An intelligent approach for estimating aeration efficiency in stepped cascades: optimized support vector regression models and mutual Information theory

Seyed Babak Haji Seyed Asadollah[1], Ahmad Sharafati [1*], Masoud Haghbin[2], Davide Motta[3], Mohamadreza Hosseinian Moghadam Noghani[1]

1. Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

2. Department of Structural Mechanics & Hydraulics Engineering, University of Granada, 18001, Spain

3. Department of Mechanical and Construction Engineering, Northumbria University, Wynne Jones Building, Newcastle upon Tyne, NE1 8ST, United Kingdom

Corresponding author: Ahmad Sharafati

Corresponding email: asharafati@gmail.com, asharafati@srbiau.ac.ir

Abstract

Soft Computing (SC) methods have increasingly been used to solve complex hydraulic engineering problems, especially those characterized by high uncertainty. SC approaches have previously proved to be an accurate tool for predicting the Aeration Efficiency coefficient (E20) in hydraulic structures such as weirs and flumes. In this study, the performance of the standalone Support Vector Regression (SVR) algorithm and three of its hybrid versions, Support Vector Regression -Firefly Algorithm (SVR-FA), -Grasshopper Optimization Algorithm (SVR-GOA), and -Artificial Bee Colony (SVR-ABC), is assessed for the prediction of E20 in stepped cascades. Mutual Information theory is used to construct input variable combinations for prediction, including the parameters unit discharge (q), the total number of steps (N), step height (h), chute overall length (L), and chute inclination (α). Entropy indicators, such as Maximum Likelihood, Jeffrey, Laplace, Schurmann-Grassberger, and Minimax, are computed to quantify the epistemic uncertainty associated with the models. Four indices - Correlation Coefficient (R), Nash-Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) - are employed for evaluating the models' prediction performance. The models' outputs reveal that the SVR-FA model (with $R = 0.947,\ NSE = 0.888,\ RMSE = 0.048\ and\ MAE = 0.027$ in testing phase) has the best performance among all the models considered. The input variable combination, including q, N, h, and L, provides the best predictions with the SVR, SVR-FA, and SVR-GOA models. From the uncertainty analysis, the SVR-FA model shows the closest entropy values to the observed ones (3.630 vs. 3.628 for the "classic" entropy method and 3.647 vs. 3.643 on average for the Bayesian entropy method). This study proves that SC algorithms can be highly accurate in simulating aeration efficiency in stepped cascades and provide a valid alternative to the traditional empirical equation.

Keywords: Aeration Efficiency, Stepped Cascades, Support Vector Regression, Optimization Algorithm

1. Introduction

Stepped cascades have many applications in river engineering, as dam structures or aeration cascades. They generally are efficient aeration structures based on air bubble entrainment, large residence time, and significant turbulent mixing (Toombes and Chanson, 2020), with aeration being a physical process (Baylar et al., 2007c) that brings water and air in close contact to increase the amount of Dissolved Oxygen (DO) (Izadi et al., 2021). Aeration cascades in waterways (Baylar et al., 2009) are common in water treatment (Baylar et al., 2007c; Hanbay et al., 2009a) for re-oxygenation and chlorine elimination (Baylar et al., 2010), denitrification and volatile organic compound removal (Baylar et al., 2007c), and volatile organic compounds stripping (Toombes and Chanson, 2005).

Stepped cascades have been used for a long time, at least for the last 3500 years, since in Roman settlements, stepped cascades were used to prevent erosion and damage to dams (Chanson, 2000). In relatively recent times, this kind of structure was progressively abandoned (Felder and Chanson, 2009) until the 1980s, with the development of new and efficient construction techniques that boosted the use of these structures (Jiang, Diao, Xue, & Sun, 2018, Salmasi, Sattari, & Nurcheshmeh, 2021). As energy dissipators, stepped cascades have multiple advantages: reduction of the required depth and size of the stilling basins (Peyras et al., 1992), significant energy loss rate, and compatibility with roller-compacted concrete (RCC) dams (Sengun et al., 2021).

The flow regime on stepped cascades is classified into three types: nappe flow, transition flow, and skimming flow (Salmasi et al., 2021). Nappe flow is characterized by a series of free-falling nappes plunging from one step to the next and is typical of low discharges (Toombes and Chanson, 2008). As the flow rate increases, the nappe impact near the steps disappears, creating a situation similar to stagnation. This regime, called transition flow, is characterized by significant aeration, water splashing, and chaotic appearance, with the flow characteristics varying from step to step. For even larger discharge, skimming flow is established: the nappe impact caused by the spillway disappears completely, and the water flows as a stream over the pseudo-bottom.

Many studies were carried out to investigate various aspects of aeration and stepped cascades, most of them via experiments. (Gameson, 1957) was the first to research aeration by stepped cascades and the use of weirs to accelerate the aeration process; (Tebbutt, 1972) measured the Aeration Efficiency in a laboratory stepped cascade; (Essery et al., 1978) proposed a formulation to predict the Aeration Efficiency of cascades for discharges between 1.5 liters per second and 22 liters per second, with step heights between 0.025 and 0.5 meters; (Toombes and Chanson, 2005) studied the oxygenation on a stepped cascade with low chute slope and for high discharges between 19 and 300 liters per second. In recent years, different research groups (Baylar et al., 2007a, 2007b, 2007c, 2006; Ahmet Baylar and Emiroglu, 2003; Emiroglu and Baylar, 2003; Hanbay et al., 2009b, 2009a) performed a number of water aeration experiments on laboratory stepped cascades. Although these research groups expressed oxygenation with various correlations, it should be noted that these correlations are purely empirical, which may therefore ignore the effect of some important parameters and may be applicable for a limited number of conditions (Khdhiri et al., 2014).

In recent years, advances have been made in developing and applying Soft Computing (SC) methods in engineering. These methods have been utilized in previous studies of hydraulic structures and have shown high performance in simulating discharge coefficient in a stepped morning glory spillway (Haghbin et al., 2022), scouring depth (Sharafati et al., 2020), and aeration efficiency in a Parshall flume (Sangeeta et al., 2021). Among the latest contributions to the field, (Sammen et al., 2020) compared three versions of Artificial Neural Network (ANN) algorithms hybridized with Harris Hawks Optimization (HHO), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA) to simulate the ski-jump scouring depth downstream of a spillway. These three hybrid algorithms were then evaluated based on several performance metrics, and the ANN-HHO model was ranked as the best among the models considered. (Sihag et al., 2021) forecasted the aeration efficiency in Parshall and Venturi flumes using various SC algorithms such as Random Forest (RF), tree-based M5P, Group Method of Data Handling (GMDH), and Multivariate Adaptive Regression Splines (MARS). Their results showed the MARS model to be the better predictor. So far, just a few researchers have applied SC methods for stepped cascades and have focused mainly on energy dissipation (Jiang et al., 2018; Salmasi et al., 2021).

One of the most popular SC algorithms is Support Vector Regression (SVR), which was developed based on the work by (Cortes and Vapnik, 1995) to model and control complex engineering systems (Panahi et al., 2020). SVR provides a non-linear mapping function that maps the training dataset in a high-dimension feature space and identifies the connections between input and output (Panahi et al., 2020). This algorithm has been adopted as a predictive tool in various applications, such as in meteorological drought (Malik et al., 2021), solar energy (Lima et al., 2022), and flood susceptibility (Saha et al., 2021), proving to be an efficient forecasting technique. In some cases, the SVR model was outperformed by other SC algorithms such as ANNs, GAs, MARS, and RF (Al-Musaylh et al., 2018; Chen, 2007; İskenderoğlu et al., 2020; Mirarabi et al., 2019). While the standalone SVR limitations in prediction performance are evident, the recent use of SVR hybridization with other optimization algorithms has shown promise.

To the authors' knowledge, there are very few previous studies that evaluated the application of hybrid SC algorithms in hydraulic engineering, especially in energy dissipation. The present study aims to bridge this scientific gap, applying for the first time several hybrid SVR models - Artificial Bees Colony (ABC), Grasshopper Optimization Algorithm (GOA), and Firefly Algorithm (FA) - to predict the Aeration Efficiency in stepped cascades. An additional element of novelty of this paper is the use of Mutual Information theory to pre-process the data and structure the best combination of input variables for prediction. The results are assessed based on several statistical and graphical evaluators, and the associated uncertainty is evaluated using "classic" and Bayesian types of entropy indicators. Overall, this study shows that SC algorithms are a valid alternative to empirical formulations and numerical simulations for the prediction of Aeration Efficiency in stepped cascades and possibly in other applications in hydraulic engineering.

2. Material and Methods

2.1 Support Vector Regression

Support Vector Regression (SVR) is one of the supervised machine learning algorithms developed by (Vapnik et al., 1997) for classification and regression tasks. With SVR, training data are used to obtain a predictive model to test against testing data. One of the key parameters in SVR models is the Structural Risk Minimization (SRM), which determines the relationship between input and output variables (Rastogi and Sharma, 2021), and is calculated as follows

$$y = k(z) = v\emptyset(z) + c \tag{1}$$

where k denotes the kernel function, z is the input data, v is a weight factor, c is a constant, and $\emptyset(z)$ represents the feature function. The following equations are used to define v and c

$$\text{Minimize: } [\frac{1}{2}\left||v|\right|^2 + P\sum_{b=1}^{n}(\vartheta_b + \vartheta^*{}_b)] \tag{2}$$

$$\text{Subject to: } \begin{cases} y_b - (v\emptyset(z_b) + c_b) \leq \varepsilon + \vartheta_b \\ (v\emptyset(z_b) + c_b) - y_b \leq \varepsilon + \vartheta^*{}_b \end{cases}$$

$$\vartheta_b, \vartheta^*{}_b \geq 0 \tag{3}$$

where P is the penalty factor, ϑ_b *and* $\vartheta^*{}_b$ are the loose variables, and ε is the optimized performance of the model (Su et al., 2018; Wang et al., 2012). The following equation is used to solve the optimization problem

$$L(v, c, \vartheta_b, \vartheta^*{}_b, B_b, B^*{}_b, \delta_b, \delta^*{}_b) \tag{4}$$

$$= \frac{1}{2} \left\| |v|^2 + P \sum_{b=1}^{1} (\vartheta_b + \vartheta^*{}_b \right.$$

