Water Supply

Water Supply Vol 22 No 5, 5355 doi: 10.2166/ws.2022.166

Extreme Learning Machine model for assessment of stream health using the Qualitative Habitat Evaluation Index

Ahmed S. Aredah [a,*], Omer Faruk Ertugrul[b], Ahmed A. Sattar[c,d], Hossein Bonakdari[e] and Bahram Gharabaghi[f]

[a] Civil & Environmental Engineering Department, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[b] Department of Electrical and Electronics Engineering, Batman University, Batman, Turkey
[c] Faculty of Engineering, Cairo University, Giza, Egypt
[d] German University in Cairo, Cairo, Egypt
[e] Department of Civil Engineering, , University of Ottawa, 161 Louis Pasteur Drive, Ottawa K1N 6N5, Canada
[f] School of Engineering, University of Guelph, NIG 2W1, Guelph, Ontario, Canada
*Corresponding author. E-mail: ahmedaredah@vt.edu

ASA, 0000-0003-0186-3783

ABSTRACT

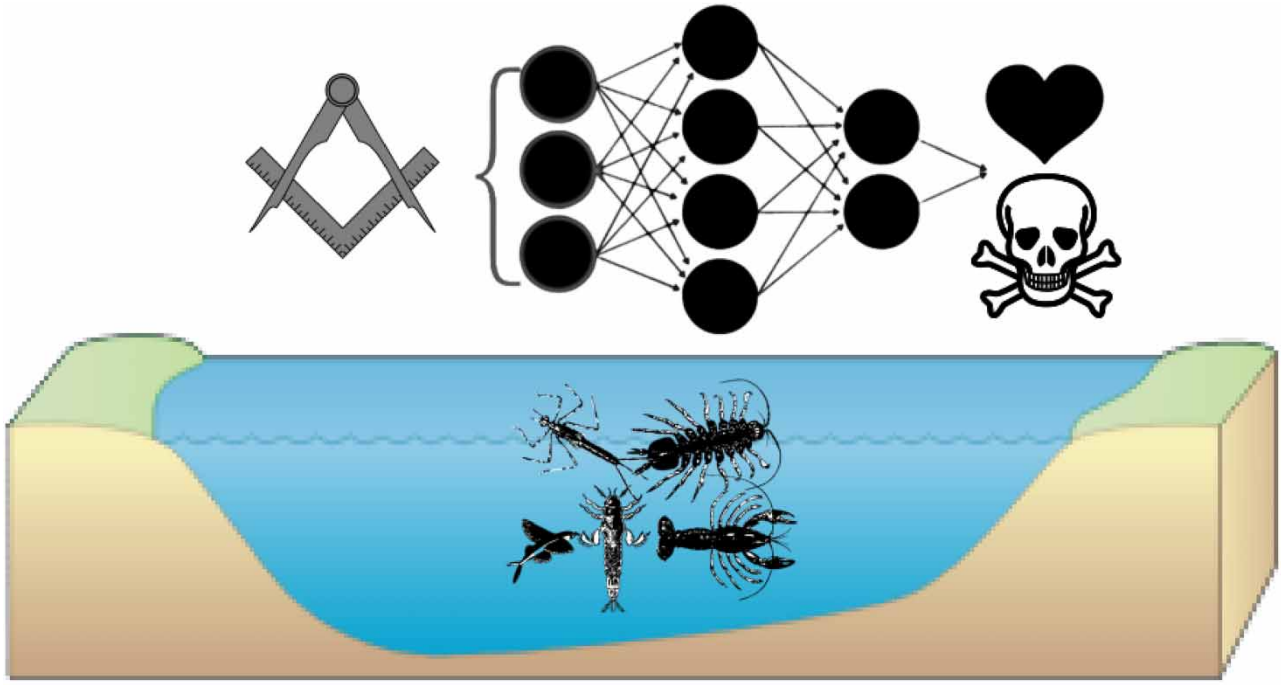
The Extreme Learning Machine (ELM) approach was used to predict stream health with a Qualitative Habitat Evaluation Index (QHEI), and watershed metrics. A dataset of 112 sites in Ontario, Canada with their Hilsenhoff Biotic Index (HBI) and richness values was used in the development of two ELM models. Each model used 70 and 30% of the dataset for training and testing respectively. The models show a great fit with Root Mean Square Error (RMSE)=0.12 and 0.33 for HBI and richness test models, respectively. Then, features elimination based on ELM coefficients and coefficient of variation showed a slight increase in the models' RMSE to reach 0.09 and 0.33 correspondingly. Accordingly, this high predictability of the models in this research provide better insights into which factors influence HBI or richness, and suggests that ELM has a better architecture than other machine learning models and ANN to learn complex non-linear relationships. Also, sensitivity analysis expressed channel slope as the most affecting stream-health parameter for stream health.

Key words: extreme learning machine (ELM), sensitivity analysis, stream restoration, watershed metrics

HIGHLIGHTS

- Evaluating stream health using the QHBI is an uncertain process.
- AI models capture this uncertainty and predict the affecting parameters on stream health.
- ELM-based models prove they can predict the most affecting QHEI parameters in stream health evaluation.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Rivers endure degradation during their lifespan as a result of direct and indirect stream stressors (Maddock 1999; Booth 2005; Nguyen *et al.* 2017). Those stressors that affect a stream are hierarchical, beginning with the watershed and progressing to the microhabitat scale. Hierarchical – physical, chemical, and biological – characteristics have been proved to affect stream habitat and fish communities (Booth 2005; Gazendam *et al.* 2011; Gorney *et al.* 2012). Commitment to minimizing the impacts of stream degradation on stream health is made to restore and mitigate systems to a near-natural state and improve the functionality of freshwater ecosystems (Kim *et al.* 2019). This is mainly required to support living systems (Karr 1999). Considering this, degraded streams should be restored to their original hydrological properties and their associated chemical, physical, and biological condition (Mrozińska *et al.* 2018).

Hence, stream restoration planning, testing, construction practices, and performance assessment strategies are being developed as a result of research and engineering activities to monitor and restore, if needed, the health of stream systems (Mrozińska *et al.* 2018; Schwindt *et al.* 2019). Accordingly, the development of indices was necessary to evaluate stream health and their ecological integrity. Indices may be biological, physical, chemical, or a mix of any of these (Gazendam *et al.* 2011). Herman & Nejadhashemi (2015) stated that chemical integrity has traditionally been the most widely used method for evaluating stream health; however, it has recently been realized that using biological integrity can lead to a better knowledge of what is happening in the ecosystem and identify the source of degradations.

Several studies have suggested that the presence of various types of aquatic organisms is a dominant component in assessing a system (Davies *et al.* 2000; Suen 2009; Gazendam *et al.* 2011; Goetz & Fiske 2013; Lee & An 2014; Roy *et al.* 2014; Woznicki *et al.* 2015; Nguyen *et al.* 2017; Mrozińska *et al.* 2018). Also, biological indicators can account for physical and chemical properties as well as biological aspects. Further, the presence of fish communities and macroinvertebrates can reflect various anthropogenic disturbances, including nutrient enrichment, toxic pollution, flow regime, and physical habitat alteration and fragmentation (Lee & An 2014). In addition, Argerich *et al.* (2004) discussed that changing environmental conditions rapidly change the biotic community, and hence key environmental variables that characterize environmental conditions should be quantified.

In addition, as stated by Karr (1999) and summarized by Hernandez-Suarez & Nejadhashemi (2018), stream health indices are categorized into biotic, multimetric indices, and multivariate methods. For biotic indices, one metric is involved to explain

stream health, while multimetric indices are leveraged by several metrics (Herman & Nejadhashemi 2015). Biotic metrics, such as the Hilsenhoff Biotic Index (HBI) and richness, rely on an individual metric to describe the abundance, biomass, and condition of macroinvertebrates and fish species, wealth, and composition of species, or the trophic composition. This accordingly could be used to show organisms' tolerance to pollution and thus allow for the identification of regional degradations (Ollis *et al.* 2006; Herman & Nejadhashemi 2015).

Macroinvertebrate-based metrics are better than fish-based metrics in local site water health control because of macroinvertebrates' limited mobility in the stream channel. Accordingly, they are beneficial for determining local causes of degradation. Also, macroinvertebrates are sensitive to low contamination levels which allows for early detection of stream degradation (Herman & Nejadhashemi 2015). The HBI, a macroinvertebrate-based metric, (Hilsenhoff 1988) is an abundance-weighted tolerance index calculated using each macroinvertebrate taxon's tolerance value. The index values range from 0 to 10, with higher measurements indicating more degraded water quality (Wang *et al.* 2007; Gazendam *et al.* 2011; Einheuser *et al.* 2012; Herman & Nejadhashemi 2015; Woznicki *et al.* 2015).

On the other hand, multimetric indices, such as the Qualitative Habitat Evaluation Index (QHEI), use multiple metrics to evaluate stream health. By accounting for multiple metrics and increased complexity, a more comprehensive view of what is happening within streams can be made (Rakocinski 2012; Herman & Nejadhashemi 2015). The QHEI is a physical-geomorphic integrity measurement for a stream. Quality of habitat is shown in the QHEI as the sum of a series of visually assessed measurements that take account of the integrity of six stream components: substrate, in-stream cover, channel morphology, riparian zone, bank erosion, riffle quality, and stream gradient. Because calculating the QHEI depends solely on visual inspection, unlike biological indices, it is recognized as a quick and comprehensive tool that allows for rapid evaluation of stream systems at a reach scale (Gazendam *et al.* 2011).

Biotic and multimetric indices are generally restricted to a small area of a watershed where data are collected (Woznicki *et al.* 2015). Multivariate approaches are introduced to model and tie observable organisms to the physical and chemical stream characteristics. After the models are created, they could be used to expand the study beyond the sample sites in the stream. For this, multivariate approaches are beneficial to identify deteriorated areas. The resultant model can be complex and its predictions are uncertain. This is because stream habitat communities respond to multiple complex and nonlinearly related stressors. Therefore, in conjunction with multimetric and biotic indices, multivariate approaches are recommended for the assessment of stream quality (Wang *et al.* 2008; Herman & Nejadhashemi 2015; Woznicki *et al.* 2015).

Because ecosystem and watershed managers need simplified and standardized methods for assessing ecological conditions (Angermeier & Karr 2018), an integrated multimetric model approach is used to assess how far a system's condition diverges from integrity (Woznicki *et al.* 2015; Angermeier & Karr 2018; Simões *et al.* 2020). Research has examined the interrelationships between stream habitat and physical parameters utilizing macroinvertebrates, and fish indices. The development of such has been investigated through the involvement of stream habitats and physical-geomorphic factors. Barbour *et al.* (1999), Rankin *et al*. (1999), and Wang *et al*. (2007) concluded that there is a strong correlation between the quality of habitat and certain biologic indices. Gazendam *et al.* (2011) proved that there is a good correlation between QHEI and HBI and taxa richness using Multilinear regression (MLR). The resultant MLR model fit could reach a maximum of 64%. Despite the straightforward implementation and interpretation of MLR, the model parameter estimation is unstable under the multicollinearity and strongly correlated variables in the inspected metrics, and the prediction power is very low. Further, because MLR's main assumption is an independent and identically distributed (IID) error, this assumption is barely met, hence other linear statistical methods with some flexibility (Generalized Linear Models, GLM, and Generalized Adaptive Models, GAM) have been introduced and the resultant model predictivity has been improved (Sauer *et al.* 2011; Clapcott *et al.* 2012; Damanik-Ambarita *et al.* 2016). In the GLM model developed by Damanik-Ambarita *et al.* (2016), the authors could reach a goodness of fit of 57%.

