AI in Civil Engineering

ORIGINAL ARTICLE

Open Access

Simulation of reservoir outflows using regression tree and support vector machine

Vijay Kaushik[1*] and Noopur Awasthi[1]

Abstract

Water stored in reservoirs has a lot of crucial function, including generating hydropower, supporting water supply, and relieving lasting droughts. During floods, water deliveries from reservoirs must be acceptable, so as to guarantee that the gross volume of water is at a safe level and any release from reservoirs will not trigger flooding downstream. This study aims to develop a well-versed assessment method for managing reservoirs and pre-releasing water outflows by using the machine learning technology. As a new and exciting AI area, this technology is regarded as the most valuable, time-saving, supervised and cost-effective approach. In this study, two data-driven forecasting models, i.e., Regression Tree (RT) and Support Vector Machine (SVM), were employed for approximately 30 years' hydrological records, so as to simulate reservoir outflows. The SVM and RT models were applied to the data, accurately predicting the fluctuations in the water outflows of a Bhakra reservoir. Different input combinations were used to determine the most effective release. For cross-validation, the number of folds varied. It is found that quadratic SVM for 10 folds with seven different parameters would give the minimum RMSE, maximum R^2, and minimum MAE; therefore, it can be considered as the best model for the dataset used in this study.

Keywords Reservoir outflow, Regression tree, Support vector machine, Error analysis

1 Introduction

As tangible constructs (either artificial or natural), reservoirs are used as water storage for supervision, monitoring and maintenance of water supply (Hussain et al., 2011), forming the most valuable element in water resources systems. Due to environmental issues, however, construction of new dams is not an easy task; therefore, it is important that active reservoirs be boosted for the maximum effectiveness, so as to handle any present and future water-related challenges. Reservoirs are built with a dam across a flow. The major feature of a reservoir is the rule of herbal streamflow with the aid of storing surplus water within moist seasons and liberating the saved water in destiny dry seasons, thus complementing the discount in river flows. The intention is to balance streamflows and change the sequential and three-dimensional water availability. The water stored in a reservoir can be distributed later for advantageous uses, giving rise to sequential changes or reroutes through waterways or pipelines to outlying locations, and thus resulting in three-dimensional changes. Reservoir outflow projection is guided by various potential constraints, such as water storage, inflow of water, water level, evaporation, infiltration, geomorphology, and other factors, all of which need to be considered, so as to understand the ambiguity. Plentiful methods have been used in forecasting hydrological practices over the past years. Traditional tactics used are of linear mathematical relationships based on the capability of machinists, a simple set of curve fitment, and standards employed to quote reservoir water outflows (Tokar & Markus, 2000). However, poor performances are found in numerical models, due to their unavailability and complexity of statistics, missing datapoints and overemphasized constraints. Various

*Correspondence:
Vijay Kaushik
vijaykaushik_2k20phdce01@dtu.ac.in
[1] Research Scholar, Department of Civil Engineering, Delhi Technological University, Delhi 110042, India

machine learning algorithms have been used in previous research, with an intention to overcome these concerns and estimate reservoir water outflows (Mokhtar et al., 2014; Seckin et al., 2013). Subsequently, many Machine Learning (ML) models, including Artificial Neural Networks (ANNs), Radial Basis Neural Networks (RBNN), Support Vector Machines (SVMs), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), Logistic Regression (LR), etc., have been deployed in water management systems progressively, so as to improve the consistency and precision of the estimation models (Ahmadlou et al., 2019; Bowden et al., 2002; Naghibi & Pourghasemi, 2015). Modelling a machine to work and improvise on its own without explicit programming each time is called ML. In intellectual studies, ML shows the capability to solve complex problems at a high level of accuracy and can make predictions as demanded for certain future periods (Mullainathan & Spiess, 2017). Nowadays, AI models have been extended successfully to the field of reservoir operation. Compared to conventional physical prediction models, ML models can, with the help of historical dataset, learn numerous hydrological operations independently at acceptable correct operating rates. Advantages of such modelling lie in the capability of their software system to map input–output models (Hejazi & Cai, 2009; Hipni et al., 2013; Najah et al., 2011). To forecast daily water levels, five different ANN models were tested each with an increasing number of inputs, finding that the accuracy began to decrease with the addition of many inputs. The reason for this is that the network started to be obsolete and irrelevant, as explained in the research (Nwobi-Okoye & Igboanugo, 2013). By comparing the performance of SVM and multilayer perceptron (MLP), it is found that due to the optimization algorithms, SVM has a great deal of capacity to resolve a linearly constrained quadratic programming function, and the optimum kernel function in this case is a radial basis kernel function (Khan & Coulibaly, 2006)**.** During the process of creating fuzzy membership functions, a study on the ANFIS technique observed that triangular and trapezoid membership functions were deemed to be more suitable than bell-shaped membership functions (Shafaei & Kisi, 2016). In addition, a genetic algorithm (GA) was successfully utilized in optimizing reservoir operations; and by using data collected in a longer period of time, the GA model could be further improved for reservoir water levels (Onur Hınçal et al., 2011). Many more AI methods, such as the adaptive network-based fuzzy inference system (ANFIS), genetic algorithm (GA) and decision tree, have been effectively applied in the reservoir operation field, in addition to these AI algorithms . In fact, many reservoirs in California have used the improved decision tree (DT) algorithms, classification methods and regression trees to estimate water storage or release (Yang et al., 2016).

In this study, five models of the SVM algorithm and Regression tree (RT) algorithm were compared with an increasing number of inputs by using 5-, 10- and 20-fold cross-validation for original data, so as to accurately forecast fluctuations of the water level of Rupnagar's Bhakra reservoir (Ropar). A lot of quantitative metrics, including root mean square error (RMSE), correlation coefficient (R^2), and mean absolute error (MAE), were used to validate and compare these models; and MATLAB R2021b was used to design and code the modelling and data procedures.

2 Materials and methods

2.1 The study area

Bhakra Dam on the Sutlej River (Bilaspur, Himachal Pradesh) is a concrete gravity dam in the northern part of India, with geographic coordinates of 31°24′39″N latitude and 76°26′0″ E longitude. The dam is considered to be the highest gravity dam in the world. The Sutlej River, a major tributary of the Indus River, originates in Tibet and flows into the Indo-Gangetic plains near Bhakra. The overall upriver catchment area of the Bhakra River is 56,980 km^2. The precipitation in the catchment changes around an annual average of about 875 mm. Situated in a canyon near the (now submerged) upstream Bhakra community in the Himachal Pradesh district of Bilaspur, the dam is 226 m high, 518.25 m long and 9.1 m wide. Its "Gobind Sagar" reservoir can hold up to 9.34 billion cubic metres of water. The Bhakra dam generates a 90-km-long reservoir, covering 168.35 square kilometres and forming India's third-largest reservoir in terms of water storage capacity. The Bhakra Beas Management Board (BBMB) is in charge of the dam's operation and maintenance.