$$- \sum_{b=1}^{1} B_b (\vartheta_b + \varepsilon - y_b + \mathrm{v}\emptyset(\mathrm{z}z_b) + \mathrm{c})$$

$$- \sum_{b=1}^{1} B^*{}_b (\vartheta^*{}_b + \varepsilon + y_b - \mathrm{v}\emptyset(z_b z) - \mathrm{c})$$

$$\left. - \sum_{b=1}^{1} (\delta_b \vartheta_b + \delta^*{}_b \vartheta^*{}_b)) \right\|$$

where $B_b . B^*{}_b . \delta_b$ and $\delta^*{}_b$ are the Lagrange multipliers, and L is the Lagrangian function. So, SVR can be calculated as follows

$$k(z) = \sum_{b=1}^{1} (B_b - B^*{}_b) m(z.z_b) + c \qquad (5)$$

where the kernel function is expressed as follows:

$$m(z, z_b) = \langle \emptyset(z) . \emptyset(z_b) \rangle \qquad (6)$$

2.2 Firefly Algorithm

The Firefly Algorithm (FA) is an optimization technique inspired by the characteristics of fireflies in nature. This algorithm was developed by (X. Yang, 2010) and is based on a metaheuristic approach that uses a specific repetitive generation procedure to solve optimization problems (Johari et al., 2013). The FA originated from formulating the fireflies' flashing demeanor and attraction based on their bioluminescence (Wu et al., 2020; X. S. Yang, 2010). The algorithm was structured based on the following assumptions:

1- the attraction process between the fireflies is not biased by sexuality and is considered unisexual.
2- The attraction level directly relates to the firefly brightness, so subsequently will become lesser with increasing distance between the two species.

3- If the level of bioluminescence is considered to be equal in two fireflies, their movement will be random.

Each firefly has a light intensity I that is calculated as follows

$$I = I_0 e^{-\gamma r_{ij}^2} \tag{7}$$

where r represents the distance of the observer from the source, I_0 is the firefly brightness intensity at $r = 0$, γ is the light absorption coefficient, and r_{ij} is the Euclidean distance between the firefly individuals i and j that is calculated as follows:

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{d=1}^{D} (x_{id} - x_{jd})^2} \tag{8}$$

where x_{id} is the $d-th$ component of the spatial coordinate x_i of the $i-th$ firefly, and D is the total number of dimensions.

The attractiveness β is calculated as follows

$$\beta = \beta_0 e^{-\gamma r_{ij}^2} \tag{9}$$

where β_0 is the attractiveness at $r = 0$.

The position update of firefly i, moving towards firefly j, is computed as follows

$$x_{id}(t+1) = x_{id}(t) + \beta \left(x_{jd}(t) - x_{id}(t) \right) + \alpha\varepsilon \tag{10}$$

where x_{id} and x_{jd} are the positions of firefly i and j in the d dimension, α is the step size factor, and ε is a random number with uniform distribution in the range $[-0.5, 0.5]$.

2.3 Grasshopper Optimization Algorithm

The Grasshopper Optimization Algorithm (GOA) was developed by (Saremi et al., 2017) and is a recently developed nature-inspired, population-based technique that mimics the behavior of grasshopper swarms. Grasshoppers are harmful pests that endanger agricultural production. Their life consists of two consecutive periods, known as “nymph” and “maturity”, respectively. The former phase is characterized by small and gradual movement, while the latter is characterized by long and fast motion. The movement in these two phases determines diversification and intensification in GOA (Ewees et al., 2020; Meraihi et al., 2021). One major assumption of the algorithm is that the gravity force does not affect the movement of grasshoppers, which leads to a faster speed of results convergence (Qin et al., 2021).

The position of the $i-th$ grasshopper Y_i is expressed as follows

$$Y_i = So_i + Gr_i + Aw_i \tag{11}$$

and the random swarm behavior is computed as follows

$$Y_i = r_1 * So_i + r_2 * Gr_i + r_3 * Aw_i \tag{12}$$

where r_1 , r_2 , and r_3 are random numbers between 0 to 1, Gr_i is the gravitational force on the $i-th$ grasshopper, Aw_i is the advection of wind and So_i is defined as follows

$$So_i = \sum_{\substack{j=1 \\ i \neq j}}^{n} sf(D_{ij})\widehat{D_{ij}} \tag{13}$$

where D_{ij} is the distance value between the $i-th$ and the $j-th$ grasshopper, sf is a mathematical function to determine the power of social organizations, and $\widehat{D_{ij}}$ is a unity vector from the $i-th$ grasshopper to the $j-th$ grasshopper. The sf function is expressed as follows

$$sf(r) = Ie^{-r/ln} - e^{-r} \tag{14}$$

where I is the attraction intensity, r is a random number between 0 to 1, and ln is the attractive length scale. The Gr_i and Aw_i components are calculated as follows

$$Gr_i = -g\widehat{e_{gr}} \tag{15}$$

$$Aw_i = d\widehat{e_w} \tag{16}$$

where g is the gravitational constant, $\widehat{e_{gr}}$ is the unit vector towards the center of the earth, d is a constant value and $\widehat{e_w}$ is the unit vector in the direction of the wind. So

$$Y_i = \sum_{\substack{j=1 \\ i\neq j}}^{n} sf\left(\left|Y_j - Y_i\right|\right)\left(\left|Y_j - Y_i\right|/D_{ij}\right) - gr\widehat{e_{gr}} + d\widehat{e_w} \tag{17}$$

$$Y_i^d = cz * \left\{\sum_{\substack{j=1 \\ i\neq j}}^{n} cz(|ul_d - ll_d|/2) sf\left(\left|Y_j{}^d - Y_i{}^d\right|\right)\left(\left|Y_j - Y_i\right|/D_{ij}\right)\right\} + \widehat{T_d} \tag{18}$$

where n is the number of grasshoppers, ul_d and ll_d are the upper and lower limits in the $D-th$ dimension, $\widehat{T_d}$ is the value of the target (current best solution) in the $D-th$ dimension, and the coefficient cz decreases the comfort zone proportional to the number of iterations. The calculation of cz is as follows

$$cz = cz_{max} - (t \times ((cz_{max} - cz_{min})/t_{max})) \tag{19}$$

where cz_{max} is the maximum value, cz_{min} is the minimum value, t is the current iteration, and t_{max} is the maximum number of iterations.

2.4 Artificial Bee Colony

The Artificial Bee Colony (ABC) algorithm is inspired by the honeybee swarm intelligent conduct. The ABC algorithm was introduced by (Teodorovic et al., 2006) to train neural networks. In the algorithm, the location of a food source (FS) represents a problem’s possible solution, and the amount of nectar in that specific source indicates the appropriateness of that solution. The value of employed bees (EBs) equals the number of FSs, and the EBs initially search for a food source. When found, they evaluate its fitness. FSs with a low amount of nectar are eliminated, and the FS search procedure is repeated until the set criteria are satisfied.

In the initial phase, the population of food sources is generated within boundaries that are delineated by x_j^{min} and x_j^{max}, where x_j^{min} is the lower bound in the jth dimension and x_j^{max} is the upper bound in the same dimension.

$$x_i^j = x_j^{min} + r_i^j \times (x_j^{max} - x_j^{min}) \\ i = 1,2,\dots,N \\ j = 1,2,\dots,D \tag{20}$$

where x is the food source position, i is the food source index, N is the number of food sources, D is the dimensionality of the optimization problem, and r_i^j represents a uniform distribution of real numbers in the range [0,1].

In the employed phase, each employed bee is allocated to one food source in the entire search space

$$cx_{s,d} = x_{s,d} + \varphi\left(x_{s,d} - x_{t,d}\right) \\ t \neq s \\ t \in \{1,2,\dots,N\} \\ d \in \{1,2,\dots,D\} \tag{21}$$

where cx is the candidate position, s is the selected bee index, t is the target bee index, N is the employed bees' number, and φ is a uniformly distributed number in the range [-1,1].

In the onlooker phase, the onlooker bees search for new food sources. The probability of being selected for a food source depends on the fitness value, which can be calculated as follows

$$p_s = fit_s / \sum fit_s \tag{22}$$

where p_s is the selection probability for the food source, and fit_s is the corresponding fitness value, which is calculated as follows

$$fit_s = \begin{cases} \dfrac{1}{fv(x_s)+1}, & fv(x_s) \geq 0 \\ 1 + |fv(x_s)|, & fv(x_s) < 0 \end{cases} \tag{23}$$

where fv is the function value of the objective function and fit_s is designed for minimization. In the scouting phase, the counter of food sources is checked. If the counter with the highest value is larger than the predefined parameter limit, the corresponding food source is considered exhausted, and the associated employed bee becomes a scout bee to make a new food source through the x_{ij} equation (Zhou et al., 2021).

2.5 Hybrid SVR Models

Optimizing the SVR model parameters - regularization parameter C, error margin ε, and RBF kernel parameter σ - is not straightforward, especially for highly non-linearity problems. The conceptual flowchart for optimizing the SVR model parameters is shown in Figure 1.

[Figure 1]

This study uses the FA, GOA, and ABC algorithm to optimize the SVR model parameters. The resulting models are identified in this paper as SVR-FA, SVR-GOA, and SVR-ABC. Table 1 reports the value of the relative optimization parameters.

[Table 1]

2.6 Laboratory Data for Aeration Efficiency Evaluation

The DO is a key parameter in the aquatic ecosystem that has a direct impact on the life of aquatic species, especially when DO concertation decreases to less than 5 $\frac{mg}{l}$ (Asadollah et al., 2021). Turbulent conditions and bubble formation directly influence the DO and its spatial gradients in water (Sangeeta et al., 2021).

Where the oxygen concentration varies between an upstream and a downstream section, a mathematical expression for oxygen Aeration Efficiency (E) can be introduced as follows (Gulliver et al., 1998):

$$E = \frac{C_D - C_U}{C_S - C_U} \tag{24}$$

In the above relation, C_D denotes the DO concentration in the downstream section, C_U is concentration section in the upstream section, and C_S is the saturated concentration.

