Water and Soil Management and Modeling

Online ISSN: 2783 - 2546

University of Mohaghegh Ardabili

Predicting the electrical conductivity of water using the Hicking CEEMD-LSSVM optimization algorithm

Elham Ghanbari-Adivi [1]*, Jalil Kermannezhad [2], Ali Raeisi[1]

[1] Department of Water Engineering, Faculty of Agriculture, Shahrekord University, Shahrekord, Iran
[2] Chaharmahal & Bakhtiari Water and Waste Water Company, Shahrekord, Iran

Abstract

Accurate prediction of electrical conductivity (EC) concentrations in river water is essential for effective water quality management and environmental protection. This study develops a novel hybrid model, named HOA-CEEMD-LSSVM, that integrates the hiking optimization algorithm (HOA), complementary ensemble empirical mode decomposition (CEEMD), and least square support vector machine (LSSVM) to forecast daily EC concentrations in the Aidoghmoush River, Iran. HOA simultaneously optimizes key parameters of CEEMD and LSSVM to enhance prediction accuracy. CEEMD decomposes complex time series into intrinsic mode functions (IMFs), which exhibitmore predictable patterns, serving as inputs to the LSSVM predictor. The model's performance is evaluated through multiple metrics, demonstrating significant improvements over benchmark models in terms of R^2 and Kling-Gupta Efficiency (KGE). The proposed model enhances the R^2 and KGE values of other prediction models by 1%-10 % and 3.17%-17%, respectively. Our findings show that the HAO-CEEMD-LSSVM model can precisely forecast EC concentration. This approach provides a robust framework for capturing the nonlinear, nonstationary characteristics of EC time series data. The model is applicable in water resource planning, pollution control, and river ecosystem management. While showing high forecasting accuracy, its computational complexity and black-box nature present limitations. Future work should explore parallel computing and explainable artificial intelligence techniques to enhance efficiency and interpretability.

Keywords: EC concentrations, Hybrid models, Optimization algorithms, Water quality parameters.

Article Type: Research Article

*Corresponding Author, E-mail: ghanbariadivi@sku.ac.ir
Citation: Ghanbari-Adivi, E., Kermannezhad, J., & Raeisi, A. (2025). Predicting the electrical conductivity of water using the Hicking CEEMD-LSSVM optimization algorithm, Water and Soil Management and Modeling, 5(4), 323-348.
doi: 10.22098/mmws.2025.18187.1657
Received: 25 August 2025, Received in revised form: 08 October 2025, Accepted: 22 October 2025, Published online: 07 November 2025
Water and Soil Management and Modeling, Year 2025, Vol. 5, No. 4, pp. 323-348.
Publisher: University of Mohaghegh Ardabili

1. Introduction

Electrical conductivity (EC) constitutes a fundamental parameter in water engineering, serving as an indicator of water suitability for diverse applications such as irrigation, industrial processes, and environmental management (Ekemen Keskin et al., 2020). EC quantitatively represents the concentration of dissolved ionic species, including sodium, calcium, chloride, and sulfate ions (Ali Khan et al., 2022). These ions critically influence water characteristics such as hardness, salinity, and electrical conductance (Ali Khan et al., 2022).

The phenomenon of EC originates from the mobility of dissolved ions, which carry electric charges under an applied electric potential, thereby facilitating the conduction of electrical current through the aqueous medium (Shah et al., 2021; Muhammad et al., 2023). EC in natural water bodies exhibits variability as a consequence of seasonal fluctuations, precipitation dynamics, geological substrate, and anthropogenic influences, including agricultural runoff and industrial effluents (Kadkhodazadeh and Farzin, 2021; Karbasi et al., 2024). Analogously, EC variations in engineered systems, such as desalination plants, are influenced by feedwater composition variability, operational efficiency, and membrane performance.

Accurate forecasting of EC concentrations is imperative for effective management of water resources, mitigation of pollution, and strategic planning in irrigation and industrial contexts (Ahmadianfar et al., 2020). Conventional predictive models, however, are often inadequate in addressing the complex, nonlinear, and nonstationary behavior intrinsic to EC temporal datasets.

Precision in EC prediction is indispensable for multiple stakeholders involved in water resource management. Agricultural authorities utilize EC forecasts to optimize irrigation protocols and soil management interventions. Concurrently, municipal water utilities and industrial operators rely on EC data to enhance water treatment efficacy and maintain operational stability. This necessity has driven the development and application of advanced modeling techniques, including artificial intelligence (AI) methodologies, aimed at improving the reliability and accuracy of EC concentration predictions.

Artificial intelligence (AI) models exhibit exceptional capabilities in capturing complex and nonlinear relationships inherent in water systems (Karbasi et al., 2024). In contrast to traditional statistical approaches, AI techniques effectively manage large, heterogeneous datasets with superior precision. These models learn from historical patterns and dynamically adapt to new data, thereby enabling accurate forecasting of electrical conductivity (EC) fluctuations across diverse environmental and operational conditions (Karbasi et al., 2024). Furthermore, AI models continuously monitor changes in water quality parameters, facilitating the projection of future EC trends under various climate change scenarios (Kumar et al., 2023). By simulating the effects of temperature variations, precipitation changes, and sea-level rise on salinity, AI-driven predictions support decision-makers in formulating adaptive management strategies to uphold water quality standards in agricultural, industrial, and municipal domains.

AI methodologies confer several advantages, including enhanced predictive accuracy and robustness against missing or noisy data (Muhammad et al., 2023). Among the diverse AI variants employed for EC prediction are K-nearest neighbors (KNN), adaptive neuro-fuzzy inference systems (ANFIS), long short-term memory networks (LSTM), artificial neural networks (ANN), support vector machines (SVM), and decision tree-based algorithms.

A comprehensive review of prior research underscores the high potential of AI models in accurately estimating EC concentrations by leveraging historical data and modeling complex input-output relationships. Nevertheless, these studies reveal limitations, particularly the restricted range of AI models investigated, which impairs predictive performance under variable environmental conditions (Jamshidzadeh et al., 2024). Notably, the least-square support vector machine (LSSVM) model emerges as a promising alternative due to its robust generalization capacity and proficiency in modeling nonlinear dynamics within water quality data (Kadkhodazadeh and Farzin, 2021).

The LSSVM model combines high prediction accuracy with computational efficiency.

The LSSVM approach is mathematically characterized by an equation comprising a bias term, a weight vector, and a nonlinear mapping function. Since the weight and bias parameters are initially unknown, the model employs an optimization framework to estimate these values. A pivotal aspect of this optimization is the inclusion of a regularization term, which effectively controls model complexity to enhance generalizability (Ehteram and Soltani-Gerdefaramarzi, 2025). Moreover, the LSSVM model replaces the explicit mapping function with a kernel function, allowing it to efficiently capture complex nonlinear relationships in the data. Through this procedure, the LSSVM model attains a final predictive formulation capable of reliable EC concentration estimation (Ehteram and Soltani-Gerdefaramarzi, 2025).

The least-square support vector machine (LSSVM) model has been extensively utilized in forecasting various environmental variables. For example, Song et al. (2021) developed an SSA-LSSVM model that improved dissolved oxygen (DO) predictions by optimizing model parameters using the sparrow search algorithm. Chia et al. (2022) employed different optimization techniques to enhance LSSVM performance for forecasting the water quality index (WQI), achieving high precision. Zhou et al. (2022) combined evolutionary algorithms with LSSVM to predict biochemical oxygen demand (BOD) and ammonia nitrogen (NH_3-N), significantly improving accuracy. Xu et al. (2024) introduced both standalone and hybrid LSSVM models for runoff prediction, where the hybrid approach incorporated data preprocessing to stabilize inputs. Similarly, Ehteram and Soltani-Gerdefaramarzi (2025) found that optimized LSSVM models outperform standalone versions in water quality forecasting.

Despite its broad applications in water resources and environmental engineering, the LSSVM model's potential for predicting electrical conductivity (EC) remains unexplored. Moreover, inherent drawbacks of LSSVM, often neglected in prior studies, may limit its predictive accuracy. Notably, the model struggles with unpredictable time series patterns characterized by complex, nonlinear, and nonstationary behaviors, which introduce uncertainty and hinder trend detection (Ehteram and Soltani-Gerdefaramarzi, 2025). Such challenges are known to degrade AI model performance, especially in the presence of abrupt changes, seasonal variability, and random noise.

To address these limitations, the current study proposes the integration of an advanced data processing method, complementary ensemble empirical mode decomposition (CEEMD). CEEMD decomposes time series into intrinsic mode functions (IMFs) representing distinct frequency components. These IMFs form sub-time series with more regular and predictable patterns, enhancing the capacity of AI models like LSSVM to effectively capture underlying trends and nonlinear characteristics of the original data (Yahia Ahmed Abuker et al., 2025).

This approach aims to overcome the LSSVM model’s challenges with irregular patterns, improving EC concentration forecasting through improved input data structure and model sensitivity.

The complementary ensemble empirical mode decomposition (CEEMD) algorithm transforms time series into intrinsic mode functions (IMFs), thereby reducing the complexity and nonstationarity of the original data. This transformation significantly enhances the forecasting accuracy of artificial intelligence models such as the least-square support vector machine (LSSVM) (Shin et al., 2025). Recently, CEEMD has been widely integrated with AI models to improve their predictive performance. For instance, Zhang et al. (2021) developed a hybrid model combining CEEMD and long short-term memory (LSTM) networks for monthly precipitation forecasting. In their approach, the original precipitation data were decomposed into stable IMFs and a residual term by CEEMD, each predicted separately by the LSTM network, and the final forecast was obtained by reconstructing these components. This hybrid method markedly enhanced prediction accuracy. Similarly, Zhao and Zhou (2024) applied CEEMD in conjunction with the kernel extreme learning machine (KELM) to improve wind power forecasting accuracy and stability. The non-smooth wind power series were decomposed into stationary

IMFs and a residual, which were individually predicted by KELM, and then aggregated. Their results demonstrated that the CEEMD-KELM hybrid model significantly outperformed traditional forecasting techniques in terms of accuracy and stability.

In summary, combining CEEMD with AI models provides a robust framework for improving the prediction accuracy of EC concentrations. This approach enables more accurate and reliable predictions, which are required for effective water resources management and environmental monitoring. Thus, our study combines the CEEMD algorithm with LSSVM to develop the hybrid CEEMD-LSSVM model, which produces intrinsic mode functions (IMFs) and utilizes them to predict EC concentrations accurately. The CEEMD-LSSVM model is a powerful and efficient tool for predicting EC concentrations. Its ability to preprocess complex time series data and accurately capture dynamic patterns makes it a promising alternative to traditional modeling approaches. However, it is essential to note that CEEMD-LSSVM can achieve accurate predictions only if its parameters, such as the LSSVM and CEEMD parameters, are appropriately adjusted. The LSSVM parameters, including the regularization and kernel parameters, significantly affect the accuracy of the forecasts (Ehteram and Soltani-Gerdefaramarzi, 2025). The regularization parameter controls the model complexity, while the kernel parameter influences the accuracy of predictions. The number of IMFs is another parameter that affects the overall forecasting accuracy and computational efficiency of the CEEMD-LSSVM model. For example, if the number of IMFs is too low, crucial frequency components may be missed, leading to under-decomposition and a failure to capture critical variations in the time series of water quality parameters. On the other hand, if the number of IMFs is too high, the training time increases significantly, and the model may become computationally expensive. Thus, it is essential to adjust and optimize the parameters of the CEEMD-LSSVM model properly.

Optimization algorithms have been developed to fine-tune AI model parameters, such as those in the CEEMD-LSSVM hybrid framework, to maximize performance by efficiently exploring the parameter space and minimizing prediction errors (Ehteram and Soltani-Gerdefaramarzi, 2025). Proper adjustment of intrinsic mode functions (IMFs) and LSSVM hyperparameters via these algorithms enhances the model's accuracy and robustness in capturing complex, nonlinear, and nonstationary patterns.

One notable optimization method is the hiking optimization algorithm (HOA), which mimics hikers ascending a mountain, with velocities influenced by terrain features like elevation and slope (Oladejo et al., 2024; Sag, 2024). HOA initializes hikers in the search space and iteratively updates their positions and velocities toward a global optimum, with these positions representing candidate solutions. HOA's advantages include fewer control parameters, making it easier to implement and less sensitive to tuning, alongside high precision and flexibility. Consequently, HOA has been successfully applied in diverse areas such as gene selection (Pashaei et al., 2025), low-carbon economic optimization (Wu et al., 2025), feature selection (Abdel-salam et al., 2025), and engineering design optimization (Özcan et al., 2025).