Still, linear statistical models are used as comparison benchmarks with other statistical methods and machine learning approaches (Hernandez-Suarez & Nejadhashemi 2018). Other researchers went beyond MLR and used advanced models with multiple physical and chemical stressors to predict benthic macroinvertebrates abundance and therefore stream quality. This model advancement is to increase the model predictability and reduce its uncertainty. Hernandez-Suarez & Nejadhashemi (2018) summarized all the models used to investigate the effect of physical and environmental factors on ecosystem communities and hence their health.

Because linear statistical models do not eliminate less significant variables, their interpretability is often reduced when there are huge numbers of predictors and model parameters. This might cause redundancy, low bias, significant variation, and reduced model accuracy. A decent way to reduce model parameters is the shrinkage variable selection. The Least Absolute

Selection Shrinkage Operator (LASSO) is a shrinkage technique used to select significant variables and reduce other weights (regularization) (Hastie *et al.* 2009). LASSO was employed by Berger *et al.* (2017) to characterize the link between six macroinvertebrate indices and environmental factors such as water quality, land use, and wastewater exposure. Results showed a great reduction in variables yet kept good predictability. The resultant R^2 of these models reached a maximum value of 60%.

Decision-tree models, boosting regression trees (BRT) and random forest (RF), are another approach for investigating interrelationships between variables. In BRT, boosting is an optimizer which reduces the difference between calculated and observed values and adds a new tree at each stage to minimize the difference. Many researchers used BRT to predict macroinvertebrate taxa richness and composition (Booker *et al.* 2015; Álvarez-Cabria *et al.* 2017). Booker *et al.* (2015) studied using RF biological distribution modeling methods to examine how the structure of macro-invertebrate communities is connected to hydrological regimes at the new Zealand nationwide scale. The model goodness of fit could reach as high as 63.9% by taxon richness, macroinvertebrate community index, and percentage Ephemeroptera, Plecoptera, and Trichoptera.

Because decision trees consume too much time and their predictivity is not even very high, Artificial Neural Networks (ANN) are introduced. ANN have been applied to model and predict complex and nonlinear environmental systems. Gazendam *et al.* (2016) used an ANN model. The research obtained a relationship between the physical characteristics and HBI and richness metrics. The resultant model reached an accuracy as high as 92%. Other advanced models have been considered such as fuzzy logic models, Support Vector Regression, and Bayesian belief networks models. Van Broekhoven *et al.* (2006) developed a fuzzy model to study the stability of macroinvertebrates, and the percentage of correctly fuzzy classified instances (% CFCI), a performance measure similar to R^2, reached a value of 66%. In addition, Hoang *et al.* (2010) studied the presence of macroinvertebrates within multiple stressors; the authors could reach a percentage of Correctly Classified Instances (CCI) of 83%. Mantyka-Pringle *et al.* (2017) evaluated, using traditional and scientific knowledge, the cumulative environmental effects of multiple stressors on ecosystem health. The research included biotic variables defining the health of wildlife, food webs, fish, and macro-invertebrates (density, richness, and diversity).

Although the presence and abundance of biological communities species are known to have many environmental determinants, the close relationship between factors makes it difficult to identify which factors play a dominant role in the preservation of the structure of a fish community (Suen 2009). This was a motivation to many scholars to study different modeling approaches to predict stream health indices in water systems. Previous studies tried to apply different mathematical models to physical or chemical variables to reach a high predictivity value with low robustness and less variance output. In some studies, the high number of model parameters has been reduced by regularization. Accordingly, some of these models did not show a very high predictivity power.

The Extreme Learning Machine (ELM) approach, a relatively new approach adopted in machine learning, (Huang *et al.* 2006) has acquired tremendous interest. Including hyperspectral and uncertain data and features selection, it could show efficient capabilities. ELM is robust and rapid since it follows the framework of a single hidden layer multi-layered perceptron. The learning stage of conventional artificial neural networks focuses on weight and bias optimization utilizing different gradient-based or other sophisticated learning methods. Contrary to this optimization methodology, in which it may be trapped in local minima and hence have difficulties converging, ELM initiates its weights randomly between the input and the hidden layers. Then, it optimizes the weight vector between the hidden and the output layers as a simple mathematical problem. The key gain of such an algorithm is the training speed and the capability to learn complex non-linear behavior. This approach helps the system to learn any data with high-precision output (Huang *et al.* 2006; Leuenberger & Kanevski 2015).

This paper aims at applying ELM to develop new predictive empirical models for the prediction of HBI and richness in watersheds. First, a brief overview of the QHEI, HBI, and richness indices and adopted models in predicting taxa richness is presented in addition to describing the parameters of the metrics parameters and categories included in the collected database. Then, ELM is applied to the database and four models are developed. Also, error and uncertainty analyses are performed on the developed models to assess their reliability and robustness. Finally, the developed models undergo sensitivity analysis to test their physical behavior.

2. METHODOLOGY

2.1. Data set description

A dataset gathered by Gazendam *et al.* (2016) was utilized to investigate the influence of various variables on HBI and richness. The dataset contains 112 sites in Southern Ontario with different watershed characteristics. The site areas vary within

the range of 1 to 1,000 ha and the bankfull width ranged from 1 to 50 m. Adjacent land uses to the sites were diversified (e.g. agricultural and urban). The local stream reach was examined with reach lengths from 25 to 500 m.

A Qualitative Habitat Evaluation Index (QHEI) was used and calculated at each site to measure the stream health based on 6 metrics (and 31 metric components) (as shown in Table 1): (1) substrate; (2) stream cover; (3) channel morphology; (4) riparian zone and bank erosion; (5) pool and riffle quality; (6) stream gradient. These metrics and measurements show almost all types of site characteristics have been captured in this study and thus this research hypothetically applies to all stream conditions.

As a measure of these parameter changes in stream system health, various benthic metrics have been used by different researchers. These metrics include richness, percentage Ephemeroptera, Plecoptera and Trichoptera (%EPT), diversity, and HBI. In this research, only HBI and richness benthic metrics were used, as studied in Gazendam *et al.* (2016).

To examine the data and its distribution, the violin plot is used in this research to describe the dataset used visually. The violin plots produced are a combination of a box plot and the density distribution and, accordingly, they reveal the data structure (Hintze & Nelson 1998).

2.2. ELM model development

Since stream health monitoring involves many aspects and parameters to consider and collecting these measurements is not always easy to obtain, easier methods are proposed. These methods test the relationship between different parameters and obtain a prediction of the stream health with fewer yet effective variables.

Table 1 | Utilized QHEI metrics and components summary

Metric	Metric Component	Mean	Std Dev	Min	Max
Substrate	Substrate Type (X1)	11.8	3.1	1.5	17.0
	Best Types (X2)	0.7	1.0	0.0	2.0
	Substrate Origin (X3)	0.7	0.3	0.0	1.0
	Substrate Quality (X4)	−1.1	1.5	−4.0	2.0
Stream Cover	Cover Type (X5)	5.4	1.8	1.0	10.0
	Cover Amount (X6)	5.4	2.5	1.0	11.0
Channel Conditions	Sinuosity (X7)	2.6	0.7	1.0	4.0
	Channel Development (X8)	4.0	1.5	1.0	7.0
	Channelization (X9)	4.9	1.5	1.0	6.0
	Channel Stability (X10)	1.9	0.7	1.0	3.0
	Erosion (X11)	2.5	0.5	1.0	3.0
Floodplain	Riparian Width (X12)	2.5	1.4	0.0	4.0
	Floodplain Quality (X13)	1.5	1.2	0.0	3.0
Geomorphic	Pool Depth (X14)	2.7	1.6	0.0	6.0
	Channel Width (X15)	1.7	0.5	0.0	2.0
	Velocity Types (X16)	2.7	0.9	1.0	5.0
	Riffle Depth (X17)	1.5	0.6	0.0	2.0
	Run Depth (X18)	1.1	0.5	0.0	2.0
	Riffle/Run Substrate (X19)	1.0	0.8	0.0	2.0
	Riffle/Run Embeddedness (X20)	0.8	0.9	−1.0	2.0
	Channel Slope (X21)	6.6	2.6	2.0	10.0
Watershed	Drainage Area (X22)	4,197	8,023	47	61,311
	Shape Factor (X23)	11.7	6.9	3.1	37.6
	Slope of Main Channel (X24)	0.7	0.6	0.1	3.1
	% Water (X25)	8.8	7.9	0.0	34.7
	% Treed (X26)	14.5	15.8	0.2	81.1
	% Community (X27)	11.0	18.5	1.2	88.2
	% Agriculture/Rural (X28)	65.2	25.0	4.8	96.7
Hydrology	BFI (X29)	0.5	0.2	0.2	0.7
	Flood Flow 1:2 Yr (X30)	9.2	12.9	0.2	93.1
	Low Flow - 7Q2 (X31)	0.3	0.2	0.0	1.0

This research suggests using ELM over ANN to improve the linkage between the 31 different factors to see which parameters most influence HBI or richness, and compares it to early studies by Gazendam *et al.* (2016). The ELM model was developed in Matlab (Figure 1).

Other models are also used as a bottom line to test our main models. These models include Linear Regression (LR), Support Vector Regression (SVR), nu-Support Vector Machine (nu SVM), RIDGER, LASSOR, PINVR, and Partial Least Square Regression (PLSR). These selected models have been utilized due to their simplicity compared to ELM. This simplicity may not reflect non-linearity capture in the data yet is still used as an efficiency indicator for the ELM models.