As a straight gravity cum concrete dam, Bhakra Dam has four radial spillway gates and an 8212 cumec designed overflow capacity. The location of the study area is shown on the map in Fig. 1. The Nangal reservoir is built with a 28–95 m high dam, situated at about 11 km downstream of the Bhakra dam. It controls irrigation releases by acting as a head regulator. During the monsoon, the dam would retain extra water; and then, it would release the water gradually throughout a year. It also prevents flood damage caused by monsoon rains. This dam feeds the Bhakra canal, which irrigates 10 million acres (40,000 km^2) of land in Haryana Punjab and Rajasthan. Table 1 shows the characteristics of the Bhakra Nagal dam and reservoir.

2.2 Data collection

A total of 2976 historical data points (for 30 years) were used, including: the reservoir level (M), the monthly

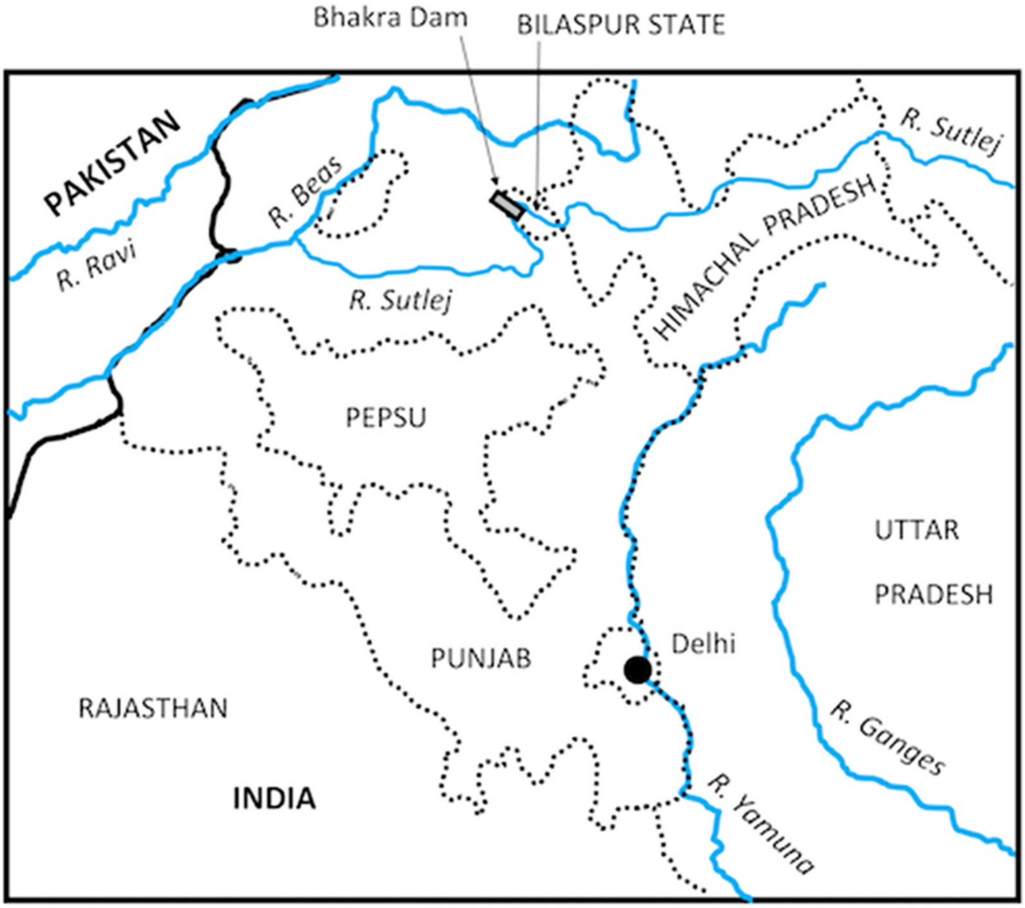


Fig. 1 Bhakra Dam's location

Table 1 Characteristics of Bhakra Nagal Dam and Reservoir

Item	Description
Catchment area	56,980 square kilometers
Dead storage level	EL. 512.06 m (EL.1680 feet)
Dead storage level	EL.445.62 m
New area irrigated	60 lakh acres
Area of the reservoir	162.48 square kilometers (62.78 sq. miles)
Length of the reservoir	96.56 km
Live storage capacity at EL.1680 ft	6911 million cum (5.60 MAF)
Gross storage capacity at EL.1680 ft	9340 million cum (7.57 MAF)
Dead storage capacity	2430 million cum (1.97 MAF)

reservoir storage (BCM), the previous inflow of reservoir (MCM), the current inflow of reservoir (MCM), the evaporation of reservoir (MCM), the previous outflow of the reservoir (MCM), and time (months) and release of the reservoir. All the data were acquired from the following websites: "UK Centre for Ecology and Hydrology", "Bhakra Beas Management Board" and "India Meteorological Department". The range of the reservoir's water level is determined by the hydraulic features of the Bhakra dam, with the maximum water level at 512.06 m and the minimum operating level at 450.45 m. Table 2 shows the essential statistical properties of the inputs, such as the minimum, maximum and total count values.

Table 2 Data acquired with descriptive statistics

Input	Unit	Minimum	Maximum	Average	Count
Inflow	MCM	352.04	10,267.8	1764.88	372
Reservoir level	Meter	450.45	512.06	489.01	372
Monthly reservoir storage	BCM	0.2	6.23	3.13	372
Evaporation	MCM	− 28.1	41.2	− 6.8	372
Previous inflow	MCM	352.04	10,267.8	1764.62	372
Previous outflow	MCM	402.3	8942.3	1819.95	372

One of the tasks during modelling nonlinear hydrological processes was to select the most significant variables from the whole set of input variables (Bahrami & Wigand, 2018; Hu & Wan, 2009). The major goal of data collection in this study is to choose appropriate input variables, depending on the data available. Also known as feature selection, the choice of the best subset of the inputs in the model was made based on certain defined governing rules (Sharafati et al., 2019), so as to increase the model's accuracy and efficiency. Therefore, during the modelling phase of this study, various combinations of input variables were used. For this study, five scenarios are initially defined at different folds, as shown in Table 3, so as to find out the most effective output. Then, the prediction accuracy was evaluated for each scenario.