Thus, our study combines the HOA algorithm with the CEEMD-LSSVM model to adjust CEEMD and LSSVM parameters. The resulting hybrid model, called the HOA-CEEMD-LSSVM model, is then used to predict EC concentrations. The novelties of the paper are described as follows:

1) Development of an innovative optimization method—HOA—for simultaneous tuning of CEEMD and LSSVM parameters to improve forecasting performance.
2) Integration of CEEMD with LSSVM to handle nonlinear and nonstationary characteristics of EC time series effectively
3) Application of the hybrid model to a real-world case study, demonstrating superior accuracy and robustness compared to conventional approaches.

2. Materials and Methods

2.1. Study Area:

2.1.1 Aidoghmoush River

There are crucial rivers in the northwest region of Iran, such as the Aidoghmoush River. The

Aidoghmoush River has a catchment area of approximately 1,802 km², and its length is nearly 80 km.

The Aidoghmoush River plays a crucial role in supporting agricultural, industrial, and domestic water demands in the northwest of Iran. However, in recent years, different factors such as climate change, land-use changes, and anthropogenic activities have significantly increased the EC concentration of the river. High EC concentrations can cause soil degradation, reduce crop yields, damage infrastructure, and disrupt aquatic ecosystems. However, to address these challenges, local decision-makers and policymakers need accurate predictions of EC concentrations. Thus, our study develops the HOA-CEEMD-LSSVM model to predict the daily concentration of EC.

Figure 1 shows the location of the case study.

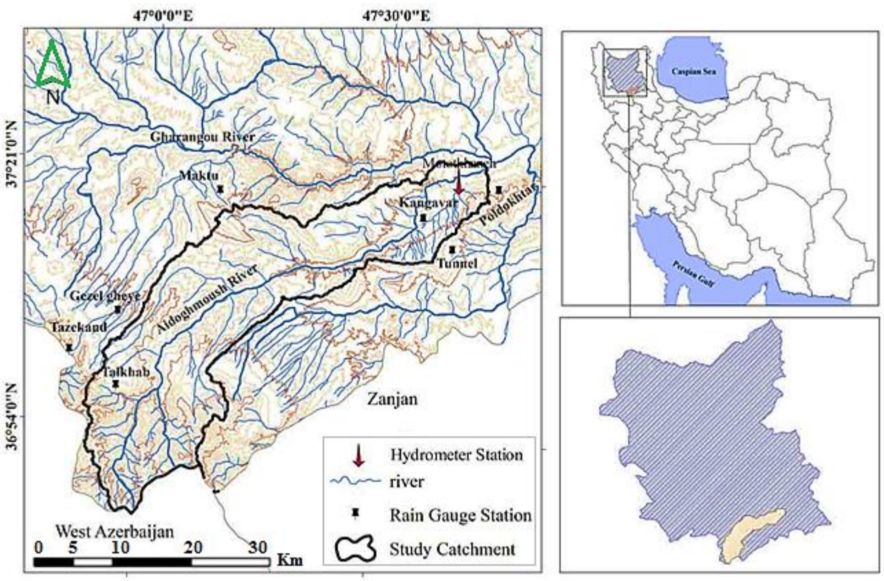


Figure 1. Geographic location of the Aidoghmoush River Basin

2.2. CEEMD

The CEEMD algorithm has emerged as a powerful tool for decomposing complex, nonlinear, and nonstationary time series into components with more predictable patterns called IMFs (Zhao and Zhou, 2024). It is an enhanced version of the ensemble empirical mode decomposition (EEMD), which can overcome issues such as mode mixing and noise instability. The CEEMD algorithm has various benefits, which are mentioned below:

- Preservation of Signal Integrity
- Broad Applicability in Environmental and Hydrological Forecasting
- Computational efficiency

The CEEMD algorithm is implemented as follows:

- First, the algorithm adds a white noise sequence with positive and negative signs to the original time series. The outputs of this operation are two perturbed versions of the original time series (Eqs. 1 and 2) (Zhao and Zhou, 2024).

$$x^{+}(t)=x(t)+\varepsilon.\omega(t) \tag{1}$$

$$x^{-}(t)=x(t)-\varepsilon.\omega(t) \tag{2}$$

Where $x^{+}(t)$ and $x^{-}(t)$: Two perturbed versions of the original time series, $\omega(t)$: The white noise sequence, and 9ε: The noise amplitude coefficient, which controls the strength of the added noise.

In this study, the noise amplitude coefficient (ε) was set to a typical value commonly used in the literature, usually around 0.2 times the standard deviation of the original signal, to ensure effective noise-assisted decomposition without overwhelming the signal. The white noise sequence was generated using a standard Gaussian distribution with zero mean and unit variance, which is a standard approach in CEEMD implementations.

- Next, CEEMD uses the empirical mode decomposition (EMD) technique to separately decompose two perturbed time series into IMFs. Equations 3 and 4 show the decomposition results for each perturbed time series (Zhao and Zhou, 2024):

$$x^{+}(t)=\sum_{i=1}^{n} IMF_i^{+}(t)+r^{+}(t) \tag{3}$$

$$x^{-}(t)=\sum_{i=1}^{n} IMF_i^{-}(t)+r^{-}(t) \tag{4}$$

Where $IMF_i^{+}(t)$ and $IMF_i^{-}(t)$: represent the i-th IMF obtained from the positively and negatively perturbed time series, respectively, $r^{+}(t)$ and $r^{-}(t)$: the corresponding residual components. This dual decomposition approach helps reduce mode mixing and improves the stability and accuracy of the IMF extraction process.

- The final IMFs are obtained by averaging the corresponding IMFs obtained from $x^{+}(t)$ and $x^{-}(t)$. Equation 5 shows this averaging process:

$$IMF_i(t)=\frac{IMF_i^{+}(t)+IMF_i^{-}(t)}{2} \tag{5}$$

Where $IMF_i(t)$: the i-th refined IMF.

- Finally, the residual components are averaged to obtain the final residual (Eq. 6) (Zhao and Zhou, 2024).

$$r(t)=\frac{r^{+}(t)+r^{-}(t)}{2} \tag{6}$$

Where: $r(t)$ the final residual component, which reflects the overall trend of the time series after all intrinsic mode functions have been extracted and averaged.

The CEEMD algorithm effectively decomposes the original time series into a set of IMFs and a final residual component. The number of IMFs is a crucial parameter that affects the precision of the forecasts. Thus, our study utilizes HOA to determine the number of IMFs accurately.

2.3. LSSVM model

LSSVM is a modified version of the standard SVM, which can predict different variables such as EC concentrations (Zhou et al., 2024; Ghanbari-Adivi and Ehteram, 2025).

The LSSVM model has a basic equation, which provides a relationship between the input sequences and the output variable (Eq. 7) (Zhou et al., 2024):

$$T(x)=\rho^{T}\tau(x)+b \tag{7}$$

Where ρ: The weight coefficient, b: bias term, $T(x)$: The predicted output (e.g., EC concentration), and $\tau(x)$: A mapping function.

However, as the weight and bias values are unknown, Equation 7 cannot be directly used to produce predictions. Moreover, the mapping function cannot efficiently capture complex patterns in the data. Thus, the LSSVM model performs three key operations to overcome these limitations and enable accurate prediction of variables such as EC concentrations.

First, the model formulates an optimization problem, which can be solved to determine the bias and weight values. However, since this problem is constrained and involves complex computations in high-dimensional space, the model cannot solve it directly (Kadkhodazadeh and Farzin, 2021). Equations 8 and 9 present the optimization problem and its associated constraint.

$$\min_{\rho,b,e}=\left[\frac{1}{2}\rho^{T}\rho+\frac{\gamma}{2}\sum_{i=1}^{N}e_i^{2}\right] \tag{8}$$

$$T_i=\rho^{T}\tau(x_i)+b+e_i \tag{9}$$

Where T_i : The output (e.g., EC concentration), x_i : Input vector, ρ : Weight value, b: Bias term, e_i : error terms, γ : A regularization parameter, and N: Number of training samples.

In the second step, the model uses Lagrange multipliers to convert the constrained optimization problem into an unconstrained one, which is explained as follows (Kadkhodazadeh and Farzin, 2021).

$$L\left(\rho,b,e,\alpha\right)=\frac{1}{2}\rho^{T}\rho+\frac{\gamma}{2}\sum_{i=1}^{N}e_i^2-\sum_{i=1}^{N}\alpha_i\left[\rho^{T}\tau\left(x_i\right)+b+e_i-T_i\right] \tag{10}$$

Where Lagrangian function and Lagrange multipliers associated with each training sample. Finally, the model replaces the nonlinear mapping function with a kernel function and solves Equation 10. This procedure yields the final formulation of the LSSVM model, which is presented in Equation 11 (Kadkhodazadeh and Farzin, 2021). This formulation produces the final predictions.

$$T\left(x\right)=\sum_{i=1}^{N}\alpha_i K\left(x,x_i\right)+b \tag{11}$$

$$K\left(x,x_i\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right) \tag{12}$$

Where σ : A kernel parameter and $K\left(x,x_i\right)$: A kernel function. The kernel parameter and regularization term are key hyperparameters that significantly influence the efficiency of the LSSVM model. Thus, our study utilizes HOA to adjust these parameters properly.

2.4. Hiking optimization algorithm

HOA mimics the behavior of hikers during their journey. At the beginning of a hike, hikers gather information about terrain characteristics and explore the most effective paths toward their destination (exploration phase). Subsequently, they exploit these paths to reach the destination efficiently (exploitation phase) (Oladejo et al., 2024). Similarly, HOA initially explores the search space broadly to identify the most promising regions. These regions are then exploited to obtain optimal or near-optimal solutions. This adaptive transition between exploration and exploitation is a key feature of HOA, enabling it to efficiently navigate complex search spaces and converge toward optimal solutions. HOA is executed as follows:

First, the algorithm uses Tobler's Hiking Function (THF) to define an initial velocity for each hiker (agent). THF is an exponential function that can determine a hiker's velocity based on the slope of a terrain (Eq. 13) (Oladejo et al., 2024).

$$V_{ij}=6e^{-3.5\left|Si,t+0.05\right|} \tag{13}$$

Where V_{ij} : The velocity of the i-th hiker in the j-th dimension and $S_{i,t}$: The slope of the terrain. The slope is determined based on the following equation:

$$S_{i,t}=\frac{dh}{dx}=\tan\theta_{i,t} \tag{14}$$

Where dh : elevation difference, dx the distance covered by the hiker, and $\theta_{i,t}$ the angle.

The slope in the Hiking Optimization Algorithm (HOA) represents the gradient or rate of change of the objective function within the parameter search space. In the context of hyperparameter optimization for the CEEMD-LSSVM model, the slope is calculated based on the variation of the model's validation error with respect to changes in parameter values. Specifically, the slope is derived from evaluating incremental changes in the objective function (e.g., prediction error or loss) as the algorithm explores the parameter space, which guides the hikers' velocity updates toward the global optimum.

- In this step, the algorithm updates the velocity of each hiker (Eq. 15).

$$V_{i,t}=V_{i,t-1}+\eta_{i,t}\left(L_{best}-\mu L_{i,t}\right) \tag{15}$$

Where μ : The sweep factor (SF) of the i-th hiker, $V_{i,t}$: The velocity of the i-th hiker at iteration t, $V_{i,t-1}$: The velocity of the i-th hiker at iteration t-1, L_{best} : The position of the best hiker, and $L_{i,t}$: The position of the i-th hiker at iteration t.

- Finally, the algorithm updates the location of each hiker. This updated location represents the current candidate solution in the search space (Eq. 16) (Oladejo et al., 2024).

$$L_{i,t+1}=L_{i,t}+V_{i,t+1} \tag{16}$$

Where $L_{i,t+1}$: The position of the i-th hiker. The Hiking optimization algorithm is executed as follows:

1- First, the algorithm determines an initial velocity for each hiker in the search space. This operation is performed using Equation 13, which is based on Tobler's Hiking Function (THF). This function simulates how a hiker adjusts their speed according to the slope of the terrain, enabling the algorithm to mimic realistic exploration behavior during the search process.

2- The algorithm determines an initial location for each hiker in the search space. This operation is performed using Equation 17, which ensures that the search starts from a diverse and well-distributed set of solutions across the search space. This step enhances the exploration capability of the Hiking optimization algorithm (HOA) and reduces the risk of early convergence to local optima.

$$L_{i,t} = \eta_j^1 + \kappa\left(\eta_j^2 - \eta_j^1\right) \tag{17}$$

Where $L_{i,t}$: The initial positions of hikers, η_j^2 and η_j^1 : The upper and lower values of the decision variables, and κ : A random number.

1- The algorithm updates the velocity of each hiker using equation 15.

2- The algorithm updates the location of each hiker using equation 16. The updated location shows the current candidate solution in the search space.

3- Once the stopping criterion is met, the optimization process terminates.

