Journal of Water Process Engineering 64 (2024) 105598

Contents lists available at ScienceDirect

Journal of Water Process Engineering

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

Hybridized machine learning models for phosphate pollution modeling in water systems for multiple uses

Tales H.A. Boratto [a], Deivid E.D. Campos [a], Douglas L. Fonseca [a], Welson Avelar Soares Filho [a], Zaher M. Yaseen [b,c], Angela Gorgoglione [d], Leonardo Goliatt [e,*]

[a] Computational Modeling Program, Federal University of Juiz de Fora, Juiz de Fora, MG 36036-900, Brazil
[b] Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[c] Interdisciplinary Research Center for Membranes and Water Security, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[d] Department of Fluid Mechanics and Environmental Engineering (IMFIA), School of Engineering, Universidad de la República, Julio Herrera y Reissig 565, Montevideo 11300, Uruguay
[e] Department of Computational and Applied Mechanics, Federal University of Juiz de Fora, Juiz de Fora, MG 36036-900, Brazil

ARTICLE INFO

Editor: Guangming Jiang

Keywords:
Harris hawks optimizer
Phosphate
Hybrid machine learning

ABSTRACT

Phosphate pollution in water bodies is a significant environmental concern, especially in regions with extensive agricultural practices. Hence, a tool for accurately assessing the phosphate concentration is essential. This research paper explores the effectiveness of machine learning (ML) models combined with nature-inspired optimization algorithms for predicting phosphate levels in water systems. The novelty consists of integrating the power of machine learning models, which have been presenting excellent performance at capturing complex relationships in environmental pollution data, with the Harris Hawks Optimizer (HHO) optimization capabilities inspired by hawks' hunting behavior. Four hybrid implementations combining the HHO were evaluated, and four feature subsets were assessed to identify the most influential variables in the modeling process. Using water quality data from Brazilian upstream watersheds, the hybrid models were trained and validated, enabling accurate and robust predictions of phosphate concentrations. The elastic net (EN) model optimized by Harris Hawk Optimizer (HHO-EN) produced the best-averaged performance among all experiments ($R = 0.825$, $R^2 = 0.670$, Root Mean Square Error (RMSE) = 0.049 mg/L, Mean Absolute Error (MAE) = 0.037 mg/L). A parametric and feature importance analysis identified the most influential parameters in the contamination modeling process. Hybrid machine learning models represent a novel and efficient strategy for water quality monitoring and environmental management, supporting the preservation of aquatic ecosystems.

1. Introduction

Intensive agricultural activities are recognized as significant contributors to nonpoint source pollution, posing a considerable threat to the quality of water bodies and overall ecosystem health [1–4]. Surface water, which serves as a primary drinking source in many countries [5], necessitates the assessment of water quality and the development of reliable tools for effective water resource management [6].

During the growing season, most agricultural practices often rely on high application rates of fertilizers to maximize crop yields. These fertilizers contain substantial amounts of nutrients, such as phosphate (PO_4^{3-}), which are essential for enhancing crop production [7]. However, during storm events, nonpoint source pollutants, particularly nutrients, can enter surface water bodies through surface runoff [8]. The disproportionate concentrations of PO_4^{3-} in surface waters are often major contributors to eutrophication. The latter is considered a global concern since it promotes rapid algae and aquatic plant growth, which leads to oxygen depletion and adversely affects fish and other aquatic organisms. Additionally, the accumulation of organic matter from algal blooms can release toxins and disrupt the ecological balance [9]. This further highlights the significance of monitoring and regularly assessing water quality parameters.

* Corresponding author.
E-mail addresses: tales.boratto@engenharia.ufjf.br (T.H.A. Boratto), deivid.campos@engenharia.ufjf.br (D.E.D. Campos), douglas.lima@estudante.ufjf.br (D.L. Fonseca), wfilho@ice.ufjf.br (W.A. Soares Filho), z.yaseen@kfupm.edu.sa (Z.M. Yaseen), agorgoglione@fing.edu.uy (A. Gorgoglione), leonardo.goliatt@ufjf.br (L. Goliatt).

https://doi.org/10.1016/j.jwpe.2024.105598
Received 27 February 2024; Received in revised form 9 May 2024; Accepted 3 June 2024
Available online 10 June 2024

Monitoring the PO_4^{3-} concentration in water bodies presents both economic and technical challenges. Traditional monitoring methods, which involve extensive sampling, laboratory processing, and data interpretation, can be costly and time-consuming [10]. The requirement for specialized equipment, skilled personnel, and regular site visits adds to the financial burden. Furthermore, the dynamic and fluctuating nature of PO_4^{3-} levels in water bodies necessitates frequent and continuous monitoring, which is logistically challenging. These economic and technical difficulties highlight the importance of developing alternative approaches, such as numerical models, that can provide accurate and efficient predictions of PO_4^{3-} concentrations, reducing the reliance on costly and resource-intensive monitoring methods.

In recent decades, physical-based and statistical models have been successfully used to simulate phosphorus dynamics at the watershed scale [11–13]. However, physically based models rely on complex mathematical equations that describe the transport and transformation processes of phosphates. However, these methods often require a significant amount of input data (e.g., catchment physical characteristics, meteorological information, land use/land cover, and soil type), which can be uncertain or unavailable [14–16].

The limitations of conventional physical-based models emphasize the need for advanced modeling approaches, such as machine learning, to capture better the complex and nonlinear nature of phosphate dynamics in water bodies. ML has become a powerful tool for capturing complex and nonlinear relationships between input and output data. Its effectiveness in predicting nonlinear systems has been widely acknowledged, leading to its adoption by researchers for addressing complex water quality problems [8,17]. With a diverse range of algorithms available, ML has proven to be a reliable approach, offering valuable insights and solutions to complex challenges in the field of environmental engineering.

Although several studies have employed ML techniques for monitoring and mitigating diverse pollutants in rivers and water reservoirs [17–20], this research specifically focused on phosphate due to its well-established agricultural system and developed industrial park. The extensive application of phosphate-based fertilizers in this area can significantly increase phosphate concentrations in rivers, necessitating effective control and management strategies. Considering the faster accumulation of phosphorus than of nitrogen in human-impacted freshwater ecosystems, phosphorus accumulation can have serious implications for trophic webs and biogeochemical cycles in estuaries and coastal areas, as it is influenced mainly by freshwater loadings [21].

Considering specifically phosphate as a target pollutant, recent studies have employed ML as a tool to aid decision-making. Artificial neural networks, random forests, and support vector machines were extensively used for phat contaminant modeling [22–27]. Other models also include gradient boosting (GB) [6], linear models [23], and partial least squares [28]. Recent studies have attempted to measure phosphate in water without sending a sample to the laboratory using spectroscopy [29] and other techniques [30,31]. However, it is possible to use ML algorithms to develop models that predict pollutant concentrations based on other water characteristics, such as temperature, pH, turbidity, and apparent color. These models can be trained on historical measurement data and evaluated on new data to validate the accuracy of predictions.

Applying ML techniques can also contribute to the development of automatic phosphate monitoring systems present in water, allowing a faster and more effective response in cases of contamination. In this sense, this work aims to contribute to the first study of the use of optmized ML techniques with nature-inspired algorithms in the measurement of phosphate in surface water bodies, exploring the potential of this approach to improve the efficiency of phosphate analysis and assisting in decision-making for environmental management and public health protection.

Based on these considerations, this work aims to investigate the effectiveness of hybrid ML models incorporating the HHO algorithm for predicting phosphate levels in three major rivers in Brazil. Additionally, we propose evaluating four feature subsets to identify the most critical variables influencing prediction accuracy. A feature importance analysis using Shapley's method and a model's parametric analysis were also conducted to find the most relevant features for producing accurate predictions. By integrating the power of ML and the optimization capabilities of the HHO algorithm and evaluating the feature subsets, this paper presents the development of accurate and robust models to enhance the monitoring and management of phosphate pollution in these river systems, supporting the preservation of aquatic ecosystems.

Specifically, this study uses water quality data from Brazilian headwaters, retrieved from Taffarello et al. [32], that supply the Cantareira reservoir in the state of São Paulo, Brazil. The Cantareira system is one of the most important systems in Brazil, as it is responsible for supplying water to more than 9 million people in the São Paulo metropolitan region, and it is also important for supplying water to planting regions. It is important to note that this region suffers from water crises; thus, water quality monitoring becomes even more relevant when supplies become scarce.

The main contributions of this paper are described as follows: first, this work proposes a water quality modeling tool that synergistically integrates ML models with nature-inspired optimization techniques to predict phosphate contamination in water systems. Secondly, a comprehensive parametric analysis is conducted on the optimized models. This analysis utilizes data collect from a region with combined industrial and agricultural characteristics to identify the most relevant variables influencing phosphate levels. The findings offer valuable insights for both future scientific endeavors and practical water quality management applications. Finally, the paper employs feature importance analysis to identify the parameters most crucial for understanding phosphate pollution in water bodies.

This paper is organized as follows. In Section 2, the dataset used in this paper, the ML models, the metaheuristic optimization implemented, and the evaluation methods are described. Section 3 addresses the model implementation, the results, the sensitivity analysis, and the discussion. Finally, Section 4 concludes this paper by presenting the final observations.

2. Material and methods

2.1. Methodology conceptualization

Searching for the most appropriate ML hyperparameters can be a challenging and time-consuming task, especially when the performance of the ML model relies heavily on selecting adequate internal parameters [33,34]. Hyperparameters are parameters that are not learned during the training process but need to be set before training the model; these parameters include the learning rate, number of hidden layers, and number of neurons in a neural network or the maximum depth and minimum samples per leaf in a decision tree. The hyperparameter search procedure involves finding the optimal combination of hyperparameters that yields the best performance of the ML model in predicting phosphate levels.

With the proposed model, we address this challenge by conducting a hyperparameter search using an evolutionary algorithm, specifically the HHO. The HHO algorithm is then employed to efficiently explore the hyperparameter space and search for the most suitable combination of hyperparameters. In the hyperparameter search procedure, the HHO algorithm starts by initializing a population of hawks, where each hawk represents a potential combination of hyperparameters for the ML model. These hyperparameters are considered the spatial coordinates of the hawks. The HHO algorithm then proceeds through multiple iterations, simulating the hunting process of the hawks to find the best hyperparameter setting.

During the exploration phase of HHO, hawks search for promising regions in the hyperparameter space, evaluating the performance of the ML model using their respective hyperparameter combinations. This is akin to hawks searching for elevated locations to spot potential prey during hunting. The exploration is guided by mathematical equations that determine the next position of each hawk based on its current position and the position of the prey, which corresponds to a promising hyperparameter combination with good model performance.

