Neural Comput & Applic
DOI 10.1007/s00521-017-3112-7

ORIGINAL ARTICLE

Hybrid soft computing approach for determining water quality indicator: Euphrates River

Jing Li[1,2] · Husam Ali Abdulmohsin[3] · Samer Sami Hasan[3] · Li Kaiming[4] · Belal Al-Khateeb[5] · Mazen Ismaeel Ghareb[6,7] · Muamer N. Mohammed[8,9]

Received: 21 December 2016 / Accepted: 15 June 2017

Abstract Recent approaches toward solving the regression problems which are characterized by dynamic and nonlinear pattern such as machine learning modeling (including artificial intelligence (AI) approaches) have proven to be useful and successful tools for prediction. Approaches that integrate predictive model with optimization algorithm such as hybrid soft computing have resulted in the enhancement of the accuracy and preciseness of models during problem predictions. In this research, the implementation of hybrid evolutionary model based on integrated support vector regression (SVR) with firefly algorithm (FFA) was investigated for water quality indicator prediction. The monthly water quality indicator (WQI) that was used to test the hybrid model over a period of 10 years belongs to the Euphrates River, Iraq. The use of the WQI as an application for this research was stimulated based on the fact that WQI is usually calculated using a manual formulation which takes much time, efforts and occasionally may be associated with errors that were not intended during the subindex calculations. The parameters considered during the formulation of the prediction model were water quality parameters as input and WQI as output. The SVR model was used to verify the accuracy of the inspected SVR–FFA model. Different statistical metrics such as best fit of goodness and absolute error measures were used to evaluate the model. The performance of the hybrid model in recognizing the dynamic and nonlinear pattern characteristics was high and remarkable compared to the pure model. The SVR–FFA model was also demonstrated to be a good and robust soft computing technique toward the prediction of WQI. The proposed model enhanced the absolute error measurements (e.g., root mean square error and mean absolute error) over the SVR-based model by 42 and 58%, respectively.

Keywords Support vector regression · Firefly algorithm · Regression problem · River water quality

✉ Li Kaiming
jlleaf@lzcu.edu.cn

1 Business School, Lanzhou City University, Lanzhou, Gansu, China

2 College of Earth and Environmental Science, Lanzhou City University, Lanzhou, Gansu, China

3 Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq

4 Geography and Planning School, Lanzhou City University, Lanzhou, Gansu, China

5 Computer Science Department, College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq

6 Department of computer science, College of Science and Technology, University of Human Development, Sulaymaniyah, Iraq

7 Department of Informatic, School of Computing and Engineering, University of Huddersfield, Huddersfield, England, UK

8 Faculty of Computer Systems and Software Engineering, University Malaysia Pahang, 26300 Kuantan, Pahang, Malaysia

9 IBM Center of Excellence, University Malaysia Pahang, 26300 Kuantan, Pahang, Malaysia

1 Introduction

1.1 Background

Soft computing techniques such as the AI techniques belong to the mathematical computational systems which usually involves a mimic process. AI models are

Published online: 05 July 2017

computational tools which analyze and understand the complex problems of nature by mimicking the naturally occurring nervous biological system of the human being [1]. In addition to this, they are also capable of generating optimal mathematical modeling in order to simulate the stochasticity which exists within the input and output of a systematical problem through the offering of a remarkable machine learning process [2–4]. Several data-driven models such as the linear regression (LR) and the autoregressive integrated moving average (ARIMA) have been used for environmental and ecological applications [5–7]. In fact, such models are linear and assumed the stationary state of a data set. Hence in hydrological processes, these typical models cannot efficiently handle the non-stationarity and nonlinearity of the data set involved [8–10]. Therefore, efforts have been channeled by scholars on the development and exploring of models which are intelligent enough to handle modeling processes involving nonlinear and non-stationary processes. The AI and its data-driven methods have demonstrated to be promising in the modeling and forecasting of the nonlinear environmental and hydrological processes [11]. It has also shown great progress in handling the large dynamicity of data sets, as well as data set noise concealing. These attributes make AI-based models well suited for modeling problems in hydrology [12, 13]. Numerous AI tools or techniques such as mathematical optimization, search optimization, as well as logics, statistical learning, classifications and probability-based methods have been reportedly used in hydrological studies [14].

In order to solve deficiencies of the NN methods, one among several AI approaches called support vector machine (SVM) was developed in 1995 by [15]. The principle of empirical risk minimization (ERM) was used to reduce errors during training the network of SVM. Furthermore, in order to reduce the upper bound of a generalization error, the SVM employed structural risk minimization (SRM) algorithm which is purely based on the principle that the boundaries of a generalization error are dependent on the generalization of the empirical errors and a confidence interval term. Avoiding the local optimums and obtaining the global optimum are the major objectives of the SVMs, and to realize this, the nonlinear problems are solved linearly at a higher rate compared to their initial dimensional feature space. The SVM has been employed in many applications such as pattern recognition, text categorization problems and nonlinear regression estimation [16, 17]. The SVM model which was previously proposed for clustering purposes has been expanded to cover the nonlinear regression problems known as the support vector regression (SVR) [18]. In a range of applications/prediction fields including atmospheric science prediction [19, 20], financial applications [21], as well as the hydrological and environmental applications [22–24], the SVR models have shown efficiency in their performances. The SVR model within a short period of time has shown efficiency in the areas of science and engineering where it has been reportedly applied [25].

Theoretically, there are three parameters which are (C, δ and ε) the trade-off between the training error and the regression function flatness, kernel function and the constant value which determines the width of the loss function in the SVR characterize SVR models. These parameters have a high influence on the forecasting accuracy, and the proper identification of these three parameters formulates serious issues with the SVR models. No proper guide or rule regarding the setting of the SVR model parameters exists even though many studies have suggested various approaches for setting them up [26, 27]. Having discussed the challenges of the SVR model within their tuning parameters, the aim of this study is to propose a novel, robust and accurate model based on the SVR model. Up to date, enough evidence to support the existence of an accurate and powerful tool for setting up these criteria (internal parameters for SVR model) is still limited [28]. In addition, the effect of the interaction of the parameters has not been sufficiently justified by the existing methods. The result of the aforementioned is an increased tendency in the utilization of evolutionary algorithms for the adjustment of the SVR parameters. In this paper, a novel optimization approach based on evolutionary facts called firefly algorithm (FFA) is adopted in this study for the optimization processes of the internal parameters SVR model. The result of the proposed algorithm was verified by comparing with those of the SVR-based model.