2.3 Support vector machine (SVM)

Support Vector Machine has gained popularity as a novel statistical learning method over recent two decades. Used for both classification and regression, it proves an efficient and reliable approach (Collobert & Bengio, 2001; Drucker et al., 1996; Vapnik, 1995). Unlike the traditional chaotic methods, the SVM method is based on the idea of mapping input data into a high-dimensional feature space, so as to help with classification and simulate unknown relationships between the set of input variables and the set of output variables. Based on the mechanism's simplicity, two advantages of this method are that it is sufficiently known by scientists and that it dominates prediction. It has a level of precision that sets it apart from several other approaches. SVM is a strategy that uses a kernel trick to understand an issue, while simultaneously lowering the complexity and prediction errors of models. SVM classification is the first step in making a decision limitation for the feature space, which is delivered by generating an ideal separation hyperplane between two classes, so as to maximize the margin by minimizing the generalization error. In theory, SVM classification has the potential to predict outcomes, which can be comprehended with three essential concepts: (1) Function of the kernel; (2) The soft-margin; and (3) The separation hyperplane (Cristianini & Shawe-Taylor, 2000; Schwefel, 1981). Polynomial, radial basis and sigmoid, functions are exemplary kernel functions. Algorithms like SVM are mostly used to forecast classification problems and support vector regression (SVR) is an expansion of SVM by adding an insensitive loss function, so that it can be used in regression analysis (Drucker et al., 1996; Kim et al., 2012). In other words, in a classification problem, SVM is utilised to partition data into "+1" and "− 1" classes. On the other hand, SVR is a generalized SVM approach to predict random real values (Basak et al., 2007; Gunn, 1988). To improve forecasting of reservoir inflows, a modified SVM-based prediction system was created (Li et al., 2010). Climatic data from previous time periods were used, in addition to highly connected climate precursors. To understand non-linear patterns underlying climatic systems more flexibly, SVM parameters were determined in a genetic algorithm-based parameter determination approach. The median of forecasts from the created models was then used to reduce the variation in the prediction by using bagging to construct several SVM models. In terms of the predictive ability, the suggested modified SVM-based model outperformed a bagged multiple linear regression (MLR), a simple SVM, and a simple MLR model.

Regression with an alternate loss function is an example of SVM. Loss functions are frequently used in estimation, model selection, and prediction; and they are critical in determining any disparities between the null and nonparametric models' fitted values (Hong & Lee, 2009). In terms of hydrology, researchers must consider loss functions when making predictions. In this study, a hydrologic loss function is used to link two primary variables: rainfall and runoff. A distance measure must be supplied, which necessitates a change in the loss function (Smola & Scholkopf, 2004). SVR's main notion is to nonlinearly translate the initial data into a higher-dimensional feature space, so as to solve the linear regression issue (Fig. 2). As a result, SVR is usually required to construct a suitable function $f(x)$ to reflect the non-linear relationship between feature $\mathrm{x_i}$ and target value y_i, as demonstrated in Eq. (4).

$$f(x_i) = w\varphi(x_i) + b\|w\|^2 \tag{1}$$

where w denotes the coefficient vector, $\varphi\ (x_i)$ denotes the transformation function, and w and b denote the weight

Table 3 The selected scenarios for input combinations

Number	Different input combinations	Output
1	Inflow, Evaporation	Output
2	Inflow, Reservoir level, Evaporation	Output
3	Inflow, Reservoir level, Evaporation, Monthly reservoir storage	Output
4	Inflow, Reservoir level, Evaporation, Monthly reservoir storage, Previous inflow	Output
5	Inflow, Reservoir level, Evaporation, Monthly reservoir storage, Previous inflow, Previous outflow	Output

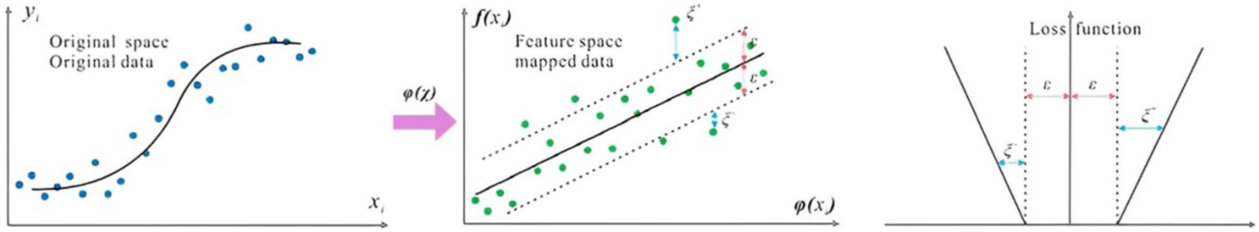


Fig. 2 Schematic diagram of SVR (Zhang et al., 2018)

and bias, respectively, which are calculated by minimising the so-called regularised risk function, as shown in Eq. (5).

$$\mathrm{R(w)} = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n} L_\varepsilon(y_i, \mathrm{f}(x_i)) \tag{2}$$

where $\frac{1}{2}\|w\|^2$ is the regularization term; c is the penalty coefficient; and $L\varepsilon(y_i, \mathrm{f}(x_i))$ is the ε-insensitive loss function, which is calculated according to Eq. (6).

$$L\varepsilon\left(y_i, \mathrm{f}(x_i)\right) = \max\{0, \left|y_i - \mathrm{f}(x_i)\right| - \varepsilon\} \tag{3}$$

where ε signifies the allowed error threshold, which will be ignored if the projected value is within the scope of the threshold; otherwise, the loss will equal to a number greater than ε.

To find out the optimization boundary, two slack factors, ξ^+ and ξ^-, are introduced:

$$\min f\left(\mathrm{w}, \xi^+, \xi^-\right) = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}(\xi^+, \xi^-) \tag{4}$$

Subject to

$$y_i - [\mathrm{w}.\varphi(x_i)] - \mathrm{b} \leq \varepsilon + \xi^-, \xi^- \geq 0$$

$$[\mathrm{w}.\varphi(x_i)] + \mathrm{b} - y_i \leq \varepsilon + \xi^+, \xi^+ \geq 0 \tag{5}$$

The minimization of a Lagrange function, which is formed from the objective function and the problem constraints, yields the dual version of this optimization problem:

$$\begin{aligned}\max\alpha, \alpha^* =& \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha^*{}_i), (\alpha_j - \alpha^*{}_j)\mathrm{K}(x_i, x_j) \\ &- \varepsilon\sum_{i=1}^{N}\left(\alpha_i + \alpha^*{}_i\right) + \varepsilon\sum_{i=1}^{N} y_i\left(\alpha_i - \alpha^*{}_i\right),\end{aligned} \tag{6}$$

$$\begin{aligned}\text{s.t}\quad & \sum_{i=1}^{l}(\alpha_i - \alpha^*{}_i) = 0, \quad \mathrm{i} = 1 \text{ to } N, \\ & \alpha_i, \alpha^*{}_i \geq 0 \quad i = 1 \text{ to } N, \\ & -\alpha_i, -\alpha^*{}_i - \geq -C \quad i = 1\text{to } N.\end{aligned} \tag{7}$$

The inner product$\{\varphi(x_i), \varphi\left(x_j\right)\}$ in the feature space is denoted by the function *K*(*x i, x j*) in the dual formulation of the issue.