In the Hiking Optimization Algorithm (HOA) applied to hyperparameter tuning for the CEEMD-LSSVM model, the population consists of multiple "hikers," each representing a candidate solution in the parameter search space. Specifically, each hiker is modeled as a vector containing the hyperparameters under optimization, structured as $[\gamma,\sigma,n_{IMFs}]$, where γ is the regularization parameter, σ the kernel width, and n_{IMFs} the number of intrinsic mode functions.

2.4.1 Hybrid HOA-MLP Model

The multilayer perceptron (MLP) model employed in this study initially consisted of a feedforward neural network with two hidden layers. The first hidden layer contained 16 neurons, and the second hidden layer contained 8 neurons. Both layers used the Rectified Linear Unit (ReLU) activation function to introduce nonlinearity. This baseline architecture was established based on prior research and preliminary tuning to provide a standard starting point.

To enhance predictive performance, the Hiking Optimization Algorithm (HOA) was integrated to optimize key hyperparameters, including the number of neurons in each hidden layer, learning rate, and regularization coefficients. Within the HOA framework, each candidate solution ("hiker") was represented as a vector encompassing these hyperparameters. The algorithm systematically explored the hyperparameter search space by iteratively updating the hikers' positions and velocities, seeking to minimize the validation error of the MLP.

2.4.2. Hybrid HOA-RNN Model

The recurrent neural network (RNN) model used in this research initially featured a single hidden layer architected with 20 neurons employing the hyperbolic tangent (tanh) activation function to capture temporal dependencies in the data. This initial design was chosen based on established practices for time series modeling.

Similar to the MLP model, the Hiking Optimization Algorithm (HOA) was deployed to optimize crucial hyperparameters, including the number of neurons in the hidden layer, learning rate, and regularization strength. Each hiker within the HOA population encoded these hyperparameters as a vector, and the algorithm conducted iterative searches by updating the hikers' velocities and positions.

The HOA-driven hyperparameter tuning enabled the RNN model to better capture complex temporal dynamics and nonlinearities inherent in the dataset, thereby improving forecast accuracy and robustness relative to the initial RNN baseline.

2.5. HOA-CEEMD-LSSVM model The HOA-CEEMD-LSSVM model is applied to predict daily EC concentrations. The hybrid model is constructed as follows (Fig. 2).

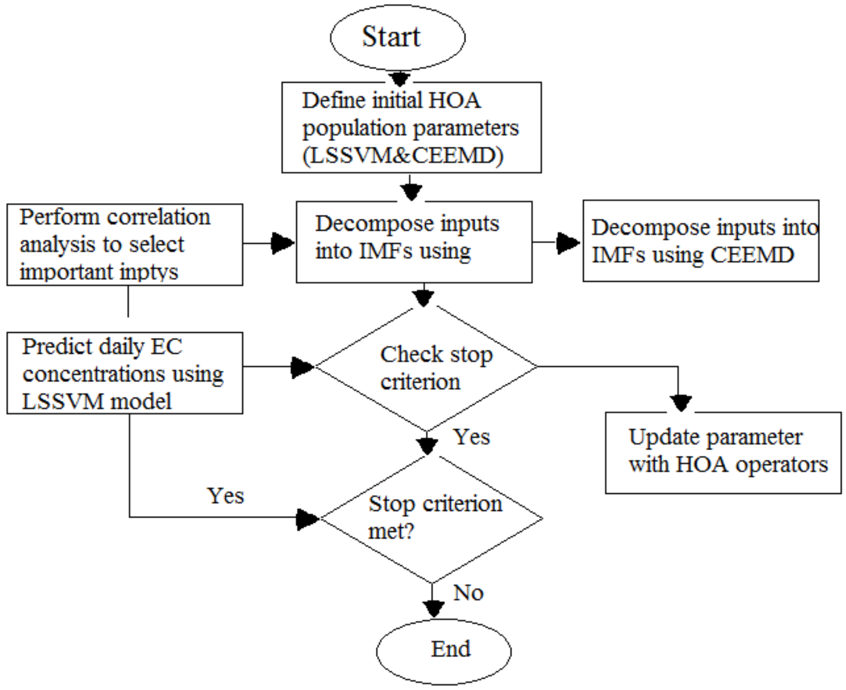


Figure 2. Flowchart for HOA-CEEMD-LSSVM

2.6. Comparative models

In this study, the HOA-CEEMD-LSSVM model is used to predict EC concentrations. However, the proposed model should be compared with several comparative models to evaluate its effectiveness and superiority in predicting EC concentrations. These models include the recurrent neural network (RNN), multilayer perceptron (MLP), and multiple linear regression (MLR), which have been widely utilized for predicting the concentration of water quality parameters.

2.6.1. MLP

The MLP model is a class of ANN models that consists of at least three layers: an input layer, one or more middle layers, and an output layer. Moreover, each layer contains neurons that effectively process input information (Reza et al., 2024).

The MLP model is executed as follows:

1) First, the input layer receives the input variables (e.g., water temperature, pH, dissolved oxygen, or other relevant water quality parameters) that significantly affect EC concentrations. Then, it passes these inputs to the hidden layer(s).

2) In the hidden layers, each neuron computes a weighted sum of the inputs and applies an activation function to produce an output (Talebzadeh et al., 2024).

3) The output layer receives the outputs of the last hidden layer, computes a weighted sum of these values, and applies an activation function (if required) to produce the final output.

However, it is essential to note that the MLP model can achieve accurate predictions only if its parameters, such as weights and biases, are correctly adjusted. Thus, our study combines HOA with the MLP model to set its parameters. The resulting hybrid model, called HOA-MLP, is then utilized to forecast EC concentrations accurately.

2.6.2. MLR

MLR is a statistical modeling technique that can establish a relationship between a dependent variable (also called the response or output variable) and two or more independent variables (predictors or input variables) (Dulger Altıner et al., 2024). The general form of the MLR model can be defined as follows:

$$Y = \psi_0 + \psi_1 X_1 + \psi_2 X_2 + .. + \psi_n X_n \tag{18}$$

Where Y : The output variable ψ_0, ψ_1, ψ_2, and ψ_n : Regression coefficients. However, the precision of the MLR model relies on properly adjusting its regression coefficients. Thus, our study combines HOA with the MLR model to set its parameters properly. Then, the resulting model, called HOA-MLR, is applied to forecast EC concentrations precisely.

2.6.3. RNN model

The RNN model is a deep learning model that can process sequential data. Unlike feedforward networks such as MLP, RNNs can handle temporal dependencies, making them particularly suitable for time-series prediction (Mienye et al., 2024).

The RNN model produces outputs at several steps. First, the model receives an input sequence at each time step. Then, it updates its hidden state, which stores crucial information. Finally, the model uses the updated hidden state to generate the final output at each time step (Mienye et al., 2024). Equations 19 and 20 show the mathematical formulation of the RNN model.

$$\rho_t = f\left(\kappa_{hh} h_{t-1} + \kappa_{xh} x + b_h\right) \tag{19}$$

$$y_t = \kappa_{hy} h_t + b_y \tag{20}$$

Where ρ_t : Hidden state at time step t, κ_{hh} and κ_{xh} : Weight matrices associated with the hidden and input connections, b_h : Bias term, and f: Activation function (e.g., hyperbolic tangent or ReLU). The RNN model can produce accurate predictions only if its parameters, such as weights, bias, and the number of hidden units, are appropriately adjusted. Thus, our study combines HOA with the RNN model to adjust its parameters properly. Then, the resulting model, called HOA-RNN, is applied to forecast EC concentrations precisely.

2.7. Evaluation metrics

In this study, multiple evaluation metrics are employed to evaluate the precision of the models. Equations 21-26 explain these metrics:

- **Absolute Percentage Bias**: APB quantifies the average magnitude of the bias between predicted values and actual observed values. A lower APB indicates better model performance.

$$APB = \left| \frac{\sum_{i=1}^{n}\left(EC^{pr} - EC^{ob}\right)}{\sum_{i=1}^{n} EC^{ob}} \right| \tag{21}$$

- **Legates and McCabe Index (LMI)**: LMI can quantify the agreement between observed and predicted data. It ranges from 0 to 1, where 1 indicates a perfect fit, and 0 indicates no agreement between observed and predicted values.

$$LMI = 1 - \frac{\sum_{i=1}^{N}|EC^{pr} - EC^{ob}|}{\sum_{i=1}^{N}|EC^{pr} - \overline{EC}^{ob}| + |EC^{ob} - \overline{EC}^{ob}|} \tag{22}$$

- **t-statistic (TS):** TS acts as a benchmark for evaluating the reliability of predictive models. A higher TS value shows that prediction errors are significant.

$$TS = \sqrt{\frac{(n-1) * MBE^2}{RMSE^2 - MBE^2}} \tag{23}$$

- **Root mean square error (RMSE):** This index evaluates the differences between predicted and observed values. A lower RMSE value indicates greater accuracy.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(EC^{ob} - EC^{pr}\right)} \tag{24}$$

- **Uncertainty at 95% (U95):** U95 quantifies the width of the 95% prediction interval around the model's forecasts. A higher U95 value suggests greater uncertainty.

$$U95 = 1.96\left(SD^2 - RMSE^2\right)^{0.50} \tag{25}$$

- **Kling–Gupta Efficiency (KGE)**: KGE can assess model performance by simultaneously considering three key components of model behavior: correlation, bias, and variability. A KGE value of 1 represents perfect agreement between forecasted and observed data.

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{E\bar{C}^{pr}}{E\bar{C}^{ob}} - 1\right)^2 + \left(\frac{CV^{pr}}{CV^{ob}}\right)} \tag{26}$$

Where r is the correlation coefficient, $E\bar{C}^{pr}$: average predicted EC concentrations, $E\bar{C}^{ob}$: average observed EC concentrations, SD: standard deviation, MBE : mean bias error, EC^{pr} : predicted EC concentrations, and EC^{ob} : observed EC concentrations.

2.7. Data set

In this study, calcium (Ca^{2+}), chloride (Cl^-), sodium (Na^+), Sulfate (SO_4^{2-}), pH, and Total Dissolved Solids (TDS) are used to predict EC concentrations in the Aidoghmoush River. The concentrations of water quality parameters are measured at the hydrometric station on the Aidoghmoush River. The statistical characteristics of these parameters are displayed in Table 1 and Figure 3. It is crucial to note that the study period is from 2018 to 2023.

Table 1. Statistical characteristics of input and output data **(**pH is dimensionless, Ec is µS/cm, and the unit for all other water quality parameters is mg/L**)**

Parameter	Average	Maximum	Minimum
Ca^{2+}	3.25	75.23	2.45
Cl^-	9.12	74.12	1.12
Na^+	8.76	72.12	1.90
SO_4^{-2}	8.98	43.85	5.12
pH	5.50	8.30	3.40
TDS	467.23	900.21	145.34
EC	430.4	900.23	125.65

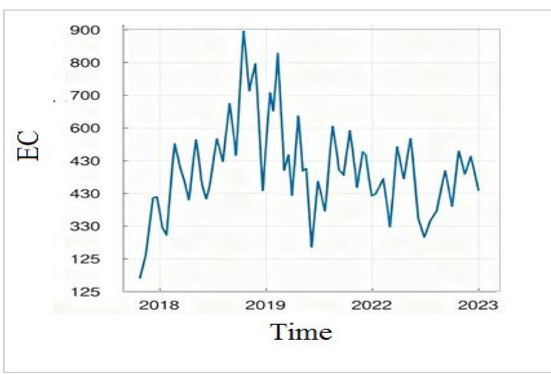


Figure3. The EC time series

4. Results

4.1. Choice of the optimal data size for training models

EC concentrations can be accurately predicted if the optimal size of the data is used to train the HOA-CEEMD-LSSVM model. Without enough training data, the model's predictive performance may degrade significantly, leading to unreliable estimates of EC concentrations. Thus, the choice of the optimal data size is necessary to ensure the prediction accuracy of the model. The current study addresses this need by systematically varying the data size and evaluating its impact on model performance metrics, such as the root mean square error (RMSE). The lowest RMSE values are generated by the optimal data size.

4.2. Determination of the most critical inputs

This study uses six water quality parameters to forecast EC concentrations. These parameters are selected based on their significant influence on electrical conductivity (EC) concentration. For each of these parameters, lag times of 1-30 days are considered to capture the temporal dependencies and dynamic relationships between the input variables and EC concentrations. This approach allows the model to account for the delayed effects of changes in water quality parameters on EC levels. Thus, the total number of inputs is 180, which is calculated by multiplying the six water quality parameters by the 30 lag times (1–30 days). These inputs may increase computational complexity and training time. Therefore, determining the most critical inputs becomes a crucial step in model development. The selection of these inputs helps to reduce computational complexity, improve model generalization, and enhance interpretability. By identifying and retaining only the most relevant variables, unnecessary noise and redundancy in the data are minimized, leading to a more efficient and robust predictive model.