The current study focuses on the Aeration Efficiency in stepped cascades. Here E depends on the flow and geometric parameters. The following general expression (Baylar et al., 2006) is considered

$$E_{20} = f(q, h, L, N, \alpha) \tag{25}$$

where E_{20} is the Aeration Efficiency at 20 ℃ and q, h, L, N, and α denote unit discharge (discharge per unit width), step height, total length of the steps (chute length), total number of steps and chute inclination angle, respectively.

A dataset with 126 laboratory tests collected in a previous study (Baylar et al., 2006) is considered here. Figure 2 shows a generic sketch of a laboratory set up to measure Aeration Efficiency. The laboratory experiments were carried out in a rectangular flume with dimensions $0.30m\ wide \times 0.50\ m\ depth \times 5.0\ m\ length$. Three types of step height (0.05 m, 0.10 m, and 0.15 m) and different values of unit discharge, varying between $16.67\ \frac{litres}{second}\ and\ 166.67\ \frac{litres}{second}$ were considered, with nappe, transition or skimming flow conditions. In this study the dataset was divided into a training dataset (70% of the data) and a testing dataset (30%). Table 2 shows the input variable ranges for both training and testing stages, including maximum (Max) and minimum (Min) values as well as average (Mean), standard deviation (STD), and skewness (SKW) values.

[Table 2]

[Figure 2]

2.7 Input Variable Combinations

To predict the Aeration Efficiency, a combination of input variables selected among those introduced earlier (q, h, L, N, and α) must be identified. To establish proper combinations, it is essential to rank the input variables in terms of their relevance to the output variable (E_{20}) by quantifying its degree of dependency on each input variable. This was carried out in this study using Mutual Information theory, a widely used approach for quantifying the flow of information among variables that originated from entropy theory (Singh, 2016; Nourani, Andalib, & Dąbrowska, 2017; Sang, Singh, Hu, Xie, & Li, 2018).

To quantify the flow of information between each parameter and the Aeration Efficiency, the Mutual Information between input parameter X and output Y is calculated as

$$\mathrm{I}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X}|\mathrm{Y}) \tag{26}$$

where $\mathrm{I}(\mathrm{X},\mathrm{Y})$ represents the conditional entropy for Y given X. $\mathrm{H}(\mathrm{X})$ denotes the Shannon entropy of the variable X, which quantifies the amount of inherent uncertainty in the variable X (Nourani et al., 2017) and is determined as follows

$$\mathrm{H}(\mathrm{X}) = -\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{P}(\mathrm{X_i}) \log \mathrm{P}(\mathrm{X_i}) \tag{27}$$

where $\mathrm{P}(\mathrm{X_i})$ represents the probability values associated with the values $\mathrm{X_i}$.

The following relation gives the term $\mathrm{H}(\mathrm{X}|\mathrm{Y})$

$$H(X|Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(X_i, Y_j) \log \mathrm{P}(X_i|Y_j) \tag{28}$$

where $\mathrm{P}(X_i|Y_j), N\ and\ M$ are the conditional probability of Y fitted on X, number of input (X), and number of output (Y) parameters, respectively.

2.8 Uncertainty Analysis

The aleatory uncertainty is due to the inherent randomness in physical phenomena. In contrast, epistemic uncertainty is the uncertainty in modeling the physical processes associated with the "idealization on which models rely." Several types of entropy indicators, one "classic" - Maximum Likelihood (ML) – and the other four of Bayesian type - Jeffrey, Laplace, Schurmann-Grassberger (SG), and Minimax, were used to determine which of the models include the same amount of information of the observed dataset. The Bayesian type indicators employ Dirichlet-based probabilities with prior and posterior stages (Archer et al., 2013; Hutter and Zaffalon, 2002). The mathematical expressions of prior and posterior stages of the Dirichlet based probabilities are expressed as follows

$$\mathrm{Dir}(\alpha) \triangleq \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, a_K) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} \pi_i^{\alpha-1}$$

$$\mathrm{Dir}(\alpha) \triangleq \mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha + n_K) = \Gamma(Ka + N) = \prod_{i=1}^{K} \frac{\pi_i^{n_{i+\alpha-1}}}{\Gamma(a + n_i)} \tag{29}$$

where α, K, π_i, N and n_K denote the Dirichlet concentration coefficient, the number of identified bins in the fitted distribution over parameters, the calculated probability that one of the dataset input parameters X is placed in the i_{th} bin, the number of total dataset input parameters, and the number of input parameters that are saved in the i_{th} bin.

The main difference between the aforementioned Bayesian entropy indicators is associated with the value of the parameter α in the Dirichlet priors. The α value equal 0, 0.5, and 1 for ML, Jeffrey, and Laplace methods, respectively. The following equations are used to compute the value of α for SG and Minimax methods, respectively:

$$\frac{1}{Size\ of\ vector\ of\ generated\ results} \tag{30}$$

$$\sqrt{\frac{Sum\ of\ generated\ results}{Size\ of\ vector\ of\ generated\ results}} \quad (31)$$

3. Results and Discussion

3.1 Input Variable Combinations

Table 3 shows the Mutual Information between the input variables considered and the Aeration Efficiency. The parameters step height (h), and chute length (L) are the most relevant to the Aeration Efficiency, while the chute inclination angle (α) is the least relevant.

[Table 3]

Using the results in Table 3, five input variable combinations for Aeration Efficiency prediction were constructed (Table 4) in such a manner that moving from the C1 combination to the C5 combination; the least relevant parameters are progressively eliminated from the set of inputs.

[Table 4]

3.2 Model Prediction Performance

In this study, the prediction performance of standalone and hybridized SVR models was evaluated using four indices, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Nash-Sutcliffe Efficiency (NSE), and Correlation Coefficient (R). These four indices have been widely employed in the literature for predictions using SC algorithms (Asadollah et al., 2022; Ehteram et al., 2021; Mokhtari et al., 2022; Sharafati et al., 2021). The prediction performance based on these metrics for both training and testing phases is shown in Tables 5 and 6, respectively. Regarding the NSE, its value ranges from $-\infty$ to 1.0 (best value); values below zero indicate unacceptable accuracy, so they have been noted as zero in the tables (Gupta and Kling, 2011).

[Table 5]

[Table 6]

Tables 5 and 6 show the SVR-FA model and the standalone SVR model, respectively, show the highest and lowest prediction performance for both testing and training phases. Also, the C2 input variable combination for SVR-FA and SVR-GOA, and the C1 combination, for SVR-ABC and standalone SVR provide the most accurate Aeration Efficiency predictions. Combinations C4 and C5, characterized by the few parameters, produce the weakest predictions.

A few graphical performance plots were produced to identify the model providing the most accurate predictions. Figure 3 shows a radar chart of the normalized performance indices R, RMSE, MAE, and NSE (the scale of their values was modified to between 0 and 1, with 1 corresponding to the best performing model). For the testing phase, the best-performing model with the closest pattern to the square boundary is the SVR-FA-C1 model, which significantly outperforms the other models. In the testing phase, the SVR-GOA-C2 model performance significantly improves compared to its training phase performance; however, the SVR-FA-C2 model remains the most accurate model. The standalone SVR was a model with the lowest performance in both phases and achieved the normalized value of 0 for all four indices. To make a better comparison in Figure 3, the results of the SVR model were eliminated.

[Figure 3]

Figures 4 and 5 show the performance of the four models, for their best input variable combination, in terms of observed vs. predicted Aeration Efficiency value for training and testing phases, respectively. The Aeration Efficiency values predicted by the hybrid SVR models are very close to the observed ones, while the standalone SVR poorly reproduces the observed data. The value of the coefficient of determination R^2, quantifying the similarity between predicted and observed data, is the highest for the SVR-FA-C2 model $\left(R^2_{Training} = 0.978.\ R^2_{Testing} = 0.899\right)$.

[Figure 4]

[Figure 5]

Finally, Figure 6 shows the Taylor diagram for the training and testing phases, which considers RMSE, R, and normalized standard deviation. In the diagram, the model that better reproduces the observed values is the closest to the point labeled "observed", characterized by RMSE = 0, R = 1, and normalized standard deviation = 1; in this case, all three hybrid SVR models plot very close to the "observed" point (as opposed to the standalone SVR model), with SVR-FA-C2 again being the nearest.

[Figure 6]

3.3 Comparison with Empirical Formulations

As shown above, the proposed SVR-FA-C2 model is an excellent predictive model. It is compared here with two empirical, experiment-derived formulations from the literature. These are the expression by (Baylar et al., 2007c)

$$E_{20} = 1 - exp[-5.730 \times q^{-0.035} \times (\cos\ \alpha)^{12.042} \times (\sin\ \alpha)^{1.594}] \qquad (32)$$

and (Essery et al., 1978)

$$E_{20} = 1 - exp\left(-\frac{H}{\sqrt{gh}}\left(0.427 + 0.31\left(\frac{y_c}{h}\right)\right)\right) \qquad (33)$$

In the above equation, $y_c = (\sqrt[3]{q^2/g})$ is the critical depth, and H denotes the total height of the cascade (product of N and h) and g is the acceleration of gravity.

The predictions by the empirical equations were compared with those provided by the SVR-FA-C2 model for the testing dataset, using the R, RMSE, MAE, and NSE indices. The results in Table 7 show that the SVR-FA-C2 model significantly outperforms the previously derived empirical formulations from the literature.

[Table 7]

3.4 Uncertainty Analysis

The uncertainty associated with the models employed for Aeration Efficiency prediction (standalone SVR, SVR-ABC, SVR-GOA, and SVR-FA) was calculated, for their best input variable combination, using the R software and the Entropy Package. As discussed earlier, different methods - Maximum Likelihood, Jeffrey, Laplace, SG, and Minimax - were used to compute entropy (Table 8 and Figure 7). The results show that the entropy indicators for the predictions provided by the hybrid models are in line with the observed data, especially for the SVR-FA model, unlike the predictions by the standalone SVR model. For instance, in the training stage, the calculated difference between epistemic uncertainties of observed data and generated data from the best predictor (SVR-FA) using the aforementioned entropies is 0.039%, 0.01%, 0.004%, 0.0038%, and 0.029%, respectively. The calculated difference of epistemic uncertainties for the weakest model (standalone SVR) are 0.765%, 0.213%, 0.099%, 0.728% and 0.578%. For the testing stage, the percent differences are 0.0567%, 0.0192%, 0.009%, 0.053% and 0.0398% for the SVR-FA model and 0.868%, 0.243%, 0.113%, 0.793% and 0.580% for the standalone SVR model. This further confirms the generally better performance of the hybrid SVR models than the standalone SVR model.