Any function $K(x_i, x_j)$ can become a kernel function, if it satisfies the inner product criteria. Hence, the regression function can be expressed as follows:

$$f(\mathrm{x}) = \sum_{i=1}^{N}(\alpha_i - \alpha^*{}_i)\mathrm{K}(x_i, x_j) + \mathrm{b} \tag{8}$$

2.4 Regression tree (RT)

As a machine-learning algorithm for building prediction models from datasets, Regression Tree employs a clustering tree with post-pruning processing. The clustering tree algorithm is often referred to as the forecasting clustering tree and the monothetic clustering tree (Chavent, 1988; Vens et al., 2010). Regression Tree is used for model-dependent variables having a finite number of values which are not arranged in order, with prediction errors commonly assessed as the squared difference between the predicted and observed values (Loh, 2011). Clustering tree algorithms are based on the top-down induction technique of decision trees (Quinlan, 1986). Regression Tree algorithms take a collection of data for training and create a new internal node as good as possible. Based on the decreased variance, such algorithms choose the top test scores. The lower the variance, the more homogeneous the cluster and the more accurate the forecast. If none of the tests significantly reduce variance, a leaf will be generated and marked as data representative (Chavent, 1988; Vens et al., 2010). By recursively splitting the data space and fitting a prediction model within each partition, a hierarchical tree-like division of the input space can be created (Breiman, 2017). A sequence of recursive splits divides the input space into local regions, which are designated by a series of recursive splits. Internal decision nodes and terminal leaves make up the tree finally. Starting at the root node, a sequence of tests and decision nodes will determine the path through the tree, till it approaches a terminal node, providing a test data point. A prediction is made at the terminal node based on the model linked to that node locally.

2.5 K-fold cross-validation

The holdout approach can be employed when the amount of data available for training and testing are limited. In this approach, a subset of data is saved for validation, while the rest is for training. It is a common practice in engineering to keep one-third data for validation and utilize the other two-thirds for training and testing (Witten & Frank, 2000). By dividing the obtained data into a specified number of equally sized observations or folds (k), the holdout approach can be further improved. The dataset used for testing is chosen from these (k) folds, whereas the rest (k-1) is employed in the training process. This procedure will be repeated for k times, with a different fold being tested in each time and the remaining folds (k-1) serving as the training dataset. As a result, the approach would generate k different degrees of accuracy. The variance of the resulting estimate diminishes as (k) will be increased. Consider a fivefold cross-validation scenario ($k=5$). Figure 3 shows how the dataset is divided into five folds. The first fold is used to test the model, while the others are used to train the model in the first iteration. Then, the second iteration uses the second fold as the testing set and the rest as the training set. This procedure will be repeated until each of the five folds has served as a testing set.

3 Results and discussions

The best value of each grouping would result in the most precise estimating model; and later, the unsurpassed grouping combination would be chosen. The data were divided into two categories: Regression tree and SVM analysis. Following this process, two models were employed to make data projections based on five diverse scenarios. The most appropriate and exact prediction scenario was determined by each model's best estimation. To decide whether a model was best, the estimating power of both models was examined. To evaluate the suggested model's execution in varied in preparing, checking and testing information, three types of measurable assessments were used, i.e., RMSE, MAE and R^2. Figures 4 and 5 show the comparison of the observed values with the predicted values of outflows by using SVM and RT models for 5-, 10- and 20-fold cross-validation. It can be clearly seen from the figures that the predicted values are much closer to the observed ones for Scenario 5 with tenfold cross-validation using SVM model. Figures 6 and 7 depict the residuals obtained by using SVM and RT models for different cross-validation conditions. Residuals obtained are the minimum for SVM models as compared to RT models and are best for tenfold cross-validation condition. Table 4 contrasts the calculations statistically for the model SVM computed by using fivefold cross-validation. Data observed from the table clearly show that Scenario 5 relates to the lowest RMSE, MAE values and a maximum R^2 value amid all the situations. Moreover, data forecasting in Scenario 5 offers the most accurate results, while Scenario 4 provides the second lowest result. Besides, Scenario 3 has erratic forecasts, compared to all other scenarios. The lowest value for authentication oversights RMSE is also held by Scenario 5. Continuous development can be seen in the results from Scenarios 1–2 in SVM Regression. A big inaccuracy from Scenarios 2 to 3 of the SVM was recorded in the values of RMSE; MAE would increase, while the coefficient of correlation decreased from 0.87 to 0.85.

Table 5 compares and contrasts the statistical evaluations for the SVM model using tenfold cross-validation. It can be observed that Scenario 5 for SVM using tenfold cross-validation does necessarily produce better results than fivefold cross-validation. The RMSE and

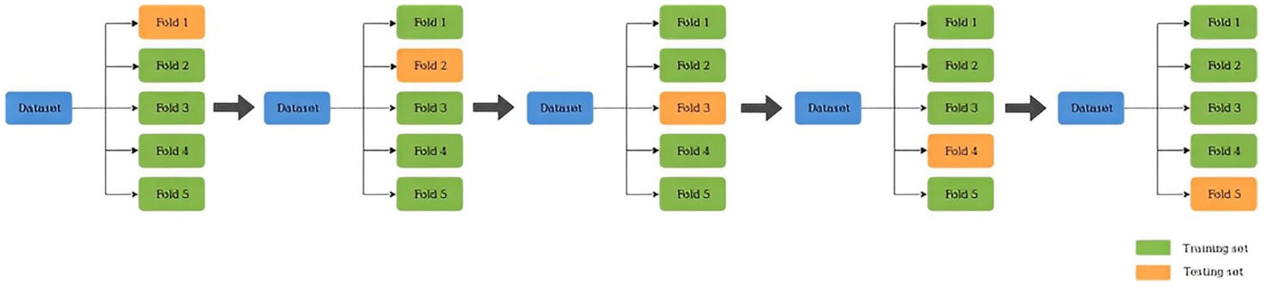
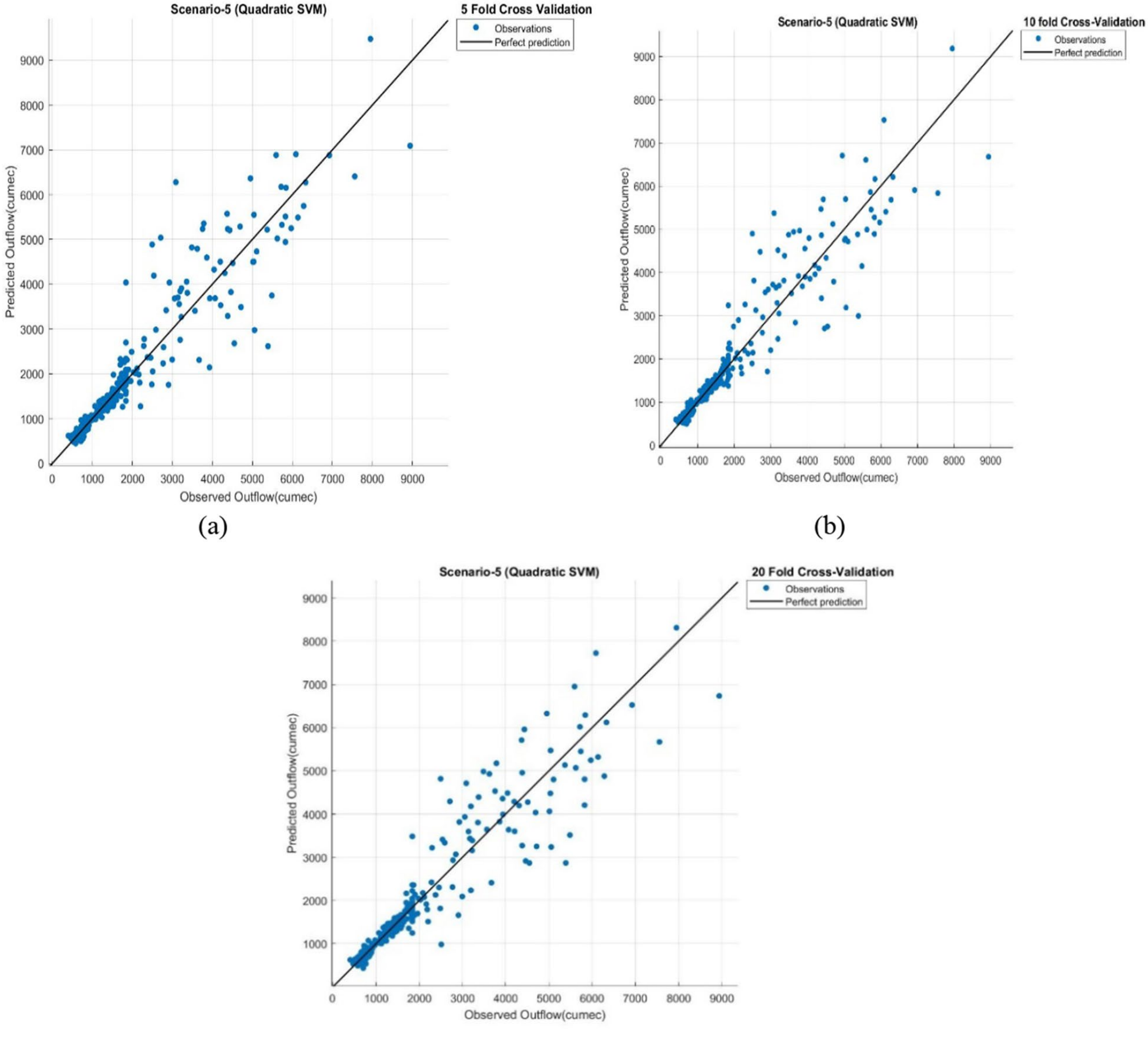