The correlation analysis produces a correlation coefficient value for each lagged water quality parameter, which indicates the strength and direction of the linear relationship between that parameter (at a specific time lag) and the target variable. If the correlation value of a water quality parameter varies from ±0.80 to ±1.00, it is identified as the most critical input (Khatti et al., 2024).

Figure 3 displays the correlation values of different water parameters. The correlation values of Ca^{2+}(t-1), Na^{+}(t-1), TDS (t-1), Ph (t-1), Cl^{-} (t-1), and SO4-2 (t-1) vary from ±0.80 to ±1.00. Therefore, these parameters are selected as the most important inputs.

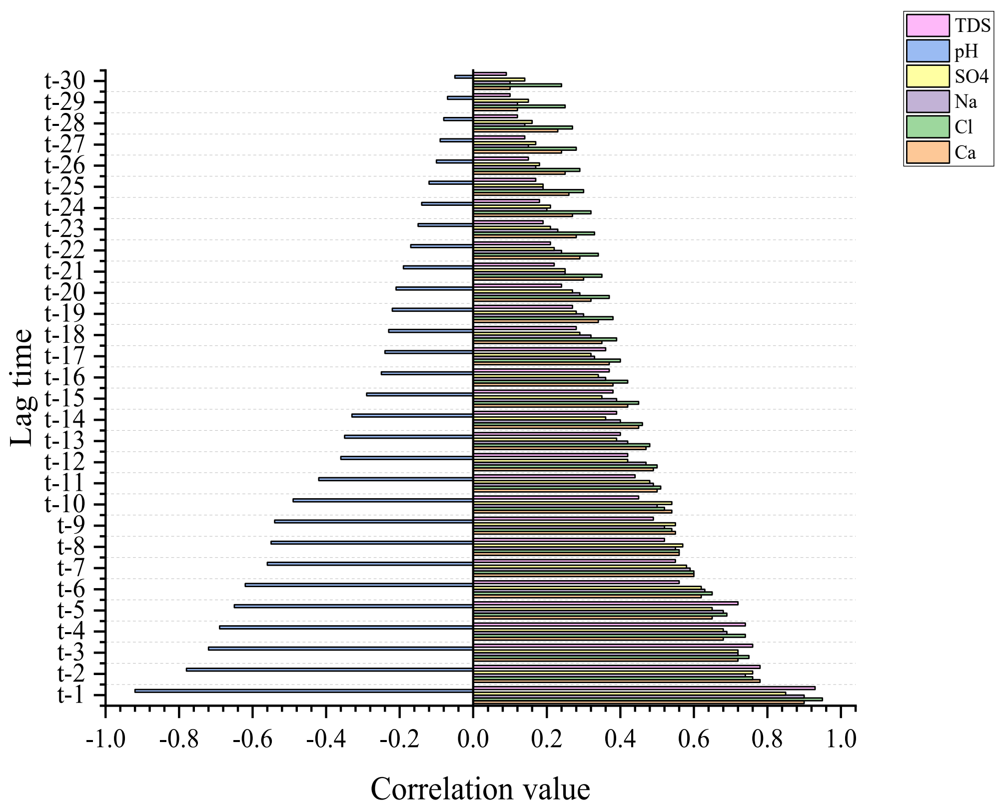


Figure 4. Correlation values between inputs and EC concentrations

4.3. Optimal values of the model parameters

In this paper, HOA is utilized to fine-tune the model parameters. In Table 2, the optimal values of model parameters are listed.

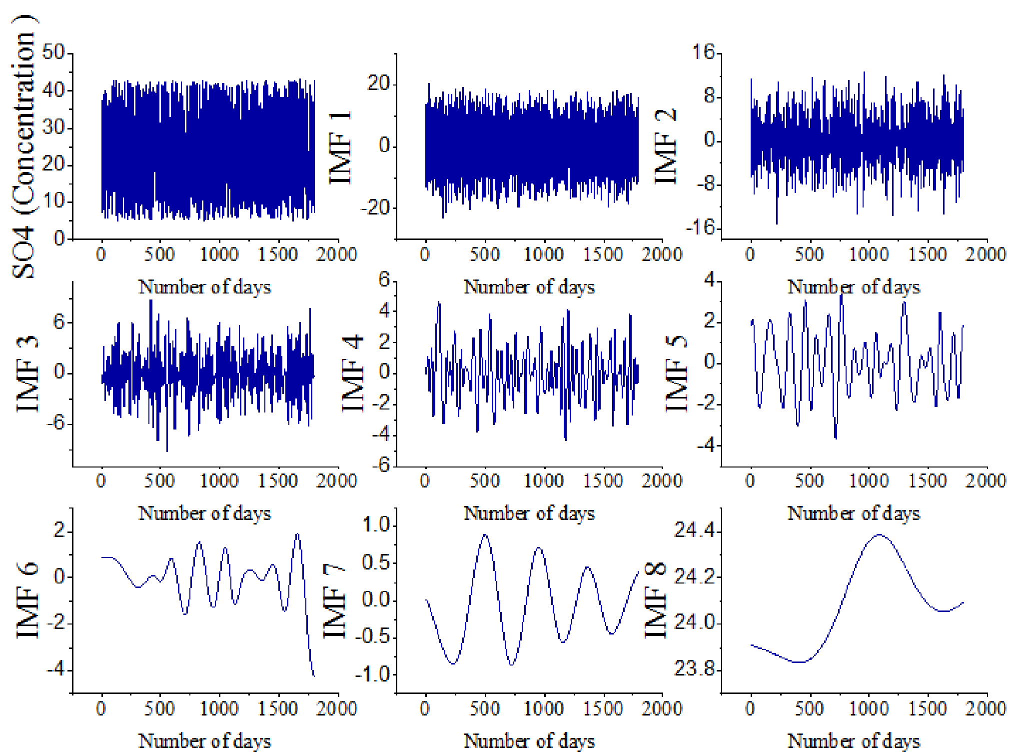
Table 2. Optimal values of the parameters of prediction models

Model or algorithm	Parameter values
LSSVM	σ =6.25 and $\gamma = 938.25$
CEEMD	Number of IMFs=8
MLP	Learning arte:0.001, batch size:20, and Number of hidden layers:1
RNN	Number of hidden units:20, learning rate:0.001 and batch size:60
HOA	Population size: 20 and maximum number of iterations:60

4.4. Production of IMFs

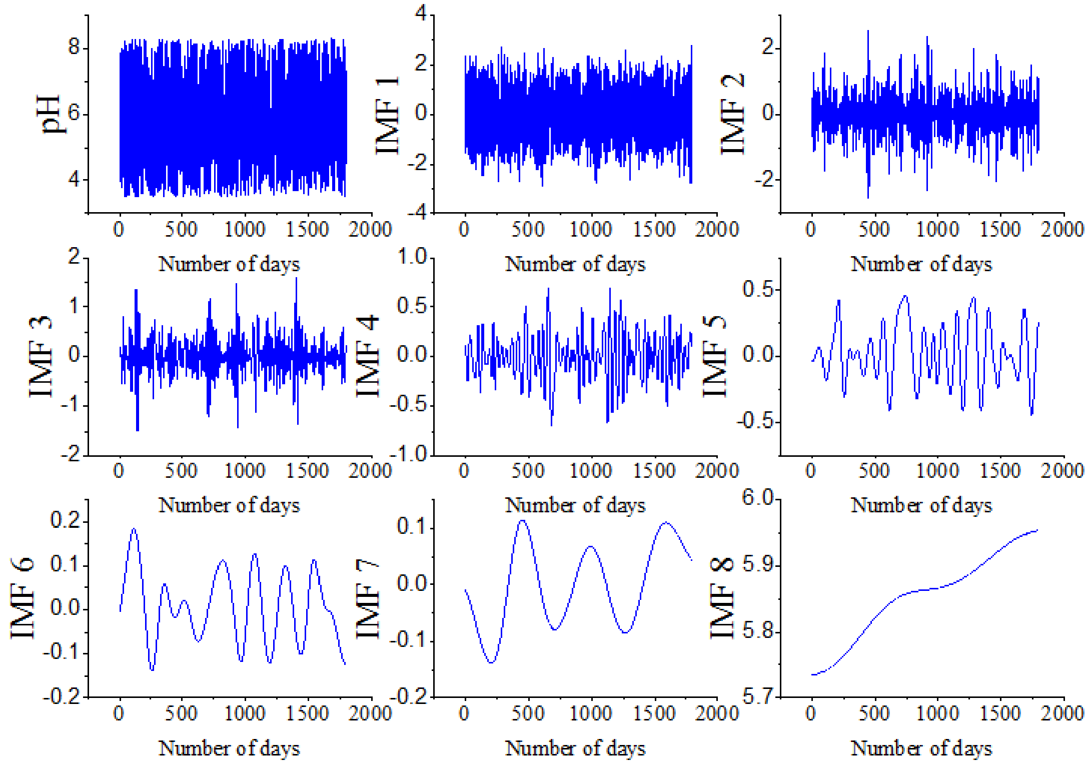
The CEEMD algorithm is applied to break down the time series of chosen inputs into IMFs. While time series of water quality parameters have unpredictable patterns, IMFs have more predictable patterns that can be used to improve the forecasting accuracy of AI models. However, the number of IMFs can significantly affect the performance of AI models. An excessive number of IMFs may increase computational overhead, while a low number of IMFs may not be sufficient to enhance the accuracy of forecasts. Therefore, determining the optimal number of IMFs is essential for achieving high forecasting accuracy. In this study, HOA is used to determine the optimal number of IMFs. This optimization ensures a balance between capturing essential data features and minimizing computational complexity.

Table 2 shows that the optimal number of IMFs is 8. Therefore, the timeseries of chosen inputs are decomposed into 8 IMFs. Figure 5 shows IMFs, which are produced from key inputs. These IMFs are fed into prediction models such as HOA-CEEMD-LSSVM to forecast EC concentrations.



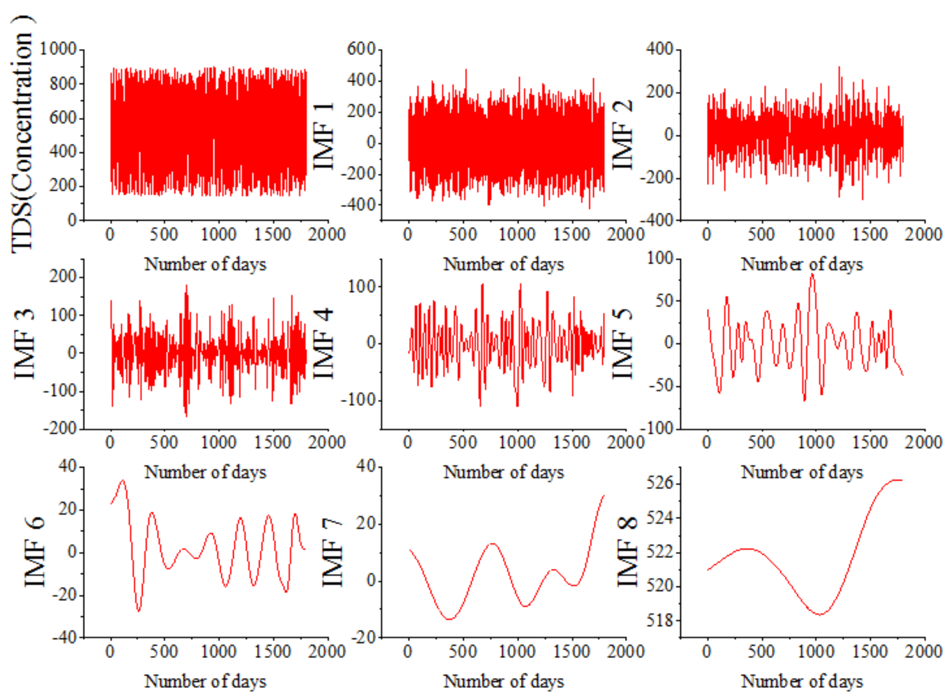
SO4^{-2} (t-1)

