Gorgan University of Agricultural Sciences and Natural Resources

Environmemtal Resources Research (ERR)

Print ISSN: 2783-4832
Online ISSN: 2783-4670

Concentration prediction of dissolved oxygen using meta-heuristic models

Reza Dehghani[1*], Taher FarhadiNejad[2], Iraj Veyskarami[3], Reza Chaman Pira[4]

[1] PhD in Water Sciences and Engineering, Department of Soil Conservation and Watershed Management, Lorestan Province Agriculture and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization, Khorramabad, Iran

[2] Research Assistant Professor, Department of Soil Protection and Watershed Management, Lorestan Province Agriculture and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization, Khorramabad, Iran

[3] Research Assistant Professor, Department of Soil Protection and Watershed Management, Lorestan Province Agriculture and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization, Khorramabad, Iran

[4] Research Assistant Professor, Department of Soil Protection and Watershed Management, Lorestan Province Agriculture and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization, Khorramabad, Iran

Article Info

Article type:
Research Article

Article history:
Received: January 2024
Accepted: March 2024

Corresponding author:
r.kh72777@gmial.com

Keywords:
Dissolved oxygen
Hybrid model
Cumberland river

Abstract

Water is one of the most essential elements in nature that forms the basis of human life and contributes to the economic growth and development of societies. Safe water is closely related to environmental health and activities. The lives of all the animals on our planet depend on water and oxygen. Moreover, sufficient Dissolved Oxygen (DO) is crucial for the survival of aquatic animals. In the present research, temperature (T) and flow (Q) variables were used to predict DO. We used monthly time series and data were related to the Cumberland River in the southern United States from 2012 to 2022. Support Vector Regression (SVR) was employed for prediction of the model in both standalone and hybrid forms. The employed hybrid models consisted of SVR combined with metaheuristic algorithms of Chicken Swarm Optimization (CSO), Social Ski-Driver (SSD) optimization, and the Algorithm of the Innovative Gunner (AIG). Pearson Correlation Coefficient (PCC) was utilized to select the best input combination. Box plots and Taylor diagrams were employed in the interpretation of the results. We observed that all the four hybrid models achieved good results. Also, according to the evaluation criteria, among the models used, SVR-AIG performed better with the coefficient of determination (R2 = 0.963), the root mean square error (RMSE =0.644 mg/l), the mean absolute value of error (MAE = 0.568 mg/l), the Nash-Sutcliffe coefficient (NS = 0.864), and bias percentage (BIAS = 0.001).

Cite this article: Dehghani, Reza; FarhadiNejad, Taher; Veyskarami, Iraj; Chaman Pira, Reza. 2024. Concentration Prediction of Dissolved Oxygen Using meta-heuristic Models. *Environmental Resources Research*, 12(1), 31-46.

DOI: 10.22069/IJERR.2024.22119.1423
Publisher: Gorgan University of Agricultural Sciences and Natural Resources

Introduction

The development of agricultural and industrial activities and the increase in the urban wastewater volume have polluted rivers so that the quality of these vital resources is seriously endangered. Also, we sometimes consume water that is contaminated to a certain extent (Krishna et al. 2020; Forstinus et al. 2016; Ighalo et al. 2020; Khalil et al. 2019; Dizaji et al. 2020; Kisi and Ay, 2012). Dissolved oxygen (DO) is one of the most important qualitative indicators for river health assessment (Dogan et al. 2009). DO is the amount of dissolved oxygen in the water as an important and effective factor in the life of aquatic organisms and indicates water pollution (Chapman.1992). High levels of DO also cause unfavorable living conditions for riverine plants and animals (Radwan et al. 2003). Today, small mobile devices equipped with membrane electrodes are used to measure dissolved oxygen at the sampling site. The membrane electrode is made of a membrane-based on the penetration rate of oxygen molecules. This physical method is simple and fast. On the other hand, the most accurate method of measuring dissolved oxygen is the iodometric method. This method is a titration method based on the oxidizing properties of dissolved oxygen (Ahmed and Shah 2017; Yaseen et al. 2018; Diaz and Rosenberg.2008; Salcedo-Sanz et al. 2016; Afan et al. 2015). Because river water quality is affected by various characteristics that have complex and nonlinear behavior, mathematical models may not perform well. Recently, hybrid models that are a subset of artificial intelligence (AI) are used to estimate river water quality. These AI techniques are simple, powerful, and can easily control complex nonlinear processes. Since these models are non-parametric, their main advantage is the lack of need for the concept of prediction and the relationship between input variables and output data (Gocić et al. 2015). A classic feature of artificial intelligence is that these models are capable of stochastic analysis of dynamics, patterns, and features in input variables used to simulate groundwater variables. Therefore, they are more feasible than other conceptual and statistical methods (such as experimental approaches and physics-based models). In general, AI-based models can be used for local applications. Therefore, models based on artificial intelligence have great potential for various applications, including hydrological and hydrogeological phenomena. Many researchers have confirmed the potential usefulness of AI techniques for simulating river water quality (Ross and Stock 2019; Shi et al. 2019; Li et al. 2020a, b; Adhaileh and Alsaade 2021; Asadollah et al. 2021; Ahmed and Lin 2021; Guo et al. 2021; Zhu et al. 2021; Tiyasha et al. 2021; Huang et al. 2021; Liu et al. 2021).

Alizadeh and Kavianpour (2015) used the combined wavelet artificial neural network model to predict qualitative parameters in Hilo Bay, Pacific and concluded that the combined wavelet and artificial neural network model performs better than the artificial neural network alone. Rajaee et al. (2020) combined ANN–ARIMA, GA–ANN, WANN, WNF, WSVR, and WLGP models to estimate the dissolved oxygen parameter in river water. In their study, Rajaee et al. (2020) gathered information, and statistics from 51 scientific articles during 2016–2000. The results showed that the models based on wavelet transform (WT) demonstrated better performance than the other models under study. In addition, the WDVR model had more accuracy than other models.

Zhu et al. (2021) used WT–ANN, WT–SVM, WT–MLR, and WT–RF hybrid models to predict the dissolved biological oxygen in the water of China's Dongjiang River. The results showed that hybrid models performed better than wavelet models. These models covered single and hybrid patterns due to the increase of model memory that improved their performance.

In general, according to the above research, the reduction in Cumberland River water quality—primarily due to the presence of regenerative chemicals, especially organic matter, and wastewater discharge—poses a significant concern, as the river is the primary water source for various regions, including Tennessee and

adjacent areas. In these areas, industrial and domestic effluents have caused many problems. Therefore, the need to model water-soluble oxygen in this river is very important to improve its quality. In this study, AIG–SVM, SKI–SVM, and CSO–SVM models were used to estimate dissolved oxygen in the Cumberland River, Tennessee, based on measured variables at the Cumberland station, such as dissolved oxygen, flow rate, and temperature. This estimation was done based on a monthly time scale.

Material and Method

Case study: Cumberland catchment

Cumberland is one of the most important rivers in the southern United States with a length of 1,106 km. The river is in the catchment region of southern Kentucky in an area of 4700 square kilometers. The Cumberland River originates in Lecher County in eastern Kentucky on the Cumberland Plain and flows into the Ohio River in Smithland. The river is located between latitudes ″57 ′51 °82 ″22 ′36 °84 East and latitudes ″50 ′7 °37 ″28 ′52 °38 N in Tennessee. Figure 1 shows the position of the study station at longitude ″56 ′49 °86 and latitude ″59 ′10 °36. In this study, for the estimations, the data on water-soluble oxygen (mg/l), precipitation (mm), flow rate (cubic meters per second), and temperature (degrees Celsius) collected by the Cumberland station during 2006-2016 were obtained on a monthly basis from the Geological Survey of the United States.

