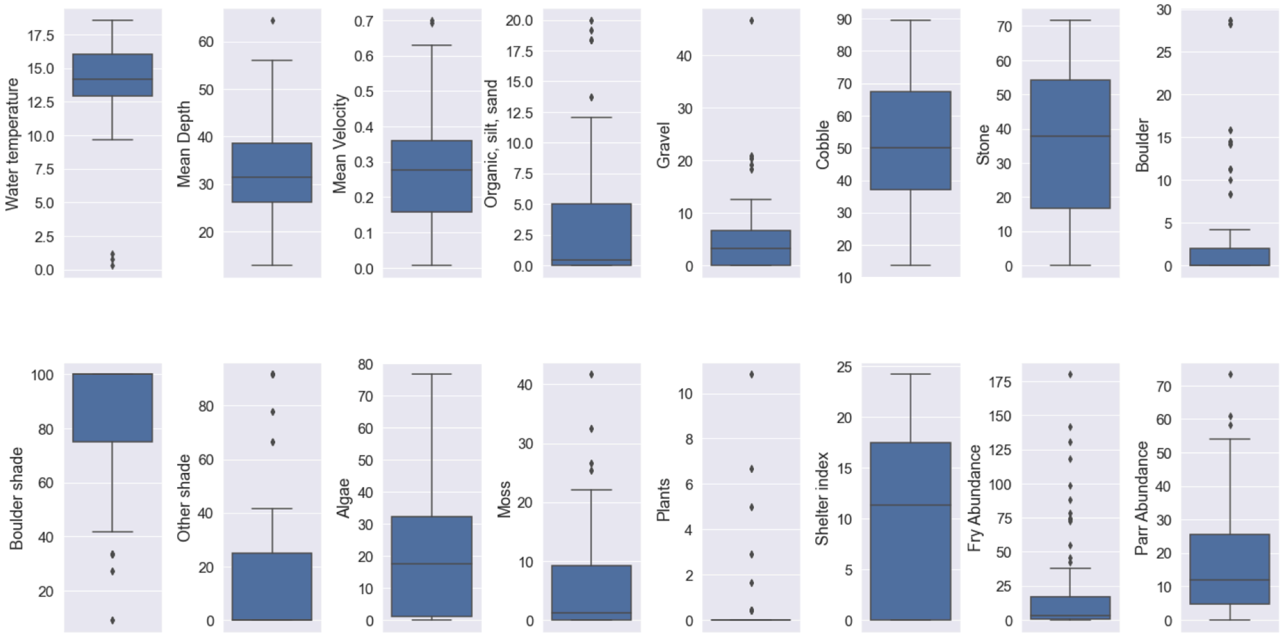


Figure A1. Boxplots of the dataset to recap the skewness, dispersion, and outliers in the habitat data and juvenile salmon abundance.

A2: Distribution of various substrate types and the corresponding juvenile salmon abundance in each sampling site in the study area – cobble and stone are the most observed substrates.