Fig. 3 Cross-validation in different folds



(a)

(b)

(c)

Fig. 4 Relationship between the observed outflow and the predicted outflow by using SVM, fivefold (**a**), tenfold (**b**) and 20-fold (**c**)

MAE values obtained with tenfold cross-validation seem to be more accurate than those obtained with fivefold cross-validation, as shown in Table 4. On the other hand, R^2 has the same value as fivefold cross-validation. Except for the R square value, i.e., 0.9, tenfold cross-validation yields better results than fivefold cross-validation for all parameters, which is the same as fivefold cross-validation. In the same way, Table 6 shows a comparison of the statistical assessments for the SVM model using 20-fold cross-validation, purposed to determine whether a greater cross-validation number may minimise the predicting errors. It shows that SVM model Scenario 5, which uses 20-fold cross-validation, may not always deliver better results than fivefold and tenfold cross-validation. Scenario fivefold and tenfold cross-validation yields lower RMSE values than 20-fold cross-validation. Fivefold and tenfold cross-validation has R square values closer to 1 as compared to 20-fold cross-validation. Table 7 displays the results of fivefold cross-validation using regression tree models under various assessment criteria. It can be observed that all of the scenarios have extremely strong prediction ability ($R^2 > 0.77$), according to the statistical assessment standards of this study. Scenario