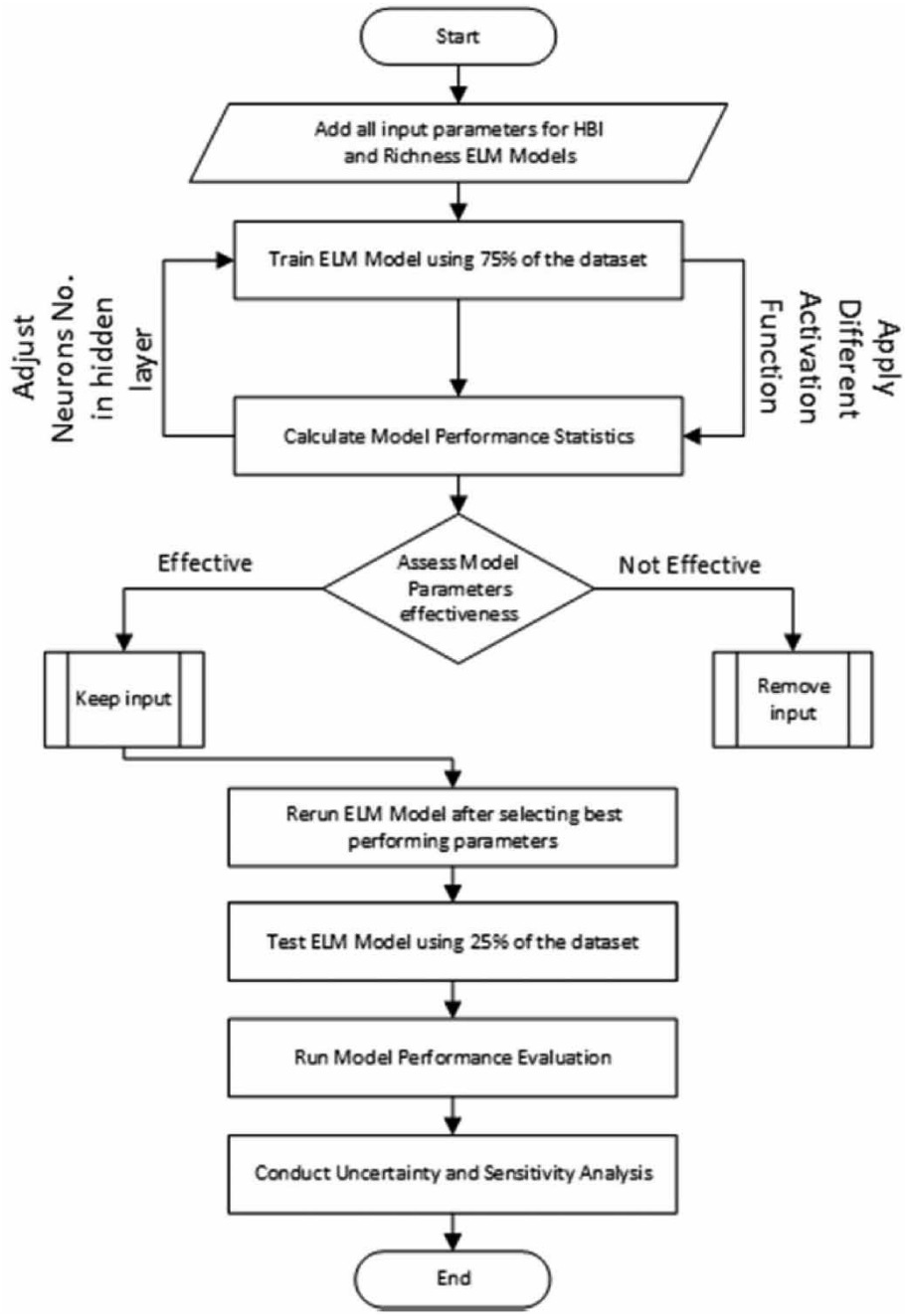


Figure 1 | ELM Model Development And Followed Methodology.

2.2.1. ELM neural network

In this section, we discuss the ELM neural network introduced by (Huang *et al.* 2006). This approach has lately been widely employed in engineering research due to its great ability to depict non-linear relationships. The ELM method utilizes the least-square training technique for the development of a single layer forward network, unlike the traditional gradient optimization approach used by other ANN. The ELM is made of three layers: one input layer, one hidden layer, and one output layer.

The main benefits of utilizing the least-square training technique are: (1) very fast algorithm compared to the traditional gradient-based algorithms, (2) it has better generalization performance than gradient-based learning, and (3) traditional artificial neural network drawbacks have been resolved such as local minima, improper learning rate and overfitting, and ELM tends to reach solutions straightforwardly.

In the ELM procedure, the weights of the hidden layer (w_{ij}) are randomly initialized, so that the training process just determines the weights of the output layer (β_{jk}) (Huang *et al.* 2006). Therefore, ELM is a fast neural training network that is particularly well suited to the problem involved. Similar to traditional ANN, ELM has a fully connected layer structure between the input-to-hidden and hidden-to-output layers.

The number of input attributes (n) is mapped to input layer nodes and likewise, the output nodes (m) represent the variables of the considered problem. In between the input and output layers, the ELM structure has a hidden layer in which the number of neurons varies. There is no rule to determine its count and perhaps the difficulty of the problem may determine it. In the current study, the trial and error method is adopted to identify the optimum hidden layer's neuron number (l).

The input-to-hidden weight matrix $w_{ij} \in \mathbb{R}^{n\times l}$ where each entry of this matrix is a coefficient that connects the i^{th} input layer neuron to the j^{th} hidden layer neuron. The weight matrix that connects the hidden layer to the output layer is $\beta_{ij} \in \mathbb{R}^{l\times m}$. This matrix maps the corresponding weight between the j^{th} hidden layer neuron to the k^{th} output layer neuron. In addition, the input matrix is $X \in \mathbb{R}^{n\times Q}$, and the target is $Y \in \mathbb{R}^{m\times Q}$ matrices of the problem, where Q is the input samples number. The results of the ELM are shown by $\mathrm{T}=(\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_Q)_{\mathrm{m}\times\mathrm{Q}}$, where each output is defined as follows:

$$\mathrm{t}_j = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{mj} \end{bmatrix}_{m\times 1} = \begin{bmatrix} \sum_{i=1}^{l} \beta_{i1} g(\mathrm{w}_i \mathrm{x}_i + b_i) \\ \sum_{i=1}^{l} \beta_{i2} g(\mathrm{w}_i \mathrm{x}_i + b_i) \\ \vdots \\ \sum_{i=1}^{l} \beta_{im} g(\mathrm{w}_i \mathrm{x}_i + b_i) \end{bmatrix}_{m\times 1}, j = 1, 2, \ldots, Q \tag{1}$$

where $g(x)$ is the neural network activation function, so that the ELM could be shown by the following equation:

$$\mathrm{H}\beta = \mathrm{T}^T \tag{2}$$

where

$$H = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_l \cdot \mathbf{x}_1 + b_l) \\ \vdots & \cdots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_Q + b_1) & \cdots & g(\mathbf{w}_l \cdot \mathbf{x}_Q + b_l) \end{bmatrix}_{Q\times l} \tag{3}$$

The results of ELM are exact when the hidden layer neurons number (l) is equal to the number of problem samples (Q). However, this research model is probably trapped in the over-fitting. Thus, the model has an error that is defined by:

$$\sum_{j=1}^{Q} \left\| \mathrm{t}_j - y_j \right\| < \varepsilon \tag{4}$$

where ε is always bigger than zero.

In the training stage, after randomly selecting w's and b's, the β matrix is obtained by using the $\min_{\beta} \left\| \mathrm{H}_{Q\times l}\beta_{l\times m} - \mathrm{T}^T_{Q\times m} \right\|$. The solution to this equation is obtained by $\hat{\beta} = \mathrm{H}^{\dagger}\mathrm{T}^T$ where $\mathrm{H}^{\dagger}$ is the Moore–Penrose generalized inverse of matrix H.

Other machine learning models have been adopted as a bottom line for our model's performance. Linear and non-linear relationships between variables and HBI and richness have been considered through the following Machine Learning models: Linear Regression (LR), Support Vector Regression (SVR), nu-Support Vector Machine (nu SVM), RIDGER, LASSOR, PINVR, and Partial Least Square Regression (PLSR). These models have been utilized due to their simplicity and their reputation among scholars.

The reason for choosing these methods for comparison is because they are known for simplicity. LR for instance is used for its simplicity to capture the direct-linear relationship between features and the target using the least square methodology and hence scholars are using it as the bottom line for performance measurement. In contrast to that, SVR and its derived models (nu SVM) are more flexible in terms of how much error is acceptable in the model, as described in the literature section.

2.2.2. Features selection in ELM

Variable and feature selection can help with data visualization and interpretation, as well as minimizing measurement and storage needs, training and utilization times, and overcoming the curse of dimensionality to increase prediction performance (Guyon & Elisseeff 2003). Features selection approaches are classified into filtering, wrapping, and embedding methods. Filtering methods depend on sorting variables based on a calculated score far from any adopted model (e.g. regression model) and thus do not provide the best-affecting features. Wrapping methods, on the other hand, measure how each feature affects the output and hence can provide the best-describing features; however, their computational power is very high. In embedded methods, the feature is determined based on it's coefficients in the adopted models. This method's selected features are highly dependent on the adopted model itself.

Ertugrul & Taugluk (2017) proposed a method to overcome all these challenges. The authors' method is to make use of the high-accuracy-ELM-model speed to compute each feature weight and then normalize it by the coefficient of variation. The coefficients used in this method are altered by applying the activation function to both terms of the ELM predicted value in Equation (1) to be

$$\mathrm{t}_j = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{mj} \end{bmatrix}_{m\times 1} = \begin{bmatrix} \sum_{i=1}^{l} \beta_{i1}(\mathrm{w}_i g(\mathrm{x}_i) + g(b_i)) \\ \sum_{i=1}^{l} \beta_{i2}(\mathrm{w}_i g(\mathrm{x}_i) + g(b_i)) \\ \vdots \\ \sum_{i=1}^{l} \beta_{im}(\mathrm{w}_i g(\mathrm{x}_i) + g(b_i)) \end{bmatrix}_{m\times 1}, j = 1, 2, \ldots, Q \tag{5}$$

Thus this equation can be rewritten as

$$\mathrm{t}_j = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{mj} \end{bmatrix}_{m\times 1} = \begin{bmatrix} \sum_{i=1}^{l} \beta_{i1}\mathrm{w}_i(g(\mathrm{x}_i)) + \psi_i \\ \sum_{i=1}^{l} \beta_{i2}\mathrm{w}_i(g(\mathrm{x}_i)) + \psi_i \\ \vdots \\ \sum_{i=1}^{l} \beta_{im}\mathrm{w}_i(g(\mathrm{x}_i)) + \psi_i \end{bmatrix}_{m\times 1}, j = 1, 2, \ldots, Q \tag{6}$$

Then, Equation (6) can be presented in compact form as

$$y_k = \alpha_{1,k}g(x_1) + \cdots + \alpha_{i,k}g(x_i) + \cdots + \alpha_{n,k}g(x_n) + \psi_k \tag{7}$$

The coefficients denoted by $\alpha_{i,k}$ are employed in ranking the feature of the desired output of the system. The authors concluded that this method is far superior to the others because it possesses a significant feature reduction ratio and faster processing time.

2.3. Sensitivity analysis

To test which model parameters have a great impact on the modeling outcome and compare that to Gazendam *et al.* (2016) results, a sensitivity analysis was performed on each ELM model. This was performed by fluctuating the analysis around the median value with an offset of 10% maximum for each parameter. The marginal sensitivity (S_c) and normalized sensitivity (S_n) are calculated by:

$$S_c = \frac{\Delta\theta}{\Delta E} \tag{8}$$

$$S_n = S_c \frac{\bar{E}}{\bar{\theta}} \tag{9}$$

where: $\Delta\theta$: Change in model output; ΔE: Change in input parameter. $\bar{\theta}$: Average of model output; $\bar{E}$: Average of input parameters.

3. RESULTS AND DISCUSSION

3.1. Dataset and initial ELM inferences

At each site, metrics of the QHEI were calculated and recorded in the field. This was done based on field observations. As the standardized QHEI scoring scheme suggests, scoring of reach-scale metrics is sorted under substrate quality, riffle-pool quality, bank, and riparian quality, channel morphology development, and in-stream cover. Table 1 shows the QHEI metric and its components' categories and summary. This summary includes mean, standard deviation, minimum, and maximum values for each sub metric in the QHEI. Table 1 with Figure 2 helps in understanding how each sub-metric is distributed compared to others.

The violin plot in Figure 2 shows a detailed summary and explanation of descriptive statistics for each attribute on the graph. For example, from both the table and the graph, the Substrate Type (X1) minimum and maximum values are 1.5 and 17.0. The first, second, and third quartile values are indicated by the dashed lines on the graph from the box plot inside the violins, which in this case are 10.0, 13.0, and 14.0 respectively. Also, the data distribution of each parameter value is indicated in each plot. For instance, Substrate Type (X1), and Cover Type (X5) are examples of unimodal distributions; however, Best Types (X2), and Channel Width (X15) are of bimodal distributions.