[Figure7]

[Table 8]

3.5 Comparison with Previous Studies

Regarding the split of 70% / 30% of the set of 126 laboratory tests for the training and testing phases, respectively, this proportion was appropriate for regression and classification purposes (Nguyen et al., 2021; Vrigazova, 2021).

Using the five selected parameters for predicting E20 - unit discharge, the total number of steps, step height, overall length, and chute inclination – is consistent with previous studies (see sections 2.6 and 3.3). The application of Mutual Information theory to identify input variable combinations is also in line with other investigations showing it to be an excellent detector of correlation, especially for datasets characterized by a high level of non-linearity (Baboukani et al., 2021; Laarne et al., 2021) and applying it for prediction of river sinuosity, drought, and landslides (Haghbin et al., 2021; Li et al., 2022; Ma et al., 2022).

The finding of FA as the best optimization algorithm among those considered in this study is similar to that of other scholars showing it to be better performing than other algorithms such as Particle Swarm Optimization (PSO) and Real Coded Genetic Algorithm (RGA) (Dash et al., 2020; Su et al., 2017; Yang, 2009).

The methods presented in this paper could be replicated using other datasets; other techniques for pre-processing and input sensitivity analysis, such as the Gamma test, could be compared with Mutual Information theory; and alternative machine learning approaches based on boosting and bagging algorithms or metaheuristic techniques such as Bat algorithm, or Cuckoo search could be employed to predict aeration efficiency.

4. Conclusion

In this paper, the "classic" standalone Support Vector Regression (SVR) algorithm and its hybrid versions coupled with different evolutionary optimization algorithms (Artificial Bee Colony (ABC), Grasshopper Optimization Algorithm (GOA), and Firefly Algorithm (FA)) were employed to predict Aeration Efficiency (E20) in stepped cascades. Five different parameters, namely unit discharge (q), step height (h), chute slope (angle α), chute length (L), and total number of steps (N), were obtained from laboratory experiments conducted by (Baylar et al., 2006), were considered as inputs of the predictive algorithms. To analyze the level of dependency of the predictive variable (E20) on the different inputs, Mutual Information theory was used, which showed that E20 has the highest degree of dependency on step height and chute length, while the lowest correlation is with the chute inclination angle. Five different input variable combinations were considered based on this "pre-processing" stage. The results revealed that the prediction performance of the SVR-FA-C2 model ($R_{training} = 0.988,\ MAE_{training} = 0.018,\ R_{testing} = 0.947,\ MAE_{testing} = 0.0270$), where C2 means that the input variables used are q, h, L, and N, was significantly better than that of the standalone SVR model ($R_{training} = 0.741$, $MAE_{trainint} = 0.129$, $R_{testing} = 0.738$, $MAE_{testing} = 0.126$). Although based on a single experimental dataset, the findings of this study suggest that hybrid SVR models, in which the SVR parameters are optimized through evolutionary algorithms, significantly outperform the classic SVR model, as well as previously developed empirical equations when predicting E20. To the authors' knowledge, the current study is the first that specifically focuses on the application of several hybrid algorithms for E20 prediction in stepped cascades and shows them to be a viable, time-efficient, and cost-effective alternative for future applications.

5. Declarations

Funding: No funding.

Competing interests: The authors declare that they have no competing interests.

Availability of data and materials: Please contact the corresponding author for data requests.

Code availability: Please contact the corresponding author for code requests.

Ethics approval: Not applicable**.**

Consent to participate: Not applicable**.**

Consent for publication: Not applicable**.**

References

Al-Musaylh, M.S., Deo, R.C., Adamowski, J.F., Li, Y., 2018. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Adv. Eng. Informatics 35, 1–16.

Archer, E., Park, I.M., Pillow, J.W., 2013. Bayesian and quasi-Bayesian estimators for mutual information from discrete data. Entropy 15, 1738–1755.

Asadollah, S.B.H.S., Sharafati, A., Motta, D., Yaseen, Z.M., 2021. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. J. Environ. Chem. Eng.

9, 104599.
Asadollah, S.B.H.S., Sharafati, A., Neshat, A., Hemmati, N., 2022. A robust stochastic approach in correcting the TRMM precipitation product and simulating flood features. Environ. Monit. Assess. 194, 1–15.
Baboukani, P.S., Graversen, C., Østergaard, J., 2021. Estimation of Directed Dependencies in Time Series Using Conditional Mutual Information and Non-linear Prediction, in: 2020 28th European Signal Processing Conference (EUSIPCO). IEEE, pp. 2388–2392.
Baylar, A., Bagatur, T., Emiroglu, E., 2007a. Aeration efficiency with nappe flow over stepped cascades. Proc. Inst. Civ. Eng. Water Manag. https://doi.org/10.1680/wama.2007.160.1.43
Baylar, A., Bagatur, T., Emiroglu, M.E., 2007b. Prediction of oxygen content of nappe, transition, and skimming flow regimes in stepped-channel chutes. J. Environ. Eng. Sci. https://doi.org/10.1139/S06-048
Baylar, Ahmet, Emiroglu, M.E., 2003. Study of aeration efficiency at stepped channels. Proc. Inst. Civ. Eng. Water Marit. Eng. https://doi.org/10.1680/wame.2003.156.3.257
Baylar, A, Emiroglu, M.E., 2003. Study of aeration efficiency at stepped channels, in: Proceedings of the Institution of Civil Engineers-Water and Maritime Engineering. Thomas Telford Ltd, pp. 257–263.
Baylar, A., Emiroglu, M.E., Bagatur, T., 2009. Influence of Chute slope on oxygen content in stepped waterways. Gazi Univ. J. Sci.
Baylar, A., Emiroglu, M.E., Bagatur, T., 2006. An experimental investigation of aeration performance in stepped spillways. Water Environ. J. 20, 35–42.
Baylar, A., Hanbay, D., Ozpolat, E., 2007c. Modeling aeration efficiency of stepped cascades by using ANFIS. Clean - Soil, Air, Water. https://doi.org/10.1002/clen.200700019
Baylar, A., Unsal, M., Ozkan, F., 2010. Hydraulic structures in water aeration processes. Water. Air. Soil Pollut. https://doi.org/10.1007/s11270-009-0226-2
Chanson, H., 2000. Forum article. Hydraulics of stepped spillways: current status. J. Hydraul. Eng. 126, 636–637.

Chen, K.-Y., 2007. Forecasting systems reliability based on support vector regression with genetic algorithms. Reliab. Eng. Syst. Saf. 92, 423–432.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Dash, J., Dam, B., Swain, R., 2020. Improved firefly algorithm based optimal design of special signal blocking IIR filters. Meas. J. Int. Meas. Confed. https://doi.org/10.1016/j.measurement.2019.106986

Ehteram, M., Sharafati, A., Asadollah, S.B.H.S., Neshat, A., 2021. Estimating the transient storage parameters for pollution modeling in small streams: a comparison of newly developed hybrid optimization algorithms. Environ. Monit. Assess. 193, 1–16.

Emiroglu, M.E., Baylar, A., 2003. An investigation of effect of stepped chutes with end sill on aeration performance. Water Qual. Res. J. 38, 527–539.

Essery, I.T.S., Tebbutt, T.H.Y., Rasaratnam, S.K., 1978. Design of spillways for re-aeration of polluted waters. Construction Industry Research and Information Association.

Ewees, A.A., Abd Elaziz, M., Alameer, Z., Ye, H., Jianhua, Z., 2020. Improving multilayer perceptron neural network using chaotic grasshopper optimization algorithm to forecast iron ore price volatility. Resour. Policy 65, 101555.

Felder, S., Chanson, H., 2009. Energy dissipation, flow resistance and gas-liquid interfacial area in skimming flows on moderate-slope stepped spillways. Environ. Fluid Mech. https://doi.org/10.1007/s10652-009-9130-y

Gulliver, J.S., Wilhelms, S.C., Parkhill, K.L., 1998. Predictive capabilities in oxygen transfer at hydraulic structures. J. Hydraul. Eng. 124, 664–671.

Gupta, H.V., Kling, H., 2011. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. Water Resour. Res. 47.

Haghbin, M., Sharafati, A., Aghamajidi, R., Asadollah, S.B.H.S., Noghani, M.H.M., Jalón, M.L., 2022. Determination of discharge coefficient of stepped morning glory spillway using a hybrid data-driven method. Flow Meas. Instrum. 85, 102161.

Haghbin, M., Sharafati, A., Motta, D., 2021. Prediction of channel sinuosity in perennial rivers using

Bayesian Mutual Information theory and support vector regression coupled with meta-heuristic algorithms. Earth Sci. Informatics 14, 2279–2292.

Hanbay, D., Baylar, A., Batan, M., 2009a. Prediction of aeration efficiency on stepped cascades by using least square support vector machines. Expert Syst. Appl. https://doi.org/10.1016/j.eswa.2008.03.003

Hanbay, D., Baylar, A., Ozpolat, E., 2009b. Predicting flow conditions over stepped chutes based on ANFIS. Soft Comput. https://doi.org/10.1007/s00500-008-0343-7

Hutter, M., Zaffalon, M., 2002. Distribution of mutual information for robust feature selection.

İskenderoğlu, F.C., Baltacioğlu, M.K., Demir, M.H., Baldinelli, A., Barelli, L., Bidini, G., 2020. Comparison of support vector regression and random forest algorithms for estimating the SOFC output voltage by considering hydrogen flow rates. Int. J. Hydrogen Energy 45, 35023–35038.

Izadi, Parnian, Izadi, Parin, Eldyasti, A., 2021. Understanding microbial shift of Enhanced Biological Phosphorus Removal process (EBPR) under different Dissolved Oxygen (DO) concentrations and Hydraulic Retention Time (HRTs). Biochem. Eng. J. https://doi.org/10.1016/j.bej.2020.107833

Jiang, L., Diao, M., Xue, H., Sun, H., 2018. Energy Dissipation Prediction for Stepped Spillway Based on Genetic Algorithm–Support Vector Regression. J. Irrig. Drain. Eng. https://doi.org/10.1061/(asce)ir.1943-4774.0001293

Johari, N.F., Zain, A.M., Noorfa, M.H., Udin, A., 2013. Firefly algorithm for optimization problem. Appl. Mech. Mater. 421, 512–517.