1.2 Related work

Machine learning is the application that imitating human brain features in solving natural problems [29]. These machine learning methods such as the neuro-fuzzy, artificial neural network, evolutionary computing and support vector machines possess the ability to respond to stimulus from input and generate the corresponding response [8]. This is achieved through recalling the memory of previous experiences in order to reproduce/generalize the data. Within the last two decades, various studies have been carried out in the field of AI and its implementation, especially in environmental studies [30–33], with greater attention to the surface water quality [12, 13, 34–36]. Being that AI can rapidly map a given input to generate the desired output, this attribute over the conventional computing paradigm has led to outcomes which can be achieved in a few clock cycles. The overall benefit of this is the enhancement of the efficiency of the analysis approach when it is compared with the conventional processes.

However, the methods of soft computing have been used to replace other procedures in engineering applications which are not time efficient. Furthermore, its simplicity and reliability in the analysis of problems at a near-perfect performance rating projected soft computing methods as a handy tool for solving engineering problems.

The earliest research was conducted for water quality determination using soft computing techniques by [37], in 2000. Gümrah et al. estimated the pollutant concentration of ground water using the application of artificial neural network. The same application has been utilized to predict dissolved oxygen concentration for river based on daily and monthly timescale historical data [38]. In 2004, Juahir et al. [39] studied the water quality indicator of tropical environment using ANN machine learning approaches. The modeling was established based on the region water quality parameters to construct the model. Another study was conducted to estimate river water quality with integrating other significant hydrological parameters like river flow [40]. Several studies proved the capability of ANN approach in river water quality modeling [34, 41–44]. On the other hand, there was an attempt to predict water quality COD and DO using least-square support vector machine integrated with particle swarm optimization (PSO) algorithm. The results showed an outperformance for the prediction over the multiple linear perceptron [45]. On the same manner, several water quality indicators including water temperature, plumbum and dissolved oxygen estimated using hybrid SVR model with PSO algorithm. The results indicated excellent performance of the proposed model [46]. In 2011, Singh and his co-researchers determined the biochemical oxygen demand parameter using the application of support vector machine. The SVR model internal parameters were optimized using grid search algorithm [47]. In the last decade, the evolutionary optimization algorithm like firefly algorithm showed a very creditable performance in optimizing SVR model [48–51]. Thus, the current study investigates the applicability of the hybrid SVR–FFA for environmental application.

1.3 Research objectives

The objectives of this research are as follows:

1. A novel approach based on hybrid SVR–FFA was inspected for water quality indicator prediction.
2. Exploring an intelligent predictive model that is characterized by a high capability of capturing the high dynamic and stochastic pattern of the river water quality.
3. Validate the hybrid SVR–FFA model against the SVR-based model in terms of prediction accuracy.

The article is structured as follows: Sect. 2 summaries the methods and material including the predictive, optimization algorithm, case study and the performance skills evaluators. Section 3 discusses the application and analysis. Finally, Sect. 4 outlines the conclusions and remarks of the study.

2 Theoretical overview

2.1 Support vector regression

The SVR as a soft computing learning algorithm recently has been used in so many fields such as environmental researches, soft computing and engineering application [52–54]. Compared to other statistical methods such as the neural network, it has been proved to show better performance accuracies in terms of prediction and forecasting [55]. The theory and process of the development of the SVR as developed by Vapnik have been documented in the literature [56].

The development of the SVR was based on the statistical machine learning development and also on structural risk minimization. This is aimed at reducing the error at the upper bound when compared to the local training error that is one of the commonly used technique machine learning methodologies. Based on the surveyed state-of-the-art, SVR has reported several improvements compared to other soft computing learning algorithms: (1) The implementation of a high-dimensional spaced set of kernel equations which specifically involve nonlinear transformations, thereby having no room for assumption in the functional transformation that makes data to be linearly separable and indispensable and (2) the second benefit of the method is the uniqueness of its solution because of the convex nature of the optimal problem.

Mathematically, the approximation function of SVR can be donated based on the theory stated by Vapnik's as follows:

$$f(x) = w.\varphi(x) + b \tag{1}$$

$$C = 0.5\|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n} L(x_i, d_i) \tag{2}$$

Let us assume we are given a range of data set $R = \{x_i, d_i\}_i^n$. In formula (1), the term $\varphi(x)$ represents the high-dimensional space that deals with the input candidates, whereas w and b are the normal vector and scalar. Equation (2) contains the following terms $0.5\|w\|^2$ and $C\frac{1}{n}\sum_{i=1}^{n} L(x_i, d_i)$ which indicate the stands error and regularization term, respectively. Formula (1) parameters including w and b are computed via the minimization function [57]:

$$\text{Min } R_{\text{SVR}}\left(w, \xi^{(*)}\right) = 0.5\|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n}\left(\xi_i, \xi_i^*\right) \tag{3}$$

$$\text{Subject to} \begin{cases} d_i - w.\varphi(x_i) + b_i \le \varepsilon + \xi_i \\ w.\varphi(x_i) + b_i - d_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \ i = 1, \ldots, l \end{cases} \tag{4}$$

where ξ_i and ξ_i^* are the positive slack variables that denote the upper and lower excess deviation. C refers to the error penalty used to control the trade-off between empirical error and regularization term. Finally, the ε defines the loss function connected to the approximation accuracies of the training data set.

By recalling formula (1), the constrains and the Lagrange can optimally be solved via the generic function [49] and best described as follows:

$$f\left(x, a_i, a_i^*\right) = \sum_{i=1}^{n}\left(a_i - a_i^*\right)K(x, x_i) + b \tag{5}$$

where $K(x, x_i)$ represents the kernel function. Here, the determination of the data correlation using nonlinear mapping methodology is the main goal of the SVR. Nonlinear learning machine can be generated using the defined kernel function represented as K, which is a straightforward computation technique. The inner product in a feature space that serves as a function to original input points was calculated using this method. The suitability of the SVR to kernel functions is important since it can subtly alter the information and transform the same into a higher-dimensional feature space. The outcome of the lower-dimensional original input space can be typified in the obtained results in such a space.

The four main kernel functions which are obtainable with the SVR model are the sigmoid, lineal, polynomial and the radial basis functions [55]. The ideal function in this category over the years has been the radial basis function (RBF) due to its efficient, simple and reliable abilities as well as an adaptable computation which is for optimization, especially in handling complex parameters [58, 59]. To train the RBF kernel equation, only a set of linear functions are required instead of the quadratic programming which is lengthy and complicated. Accordingly, the radial basis equation (with parameter σ) was adopted and the nonlinear radial basis kernel function defined as the accuracy of predictions using the RBF kernel function which depends on the selection of its three factors (γ, ε and C). The firefly algorithm was used to establish the optimal values of these factors in this study, and the flow chart of the hybrid model is displayed in Fig. 1.