In addition, the plot also indicates that the distribution for several parameters is very skewed, and this suggests that the model may not capture the trend perfectly. For example, the drainage area (X22) shows a very skewed distribution with a mean of 4,197 and a standard deviation of 8,023. Knowing that the maximum value for this attribute is 61,311 indicates that the distribution is much skewed to the maximum direction. In contrast, other attributes showed a very different distribution. Floodplain quality (X13), for instance, showed almost a constant distribution. Erosion (X11) showed a conical distribution

Two separate ELM prediction models were then used. Firstly, all the physical parameters were considered in the first model, and only the selected variables were considered in the second model. In each model, the dataset used was split into two; the first part was for model training using 70% of the total observations chosen randomly and the other part using the remaining 30% for testing the trained model. Once the ELM and the other AI models were developed, they were tested against the following statistical error indicators (Table 2): mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), and scatter index (SI).

3.2. ELM model development

ELM was analyzed with different activation functions and numbers of neurons in the hidden layer to test its potential for understanding the physical-HBI-richness relationship. Different activation functions allow for different non-linearities which might work better for solving such a relationship. Furthermore, the number of neurons in each layer affects how the model behaves with the data variance. Hence, more neurons indicate a better fit for the data. Failing to provide a sufficient number of neurons compromises the model's precision. Conversely, exceeding the optimum number of neurons overfits the data and also compromises the model precision. This is examined through testing the model on the test dataset.

Furthermore, to select significant parameters for the second model, correlation between variables was tested on the features based on the correlation matrix. Removing correlated features would result in a faster learning algorithm. Due to the curse of

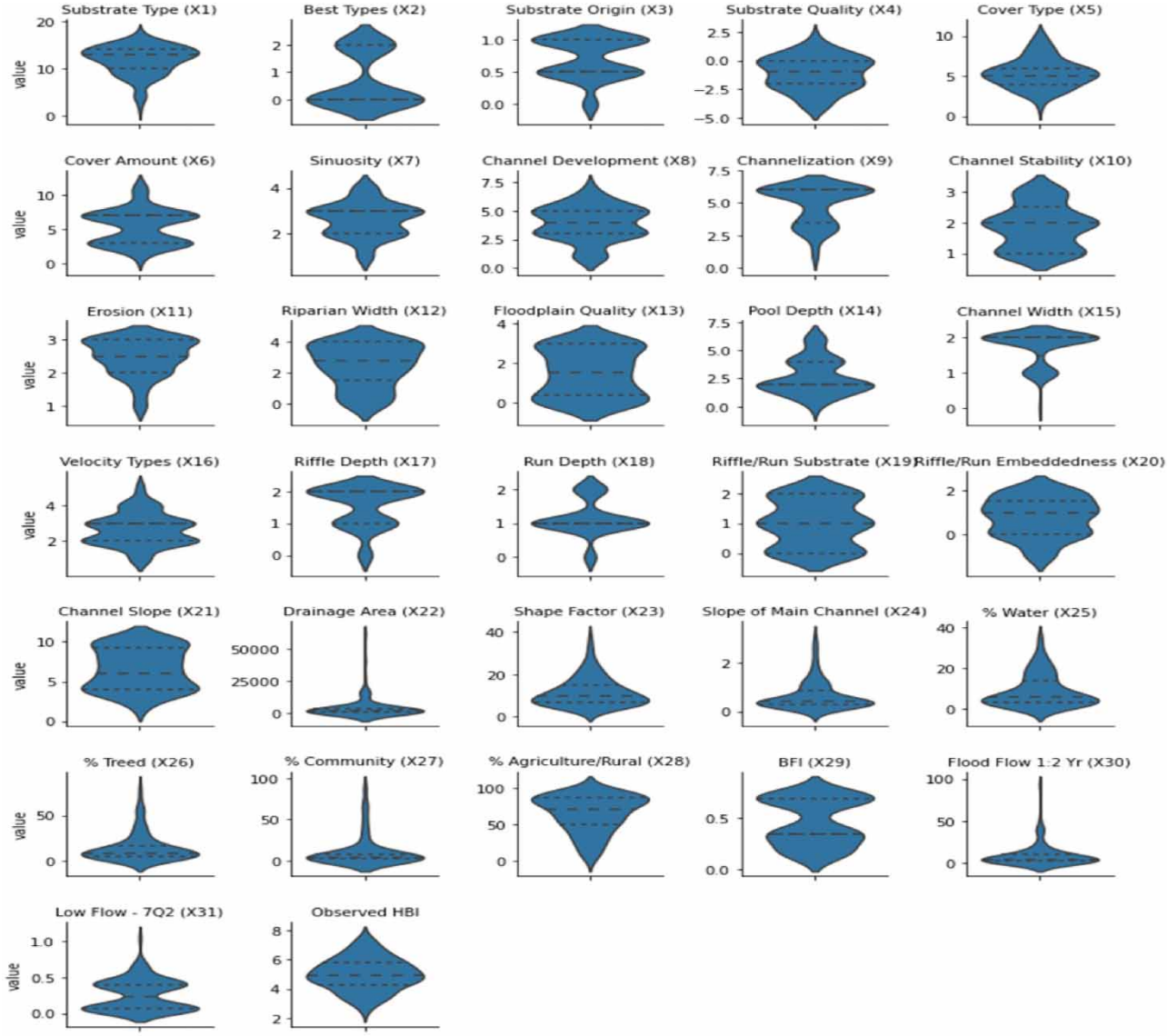


Figure 2 | Data distribution for all features ($X_{1\sim31}$) for the ELM models input.

Table 2 | Performance measurement equations

Error Metric Name	Equation	
Mean Absolute Error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert O_i - S_i\rvert$	Equation 10
Mean Absolute Percentage Error	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{O_i - S_i}{O_i}\right\rvert$	Equation 11
Root Mean Square Error	$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(O_i - S_i)^2}$	Equation 12
Scatter index	$\mathrm{SI} = \frac{RMSE}{\frac{1}{N}\sum_{i=1}^{N}O_i}$	Equation 13

dimensionality (31 variables), fewer features mean high improvement in terms of speed. In such a case, since speed is not an issue because of the small dataset, the ELM model itself will select which parameter to keep and which to throw out internally based on each node weight. Nonetheless, significant variables are selected based on their relative importance factor (relative contribution of the variable to the final output as explained in Pires dos Santos *et al.* (2019) and as described in Section 2.2.2).

3.3. ELM model with all dataset parameters (model 1)

ELM was carried out on the dataset. Diverse activation functions and neurons number in the hidden layer were adopted to test which function is better in this situation. As shown in Figure 3, the sigmoid function exhibited a dominant superiority over the other activation functions in accordance with the number of neurons.

The results show there is a strong correlation between the model's predictions and its actual results. Therefore, it determines a good fit of the ELM model and depicts the randomness (Figure 4) in the residuals indicating that the variance is almost constant for all the input parameters. Figure 5 shows the distribution of the residuals, which tends to be normally distributed around the mean (zero value); the normality test is significant at α=0.05.

Furthermore, the ELM model output was compared with other potential models, as indicated in Table 3. Table 3 indicates all the models utilized for the analysis, the old model proposed in Gazendam *et al.* (2016), and their corresponding residual

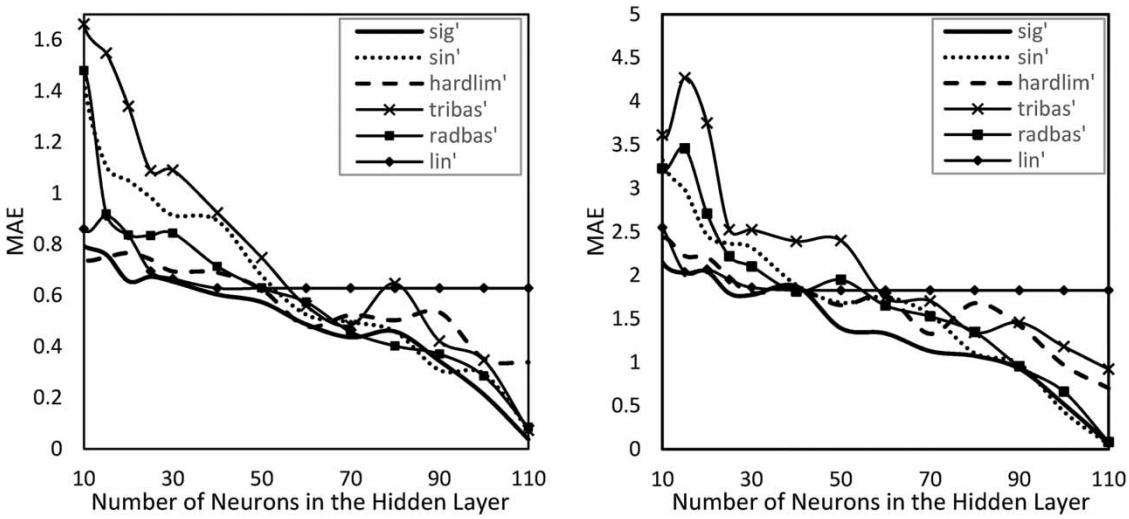


Figure 3 | Mean Absolute Error for all utilized activation functions for HBI (Left) and richness (Right) ELM Models with all features ($X_{1\sim31}$).

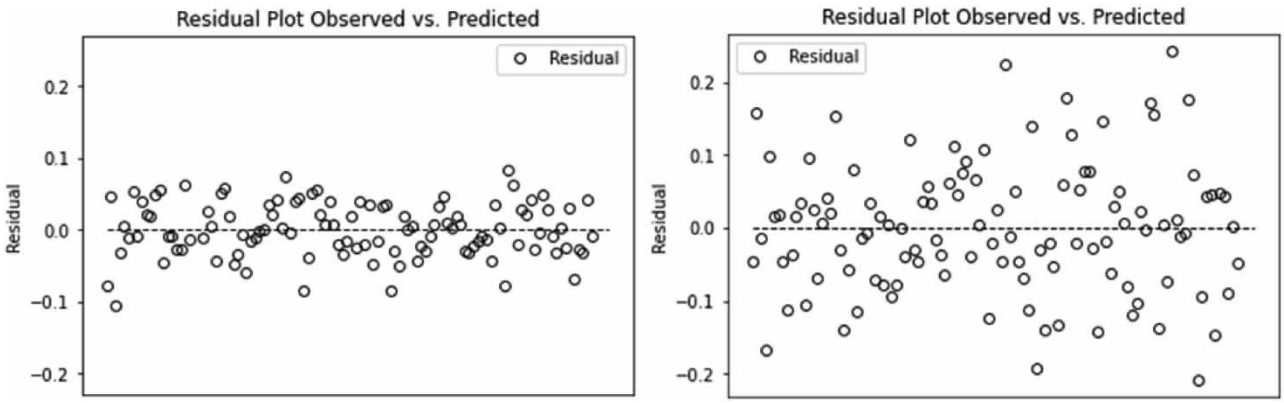


Figure 4 | HBI (left) and richness (right) ELM Model Residuals for all features ($X_{1\sim31}$).