Also, for modeling, the parameters of monthly flow rate (Q), temperature (T), and Dissolved Oxygen (DO) in river water, also available in the US Geological Survey, for the period of 2008-2018 were used. An amount of 70% of the data were employed for model construction (training course) and 30% for model validation or evaluation (Khosravi et al, 2018). Table (1) shows the descriptive statistics of the available variables (minimum, maximum, average, standard deviation, and skewness) in the total dataset used.

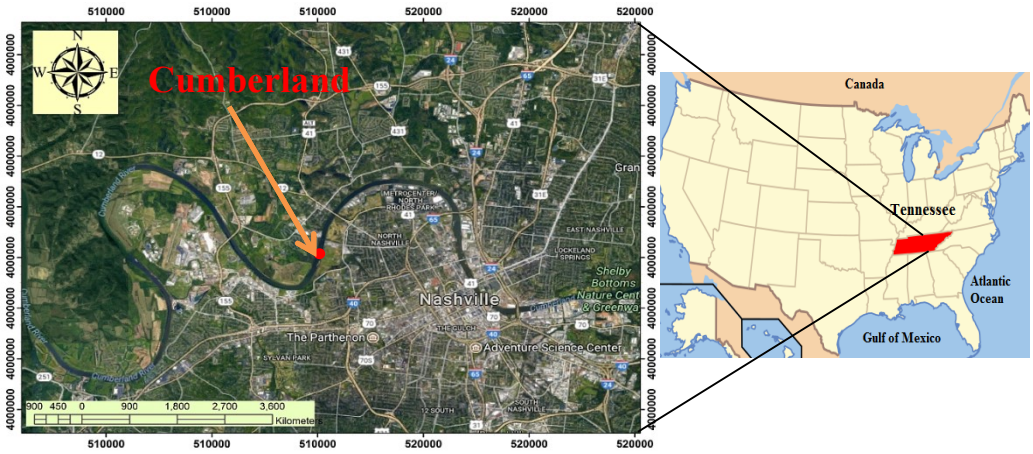

Figure 1. The studied region

Table 1. Statistical specifications of the parameters used

Parameter	Training			Testing		
	Minimum	Mean	Maximum	Minimum	Mean	Maximum
Q(m^3/s)	1.203	25.692	156.139	4.295	32.347	70.593
T(^0C)	5.33	17.204	28.41	5.62	16.272	26.18
DO(mg/l)	5.53	9.588	14.96	6.03	9.424	13.02

Methods

In this study, to simulate the amount of dissolved oxygen in the water of the Cumberland River, the support vector regression model approach was used. Given that based on recent findings this model has

some errors, model adjustment parameters were optimized by metaheuristic algorithms to reduce the model error. In recent years, several studies have investigated the SVR hybrid model with meta-heuristic algorithms; however, this study employed new algorithms that have not been studied in hydrological processes aiming to address the challenges and limitations of the existing model. In addition, a new algorithm was introduced to facilitate the simulation process and predict the river water quality using dependent parameters so that the decline of water quality, which will cause irreparable damage to surface water resources, can be prevented. Given that this is one of the most fundamental problems in world water issues, this study sets out to apply new algorithms including creative rifle, black widow spider, ski, and chicken swarm to simulate and estimate the amount of dissolved oxygen in river water.

According to the structure of artificial intelligence networks, the most basic step is to determine the modeling parameters. These parameters are usually determined through trial and error in artificial intelligence models such as SVR. Many factors affect the outcome of trial and error and the accuracy of the model prediction. Since these parameters are determined through trial and error, they generally reduce the predictive power of the model. Numerous solutions have been proposed by various researchers to address this fundamental weakness. One of these solutions is to combine backup vector regression with fuzzy logic. Researchers employ another solution to calculate the parameters and optimize these parameters by meta-heuristic algorithms. Many decision problems can be expressed as finite optimization problems which feature several decision variables subject to a few constraints. Hybrid optimization problems are usually easy to articulate, but difficult to solve. Two categories of algorithms employed for solving hybrid problems include exact and approximate algorithms. Accurate algorithms ensure finding the most optimal solution, but the problem is that these algorithms do not apply to difficult problems and the time required to find solutions to difficult problems will increase exponentially. Moreover, for most difficult problems, the algorithm accuracy is not satisfactory. If the optimal answer is not achievable using the exact algorithm in practice, we turn to the approximate algorithm. The approximation algorithm, commonly known as heuristic methods, seeks an appropriate and near-optimal solution. This method shortens the computation time compared to the previous method but does not ensure providing the most optimal solution. Meta-innovation is a general framework of algorithms that can provide solutions to the same problem with minor variation in different problems. Many meta-innovative algorithms are available such as Genetic Algorithm, Forbidden Search Simulation, Ant Society, Particle Swarm, Differential Evolution, Harmony Search, Artificial Bee Society, Firefly, Cuckoo or Frog, Frog Mutation, Invasive Weeds, and Insect Competition along with gravity, bats, spirals, pollinators, gray wolves, social spiders, lions, whales, locusts, and so on. Therefore, to optimize the model parameters in SVR, this research applied new innovative optimizer algorithms including creative rifleman, black widow spider, ski, and chicken swarm, presented in 2021, to solve hydrology and water issues for the first time.

Support Vector Regression

The support vector machine is an efficient learning system based on the theory of constrained optimization that uses the inductive principle of structural error minimization and leads to an overall optimal solution (Vapnik, 1995). In the SVR model, a function related to the dependent variable Y, which is itself a function of several independent variables x, is estimated. Like other regression problems, it is assumed that the relationship between independent and dependent variables with an algebraic function such as f (x) with some perturbation (allowable error (ε)) is determined (Vapnik, 1998) as:

$$f(x)=W^{T}.\emptyset(x)+b \quad (1)$$

$$y=f(x)+noise \quad (2)$$

where W^T is a transcript of the coefficients, constant *b* belongs to the properties of the regression function, and ∅ is the kernel function based on which the goal is to find a functional form for f (x). This is achieved by teaching the SVM model through a set of data (training set) (Misra et al., 2009). To calculate W and b, it is necessary to minimize the error function (Equation 3) in the SVM-ε model by considering the conditions (constraints) in Equations (4) and (5) (Hamel, 2009).