Figure 5. Generated IMFs for predicting EC concentrations



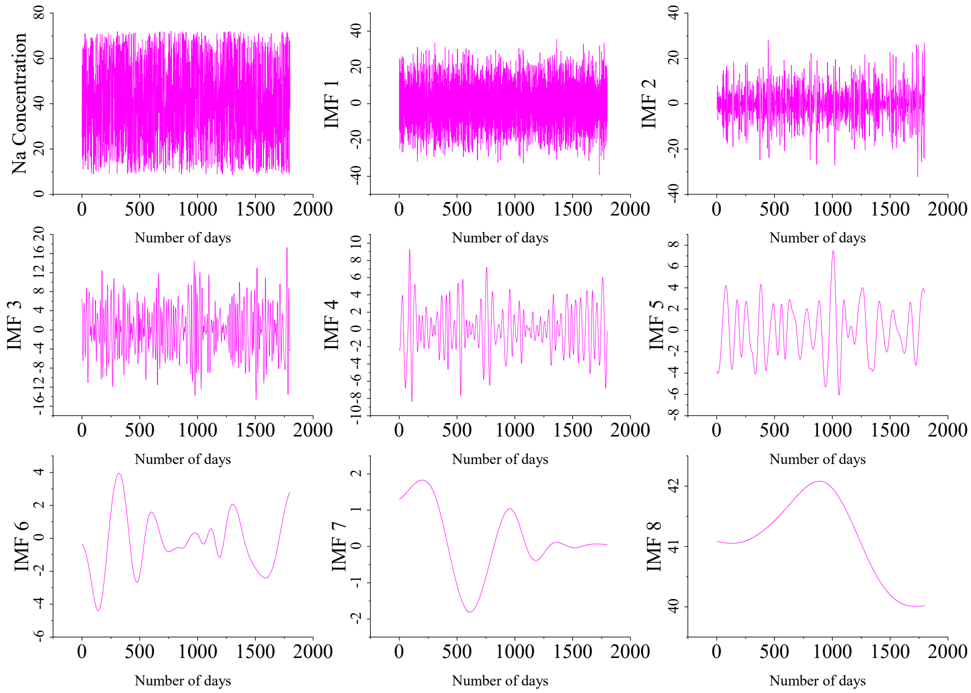
pH (t-1)

Figure 5. cont. Generated IMFs for predicting EC concentrations



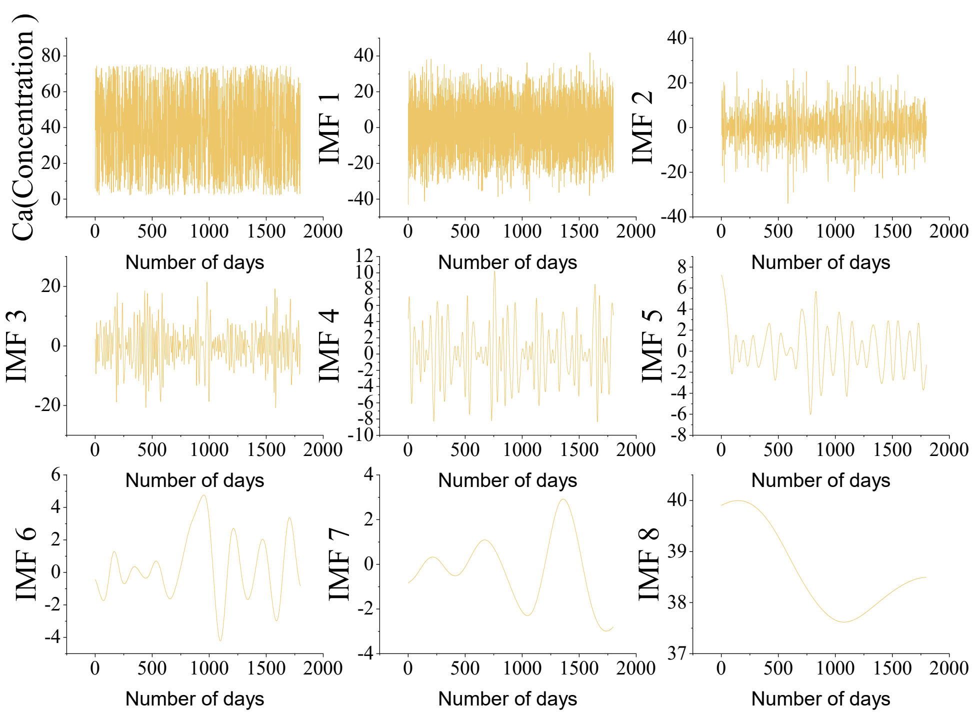
TDS (t-1)

Figure 5. cont. Generated IMFs for predicting EC concentrations



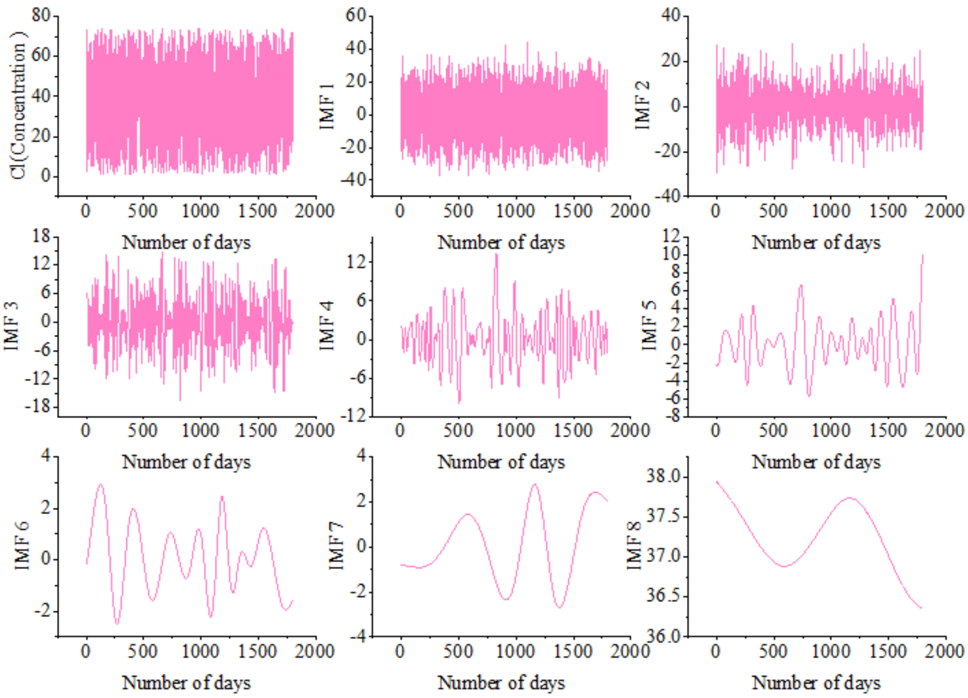
Na^{+}(t-1)

Figure 5. cont. Generated IMFs for predicting EC concentrations



Ca^{2+}(t-1)

Figure 5. cont. Generated IMFs for predicting EC concentrations



Cl⁻ (t-1)

Figure 5. cont. Generated IMFs for predicting EC concentrations

4.5. Assessment of the accuracy of forecasts

The HOA-CEEMD-LSSVM model is compared with other hybrid models to assess its overall predictive capability. These comparisons are described as follows:

A comparison between LSSVM and other standalone models: In this paper, the LSSVM model is developed to predict EC concentrations. However, it is benchmarked against other standalone models to assess its baseline performance in forecasting electrical conductivity (EC) concentrations. Table 3 shows the APB, LMI, TS, and KGE values of models during the testing phase. The LSSVM model produces an APBI of 20.21% and a TS of 0.034. LSSVM also shows an improvement of 5% in TS compared to the MLR model.

Table 3 shows that the KGE value of the LSSVM model is 3.9% and 1.3% lower than that of the RNN and MLP. These results highlight that the LSSVM model has poorer performance compared to the MLP and RNN models. This underperformance of LSSVM can be attributed to improper adjustment of its parameters, which may not fully capture the temporal dependencies in the input time series.

Table 3 also shows that the RNN model achieves the lowest APB value of 16.78% and the highest KGE value of 0.812 among the standalone models. The ability of RNN to retain information from previous time steps enhances its predictive accuracy, allowing it to better track the dynamic changes in water quality parameters over time.

A comparison between HOA-LSSVM and LSSVM models: The HOA-LSSVM model shows an improvement of 25% in KGE values and 35% in LMI values compared to the LSSVM model. These results suggest that metaheuristic optimization techniques like HOA can significantly boost the predictive capability of traditional machine learning models such as LSSVM. By fine-tuning hyperparameters such as the regularization factor and kernel width, HOA enables LSSVM to better adapt to the nonlinear and dynamic characteristics of EC concentration data.

A comparison between CEEMD-LSSVM and LSSVM models: CEEMD-LSSVM produces an LMI of 0.921 and an APB of 7.38, while the LSSVM model produces an LMI of 0.798 and an APB of 20.21. These results demonstrate that the CEEMD-LSSVM hybrid model significantly outperforms the standalone LSSVM model in terms of both accuracy and stability. By breaking down the input signal into multiple frequency bands, CEEMD minimizes the noise and irregularities in raw data, which can enhance the overall learning efficiency of the LSSVM model.

A comparison between HOA-CEEMD-LSSVM and all other models: The HOA-CEEMD-LSSVM model enhances TS and KGE values of all other models by 24%-83% and 3.7%-17%. These results demonstrate that the HOA-CEEMD-LSSVM hybrid model significantly outperforms all other standalone and hybrid models in terms of predictive accuracy, correlation, and overall model efficiency. The substantial improvements in TS and KGE values indicate that integrating HOA, CEEMD, and LSSVM leads to a more robust and adaptive forecasting framework. Our findings suggest that the combination of decomposition, optimization, and machine learning techniques enhances the model's ability to capture complex nonlinear patterns and temporal dependencies in water quality parameters.

Table 3. Evaluation metrics of the different models at the testing level

Model	APB	LMI	TS	KGE
HOA-CEEMD-LSSVM	4.13	0.934	0.034	0.971
CEEMD-LSSVM	7.38	0.921	0.045	0.940
HOA-RNN	9.10	0.900	0.089	0.920
HOA-MLP	11.25	0.877	0.138	0.890
HOA-LSSVM	12.34	0.867	0.145	0.876
HOA-MLR	14.56	0.845	0.155	0.865
RNN	16.78	0.834	0.178	0.845
MLP	18.90	0.810	0.189	0.823
LSSVM	20.21	0.798	0.190	0.812
MLR	22.21	0.786	0.200	0.800

In Figure 6, heat scatterplots are displayed. The HOA-CEEMD-LSSVM model produces an R^2 value of 0.9987, which indicates its successful performance in forecasting concentrations. This superior performance can be attributed to the effective integration of the HOA, CEEMD, and LSSVM in the HOA-CEEMD-LSSVM hybrid model.

Heat scatterplots display that LSSVM and MLR produce R^2 values of 0.9034 and 0.8895, respectively. These results show that the HOA-CEEMD-LSSVM hybrid model outperforms these two models. As a linear regression technique, MLR struggles to capture the complex, nonlinear dynamics inherent in EC time series data. Its heat scatterplot reveals greater deviation from the 1:1 line.