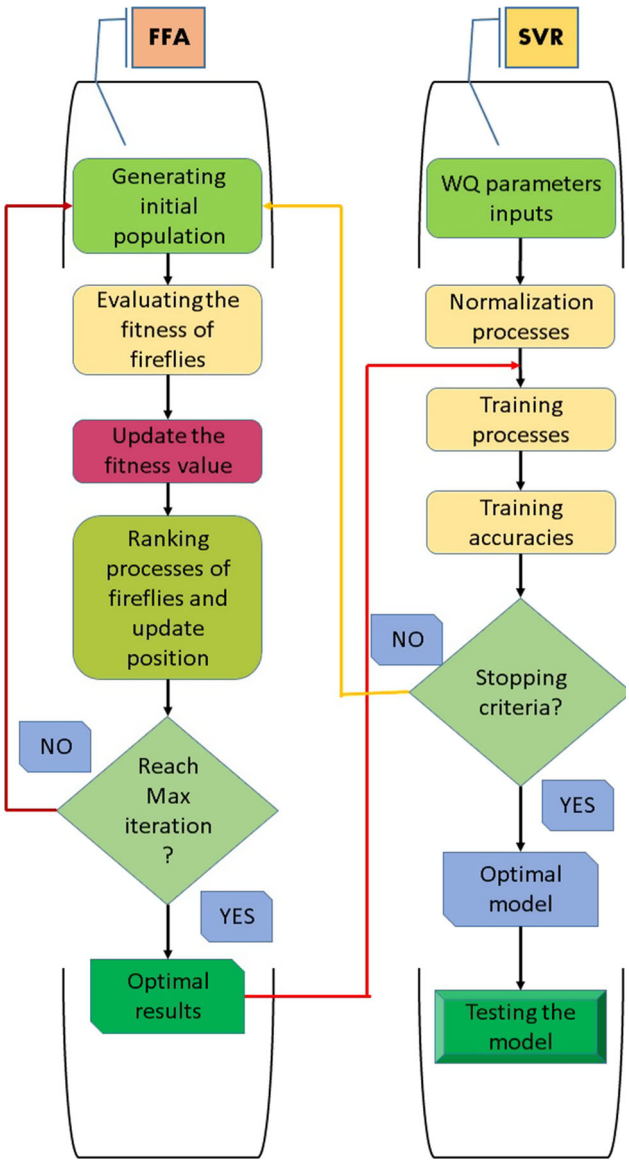

Fig. 1 The hybrid predictive model flowchart including support vector regression processes and the firefly optimization algorithm

2.2 The hybrid support vector regression–firefly algorithm

The metaheuristic optimization algorithms such as genetic algorithm (GA), ant colony optimization (ACO), cuckoo search (CS), FFA and particle swarm optimization (PSO) which are biologically inspired have been applied over the years for wide applications in optimization studies [50, 60–63]. The FFA developed by Yang et al. (2010) is one of the most recent approaches in metaheuristic optimization algorithms that are biologically inspired, and it is dependent on certain behavioral pattern, especially the light flashing characteristic of fireflies [64]. A firefly is an insect that attracts mates or prey through the principle of bioluminescence. The luminance from a firefly helps other fireflies to trace their paths in search of their prey. The development of many optimization algorithms stemmed from this firefly luminance production concept. The FFA has shown to be interesting, promising, efficient and robust,

when matched with the conventional metaheuristic algorithms in achieving both local and global optimization [65].

For the development of the FFA, basic fundamental rules are observed which stemmed from the following attributes of the fireflies. Firstly, all fireflies are unisex, thereby conferring with the ability to attract other fireflies irrespective of the sex. Secondly, the luminous intensity of a firefly determines the degree of its attractiveness; this intensity tends to decrease as the distance between the fireflies increases, prompting the ones with lesser luminous intensity to be attracted to the ones with better intensity. Thirdly, there is the fact that the nature of the cost function which is encoded affects the luminosity of an individual firefly. More technically, the brightness of the light from a firefly is dependent on the value of the objective function [66]. The formulation of the attractiveness, which is the objective function, and the variation of the luminosity are the major issues in the development of the FFA. For instance, the fitness function ought to be relatively proportional to the luminosity or the quantity of light emitted by the firefly during the design of functions for an optimal problem that involves the maximization of the objective function. Therefore, a reduction in the luminosity due to increasing distance between the fireflies will lead to changes in the intensity which will reduce the degree of attractiveness among them. The light intensity with the varying can be expressed mathematically as follows:

$$I(r) = I_o \exp\left(-\gamma r^2\right) \tag{6}$$

here I indicates the light intensity at a distance of r from a firefly, while the I_o donates the initial light intensity. The γ parameter was taken as a constant value ranging between 0.1 and 10, and following Sudheer et al. (2014) study [50], in the case of r it is equal to zero. The attractiveness term β at a distance r from the firefly defined as:

$$\beta(r) = \beta_o \exp\left(-\gamma r^2\right) \tag{7}$$

Cartesian distance between two fireflies i and j is indicated in the following formula:

$$r_{ij} = \left\|x_i + x_j\right\| = \sqrt{\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k}\right)} \tag{8}$$

The attraction movement between two fireflies i and j can be presented by:

$$\Delta x_i = \beta_o \exp\left(-\gamma r^2\right) \cdot \left(x_i - x_j\right) + \alpha \varepsilon_i \tag{9}$$

2.3 Modeling performance metrics

The modeling accuracy skills were validated using different statistical performance metrics such as correlation of coefficient (r^2), degree of agreement (d), root mean square error (RMSE), mean absolute error (MAE) and relative error distribution [67–69].

3 Application description

In Iraq region, there is an abundance of renewable and non-renewable water resources, but within the past three decades, there has been a shift where Iraq has moved from being water secured to becoming a water-stressed country [70]. Surface water, ground water, marshlands, lakes, rain, snowfall, reservoirs and drainage water formed the water resources in Iraq. One of the main two rivers in Iraq is Euphrates River. The Euphrates River which has its origins in Turkey flows through Syria and entered Iraq from the western border before discharging into the Shatt al-Arab that is one of the only two major rivers and flows through Iraq region supplying several cities in its way. Before the river flows into the Arab Gulf, it has traveled a distance of about 2700 km and the water from the river can be used for irrigation purposes, drinking, recreation and fishing. But unfortunately, the irrigation requirements of the basin are not favored by the seasonal distribution and availability of water because they do not coincide. In this study, the data set has been obtained from four sampling stations which are used for monitoring the quality of the river flow in Babylon governorate.