(3)

$$\frac{1}{2}W^T .W+C \sum_{i=1}^{N} \varepsilon_i +C\sum_{i=1}^{N} \varepsilon_i^*$$

(4)

$$W^T. \emptyset (X_i)+b-y_i \leq \varepsilon+ \varepsilon_i^* ,$$

(5)

$$y_i-W^T. \emptyset (X_i)-b \leq \varepsilon+ \varepsilon_i , \varepsilon_i ,$$
$$\varepsilon_i^* \geq 0 , i=1,2,\ldots,N$$

In the above equations, C is an integer and a positive number, which determines the penalty when an error occurs in model training. The kernel function is N (the number of instances and the two properties ε_i and ε_i^* are deficient variables. Finally, the SVM regression function can be rewritten as follows:

$$f(x)= \sum_{i=1}^{N} \bar{\alpha}_i \emptyset(x_i)^T. \emptyset(x)+b \quad (6)$$

In Equation 6, $\bar{\alpha}_i$ is the mean of the Lagrangian coefficients. Calculating ∅ (x) in its characteristic space can be very complex (Yoon et al, 2011). To solve this problem, the usual procedure in the SVM regression model is to select a kernel function as follows:

$$K(X_J ,X)=\emptyset(X_i)^T\sqrt{b^2-4ac} \quad (7)$$

Different kernel functions can be used to build different types of SVM-ε. The types of kernel functions that can be used in the SVM regression model are polynomial kernel 1, Radial Base Function (RBF) kernel, and linear kernel that are respectively calculated by the equations given below. Figure 2 shows the structure of the backup vector machine model. Given that the most widely used kernel functions are radial, linear, and polynomial (Basak et al, 2007; Vapnik and Chervonenkis, 1991), in this study, these three kernel functions were used to constitute the vector regression model. The parameters of the studied kernels C, t, and d converged to the optimal values using new ultra-exploration algorithms and then, the hybrid model was investigated. It is noteworthy that the process of backup vector machine calculations was performed based on coding in MATLAB environment and the kernel functions were optimized through trial and error.

$$k(x,x_j)=(t+x_i.x_j)^d \quad (8)$$

$$K(x,x_i)=\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \quad (9)$$

$$k(x,x_j)=x_i.x_j \quad (10)$$

Chicken Swarm Optimization

Chicken swarm optimization is a bio-inspired algorithm used for single-objective optimization (Zvache et al., 2019). This algorithm was proposed by Meng et al. (2014). In a group of N chickens, we distinguish the following numbers: RN, HN, CN, and MN, which represent the number of roosters, chickens, chickens, and hens, respectively. Figure (2) shows the flowchart of the chicken swarm algorithm.

The position of each chicken in a D-dimensional space is expressed according to Equation (11):

$$x_{i,j} (i \epsilon [1 , \ldots , D]), j\epsilon[1 , \ldots , D] \quad (11)$$

There are three types of chickens in the CSO algorithm. Each type of equation has its own proper motion. Roosters with the best proportions can find food in a wider area than those with worse proportions. The movement of roosters is obtained through Equations (12) and (13):

$$x_{i,j}^{t+1} = x_{i.j}^t * \left(1 + randn\,(0,\sigma^2)\right) \quad (12)$$

$$\sigma^2 = \begin{cases} 1 & ,if\ f_i \leq f_k \\ \exp\left(\frac{f_k-f_i}{|f_i|+\varepsilon}\right) & ,otherwise \end{cases} \quad (13)$$

In this relation, Randn (2,0) is a Gaussian distribution with mean 0 and standard deviation 2, which are very small constants used to avoid the error of dividing by zero. K is the rooster index that is randomly assigned between groups. Roosters are

selected and Fi is the proportion of rooster Xi. Chickens also follow their group in search for food. In addition, they accidentally steal food found by other chickens, although prohibited by them. The superior and dominant chickens in the competition for food have advantage over more obedient chickens. Mathematically, the movement of chickens can be obtained using Relations (14), (15), and (16):

$$x_{i,j}^{t+1} = x_{i,j}^{t} + S_1 * Randn * \left(x_{r_1,j}^{t} - x_{i,j}^{t}\right) + S_2 * Randn * \left(x_{r_2,j}^{t} - x_{i,j}^{t}\right) \quad (14)$$

$$S_1 = exp\left(\frac{f_i - f_{r_1}}{|f_i| + \varepsilon}\right) \quad (15)$$

$$S_2 = exp\left(f_{r_2} - f_i\right) \quad (16)$$

where Rand is a random number that is evenly distributed between 0 and 1. r1 is a rooster indicator and r2 is a chicken (rooster or chicken) indicator, both randomly selected from the crowd ($r_1 \neq r_2$). Chickens explore food around their mother. The movement of the chickens is obtained by Equation (17).

$$x_{i,j}^{t+1} = x_{i,j}^{t} + FL * \left(x_{m,j}^{t} - x_{i,j}^{t}\right) \quad (17)$$

where $x_{m,j}^{t}$ is the position of the mother of the i-th chicken so that m ϵ [1, N] and FL is a parameter that indicates how fast the chicken follows its mother. To consider the differences between different chickens, FL is randomly selected in the range [0, 2].

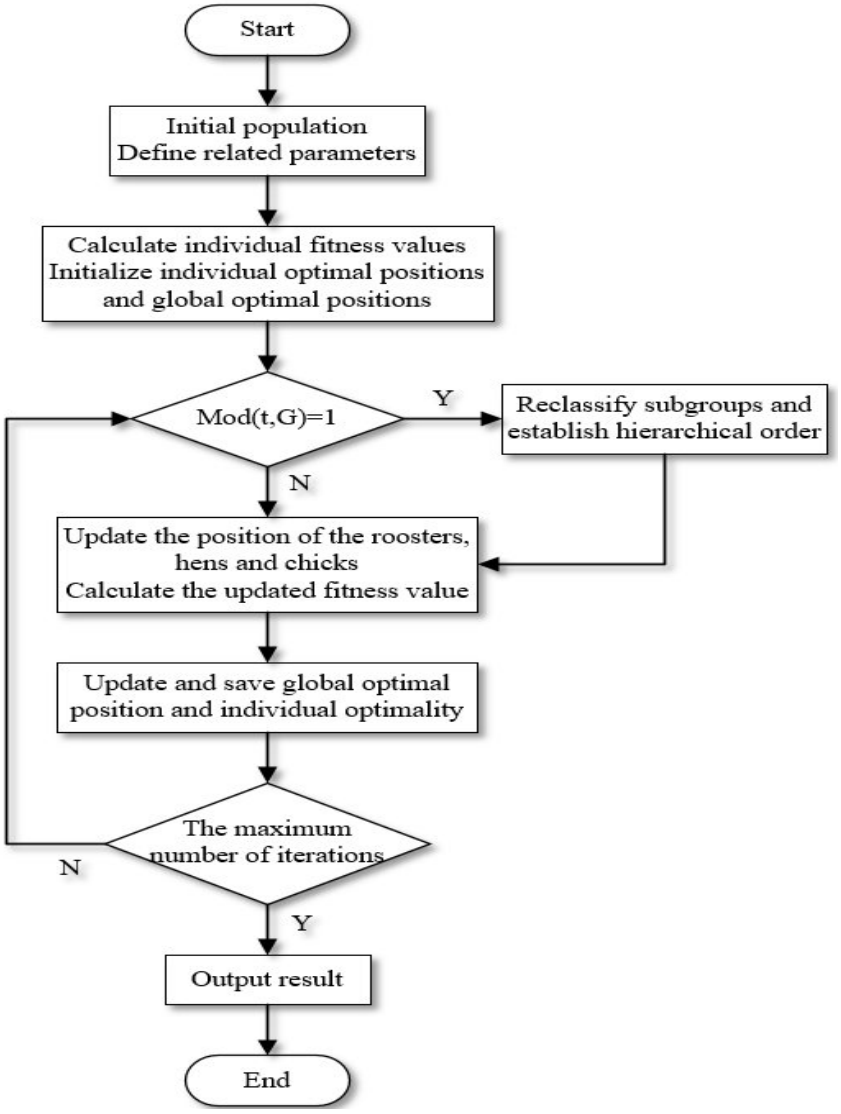

Figure 2. Flowchart of chicken swarm algorithm (Priyadarshi et al., 2017)

The Algorithm of the Innovative Gunner (AIG)

The algorithm of the innovative gunner is one of the very new meta-innovative optimization algorithms proposed by Pijarski & Kacejko (2019). The steps of this algorithm are summarized as follows:

1-Start the model at a starting point (the initial value for the first bullet determined randomly);

2-Determine the firing distance (firing distance of the bullet from the gun to the target point);

3-Calculate the produced bullet (the second bullet in the third stage taken from the first bullet);

4-Check the possibility of a bullet hitting the target (the location shot - did thebullet hit the target correctly?);

5-Select N random bullets as the main bullets (in case of hitting the target correctly);

6-Check and update the position where the bullet has hit the target (if the bullet hits the center of the target, the termination condition will be fulfilled and the work will be completed; but if it does not hit the target, the initial value must be redetermined);

7- Determine the best registered position;

8- Finish.