As the algorithm progresses, it transitions from exploration to

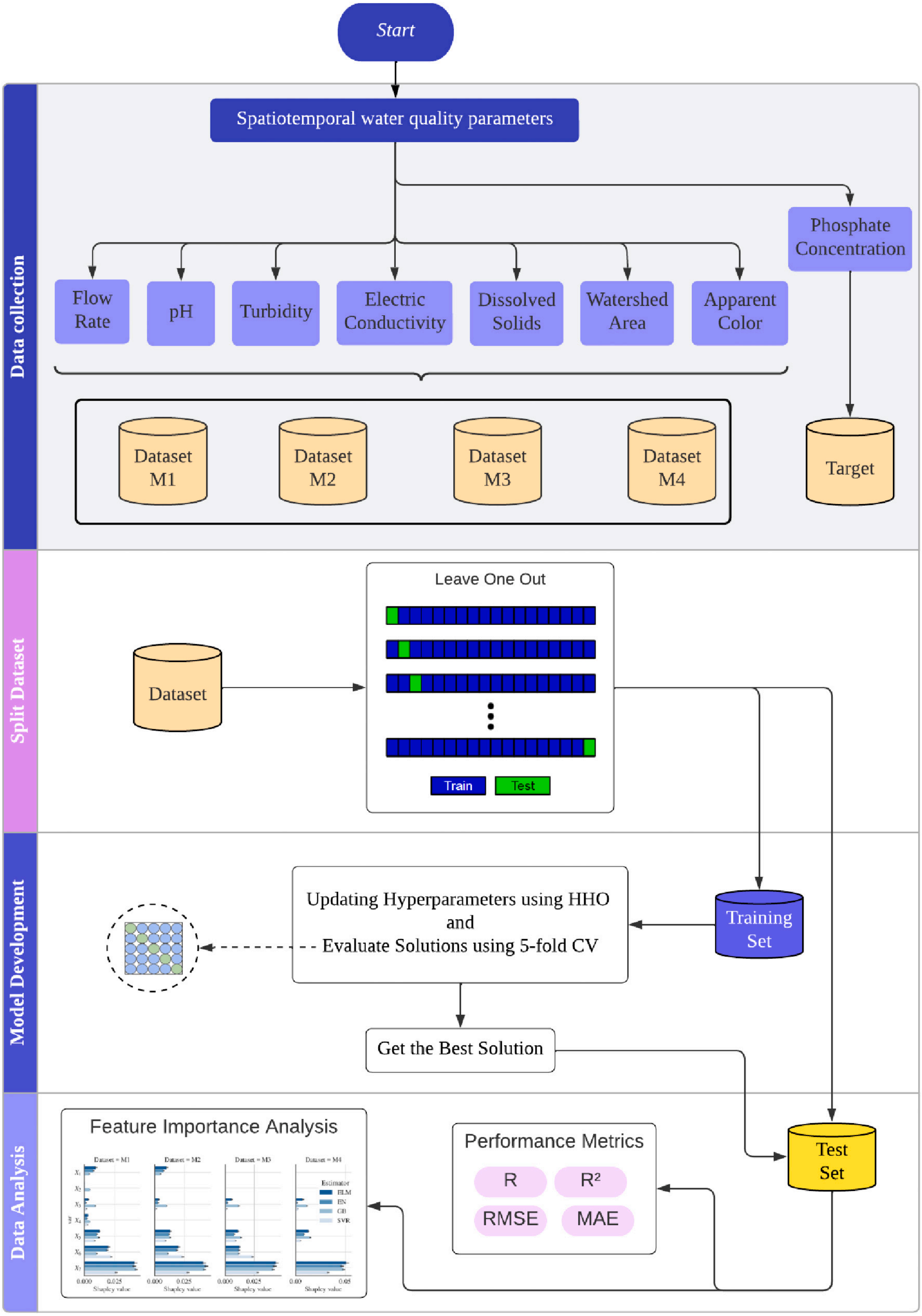

Fig. 1. Schematic flowchart of the logic of the implemented computational model.

exploitation, imitating how hawks shift their hunting strategy based on the prey's energy. In this phase, the HHO algorithm focuses on intensifying the search around promising hyperparameter combinations with the highest potential for improving model performance. The prey's energy, represented as K in the equations, influences the extent of exploration and exploitation in the hyperparameter search.

By effectively combining exploration and exploitation, the HHO algorithm efficiently navigates through the hyperparameter space, gradually refining the population of hawks to converge toward the best hyperparameter setting that optimizes the ML model's performance in predicting phosphate levels.

The flowchart shown in Fig. 1 illustrates the logic of the implemented computational model. First, the original dataset was divided according to the following feature selection procedure: The degree of correlation was manually determined by subdividing the data into four subsets (M1, M2, M3, and M4). Then, each dataset was split into training and test sets using the Leave One Out (LOO) method. For each iteration of the LOO algorithm, the model's hyperparameters were updated, considering cross-validation with five folds in the current training set as a form of evaluation. After finding the best parameter setting, the test set was used as input for the assessment model. Table 3 shows the search space for each ML model.

2.2. Datasets

The original dataset used in this paper was extracted and adapted from Taffarello et al. [32] and consists of fresh-water quality information mostly collected in three Brazilian cities: Extrema (MG), Joanópolis (SP) and Nazaré Paulista (SP), as shown in Fig. 2. The original dataset was filtered, and for this study, 8 variables, 7 of which were features ($X_1 \ldots X_7$) and 1 of which was the model target (y) were selected. A total of 62 samples were used for the experiments. These variables are described in Table 1.

Even though the dataset comprises a relatively small number of input variables, it is crucial to acknowledge that in an ML context, incorporating more variables does not necessarily yield improved results [35,36]. Since the attributes describe or concentrate information about a given system or problem, it is common that these variables have different degrees of importance and may be highly relevant, irrelevant, or even not very relevant, as is the case for variables whose information has already been contemplated to a large extent by other attributes, becoming redundant. In this case, a manual feature selection approach was implemented to assess the degree of importance of certain variables in the proposed regression problem [37]. This procedure was based on the variable's correlation, a statistical measure expressing the degree of a linear relationship between two variables ranging from −1 (perfectly negative correlation) to 1 (perfectly positive correlation) such that 0 indicates no correlation. The correlation levels and relationship behaviors between the variables in this problem can be found in Figs. 3 and 4.

Table 1
Variable description and basic statistics: $X_1 \ldots X_7$ are features, and y is the model target.

	Variable	Unit	Mean	Std	Min	Max
X_1	Flow rate	m^3/s	1.455	2.941	0.0001	20.689
X_2	Potential of hydrogen (pH)		6.879	0.178	6.17	7.240
X_3	Turbidity	NTU	18.963	36.309	1.50	180.000
X_4	Electric conductivity	[μS/cm]	42.710	21.475	12.96	133.940
X_5	Dissolved solids	mg/L	54.817	51.032	0.00	242.000
X_6	Watershed area	km^2	139.070	221.819	0.66	925.300
X_7	Apparent color		83.016	66.079	12.00	323.000
y	Phosphate concentration (PO_4^{3-})	mg/L	0.132	0.086	0.02	0.390

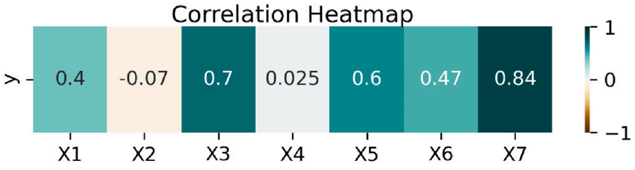

Fig. 3. Correlations between the input features and the target variable.

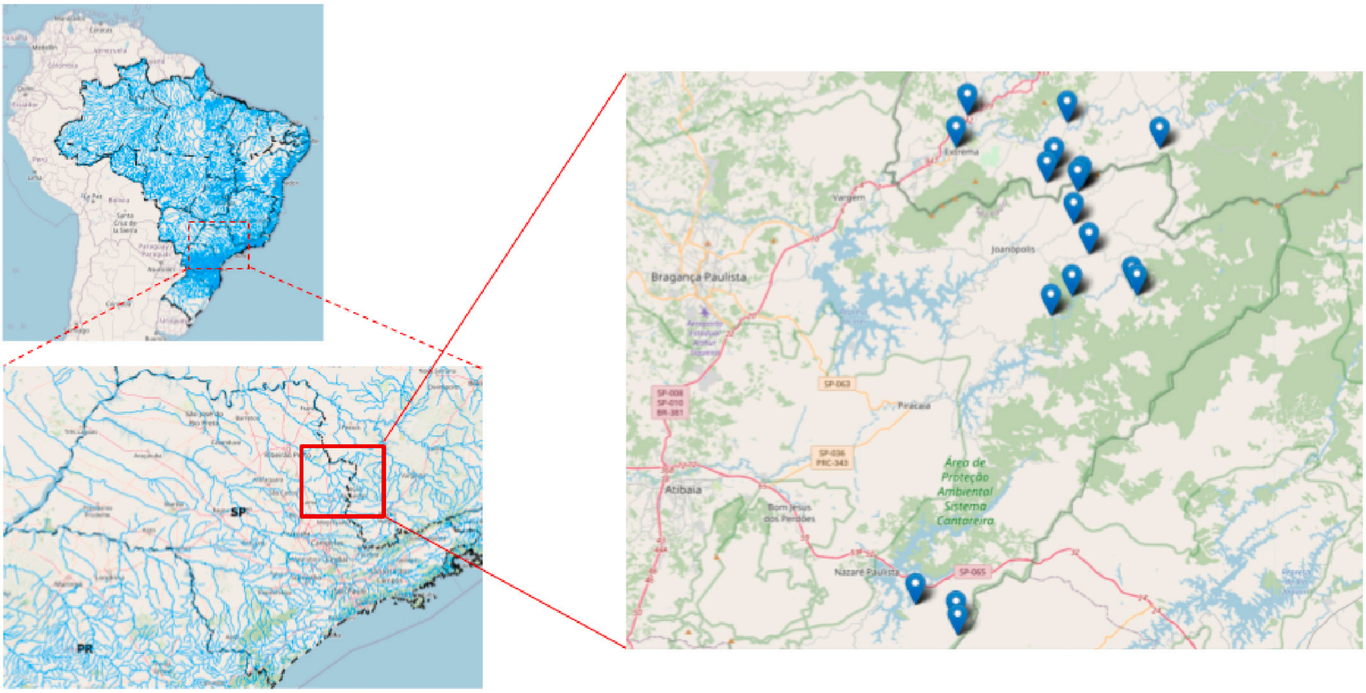

Fig. 2. Map depicting the collection area and highlighting the geographical distribution of the data collected [32].