The Ministry of Environment, Department of Protect and Improve the Environment in the Middle Euphrates Region, provided the data which were used for the water quality study. The data covered a period of 10 years for the time period (2004–2013) and were accompanied by the values for the monthly average results of 14 water quality parameters. The main water quality parameters used in this study to compute the WQI are similar to several other regions such as Brazil [71], India [72], Portugal [73], Korea [74], USA [75] and many other countries. The water quality parameters are tabulated in Table 1 [76].

4 Analysis and discussion

In this study and for the first time, the implementation of hybrid model called SVR–FFA for monthly WQI prediction was inspected. The main modeling concept is to utilize the efficiency of firefly optimization algorithm to tune the main internal parameters of the kernel function of the SVR model including (γ, ε and C). Figure 2 shows the performance of the prediction models over the testing phase. The figure displays the actual and predicted (2012–2013) values for each single observation (24 months). The prediction pattern was revolved around the actual values. On a closer

Table 1 The water quality parameters standards for Iraq region [70]

Water quality parameters	Unit	Limitations
Total dissolved solids (TDS)	(mg/l)	2500
Biological oxygen demand (BOD)	(mg/l)	40
Chloride (Cl)	(mg/l)	250
Potassium (K)	(mg/l)	100
Dissolved oxygen	(mg/l)	5
Electrical conductivity (EC)	μS/cm	250
Sodium (Na)	(mg/l)	250
Magnesium (Mg)	(mg/l)	80
Alkalinity	(mg/l)	200
pH	–	4–8.5
Calcium (Ca)	(mg/l)	450
Phosphate (PO_4)	(mg/l)	25
Nitrate (NO_3)	(mg/l)	50
Total hardness (TH)	(mg/l)	300
Sulfate (SO_4)	(mg/l)	200

look, the hybrid SVR–FFA model presented closer prediction when compared with the standalone SVR model.

Figure 3a, b shows the scatter plots of the data belonging to the actual and the predicted responses of WQI for the testing data set covering the time period (2012–2013), for SVR and SVR–FFA models, respectively. According to those figures, it was observed that the two models had a good correlation with the line of best fit based on the correlation coefficient R [which is the root square of the coefficient of determination (r^2)] and the line of fit formula (assumed to be given as $y = a_o x + a_1$) in the scatter plots. However, it can be seen that SVR–FFA model possessed more accuracy and relatively outperformed the SVR model. In measurable terms, the r value (correlation coefficient) of the SVR–FFA and SVR models for the prediction of WQI was 0.94 and 0.90, respectively. Furthermore, coefficients of the a_o and a_1 for the SVR–FFA approach were closer to the 1 and 0, respectively, than the one obtained for the SVR model. Also, the ability of the hybrid support vector regression algorithm to capture the nonlinearity which exists between the water quality predictors and the predicted (WQI) was clearly explained with this observation.

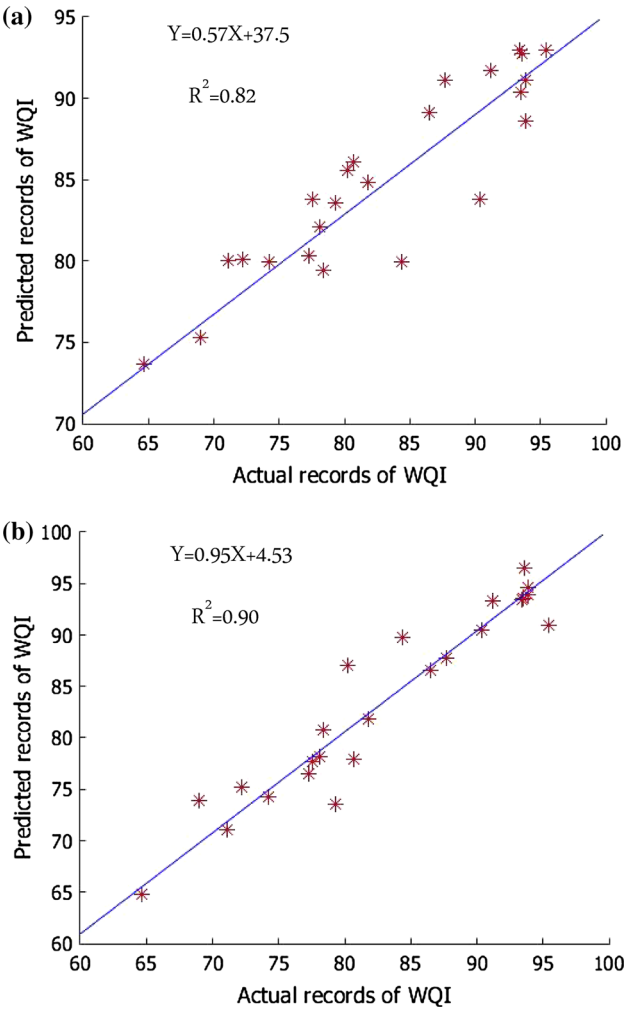

Fig. 3 Scatter plots between the actual and prediction models and the correlation coefficients **a** SVR model and **b** SVR–FFA model

Another excellent presentation for the modeling accuracies, the RE of both models (SVR–FFA and SVR) was also calculated in order to assess the performances of the