Khdhiri, H., Potier, O., Leclerc, J.P., 2014. Aeration efficiency over stepped cascades: Better predictions from flow regimes. Water Res. https://doi.org/10.1016/j.watres.2014.02.022

Laarne, P., Zaidan, M.A., Nieminen, T., 2021. ennemi: Non-linear correlation detection with mutual information. SoftwareX 14, 100686.

Li, Q., Han, X., Liu, Z., He, P., Shi, P., Chen, Q., Du, F., 2022. A novel information changing rate and conditional mutual information-based input feature selection method for artificial intelligence drought prediction models. Clim. Dyn. 1–21.

Lima, M.A.F.B., Fernández Ramírez, L.M., Carvalho, P., Batista, J.G., Freitas, D.M., 2022. A comparison between deep learning and support vector regression techniques applied to solar forecast in Spain. J. Sol. Energy Eng. 144.

Ma, J., Wang, Y., Niu, X., Jiang, S., Liu, Z., 2022. A comparative study of mutual information-based input variable selection strategies for the displacement prediction of seepage-driven landslides using optimized support vector regression. Stoch. Environ. Res. Risk Assess. 1–21.

Malik, A., Tikhamarine, Y., Souag-Gamane, D., Rai, P., Sammen, S.S., Kisi, O., 2021. Support vector regression integrated with novel meta-heuristic algorithms for meteorological drought prediction. Meteorol. Atmos. Phys. 133, 891–909.

Meraihi, Y., Gabis, A.B., Mirjalili, S., Ramdane-Cherif, A., 2021. Grasshopper optimization algorithm: theory, variants, and applications. IEEE Access 9, 50001–50024.

Mirarabi, A., Nassery, H.R., Nakhaei, M., Adamowski, J., Akbarzadeh, A.H., Alijani, F., 2019. Evaluation of data-driven models (SVR and ANN) for groundwater-level prediction in confined and unconfined systems. Environ. Earth Sci. 78, 1–15.

Mokhtari, S., Sharafati, A., Raziei, T., 2022. Satellite-based streamflow simulation using CHIRPS satellite precipitation product in Shah Bahram Basin, Iran. Acta Geophys. 1–14.

Nguyen, Q.H., Ly, H.-B., Ho, L.S., Al-Ansari, N., Le, H. Van, Tran, V.Q., Prakash, I., Pham, B.T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Math. Probl. Eng. 2021.

Nourani, V., Andalib, G., Dąbrowska, D., 2017. Conjunction of wavelet transform and SOM-mutual information data pre-processing approach for AI-based Multi-Station nitrate modeling of watersheds. J. Hydrol. 548, 170–183.

Panahi, M., Sadhasivam, N., Pourghasemi, H.R., Rezaie, F., Lee, S., 2020. Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). J. Hydrol. 588, 125033.

Peyras, L., Royet, P., Degoutte, G., 1992. Flow and Energy Dissipation over Stepped Gabion Weirs. J. Hydraul. Eng. https://doi.org/10.1061/(asce)0733-9429(1992)118:5(707)

Qin, P., Hu, H., Yang, Z., 2021. The improved grasshopper optimization algorithm and its applications. Sci. Rep. 11, 1–14.

Rastogi, R., Sharma, S., 2021. Ternary tree-based structural twin support tensor machine for clustering. Pattern Anal. Appl. https://doi.org/10.1007/s10044-020-00902-8

Saha, A., Pal, S.C., Arabameri, A., Blaschke, T., Panahi, S., Chowdhuri, I., Chakrabortty, R., Costache, R., Arora, A., 2021. Flood susceptibility assessment using novel ensemble of hyperpipes and support vector regression algorithms. Water 13, 241.

Salmasi, F., Sattari, M.T., Nurcheshmeh, M., 2021. Genetic Programming Approach for Estimating Energy Dissipation of Flow over Cascade Spillways. Iran. J. Sci. Technol. - Trans. Civ. Eng. https://doi.org/10.1007/s40996-020-00541-3

Sammen, S.S., Ghorbani, M.A., Malik, A., Tikhamarine, Y., AmirRahmani, M., Al-Ansari, N., Chau, K.-W., 2020. Enhanced artificial neural network with Harris hawks optimization for predicting scour depth downstream of ski-jump spillway. Appl. Sci. 10, 5160.

Sangeeta, Haji Seyed Asadollah, S.B., Sharafati, A., Sihag, P., Al-Ansari, N., Chau, K.-W., 2021. Machine learning model development for predicting aeration efficiency through Parshall flume. Eng. Appl. Comput. Fluid Mech. 15, 889–901.

Saremi, S., Mirjalili, S., Lewis, A., 2017. Grasshopper Optimisation Algorithm: Theory and application. Adv. Eng. Softw. 105, 30–47. https://doi.org/https://doi.org/10.1016/j.advengsoft.2017.01.004

Sengun, E., Alam, B., Shabani, R., Yaman, I.O., 2021. Strength and fracture properties of roller compacted concrete (RCC) prepared by an in-situ compaction procedure. Constr. Build. Mater. https://doi.org/10.1016/j.conbuildmat.2020.121563

Sharafati, A., Asadollah, S.B.H.S., Shahbazi, A., 2021. Assessing the impact of climate change on urban water demand and related uncertainties: a case study of Neyshabur, Iran. Theor. Appl. Climatol. 145, 473–487.

Sharafati, A., Haghbin, M., Haji Seyed Asadollah, S.B., Tiwari, N.K., Al-Ansari, N., Yaseen, Z.M., 2020. Scouring Depth Assessment Downstream of Weirs Using Hybrid Intelligence Models. Appl. Sci. 10, 3714.

Sihag, P., Dursun, O.F., Sammen, S.S., Malik, A., Chauhan, A., 2021. Prediction of aeration efficiency of parshall and modified venturi flumes: application of soft computing versus regression models. Water Supply 21, 4068–4085.

Su, H., Cai, Y., Du, Q., 2017. Firefly-Algorithm-Inspired Framework with Band Selection and Extreme Learning Machine for Hyperspectral Image Classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. https://doi.org/10.1109/JSTARS.2016.2591004

Su, H., Li, X., Yang, B., Wen, Z., 2018. Wavelet support vector machine-based prediction model of dam deformation. Mech. Syst. Signal Process. https://doi.org/10.1016/j.ymssp.2018.03.022

Tebbutt, T.H.Y., 1972. Some studies on reaeration in cascades. Water Res. https://doi.org/10.1016/0043-1354(72)90007-3

Teodorovic, D., Lucic, P., Markovic, G., Dell'Orco, M., 2006. Bee colony optimization: principles and applications, in: 2006 8th Seminar on Neural Network Applications in Electrical Engineering. IEEE, pp. 151–156.

Toombes, L., Chanson, H., 2020. Air-water flow and gas transfer at aeration cascades: A comparative study of smooth and stepped chutes, in: Hydraulics of Stepped Spillways. https://doi.org/10.1201/9781003078609-13

Toombes, L., Chanson, H., 2008. Flow patterns in nappe flow regime down low gradient stepped chutes. J. Hydraul. Res. https://doi.org/10.1080/00221686.2008.9521838

Toombes, L., Chanson, H., 2005. Air–Water Mass Transfer on a Stepped Waterway. J. Environ. Eng. https://doi.org/10.1061/(asce)0733-9372(2005)131:10(1377)

Vapnik, V., Golowich, S.E., Smola, A.J., 1997. Support vector method for function approximation, regression estimation and signal processing, in: Advances in Neural Information Processing Systems. pp. 281–287.

Vrigazova, B., 2021. The proportion for splitting data into training and test set for the bootstrap in classification problems. Bus. Syst. Res. Int. J. Soc. Adv. Innov. Res. Econ. 12, 228–242.

Wang, J., Li, L., Niu, D., Tan, Z., 2012. An annual load forecasting model based on support vector regression with differential evolution algorithm. Appl. Energy. https://doi.org/10.1016/j.apenergy.2012.01.010

Wu, J., Wang, Y.G., Burrage, K., Tian, Y.C., Lawson, B., Ding, Z., 2020. An improved firefly algorithm for global continuous optimization problems. Expert Syst. Appl. https://doi.org/10.1016/j.eswa.2020.113340

Yang, X., 2010. Nature-Inspired Metaheuristic Algorithms, Nature-Inspired Metaheuristic Algorithms Second Edition.

Yang, X.S., 2010. Firefly algorithm, stochastic test functions and design optimization. Int. J. Bio-Inspired Comput. https://doi.org/10.1504/IJBIC.2010.032124

Yang, X.S., 2009. Firefly algorithms for multimodal optimization, in: Stochastic Algorithms: Foundations and Applicaions. Lect. Notes Comput. Sci.