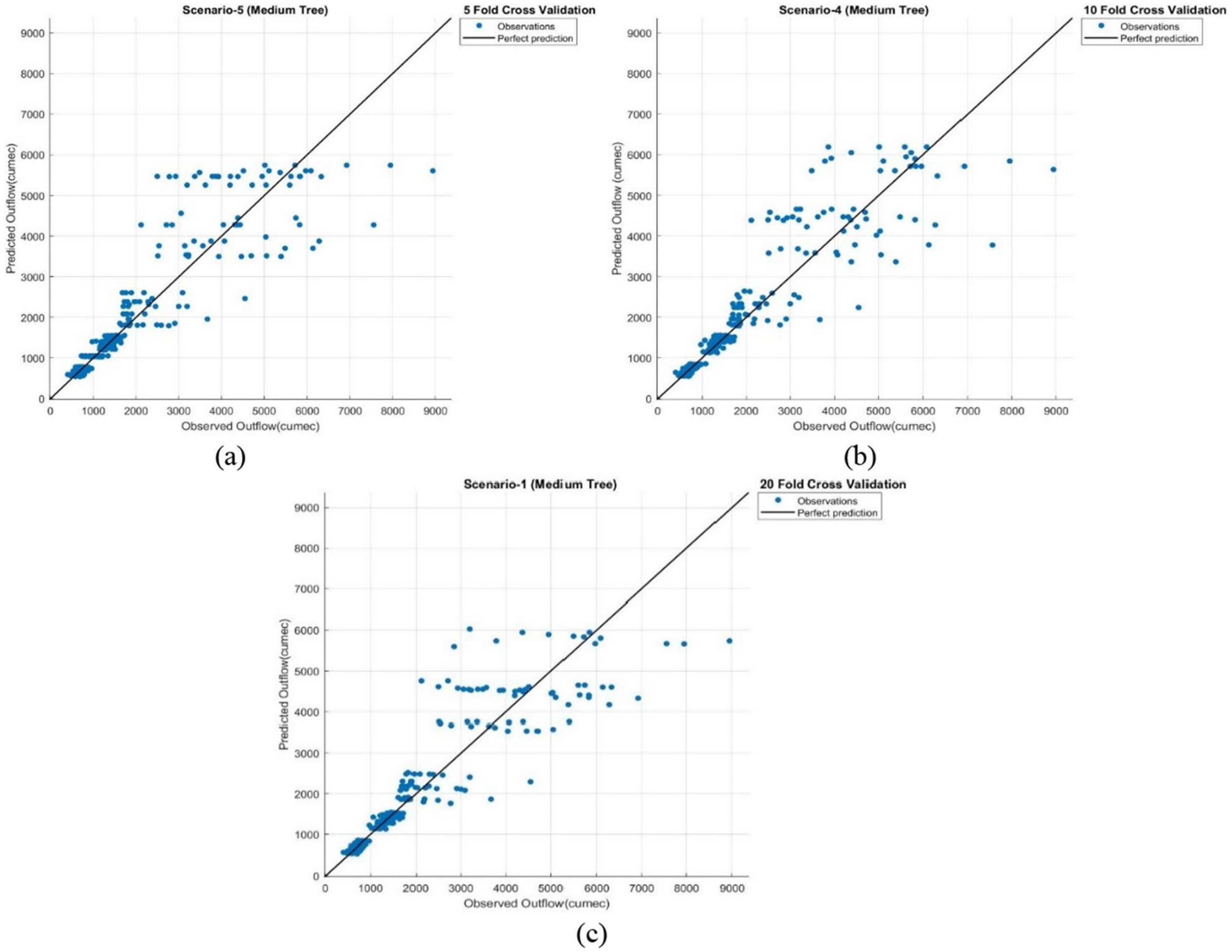
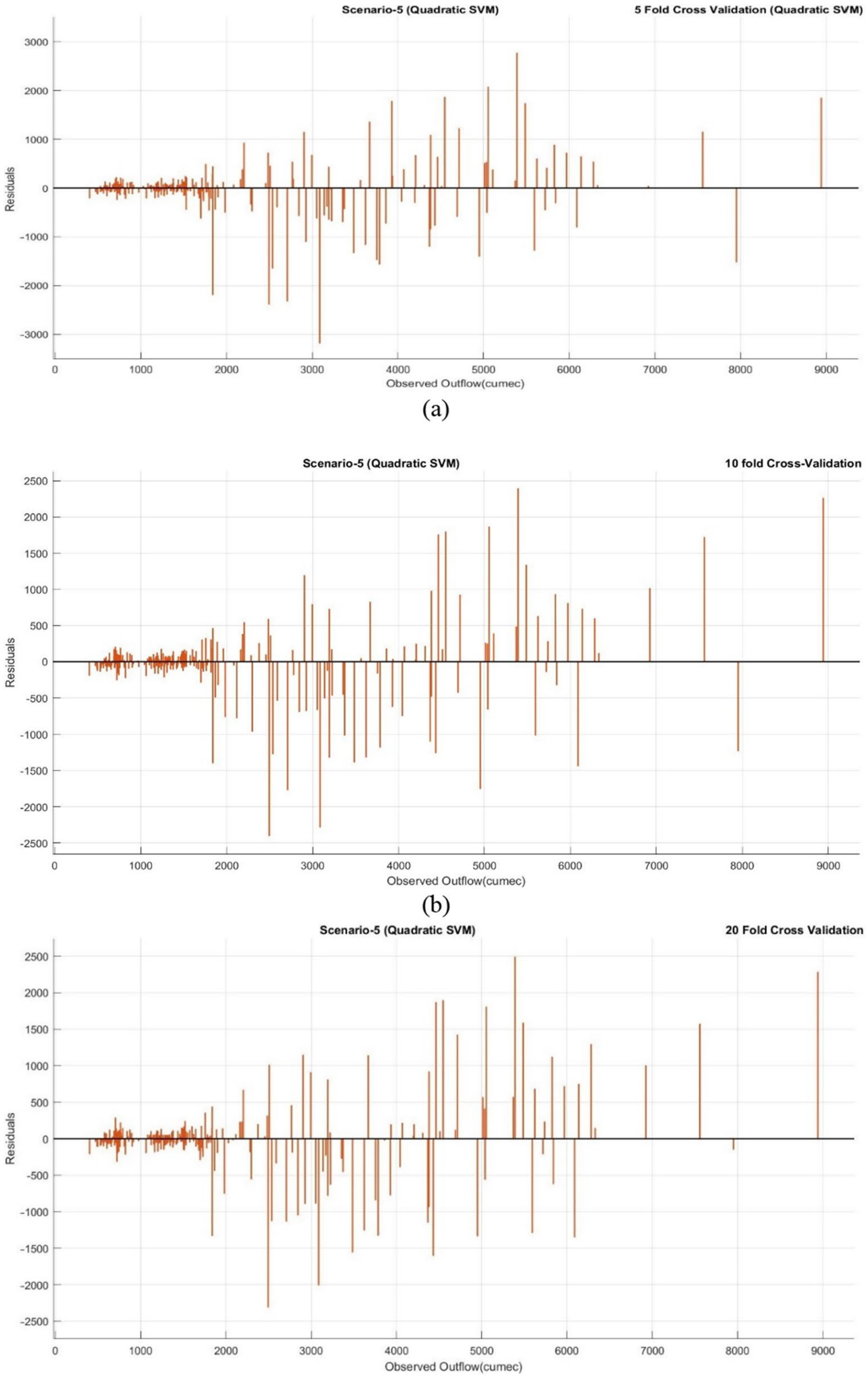


Fig. 5 Relationship between the observed outflow and the predicted outflow by using RT, fivefold (**a**), tenfold (**b**) and 20-fold (**c**)

3 achieves the best result, since it has the highest R^2 value (0.82), followed by Scenarios 2, 4, 1 and 5. In terms of RMSE, Scenario 5 gives the best predictive power (602.8), followed by Scenarios 3, 4, 2 and 1. Table 8 shows that Scenario 4 of Regression tree model using tenfold cross-validation has very good predictive ability, since it gives the best R^2 value (0.85), the lowest RMSE (557.48) and the lowest MAE (270.23), followed by Scenario 2. As clearly seen in Table 9, the results of the Regression tree of Scenario 1 using 20-fold cross-validation are overall better than those for both fivefold and tenfold cross-validation. The computed RMSE values are higher than tenfold cross-validation. Table 10 compares the predicted outcomes based on two distinct AI models, i.e., Regression Tree and Support Vector Machine, united with various parameters of the model. The findings reveal that the SVM model with tenfold cross-validation [RMSE (452.17), R^2 (0.9)] performs the best when compared to other SVM and RT models.