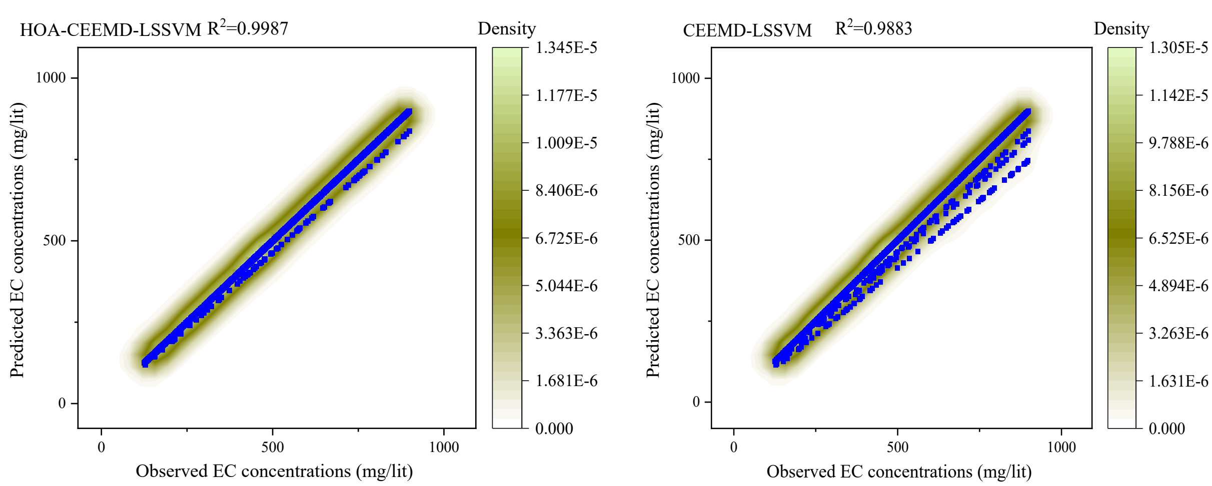


Figure 6. **cont.** Heat scatter plots of various models

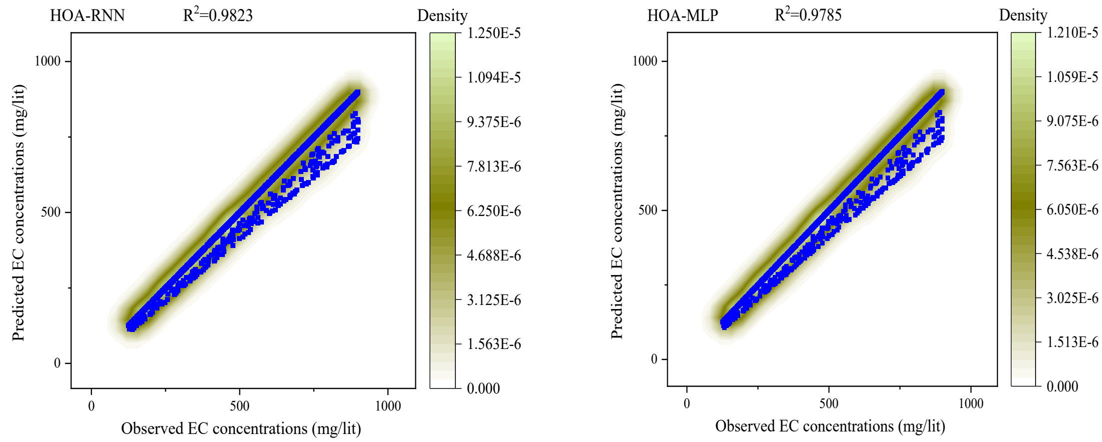


Figure 6. **cont.** Heat scatter plots of various models

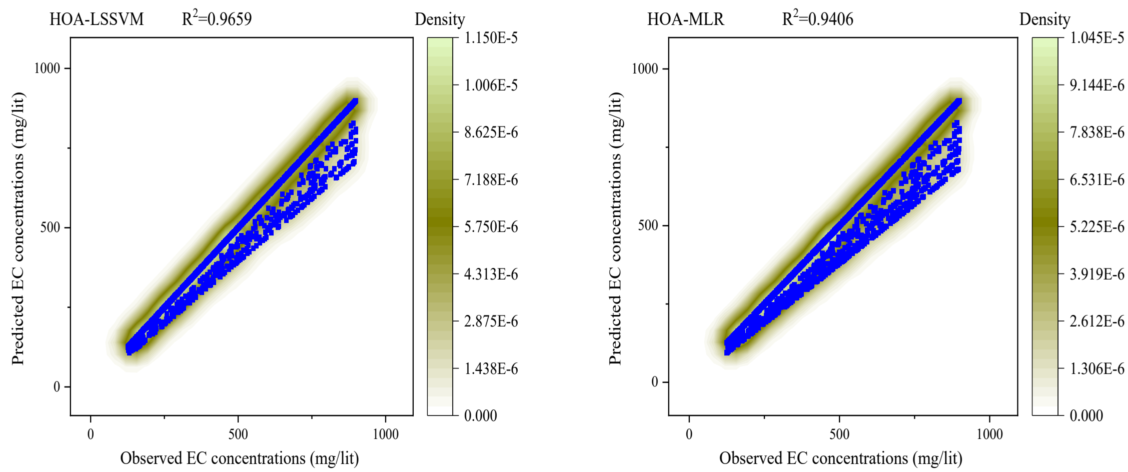


Figure 6. **cont.** Heat scatter plots of various models

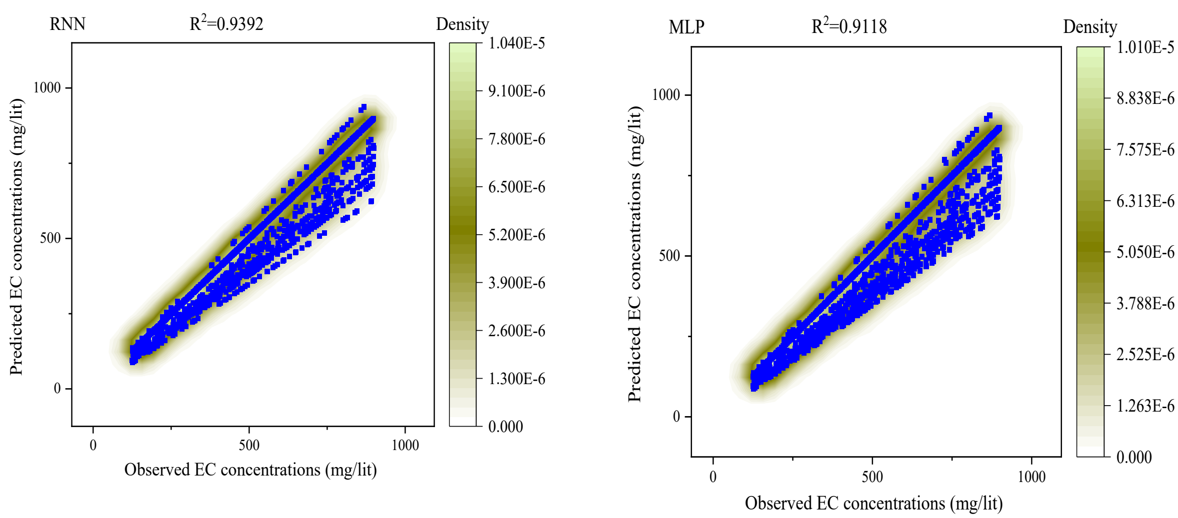


Figure 6. **cont.** Heat scatter plots of various models

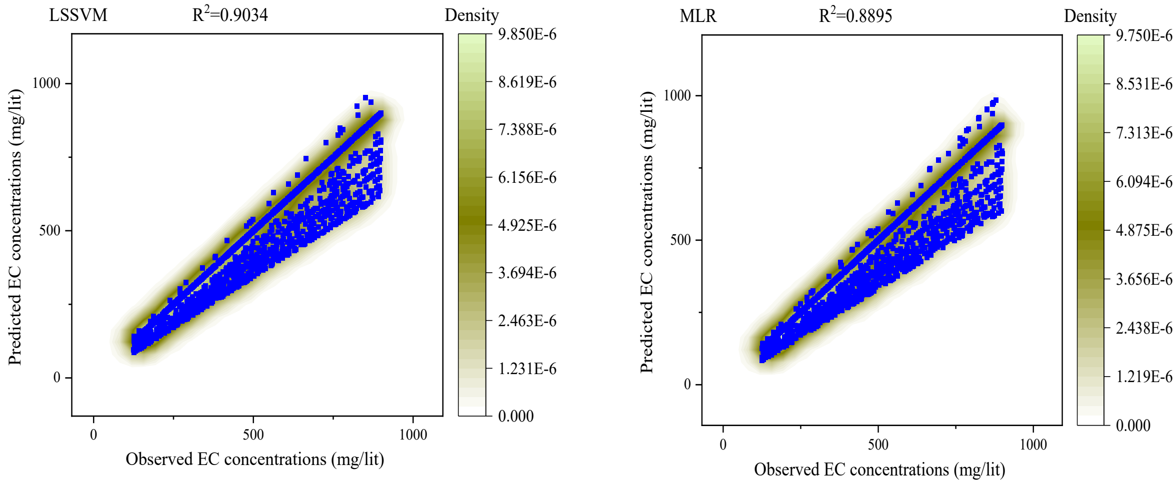


Figure 6. **cont.** Heat scatter plots of various models

In Figure 7, box plots of the models are displayed. The boxplot of the measured data shows a maximum value of 900. 23 mg/lit and a minimum value of 125.65 mg/lit. The difference between the two values is 774.58 mg/lit, which highlights the wide variability in electrical conductivity (EC) concentrations in the Aidoghmoush River over the study period. This broad range poses a significant challenge for predictive models, as they must accurately capture both low and high EC events in hydrological conditions.

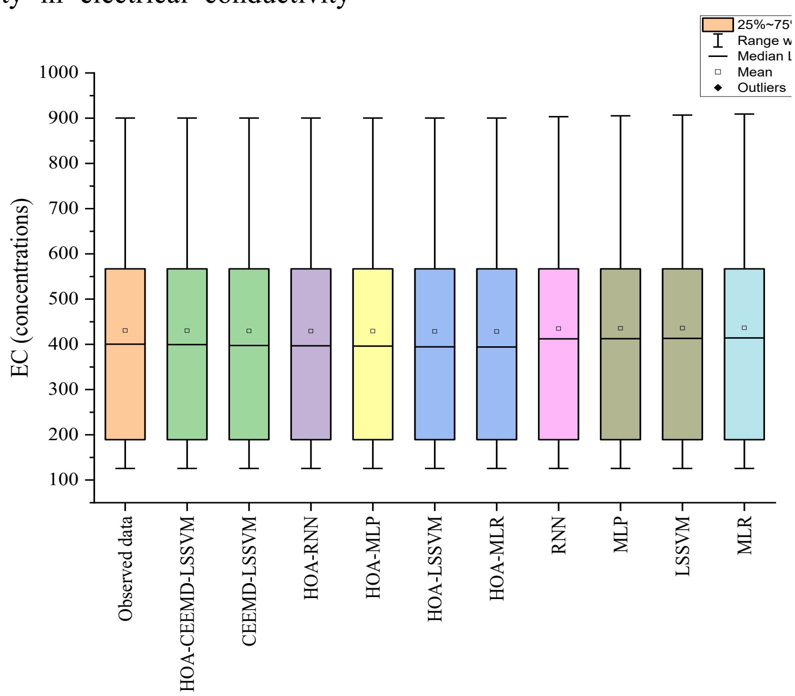


Figure 7. Boxplots of the various prediction models

The boxplot of the HOA-CEEMD-LSSVM model shows a median value of 399.285 mg/lit and an average value of 430.12 mg/lit, which are closely aligned with the median and mean values of the measured data. These results show that the HOA-CEEMD-LSSVM effectively captures the central tendency of the observed electrical conductivity (EC) data. In contrast, the boxplot of the MLR model shows that the model cannot effectively capture the central tendency of the measured EC data.

The boxplot of the LSSVM models shows a maximum value of 906.98 mg/lit, which is slightly higher than the observed maximum EC value of 900.23 mg/lit. The results indicate that the LSSVM model cannot effectively capture the central tendency of the observed EC data and tends to overestimate extreme EC concentrations in some cases. U95 is another index used to evaluate the uncertainty and reliability of predictive models. Figure 8 displays the predicted time series and their corresponding U95 values. The HOA-CEEMD-LSSVM hybrid model demonstrates a 62% to 90% enhancement in U95 values compared to other models. Thus, the HOA-CEEMD-LSSVM model produces more reliable predictions than other models.