Figure (3) shows the general flowchart of the algorithm of the innovative gunner.

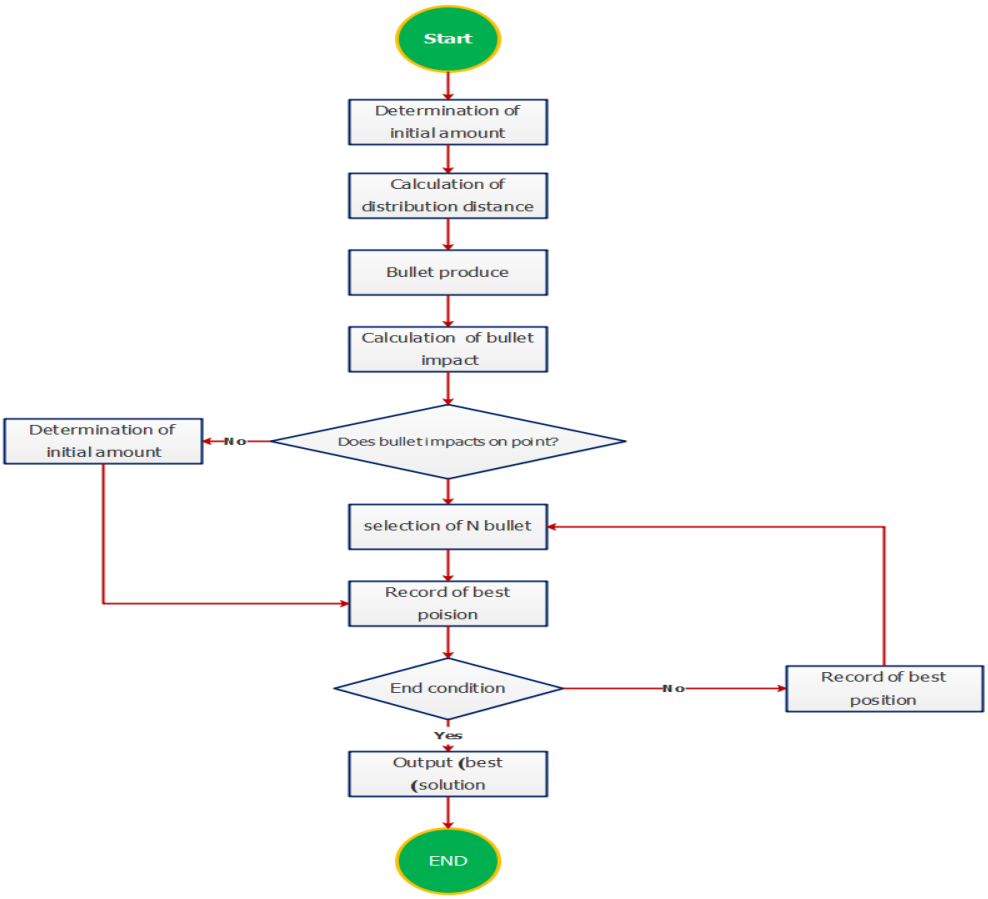

Figure 3. General flowchart of the AIG

Social Ski-Driver (SSD) optimization algorithm

In what follows, a novel optimization algorithm is proposed, which is called Social Ski-Driver (SSD) algorithm. The behavior of SSD was inspired by many different evolutionary optimization algorithms. Its name implies the fact that its stochastic exploration somehow resembles the paths that ski-drivers take downhill. SSD has some parameters a brief

description of which is given in the following:

- Positions of the agents ($X_i \in R_n$) are used to calculate the objective function at each location with n representing the dimension of the search space,
- Previous best position (Pi): The fitness value for all agents is calculated using the fitness function. The fitness value for each agent is then compared with its current position and the best position is stored. This is like the PSO algorithm (Poli et al., 2007).

The main objective of the SSD is to search in a space for optimal or near-optimal solutions. The number of parameters needed to be optimized determines the dimension of that space. In SSD, the positions (X_i) of agents are randomly initialized, where the number of agents is determined by the user.

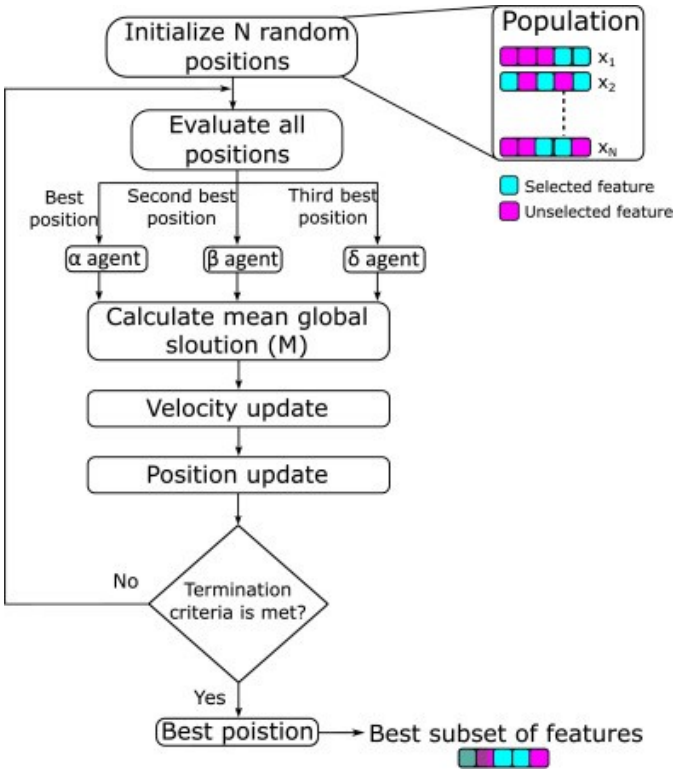

Figure 4. Flowchart of the SSD algorithm (Tharwat et al., 2020)