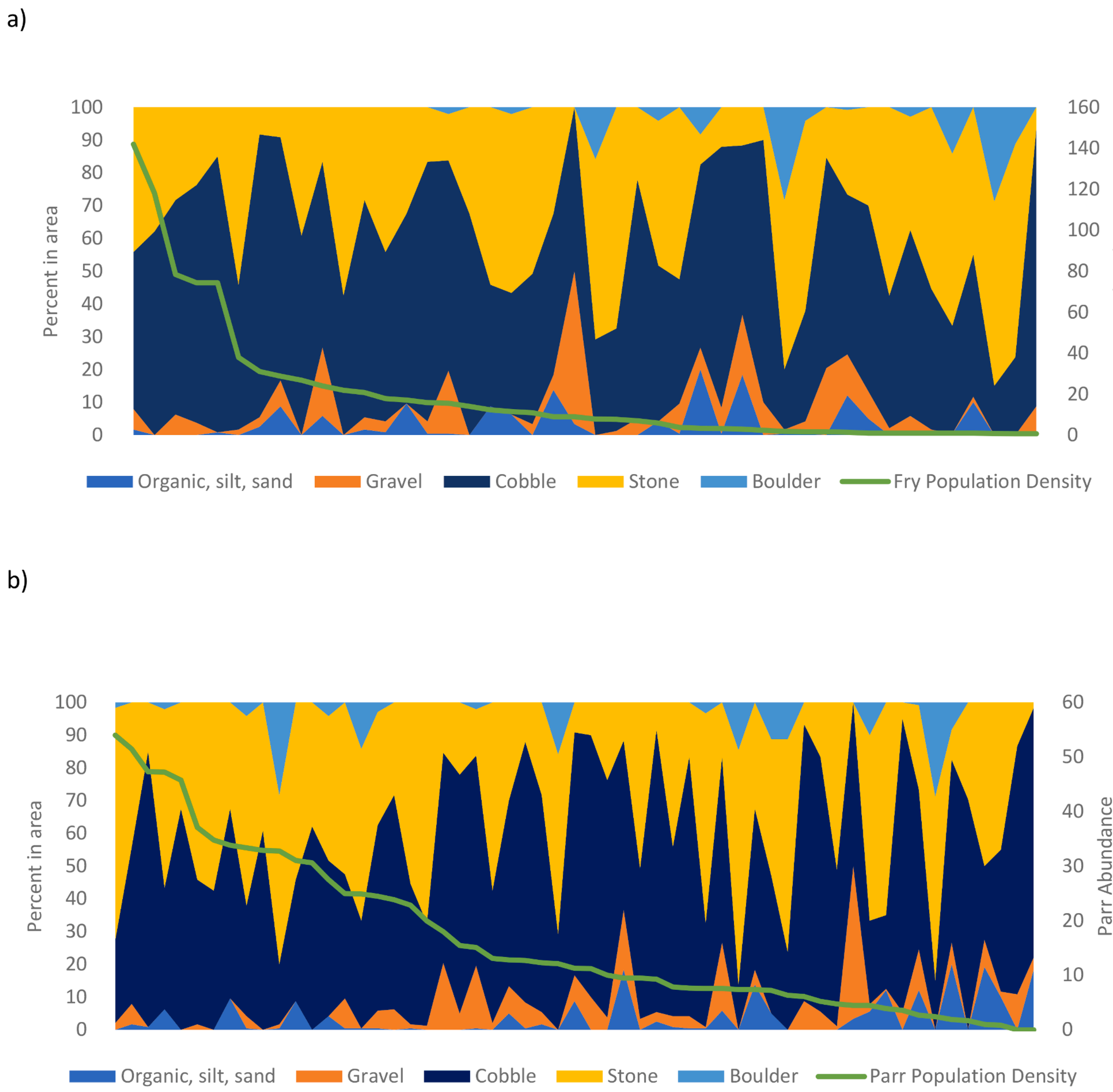


Figure A2. Substrate type distributions and juvenile salmon abundance per sampling site (a: Fry abundance, b: Parr abundance).

References

Abobakr Yahya, A.S., Ahmed, A.N., Binti Othman, F., Ibrahim, R.K., Afan, H.A., El-Shafie, A., Fai, C.M., Hossain, M.S., Ehteram, M., Elshafie, A., 2019. Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios. Water 11, 1231. https://doi.org/10.3390/w11061231.

Ackerly, D.D., Cornwell, W.K., Weiss, S.B., Flint, L.E., Flint, A.L., 2015. A geographic mosaic of climate change impacts on terrestrial vegetation: which areas are most at risk? PLoS One 10 (6), e0130629.

Adachi, N., Yoshida, Y., 1995. Accelerating genetic algorithms: protected chromosomes and parallel processing. Proceedings of the first international conference on genetic algorithms in engineering systems: innovations and applications, 1-20. 10.1049/cp:19951028.