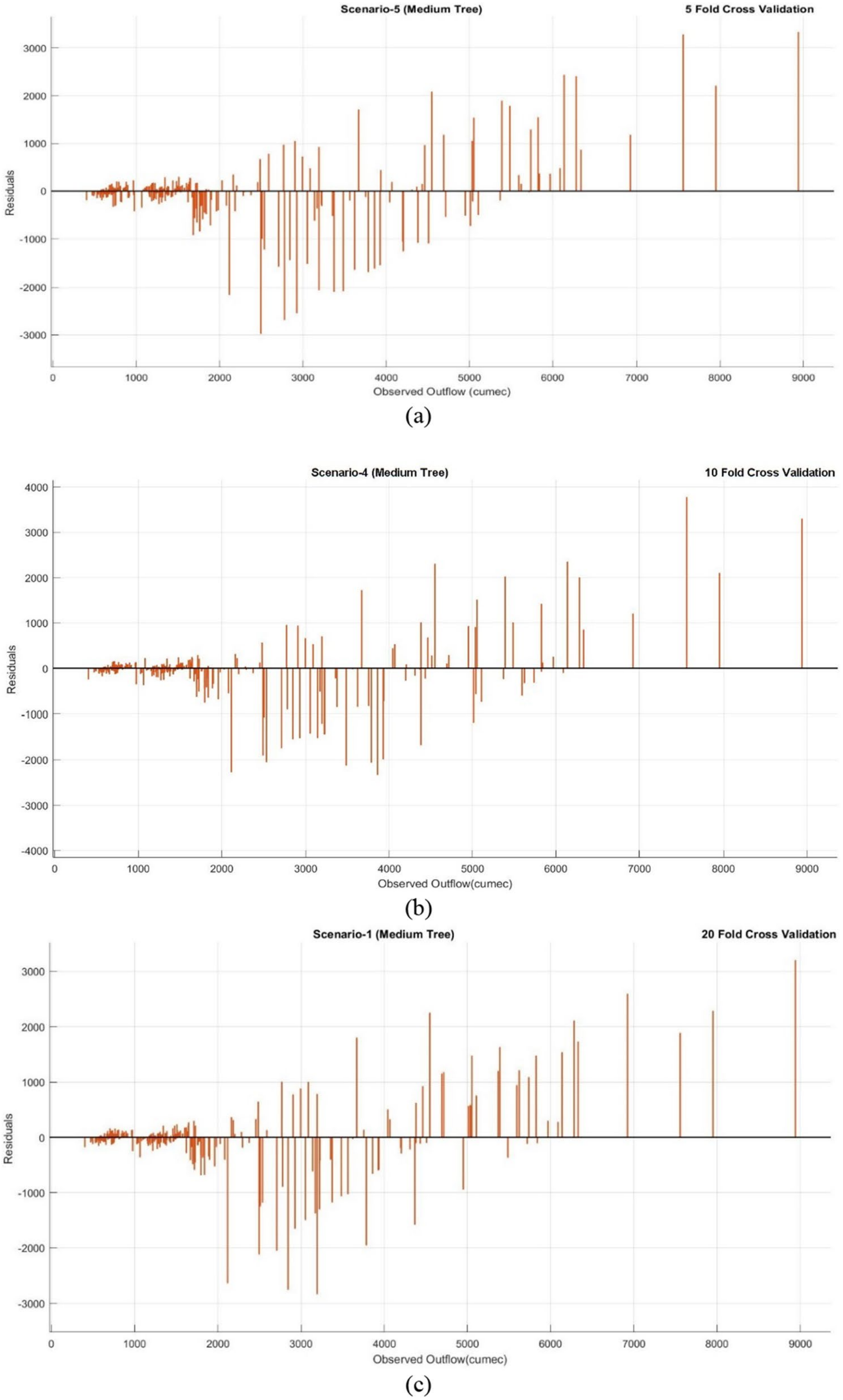
4 Conclusion

Over the past decades, traditional hydrological forecasting models have greatly changed, with SVM taking the prominence, because it can offer accurate data forecasts for a variety of hydrological processes. The ability to accurately estimate changes in reservoir water levels is beneficial for the planning and management of reservoir water usage in the long run. By examining two distinct Machine Learning approaches, i.e., Regression Tree and Support Vector Machine, this study tries to find which one is the most accurate in predicting water levels based on monthly hydrological records collected in the past 30 years, so as to simulate reservoir outflows. To get the best parameters, this study evaluated a variety of scenarios based on a variety of data



(c)

Fig. 6 Residue plots for the SVM model at the monthly scale, fivefold (**a**), tenfold (**b**) and 20-fold (**c**)



(a)

(b)

(c)

Fig. 7 Residue plots for the RT model at monthly scale, fivefold (**a**), tenfold (**b**) and 20-fold (**c**)

Table 4 Statistical evaluation of support vector machine for fivefold cross-validation

Cross-validation		Kernel function	RMSE	R^2	MAE
5-FOLD	Scenario 1	QUADRATIC	524.26	0.87	245.53
	Scenario 2	QUADRATIC	518.02	0.87	241.24
	Scenario 3	QUADRATIC	570.40	0.85	268.98
	Scenario 4	QUADRATIC	483.79	0.89	231.79
	Scenario 5	QUADRATIC	465.81	0.90	232.66

Table 5 Statistical evaluation of support vector machine for tenfold cross-validation

Cross-validation		Kernel function	RMSE	R^2	MAE
10-FOLD	Scenario 1	QUADRATIC	544.16	0.86	255.37
	Scenario 2	QUADRATIC	523.53	0.87	246.41
	Scenario 3	QUADRATIC	496.54	0.88	238.37
	Scenario 4	QUADRATIC	483.46	0.89	234.78
	Scenario 5	QUADRATIC	452.17	0.90	219.68

Table 6 Statistical calculation of Support Vector Machine for 20-fold cross-validation

Cross-validation		Kernel function	RMSE	R^2	MAE
20-FOLD	Scenario 1	QUADRATIC	545.27	0.86	254.21
	Scenario 2	QUADRATIC	527.27	0.87	246.36
	Scenario 3	QUADRATIC	513.03	0.87	247.30
	Scenario 4	QUADRATIC	487.11	0.89	235.82
	Scenario 5	QUADRATIC	471.98	0.89	226.78

Table 7 Statistical calculation of Regression tree for fivefold cross-validation

Cross-validation		Kernel function	RMSE	R^2	MAE
5-FOLD	Scenario 1	FINE	631.57	0.81	294.51
	Scenario 2	MEDIUM	633.14	0.81	289.25
	Scenario 3	FINE	610.93	0.82	282.70
	Scenario 4	FINE	623.00	0.81	292.39
	Scenario 5	MEDIUM	602.80	0.78	322.53

Table 8 Statistical calculation of regression tree for tenfold cross-validation

Cross-validation		Kernel function	RMSE	R^2	MAE
10-FOLD	Scenario 1	MEDIUM	602.59	0.83	288.27
	Scenario 2	FINE	625.43	0.81	297.57
	Scenario 3	COARSE	645.52	0.80	325.17
	Scenario 4	MEDIUM	596.34	0.84	274.07
	Scenario 5	FINE	613.99	0.82	284.99

Table 9 Statistical calculation of regression tree for 20-fold cross-validation

Cross-validation		Kernel function	RMSE	R^2	MAE
20-FOLD	Scenario 1	MEDIUM	596.2	0.83	279.73
	Scenario 2	MEDIUM	622.46	0.82	294.46
	Scenario 3	FINE	622.17	0.82	294.26
	Scenario 4	FINE	615.97	0.82	291.07
	Scenario 5	COARSE	629.35	0.81	309.94

Table 10 Statistical calculation of regression tree model and support vector machine model

Cross-validation	Model	RMSE	R^2	MAE
5-FOLD	SVM	465.81	0.9	232.66
	Regression tree	610.93	0.82	282.7
10-FOLD	SVM	452.17	0.9	219.68
	Regression tree	596.34	0.84	274.07
20-FOLD	SVM	471.98	0.89	226.78
	Regression tree	596.2	0.83	279.73

inputs. For this purpose, RMSE, MAE and R^2 indices are used to quantify the performance of the forecasting models. In summary, Scenario 5 shows the optimum combination of input data, which comprise inflow, evaporation, water level, reservoir storage, previous inflow and previous outflow. The best SVM regression is with quadratic kernel function, and the best V-fold cross-validation is tenfold, which is employed for the optimal scenario selection. In the comparative analysis of water level prediction by the two algorithms, SVM is proven to be the best algorithms for water level prediction. However, when performing water level prediction individually, Regression tree with tenfold cross-validation shows that the SVM model can make accurate predictions. This highlights its unique capabilities and benefits in detecting hydrological time series with non-linear properties. Therefore, SVM has certain generality and can be used as a model for reservoir water level prediction. More kind of hydrological data, such as infiltration rates, transpiration rates, low inflow conditions and other relevant parameters, should be added in future studies, so as to deliver more precise forecasts.