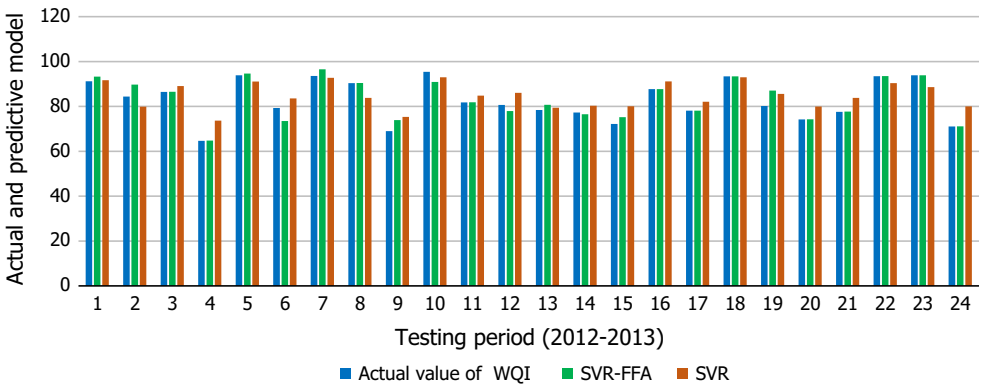

Fig. 2 Actual and predictive models for the testing period time series

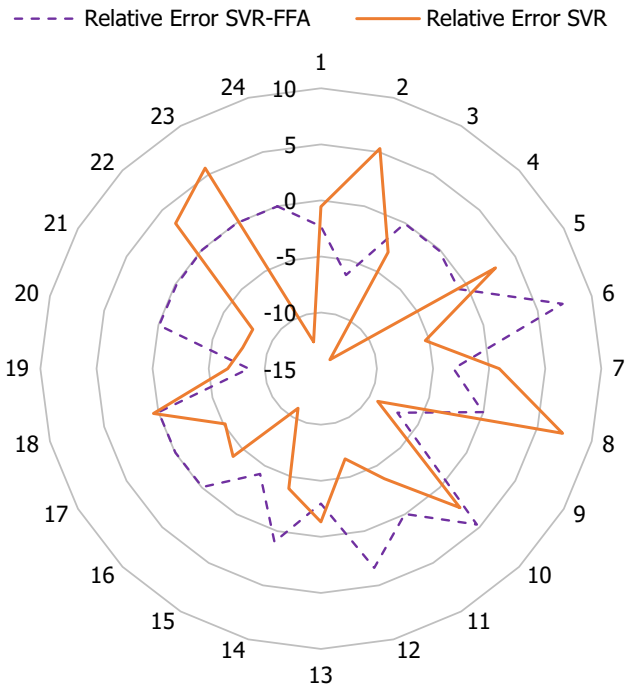

Fig. 4 The relative error distribution indicators over the testing period for both SVR and SVR–FFA models

two models. The RE (%) distribution is shown in Fig. 4. Based on the visualization of the testing phase results, the RE distribution (%) of SVR model exceeds −10%, and generally, the fluctuation of the RE ranged between 6 and −8% for more than 70% of the testing data set. On the other hand, the performance of the hybrid SVR–FFA model was quite good in comparison with SVR. The error distribution between most of the tested data (>80%) was within the range of (−5 and +5)%, while three of the cases showed a distribution error of +8, −8 and −7%, respectively. This outcome can better be explained as an improved attribute of the firefly optimization algorithm which is capable of tuning the internal variables of the predictive SVR model and leads to obtain better mimicking the pattern variables from water quality parameters which influenced the WQI.

To estimate the importance of each variable in predicting the water quality, the sensitivity analysis was carried out. In data mining, fitting or even model building, sensitivity analysis is usually referred to as the assessment of the predictors of importance in the fitted models. The sensitivity prediction usually ranks the variables of the predictor according to the decline in model performance which ensues when a variable is pulled out of the model. Also, the variables which can be safely ignored and those that must be retained in subsequent analyses can be identified [77]. The results of the analysis could be useful mainly for the sake of information or for pruning of input variable. The importance of each input parameter is reported in Fig. 5.

Table 1 presents the best performance evaluation criteria when using both predictive models. Based on Table 1, the best coefficient of determination of the models and their degree of agreement was (0.82–0.90) and (0.83–0.95), respectively, for SVR and SVR–FFA models. These indicated the hybrid model performed well as shown by the best fit goodness of the modeling process. The absolute errors which include the error of the root mean square and the mean absolute error for SVR and SVR–FFA models were (4.91–2.81) and (4.27–1.77), respectively. The SVR–FFA model augmentation over the SVR model can be expressed as 42% for RMSE and 58% for MAE. There is generally an observable enhancement in the prediction modeling using the SVR–FFA approach (Table 2).

It is worthy to state that the application of the hybrid SVR–FFA provided a better tool which can solve complex problems, including the assessment of river water quality. The uncertainties and randomness which are involved in the characterization of river water quality might at a certain time be captured by the SVR–FFA model. It could also capture the nonlinearity of the problem modeled in comparison with standalone SVR modeling.

5 Conclusions and remarks

With this in mind, conclusion can be made that the assessment of river WQI particularly in semiarid regions such as Iraq using hybrid model SVR–FFA and AVR models has advantages when compared to the routine or the manual computational methods. The classical method which was recommended to subjective empirical (subindices) requires added efforts and additional time to transform the fourteen raw data into its subindices. Furthermore, all the calculations when using this model depended on the subindices formulas that were gotten from the rating curves instead of from the original parameters. But the hybrid SVR–FFA approach uses the variables of the raw water quality for the training and testing instead of depending on the subindices which can lead to a more direct prediction of the WQI. Therefore, the soft computing models present a more direct, convenient and rapid technique for testing the WQI rather than the conventional methods.

This research accordingly highlighted the fact that the hybrid SVR–FFA model can well be a valuable tool for predicting water quality, especially in semiarid riverine environments since the calculation of the WQI is so simplified, thereby reducing substantially the time and efforts required for the optimization of the computations. These forms of approaches can be used worldwide in any aquatic