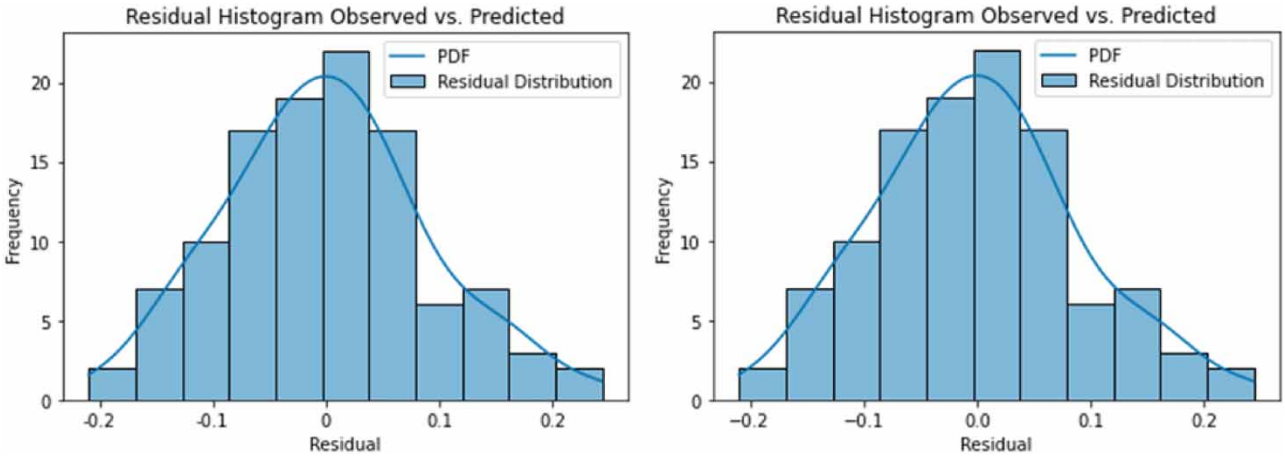


Figure 5 | HBI (left) and richness (right) ELM Model Residuals Distribution for all features ($X_{1\sim31}$).

Table 3 | Performance of ELM model 1 vs. the Gazendam *et al.* (2016) model vs. other studied AI models for all features ($X_{1\sim31}$)

	ELM (Train)	ELM (Test)	Gazendam *et al.* (2016)	LR	SVR	nu-SVR	RIDGER	LASSOR	PINVR	PLSR
(A) HBI Model.										
MAE	0.02	0.07	0.08	0.63	0.86	0.38	0.63	0.63	0.63	0.63
MAPE	0.01	0.02	1.57	13.62	18.43	8.40	13.61	13.61	13.62	13.68
RMSE	0.02	0.12	0.84	1.09	0.52	0.84	0.84	0.84	0.84	0.07
SI	0.00	0.00	1.53	0.17	0.22	0.10	0.17	0.17	0.17	0.17
(B) Richness Model.										
MAE	0.04	0.29	0.57	1.83	2.33	1.59	1.83	1.83	1.83	1.82
MAPE	0.00	0.03	5.31	17.39	23.01	15.91	17.40	17.40	17.39	17.37
RMSE	0.06	0.33	2.28	2.91	2.25	2.29	2.29	2.28	2.29	0.09
SI	0.00	0.00	0.34	0.20	0.25	0.19	0.20	0.20	0.20	0.20

analysis. Mean Absolute Error (MAE) is the most intuitive residual analysis metric since the absolute difference between the data and the model's predictions does not indicate underperformance or overperformance of the model. Each residual (regardless of outliers' presence) contributes proportionally to the total error. This means larger errors will contribute linearly to the overall error, which makes it very robust to outliers. A small error metric suggests the model is great at prediction, while a large value suggests that the model may have trouble in certain areas.

In Table 3, ELM shows lower error values compared to the other models, including the proposed model in Gazendam *et al.* (2016). The test ELM RMSE showed an improvement of almost 7 to 8 times compared to the Gazendam *et al.* (2016) model for the HBI and richness models. The scatter index (SI) improved dramatically from 1.53 and 0.34 to 0.00 in both the HBI and richness models. MAE and MAPE showed a similar trend which indicates a better prediction.

It is also notable that the less-significance models (e.g. LR and SVR) are based on the regression concept. They could not capture the nonlinearity part of the relationship in the dataset. Also perhaps those other models showed this low potential compared to the ELM since there is collinearity between the model parameters. When a correlation exists among parameters, the standard error of parameters' coefficients will increase and the variance of predictor's coefficients is inflated (Daoud 2017).

To test this, Variance Inflation Factors (VIF) analysis was undertaken on the model parameters as shown in Table 4. The results suggest that Drainage Area (X_{22}), % Water Cover (X_{25}), % Treed (X_{26}), % Community (X_{27}), % Agricultural/Rural (X_{28}), and Flood Flow (X_{30}) are highly correlated. Therefore, the relationship between the independent variables and the

Table 4 | VIF analysis for ELM model 1 with all features ($X_{1\sim31}$)

Features	VIF Factor	Features	VIF Factor
X1~X21,	<5	% Water (X25)	108.72
Shape Factor (X23),	<5	% Treed (X26)	441.98
BFI (X29), and	<5	% Community (X27)	588.30
Low Flow – 7Q2 (X31)	<5	% Agriculture/Rural (X28)	1,114.38
Drainage Area (X22)	16.55	Flood Flow 1:2 Yr (X30)	18.32

dependent variables is distorted by the strong relationship between these related independent variables. In return, this will lead to the possibility that our interpretation of relationships will be inappropriate (Daoud 2017).

Another possible explanation that models do not show great potential is the outliers in the dataset. As cited by Khamis *et al.* (2005) outliers in a dataset will influence the modeling accuracy and the estimated statistics. In addition, Khamis *et al.* (2005) concluded that as the percentage of outliers increases in a dataset, models do not optimize and they deviate from the desired accuracy. Consequently, outliers have been identified in the dataset and because the dataset has a slight number of outliers, models with no outliers may not reflect a dramatic improvement in accuracy.

3.4. ELM model with selected parameters (model 2)

Parameters found to be less significant to the model output were removed from the model. Significant variables were selected based on their relative importance factor (indicated in Section 2.2.2). Only Surface Type (X_1), Cover Amount (X_6), Channel Slope (X_{21}), Drainage Area (X_{22}), Shape Factor (X_{23}), % Water Cover (X_{25}), % Treed (X_{26}), % Community (X_{27}), % Agricultural/Rural (X_{28}) and Flood Flow (X_{30}) were kept in the model (Table 5), while the rest were ignored. Accordingly, the model was processed again. It is noteworthy that in the ELM model 2, both models for HBI and richness utilized the same parameters, which is due to the relative correlation between the two indices (DeWalt *et al.* 1998). This was been confirmed with our dataset Pearson's correlation test where the correlation between HBI and richness was found to be −0.339 (correlation significant at 0.01 level). This correlation may influence the parameters selected from the QHEI yet with different trends. Also, the only difference found between the HBI and richness in this study is that the more the HBI index value is, the lower the water quality is. For the richness index, it is vice versa.

The same activation functions and neurons number in the hidden layer were adopted again in this phase to test if the sigmoid function is still dominant. The activation functions MAE followed the same pattern in Figure 3. Sigmoid function persists with the lowest MAE in both models; however, the new model showed a better adoption in terms of less MAE.

The results show there is still a good fit between the actual values and the model output in the HBI case after removing insignificant variables. Although the residuals are still noticeable in the richness model, it is still in the acceptable range with fewer parameters utilized. This in return reflects less effort on real streams measurements.

Figures 6 and 7 show the error is random and follows the normal distribution about the mean zero value. Comparing those two figures with Figures 4 and 5, the new model shows an improvement to the HBI model; however, the richness model observed fit loss. The HBI model residuals drop from 0.15 to less than 0.06; nevertheless, for the richness model, the residuals were raised from 0.2 to 0.4.

The previous models showed great potential compared to the model developed by Gazendam *et al.* (2016) even with outliers kept in the models. However, outliers were removed to improve the model accuracy. Accordingly, the models

Table 5 | Selected QHEI metric components in ELM model 2

Metric	Metric Component	Metric	Metric Component
Substrate	Substrate Type (X1)	Watershed	% Treed (X26)
Stream Cover	Cover Amount (X6)		% Community (X27)
Watershed	Drainage Area (X22)		% Agriculture/Rural (X28)
	Shape Factor (X23)	Geomorphic	Channel Slope (X21)
	% Water (X25)	Hydrology	Flood Flow 1:2 Yr (X30)

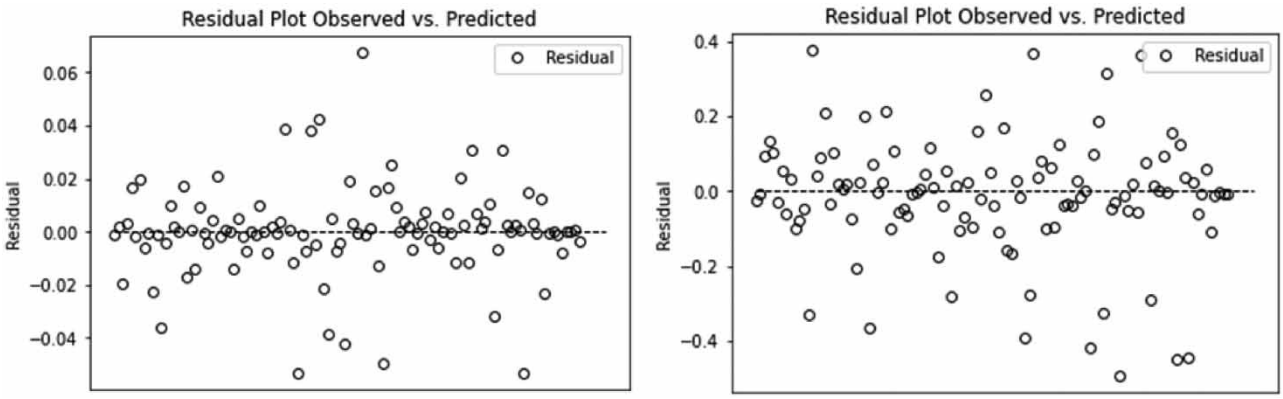


Figure 6 | HBI (left) and Richness (right) ELM Models Residuals for selected Features ($X_{1, 6, 21, 22, 23, 25, 26, 27, 28, 30}$).

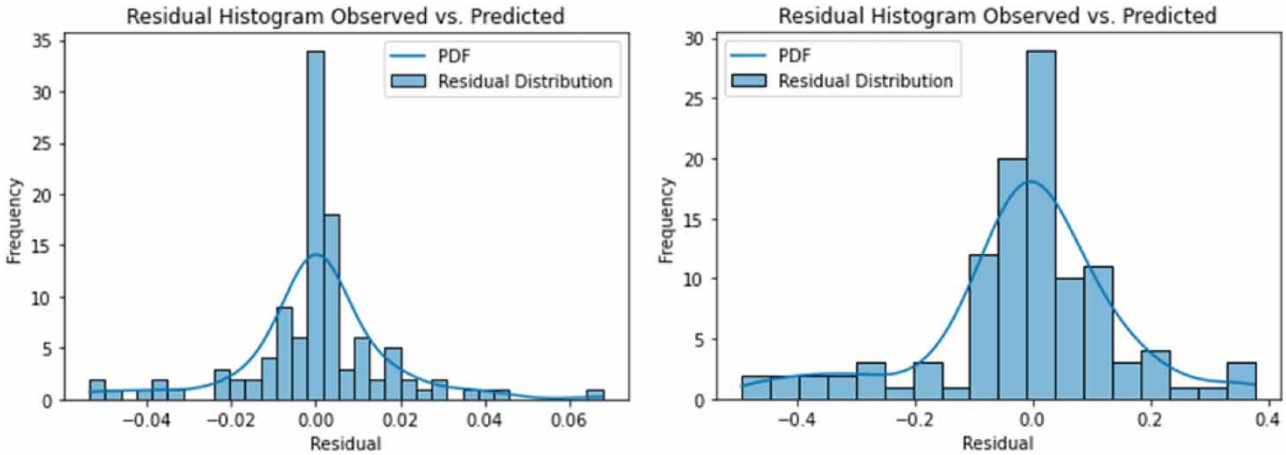


Figure 7 | HBI (left) and richness (right) ELM Models residuals distribution for selected features ($X_{1, 6, 21, 22, 23, 25, 26, 27, 28, 30}$).

showed a better fit with lower MAE. Table 6 shows the resultant MAE, MAPE, RMSE, and SI from applying the model to specific features categories. The Gazendam *et al.* (2016) model was considered in this case with the authors' selected variables.