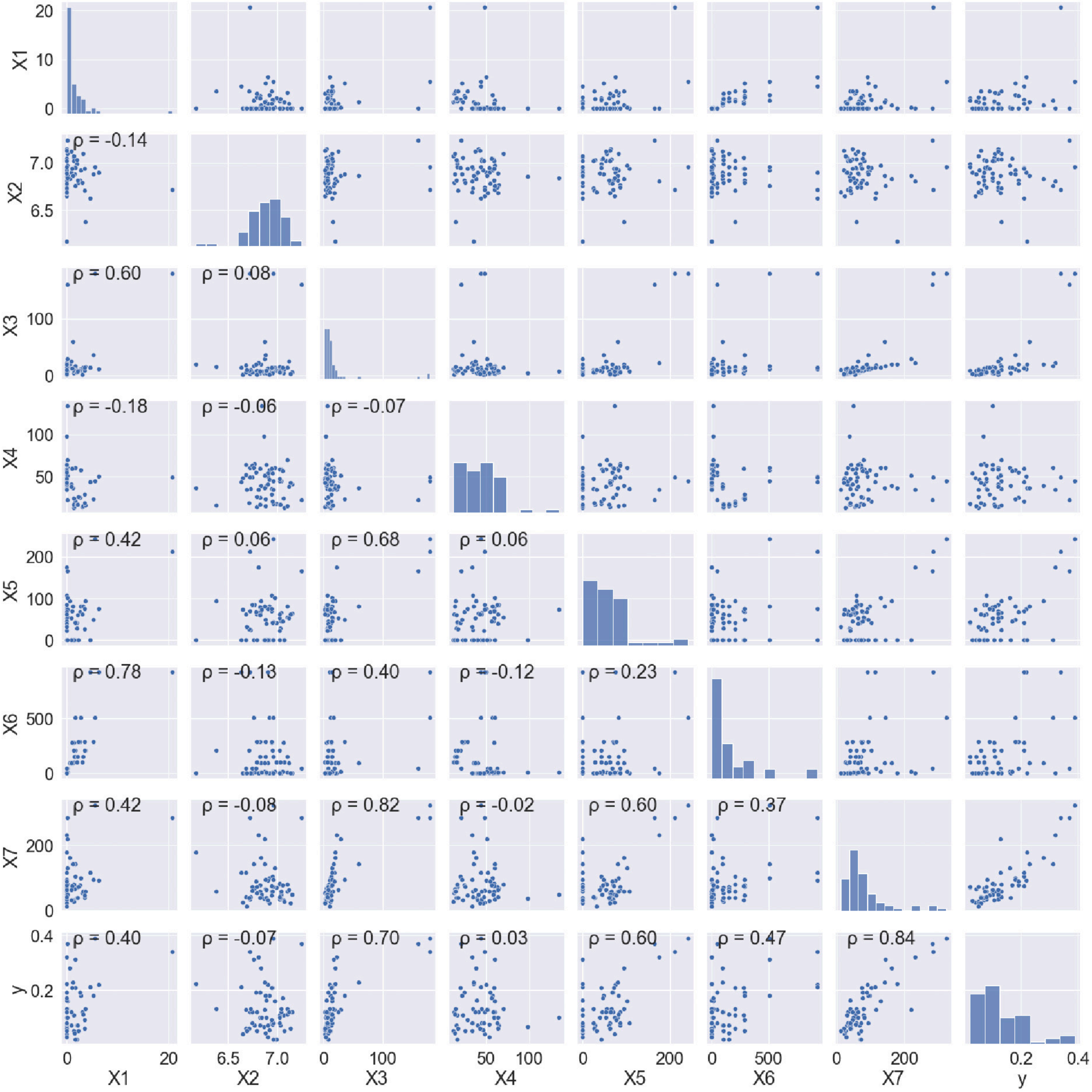

Fig. 4. Pair plot of the variables showing the relationship between them and the correlation degree to two decimal places.

To assess the performance of the proposed data-driven optimized model, four datasets (M1, M2, M3, and M4) were constructed, as detailed in Table 2. The first one (M1) comprises all the variables without any selection. To compose the second dataset (M2), attributes with very weak linear correlations were discarded: X_2 and X_4. Next, the attributes with a correlation lower than 0.5 were sequentially discarded when analyzing the previous subset. Thus, M3 emerged when disregarding the feature X_1, which presents a correlation coefficient equal to 0.4. Subsequently, when disregarding the feature X_6, the subset M4 was created.

Table 2
Feature distribution for each subset.

Dataset/variable set	Features
M1	$X_1, X_2, X_3, X_4, X_5, X_6, X_7$
M2	X_1, X_3, X_5, X_6, X_7
M3	X_3, X_5, X_6, X_7
M4	X_3, X_5, X_7

Table 3
Search space for the ML internal parameters.

Model	Parameters	Range of values
ELM	No. Hidden Neurons (HL)	$[1, 300]$
	L_2 penalization coefficient, C_2	$[0, 1000]$
	Activation Function (G)	0: Identity, 1: Gaussian, 2: Multiquadric, 3: Inverse Multiquadric, 4: ReLU, 5: Swish, 6: Sigmoid
EN	α	$[0, 100]$
	L_1 Ratio, ρ	$[0, 1]$
	Fit Intercept	0: False, 1: True
GB	Learning Rate	$[0.00001, 1]$
	No. Estimators	$[1, 300]$
	Subsamples	$[0.1, 1]$
	Maximum Depth	$[1, 12]$
SVM	Regularization Parameter, C	$[0.001, 100000]$
	Kernel coefficient, γ	$[0.001, 10]$
	Degree	$[1, 3]$
	Kernel	0: linear; 1: poly; 2: rbf; 3: sigmoid

Table 4
Performance metrics and their mathematical expression.

Performance metric	Mathematical expression
R	$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}$
R^2	$1-\frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_i)^2}$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}$
MAE	$MAE=\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\widehat{y}_i\rvert$
MAD	$\frac{1}{250000}\sum_{i=1}^{250000}\parallel\widehat{y}_i-median(\widehat{y})\parallel$
Uncertainty %	$\frac{100\times MAD}{median(\widehat{y})}$

Table 5
Average results and standard deviations for each dataset.

Dataset	Estimator	R	R^2	RMSE	MAE
M1	HHO-	0.815	0.652	0.050	0.038
	ELM	(0.018)	(0.042)	(0.003)	(0.002)
	HHO-EN	**0.824**	**0.670**	**0.049**	**0.037**
		(0.008)	(0.015)	(0.001)	(0.001)
	HHO-GB	0.791	0.619	0.053	0.041
		(0.025)	(0.044)	(0.003)	(0.003)
	HHO-SVR	0.802	0.643	0.051	0.038
		(0.005)	(0.009)	(0.001)	(0.000)
M2	HHO-	0.815	0.653	0.050	**0.037**
	ELM	(0.027)	(0.051)	(0.004)	(0.002)
	HHO-EN	**0.825**	**0.666**	**0.049**	0.038
		(0.011)	(0.021)	(0.002)	(0.001)
	HHO-GB	0.810	0.650	0.051	0.039
		(0.022)	(0.038)	(0.003)	(0.002)
	HHO-SVR	0.800	0.634	0.052	0.038
		(0.003)	(0.005)	(0.000)	(0.000)
M3	HHO-	0.798	0.620	0.053	0.038
	ELM	(0.021)	(0.043)	(0.003)	(0.002)
	HHO-EN	0.816	0.659	**0.050**	**0.037**
		(0.012)	(0.023)	(0.002)	(0.001)
	HHO-GB	**0.818**	**0.664**	**0.050**	0.039
		(0.018)	(0.032)	(0.002)	(0.002)
	HHO-SVR	0.799	0.634	0.052	0.038
		(0.003)	(0.005)	(0.000)	(0.000)
M4	HHO-	0.798	0.631	0.052	**0.037**
	ELM	(0.014)	(0.023)	(0.002)	(0.001)
	HHO-EN	**0.810**	**0.652**	**0.050**	0.038
		(0.010)	(0.018)	(0.001)	(0.001)
	HHO-GB	0.800	0.629	0.052	0.039
		(0.025)	(0.047)	(0.003)	(0.002)
	HHO-SVR	0.726	0.501	0.060	0.047
		(0.035)	(0.037)	(0.002)	(0.002)

Table 6
Approximate median values of the distribution of the parameters.

ML model	Internal parameters	Dataset			
		M1	M2	M3	M4
HHO-ELM	C_2	250	230	280	430
	HL	9	8	5	7
	G	Identity	Identity	Identity	Identity
HHO-EN	L_1 ratio	0	0	0.025	0.025
	α	0.7	0.8	7	6
	Intercept	False	True	True	True
HHO-GB	LR	0.11	0.11	0.095	0.1000
	Maximum depth	6	6	5	5
	No. estimators	143	145	135	122
	Subsample	0.50	0.50	0.40	0.45
HHO-SVR	C	26,137	28,152	27,578	0.001
	γ	2.0	2.5	2.5	0.002
	Degree	1	1	1	1
	Kernel	Poly	Poly	Poly	Linear

2.3. *Leave one-out cross-validation*

Cross-validation is a widely used technique in ML problems for evaluating the generalizability and performance of a model. The fundamental concept is to split the available data into training and validation sets. The training set is used to fit the ML model, while the validation set is used to assess it. This method is performed numerous times with different data splits to generate a more trustworthy assessment of the model's performance since the average resulting validation set performance rate estimates the test performance rate [38].

Among the possible cross-validation approaches, Leave One-Out was chosen for this work, mainly because of the size of the available database. This technique consists of taking each instance of the database to compose the validation set once, and the remaining observations are used to train the model, as illustrated in Fig. 1. In this case, the LOO algorithm was used to create training and test sets instead of training and validation sets. However, another cross-validation technique, called K-fold, was applied to each training set generated by the LOO technique to evaluate the model's hyperparameters (solutions) determined by the optimization algorithm. This procedure is also illustrated in Fig. 1.

2.4. *Machine learning models*

2.4.1. *Elastic Net*

Elastic Net is a regression method that performs, simultaneously with the selection of variables, a regularization process. Regularization is a procedure for when the model is overfitting; that is, the model is very well adjusted to the training data but does not generate good results with the test data.

Regularization can be performed using penalty L1 or penalty L2. The penalty L1 is called the ‘Lasso’ regression, and the penalty L2 is called the ‘Ridge’ regression. Lasso regression tries to adjust the model coefficients to zero, leaving only a small subset of non-zero coefficients. The ridge penalty does not seek to adjust the coefficients to 0. However, the method shrinks the coefficients as close as possible to 0. Thus, ridge regression does not remove the predictors in the model selection [39].

2.4.2. *Extreme Learning Machine*

The Extreme Learning Machine (ELM), proposed by Huang et al. [40], is an artificial feedforward neural network composed of only one hidden layer. Furthermore, its neurons have randomly initialized weights and do not need to be adjusted during training. Therefore, only the output layer is trained, making the process much faster than in other traditional neural networks. In addition to its learning speed, ELM has advantages such as its generalization capability and convenience in modeling [41]. One can enhance their comprehension of the algorithm's behavior by scrutinizing its mathematical formulation, commencing with the following Eqs. (1) and (2).

$$\sum_{i=1}^{\widetilde{N}}\beta_i g(w_i.x_j+b_i)=o_j, j=1,\dots,N \tag{1}$$

$$\sum_{i=1}^{\widetilde{N}}\beta_i g(w_i.x_j+b_i)=t_j, j=1,\dots,N \tag{2}$$

According to [40], the variables have the following meaning in this context: $\widetilde{N}$ represents the number of neurons in the hidden layer. The weight vector of neuron i in the hidden layer is represented by w_i, while the bias of neuron i in the hidden layer is denoted by b_i. Additionally, x_j and t_j are N input patterns used in the process. Finally, the weight vector between hidden neuron i and the output layer is represented by β_i.

2.4.3. *Support Vector Machine*

Support Vector Regression (SVR) is a supervised ML algorithm based on the principle of structural risk minimization and is a variant of

Table 7
Uncertainty analysis.