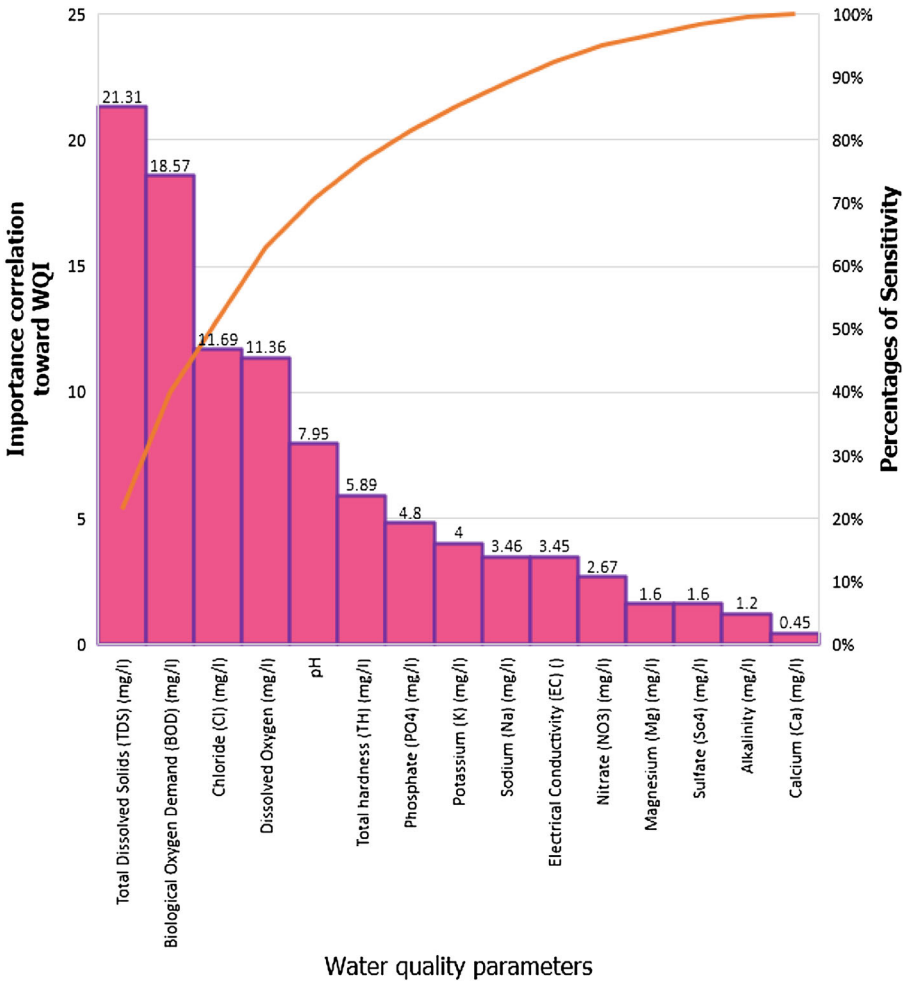

Fig. 5 The importance of each water quality parameter as an input candidate to predict the WQI

Table 2 The performance criteria indicators for both models (SVR and SVR–FFA) including coefficient of determination (r^2), degree of agreement (d), root mean square error (RMSE) and mean absolute percentage error (MAPE) over the testing phase data set

Performance indicators	r^2	d	RMSE	MAPE
SVR	0.82	0.83	4.91	4.27
SVR–FFA	0.90	0.95	2.81	1.77

system. This study, therefore, recommends that managers and water authorities should make use of the soft computing methods which are better in reliability and more direct alternatives for water quality prediction in wetlands as well as other water bodies.

It is also worthy to emphasize that the findings of the study contributed immensely to the proper identification of all the impacts of the activities of river basin development on river water quality. Further insight into the environmental consequences of land use change can also be available to decision makers from the findings of this study. The linking of the quality of river water with the land use types can allow for proper prediction of the quality of water in rivers depending on the variables used. The advantage of this is the reduction of the effort and provision of an intelligence-based mathematical model alternative. These policies will be those making policies and decisions to have a balance between the water resource usage and the sustainable developments. In view of these, an efficient soft computing technique such as AI is recommended to be used for the long-term analysis of water quality and for environmental monitoring records. The agencies responsible for water quality monitoring and their allies are encouraged by the results of this study to implement the AI-based models when assessing river water quality.

Acknowledgements This work is supported by National Natural Science foundation of China (41661014), the university scientific research project of Gansu (2016A-071) and the Urban Development Institute scientific research project of Gansu (2013-GSCFY-RW30). These supports are appreciated. Also, the authors would like to express their gratitude and appreciation to the reviewers.

Compliance with ethical standards

Conflict of interest Here is a declaration stated by all the authors that there is no conflict of interests about publishing this article.

Appendix

$$r^2 = \frac{\sum_{i=1}^{n}\left[\left(\mathrm{WQI}_a - \overline{\mathrm{WQI}}_a\right) \times \left(\mathrm{WQI}_p - \overline{\mathrm{WQI}}_p\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(\mathrm{WQI}_a - \overline{\mathrm{WQI}}_a\right)^2 \sum_{i=1}^{n}\left(\mathrm{WQI}_p - \overline{\mathrm{WQI}}_p\right)^2}}$$

$$d = 1 - \left[\frac{\sum_{i=1}^{n}\left(\mathrm{WQI}_a - \mathrm{WQI}_p\right)^2}{\sum_{i=1}^{n}\left(\left|\mathrm{WQI}_p - \overline{\mathrm{WQI}}_a\right| + \left|\mathrm{WQI}_a - \overline{\mathrm{WQI}}_a\right|\right)^2}\right]$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{n}\left(\mathrm{WQI}_a - \mathrm{WQI}_P\right)^2}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{n}\left|\mathrm{WQI}_a - \mathrm{WQI}_P\right|$$

$$\mathrm{RE}\% = \left[\frac{\mathrm{WQI}_a - \mathrm{WQI}_P}{\mathrm{WQI}_a}\right] \times 100$$

where WQI_a and WQI_P are the actual and predicted values of the water quality indicator and n is the number of observations over which the errors are predicted. $\overline{\mathrm{WQI}}$ indicates the mean values of the actual or predicted records.