Table 6 | Performance of ELM model 2 vs. Gazendam *et al.* (2016) model vs. other studied AI models for selected features ($X_{1, 6, 21, 22, 23, 25, 26, 27, 28, 30}$)

	ELM (Train)	ELM (Test)	Gazendam *et al.* (2016)	LR	SVR	nu-SVR	RIDGER	LASSOR	PINVR	PLSR
(A) HBI Model										
MAE	0.00	0.04	0.675	0.71	0.86	0.38	0.76	0.76	0.71	1.06
MAPE	4.02	4.00	18.44	15.17	18.41	8.00	16.29	16.29	15.17	22.40
RMSE	0.00	0.09	0.12	0.92	1.09	0.52	0.98	0.98	0.92	1.34
SI	0.00	0.00	0.00	0.18	0.22	0.10	0.19	0.19	0.18	0.27
(B) Richness Model										
MAE	0.04	0.29	0,720	2.13	2.33	1.59	2.24	2.24	2.13	2.19
MAPE	0.00	0.02	36.47	20.79	23.01	15.93	22.23	22.23	20.79	20.75
RMSE	0.06	0.33	0.82	2.65	2.91	2.25	2.76	2.76	2.65	2.81
SI	0.00	0.00	0.07	0.23	0.25	0.19	0.24	0.24	0.23	0.24

The ELM model output was compared with the same models utilized in Table 3 (LR, SVR, nu-SVR, RIDGER, LASSOR, PINVR, and PLSR). The selected variables for these models are the same models selected for the ELM model 2. Hence the only model that deviates from utilizing the same selected variables of ELM Model 2 is Gazendam *et al.* (2016).

Table 6 indicates all the utilized models for the analysis and their corresponding residual statistics. ELM shows a superior fit for both HBI and richness models when compared to the others. Although the nu-SVR fit is raised for HBI dramatically after the dimensionality reduction, the fit remained the same for the richness model. By keeping the model's high robustness, the selected-parameters model resulted in shrunk model parameters and features which reduce the effort to be made at the site.

3.5. Sensitivity analysis

A sensitivity analysis was made on the selected variables from the Gazendam *et al.* (2016) model (Table 7) and the ELM model 2 (Table 8). The adopted sensitivity test is similar to what was adopted in Gazendam *et al.* (2016), that is, Marginal (S_c) and Normalized Marginal (S_n) sensitivities. These are calculated by changing a parameter value each time by 10% from the median while keeping the rest constant at the median value. The HBI model shows that Velocity Types (X_{16}) and Erosion (X_{11}) are the most negatively sensitive to change with S_n=-0.641 and −0.208, respectively.

Naturally, water velocity is an important variable for macroinvertebrate communities since it determines the ecological distribution of aquatic ecosystems (Ahn *et al.* 2013; Nguyen *et al.* 2018). Similarly, erosion causes soil nutrient deprivation, degradation of land, and leads to many notable off-site environmental problems such as water siltation, and pollution which in turn affect stream health (Issaka & Ashraf 2017). In the richness model, Pool Depth (X_{14}) is the most positively sensitive to change with Sn=50.64. McCulloch (1986) stated that stream pools supported a more diverse benthic fauna because the majority of taxa were adapted to a pool environment. Plus, it is noteworthy that the stream velocity in pools is comparatively slow.

Nonetheless, Table 8 shows the sensitivity analysis performed on the selected parameters from this research and applied to the ELM model. In the HBI and the richness models, Channel Slope (X_{21}) is the most affecting parameter with S_n=13.641 and 35.763, respectively. The gradient of the stream bed is the cause of other factors such as erosion, sediment movement, and the speed of water flow (Hauer & Lamberti 2011).

3.6. Physical interpretation

It is possible to discuss the interpretation of the model selected parameters on the physical sense of stream health. A detailed discussion supported by a literature review is provided here to detail the mutual dependencies and relationships between ELM Model 2 parameters and HBI and richness. References from the literature have been included to verify the robustness of the ELM model 2 selected parameters.

The relationship between substrate type (X1) and macroinvertebrates is confirmed by many researchers. The channel morphology and habitats structure that are available to aquatic organisms are primarily derived from the distribution of sediment along a stream channel (Rempel & Church 2009). This, as a consequence, influences the benthic communities (Halwas & Church 2002). Further, Duan *et al.* (2008) studied the response of macroinvertebrate richness by different substrate types

Table 7 | Sensitivity analysis for selected variables in the Gazendam *et al.* (2016) model for HBI and richness models

HBI Model					Richness Model				
Variable	**Δϕ**	**ΔE**	**Sc**	**Sn**	**Variable**	**Δϕ**	**ΔE**	**Sc**	**Sn**
X1	0.045	1.300	0.035	−0.007	X1	−0.385	1.300	−0.296	−0.090
X4	−0.008	0.100	−0.083	−0.046	X7	0.214	0.300	0.714	0.181
X11	−0.079	0.250	−0.314	−0.208	X11	−0.032	0.250	−0.126	−0.035
X16	−0.102	0.300	−0.339	−0.641	X12	−0.023	0.275	−0.083	−0.013
X26	−0.050	0.855	−0.059	−0.061	X13	0.107	0.150	0.716	0.145
X30	0.023	0.469	0.048	0.048	X14	0.056	0.200	0.281	50.641
–	–	–	–	–	X22	0.021	177.770	0.000	0.000
–	–	–	–	–	X26	0.079	0.855	0.092	0.092

Table 8 | Sensitivity analysis for ELM model 2 variables for HBI and richness models

HBI Model					Richness Model			
Variable	**Δϕ**	**ΔE**	**Sc**	**Sn**	**Δϕ**	**ΔE**	**Sc**	**Sn**
Substrate Type (X1)	−0.091	1.300	−0.070	−0.109	−0.262	1.300	−0.202	−0.097
Cover Amount (X6)	0.022	0.700	0.031	0.041	0.374	0.700	0.535	0.220
Channel Slope (X21)	0.021	0.600	0.035	13.641	0.176	0.600	0.294	35.763
Drainage Area (X22)	0.007	177.770	0.000	0.000	0.098	177.770	0.001	0.000
Shape Factor (X23)	0.021	0.970	0.021	0.026	0.012	0.970	0.012	0.005
% Water (X25)	−0.007	0.555	−0.013	−0.024	−0.172	0.555	−0.309	−0.181
% Treed (X26)	0.033	0.855	0.039	0.028	−0.252	0.855	−0.295	−0.064
% Community (X27)	0.013	0.320	0.039	0.624	−0.110	0.320	−0.344	−1.682
% Agriculture/Rural (X28)	0.357	7.140	0.050	0.052	−2.616	7.140	−0.366	−0.118
Flood Flow 1:2 Yr (X30)	−0.004	0.469	−0.008	−0.008	−0.061	0.469	−0.129	−0.129

on the Juma River in China. The authors replaced the substrate with fine sand, coarse sand, pebbles, coarse hewn stone, and cobbles. Researchers also cleared the macrophytes from the experimental stretch so that macroinvertebrates could colonize naturally. The study results indicated that the biodiversity is comparatively maximum in pebbles, higher in cobbles and hewn stones, and low in coarse sand and fine sand. Sandy substrate is inhospitable to benthic fauna, and the benthic community within it has the lowest taxa richness. The substrata with large grain sizes are stable and hence protect macroinvertebrates.

In other words, it has been proven that the high diversity and density of macroinvertebrate communities were originally sustained by the type of coarse substrate, which in turn provide suitable habitats for most EPT taxa. This could be because the interstitial spaces are a source of habitat and food for invertebrates (Blackwell *et al.* 2006). This has been confirmed by another study undertaken by Jun *et al.* (2011). The research authors studied two mountain streams in Korea and concluded that the type of substrate (X1) is the main factor that controls the distribution of macroinvertebrates and EPT species along with other large-sized habitats (Jun *et al.* 2011). Since lower HBI values indicate higher water quality, previous literature analysis aligns with the ELM model 2 results in terms of substrate type.

Moreover, and as stated by Snyder *et al.* (2005), landcover has been acknowledged as one of the key forces affecting ecological systems by the National Academy of Sciences (Gleick 2000). Much research, e.g. Wang *et al.* (2001), has proven that freshwater streams are susceptible to land use alteration, in particular increased urbanization, which has resulted in the degradation of stream biota. Several studies have found and confirmed the relationship between land use and the degradation of the quality of streams. This has been confirmed by a case study done by Hanna *et al.* (2020) on four streams with different land covers. Their work showed the effects of watershed land use and land cover on stream and riparian ecosystem service provision and biodiversity. They also found that the ecosystem biodiversity quantity and quality were generally higher in sites with close forests (X26). Considering the importance of land cover type, it has also been confirmed that the intensity of land cover intensity (X6) has a very great impact on the stream biological conditions. Alberti *et al.* (2007) studied a sample of 42 basins and concluded that there is a linear relationship between land cover intensity and biological conditions in a stream. Further, Snyder *et al.* (2005) studied 218 small watersheds with logistic regression models to predict stream health. The authors concluded that the percentage of tree cover in the watershed and riparian buffer zone is very dominant in predicting stream health.

Also, Sawyer *et al.* (2004) studied and sampled macroinvertebrate and fish assemblages on 49 sites throughout the Choctawhatchee-Pea, a southeastern U.S. watershed, with their corresponding water quality, land-use, and qualitative and quantitative habitat assessments. The authors determined the relationship between environmental variables and dependent biological variables, and the macroinvertebrate community structure. The 'Drainage factor' for instream habitat had the highest correlation to macroinvertebrate diversity. The Drainage factor comprises stream width and catchment area (X22). This factor is found to be highly and positively correlated to EPT taxa. In addition, not only the drainage area should be considered but also the effective impervious area. Walsh (2004) raised a hypothesis that minimization of drainage connection in streams will result in greater protection of macroinvertebrate taxa. Vietz *et al.* (2014) used a sample of 17 sites in independent

watersheds to test if a stormwater drainage system's connections to streams (Effective Imperviousness) has a relationship with geomorphic attributes other than urban density (Total Imperviousness). This is because Effective Imperviousness controls storm runoff and, in return, runoff is the main driver of geomorphic impacts.

Lei *et al.* (2021) quantified effects on stream water quality between 1992 and 2019 at multiple spatial scales in Germany. The results show that stream slope (X21) explained the spatial variation in water quality and also a strong correlation with nutrient concentration which in return affected macroinvertebrates. The authors, also, concluded that the occurrence of degraded water quality was mostly found in streams with slightly steeper slopes. Further, the study of Park *et al.* (2021) concluded that the mean slope (X21) and forest area (X26) at the watershed and riparian scale have strong positive impacts on the biological indicators of streams.

In addition to slope, the hydrological regime, in general, is found responsible for many of the structural characteristics of rivers. Variations in the flow of water are frequently related to changes in the community structure of organisms (Argerich *et al.* 2004). Argerich *et al.* (2004) studied the Matarranya River macroinvertebrates density after an extraordinary flood of approximately 450 m^3 /s. They concluded that after the flood, macroinvertebrates density dropped to nearly 15–30%. This was confirmed by another study by Argerich *et al.* (2004) in the USA. The authors made a periodic analysis of the Mohawk River to accurately capture storm flow (X30) effect on taxa richness. They found that total taxa richness was significantly lower after floods than before them.