Ahmadi, F., Mehdizadeh, S., Mohammadi, B., 2021. Development of Bio-Inspired- and Wavelet-Based Hybrid Models for Reconnaissance Drought Index Modeling. Water 35, 4127–4147. https://doi.org/10.1007/s11269-021-02934-z.

Amit, Y., Geman, D., 1997. Shape Quantization and Recognition with Randomized Trees. Neural Comput. 9, 1545–1588. https://doi.org/10.1162/neco.1997.9.7.1545.

Armstrong, J.D., Kemp, P.S., Kennedy, G.J.A., Ladle, M., Milner, N.J., 2003. Habitat requirements of Atlantic Salmon and brown trout in rivers and streams. Fish. Res. 62, 143–170. https://doi.org/10.1016/S0165-7836(02)00160-1.

Auerbach, D.S., Fremie, A.K., 2022. Identification of salmon redds using RPV-based imagery produces comparable estimates to ground counts with high inter-observer variability. River Res. Appl. 39, 35–45. https://doi.org/10.1002/rra.4065.

Binns, N.A., Eiserman, F.M., 1979. Quantification of fluvial trout habitat in Wyoming. Transaction of the American Fisheries Society 108, 215–228. https://doi.org/10.1577/1548-8659(1979)108<215:QOFTHI>2.0.CO;2.

Breiman, L., 1996. Bagging Predictors. Machine Learning 24, 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Bühlmann, P., Hothorn, T., 2007. Boosting Algorithms: Regularization, Prediction and Model Fitting. Stat. Sci. 22, 477–505. https://doi.org/10.1214/07-STS242.

Buisson, L., Thuiller, W., Casajus, N., Lek, S., Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Glob. Chang. Biol. 16, 1145–1157. https://doi.org/10.1111/j.1365-2486.2009.02000.x.

Callens, A., Morichon, D., Abadie, S., Delpey, M., Liquet, B., 2020. Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. Appl. Ocean Res. 104 https://doi.org/10.1016/j.apor.2020.102339.

Cameron, E.K., Sundqvist, M.K., Keith, S.A., CaraDonna, P.J., Mousing, E.A., Nilsson, K. A., Metcalfe, D.B., Classen, A.T., 2019. Uneven global distribution of food web studies under climate change. Ecosphere 10. https://doi.org/10.1002/ecs2.2645.

Chapelle, O., Vapnik, V., Bengio, Y., 2002. Model selection for small sample regression. Mach. Learn. 48, 9–23.

Crisci, C., Ghattas, B., Perera, G., 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecol. Model. 240 https://doi.org/10.1016/j.ecolmodel.2012.03.001.

Danandeh Mehr, A., Erkinaro, J., Hjort, J., Torabi Haghighi, A., Ahrari, A., Korpisaari, M., Kuusela, J., Dempson, B., Marttila, H., 2022. Factors affecting the presence of Arctic charr in streams based on a jittered binary genetic programming model. Ecol. Ind. 142 https://doi.org/10.1016/j.ecolind.2022.109203.

Davoudi Moghaddam, D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Torabi Haghighi, A., Asadi Nalivan, O., Tien Bui, D., 2020. The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. Catena 187. https://doi.org/10.1016/j.catena.2019.104421.

De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Ding, S., Qi, B., Tan, H., 2011. An overview on theory and algorithm of support vector machines. Journal of Electron Science Technology of China 40, 2–10. https://doi.org/10.1007/s10462-012-9336-0.

Ditria, E.M., Sievers, M., Lopez-Marcano, S., Jinks, E.L., Connolly, R.M., 2020. Deep learning for automated analysis of fish abundance: the benefits of Training across multiple habitats. Environ. Monit. Assess. 192 https://doi.org/10.1007/s10661-020-08653-z.

Erkinaro, J., Czorlich, Y., Orell, P., Kuusela, J., Länsman, M., Falkegård, M., Pulkkinen, H., Primmer, C., Niemelä, E., 2019. Life history variation across four decades in a diverse population complex of Atlantic Salmon in a large subarctic river. Can. J. Fish. Aquat. Sci. 76, 42–55. https://doi.org/10.1139/cjfas-2017-0343.