References

1. Fogel LJ, Owens AJ, Walsh MJ (1967) Artifical intelligence through simulated evolution, vol 1. Wiley
2. Coppin B (2004) Artificial intelligence illuminated. Expert Syst. doi:10.1049/esn.1987.0009
3. Gevarter WB (1987) Introduction to artificial intelligence. Chem Eng Prog 83:21–37. doi:10.2207/qjjws1943.57.490
4. Russell SJ, Norvig P (1995) Artificial intelligence: a modern approach. Neurocomputing. doi:10.1016/0925-2312(95)90020-9
5. Ömer Faruk D (2010) A hybrid neural network and ARIMA model for water quality time series prediction. Eng Appl Artif Intell 23:586–594. doi:10.1016/j.engappai.2009.09.015
6. Abaurrea J, Asín J, Cebrián AC, García-Vera MA (2011) Trend analysis of water quality series based on regression models with correlated errors. J Hydrol 400:341–352. doi:10.1016/j.jhydrol.2011.01.049
7. Parmar K, Bhardwaj R (2014) Water quality management using statistical analysis and time-series prediction model. Appl Water Sci 4:1–10. doi:10.1007/s13201-014-0159-9
8. Yaseen ZM, El-shafie A, Jaafar O et al (2015) Artificial intelligence based models for stream-flow forecasting: 2000–2015. J Hydrol 530:829–844. doi:10.1016/j.jhydrol.2015.10.038
9. Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Model Softw 15:101–124. doi:10.1016/S1364-8152(99)00007-9
10. Nourani V, Kisi Ö, Komasi M (2011) Two hybrid artificial intelligence approaches for modeling rainfall-runoff process. J Hydrol 402:41–59. doi:10.1016/j.jhydrol.2011.03.002
11. Maier HR, Kapelan Z, Kasprzyk J et al (2014) Evolutionary algorithms and other metaheuristics in water resources: current status, research challenges and future directions. Environ Model Softw 62:271–299. doi:10.1016/j.envsoft.2014.09.013
12. Wang L, Li X, Cui W (2012) Fuzzy neural networks enhanced evaluation of wetland surface water quality. Int J Comput Appl Technol 44:235. doi:10.1504/IJCAT.2012.049087
13. Khalil B, Ouarda TBMJ, St-Hilaire A (2011) Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. J Hydrol 405:277–287. doi:10.1016/j.jhydrol.2011.05.024
14. Abrahart RJ, Anctil F, Coulibaly P et al (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Prog Phys Geogr 36:480–513. doi:10.1177/0309133312444943
15. Vapnik V (1995) The nature of statistical learning theory. Springer, New York, USA
16. Ji Y, Sun S (2013) Multitask multiclass support vector machines: model and experiments. Pattern Recognit 46:914–924. doi:10.1016/j.patcog.2012.08.010
17. Tabari H, Kisi O, Ezani A, Hosseinzadeh Talaee P (2012) SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. J Hydrol 444–445:78–89. doi:10.1016/j.jhydrol.2012.04.007
18. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in neural information processing systems, vol 9. MIT Press, Cambridge, MA, pp 155–161
19. Wang X, Ye M (2008) Hysteresis and nonlinearity compensation of relative humidity sensor using support vector machines. Sens Actuators B Chem 129:274–284. doi:10.1016/j.snb.2007.08.005
20. Samui P, Mandla VR, Krishna A, Teja T (2011) Prediction of rainfall using support vector machine and relevance vector machine. Earth Sci India 4:188–200
21. Kim K (2003) Financial time series forecasting using support vector machines. Neurocomputing 55:307–319. doi:10.1016/S0925-2312(03)00372-2
22. Goyal MK, Bharti B, Quilty J et al (2014) Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. Expert Syst Appl 41:5267–5276. doi:10.1016/j.eswa.2014.02.047
23. Rasouli K, Hsieh WW, Cannon AJ (2010) Short lead-time streamflow forecasting by machine learning methods, with climate variability incorporated. World Environ Water Resour Congr 2010:4608–4619. doi:10.1061/41114(371)468
24. Najah A, Karim OA, Jaafar O, El-shafie AH (2011) An application of different artificial intelligences techniques for water quality prediction. Int J Phys Sci 6:5298–5308. doi:10.5897/IJPS11.1180
25. Lingras P, Butz CJ (2010) Rough support vector regression. Eur J Oper Res 206:445–455. doi:10.1016/j.ejor.2009.10.023
26. Sudheer C, Maheswaran R, Panigrahi BK, Mathur S (2013) A hybrid SVM-PSO model for forecasting monthly streamflow. Neural Comput Appl. doi:10.1007/s00521-013-1341-y
27. Ch S, Anand N, Panigrahi BK, Mathur S (2013) Streamflow forecasting by SVM with quantum behaved particle swarm optimization. Neurocomputing 101:18–23. doi:10.1016/j.neucom.2012.07.017