Evaluation Criteria and Comparison of Models

Certain criteria are used in any project to evaluate the efficiency of modeling. In the present study, different statistical criteria were employed to evaluate the efficiency of the models, including the coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute error (MAE), Nash-Sutcliffe Efficiency (NSE) coefficient, and Bias (Chai & Draxler, 2014; Legates & McCabe, 1999). The value of R^2 is in the [0-1] range and the closer it gets to one, the higher the prediction ability of the model will be and vice versa. Zero suggests that the model does not define the variations of the response data around the mean value, and one means that it defines all of them around the mean (Nagelkerke, 1991). Nash-Sutcliffe Efficiency (NSE) coefficient is a normalized statistic that defines the relative value of residual variance in comparison with the variance of the measured data (Nash & Sutcliffe, 1970; Moriasi et al., 2007). The NSE ranges between $-\infty < NSE < 1$ and the more its value approaches one, the more optimized the answer will be. The values between zero and one are generally accepted as the acceptable performance ratings and NSE$\ll$0

suggests that the mean observational values have higher predictive power than the estimated values, implying unacceptable performance of the model (Suie et al., 2020). This criterion was recommended by ASCE (1993) and its use is very common, because it provides a vast array of information regarding the reported values (ASCE, 1993). The use of this criterion has been highly welcomed in different scientific fields and numerous researchers throughout the world are benefiting from it (Sevat & Dezetter, 1991; Kesgin et al., 2020). Percentage of bias (PBIAS) measures the orientation of computational (simulated) data to their smaller or larger observational counterparts (Dabanli & Sen, 2018). The PBIAS value can be positive, negative, or zero. Zero suggests the optimal value and low-magnitude values suggest precision of the model during the simulation process. Positive and negative values denote the underestimation and overestimation of the model, respectively (Gupta et al., 1999). This criterion is favored more by the scholars in this field and applied by nearly all of them. It is also prevalent in most hydrological, water resource management, and geological studies (Musi et al., 2019; Pengxin et al., 2019). The above-mentioned criteria are derived by Equations (19-23) presented here.

$$R^2 = \left[\frac{\sum_{i=1}^{n}(M_{oi}-\bar{M}_0)(M_{ei}-\bar{M}_e)}{\sqrt{\sum_{i=1}^{n}(M_{oi}-\bar{M}_0)^2 \cdot \sum_{i=1}^{n}(M_{ei}-\bar{M}_e)^2}}\right]^2, 0 \le R^2 \le 1 \quad (18)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_{ei}-M_{oi})^2}\ 0 \le RMSE \le +\infty \quad (19)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|M_{ei}-M_{oi}|, 0 \le MAE \le +\infty \quad (20)$$

$$\text{NSE}=1-\frac{\sum_{i=1}^{n}(M_{ei}-M_{oi})^2}{(M_{ei}-\bar{M}_e)^2}, -\infty < NSE < 1 \quad (21)$$

$$PBIAS = \frac{\sum_{i=1}^{n}(M_{oi}-M_{ei})}{\sum_{i=1}^{n}M_{ei}} \times 100, -100 \le PBIAS \le 100 \quad (22)$$

Results and Discussion

One of the most important steps in modeling is choosing the right combination of input variables. In artificial intelligence models, selecting appropriate and effective initial input in order to teach the nature of the mechanism governing the phenomenon is an essential step toward improved performance (Satari et al., 2016; Nourani et al., 2016). In this study, in order to simulate the amount of dissolved oxygen in river water, monthly data of the Cumberland station during the statistical period of 2008 to 2018 were used. For modeling, the parameters of flow rate (Q) and temperature (T) in time steps t, t-1, and t-2 were used as input and the amount of Dissolved Oxygen (DO) as the output parameter of the model. It should be noted that 80% of the data for modeling and the remaining 20% for testing were randomly selected to model a wide range of data types (Kisi and Karahan, 2006; 2002 Nagy et al). The results of the models used are presented below.

Table 2. Combinations of input variables for selecting the best model

Number	Input	Output
1	Q(t)	DO(t)
2	Q(t),T(t)	DO(t)
3	Q(t),T(t),Q(t-1)	DO(t)
4	Q(t),T(t),Q(t-1), T(t-1)	DO(t)
5	Q(t),T(t),Q(t-1), T(t-1),Q(t-2)	DO(t)
6	Q(t),T(t),Q(t-1),T(t-1),Q(t-2), T(t-2)	DO(t)

In this study, to simulate the amount of dissolved oxygen in the Cumberland River, new models and algorithms were analyzed with an observational dataset and the highest efficiency was selected for further modeling and analysis. This step had six patterns that were selected as the best patterns of input compounds, which are described in Table (2). Also, for each hybrid model including SVR-AIG, SVR-SSD, and SVR-CSO, all the six combinations were used in the training and

testing stages (Khosravi et al, 2016). Recent studies have often evaluated the performance of AI networks relative to each other based on R2 or RMSE. The main goal of artificial intelligence systems is to reduce the estimation error. In this regard, in the present research, the criteria for determining the superiority of models were RMSE and R2. These two indicators have an inverse relationship with each other and decreasing RMSE increases R2. It can generally be said that adding variables with high CC in determining the output of a model increases the predictive power. As can be seen in Table (3), the combined SVR-AIG model had lower error than the other hybrid models under study, which can be due to not getting stuck in the local optimal points and faster convergence to the most optimal training parameters of the SVR model.

Table 3. Selection of the optimal input combination based on RMSE

Model	Evaluation Criteria	Phase	1	2	3	4	5	6
AIG-SVM	RMSE(mg/l)	Training	0.784	0.753	0.736	0.723	0.708	0.693
		Testing	0.727	0.711	0.694	0.681	0.665	0.644
SKI-SVM	RMSE(mg/l)	Training	0.934	0.918	0.903	0.891	0.884	0.878
		Testing	0.901	0.886	0.871	0.863	0.852	0.843
CSO-SVM	RMSE(mg/l)	Training	0.955	0.941	0.932	0.917	0.897	0.881
		Testing	0.928	0.911	0.893	0.884	0.877	0.864

In Table (3), given the different structure of each model, the composition of the optimal input variables is different for the models. For each model, RMSE values were estimated in both testing and training phases. The lowest amount of RMSE was selected in the test to comment on the accuracy of the models. As can be seen from Table (3), the sixth model had the best performance among all with the lowest RMSE and this was due to the increase in the number of input parameters (Dehghani et al, 2020). With normalization of data in the range of zero to one, the error was calculated very accurately. Also, given that the hybrid pattern 6 included more effective parameters or variables, the error was reduced by the same amount and therefore, it was preferable to other patterns. As the temperature of the river water increased, so did the dissolved oxygen in the water. Reduction in light also reduced the amount of oxygen released by plants. The reason is that that if light does not reach the plants completely, their production of oxygen will stop, and the existing bacteria will consume oxygen. On the other hand, with increasing river discharge, water concentration due to the entry of effluents does not increase and water-soluble oxygen does not decrease.

Model performance evaluation

In this study, novel developed models (SVR-SSD, SVR-CSO, SVR-AIG) were used to simulate the amount of dissolved oxygen in the water of the Cumberland River in the United States. First, different kernels including Radial Base Function (RBF), Polynomial (Poly), and Linear (Line) were studied. RBF was selected according to the evaluation indicators in combination with modern meta-heuristic algorithms. Then, by combining the support vector regression model with meta-heuristic algorithms, hybrid models were obtained. Finally, the mentioned evaluation indicators were utilized to analyze the hybrid models in relation to each other as well as to examine the series of observational data relative to the computational models of the diagrams. Time series changes, distribution, error rate, box plot, and violin and Taylor diagrams were used. To evaluate the studied models accurately, after normalizing the observational data, optimizing the parameters, and setting up the support vector regression model (C, t,

and d), R2 and RMSE indices, MAE, NSE, and PBIAS were used.