Zhou, X., Lu, J., Huang, J., Zhong, M., Wang, M., 2021. Enhancing artificial bee colony algorithm with multi-elite guidance. Inf. Sci. (Ny). https://doi.org/10.1016/j.ins.2020.07.037

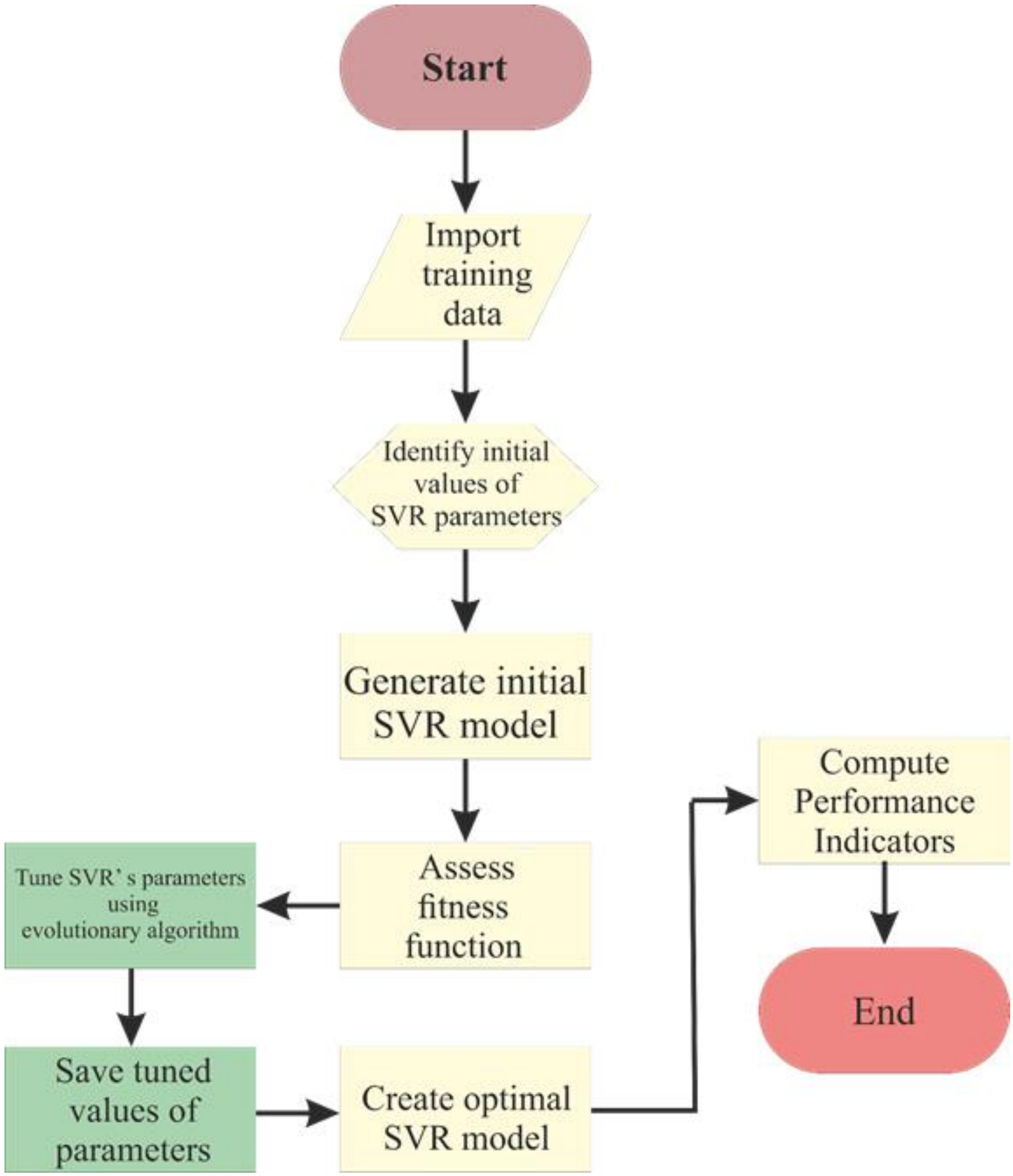

Figure 1: Flowchart of the SVR model hybridization.

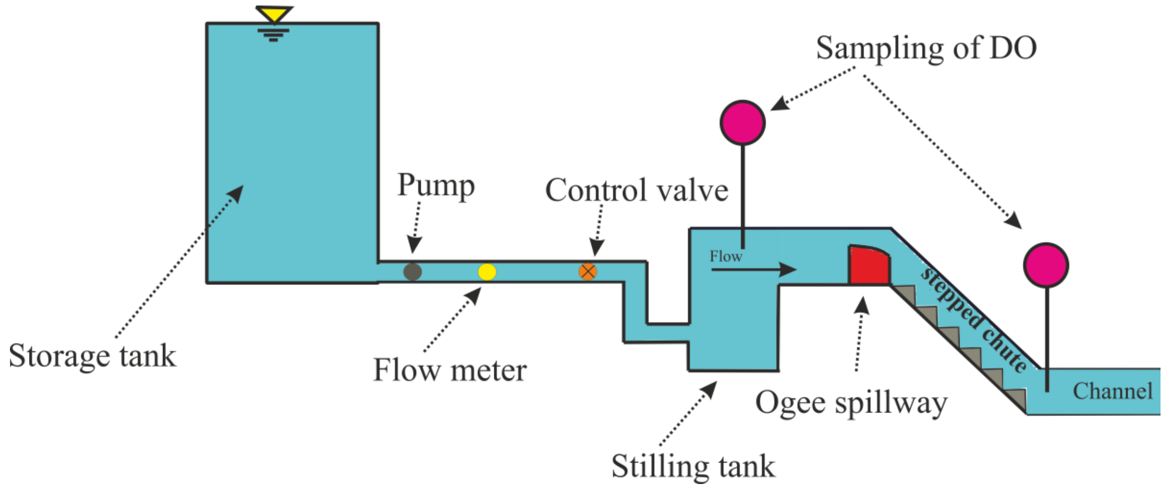

Figure 2: Generic sketch of a laboratory experimental set up to measure Aeration Efficiency in a stepped cascade.

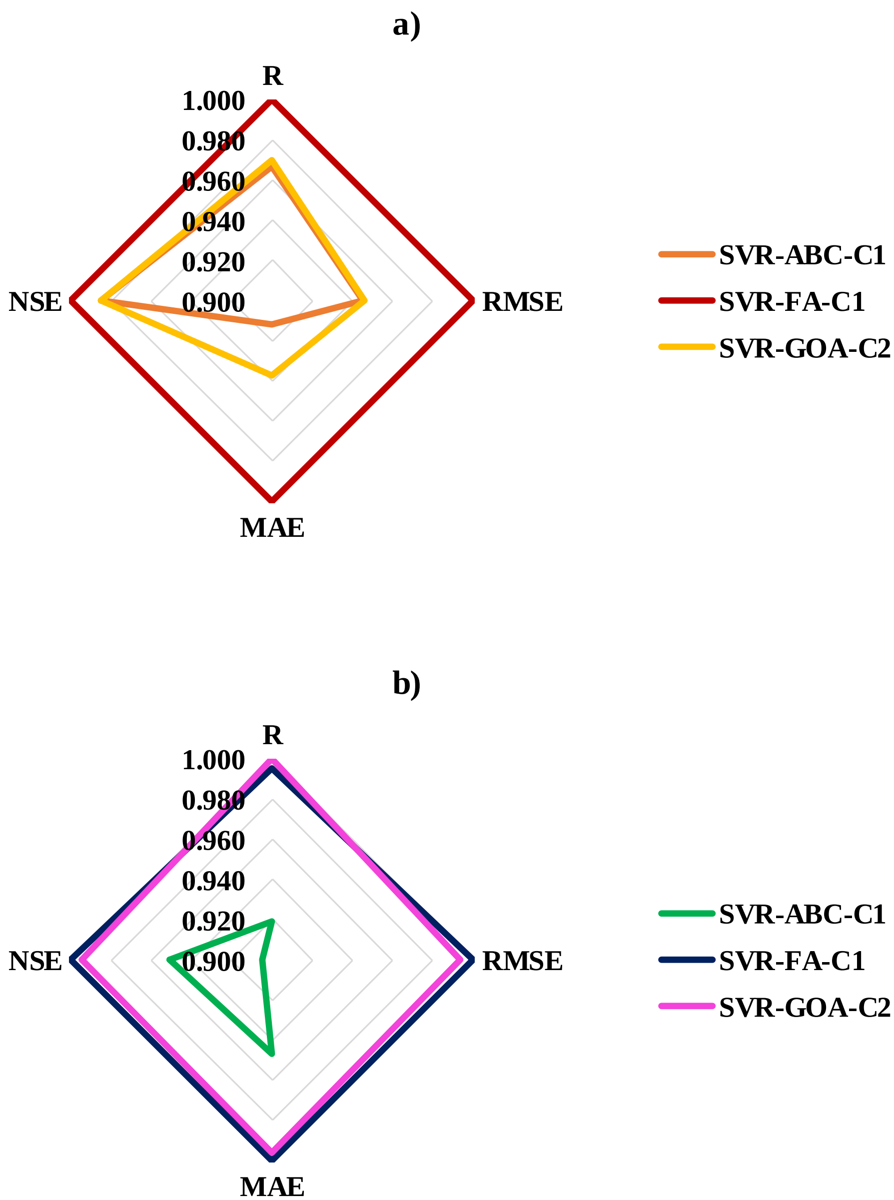

Figure 3: Radar plot of Aeration Efficiency prediction performance indices for (a) training phase and (b) testing phase.