Dataset	Model	NF	MPE	Median	MAD	Uncertainty	RMSE
M1	HHO-ELM	7	−0.0001 (0.0502)	0.227 (0.018)	0.082 (0.007)	36.66 (4.64)	0.050 (0.002)
	HHO-EN	7	+0.0003 (0.0488)	0.225 (0.014)	0.080 (0.007)	35.73 (4.90)	0.048 (0.001)
	HHO-GB	7	−0.0017 (0.0523)	0.265 (0.016)	0.059 (0.010)	22.58 (5.41)	0.052 (0.003)
	HHO-SVR	7	−0.0013 (0.0510)	0.255 (0.008)	0.054 (0.002)	21.47 (0.85)	0.051 (0.000)
M2	HHO-ELM	5	+0.0003 (0.0501)	0.216 (0.014)	0.084 (0.008)	39.41 (5.83)	0.050 (0.002)
	HHO-EN	5	−0.0019 (0.0493)	0.222 (0.014)	0.084 (0.013)	38.26 (7.48)	0.049 (0.001)
	HHO-GB	5	−0.0008 (0.0500)	0.268 (0.015)	0.062 (0.011)	23.38 (5.05)	0.050 (0.001)
	HHO-SVR	5	−0.0050 (0.0515)	0.253 (0.008)	0.058 (0.003)	23.20 (1.21)	0.051 (0.000)
M3	HHO-ELM	4	−0.0001 (0.0506)	0.239 (0.012)	0.081 (0.015)	34.37 (8.24)	0.050 (0.002)
	HHO-EN	4	−0.0011 (0.0499)	0.240 (0.005)	0.074 (0.011)	31.05 (4.54)	0.049 (0.001)
	HHO-GB	4	−0.0017 (0.0493)	0.267 (0.015)	0.061 (0.011)	23.28 (5.15)	0.049 (0.002)
	HHO-SVR	4	−0.0050 (0.0515)	0.253 (0.008)	0.058 (0.003)	23.19 (1.20)	0.051 (0.000)
M4	HHO-ELM	3	+0.0004 (0.0522)	0.220 (0.020)	0.103 (0.026)	47.67 (15.0)	0.052 (0.001)
	HHO-EN	3	−0.0012 (0.0505)	0.226 (0.004)	0.080 (0.011)	35.48 (5.15)	0.050 (0.001)
	HHO-GB	3	+0.0006 (0.0524)	0.240 (0.015)	0.062 (0.010)	26.12 (5.24)	0.052 (0.002)
	HHO-SVR	3	−0.0037 (0.0597)	0.188 (0.005)	0.054 (0.007)	29.09 (3.59)	0.059 (0.002)

Support Vector Machines (SVM) used to solve regression problems [42]. The method is considered robust and efficient, especially when the number of samples is smaller than the number of features.

SVR aims to generate the maximum number of support vectors with reduced error values to separate the data efficiently. To achieve this, regression analysis is used to find the best-separating line or surface that minimizes prediction error [43]. Unlike other regression methods, SVR seeks to reduce the upper bound on the generalization error and can produce linear or nonlinear predictions depending on the kernel function used. This means that SVR is capable of handling regression problems where the relationship between the independent variables and the dependent variable is nonlinear [44].

One of the main benefits of SVR is its ability to work well with a clear separation margin. It is effective when the number of dimensions is greater than the number of samples, making it an ideal algorithm for complex datasets. However, there are several limitations associated with the SVR. For example, this approach does not work well when dealing with large datasets due to the required training time, and it presents difficulties when dealing with noisy datasets or when target classes overlap [45].

2.4.4. *Gradient Boosting Machine*

The Gradient Boosting Machines (GBM) are ML models highly flexible ML models that can be customized for any specific data-driven task [46]. According to [47,48], *Boosting* is a technique used for generating a strong learner in an iterative way by combining weak learners. Therefore, since powerful learners are generated by combining weak learners, their performance can be improved by refining the predictions of the base learners [46]. One way to do this is through the application of gradient descent, which gives rise to the formulation of GBM. Thus, the fundamental concept of this technique is to build new base learners that have a maximum correlation with the ensemble's overall negative gradient of the loss function [46].

2.5. *Harris hawks optimizer metaheuristic*

Heidari et al. [49] presented the HHO, an algorithm inspired by Harris Hawks' hunting behavior. Each bird in the algorithm represents a potential solution, and the cooperative hunting behavior corresponds to exploring, the transition from exploration to exploitation and exploiting. We can observe Fig. 5 and have a general overview of the model. Next, we delve into each phase to comprehend its operation.

2.5.1. *Exploration*

During exploration, birds search for elevated locations to utilize their advanced eyesight for potential prey. The birds use two equiprobable strategies: analyzing the position of prey and other hawks when $b < 0.5$ or randomly searching for elevated locations while considering the range of attack members for $b \geq 0.5$.

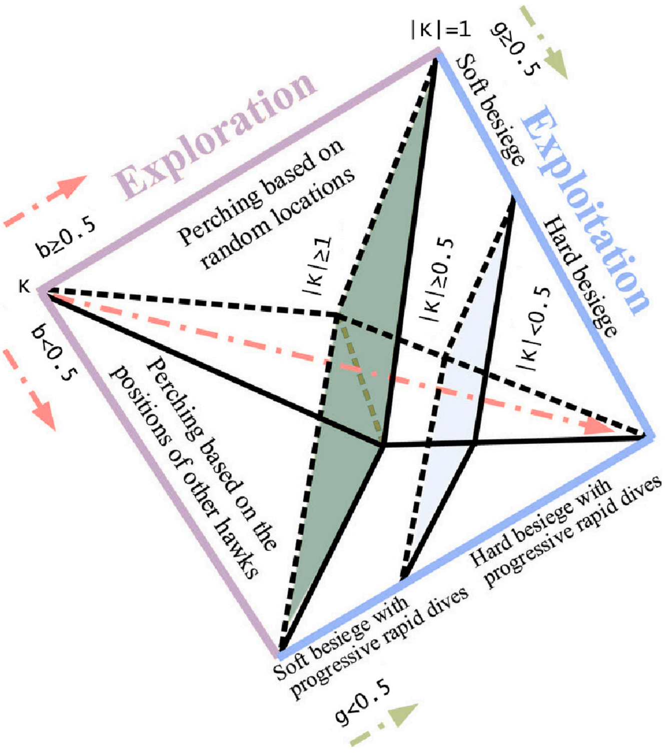
|K|=1
g≥0.5
Soft besiege
Exploitation
Hard besiege
Exploration
Perching based on random locations
b≥0.5
K
|K|≥1
|K|≥0.5
|K|<0.5
b<0.5
Perching based on the positions of other hawks
Hard besiege with progressive rapid dives
Soft besiege with progressive rapid dives
g<0.5

Fig. 5. Phases of HHO [49].

This behavior is mathematically represented in Eq. (3).

$$W(t+1) = \begin{cases} W_{rand}(t) - g_1|W_{rand}(t) - 2g_2W(t)| & b \geq 0.5 \\ (W_{prey}(t) - W_m) - g_3(LB + g_4(UB - LB)) & b < 0.5 \end{cases} \tag{3}$$

This study considers that $W(t+1)$ represents the updated location of hawks in step $t+1$. $W_{rand}(t)$ represents a hawk that has been selected randomly from the present population, while $W(t)$ represents the current position of the predator. The variables $g1$, $g2$, $g3$, $g4$, and b are updated in each iteration and represent random numbers within the range of (0,1). Additionally, UB and LB delimit the lower and upper bounds of the variables, respectively. The position of the prey is represented by $W_{prey}(t)$, while the following Eq. (4) gives the average location of the hawks:

$$W_m(t) = \frac{1}{N}\sum_{i=1}^{N} W_i(t) \tag{4}$$

The expression N represents the complete count of hawks, while $W_i(t)$ represents the spatial coordinates of each hawk at iteration t.

2.5.2. Transition from exploration to exploitation

In this phase, the action determining how hawks behave will be the prey's energy. The following Eq. (5) represents the prey's energy

$$K = K_0\left(1 - \frac{t}{T}\right) \tag{5}$$

where K represents the prey's energy, K_0 denotes the initial power, and the maximum number of iterations is represented by T. The variable K_0 can change from -1 to 1 at each iteration. We consider that the prey has energy when the energy goes from 0 to 1. When the power increases from 0 to -1, the target becomes tired. When $|K| \geq 1$, the hawks search for different areas to explore the prey's location. For $|K| < 1$, HHO tries to find the neighboring solution during the exploration phase.

2.5.3. Exploitation

In this stage, hawks attack their prey, and HHO has four possible strategies: soft besiege, soft besiege with progressive rapid dives, hard besiege, and hard besiege with progressive rapid dives. The variables g and K determine these strategies, representing the chance of escape and prey energy, respectively.

The Soft Besiege behavior of hawks is characterized by $g \geq 0.5$ and $|K| \geq 0.5$, indicating that the prey has some remaining energy but ultimately fails to escape. During this approach, hawks surround their prey gently to exhaust them before launching a surprise attack. Eqs. (6) and (7) model this behavior of Harris' Hawks.

$$\Delta W(t) = W_{prey}(t) - W(t) \tag{6}$$

$$W(t+1) = \Delta W(t) - |JW_{prey}(t) - W(t)| \tag{7}$$

In each iteration, the value of variable J, which represents the intensity of the prey's evasive jump, is modified to replicate the dynamics of its locomotion. It is expressed as $J = 2(1 - g_5)$, where g_5 is a random number within the range $(0, 1)$. On the other hand, $\Delta W(t)$ signifies the discrepancy between the prey's position vector and its current position in iteration t.

In the Hard Besiege stage, when the prey has little energy to escape, the hawks skip the surrounding area and direct attack. This happens when $g \geq 0.5$ and $|K| < 0.5$. Eq. (8) represents this scenario by updating the current positions.

$$W(t+1) = W_{prey}(t) - K|\Delta W(t)| \tag{8}$$

Soft besiege with progressive rapid dives updates hawk positions when prey can still escape ($|K| \geq 0.5$), and hawks construct a soft besiege ($g < 0.5$). Hawks choose the best dive toward the prey through multiple movements, as evaluated by Eq. (9). If deemed necessary, the team may execute rapid descent using a Levy flight (LF) strategy to enhance their ability to exploit resources, as mathematically represented by Eq. (10). LF mimics natural prey movements, particularly those of rabbits, during the escape phase, during which the hawk dives around the target.

$$H = W_{prey}(t) - K|JW_{prey}(t) - W(t)| \tag{9}$$

$$Y = H + L \times LF(D) \tag{10}$$

The random vector L is represented by the product of size one and the Levy flight function, LF(D), where D is the problem dimension. This function can be calculated as shown in Eq. (11).

$$LF(x) = 0.01 \times \frac{\sigma \times \mu}{\| \nu \|^{\frac{1}{\beta}}}, \quad \sigma = \left(\frac{\sin\frac{\pi\beta}{2} \times \Gamma(1+\beta)}{2^{\frac{\beta-1}{2}} \times \Gamma\left(\frac{1+\beta}{2}\right) \times \beta} \right)^{\frac{1}{\beta}} \tag{11}$$

The default constant β is 1.5, and the variables μ and ν are random values ranging from 0 to 1. The process of updating the spatial coordinates of hawks during the soft besiege phase involves implementing Eq. (12). Recall that to find the values of H and Y. We need to use the 9 and 10 equations.

$$W(t+1) = \begin{cases} H & \text{if} \quad F(H) < F(W(t)) \\ Y & \text{if} \quad F(Y) < F(W(t)) \end{cases} \tag{12}$$

In the Hard besiege with progressive rapid dives stage, hawks aim to catch and kill prey when $|K| < 0.5$ and $g < 0.5$. To achieve this, they use a hard besiege strategy and apply Eq. (13) to decrease the distance between themselves and the prey. The values of H and Y are obtained by calculating the rules described in Eqs. (14) and (15). Finally, to determine $W_m(t)$ from Eq. (14), we need to use Eq. (4).

$$W(t+1) = \begin{cases} H & \text{if} \quad F(H) < F(W(t)) \\ Y & \text{if} \quad F(Y) < F(W(t)) \end{cases} \tag{13}$$

$$H = W_{prey}(t) - K|JW_{prey}(t) - W_m(t)| \tag{14}$$

$$Y = H + L \times LF(D) \tag{15}$$

2.6. Feature importance

The Shapley Method (SM) is a cooperative game theory method for estimating feature importance in ML modeling [50]. SM provides a systematic way to attribute the contribution of each feature to the prediction or outcome of a model. The mathematical formulation of the SM can be represented as follows:

$$\phi_i = \frac{1}{N!}\sum_{S \subseteq N \setminus \{i\}} \binom{|S|}{|N| - |S| - 1}^{-1} [V(S \cup \{i\}) - V(S)]$$

where N is the set of features, $V(S)$ denotes the model's prediction when considering the feature set S, and ϕ_i represents the Shapley Coefficient (SC) for feature i. The summation considers all possible subsets S excluding the feature i and calculates the difference in model predictions with and without the inclusion of feature i. The division factor accounts for the number of possible orderings of the features.