In summary, after selecting the best input combination for each model, the results of hybrid models for simulating dissolved oxygen in the observed river water, as given in Table 4, show that the Support Vector Regression-Algorithm of the Innovative Gunner (SVR-AIG) had better performance than other hybrid models, which were respectively, Support Vector Regression-Chicken Swarm Optimization (SVR-CSO), and Support Vector Regression-Social Ski-Driver (SVR-SSD). For the SVR-AIG model, we had R2 = 0.963, RMSE = 0.644 (mg / l), MAE = 0.568 (mg / l), NS=0.864, and BIAS = 0.001 in the testing phase. In general, it can be said that the SVR-AIG model had the best performance, while the SVR-CSO model had the weakest performance. The BIAS value was also positive for the study area, which meant underestimation by the model.

Table 4. Performance evaluation of hybrid models in simulating dissolved oxygen

Model	Training					Testing				
	R	RMSE (mg/l)	MAE (mg/l)	NSE	PBIAS	R	RMSE (mg/l)	MAE (mg/l)	NSE	PBIAS
AIG-SVM	0.950	0.693	0.512	0.935	0.001	0.963	0.644	0.568	0.864	0.001
SKI-SVM	0.918	0.878	0.641	0.873	0.003	0.936	0.843	0.767	0.821	0.003
CSO-SVM	0.914	0.881	0.655	0.864	0.003	0.928	0.864	0.775	0.815	0.003

Figure 6 shows the time series variation diagram and the distribution of observational and computational values. It shows good accuracy in estimating minimum and not much acceptable accuracy for maximum values. In addition, according to the Y = X distribution diagram, all the four hybrid models estimated computational values close to observational values, while SVR-AIG had better performance than others due to the equality of observational and computational values.

In Figure 7, the diagram for increase in the accuracy of the studied hybrid models compared to the single model shows that the SVR-AIG, SVR-CSO, and SVR-SSD models lead to more computational accuracy by 6.52%, 1.90%, and 1.75%, respectively, proving the higher efficiency of the hybrid models.

The mentioned relative superiority of the hybrid models lies in the lower number of their outlier data, the measurement accuracy of the observational parameters, the operator's precision of measurements, and favorable quality of the data. However, the evident point is that prediction by different machine learning models may lead to different results considering their various datasets and structures of algorithm. That is, any algorithm is a different and complex structure with its own advantages and disadvantages. In other words, some models perform differently with various problems and conditions due to multiple reasons. Besides, different studies conducted worldwide illustrate that there is no global index to prove absolute superiority of a model to others. More clearly, an algorithm may lead to the best optimal solution to one problem and the worst solution to another. Moreover, such solutions may be accompanied by low or high noise due to different factors such as the nature and type of the problem; structure and configuration of the model; proper definition of the problems, existing parameters, and borderline conditions; and low or high number of the data and their different qualities. Another significant issue should be noted concerning the differences between single and hybrid models. The studies being undertaken worldwide suggest that there is no globally accepted mechanism and standard procedure regarding the superiority of hybrid models over single ones. However, different studies suggest that hybrid models generally improve the performance of standalone models, which is in line with the findings of the present study.

The box diagram of the amount of dissolved oxygen in river water in Figure 8 shows that the SVR-AIG model properly fits with the observed maximum dissolved oxygen. Also, SVR-CSO, and SVR-SSD models are behind in terms of compatibility, in order. The same result was observed in predicting the minimum dissolved oxygen. According to Figure 7 and the results of the box plot diagram, it can be stated that although SVR is one of the smart and accurate models, it cannot predict the maximum values well. However, when combined with hybrid algorithms such as AIG, its performance in predicting maximum values is greatly improved.

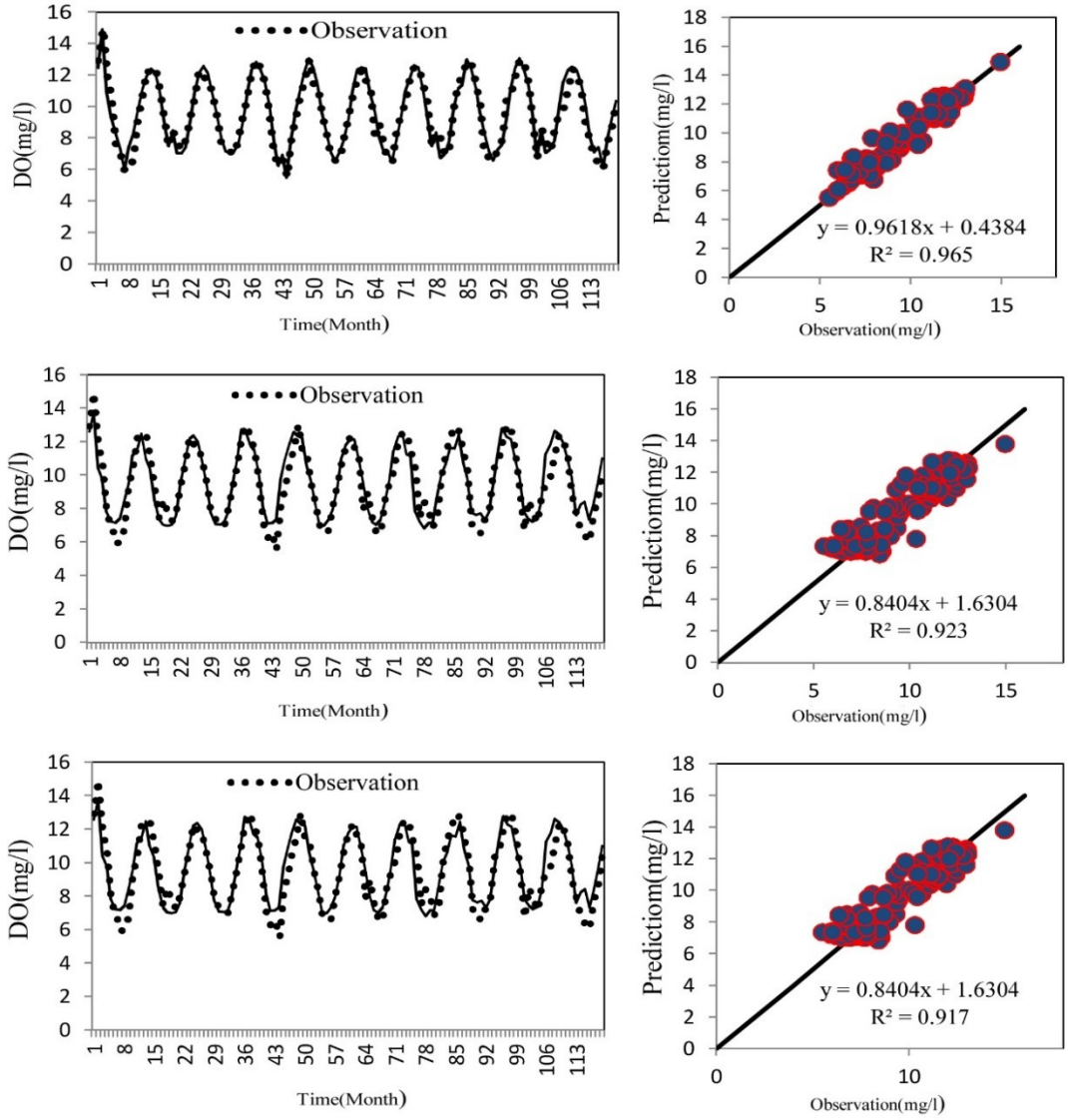

Figure 6. Scatter diagram and temporal variations of the observational and computational data for the three observation wells