Fan, J., Wu, J., Kong, W., Zhang, Y., Li, M., Zhang, Y., Meng, W., Zhang, A.M., 2017. Predicting Bio-indicators of aquatic ecosystems using the support vector machine model in the Taizi River. China. Sustainability 9, 892. https://doi.org/10.3390/su9060892.

Fayed, H., Atiya, A., 2019. Speed up grid-search for parameter selection of support vector machines. Appl. Soft Comput. 80, 202–210. https://doi.org/10.1016/j.asoc.2019.03.037.

Ficsór, M., Csabai, Z., 2023. Machine learning model ensemble based on multi-scale predictors confirms ecological segregation and accurately predicts the occurrence of net-spinning caddisfly larvae species groups (Trichoptera: Hydropsychidae) at catchment-scale. Ecol. Ind. 146 https://doi.org/10.1016/j.ecolind.2022.109769.

Finstad, A.G., Einum, S., Forseth, T., Ugedal, O., 2007. Shelter availability affects behaviour, size-dependent and mean growth of juvenile Atlantic Salmon. Freshw. Biol. 52, 1710–1718. https://doi.org/10.1111/j.1365-2427.2007.01799.x.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

García Nieto, P.J., García-Gonzalo, E., Lasheras, F.S., Alonso Fernández, J.R., Muñiz, C. D., de Cos Juez, F.J., 2018. Cyanotoxin level prediction in a reservoir using gradient boosted regression trees: a case study. Environ. Sci. Pollut. Res. 25, 22658–22671. https://doi.org/10.1007/s11356-018-2219-4.

García Nieto, P.J., García-Gonzalo, E., Alonso Fernández, J.R., Muñiz, C.D., 2021. Modeling algal atypical proliferation in La Barca reservoir using L-SHADE optimized gradient boosted regression trees: a case study. Neural Comput. & Applic. 33, 7821–7838. https://doi.org/10.1007/s00521-020-05523-0.

Giorgio, A., Bonis, S.D., Guida, M., 2016. Macroinvertebrate and diatom communities as indicators for the biological assessment of river Picentino (Campania, Italy). Ecol. Ind. 64, 85–91. https://doi.org/10.1016/j.ecolind.2015.12.001.

Granata, F., Papirio, S., Esposito, G., Gargano, R., De Marinis, G., 2017. Machine learning algorithms for the forecasting of wastewater quality indicators. Water 105. https://doi.org/10.3390/w9020105.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x.

Guo, C., Lek, S., Ye, S., Li, W., Liu, J., Li, Z., 2015. Uncertainty in ensemble modelling of large-scale species distribution: effects from species characteristics and model techniques. Ecol. Model. 306, 67–75. https://doi.org/10.1016/j.ecolmodel.2014.08.002.

Guo, C., Chen, Y., Liu, H., Lu, Y., Qu, X., Yuan, H., Lek, S., Xie, S., 2019. Modelling fish communities in relation to water quality in the impounded lakes of China's South-to-North Water Diversion Project. Ecol. Model. 397, 25–35. https://doi.org/10.1016/j.ecolmodel.2019.01.014.

Heggenes, J., Saltveit, S.J., 1990. Seasonal and spatial microhabitat selection and segregation in young Atlantic Salmon, Salmo salar L., and brown trout, Salmo trutta L., in a Norwegian river. J. Fish Biol. 36, 707–720. https://doi.org/10.1111/j.1095-8649.1990.tb04325.x.

Hiedanpää, J., Saijets, J., Jounela, P., Jokinen, M., Sarkki, S., 2020. Beliefs in Conflict: The Management of Teno Atlantic Salmon in the Sámi Homeland in Finland. Environ. Manag. 66, 1039–1058. https://doi.org/10.1007/s00267-020-01374-6.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844. https://doi.org/10.1109/34.709601.

Ho, T.K., 1995. Random decision forests. In Document analysis and recognition. Proceedings of the third international conference, Montreal, Quebec, Canada 1, 278–282. 10.1109/ICDAR.1995.598994.