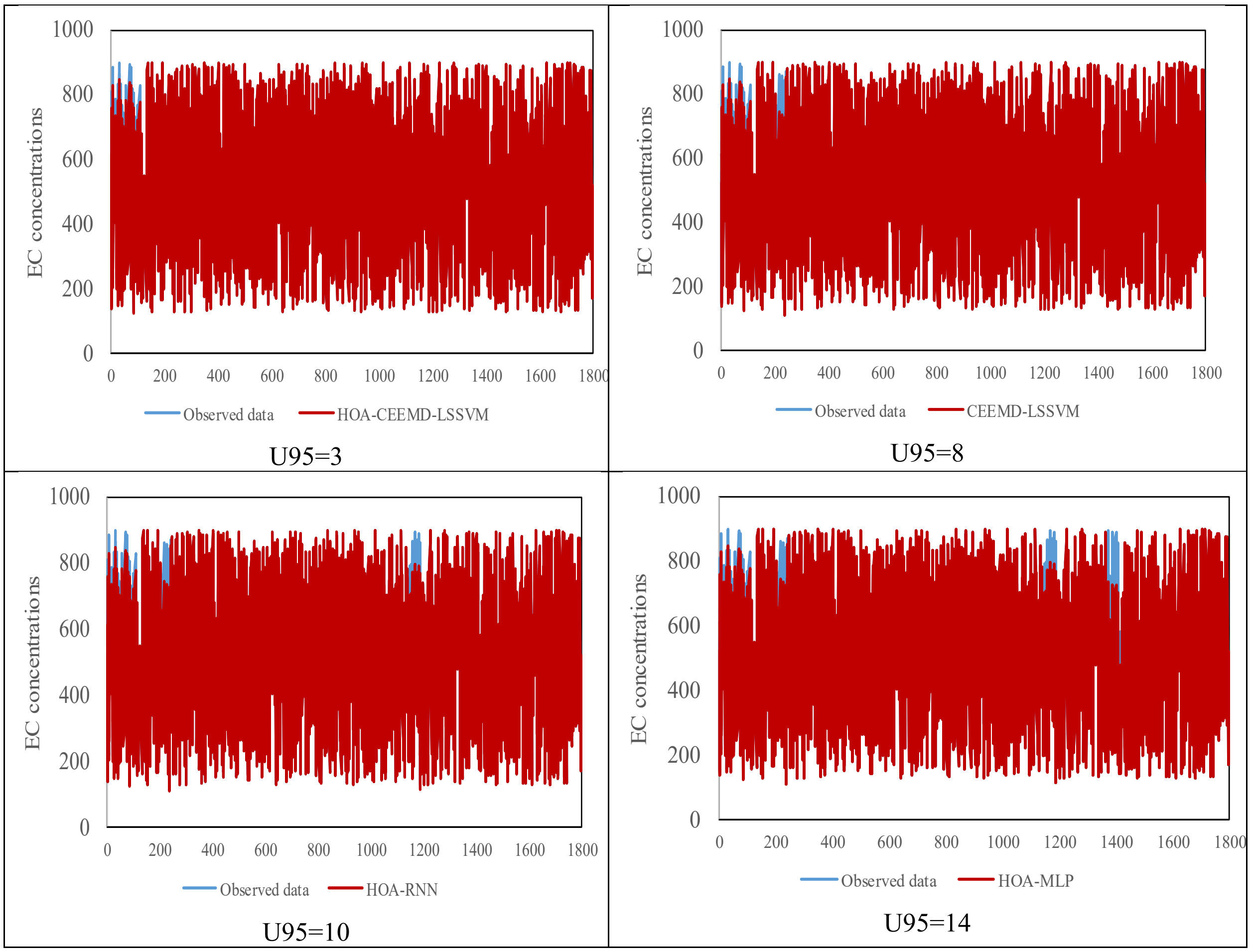


Figure 8. cont. Predicted and observed time series

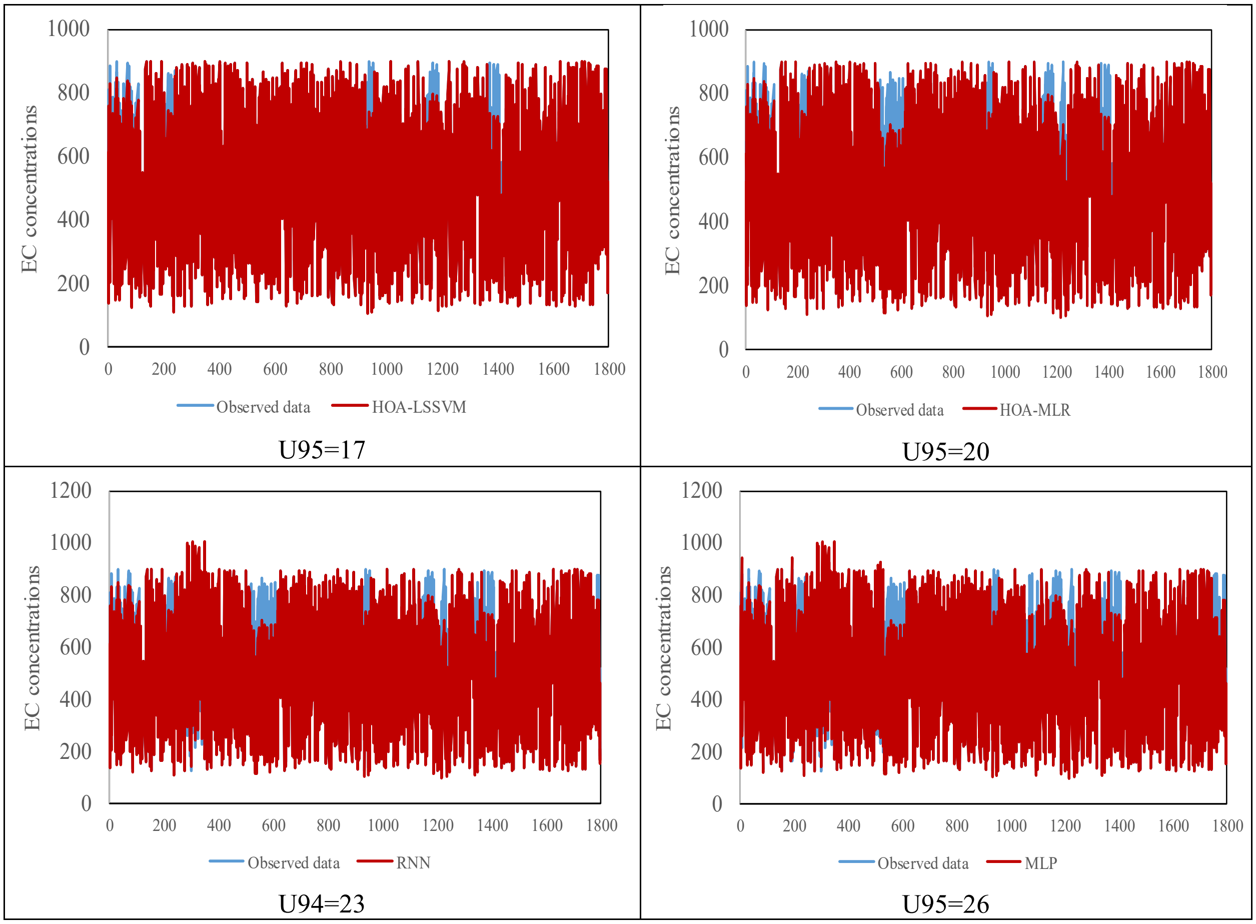


Figure 8. cont. Predicted and observed time series

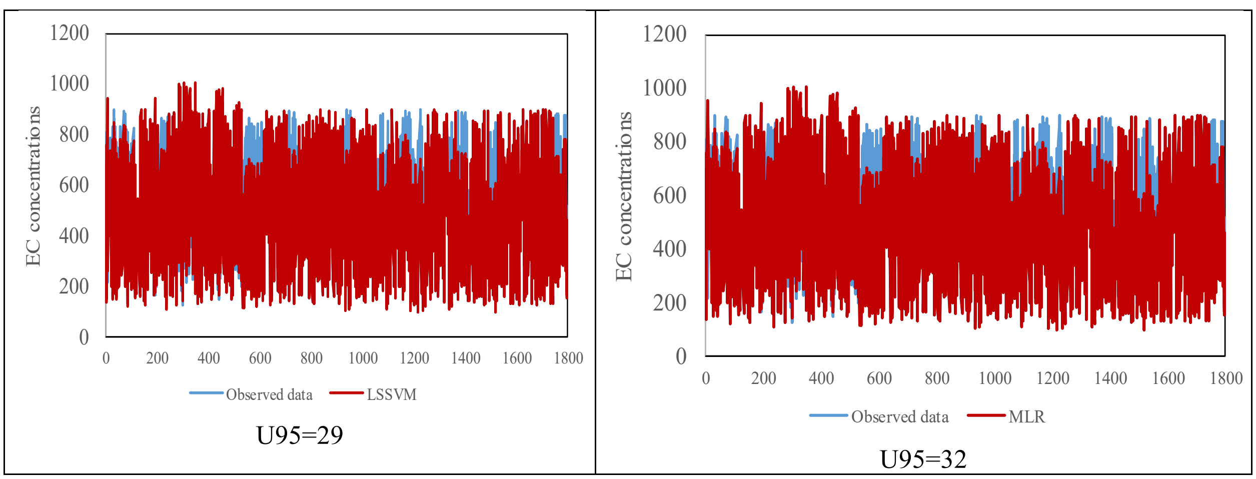


Figure 8. cont. Predicted and observed time series

4.6 Hypothetical Diebold–Mariano (DM) Test

The HOA-CEEMD-LSSVM model was treated as the benchmark and compared against all other models (Table 4).

Table 4. Hypothetical Diebold–Mariano (DM) Test Results

Comparison	Mean loss difference	DM statistic	p-value	| Interpretation
HOA-CEEMD-LSSVM vs CEEMD-LSSVM	0.012	2.06	0.040	Small but significant improvement of the benchmark
HOA-CEEMD-LSSVM vs HOA-RNN	0.025	3.1	0.002	Benchmark significantly better
HOA-CEEMD-LSSVM vs HOA-MLP	0.038	4.20	<0.001	Benchmark much better
HOA-CEEMD-LSSVM vs HOA-LSSVM	0.045	4.80	<0.001	Benchmark much better
HOA-CEEMD-LSSVM vs HOA-MLR	0.060	6.00	<0.0001	Benchmark strongly better
HOA-CEEMD-LSSVM vs RNN	0.070	6.80	<0.0001	Benchmark strongly better
HOA-CEEMD-LSSVM vs MLP	0.080	7.20	<0.0001	Benchmark strongly better
HOA-CEEMD-LSSVM vs LSSVM	0.095	8.10	<0.0001	Benchmark strongly better
HOA-CEEMD-LSSVM vs MLR	0.110	9.00	<0.0001	Benchmark strongly better

Note: Positive mean loss differences indicate that the competitor has a higher forecast loss than the benchmark (i.e., the benchmark model performs better)

4.7. A comparison between the current study and previous studies

The current study introduces the HOA-CEEMD-LSSVM model for predicting daily electrical conductivity (EC) concentrations in the Aidoghmoush River, Iran. Ekemen Keskin et al. (2020) developed various versions of ANN models to forecast EC concentrations. The model inputs included various water quality parameters that significantly affected EC concentrations. Their analysis showed that the best ANN model produced an R2 value of 0.979, which indicated a high level of predictive accuracy and a strong correlation between the predicted and observed EC values. However, the precision of the ANN model is lower than that of the HOA-CEEMD-LSSVM model, which produces an R2 value of 0.9987.

Kadkhodazadeh and Farzin (2021) used the gradient-based optimizer (GBO) algorithm-LSSVM model to predict EC concentrations. The GBO algorithm enhanced the performance of LSSVM by systematically searching for optimal hyperparameter settings. The GBO-LSSVM model produced an R2 value of 0.9491, which indicated a strong correlation between forecasted and observed EC values. However, the accuracy of the GBO-LSSVM model is lower than that of the HOA-CEEMD-LSSVM model, which achieves an exceptional R^2 value of 0.9987. Unlike the GBO-LSSVM model, which operates on raw input data, the HOA-CEEMD-LSSVM model incorporates the CEEMD algorithm. This algorithm converts the original time series into multiple IMFs that are more predictable and can enhance prediction stability and accuracy. By preprocessing the input data, the HOA-CEEMD-LSSVM model can capture temporal dependencies and nonlinear patterns in the data, leading to improved predictive performance. Jamei et al. (2023) developed the Boruta-XGBoost (BXGB)- Elman recurrent neural network (ERNN) to forecast EC concentrations. The XGBoost (BXGB) identified the most relevant input variables, while the ERNN model produced predictions. The BXGB-ERNN model produced a KGE value of 0.9675, which indicated a strong correlation between predicted and observed values. However, the precision of the BXGB-ERNN model is lower than that of the HOA-CEEMD-LSSVM model, which produces a KGE value of 0.971.

5. Discussion

In this study, the HOA-CEEMD-LSSVM model is used to forecast EC concentrations in the Aidoghmoush River, Iran. After analyzing the performance of this model and other prediction models, the following findings were obtained:

1)The HOA-CEEMD-LSSVM model demonstrates better performance in forecasting EC concentrations compared to other benchmark models such as CEEMD-LSSVM, standalone LSSVM, and ANN. The enhanced performance of the HOA-CEEMD-LSSVM model demonstrates the effectiveness of integrating optimization techniques with decomposition algorithms and artificial intelligence models for environmental time series forecasting. Specifically, the combination of HOA with the CEEMD-LSSVM framework plays a crucial role in improving prediction accuracy.

2) Our results show that HOA significantly enhances the performance of both CEEMD and LSSVM by optimizing critical parameters such as the number of IMFs, regularization factor, and kernel width. By systematically searching for optimal hyperparameter settings, HOA ensures that the model adapts effectively to the characteristics of each IMF, leading to faster convergence and more accurate predictions.

3) Our findings show that the CEEMD algorithm plays a key role in improving the forecasting accuracy of the LSSVM model. The CEEMD algorithm effectively decomposes complex, nonlinear, and nonstationary time series into multiple IMFs with more predictable patterns. These IMFs reduce noise interference, capture multiscale features of the original signal, and allow the LSSVM model to more effectively understand temporal dependencies and dynamic changes in water quality parameters.

4) The current study has significant implications for water quality management and environmental monitoring. The HOA-CEEMD-LSSVM model is a powerful and reliable tool for predicting electrical conductivity (EC) concentrations in river systems. The proposed model can continuously track changes in water quality parameters and forecast EC fluctuations under varying environmental and climate conditions. This capability is essential for supporting sustainable water resource planning, agricultural irrigation management, industrial process optimization, and ecosystem protection.