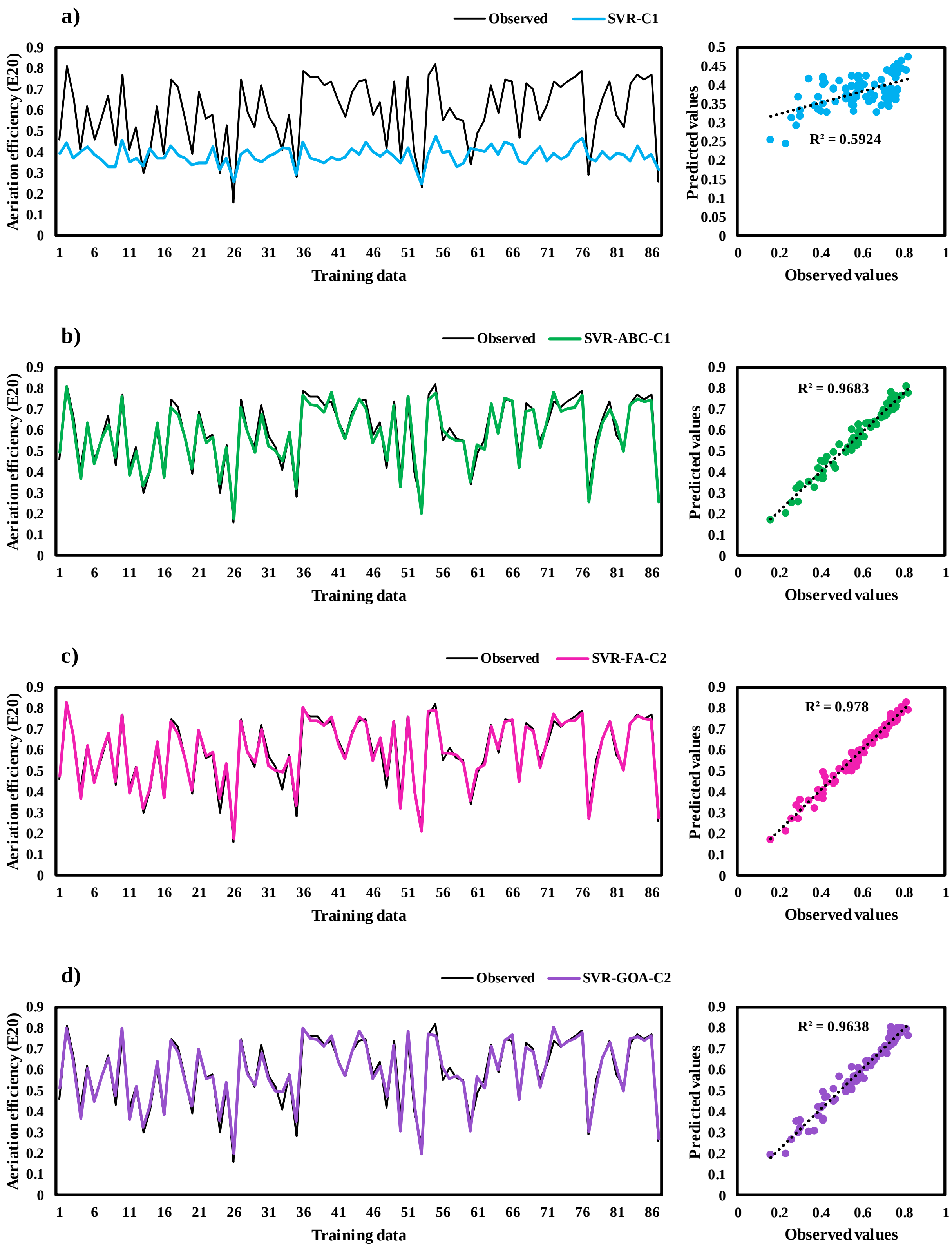

Figure 4: Predicted vs observed Aeration Efficiency for training phase.

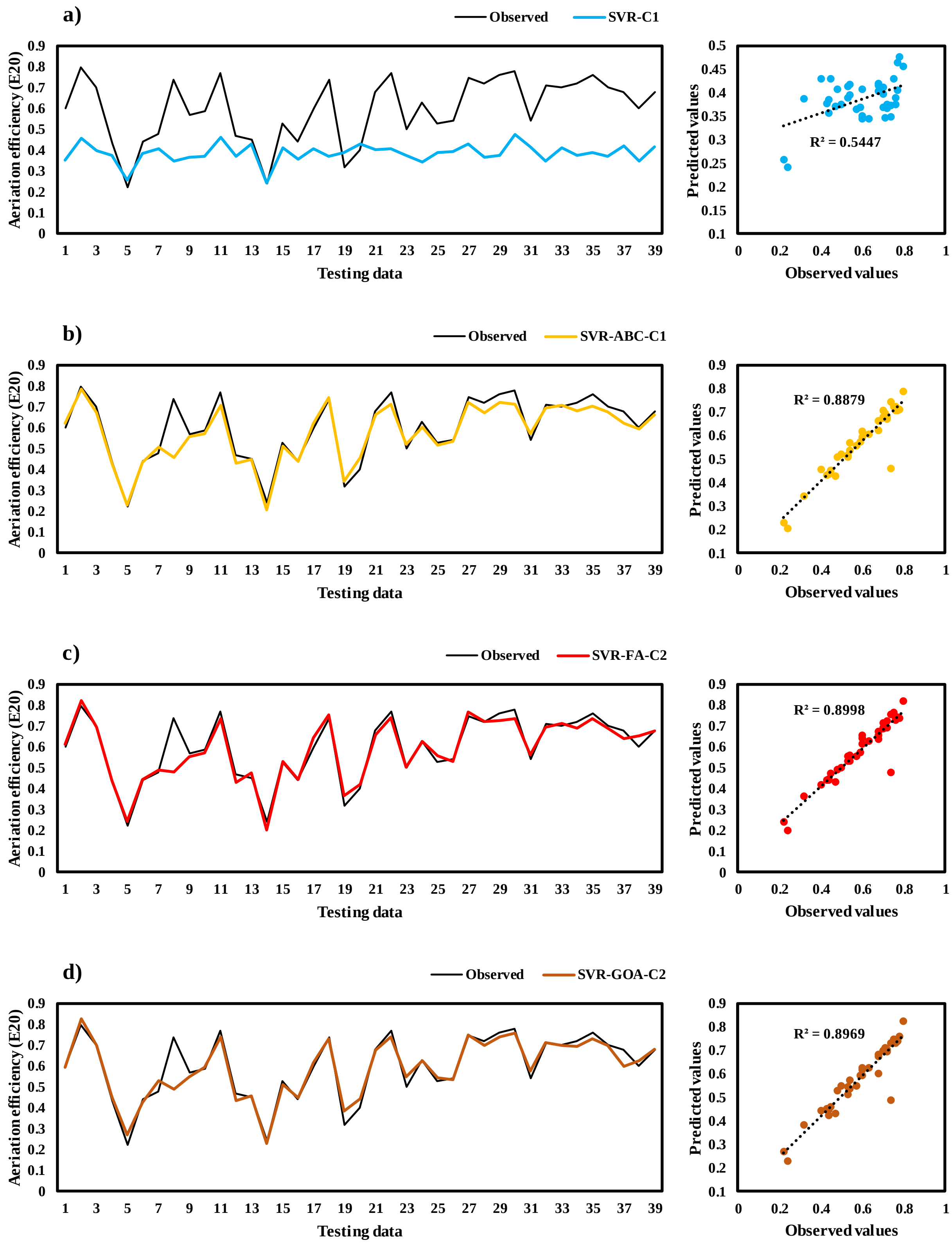

Figure 5: Predicted vs observed Aeration Efficiency for testing phase.

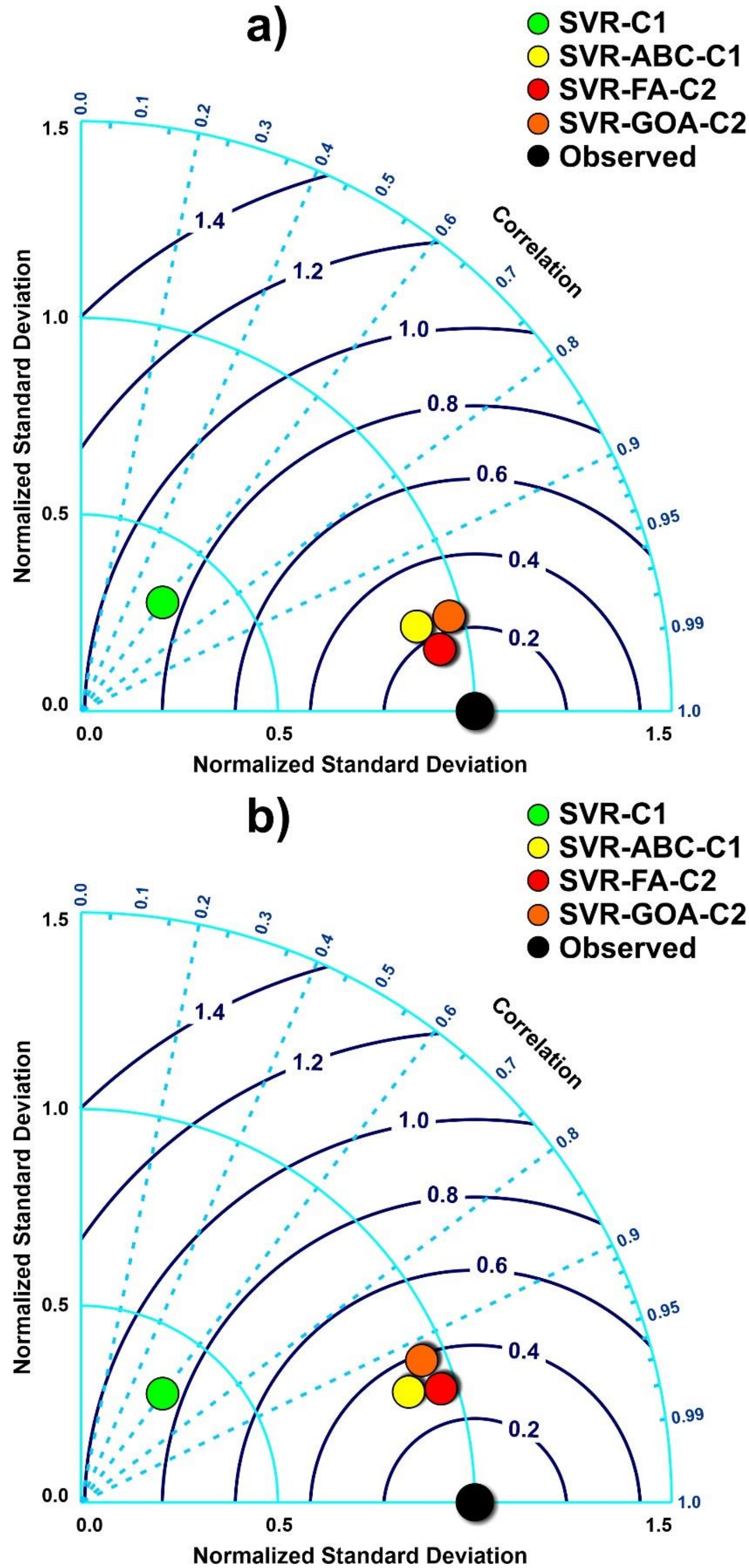

Figure 6: Taylor diagram of Aeration Efficiency prediction performance indices for (a) training phase and (b) testing phase.