4. CONCLUSION

This study scrutinized the effect of geomorphic, hydrological, and other factors as described in the QHEI metric on stream health and macroinvertebrates. The paper builds on the work done by Gazendam *et al.* (2016) that examined the applicability of ANNs on stream water quality predictability articulated by macroinvertebrates richness and HBI metrics. This study utilized the same dataset (Gazendam *et al.* 2016) used with QHEI being the main metric and another 10 additional watershed-scale metrics.

Two separate ELM Models were developed for the target parameters: HBI and richness. The ELM models were trained on a randomly selected 70% of instances from the dataset. The dataset consisted of 112 stream sites in Ontario, Canada. Further, the two models were validated on the remaining 30%. Feature selection criteria were considered to eliminate less significant features from the model resulting in only ten parameters. A comparison between this study model, the Gazendam *et al.* (2016) model and other potential models (LR, SVR, nu-SVR, RIDGER, LASSOR, PINVR, and PLSR) was made to demonstrate the superiority of the ELM model. LR, SVR, nu-SV, … etc models are used to set a bottom-line performance for these research models and the Gazendam *et al.* (2016) model.

The model with no variable elimination shows a very good fit with RMSE=0.12 for the HBI model and 0.33 for the richness model when compared to the Gazendam *et al.* (2016) model which reached 0.84 and 2.28 for HBI and richness models. Both models from this study and the Gazendam *et al.* (2016) model show a very good potential compared to the other models used for baseline comparison (LR, SVR, and nu-SVR, etc.). After applying the feature selection technique, the QHEI-Channel Conditions and Floodplain categories were eliminated from the high-significant ELM model along with sub-metrics from the other categories. Almost all watershed metrics were included after applying the features selection. This in return, increases the model error RMSE slightly to reach 0.09 and 0.33 for the HBI and richness models. This is still smaller than that of the Gazendam *et al.* (2016) model with 0.12 and 0.82 for the HBI and richness models.

Sensitivity analysis of the ELM models revealed that HBI is directly proportional to Channel Slope (X21), and % Community (X27) and inversely proportional to % Agriculture/Rural (X28) and Flood Flow 1:2 Yr (X30). Richness is directly proportional to Channel Slope (X21), and Cover Amount (X6), and inversely proportional to % Community (X27), % Agriculture/Rural (X28), and Flood Flow 1:2 Yr (X30). A physical interpretation was considered to represent the superiority of the ELM model 2 selected parameters over the other parameters. This was supported by the literature review.

The results of this research acknowledge that the data is too small and only covers very small creeks in a small geographic area of southern Ontario. Additionaly, the range of applicability of this study is the range of applicability of the used sites and therefore there might be higher uncertainties if used with other site's data. Therefore the resultant model would not be applicable anywhere else as is and watershed managers should adjust this model accordingly. Also, as discussed earlier (Gazendam *et al.* 2016), the methodology requires local data sets for model training and validation yet this study shows that ELM has a better architecture than ANN to learn the complex non-linear relationships of the 31 parameters and that

is why the new results of this paper provide better insight on which factors influence HBI or richness. As a consequence, watershed managers should test this methodology and model in their local watershed first before using it. In addition, this study recommends studying the effect of watershed parameters and metrics on stream water quality as a whole instead of site scale only since they proved a higher correlation than other metrics.

FUNDING

The research received no funds.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

CODE AVAILABILITY

There is no custom code or data used other than that used in Gazendam *et al.* (2016).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Ahn, C., Song, H., Lee, S., Oh, J., Ahn, H., Park, J.-R., Lee, J. & Joo, J. 2013 Effects of water velocity and specific surface area on filamentous periphyton biomass in an artificial stream mesocosm. *Water* **5** (4), 1723–1740. https://doi.org/10.3390/w5041723.

Alberti, M., Booth, D., Hill, K., Coburn, B., Avolio, C., Coe, S. & Spirandelli, D. 2007 The impact of urban patterns on aquatic ecosystems: an empirical analysis in Puget lowland sub-basins. *Landscape and Urban Planning* **80** (4), 345–361. https://doi.org/10.1016/j.landurbplan.2006.08.001.

Álvarez-Cabria, M., González-Ferreras, A. M., Peñas, F. J. & Barquín, J. 2017 Modelling macroinvertebrate and fish biotic indices: from reaches to entire river networks. *Science of the Total Environment* **577**, 308–318.

Angermeier, P. L. & Karr, J. R. 2018 Ecological health indicators. In: *Encyclopedia of Ecology*. Elsevier, pp. 391–401. https://doi.org/10.1016/B978-0-12-409548-9.10926-1.

Argerich, A., Puig, M. A. & Pupilli, E. 2004 Effect of floods of different magnitude on the macroinvertebrate communities of Matarranya stream (Ebro river basin, NE Spain). *Limnetica* **23** (43528), 292–293.

Barbour, M. T., Gerritsen, J., Snyder, B. D. & Stribling, J. B. 1999 *Rapid Bioassessment Protocols for use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish*, Vol. 339. US Environmental Protection Agency, Office of Water, Washington, DC.

Berger, E., Haase, P., Kuemmerlen, M., Leps, M., Schäfer, R. B. & Sundermann, A. 2017 Water quality variables and pollution sources shaping stream macroinvertebrate communities. *Science of The Total Environment* **587–588**, 1–10. https://doi.org/10.1016/j.scitotenv.2017.02.031.

Blackwell, G., Haggerty, M., Burns, S., Davidson, L., Gnanalingam, G. & Moller, H. 2006 Cleaner streams and improved stream health on north island dairy and south island sheep/beef farms. *ARGOS Research Report* **6** (03), 10.

Booker, D. J., Snelder, T. H., Greenwood, M. J. & Crow, S. K. 2015 Relationships between invertebrate communities and both hydrological regime and other environmental factors across New Zealand's rivers. *Ecohydrology* **8** (1), 13–32.

Booth, D. B. 2005 Challenges and prospects for restoring urban streams: a perspective from the pacific northwest of North America. *Journal of the North American Benthological Society* **24** (3), 724–737. https://doi.org/10.1899/04-025.1.

Clapcott, J. E., Collier, K. J., Death, R. G., Goodwin, E. O., Harding, J. S., Kelly, D., Leathwick, J. R. & Young, R. G. 2012 Quantifying relationships between land-use gradients and structural and functional indicators of stream ecological integrity. *Freshwater Biology* **57** (1), 74–90. https://doi.org/https://doi.org/10.1111/j.1365-2427.2011.02696.x.

Damanik-Ambarita, M., Everaert, G., Forio, M., Nguyen, T., Lock, K., Musonge, P., Suhareva, N., Dominguez-Granda, L., Bennetsen, E., Boets, P. & Goethals, P. 2016 Generalized linear models to identify Key hydromorphological and chemical variables determining the occurrence of macroinvertebrates in the Guayas River Basin (Ecuador). *Water* **8** (7), 297. https://doi.org/10.3390/w8070297.

Daoud, J. I. 2017 Multicollinearity and regression analysis. *Journal of Physics: Conference Series* **949**, 12009. https://doi.org/10.1088/1742-6596/949/1/012009.

Davies, N. M., Norris, R. H. & Thoms, M. C. 2000 Prediction and assessment of local stream habitat features using large-scale catchment characteristics. *Freshwater Biology* **45** (3), 343–369. https://doi.org/10.1111/j.1365-2427.2000.00625.x.

DeWalt, R. E., Webb, D. W. & Harris, M. A. 1998 Summer Ephemeroptera, Plecoptera, and Trichoptera (EPT) Species Richness and Hilsenhoff Biotic Index at Eight Stream Segments in the Lower Illinois River Basin. *The Great Lakes Entomologist* **32** (2).

Duan, X., Wang, Z. & Tian, S. 2008 Effect of streambed substrate on macroinvertebrate biodiversity. *Frontiers of Environmental Science & Engineering in China* **2** (1), 122–128. https://doi.org/10.1007/s11783-008-0023-y.

Einheuser, M. D., Nejadhashemi, A. P., Sowa, S. P., Wang, L., Hamaamin, Y. A. & Woznicki, S. A. 2012 Modeling the effects of conservation practices on stream health. *Science of the Total Environment* **435–436**, 380–391. https://doi.org/10.1016/j.scitotenv.2012.07.033.

Ertugrul, Ö. F. & Taugluk, M. E. 2017 A fast feature selection approach based on extreme learning machine and coefficient of variation. *Turkish Journal of Electrical Engineering Ü& Computer Sciences* **25** (4), 3409–3420.

Gazendam, E., Gharabaghi, B., Jones, F. C. & Whiteley, H. 2011 Evaluation of the qualitative habitat evaluation index as a planning and design tool for restoration of rural ontario waterways. *Canadian Water Resources Journal* **36** (2), 149–158. https://doi.org/10.4296/cwrj3602827.

Gazendam, E., Gharabaghi, B., Ackerman, J. D. & Whiteley, H. 2016 Integrative neural networks models for stream assessment in restoration projects. *Journal of Hydrology* **536**, 339–350. https://doi.org/10.1016/j.jhydrol.2016.02.057.

Gleick, P. H. 2000 *Water: the Potential Consequences of Climate Variability and Change for the Water Resources of the United States*. Pacific Institute for Studies in Development, Environment, and Security, Oakland, CA, USA.

Goetz, S. & Fiske, G. J. 2013 On the relationship between stream biotic diversity and exurbanization in the Northeastern USA. In: *Geospatial Tools for Urban Water Resources*, Vol. 7. Springer, Netherlands, pp. 61–78. https://doi.org/10.1007/978-94-007-4734-0_4.

Gorney, R. M., Williams, M. G., Ferris, D. R. & Williams, L. R. 2012 The influence of channelization on fish communities in an agricultural coldwater stream system. *American Midland Naturalist* **168** (1), 132–143. https://doi.org/10.1674/0003-0031-168.1.132.

Guyon, I. & Elisseeff, A. 2003 An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157–1182.

Halwas, K. L. & Church, M. 2002 Channel units in small, high gradient streams on Vancouver Island, British Columbia. *Geomorphology* **43** (3–4), 243–256. https://doi.org/10.1016/S0169-555X(01)00136-2.

Hanna, D. E. L., Raudsepp-Hearne, C. & Bennett, E. M. 2020 Effects of land use, cover, and protection on stream and riparian ecosystem services and biodiversity. *Conservation Biology* **34** (1), 244–255. https://doi.org/10.1111/cobi.13348.

Hastie, T., Tibshirani, R. & Friedman, J. 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, Stanford, CA, USA.

Hauer, F. & Lamberti, G. 2011 *Methods in Stream Ecology*. Elsevier Inc. Available from: https://books.google.com/books?hl=en&lr=&id=rlclsSCF_dQC&oi=fnd&pg=PP1&dq=Methods+in%0AStream+Ecology&ots=4yzH-YhOGh&sig=GYWj4PpaA6pezDXm2DcwnOJyrMk

Herman, M. R. & Nejadhashemi, A. P. 2015 A review of macroinvertebrate- and fish-based stream health indices. In: *Ecohydrology and Hydrobiology*, Vol. 15, (2). Elsevier, pp. 53–67. https://doi.org/10.1016/j.ecohyd.2015.04.001.