SM provides a consistent allocation of importance across features, ensuring that each feature's contribution is measured accurately and without bias. This property is particularly useful when dealing with correlated or redundant features. Shapley's approach allows interpretability, as it clearly explains the relative influence of different features in the model's output. However, this approach can be computationally expensive, especially for models with a large number of features, as evaluating the model requires evaluating multiple subsets of features [51]. Additionally, the SC is not unique, meaning that different orderings of features can yield different results, although the average value remains the same. SM is a robust framework for feature importance estimation, but careful consideration should be given to its computational complexity and potential nonuniqueness.

2.7. Performance metrics

In this paper, four performance metrics were used to evaluate the model's results: the Pearson correlation coefficient (R), which measures the degree of linear relationship between two variables. The coefficient of determination (R^2), defined as the square of the correlation, can also be understood as a statistical measurement that quantifies how well a model predicts an outcome. Root Mean Square Error and Mean Absolute

Error are different performance indicators that calculate the average difference between the predicted values by a model and their true values. These metrics are described in Table 4, which also presents their mathematical expression.

For the calculation of the Mean Absolute Deviation (MAD) and, consequently, Uncertainty metrics, the input values were synthetically generated from a normal distribution with the lower and upper bounds being the minimum and maximum values of each parameter, as shown in Table 1. For each studied scenario, 250,000 outcomes were generated and used, each representing one daily measurement.

3. Computational experiments

The results obtained with the proposed hybrid model are presented in this section. The performance of the model was evaluated using various metrics to assess its predictive accuracy for phosphate pollution indicators in rivers and watercourses.

The model's predictive capabilities were assessed by comparing its predictions with actual phosphate pollution data collected from different monitoring sites. MAE, RMSE, R, and R^2 were calculated to quantify the model's accuracy in capturing the variations in phosphate concentrations.

The computational experiments were conducted based on pandas [52,53], NumPy [54], scikit-learn framework [55], seaborn [56], scipy [57], matplotlib [58], mealpy [59] and implementations adapted from it.

3.1. Simulation results

The prediction performances of the computer models are presented in Table 5, reflecting the average results obtained from 62 simulations conducted using the LOO technique with the testing sets. These results show that the HHO-EN model exhibits superiority in 3 out of the 4 evaluated cases, as indicated by the bold values in the table. Specifically, when considering the full set of attributes (M1), the linear model, i.e. EN, yields the best average performance across all the metrics, as denoted by the bold values. Similar trends are observed in M2 and M4, except that the ELM model optimized by HHO (HHO-ELM) attains the best average MAE values. Furthermore, in the case of M3, the GB optimized by HHO (HHO-GB) algorithm yields better results for the R, R^2, and RMSE metrics, while the linear model excels in achieving the lowest values for the RMSE and MAE error assessments. Although the HHO-GB model performs better on problem M3, we analyzed whether there is a significant difference in the comparative results between these two models.

Evaluating the performance measures R and R^2, one can observe that by taking the M1 and M2 datasets, one can see that the HHO-ELM and HHO-EN models exhibit consistent performance, showing a maximum variation of 0.1% in R and 0.4% in R^2. However, their standard deviations demonstrate a more significant increase, reaching up to 0.9% in both metrics. This indicates that the attributes X_2 (pH) and X_4 (Electric Conductivity) have little importance in these models. Consequently, the observed compartmentalization of performance was to be expected since even though most of the information from X_2 and X_4 is not representative, a small portion of the importance of these variables in these models is distributed over the uncertainty of the measurements.

The HHO-GB model exhibited a substantial average increase, from 0.791 and 0.619 to 0.810 and 0.650, respectively, in R and R^2. Moreover, its standard deviations were reduced, albeit with smaller variations of 0.3% and 0.6%, respectively. This stands in contrast to the behavior of the HHO-ELM and HHO-EN models, where variables X_2 and X_4 were found to carry redundant information from other variables, adversely affecting model performance. Consequently, by excluding these variables in the M2 dataset, the performances of the HHO-GB model improved significantly. On the other hand, when the focus is on the SVR model optimized by HHO (HHO-SVR), one can observe a decrease in both the average performance and its standard deviation. This observation suggested that these variables played a crucial role in the prediction model. When the average performance was affected, the standard deviations indicated less variation.

Regarding dataset M3, but still considering the performance measures R and R^2, the removal of attribute X_1 (Flow Rate) caused a reduction in the mean performance and standard deviations in the HHO-ELM model, both in terms of R and R^2. On the other hand, HHO-EN was affected by a reduction in average performance and a slight increase in its standard deviations of 0.1% (R) and 0.2% (R^2). Conversely, the HHO-GB performed even better without X_1, with an increase of approximately 0.8% in average R and 1.4% in average R^2 and a reduction in its standard deviation. This again demonstrates that X_1 presented redundant information that impaired the performance of this model. On the other hand, the HHO-SVR held steady with virtually no change in performance, indicating that the X_1 variable was not representative of this model. Finally, taking M4, it can be seen that discarding attribute X_6 (watershed area) caused a reduction in the average performance of all the models, except for HHO-ELM, which performed better in terms of the average R^2 and held the average R constant, managing to reduce the standard deviations in both metrics. Despite the decrease in average performance, the HHO-EN model managed to reduce the variability of the results by approximately 0.2% and 0.5% for R and R^2, respectively. On the other hand, the HHO-GB and HHO-SVR methods showed a decrease in performance and an increase in variability, with these aspects being observed more intensely for the HHO-SVR. Thus, these results suggest that the variable X_6 is an important problem attribute.

In terms of the RMSE and MAE, there was no significant variation between the best models in each dataset, as indicated by the bold values in the columns of Table 5. Regarding the RMSE, the error values associated with each model show a slight deterministic increase when transitioning from M2 to M3. However, for the MAE, there appears to be no general variation, with only a minor decrease in the error value of the HHO-GB model occurring from M1 to M2. These differences are subtle and fall within the standard deviation. One notable exception to this pattern is evident when considering the HHO-SVR model in datasets M3 to M4. In this case, there was a considerable increase (exceeding the bounds of the standard deviations) in the values of both error measures, with RMSE increasing from an average value of 0.052 to 0.060 and MAE increasing from 0.038 to 0.047.

As observed in Fig. 6, the variables from the M3 dataset were found to be beneficial for the performance of the HHO-GB model while having a detrimental effect on the other models. The HHO-GB model demonstrates a strong ability to capture nonlinear relationships among the variables. Interestingly, despite having fewer variables than M1 and M2, the M3 dataset resulted in better performance for the HHO-GB model than did the other datasets. Moreover, the highest R^2 performances among the best models are associated with HHO-GB in all the datasets.

3.2. Parameteric analysis

A study of the parameters determined by the optimization algorithm for each model was also conducted to evaluate their distributions at the end of the independent runs. This analysis allowed us to identify the most recurrently determined values for the parameters by the optimizer. Table 6 presents the medians of the distributions of the internal parameters for the HHO-ELM, HHO-EN, HHO-GB, and HHO-SVR models in each dataset, considering the 62 executions sorted by LOO. It is crucial to clarify the discrepant values observed in M4 concerning the HHO-SVR model's C and γ parameters, which predominantly resulted in the selection of the linear kernel in approximately 86.8% of the cases.

The parameter γ exclusively applies to nonlinear kernels such as 'rbf', 'poly', and 'sigmoid', whereas parameter C exhibits an inverse relationship with the regularization strength. Thus, in the case of noisy

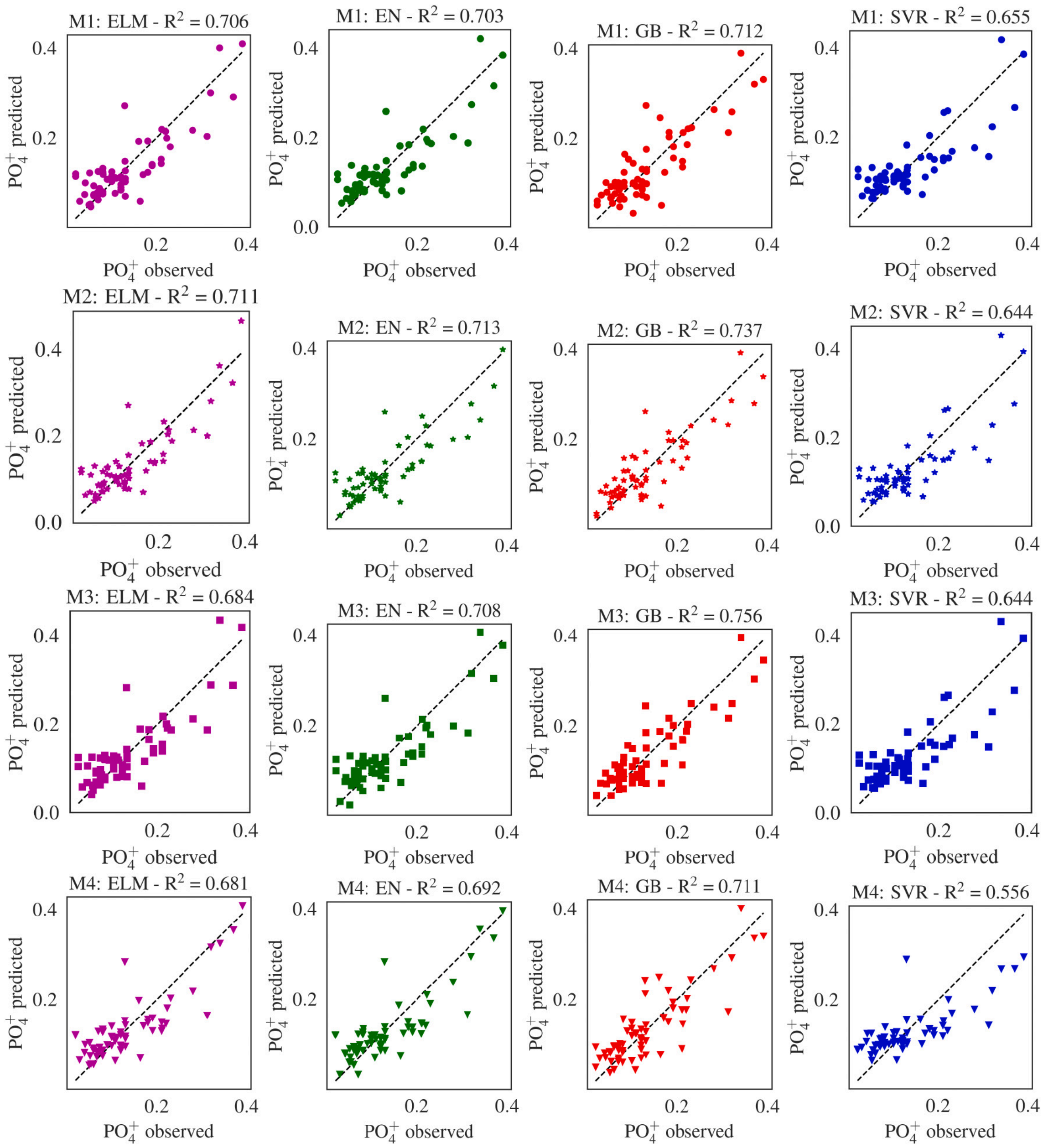

Fig. 6. Scatter plots for the best models.