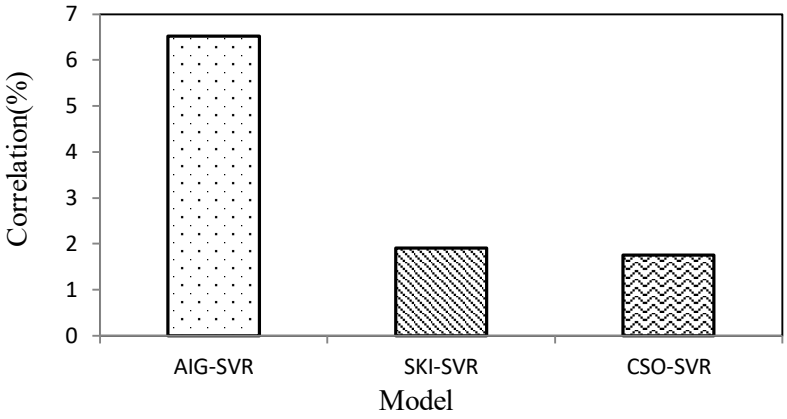

Figure 7. Correlation diagrams of the studied models compared to the single mo

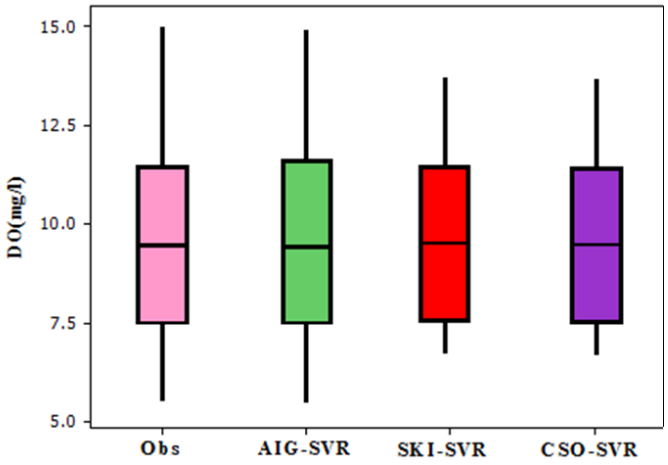

Figure 8. Box plot for the measured and predicted values

Conclusion

In general, it can be said that the developed models for simulating dissolved oxygen in the water of the Cumberland River in the United States achieved desirable prediction results. The results showed that the higher the number of effective parameters (dependent variables) in hand, the better the network performance. Also, the higher the input to the network, the higher the efficiency and accuracy of the model.

In this study, hybrid models based on support vector regression were used to simulate dissolved oxygen in river water. Studies by various researchers around the world show that support vector regression generally does not perform well for estimating hydrological phenomena due to the nature of trial and error in estimating kernel setting parameters. In other words, it is easier to calculate the parameters of setting kernels in the support vector regression model based on trial-and-error method and this is a weakness for the model. Our results showed that hybrid models had acceptable performance in increasing the estimation ability of the SVR model by 1.5 to 6.5%.

Also, according to the evaluation criteria, it was concluded that all the four models could accurately estimate the relatively high level of dissolved oxygen in river water. Meanwhile, the SVR-AIG model showed more accuracy and less error than the SVR-BWO, SVR-CSO, and SVR-SKI models.

Overall, the results of this study indicated the superiority of the AIG to other algorithms (based on correlation and RMSE criteria). This advantage goes back to the

powerful internal structure of this algorithm and the use of primary and secondary parameters, cost reduction function, and time saving and more effective convergence in achieving the optimal solution. In general, it can be said that most of the algorithms such as Chicken Swarm Optimization (CSO), and Social-Ski Driver (SSD) focus on the cost function and primary criteria, while in the AIG, in addition to the above-stated items, secondary parameters are also considered. This has a significant effect on the optimal results of the model. Also, the powerful structure of the AIG makes it possible to better converge to the optimal answer and local minima. In simpler terms, it can be said that the effect of these secondary parameters increases the speed of convergence. Also, their performance along with other factors reduces the search amplitude, resulting in better and faster convergence, because the more limited the search amplitude, the faster and more accurate the achievement of the optimal answer and the faster the convergence. Overall, this study showed that the use of SVR-AIG model could be effective in estimating dissolved oxygen in river water. This model can be useful in facilitating the development and implementation of surface water management strategies, which is a step forward in making management decisions to improve the quantity of surface water resources.

References

Afan, H. A., El-Shafie, A., Yaseen, Z. M., Hameed, M. M., Mohtar, W. H. M. W., and Hussain, A. 2015. ANN based sediment prediction model utilizing different input scenarios. Water Resources Management. 29(4), 1231–1245.

Ahmed, A.A.M., and Shah, S.M.A. 2017. Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. Journal of King Saud University-Engineering Sciences. 29(3), 237–243.

ASCE. 1993. Criteria for evaluation of watershed models. Journal of Irrigation Drainage Engineering, 119(3), 429-442.

Chai, T., and Draxler, R.R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geoscientific Model Development. 7, 1247–1250.

Dabanlı, İ., and Şen, Z. 2018. Precipitation projections under GCMs perspective and Turkish Water Foundation (TWF) statistical downscaling model procedures. Theoretical and Applied Climatology. 132, 153-166.

Dehghani, R., Torabi poudeh, H., Younesi, H., and Shahinejad, B. 2020. Daily streamflow prediction using support vector machine-artificial flora (SVM-AF) hybrid model. Acta Geophysica. 68, 1763–1778.

Deo, R. C., and Şahin, M. 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environmental Monitoring and Assessment. 188(2), 90.

Diaz, R.J., and Rosenberg, R. 2008. Spreading dead zones and consequences for marine ecosystems. Science. 321(5891), 926–929.

Dizaji, A. R., Hosseini, S. A., Rezaverdinejad, V., and Sharafati, A. 2020. Groundwater contamination vulnerability assessment using DRASTIC method, GSA, and uncertainty analysis. Arabian Journal of Geosciences, 13(14), 1–15.

Duie Tien, B., Khosravi, K., Tiefenbacher, J., Nguyen, H., and Kazakis, N. 2020.Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Journal of Science of the Total Environment 721, 136612.

Fadaee, M., Mahdavi-Meymand, A., and Zounemat-Kermani, M. 2020. Seasonal Short-Term Prediction of Dissolved Oxygen in rivers Via Nature-Inspired Algorithms. CLEAN–Soil, Air, Water, 48(2), 1900300.

Forstinus, N. O., Ikechukwu, N. E., Emenike, M. P., and Christiana, A. O. 2016. Water and waterborne diseases: A review. International Journal of Tropical Diseases and Health, 12(4), 1–14.

Gupta, H.V., Sorooshian, S., and Yapo, P.O. 1999. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. Journal of Hydrologic Engineering. 4(2), 135-143

Hamel, L. 2009. Knowledge Discovery with Support Vector Machines, Hoboken, N.J. John Wiley.

Hayyolalam, V., and Kazem, A. A. P. 2020. Black widow optimization algorithm: A novel meta-heuristic approach for solving engineering optimization problems. Engineering Applications of Artificial Intelligence, 87, 103249.

Ighalo, J. O., Adeniyi, A. G., Adeniran, J. A., and Ogunniyi, S. 2020. A systematic literature analysis of the nature and regional distribution of water pollution sources in Nigeria. Journal of Cleaner Production, 124566.

Kesgin, E., Agaccioglu, H., and Dogan, A. 2020. Experimental and numerical investigation of drainage mechanisms at sports fields under simulated rainfall. Journal of Hydrology. 580, 124251.