Hoang, T.H., Lock, K., Mouton, K., Goethals, L.M., P.,, 2010. Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. Eco. Inform. 5, 140–146. https://doi.org/10.1016/j.ecoinf.2009.12.001.

Ices, 2023. Working Group on North Atlantic Salmon (WGNAS). ICES Scientific Reports. https://doi.org/10.17895/ices.pub.22743713.

Isaak, D.J., Wollrab, S., Horan, D., Chandler, G., 2012. Climate change effects on stream and river temperatures across the northwest US from 1980–2009 and implications for salmonid fishes. Clim. Change 113, 499–524. https://doi.org/10.1007/s10584-011-0326-z.

Jelovica, B., Marttila, H., Ashraf, F.B., Kløve, B., Torabi Haghighi, A., 2022. A probability-based model to quantify the impact of hydropeaking on habitat suitability in rivers. River Res. Appl. 39, 490–500. https://doi.org/10.1002/rra.4050.

Kalita, D.J., Singh, V.P., Kumar, V., 2023. A novel adaptive optimization framework for SVM hyper-parameters tuning in non-stationary environment: A case study on intrusion detection system. Expert Syst. Appl. 213 https://doi.org/10.1016/j.eswa.2022.119189.

Kang, H., Jeon, D.J., Kim, S., Jung, K., 2022. Estimation of fish assessment index based on ensemble artificial neural network for aquatic ecosystem in South Korea. Ecol. Ind. 136 https://doi.org/10.1016/j.ecolind.2022.109769.

Koster, E., Dankers, R., Linden, S.V.D., 2005. Water balance modelling of (Sub-) Arctic rivers and freshwater supply to the Barents Sea Basin. Permafr. Periglac. Process. 16, 249–259. https://doi.org/10.1002/ppp.510.

Landauer, M., Joona, J., Keskitalo, P., 2023. Stakeholder Perceptions of Landscape Justice in the Case of Atlantic Salmon Fishing in Northern Finland. Land 12, 1174. https://doi.org/10.3390/land12061174.

Leathwick, J.R., Elith, J., Francis, M., Hastie, T., 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Mar. Ecol. Prog. Ser. 321, 267–281. https://doi.org/10.3354/meps321267.

Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modeling of coastal algal blooms. Ecol. Model. 159, 179–201. https://doi.org/10.1016/S0304-3800(02)00281-8.

Lennox, R.J., Alexandre, C.M., Almeida, P.R., Bailey, K.M., Barlaup, B.T., Bøe, K., Breukelaar, A., Erkinaro, J., Forseth, T., Gabrielsen, S.-E., Halfyard, E., Hanssen, E. M., Karlsson, S., Koch, S., Koed, A., Langåker, R.M., Lo, H., Lucas, M.C., Mahlum, S., Perrier, C., Pulg, U., Sheehan, T., Skoglund, H., Svenning, M., Thorstad, E.B., Velle, G., Whoriskey, F.G., Vollset, K.W., 2021. The quest for successful Atlantic salmon restoration – perspectives, priorities, and maxims. ICES J. Mar. Sci. 78, 3479–3497. https://doi.org/10.1093/icesjms/fsab201.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017. Feature Selection: A Data Perspective. ACM Computing Survey 50, 1–45. https://doi.org/10.1145/3136625.

Li, X., Li, L., Wang, X., Lin, Q., Wu, D., Dong, Y., Han, S., 2021. Visual quality evaluation model of an urban river landscape based on random forest.Ecological Indicators 133. https://doi.org/10.1016/j.ecolind.2021.108381.

Lin, Y.P., Lin, W.C., Wu, W.Y., 2015. Uncertainty in various habitat suitability models and its impact on habitat suitability estimates for fish. Water 7, 4088–4107. https://doi.org/10.3390/w7084088.

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., Wei, Y., 2013. A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Math. Comput. Model. 58, 458–465. https://doi.org/10.1016/j.mcm.2011.11.021.