observations, increasing the regularization by decreasing the value of the parameter C is necessary. In the context of M1, M2, and M3, the gamma values correspond to the runs when HHO determined the kernel parameter to be linear, which occurred with a frequency of 42%, 32.5%, and 13.2% of the independent runs, respectively, for these datasets.

3.3. Feature importance analysis

Fig. 7 illustrates the SC assigned to seven variables using the SM for feature importance. This figure shows the varying importance levels across the different variables and allows for a visual comparison of their impacts. The horizontal bar chart presents the magnitude of the SC for each variable, providing insights into their relative contributions to the model's predictions. Each bar represents a variable, and its height corresponds to the SC. The higher the coefficient is, the greater the influence of the respective variable on the model's output. For instance, variable X_7 exhibits the highest SC, indicating its significant contribution to the model's predictions. In contrast, variables X_2, X_3, and X_4 display relatively lower coefficients, suggesting relatively less influence.

3.4. Uncertainty analysis

Table 7 shows the uncertainty analysis for this study. The first column indicates the specific feature subset (M1, M2, M3, M4) used for training and evaluating the model. The second column shows the hybridized models. The third column displays the number of features in the dataset (NF), and the fourth column shows the Mean Prediction Error (MPE). The median values of the predicted pollutant concentration are shown in the fifth column. The sixth column is the Mean Absolute Deviation (MAD), and the seventh column displays the uncertainty, the percentile relationship between the MAD, and the median value of the predicted concentration values. The RMSE is shown in the last column.

As observed in Fig. 7, there is a general trend of an inverse relationship between RMSE and Uncertainty. Models with lower RMSE (better accuracy) tend to exhibit higher Uncertainty, highlighting the challenge of achieving both high accuracy and high confidence in

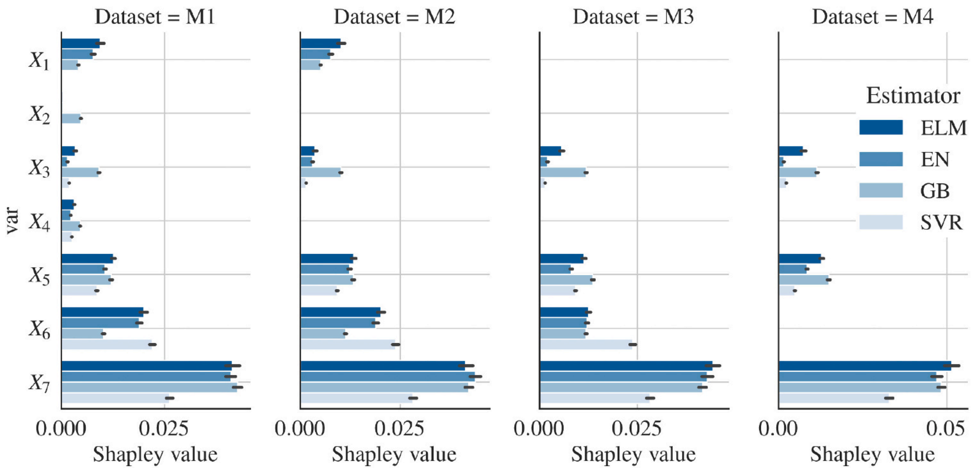

Fig. 7. Averaged Shapley values.

predictions. The simulation results show that the number of features (NF) has a varying impact on different models. The performance of HHO-EN remains relatively stable across different feature subsets, suggesting that it is robust and less sensitive to the specific choice of input variables. On the other hand, the performance of HHO-GB appears to be more dependent on the feature subset. HHO-ELM exhibits higher RMSE and Uncertainty than HHO-EN, and their performance also varies across different feature subsets. HHO-EN demonstrates a favorable balance between accuracy and uncertainty across most datasets, making it a reliable choice for phosphate prediction; while HHO-GB can achieve high accuracy with certain feature subsets, its performance is less consistent and comes with higher uncertainty.

Fig. 8 presents a scatter plot comparing the Uncertainty (%) and RMSE for all the models (HHO-ELM, HHO-EN, HHO-GB, HHO-SVR) and across different datasets (M1, M2, M3, M4, according to the number of features) used in the study. Each point represents a specific model-dataset combination, allowing for a comparison of their performance and uncertainty characteristics. The x-axis represents the uncertainty, calculated as the ratio of the MAD to the median of the predicted phosphate concentrations, expressed as a percentage. Higher uncertainty values indicate a wider spread and lower confidence in the model's predictions. The y-axis shows the RMSE values, which measure the average magnitude of prediction errors the model makes. Lower RMSE values indicate better prediction accuracy.

A roughly Pareto front can be observed from Fig. 8, demonstrating the inherent trade-off between accuracy and uncertainty in prediction models. A Pareto front represents the set of solutions where improving one objective without worsening another is not possible. In this context, the two objectives are minimizing RMSE (improving accuracy) and minimizing Uncertainty (increasing confidence). Generally, models with lower RMSE (better accuracy) tend to have higher uncertainty (lower confidence) and vice versa. This reflects the challenge of balancing the need for precise predictions with the inherent variability and uncertainty in complex environmental systems. The points representing the HHO-GB (M2 and M3) model generally appear closer to the low RMSE and low uncertainty region, suggesting a favorable balance between accuracy and confidence in its predictions. Observing Tables 1 and 2 and the variables that consist of M2 and M3, one can observe that the turbidity, dissolved solids, watershed area, and apparent color are the most relevant features impacting the phosphate prediction. While some HHO-EN and HHO-ELM models exhibit high accuracy (low RMSE), they often come with higher uncertainty, and HHO-SVR tends to have higher RMSE with low uncertainties; Fig. 8 can guide the selection of the most appropriate model based on the desired balance between accuracy and uncertainty.

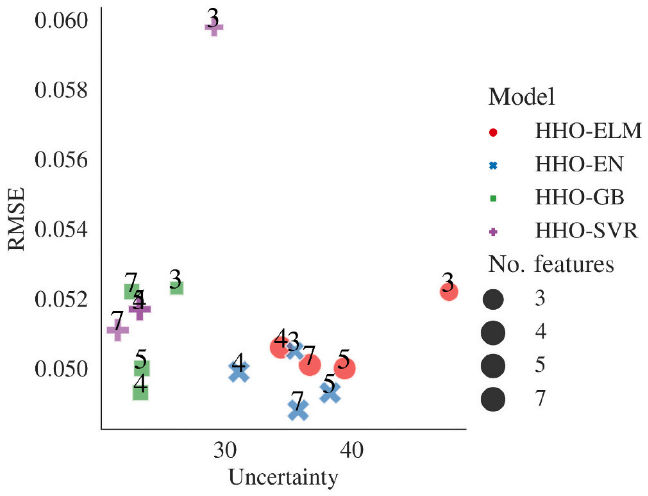

Fig. 8. Uncertainty comparison.

3.5. Discussion

The variable X_7 (Apparent Color) was the most representative of all of the variables, while X_2 (pH), X_3 (Turbidity), and X_4 (Electric Conductivity) were the least relevant for the models. In practice, the relationship between apparent color and phosphate concentration can be seen in general terms from three aspects: organic matter, sediments and algal blooms [60,61]. Elevated dissolved or particulate organic matter levels in aqueous environments can impart a yellow or brown hue, commonly denoted as “apparent watercolor” [60]. The origin of organic matter can be attributed to various sources, including the decomposition of plant matter, runoff from terrestrial regions, or the discharge of wastewater [62]. In some cases, organic matter can bind phosphates, potentially affecting their availability and influencing the correlation between phosphate concentration and watercolor. The presence of suspended sediments in aqueous systems can contribute significantly to the observed watercolor, particularly when they contain fine particles and high concentrations of iron or manganese oxides [63]. Phosphates can adsorb onto sediment particles, and their presence can influence the relationship between phosphate concentration and water coloration [61]. The occurrence of algal blooms, specifically those dominated by specific algal species such as cyanobacteria, can manifest a greenish or bluish-green color [60]. These blooms are often associated with increased nutrient concentrations, including phosphates, which can act as a fertilizer for algal growth [64]. In such cases, a positive correlation between phosphate concentration and water coloration may arise, as higher levels of phosphates provide favorable conditions for algal proliferation, consequently contributing to the observed alterations in color.

Furthermore, looking at sets M1, M2, and M3, it can be seen that variable X_6 (Watershed Area) ranks second in importance among the variables, which may be related to a diversity of factors, such as nutritional inputs, dilution effects, and retention processing. The dimensions of a watershed can influence both the potential sources and magnitude of nutrient inputs, including phosphates [65]. Larger watersheds tend to receive inputs from more extensive land areas, thereby contributing to a greater nutrient load entering the associated water body. Agricultural practices, urbanization, industrial operations, and various other human activities occurring within watersheds can release phosphates into water bodies via mechanisms such as runoff, erosion, or direct discharge [65]. In some cases, larger watersheds may experience a dilution effect, leading to lower phosphate concentrations. Dilution phenomena transpire when the aggregate quantity of phosphates entering a water body from diverse origins is dispersed across a greater volume of water within the entire watershed. The size of the watershed can influence the retention and processing of phosphates within the system. Larger watersheds offer greater prospects for natural processes, such as sedimentation, adsorption onto soils, or biological uptake. These mechanisms can eliminate or diminish phosphate concentrations as water moves through a watershed [66].