Khalil, B., Adamowski, J., Abdin, A., and Elsaadi, A. 2019. A statistical approach for the estimation of water quality characteristics of ungauged streams/watersheds under stationary conditions. Journal of Hydrology. 569, 106–116.

Khosravi, K., Nohani, E., Maroufinia, E., and Pourghasemi, H.R. 2016. A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights of evidence bivariate statistical models with multi-criteria method. Natural Hazards. 83(2), 1-41.

Kisi, O., Karahan, M., and Sen, Z. 2006. River suspended sediment modeling using fuzzy logic approach .Hydrol Process 20: 4351-4362.

Kisi, O., and Ay, M. 2012. Comparison of ANN and ANFIS techniques in modeling dissolved oxygen. Sixteenth International Water Technology Conference, IWTC-16, Istanbul, Turkey. 1–10.

Kisi, O., Alizamir, M., and Gorgij, A. D. 2020. Dissolved oxygen prediction using a new ensemble method. Environmental Science and Pollution Research. 1–15.

Krishna, R. S., Mishra, J., and Ighalo, J. O. 2020. Rising Demand for Rain Water Harvesting System in the World: A Case Study of Joda Town, India. World Scientific News. 146, 47–59.

Legates, D.R., and McCabe, G.J. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Recourse Research. 35, 233–241.

Li, W., Wu, H., Zhu, N., Jiang, Y., Tan, J., and Guo, Y. 2020. Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). Information Processing in Agriculture.8(1),185-193

Lin, J.Y., Cheng, C.T., and Chau, K.W. 2006. Using support vector machines for long-term discharge prediction. Hydrology Science Journal. 51, 3. 599–612

Lo Conti, F., Hsu, K.L., Noto, L.V., and Sorooshian, S. 2014. Evaluation and comparison of satellite precipitation estimates with reference to a local area in the Mediterranean Sea. Atmospheric Research. 138, 189-204.

Lu, H., and Ma, X. 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 249, 126169.

Meng, X., Liu, Y., Gao, X., and Zhang, H. 2014. A new bio-inspired algorithm: chicken swarm optimization. In International conference in swarm intelligence (pp. 86-94). Springer, Cham.

Misra, D., Oommen, T., Agarwal, A., Mishra, S.K., and Thompson, A.M. 2009. Application and analysis of support vector machine based simulation for runoff and sediment yield. Biosyst Eng. 103: 3. 527–535

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., and Veith, T.L. 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. American Society of Agricultural and Biological Engineers. 50(3), 885−900.

Nacar, S., Mete, B., and Bayram, A. 2020. Estimation of daily dissolved oxygen concentration for river water quality using conventional regression analysis, multivariate adaptive regression splines, and TreeNet techniques. Environmental Monitoring and Assessment, 192(12), 1–21.

Nagelkerke, N.J.D. 1991. A note on a general definition of the coefficient of determination. Biometrika. https://doi.org/10.1093/biomet/78.3.691.

Nagy, H., Watanabe, K., and Hirano, M. 2002. Prediction of sediment load concentration in rivers using artificial neural network model, Journal of Hydraulics Engineering. 128. 558-559.

Nash, J. E., and Sutcliffe, J.V. 1970. River flow forecasting through conceptual models: Part 1. A discussion of principles. Journal of Hydrology. 10(3), 282-290.

Niu, W.J., Feng, Z.K., Zeng, M., Feng, B.F., Min, Y.W., Cheng, C.T., and Zhou, J.Z. 2019. Forecasting reservoir monthly runoff via ensemble empirical mode decomposition and extreme learning machine optimized by an improved gravitational search algorithm. Applied Soft Computing Journal. 82(4),589-598.

Nourani, V. 2017. An Emotional ANN (EANN) approach to modeling rainfall-runoff process. Journal of Hydrology. 544(3), 267-277

Pengxin, D., Zhang, M., Bing, J., Jia, J., and Zhang, D. 2019.Evaluation of the GSMaP_Gauge products using rain gauge observations and SWAT model in the Upper Hanjiang River Basin. Atmospheric Research. 2191, 153-165.

Pijarski, P., and Kacejko, P. 2019. A new metaheuristic optimization method: the algorithm of the innovative gunner (AIG). Engineering Optimization. 51(12), 2049-2068.

Pincus, R., Batstone, C.P., Hofmann, R.J.P., Taylor, K.E., and Glecker, P.J. 2008. Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. Journal of Geophysical Research: Atmospheres. 113(D14), 1-10.

Poli, R., Kennedy, J., and Blackwell, T. 2007. Particle swarm optimization. Swarm Intell. 1(1), 33–57.

Priyadarshi, N., Azam, F., Solanki, S. S., Sharma, A. K., Bhoi, A. K., and Almakhles, D. 2017. A Bio-Inspired Chicken Swarm Optimization-Based Fuel Cell System for Electric Vehicle Applications. In Bio-inspired Neurocomputing (pp. 297-308). Springer, Singapore.

Salcedo-Sanz, S., Deo, R. C., Carro-Calvo, L., and Saavedra-Moreno, B. 2016. Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. Theoretical and Applied Climatology, 125(1–2), 13–25.

Sebastian, P. A., and Peter, K. V. (Eds.). 2009. Spiders of India. Universities press.

Sevat, E., and A. Dezetter. 1991. Selection of calibration objective functions in the context of rainfall-runoff modeling in a Sudanese savannah area. Hydrological Science Journal. 36(4), 307-330.

Sigaroudi, A. E., Nayeri, N. D., and Peyrovi, H. 2013. Antecedents of elderly home residency in cognitive healthy elders: A qualitative study. Global Journal of Health Science. 5, 200-2007

Taylor, K.E. 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysics Research Atmosphere. 106, 7183–7192.

Tharwat A, and Gabe T. 2019. Parameters optimization of support vector machines for imbalanceddata using social ski driver algorithm. Neural Computing and Applications. 32(3), 514-527

Tharwat, A., Darwishb, A., and Hassanien, A. 2020.Rough sets and social ski-driver optimization for drug toxicity analysis. Computer Methods and Programs in Biomedicine. 197, 1-11

Vapnik, V., and Chervonenkis, A. 1991. The necessary and sufficient conditions for consistency in the empirical risk minimization method. Pattern Recognition and Image Analysis. 1(3), 283-305.

Vapnik, V.N. 1995. The nature of statistical learning theory. Springer, New York, Pp: 250-320.

Vapnik, V.N. 1998. Statistical learning theory. Wiley, New York, Pp: 250-320.

Wehner, M.F. 2013. Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections, Climate Dynamics. 40(1-2), 59-80.

Yaseen, Z. M., Ehteram, M., Sharafati, A., Shahid, S., Al-Ansari, N., and El-Shafie, A. 2018. The integration of nature-inspired algorithms with least square support vector regression models: application to modeling river dissolved oxygen concentration. Water. 10 (9), 1124.

Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O, and Lee, K.K. 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. Journal of Hydrology. 396(4), 128–138

Zhang, Y.F., Fitch, P., and Thorburn, P. J. 2020. Predicting the Trend of Dissolved Oxygen Based on the kPCA-RNN Model. Water. 12(2), 585.

Zhu, S., and Heddam, S. 2020. Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN). Water Quality Research Journal. 55(1), 106–118.

Zouache, D., Arby, Y. O., Nouioua, F., and Abdelaziz, F. B. 2019. Multi-objective chicken swarm optimization: A novel algorithm for solving multi-objective optimization problems. Computers and Industrial Engineering, 129, 377-391.