Hernandez-Suarez, J. S. & Nejadhashemi, A. P. 2018 A review of macroinvertebrate- and fish-based stream health modelling techniques. In: *Ecohydrology*, Vol. 11 (8). John Wiley and Sons Ltd, p. e2022. https://doi.org/10.1002/eco.2022.

Hilsenhoff, W. L. 1988 Rapid field assessment of organic pollution with a family-level biotic index. *Journal of the North American Benthological Society* **7** (1), 65–68. https://doi.org/10.2307/1467832.

Hintze, J. L. & Nelson, R. D. 1998 Violin plots: a Box plot-density trace synergism statistical computing and graphics violin plots: a box plot-density trace synergism. *Source: The American Statistician* **52** (2), 181–184. Available from: http://www.jstor.org/stable/2685478%5Cnhttp://www.jstor.org/%5Cnhttp://www.jstor.org/action/showPublisher?publisherCode=astata.%5Cnhttp://www.jstor.org

Hoang, T. H., Lock, K., Mouton, A. & Goethals, P. L. M. 2010 Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecological Informatics* **5** (2), 140–146. https://doi.org/10.1016/j.ecoinf.2009.12.001.

Huang, G.-B., Zhu, Q.-Y. & Siew, C.-K. 2006 Extreme learning machine: theory and applications. *Neurocomputing* **70** (1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126.

Issaka, S. & Ashraf, M. A. 2017 Impact of soil erosion and degradation on water quality: a review. *Geology, Ecology, and Landscapes* **1** (1), 1–11. https://doi.org/10.1080/24749508.2017.1301053.

Jun, Y. C., Kim, N. Y., Kwon, S. J., Han, S. C., Hwang, I. C., Park, J. H., Won, D. H., Byun, M. S., Kong, H. Y., Lee, J. E. & Hwang, S. J. 2011 Effects of land use on benthic macroinvertebrate communities: comparison of two mountain streams in Korea. *Annales de Limnologie* **47** (S1), S35–S49. https://doi.org/10.1051/limn/2011018.

Karr, J. R. 1999 Defining and measuring river health. *Freshwater Biology* **41** (2), 221–234. https://doi.org/10.1046/j.1365-2427.1999.00427.x.

Khamis, A., Ismail, Z., Haron, K. & Mohamm, A. T. 2005 The effects of outliers data on neural network performance. *Journal of Applied Sciences* **5** (8), 1394–1398. https://doi.org/10.3923/jas.2005.1394.1398.

Kim, J. J., Atique, U. & An, K. G. 2019 Long-term ecological health assessment of a restored urban stream based on chemical water quality, physical habitat conditions and biological integrity. *Water (Switzerland)* **11** (1). https://doi.org/10.3390/w11010114.

Lee, J. H. & An, K. G. 2014 Integrative restoration assessment of an urban stream using multiple modeling approaches with physical, chemical, and biological integrity indicators. *Ecological Engineering* **62**, 153–167. https://doi.org/10.1016/j.ecoleng.2013.10.006.

Lei, C., Wagner, P. D. & Fohrer, N. 2021 Effects of land cover, topography, and soil on stream water quality at multiple spatial and seasonal scales in a German lowland catchment. *Ecological Indicators* **120**, 106940. https://doi.org/10.1016/j.ecolind.2020.106940.

Leuenberger, M. & Kanevski, M. 2015 Extreme learning machines for spatial environmental data. *Computers & Geosciences* **85**, 64–73. https://doi.org/10.1016/j.cageo.2015.06.020.

Maddock, I. 1999 The importance of physical habitat assessment for evaluating river health. *Freshwater Biology* **41** (2), 373–391. https://doi.org/10.1046/j.1365-2427.1999.00437.x.

Mantyka-Pringle, C. S., Jardine, T. D., Bradford, L., Bharadwaj, L., Kythreotis, A. P., Fresque-Baxter, J., Kelly, E., Somers, G., Doig, L. E., Jones, P. D. & Lindenschmidt, K.-E. 2017 Bridging science and traditional knowledge to assess cumulative impacts of stressors on ecosystem health. *Environment International* **102**, 125–137. https://doi.org/10.1016/j.envint.2017.02.008.

McCulloch, D. L. 1986 Benthic macroinvertebrate distributions in the riffle-pool communities of two east Texas streams. *Hydrobiologia* **135** (1), 61–70.

Mrozińska, N., Glińska-Lewczuk, K., Burandt, P., Kobus, S., Gotkiewicz, W., Szymańska, M., Bąkowska, M. & Obolewski, K. 2018 Water quality as an indicator of stream restoration effects – A case study of the kwacza river restoration project. *Water* **10** (9), 1249. https://doi.org/10.3390/w10091249.

Nguyen, T. H. T., Boets, P., Lock, K., Forio, M. A. E., Van Echelpoel, W., Van Butsel, J., Utreras, J. A. D., Everaert, G., Granda, L. E. D., Hoang, T. H. T. & Goethals, P. L. M. 2017 Water quality related macroinvertebrate community responses to environmental gradients in the Portoviejo River (Ecuador). *Annales de Limnologie* **53**, 203–219. https://doi.org/10.1051/limn/2017007.

Nguyen, T. H. T., Forio, M. A. E., Boets, P., Lock, K., Ambarita, M. N. D., Suhareva, N., Everaert, G., Van der heyden, C., Dominguez-Granda, L. E., Hoang, T. H. T. & Goethals, P. 2018 Threshold responses of macroinvertebrate communities to stream velocity in relation to hydropower dam: a case study from the Guayas River Basin (Ecuador). *Water (Switzerland)* **10** (9). https://doi.org/10.3390/w10091195.

Ollis, D. J., Dallas, H. F., Esler, K. J. & Boucher, C. 2006 Bioassessment of the ecological integrity of river ecosystems using aquatic macroinvertebrates: an overview with a focus on South Africa. *African Journal of Aquatic Science* **31** (2), 205–227. https://doi.org/10.2989/16085910609503892.

Park, S.-R., Hwang, S.-J., An, K. & Lee, S.-W. 2021 Identifying key watershed characteristics that affect the biological integrity of streams in the Han River Watershed, Korea. *Sustainability* **13** (6), 3359. https://doi.org/10.3390/su13063359.

Pires dos Santos, R., Dean, D. L., Weaver, J. M. & Hovanski, Y. 2019 Identifying the relative importance of predictive variables in artificial neural networks based on data produced through a discrete event simulation of a manufacturing environment. *International Journal of Modelling and Simulation* **39** (4), 234–245. https://doi.org/10.1080/02286203.2018.1558736.

Rakocinski, C. F. 2012 Evaluating macrobenthic process indicators in relation to organic enrichment and hypoxia. *Ecological Indicators* **13** (1), 1–12. https://doi.org/10.1016/j.ecolind.2011.04.031.

Rankin, E., Miltner, B., Yoder, C. & Mishne, D. 1999 Association between Nutrients, Habitat, and the Aquatic Biota in Ohio Rivers and Streams *Ohio EPA Technical bulletin MAS/1999-1*. Ohio EPA, Columbus, OH, USA.

Rempel, L. L. & Church, M. 2009 Physical and ecological response to disturbance by gravel mining in a large alluvial river. *Canadian Journal of Fisheries and Aquatic Sciences* **66** (1), 52–71. https://doi.org/10.1139/F08-184.

Roy, A. H., Rhea, L. K., Mayer, A. L., Shuster, W. D., Beaulieu, J. J., Hopton, M. E., Morrison, M. A. & St Amand, A. 2014 How much is enough? minimal responses of water quality and stream biota to partial retrofit stormwater management in a suburban neighborhood. *PLoS ONE* **9** (1). https://doi.org/10.1371/journal.pone.0085011.

Sauer, J., Domisch, S., Nowak, C. & Haase, P. 2011 Low mountain ranges: summit traps for montane freshwater species under climate change. *Biodiversity and Conservation* **20** (13), 3133–3146. https://doi.org/10.1007/s10531-011-0140-y.

Sawyer, J. A., Stewart, P. M., Mullen, M. M., Simon, T. P. & Bennett, H. H. 2004 Influence of habitat, water quality, and land use on macroinvertebrate and fish assemblages of a southeastern coastal plain watershed, USA. *Aquatic Ecosystem Health & Management* **7** (1), 85–99. https://doi.org/10.1080/14634980490281353.

Schwindt, S., Pasternack, G. B., Bratovich, P. M., Rabone, G. & Simodynes, D. 2019 Hydro-morphological parameters generate lifespan maps for stream restoration management. *Journal of Environmental Management* **232**, 475–489. https://doi.org/10.1016/j.jenvman.2018.11.010.

Simões, N. R., Braghin, L. S. M., Duré, G. A. V., Santos, J. S., Sonoda, S. L. & Bonecker, C. C. 2020 Changing taxonomic and functional β-diversity of cladoceran communities in Northeastern and South Brazil. *Hydrobiologia* **847** (18), 3845–3856. https://doi.org/10.1007/s10750-020-04234-w.

Snyder, M. N., Goetz, S. J. & Wright, R. K. 2005 Stream health rankings predicted by satellite derived land cover metrics. *Journal of the American Water Resources Association* **41** (3), 659–677. https://doi.org/10.1111/j.1752-1688.2005.tb03762.x.

Suen, J. P. 2009 Use of artificial neural networks for habitat unit composition modeling. *Proceedings of World Environmental and Water Resources Congress 2009 – World Environmental and Water Resources Congress 2009: Great Rivers* **342**, 3159–3166. https://doi.org/10.1061/41036(342)319.

Van Broekhoven, E., Adriaenssens, V., De Baets, B. & Verdonschot, P. F. M. 2006 Fuzzy rule-based macroinvertebrate habitat suitability models for running waters. *Ecological Modelling* **198** (1–2), 71–84. https://doi.org/10.1016/j.ecolmodel.2006.04.006.

Vietz, G. J., Sammonds, M. J., Walsh, C. J., Fletcher, T. D., Rutherfurd, I. D. & Stewardson, M. J. 2014 Ecologically relevant geomorphic attributes of streams are impaired by even low levels of watershed effective imperviousness. *Geomorphology* **206**, 67–78. https://doi.org/10.1016/j.geomorph.2013.09.019.

Walsh, C. J. 2004 Protection of in-stream biota from urban impacts: minimise catchment imperviousness or improve drainage design? *Marine and Freshwater Research* **55** (3), 317. https://doi.org/10.1071/MF03206.

Wang, L., Lyons, J., Kanehl, P. & Bannerman, R. 2001 Impacts of urbanization on stream habitat and fish across multiple spatial scales. *Environmental Management* **28** (2), 255–266. https://doi.org/10.1007/s0026702409.

Wang, L., Robertson, D. M. & Garrison, P. J. 2007 Linkages between nutrients and assemblages of macroinvertebrates and fish in wadeable streams: implication to nutrient criteria development. *Environmental Management* **39** (2), 194–212. https://doi.org/10.1007/s00267-006-0135-8.

Wang, L., Brenden, T., Seelbach, P., Cooper, A., Allan, D., Clark, R. & Wiley, M. 2008 Landscape based identification of human disturbance gradients and reference conditions for Michigan streams. *Environmental Monitoring and Assessment* **141** (1–3), 1–17. https://doi.org/10.1007/s10661-006-9510-4.

Woznicki, S. A., Nejadhashemi, A. P., Ross, D. M., Zhang, Z., Wang, L. & Esfahanian, A. H. 2015 Ecohydrological model parameter selection for stream health evaluation. *Science of the Total Environment* **511**, 341–353. https://doi.org/10.1016/j.scitotenv.2014.12.066.

First received 18 December 2021; accepted in revised form 29 March 2022. Available online 16 April 2022