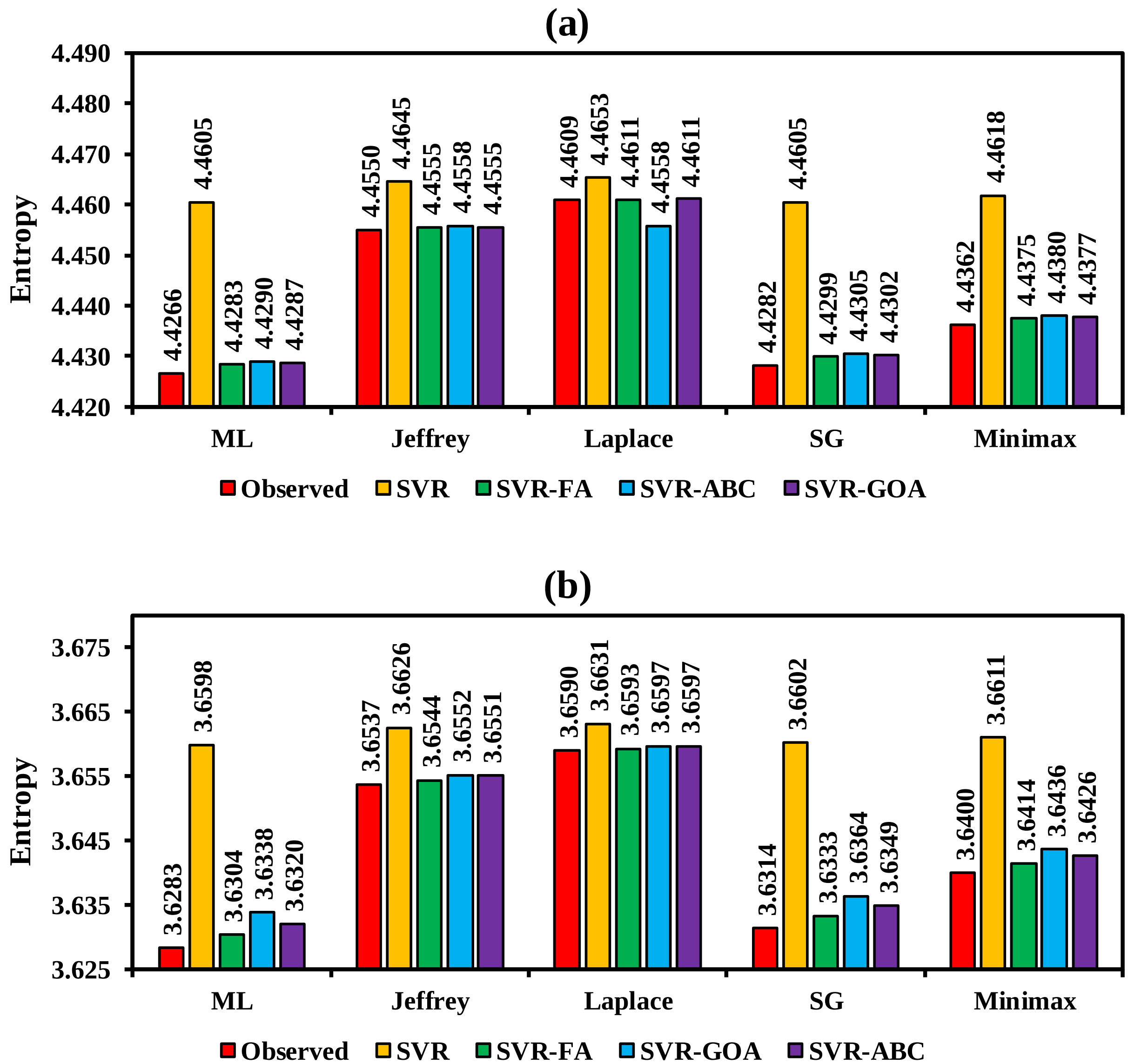

Figure 7: Entropy indicators for predicted and observed Aeration Efficiency for (a) training phase and (b) testing phase.

Table 1: Optimization parameters.

Optimizer: Firefly Algorithm	
Parameter	**Value**
Maximum Iteration Number	500
Number of Populations	50
Light Absorption Coefficient	0.01
Attraction Coefficient Base Value	2
Mutation	0.35
Mutation Coefficient Damping Ratio	0.59
Optimizer: Artificial Bee Colony	
Maximum Iteration Number	500
Number of Populations	50
Number of Onlooker Bees	50
Acceleration Coefficient	0.02
Abandonment Limit Parameter	10
Optimizer: Grasshopper Optimization Algorithm	
Maximum Iteration Number	500
Number of Populations	50
Maximum Constriction Coefficient	1
Minimum Constriction Coefficient	4E-5
Weight	0.6

Table 2: Input variable ranges for training and testing phases.

Phase	Statistical variables	q (m^2/s)	h (m)	α (°)	L (m)	N
Training	Max	166.67	0.15	50.00	5.00	50.00
	Min	16.67	0.05	14.48	3.26	8.00
	Mean	80.27	0.10	30.08	3.98	23.40
	STD	52.37	0.04	12.84	0.71	14.94
	SKW	0.47	0.02	0.40	0.51	0.91
Testing	MAX	166.67	0.15	50.00	5.00	50.00
	MIN	16.67	0.05	14.48	3.26	8.00
	Mean	82.48	0.10	27.50	4.21	21.03
	STD	48.21	0.04	11.47	0.74	10.77
	SKW	0.28	-0.05	0.64	-0.05	1.25

Table 3: Mutual Information or degree of dependency between input variables and output.

Input Variable	Mutual Information or degree of dependency between the input variable and output E_{20}
q	0.003
h	0.005
$\boldsymbol{\alpha}$	0.001
L	0.0049
N	0.0046

Table 4: Input variable combinations for Aeration Efficiency prediction.

	q	**h**	$\boldsymbol{\alpha}$	**L**	**N**
C1	✓	✓	✓	✓	✓
C2	✓	✓	-	✓	✓
C3	-	✓	-	✓	✓
C4	-	✓	-	✓	-
C5	-	✓	-	-	-

Table 5: Results of standalone SVR and SVR using FA, GOA, and ABC optimization algorithms for the training phase.

Predictive Model	R	NSE	RMSE	MAE
SVR-C1	0.749	0.000	0.147	0.129
SVR-C2	**0.741**	0.000	**0.147**	**0.129**
SVR-C3	0.111	0.000	0.141	0.124
SVR-C4	0.074	0.000	0.158	0.127
SVR-C5	0.014	0.000	0.158	0.131
SVR-ABC-C1	**0.981**	**0.960**	**0.030**	**0.026**
SVR-ABC-C2	0.807	0.011	0.093	0.084
SVR-ABC-C3	0.797	0.227	0.095	0.084
SVR-ABC-C4	0.007	0.000	0.158	0.131
SVR-ABC-C5	0.054	0.000	0.158	0.131
SVR-FA-C1	0.989	0.976	0.023	0.016
SVR-FA-C2	**0.988**	**0.977**	**0.024**	**0.018**
SVR-FA-C3	0.083	0.389	0.089	0.073
SVR-FA-C4	0.045	0.000	0.025	0.158
SVR-FA-C5	0.062	0.000	0.158	0.134
SVR-GOA-C1	0.979	0.958	0.032	0.028
SVR-GOA-C2	**0.982**	**0.962**	**0.030**	**0.023**
SVR-GOA-C3	0.832	0.537	0.087	0.071
SVR-GOA-C4	0.120	0.000	0.159	0.131
SVR-GOA-C5	0.062	0.000	0.158	0.134

Table 6: Results of standalone SVR and SVR using FA, GOA and ABC optimization algorithms for the testing phase.

Predictive Model	R	NSE	RMSE	MAE
SVR-C1	0.738	0.000	0.144	0.126
SVR-C2	0.738	0.000	0.144	0.126
SVR-C3	0.630	0.000	0.140	0.121
SVR-C4	0.446	0.000	0.144	0.130
SVR-C5	0.380	0.000	0.145	0.128
SVR-ABC-C1	0.931	0.844	0.058	0.032
SVR-ABC-C2	0.711	0.000	0.106	0.095
SVR-ABC-C3	0.712	0.241	0.106	0.830
SVR-ABC-C4	0.037	0.000	0.151	0.126
SVR-ABC-C5	0.091	0.000	0.152	0.128
SVR-FA-C1	0.942	0.877	0.050	0.026
SVR-FA-C2	**0.947**	**0.888**	**0.048**	**0.027**
SVR-FA-C3	0.753	0.358	0.099	0.075
SVR-FA-C4	0.082	0.000	0.023	0.151
SVR-FA-C5	0.226	0.000	0.147	0.127
SVR-GOA-C1	0.930	0.856	0.055	0.032
SVR-GOA-C2	0.948	0.874	0.049	0.027
SVR-GOA-C3	0.728	0.470	0.103	0.077
SVR-GOA-C4	0.091	0.000	0.152	0.126
SVR-GOA-C5	0.226	0.000	0.147	0.127

Table 7: Prediction performance comparison between the current study best predictive model and previously developed empirical formulations.

	R	**RMSE**	**NSE**	**MAE**
SVR-FA-C2	0.947	0.048	0.888	0.027
Baylar (2003)	0.515	0.466	0.348	0.406
Essers et al. (1978)	0.181	0.347	0.079	0.300

Table 8: Entropy indicators for predicted and observed Aeration Efficiency for training and testing phases.

Training Phase

Model	Classic Entropy Method	Bayesian Entropy Methods			
	ML	Jeffrey	Laplace	SG	Minimax
SVR-FA	4.428344	4.455457	4.461063	4.429865	4.437463
SVR-ABC	4.428952	4.455801	4.455801	4.430475	4.437999
SVR-GOA	4.428703	4.45553	4.461094	4.430205	4.437719
SVR	4.460481	4.464497	4.465272	4.460481	4.461814
Observed	4.426606	4.454999	4.460854	4.428203	4.436174

Testing Phase

Model	Classic Entropy Method	Bayesian Entropy Methods			
	ML	Jeffrey	Laplace	SG	Minimax
SVR-FA	3.630378	3.654396	3.659322	3.633287	3.641409
SVR-GOA	3.633822	3.655238	3.659697	3.636387	3.643607
SVR-ABC	3.632031	3.655062	3.659664	3.63486	3.642625
SVR	3.659844	3.662582	3.663118	3.660181	3.661076
Observed	3.628319	3.653693	3.658976	3.631363	3.639957