Mäki-Petäys, A., Huusko, A., Erkinaro, J., Muotka, T., 2002. Transferability of habitat suitability criteria of juvenile Atlantic Salmon (Salmo salar). Can. J. Fish. Aquat. Sci. 59, 218–228. https://doi.org/10.1139/f01-209.

Mäki-Petäys, A., Erkinaro, J., Niemelä, E., Huusko, A., Muotka, T., 2004. Spatial distribution of juvenile Atlantic Salmon (Salmo salar) in a subarctic river: size-specific changes in a strongly seasonal environment. Can. J. Fish. Aquat. Sci. 61, 2329–2338. https://doi.org/10.1139/f04-218.

Martínez-Santos, P., Aristizábal, H.F., Díaz-Alcaide, S., Gómez-Escalonilla, V., 2021. Predictive mapping of aquatic ecosystems by means of support vector machines and random forests. J. Hydrol. 595 https://doi.org/10.1016/j.jhydrol.2021.126026.

Matsuzawa, Y., Fukuda, S., Ohira, M., Baets, M.D., 2023. Modelling fish co-occurrence patterns in a small spring-fed river using a machine learning approach. Ecol. Ind. 151 https://doi.org/10.1016/j.ecolind.2023.110234.

McPherson, J., Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. Ecography 30, 135–151. https://doi.org/10.1111/j.0906-7590.2007.04823.

Mondal, R., Bhat, A., 2021. Comparison of regression-based and machine learning techniques to explain alpha diversity of fish communities in streams of central and eastern India. Ecol. Ind. 129 https://doi.org/10.1016/j.ecolind.2021.107922.

Mosavi, A., Ozturk, P., Chau, K.W., 2018. Flood Prediction Using Machine Learning Models: Literature Review. Water 10, 1536. https://doi.org/10.3390/w10111536.

Naghibi, A.A., Pourghasemi, H.R., Dixon, B., 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ. Monit. Assess. 188 https://doi.org/10.1007/s10661-015-5049-6.

Niemelä, E., Erkinaro, J., Julkunen, M., Hassinen, E., 2005. Is juvenile salmon abundance related to subsequent and preceding catches? Perspectives from a long-term monitoring programme. ICES J. Mar. Sci. 62, 1617–1629. https://doi.org/10.1016/j.icesjms.2005.07.002.

Olaya-Marín, E.J., Martínez-Capel, F., Vezza, P., 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowl. Manag. Aquat. Ecosyst. 409 https://doi.org/10.1051/kmae/2013052.

Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems 9, 181–199. https://doi.org/10.1007/s10021-005-0054-1.

Rahimian Boogar, A., Salehi, H., Pourghasemi, H.R., Blaschke, T., 2019. Predicting Habitat Suitability and Conserving Juniperus spp. Habitat Using SVM and Maximum Entropy Machine Learning Techniques. Water 11, 2049. https://doi.org/10.3390/w11102049.

Ritson, J.P., Graham, N.J.D., Templeton, M.R., Clark, J.M., Gough, R., Freeman, C., 2014. The impact of climate change on the treatability of dissolved organic matter (DOM) in upland water supplies: A UK perspective. Science of Total Environment 473–474, 714–730. https://doi.org/10.1016/j.scitotenv.2013.12.095.

Ro, K., Zou, C., Wang, Z., Yin, G., 2015. Outlier detection for high-dimensional data. Biometrika 102, 589–599. https://doi.org/10.1093/biomet/asv021.

Rosenfeld, J.S., Leiter, T., Lindner, G., Rothman, L., 2005. Food abundance and fish density alters habitat selection, growth, and habitat suitability curves for juvenile coho salmon (Oncorhynchus kisutch). Can. J. Fish. Aquat. Sci. 62, 1691–1701. https://doi.org/10.1139/f05-072.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdisciplin Rev Data Mining Knowledge Discovery 8. https://doi.org/10.1002/widm.1249.