6. Conclusion

In this paper, the HOA-CEEMD-LSSVM model is used to forecast EC concentrations in the Aidoghmoush River, Iran. The model operates in several key steps that integrate an optimization algorithm, an advanced data processing technique, and a machine learning algorithm to enhance forecasting accuracy and reliability. First, HOA is employed to optimize both CEEMD and LSSVM parameters. It systematically adjusts critical hyperparameters such as the number of IMFs, the regularization parameter (γ), and the kernel width (σ) in LSSVM to ensure optimal model performance. Then, the CEEMD algorithm decomposes the time series of water quality parameters into IMFs with more predictable patterns that can enhance the accuracy of forecasts. Finally, the LSSVM model produces predictions using IMFs. The performance of the proposed model is analyzed using multiple evaluation metrics, and the results are summarized as follows:

- The results show that the integration of HOA, CEEMD, and LSSVM significantly improves predictive accuracy, robustness, and generalization capability.
- The results demonstrate that the integration of CEEMD with the LSSVM model significantly enhances the model's predictive performance.
- The results highlight the limitations of standalone machine learning and statistical models in capturing complex, nonlinear, and nonstationary patterns in the time series.

However, the HOA-CEEMD-LSSVM model may have restrictions that can affect its generalization and applicability in different fields. For example, HOA-CEEMD-LSSVM is a block box model and does not provide a transparent or interpretable representation of the relationships between input variables and output predictions. Moreover, while the integration of HOA with CEEMD and LSSVM enhances predictive performance, it also increases computational complexity and training time, which can limit the application of the hybrid model. However, future studies can overcome these limitations by exploring parallel computing techniques and explainable AI (XAI) methods, which can reduce computational costs and improve interpretability.

Author Contributions:
Elham GhanbariAdivi: Conceptualization, methodology, formal analysis and investigation, visualization, resources, writing-original draft preparation.
Jalil Kermannezhad: Conceptualization, supervision, formal analysis, and investigation.
Ali Raeisi: manuscript editing**.**

Conflicts of interest

The authors of this article declared no conflict of interest regarding the authorship or publication of this article.

Data availability statement:

All data generated or analyzed during this study are included in this published article.

References

Abdel-salam, M., Alomari, S. A., Almomani, M. H., Hu, G., Lee, S., Saleem, K., Smerat, A., & Abualigah, L. (2025). Quadruple strategy-driven hiking optimization algorithm for low and high-dimensional feature selection and real-world skin cancer classification. Knowledge-Based Systems, 315, 113286. doi: 10.1016/j.knosys.2025.113286

Ahmadianfar, I., Jamei, M., & Chu, X. (2020). A novel Hybrid Wavelet-Locally Weighted Linear Regression (W-LWLR) Model for Electrical Conductivity (EC) Prediction in Surface Water. Journal of Contaminant Hydrology. doi: 10.1016/j.jconhyd.2020.103641

Ahmadianfar, I., Shirvani-Hosseini, S., He, J., Samadi-Koucheksaraee, A., & Yaseen, Z. M. (2022). An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction. Scientific Reports. doi: 10.1038/s41598-022-08875-w

Aldrees, A., Awan, H. H., Javed, M. F., & Mohamed, A. M. (2022). Prediction of water quality indexes with ensemble learners: Bagging and boosting. Process Safety and Environmental Protection. doi: 10.1016/j.psep.2022.10.005

Alsaedi, A. W. M., Al-Hilphy, A. R., Al-Mousawi, A. J., & Gavahian, M. (2024). Artificial Neural Network Modeling to Predict Electrical Conductivity and Moisture Content of Milk During Non-Thermal Pasteurization: New Application of Artificial Intelligence (AI) in Food Processing. Processes, 12(11), 2507. doi: 10.3390/pr12112507

Ali Khan, M., Izhar Shah, M., Faisal Javed, M., Ijaz Khan, M., Rasheed, S., El-Shorbagy, M. A., Roshdy El-Zahar, E., & Malik, M. Y. (2022). Application of random forest for modelling of surface water salinity. Ain Shams Engineering Journal. doi: 10.1016/j.asej.2021.11.004

Chia, S. L., Chia, M. Y., Koo, C. H., & Huang, Y. F. (2022). Integration of advanced optimization algorithms into least-square support vector machine (LSSVM) for water quality index prediction. Water Supply. doi: 10.2166/ws.2021.303

Dulger Altıner, D., Yıkmış, S., Şimşek, M. A., Türkol, M., Tokatlı Demirok, N., & Celik, G. (2024). Impact of thermosonication treatment on parsley juice: particle swarm algorithm (PSO), multiple linear regression (MLR), and response surface methodology (RSM). ACS omega, 9(27), 29585-29597. doi: 10.1021/acsomega.4c02749

Ehteram, M., & Soltani-Gerdefaramarzi, S. (2025). Advanced hybrid frameworks for water quality index prediction. Ain Shams Engineering Journal, 16(8), 103478. doi: 10.1016/j.asej.2025.103478

Ekemen Keskin, T., Özler, E., Şander, E., Düğenci, M., & Ahmed, M. Y. (2020). Prediction of electrical conductivity using ANN and MLR: a case study from Turkey. Acta Geophysica. doi: 10.1007/s11600-020-00424-1

Ghanbari-Adivi, E., Ehteram, M. CEEMDAN-BILSTM-ANN and SVM Models: Two Robust Predictive Models for Predicting River flow. Water Resour Manage 39, 3235–3271 (2025). doi: 10.1007/s11269-025-04105-w

Jamei, M., Ali, M., Karimi, B., Karbasi, M., Farooque, A. A., & Yaseen, Z. M. (2023). Surface water electrical conductivity and bicarbonate ion determination using a smart hybridization of optimal Boruta package with Elman recurrent neural network. Process Safety and Environmental Protection, 174, 115-134. doi: 10.1016/j.psep.2023.03.062

Jamshidzadeh, Z., Latif, S. D., Ehteram, M., Sheikh Khozani, Z., Ahmed, A. N., Sherif, M., & El-Shafie, A. (2024). An advanced hybrid deep learning model for predicting total dissolved solids and electrical conductivity (EC) in coastal aquifers. Environmental Sciences Europe. doi: 10.1186/s12302-024-00850-8

Kadkhodazadeh, M., & Farzin, S. (2021). A Novel LSSVM Model Integrated with GBO Algorithm to Assessment of Water Quality Parameters. Water Resources Management. doi: 10.1007/s11269-021-02913-4

Karbasi, M., Ali, M., Bateni, S. M., Jun, C., Jamei, M., Farooque, A. A., & Yaseen, Z. M. (2024). Multi-step ahead forecasting of electrical conductivity in rivers by using a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model enhanced by Boruta-XGBoost feature selection algorithm. Scientific reports, 14(1), 15051. doi: 10.1038/s41598-024-65837-0

Khatti, J., Khanmohammadi, M., & Fissha, Y. (2024). Prediction of time-dependent bearing capacity of concrete pile in cohesive soil using optimized relevance vector machine and long short-term memory models. Scientific Reports, 14(1), 32047. doi: 10.1038/s41598-024-83784-8

Kumar, D., Singh, V. K., Abed, S. A., Tripathi, V. K., Gupta, S., Al-Ansari, N., Vishwakarma, D. K., Dewidar, A. Z., Al-Othman, A. A., & Mattar, M. A. (2023). Multi-ahead electrical conductivity forecasting of surface water based on machine learning algorithms. Applied Water Science. doi: 10.1007/s13201-023-02005-1

Li, H., Li, X., Yang, Z., Liu, Z., Bu, J., & Wang, Y. (2024). Loaded coal-rock temperature denoising algorithm based on CEEMD and adaptive NIWT with NIWOA. Measurement, 236, 115176. doi: 10.1016/j.measurement.2024.115176

Mattas, C., Dimitraki, L., Georgiou, P., & Venetsanou, P. (2021). Use of factor analysis (Fa), artificial neural networks (anns) and multiple linear regression (mlr) for electrical conductivity prediction in aquifers in the gallikos river basin, Northern Greece. Hydrology. doi: 10.3390/hydrology8030127

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. Information, 15(9), 517. doi: 10.3390/info15090517

Muhammad, A. U., Djigal, H., Muazu, T., Adam, J. M., Ba, A. F., Dabai, U. S., Tijjani, S., Yahaya, M. S., Ashiru, A., Kumshe, U. M. M., Aliyu, S., & Richard, F. A. (2023). An autoencoder-based stacked LSTM transfer learning model for EC forecasting. Earth Science Informatics. doi: 10.1007/s12145-023-01096-3

Nakhaei-Kohani, R., Ali Madani, S., Mousavi, S. P., Atashrouz, S., Abedi, A., Hemmati-Sarapardeh, A., & Mohaddespour, A. (2022). Machine learning assisted Structure-based models for predicting electrical conductivity of ionic liquids. Journal of Molecular Liquids. doi: 10.1016/j.molliq.2022.119509

Özcan, A. R., Mehta, P., Sait, S. M., Gürses, D., & Yildiz, A. R. (2025). A new neural network–assisted hybrid chaotic hiking optimization algorithm for optimal design of engineering components. Materials Testing, (0). doi: 10.1515/mt-2024-0519

Oladejo, S. O., Ekwe, S. O., & Mirjalili, S. (2024). The Hiking Optimization Algorithm: A novel human-based metaheuristic approach. Knowledge-Based Systems, 296, 111880. doi: 10.1016/j.knosys.2024.111880

Pashaei, E., Pashaei, E., & Mirjalili, S. (2025). Binary hiking optimization for gene selection: Insights from HNSCC RNA-Seq data. Expert Systems with Applications, 268, 126404. doi: 10.1016/j.eswa.2025.126404

Reza, A. F., Singh, R., Verma, R. K., Singh, A., Ahn, Y. H., & Ray, S. S. (2024). An integral and multidimensional review on multilayer perceptron as an emerging tool in the field of water treatment and desalination processes.

Desalination, 117849. doi: 10.1016/j.desal.2024.117849

Sağ, T. (2024, September). Binary Hiking Optimization Algorithm. In International Conference on Cellular Automata for Research and Industry (pp. 231-242). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-71552-5_19

Shah, M. I., Javed, M. F., & Abunama, T. (2021). Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques. Environmental Science and Pollution Research. doi: 10.1007/s11356-020-11490-9.

Shin, J., Yoon, S., Park, N. S., & Kim, Y. (2025). Development of Water Quality Prediction Model Using LTSF-Linear and Complete Ensemble Empirical Mode Decomposition. Desalination and Water Treatment, 101254. doi: 10.1016/j.dwt.2025.101254

Song, C., Yao, L., Hua, C., & Ni, Q. (2021). A water quality prediction model based on variational mode decomposition and the least squares support vector machine optimized by the sparrow search algorithm (VMD-SSA-LSSVM) of the Yangtze River, China. Environmental Monitoring and Assessment. doi: 10.1007/s10661-021-09127-6

Talebzadeh, H., Talebzadeh, M., Satarpour, M., Jalali, F., Farhadi, B., & Vahdatpour, M. S. (2024). Enhancing breast cancer diagnosis accuracy through genetic algorithm-optimized multilayer perceptron. Multiscale and Multidisciplinary Modeling, Experiments and Design, 7(4), 4433-4449. doi: 10.1007/s41939-024-00487-3

Wu, K., Guo, Y., Wang, K., & Chen, Z. (2025). Low-carbon economic optimization for flexible DC distribution networks based on the hiking optimization algorithm. Energy Informatics, 8(1), 44. doi: 10.1186/s42162-025-00486-9

Xu, D. Mei, Li, Z., & Wang, W. Chuan. (2024). An ensemble model for monthly runoff prediction using least squares support vector machine based on variational modal decomposition with dung beetle optimization algorithm and error correction strategy. Journal of Hydrology. doi: 10.1016/j.jhydrol.2023.130558

Yahia Ahmed Abuker, Y., Liu, Z., & Mao, L. (2025). Water Management Fault Diagnosis of Polymer Electrolyte Membrane Fuel Cells Based on Complementary Ensemble Empirical Mode Decomposition and Sensitivity Analysis. Journal of Energy Engineering, 151(3), 05025001. doi: 10.1061/JLEED9.EYENG-5508

Zhang, X., Wu, X., He, S., & Zhao, D. (2021). Precipitation forecast based on CEEMD–LSTM coupled model. Water Supply. doi: 10.2166/ws.2021.237

Zhou, M., Zhang, Y., Wang, J., Shi, Y., & Puig, V. (2022). Water Quality Indicator Interval Prediction in Wastewater Treatment Process Based on the Improved BES-LSSVM Algorithm. Sensors. doi: 10.3390/s22020422

Zhao, M., & Zhou, X. (2024). Multi-Step short-term wind power prediction model based on CEEMD and improved snake optimization algorithm. IEEE Access. doi: 10.1109/ACCESS.2024.3385643

Zhou, M., Yu, J., Wang, M., Quan, W., & Bian, C. (2024). Research on the combined forecasting model of cooling load based on IVMD-WOA-LSSVM. Energy and Buildings, 317, 114339. doi: 10.1016/j.enbuild.2024.114339