The variable X_5 (Dissolved Solids) was also representative of the problem. However, its practical implications are related to the apparent color (X_7) and the flow rate (X_1), although in a less intense way. The correlation between this last variable and the phosphate concentration in aquatic systems can vary depending on several factors, including the source and type of phosphate input, hydrological conditions, and characteristics of the water body [61]. In many cases, an increase in flow rate (e.g., due to rainfall, runoff, or high discharge) can give rise to a dilution effect, leading to a reduction in phosphate concentration [17]. Heightened flow rates can amplify the water volume, thereby diminishing the relative concentration of phosphates within the system. Concerning the transport and redistribution of phosphates within aquatic systems, the flow rate plays a significant role [9]. An increased flow can enhance the movement of phosphates by carrying them downstream or through mixing and stirring within the water column. This dynamic can culminate in heightened dispersion and mixing of phosphates, potentially inducing alterations in their concentration patterns along the flow path [15]. Finally, elevated flow rates can instigate the resuspension of sediments, especially in rivers and streams [67]. Phosphates that are adsorbed or bound to sediment particles can be released into the water column during such events, potentially temporarily increasing phosphate concentrations. This phenomenon can occur in regions characterized by the accumulation of phosphate-rich materials within sediments or in locations where erosion and sedimentation processes occur.

3.6. Model strengths and limitations

The experiments with the M3 dataset demonstrate that the appropriate selection of variables can significantly enhance the performance of HHO-GB models. In this regard, we observed that hybrid models allow for the inclusion of feature selection processes in addition to the automatic adjustment of hyperparameters [68]. Hybrid models with internal feature selection processes can be used to better explore the potential of ML models [69]. When combined with suitable variables, remarkable performances can be achieved [70]. A large number of variables provide additional information for ML modeling. However, if the variables are uninformative, this additional information may deteriorate the predictive capacity of the ML models.

A strength of the proposed model is its ability to integrate data from diverse sources. Incorporating satellite data, such as land use and land cover, vegetation indices, and meteorological data from stations or sensors along watercourses or urban areas, can yield a more comprehensive set of variables for pollutant concentration modeling, leading to a more accurate model. In addition, the model's flexibility allows for incorporating additional pollutants. This can be achieved by simply adjusting the target variable and retraining the model using the framework proposed in this paper. With a robust data collection and technology integration process, the model can be applied to various watercourses to assess diverse pollutant levels. The model can potentially serve as a component within an integrated solution to support decision-making by environmental managers responsible for water resource management and pollution control.

The results presented in this paper show that ML techniques can substantially contribute to predicting phosphate pollution indicators in rivers and watercourses. Using the capabilities of synergistic ML algorithms, environmental engineers can develop accurate models that predict and identify potential instances of phosphate pollution. This predictive capability allows for the implementation of proactive measures, including targeted monitoring programs, appropriate land management practices, and effective mitigation strategies [71].

The proposed model for predicting phosphate concentration in water bodies works primarily as a simulation tool to aid experts in evaluating potential outcomes in different scenarios. The model does not make decisions independently; instead, it assists the specialist in exploring different possibilities, allowing changes to the input variables and thus analyzing the corresponding predictions. The model can also be integrated into broader decision support systems, providing real-time or near-real-time predictions to inform adaptive management strategies. It has the potential to be used as a component within an integrated solution for supporting decision-making by environmental managers responsible for water resource management and pollution control.

The model presented in this paper holds significant promise in environmental management, aiding environmental agencies and policymakers in identifying and prioritizing areas susceptible to phosphate pollution. Additionally, it can assist in optimizing nutrient removal processes in wastewater treatment plants and minimizing the impact of agricultural runoff on water bodies, thus contributing to sustainable water resource management practices. Furthermore, the model can be utilized in risk assessment and early warning systems, enabling the development of predictive tools for the early detection of potential phosphate pollution events. This capability allows for proactive measures to safeguard water quality and public health. Moreover, it serves as a valuable decision-support system for stakeholders involved in various water-related activities, such as drinking water supply, recreational water use, and aquatic ecosystem management. The model allows for building scenarios on phosphate pollution trends and patterns while empowering stakeholders to make informed decisions to mitigate pollution and ensure the sustainability of water resources.

Further research involves using Genetic Programming (GP) models that automatically create mathematical expressions. These models can be computationally simpler and, depending on the complexity of the expressions, can be more interpretable. This approach can be used in future research. Since the dataset is limited in the number of samples collected at each of the different locations, the information provided carries a significant amount of uncertainty, making the problem even more challenging [72]. For this reason, employing simpler regression models may be an advantageous strategy compared to using complex models, as they can extract less detail from the information and introduce less complexity to an already complicated problem. Nevertheless, the results indicate that this approach can be a promising alternative for estimating pollutants in rivers and thus assist in quantifying water quality.

4. Conclusion

This paper proposed an integrated approach using ML models

(Elastic Net, Gradient Boosting, Extreme Learning Machines, and Support Vector Regression) hybridized with the nature-inspired Harris Hawk optimization algorithm to predict phosphate pollution in the Cantareira water system in Brazil. The hybridization of these techniques resulted in a more robust and efficient optimization process, leading to significantly improved accuracy in predicting phosphate pollution levels. The enhanced predictive capabilities empower environmental authorities and water resource managers to make better-informed decisions, strengthening their efforts to protect and restore the integrity of the Cantareira water system.

This study provides an example of an application in the domain of water quality management while highlighting the significance of interdisciplinary approaches. The proposed model addressed the complex challenge of phosphate pollution more comprehensively and effectively by synergizing domain knowledge related to water systems and environmental factors with ML and optimization expertise. The main findings are summarized below.

1. In general, the hybrid HHO-EN model achieved the best performance metrics for all the datasets except for M3. With this dataset, the hybrid Harris Hawk Gradient Boosting (HHO-GB) algorithm yielded the best performance, suggesting that the proper choice of variables can improve the ability of HHO-GB models.
2. The HHO-EN model produced the best-averaged performance metrics among all the experiments ($R = 0.825$, $R^2 = 0.670$, RMSE $=$ 0.049 mg/L, MAE $=$ 0.037 mg/L).
3. The results suggest that HHO-EN is a robust model that performs well on the majority of the datasets tested in this paper.
4. Independent of the dataset, the feature importance analysis reveals that the apparent color is the most relevant feature, followed by the watershed area. Dissolved solids are also representative of the modeling task.

Using ML techniques can improve phosphate measurements in water bodies. These techniques can analyze large datasets faster and more accurately than traditional methods of chemical analysis.

The results validate the effectiveness of the proposed hybrid model in predicting phosphate pollution indicators in rivers and watercourses. Its robust performance, superior predictive accuracy, and ability to identify potential pollution hotspots make it a valuable tool for environmental engineers and water resource managers in mitigating phosphate pollution and safeguarding freshwater quality.

Abbreviations and acronyms

ELM	Extreme Learning Machine
EN	Elastic Net
GB	Gradient Boosting
GBM	Gradient Boosting Machine
GP	Genetic Programming
HHO	Harris Hawks Optimizer
HHO-ELM	Hybrid Harris Hawks Optimizer integrated to Extreme Learning Machine
HHO-EN	Hybrid Harris Hawks Optimizer integrated to Elastic Net
HHO-GB	Hybrid Harris Hawks Optimizer integrated to Gradient Boosting
HHO-SVR	Hybrid Harris Hawks Optimizer integrated to Support Vector Regression
LF	Levy Flight
LB	Lower Bound
LOO	Leave One Out
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
ML	Machine Learning
MPE	Mean Prediction Error
NF	Number of Features
NTU	Nephelometric Turbidity Units
PO_4^{3-}	Phosphate
pH	Potential of Hydrogen
R	Pearson Correlation Coefficient
R^2	Coefficient of Determination
RBF	Radial Basis Function
RMSE	Root Mean Square Error
SC	Shapley Coefficient
SM	Shapley Method
SVM	Support Vector Machine
SVR	Support Vector Regressor
UB	Upper Bound

Funding

The authors acknowledge the support of CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (finance code 001). L.G. acknowledges the support of the funding agency CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (grants 307688/2022-4 and 401796/2021-3).

Code availability

The code can be obtained upon request from the authors.

CRediT authorship contribution statement

Tales H.A. Boratto: Writing – original draft, Software, Data curation. **Deivid E.D. Campos:** Writing – original draft, Visualization, Methodology. **Douglas L. Fonseca:** Writing – original draft, Visualization, Data curation. **Welson F.A. Soares:** Validation, Software, Data curation. **Zaher M. Yaseen:** Validation, Methodology, Investigation, Formal analysis. **Angela Gorgoglione:** Writing – review & editing, Validation, Methodology. **Leonardo Goliatt:** Writing – review & editing, Supervision, 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.

References

[1] A. Ullrich, M. Volk, Application of the soil and water assessment tool (swat) to predict the impact of alternative management practices on water quality and quantity, Agric. Water Manag. 96 (2009) 1207–1217.

[2] A.E. Ulrich, M. Stauffacher, P. Krütli, E. Schnug, E. Frossard, Tackling the phosphorus challenge: time for reflection on three key limitations, Environ. Dev. 8 (2013) 137–144.

[3] H. Tiessen, Framing a rational debate on phosphate use, Environ. Dev. (2013) 145–146.

[4] A.E. Ulrich, M. Stauffacher, P. Krütli, E. Schnug, E. Frossard, Response to the comments on "tackling the phosphorus challenge: time for reflection on three key limitations", Environ. Dev. 8 (2013) 149–151.

[5] T.S. Hussain, A.H. Al-Fatlawi, Remove chemical contaminants from potable water by household water treatment system, Civ. Eng. J. 6 (2020) 1534–1546.

[6] S.D. Latif, A.H. Birima, A.N. Ahmed, D.M. Hatem, N. Al-Ansari, C.M. Fai, A. El-Shafie, Development of prediction model for phosphate in reservoir water system based machine learning algorithms, Ain Shams Eng. J. 13 (2022) 101523.

[7] I.A. Guiamel, H.S. Lee, Watershed modelling of the Mindanao river basin in the Philippines using the swat for water resource management, Civ. Eng. J. 6 (2020) 626–648.

[8] C. Russo, A. Castro, A. Gioia, V. Iacobellis, A. Gorgoglione, Improving the sediment and nutrient first-flush prediction and ranking its influencing factors: an integrated machine-learning framework, J. Hydrol. 616 (2023) 128842.

[9] A. Gorgoglione, A. Gioia, V. Iacobellis, A framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds, Sustainability 11 (2019) 4933.

[10] R. Rodríguez Núñez, M. Pastorini, L. Etcheverry, C. Chreties, M. Fossati, A. Castro, A. Gorgoglione, Water-quality data imputation with a high percentage of missing values: a machine learning approach, Sustainability 13 (11) (2021) 1–17, jun 2021.

[11] W. Shi, M. Huang, Predictions of soil and nutrient losses using a modified swat model in a large hilly-gully watershed of the Chinese Loess Plateau, Int. Soil Water Conserv. Res. 9 (2021) 291–304.

[12] Y. Yuan, L. Koropeckyj-Cox, Swat model application for evaluating agricultural conservation practice effectiveness in reducing phosphorous loss from the western Lake Erie basin, J. Environ. Manag. 302 (2022) 114000.

[13] A. Ahsan, S.K. Das, M.H.R.B. Khan, A.W.M. Ng, N. Al-Ansari, S. Ahmed, M. Imteaz, M.A.U.R. Tariq, M. Shafiquzzaman, Modeling the impacts of best management practices (bmps) on pollution reduction in the Yarra river catchment, Australia, Appl Water Sci 13 (2023).

[14] G.K.G. Cunha, K.P.V. da Cunha, Effects of land use changes on the potential for soil to contribute phosphorus loads in watersheds, Environ. Dev. 45 (2023) 100825.