Acknowledgements

The authors would like to express their deepest appreciation to the anonymous reviewers for their time spent on this paper.

Author contributions
Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The data used in this study is available in digital platform.

Declarations

Competing interests
The authors reported no potential conflicts of interest related to the work submitted for publication.

Received: 19 October 2022 Revised: 12 November 2022 Accepted: 21 March 2023
Published online: 28 April 2023

References

Ahmadlou, M., Karimi, M., Alizadeh, S., Shirzadi, A., Parvinnejhad, D., Shahabi, H., & Panahi, M. (2019). Flood susceptibility assessment using integration of adaptive network-based fuzzy inference system (anfis) and biogeography-based optimization (bbo) and bat algorithms (ba). *Geocarto International, 34*, 1252–1272.

Bahrami, S., & Wigand, E. (2018). Daily streamflow forecasting using nonlinear echo state network. *International Journal Advance Research Science Engineering Technology, 5*, 3619–3625.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews, 11*, 203–224.

Bowden, G. J., Maier, H. R., & Dandy, G. C. (2002). Optimal division of data for neural network models in water resources applications. *Water resources research, 38*(2), 2–1.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Chavent, M. A. (1988). Monothetic clustering method. *Pattern Recognition Letters, 19*, 989–996.

Collobert, R., & Bengio, S. (2001). SVM Torch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research, 1*, 143–160.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other Kernel-based learning methods*. Cambridge University Press.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V. (1996). Support vector regression machines. *Advance in Neural Information Processing System* 155–161.

Gunn, S. R. (1988). Support vector machines for classification and regression. *ISIS Technical Report, 14*, 5–16.

Hejazi, M. I., & Cai, X. M. (2009). Input variable selection for water resources systems using a modified minimum redundancy maximum relevance algorithm. *Advances in Water Resources, 32*(4), 582–593.

Hipni, A., El-shafie, A., Najah, A., Karim, O. A., Hussain, A., & Mukhlisin, M. (2013). Daily forecasting of dam water levels: Comparing a Support Vector Machine (SVM) Model with Adaptive Neuro Fuzzy Inference System (ANFIS). *Water Resour. Manag., 27*, 3803–3823.

Hong, Y., Lee, Y. J. (2009). Citing Websites. A loss function approach to model specification testing and its relative efficiency to the GLR test. Retrieved 3rd February 2013.

Hu, C., Wan, F. (2009). Input Selection in Learning Systems: A Brief Review of Some Important Issues and Recent Developments. *In Proceedings of the IEEE International Conference on Fuzzy Systems*, Jeju Island, Korea, 20–24 August, pp. 530–535.

Hussain, H. W., Ishak, W., Ku-mahamud, K. R., & Norwawi, N. (2011). Neural network application in reservoir water level forecasting and release decision. *International Journal of New Computer Architectures and Their Applications, 1*, 265–274.

Khan, M. S., & Coulibaly, P. (2006). Application of support vector machine in lake water level prediction. *Journal of Hydrologic Engineering, 11*(3), 199–205.

Kim, S. J., Ryoo, E. C., Jung, M. K., Kim, J. K., & Ahn, H. C. (2012). Application of support vector regression for improving the performance of the emotion prediction model. *Journal of Intelligent Information System, 18*, 185–202.

Li, P. H., Kwon, H. H., Sun, L., Lall, U., & Kao, J. J. (2010). A modified support vector machine based prediction model on streamflow at the Shihmen Reservoir, Taiwan. *International Journal of Climatology, 30*, 1256–1268.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery., 1*(1), 14–23.

Mokhtar, A. S., Wan Ishak, W. H., Md Norwawi, N. (2014). Modelling of reservoir water release decision using neural network and temporal pattern of reservoir water level. *In: Proceedings of the fifth international conference on intelligent systems, modelling and simulation*, pp 127–130.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspective, 31*, 87–106.

Naghibi, S. A., & Pourghasemi, H. R. (2015). A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. *Water Resources Management., 29*, 5217–5236.

Najah, A. A., El-Shafie, A., Karim, O. A., & Jaafar, O. (2011). Integrated versus isolated scenario for prediction dissolved oxygen at progression of water quality monitoring stations. *Hydrology and Earth System Sciences Discuss, 8*, 6069–6112.

Nwobi-Okoye, C. C., & Igboanugo, A. C. (2013). Predicting water levels at Kainji dam using artifcial neural networks. *Nigerian Journal of Technology, 32*(1), 129–136.

Onur Hınçal, A., Burcu Altan-Sakarya, A., & Ger, M. (2011). Optimization of multireservoir systems by genetic algorithm. *Journal Water Resource Management, 25*, 1465–1487.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning, 1*, 1–81.

Schwefel, H. P. (1981). *Numerical optimization of computer models*. John Wiley & Sons Inc.

Seckin, N., Cobaner, M., Yurtal, R., & Haktanir, T. (2013). Comparison of artifcial neural network methods with l-moments for estimating food fow at ungauged sites. *Water Resour Manage J, 27*(7), 2103–2124.

Shafaei, M., & Kisi, O. (2016). Lake level forecasting using Wavelet-SVR, Wavelet-ANFIS and Wavelet-ARMA conjunction models. *Water Resources Management, 30*(1), 79–97.

Sharafati, A., Khosravi, K., Khosravinia, P., Ahmed, K., Salman, S. A., Yaseen, Z. M., & Shahid, S. (2019). The potential of novel data mining models for global solar radiation prediction. *International Journal of Environmental Science and Technology, 16*, 7147–7164.

Smola, A. J., & Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 1–24.

Tokar, A. S., & Markus, M. (2000). Precipitation-runoff modeling using artificial neural networks and conceptual models. *Journal of Hydrologic Engineering, 5*(2), 156–161.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Vens, C., Schietgat, L., Struyf, J., Blockeel, H., Kocev, D., & Džeroski, S. (2010). *Inductive databases and constraint-based data mining* (pp. 365–387). Springer.

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementations*. Morgan Kaufmann Publishers.

Yang, T., Gao, X., Sorooshian, S., & Li, X. (2016). Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled crossvalidation scheme. *Water Resources Research, 52*, 1626–1651.

Zhang, D., et al. (2018). Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. *Journal of Hydrology, 565*, 720–736.

Publisher's Note