28. Hong W-C (2009) Hybrid evolutionary algorithms in a SVR-based electric load forecasting model. Int J Electr Power Energy Syst 31:409–417. doi:10.1016/j.ijepes.2009.03.020
29. Adhikary BB, Mutsuyoshi H (2004) Artificial neural networks for the prediction of shear capacity of steel plate strengthened RC beams. Constr Build Mater 18:409–417. doi:10.1016/j.conbuildmat.2004.03.002
30. Bayram A, Kankal M, Tayfur G, Önsoy H (2013) Prediction of suspended sediment concentration from water quality variables. Neural Comput Appl 24:1079–1087. doi:10.1007/s00521-012-1333-3
31. Szemis JM, Maier HR, Dandy GC (2012) A framework for using ant colony optimization to schedule environmental flow management alternatives for rivers, wetlands, and floodplains. Water Resour Res 48:1–21. doi:10.1029/2011WR011276
32. Paleologos EK, Skitzi I, Katsifarakis K, Darivianakis N (2013) Neural network simulation of spring flow in karst environments. Stoch Environ Res Risk Assess 27:1829–1837. doi:10.1007/s00477-013-0717-y
33. May RJ, Maier HR, Dandy GC, Fernando TMKG (2008) Non-linear variable selection for artificial neural networks using partial mutual information. Environ Model Softw 23:1312–1326. doi:10.1016/j.envsoft.2008.03.007
34. Palani S, Liong SY, Tkalich P (2008) An ANN application for water quality forecasting. Mar Pollut Bull 56:1586–1597. doi:10.1016/j.marpolbul.2008.05.021
35. May DB, Sivakumar M (2009) Prediction of urban stormwater quality using artificial neural networks. Environ Model Softw 24:296–302. doi:10.1016/j.envsoft.2008.07.004
36. Niroobakhsh M (2012) Prediction of water quality parameter in Jajrood River basin: application of multi layer perceptron (MLP) perceptron and radial basis function networks of artificial neural networks (ANNs). Afr J Agric Res 7:4131–4139. doi:10.5897/AJAR11.1645
37. Gümrah F, Öz B, Güler B, Evin S (2000) The application of artificial neural networks for the prediction of water quality of polluted aquifer. Water Air Soil Pollut 119:275–294. doi:10.1023/A:1005165315197
38. Rounds SA (2002) Development of a neural network model for dissolved oxygen in the Tualatin River, Oregon. In: Proceedings of the second federal interagency hydrologic modeling conference, Las Vegas, Nevada
39. Juahir H, Zain SM, Toriman ME, Mokhtar M, Man HC (2004) Application of artificial neural network models for predicting water quality index. Malays J Civ Eng 16(2):42–55
40. Diamantopoulou MJ, Papamichail DM, Antonopoulos VZ (2005) The use of a neural network technique for the prediction of water quality parameters. Oper Res 5:115–125. doi:10.1007/BF02944165
41. Singh KP, Basant A, Malik A, Jain G (2009) Artificial neural network modeling of the river water quality—a case study. Ecol Model 220:888–895. doi:10.1016/j.ecolmodel.2009.01.004
42. Najah A, Elshafie A, Karim OA, Jaffar O (2009) Prediction of Johor River water quality parameters using artificial neural networks. Eur J Sci Res 28:422–435
43. Najah AA, El-Shafie A, Karim OA, Jaafar O (2012) Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation. Neural Comput Appl 21:833–841. doi:10.1007/s00521-010-0486-1
44. Najah A, El-Shafie A, Karim OA, Jaafar O (2011) Integrated versus isolated scenario for prediction dissolved oxygen at progression of water quality monitoring stations. Hydrol Earth Syst Sci 15:2693–2708. doi:10.5194/hess-15-2693-2011
45. Yunrong X, Liangzhong J (2009) Water quality prediction using LS-SVM with particle swarm optimization. In: Proceedings of the WKDD 2009 second international workshop on knowledge discovery and data mining, pp 900–904
46. Xuan W, Jiake L, Deti X (2010) A hybrid approach of support vector machine with particle swarm optimization for water quality prediction. In: 2010 5th International conference on computer science and education, pp 1158–1163. doi:10.1109/ICCSE.2010.5593697
47. Singh KP, Basant N, Gupta S (2011) Support vector machines in water quality management. Anal Chim Acta 703:152–162. doi:10.1016/j.aca.2011.07.027
48. Moghaddam TB, Soltani M, Shahraki HS et al (2016) The use of SVM-FFA in estimating fatigue life of polyethylene terephthalate modified asphalt mixtures. Meas J Int Meas Confed 90:526–533. doi:10.1016/j.measurement.2016.05.004
49. Shamshirband S, Mohammadi K, Tong CW et al (2016) A hybrid SVM-FFA method for prediction of monthly mean global solar radiation. Theor Appl Climatol 125:53–65. doi:10.1007/s00704-015-1482-2
50. Ch S, Sohani SK, Kumar D et al (2014) A support vector machine-firefly algorithm based forecasting model to determine malaria transmission. Neurocomputing 129:279–288. doi:10.1016/j.neucom.2013.09.030
51. Gocić M, Motamedi S, Shamshirband S et al (2015) Soft computing approaches for forecasting reference evapotranspiration. Comput Electron Agric 113:164–173. doi:10.1016/j.compag.2015.02.010
52. Yang H, Chan L, King I (2002) Support vector machine regression for volatile stock market prediction. Intell Data Eng Autom. doi:10.1007/3-540-45675-9_58
53. Liu D, Chen Q (2013) Prediction of building lighting energy consumption based on support vector regression. In: 2013 9th Asian control conference ASCC 2013. doi:10.1109/ASCC.2013.6606376
54. Zhao W, Tao T, Zio E (2015) System reliability prediction by support vector regression with analytic selection and genetic algorithm parameters selection. Appl Soft Comput J 30:792–802. doi:10.1016/j.asoc.2015.02.026
55. Raghavendra S, Deka PC (2014) Support vector machine applications in the field of hydrology: a review. Appl Soft Comput J 19:372–386. doi:10.1016/j.asoc.2014.02.002
56. Vapnik VN (2000) The nature of statistical learning theory. Springer, Berlin. doi:10.1109/TNN.1997.641482
57. Vapnik VN (1998) Statistical learning theory. Wiley, London. doi:10.2307/1271368
58. Yaseen ZM, Kisi O, Demir V (2016) Enhancing long-term streamflow forecasting and predicting using periodicity data component: application of artificial intelligence. Water Resour Manag. doi:10.1007/s11269-016-1408-5
59. Wu KP, Wang SD (2009) Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Pattern Recognit 42:710–717. doi:10.1016/j.patcog.2008.08.030
60. Gromov VA, Shulga AN (2012) Chaotic time series prediction with employment of ant colony optimization. Expert Syst Appl 39:8474–8478. doi:10.1016/j.eswa.2012.01.171
61. Walton S, Hassan O, Morgan K, Brown MR (2011) Modified cuckoo search: a new gradient free optimisation algorithm. Chaos Solitons Fractals 44:710–718. doi:10.1016/j.chaos.2011.06.004
62. Dimatteo A, Vannucci M, Colla V (2014) Prediction of mean flow stress during hot strip rolling using genetic algorithms. ISIJ Int 54:171–178. doi:10.2355/isijinternational.54.171

63. Zhao L, Yang Y (2009) PSO-based single multiplicative neuron model for time series prediction. Expert Syst Appl 36:2805–2812. doi:10.1016/j.eswa.2008.01.061
64. Yang X-S (2010) Firefly algorithm, stochastic test functions and design optimization. Int J Bioinspir Comput 2(2):78–84. doi:10.1504/IJBIC.2010.032124
65. Mohammadi S, Mozafari B, Solimani S, Niknam T (2013) An adaptive modified firefly optimisation algorithm based on hong's point estimate method to optimal operation management in a microgrid with consideration of uncertainties. Energy 51:339–348. doi:10.1016/j.energy.2012.12.013
66. Olatomiwa L, Mekhilef S, Shamshirband S et al (2015) A support vector machine-firefly algorithm-based model for global solar radiation prediction. Sol Energy 115:632–644. doi:10.1016/j.solener.2015.03.015
67. Legates DR, Mccabe GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 35:233–241
68. Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82. doi:10.3354/cr030079
69. Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250. doi:10.5194/gmd-7-1247-2014
70. Zolnikov TR (2013) The maladies of water and war: addressing poor water quality in Iraq. Am J Public Health 103:980–987. doi:10.2105/AJPH.2012.301118
71. Abrahão R, Carvalho M, Da Silva WR et al (2007) Use of index analysis to evaluate the water quality of a stream receiving industrial effluents. Water SA 33:459–465. doi:10.4314/wsa.v33i4.52940
72. Sargaonkar A, Deshpande V (2003) Development of an overall index of pollution for surface water based on a general classification scheme in Indian context. Environ Monit Assess 89:43–67. doi:10.1023/A:1025886025137
73. Bordalo AA, Teixeira R, Wiebe WJ (2006) A water quality index applied to an international shared river basin: the case of the Douro River. Environ Manag 38:910–920. doi:10.1007/s00267-004-0037-6
74. Song T, Kim K (2009) Development of a water quality loading index based on water quality modeling. J Environ Manag 90:1534–1543. doi:10.1016/j.jenvman.2008.11.008
75. Cude CG (2001) Oregon water quality index: a tool for evaluating water quality management effectiveness. J Am Water Resour Assoc 37:125–137
76. Abbood DW, Gubashi KR, Abbood HH (2014) Evaluation of water quality index in the main Drain River in Iraq by application of CCME water quality. Civ Environ Res 6:19–24
77. Olszewski T, Ryniecki A, Boniecki P (2008) Neural network development for automatic identification of the endpoint of drying barley in bulk. J Res Appl Agric Eng 53(1) 26–31