[15] C. Russo, A. Castro, A. Gioia, V. Iacobellis, A. Gorgoglione, A stormwater management framework for predicting first flush intensity and quantifying its influential factors, Water Resour. Manag. 37 (2023) 1437–1459.

[16] S.K. Gurjar, S. Shrivastava, S. Suryavanshi, V. Tare, Assessment of the natural flow regime and its variability in a tributary of Ganga river: impact of land use and land cover change, Environ. Dev. 44 (2022) 100756.

[17] A. Gorgoglione, A. Castro, V. Iacobellis, A. Gioia, A comparison of linear and non-linear machine learning techniques (pca and som) for characterizing urban nutrient runoff, Sustainability 13 (2021) 2054.

[18] X. Li, J. Yang, Y. Fan, M. Xie, X. Qian, H. Li, Rapid monitoring of heavy metal pollution in lake water using nitrogen and phosphorus nutrients and physicochemical indicators by support vector machine, Chemosphere 280 (2021) 130599.

[19] Z.M. Yaseen, The next generation of soil and water bodies heavy metals prediction and detection: new expert system based edge cloud server and federated learning technology, Environ. Pollut. 313 (2022) 120081.

[20] L. Li, S. Rong, R. Wang, S. Yu, Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: a review, Chem. Eng. J. 405 (2021) 126673.

[21] Z. Yan, W. Han, J. Peñuelas, J. Sardans, J.J. Elser, E. Du, P.B. Reich, J. Fang, Phosphorus accumulates faster than nitrogen globally in freshwater ecosystems under anthropogenic impacts, Ecol. Lett. 19 (2016) 1237–1246.

[22] N.-T. Ha, H.Q. Nguyen, N.C.Q. Truong, T.L. Le, V.N. Thai, T.L. Pham, Estimation of nitrogen and phosphorus concentrations from water quality surrogates using machine learning in the Tri An Reservoir, Vietnam, Environ. Monit. Assess. 192 (2020) 1–20.

[23] A. Bhattarai, S. Dhakal, Y. Gautam, R. Bhattarai, Prediction of nitrate and phosphorus concentrations using machine learning algorithms in watersheds with different landuse, Water 13 (2021) 3096.

[24] T. Paepae, P.N. Bokoro, K. Kyamakya, A virtual sensing concept for nitrogen and phosphorus monitoring using machine learning techniques, Sensors 22 (2022) 7338.

[25] J.-S. Chou, C.-C. Ho, H.-S. Hoang, Determining quality of water in reservoir using machine learning, Ecol. Inform. 44 (2018) 57–75.

[26] M.A.M. Yunus, M. Faramarzi, S. Ibrahim, W.A.H. Altowayti, G.P. San, S. C. Mukhopadhyay, Comparisons between radial basis function and multilayer perceptron neural networks methods for nitrate and phosphate detections in water supply, in: 2015 10th Asian Control Conference (ASCC), 2015, pp. 1–6, https://doi.org/10.1109/ASCC.2015.7244593.

[27] L.Q. Shen, G. Amatulli, T. Sethi, P. Raymond, S. Domisch, Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework, Sci. Data 7 (2020) 161.

[28] N. Wang, L. Xie, Y. Zuo, S. Wang, Determination of total phosphorus concentration in water by using visible-near-infrared spectroscopy with machine learning algorithm, Environ. Sci. Pollut. Res. 30 (2023) 58243–58252.

[29] X. Zhu, J. Ma, Recent advances in the determination of phosphate in environmental water samples: insights from practical perspectives, TrAC Trends Anal. Chem. 127 (2020) 115908.

[30] Y. Yokoyama, T. Danno, M. Haginoya, Y. Yaso, H. Sato, Simultaneous determination of silicate and phosphate in environmental waters using pre-column derivatization ion-pair liquid chromatography, Talanta 79 (2009) 308–313.

[31] H.P. Jarvie, J.A. Withers, C. Neal, Review of robust measurement of phosphorus in river water: sampling, storage, fractionation and sensitivity, Hydrol. Earth Syst. Sci. 6 (2002) 113–131.

[32] D. Taffarello, R. Srinivasan, G.S. Mohor, J.L.B. Guimarães, M. do Carmo Calijuri, E. M. Mendiondo, Modeling freshwater quality scenarios with ecosystem-based adaptation in the headwaters of the Cantareira system, Brazil, Hydrol. Earth Syst. Sci. 22 (2018) 4699–4723.

[33] R.O. Silva, C.M. Saporetti, Z.M. Yaseen, E. Pereira, L. Goliatt, An approach for total organic carbon prediction using convolutional neural networks optimized by differential evolution, Neural Comput. & Applic. 35 (2023) 20803–20817.

[34] L. Goliatt, C. Saporetti, E. Pereira, Super learner approach to predict total organic carbon using stacking machine learning models based on well logs, Fuel 353 (2023) 128682.

[35] T.H. Boratto, A.A. Cury, L. Goliatt, Machine learning-based classification of bronze alloy cymbals from microphone captured data enhanced with feature selection approaches, Expert Syst. Appl. 215 (2023) 119378.

[36] T.H.A. Boratto, C.M. Saporetti, S.C.A. Basilio, A.A. Cury, L. Goliatt, Data-driven cymbal bronze alloy identification via evolutionary machine learning with automatic feature selection, J. Intell. Manuf. 35 (2024) 257–273.

[37] F. Vilaseca, A. Castro, C. Chreties, A. Gorgoglione, Assessing influential rainfall–runoff variables to simulate daily streamflow using random forest, Hydrol. Sci. J. 0 (2023) 1–16.

[38] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, Springer Publishing Company, Incorporated, 2014.

[39] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat Methodol. 67 (2005) 301–320.

[40] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: a new learning scheme of feedforward neural networks, in: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), Volume 2 vol. 2, 2004, pp. 985–990, https://doi.org/10.1109/IJCNN.2004.1380068.

[41] T.L. Fonseca, L. Goliatt, Extreme learning machine based model improved with adaptive activation functions, in: Computational Intelligence in Information Systems: Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2020), Springer, 2021, pp. 119–128.

[42] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.

[43] A.-L. Balogun, F. Rezaie, Q.B. Pham, L. Gigović, S. Drobnjak, Y.A. Aina, M. Panahi, S.T. Yekeen, S. Lee, Spatial prediction of landslide susceptibility in western Serbia using hybrid support vector regression (svr) with gwo, bat and coa algorithms, Geosci. Front. 12 (2021) 101104.

[44] C. Saporetti, D. Fonseca, L. Oliveira, E. Pereira, L. Goliatt, Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields, Mar. Pet. Geol. 143 (2022) 105783.

[45] M. Awad, R. Khanna, Support Vector Regression, Apress, Berkeley, CA, 2015, pp. 67–80.

[46] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobot. 7 (2013).

[47] C. Bentéjac, A. Csörgö, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artif. Intell. Rev. 54 (2020) 1937–1967.

[48] R.E. Schapire, A brief introduction to boosting, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 1401–1406.

[49] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, H. Chen, Harris hawks optimization: algorithm and applications, Futur. Gener. Comput. Syst. 97 (2019) 849–872.

[50] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Proces. Syst. 30 (2017).

[51] C. Saporetti, D. Fonseca, L. Oliveira, E. Pereira, L. Goliatt, Machine learning with model selection to predict toc from mineralogical constituents: case study in the Sichuan basin, Int. J. Environ. Sci. Technol. 20 (2023) 1585–1596.

[52] Wes McKinney, Data structures for statistical computing in Python, in: Stéfan van der Walt, Jarrod Millman (Eds.), Proceedings of the 9th Python in Science Conference, 2010, pp. 56–61, https://doi.org/10.25080/Majora-92bf1922-00a.

[53] T. pandas development team, pandas-dev/pandas: Pandas, 2020, https://doi.org/10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[54] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Rio, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[56] M.L. Waskom, Seaborn: statistical data visualization, J. Open Source Softw. 6 (2021) 3021.

[57] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods 17 (2020) 261–272.

[58] J.D. Hunter, Matplotlib: a 2d graphics environment, Comput. Sci. Eng. 9 (2007) 90–95.

[59] N. V. Thieu, S. Mirjalili, MEALPY: a Framework of The State-of-The-Art Meta-Heuristic Algorithms in Python, 2022. URL: https://doi.org/10.5281/zenodo.6684223. doi: 10.5281/zenodo.6684223.

[60] D. Blondeau-Patissier, J.F. Gower, A.G. Dekker, S.R. Phinn, V.E. Brando, A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans, Prog. Oceanogr. 123 (2014) 123–144.

[61] H. Yin, P. Yin, Z. Yang, Seasonal sediment phosphorus release across sediment-water interface and its potential role in supporting algal blooms in a large shallow eutrophic lake (Lake Taihu, China), Sci. Total Environ. 896 (2023) 165252.

[62] W. Wang, B. Zheng, X. Jiang, J. Chen, S. Wang, Characteristics and source of dissolved organic matter in lake hulun, a large shallow eutrophic steppe lake in northern China, Water 12 (2020).

[63] G. Bilotta, R. Brazier, Understanding the influence of suspended solids on water quality and aquatic biota, Water Res. 42 (2008) 2849–2861.

[64] A. Gorgoglione, J. Gregorio, A. Rios, J. Alonso, C. Chreties, M. Fossati, Influence of land use/land cover on surface-water quality of Santa Luca River, Uruguay, Sustainability 12 (2020) 4692.

[65] A. Lintern, J. Webb, D. Ryu, S. Liu, U. Bende-Michl, D. Waters, P. Leahy, P. Wilson, A.W. Western, Key factors influencing differences in stream water quality across space, WIREs Water 5 (2018) e1260.

[66] S. Narbondo, A. Gorgoglione, M. Crisci, C. Chreties, Enhancing physical similarity approach to predict runoff in ungauged watersheds in sub-tropical regions, Water 12 (2020).

[67] A. Gorgoglione, F.A. Bombardelli, B.J. Pitton, L.R. Oki, D.L. Haver, T.M. Young, Uncertainty in the parameterization of sediment build-up and wash-off processes in the simulation of sediment transport in urban areas, Environ. Model. Softw. 111 (2019) 170–181.

[68] L. Goliatt, S.O. Sulaiman, K.M. Khedher, A.A. Farooque, Z.M. Yaseen, Estimation of natural streams longitudinal dispersion coefficient using hybrid evolutionary machine learning model, Eng. Appl. Comput. Fluid Mech. 15 (2021) 1298–1320.

[69] A.D. Martinho, C.M. Saporetti, L. Goliatt, Approaches for the short-term prediction of natural daily streamflows using hybrid machine learning enhanced with grey wolf optimization, Hydrol. Sci. J. 0 (2022) 1–18.

[70] L. Goliatt, R.S. Mohammad, S.I. Abba, Z.M. Yaseen, Development of hybrid computational data-intelligence model for flowing bottom-hole pressure of oil wells: new strategy for oil reservoir management and monitoring, Fuel 350 (2023) 128623.

[71] S. Sauvé, S. Lamontagne, J. Dupras, W. Stahel, Circular economy of water: tackling quantity, quality and footprint of water, Environ. Dev. 39 (2021) 100651.

[72] A. Gorgoglione, A. Castro, C. Chreties, L. Etcheverry, Overcoming data scarcity in earth science, Data 5 (2020).