Salman, R., Kecman, V., 2012. Regression as classification. 2012 Proceedings of IEEE Southeastcon, Orlando, FL, USA, 1-6. doi: 10.1109/SECon.2012.6196887.

Sanz-Garcia, A., Fernandez-Ceniceros, J., Antonanzas-Torres, F., Pernia-Espinoza, A.V., Martinez-de-Pison, F.J., 2015. GA-PARSIMONY: A GA-SVR approach with feature selection and parameter optimization to obtain parsimonious solutions for predicting temperature settings in a continuous annealing furnace. Appl. Soft Comput. 35, 13–28. https://doi.org/10.1016/j.asoc.2015.06.012.

Singh, K.P., Basant, N., Gupta, S., 2011. Support vector machines in water quality management. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2011.07.027.

Sivapalan, M., Blöschl, G., Merz, R., Gutknecht, D., 2005. Linking flood frequency to long-term water balance: Incorporating effects of seasonality. Water Resour. Res. 51 https://doi.org/10.1029/2004WR003439.

Torgo, L., Gama, J., 1996. Regression by classification. Advances in Artificial Intelligence - Lecture Notes in Computer Science 1159, 51–60. https://doi.org/10.1007/3-540-61859-7_6.

Tsai, L.T., Yang, C.C., 2012. Improving measurement invariance assessments in survey research with missing data by novel artificial neural networks. Expert Syst. Appl. 39, 10456–10464. https://doi.org/10.1016/j.eswa.2012.02.048.

Vähä, J.P., Erkinaro, J., Falkegård, M., Orell, P., Niemelä, E., 2017. Genetic stock identification of Atlantic Salmon and its evaluation in a large population complex. Can. J. Fish. Aquat. Sci. 74, 327–338. https://doi.org/10.1139/cjfas-2015-0606.

Vapnik, V., Chervonenkis, A., 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and Its Application 16, 264–280. https://doi.org/10.1137/1116025.

Vapnik, V.N., 1998. Statistical Learning Theory; Wiley: New York, NY, USA.

Vorpahl, P., Elsenbeer, H., Märker, M., Schröder, B., 2012. How can statistical models help to determine driving factors of landslides? Ecol. Model. 239, 27–39. https://doi.org/10.1016/j.ecolmodel.2011.12.007.

Welchowski, T., Maloney, K.O., Mitchell, R., Matthias, S., 2022. Techniques to improve ecological interpretability of Black-Box machine learning models. J. Agricultural, Biological and Environmental Statistics 27, 175–197. https://doi.org/10.1007/s13253-021-00479-7.

Wellmann, T., Lausch, A., Scheuer, S., Haase, D., 2020. Earth observation based indication for avian species distribution models using the spectral trait concept and machine learning in an urban setting. Ecol. Ind. 111 https://doi.org/10.1016/j.ecolind.2019.106029.

Woo, S.Y., Jung, C.G., Lee, J.W., Kim, S.J., 2019. Evaluation of watershed scale aquatic ecosystem health by SWAT modeling and random Forest technique. Sustainability 11, 3397. https://doi.org/10.3390/su11123397.

Xu, Y., Zhang, D., Lin, J., Peng, Q., Lei, X., Jin, T., Wang, J., Yuan, R., 2024. Prediction of phytoplankton biomass and identification of key influencing factors using interpretable machine learning models. Ecol. Ind. 158 https://doi.org/10.1016/j.ecolind.2023.111320.

Yang, Y., Xiong, Q., Wu, C., Zou, Q., Yu, Y., Yi, H., Gao, M., 2021. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. Environ. Sci. Pollut. Res. 28, 55129–55139. https://doi.org/10.1007/s11356-021-14687-8.

Yang, Z., Zhu, D., Zhu, Q., Hu, L., Wan, C., Zhao, N., Liu, H., Che, X., 2020. Development of new fish-based indices of biotic integrity for estimating the effects of cascade reservoirs on fish assemblages in the upper Yangtze River. China. Ecological Indicators 119. https://doi.org/10.1016/j.ecolind.2020.106860.