Application of Hybrid Support Vector Machine model for Streamflow Prediction in Barak valley, India

A Sahoo¹, and D K Ghose²

¹Research Scholar, Department of Civil Engineering, National Institute of Technology Silchar, Fakiratilla, Silchar - 788010, Assam, India
²Assistant Professor, Department of Civil Engineering, National Institute of Technology Silchar, Fakiratilla, Silchar - 788010, Assam, India

Email: ¹bablusahoo1992@gmail.com; ²dillipghose_2002@yahoo.co.in

Abstract. Forecasting streamflow (Q<sub>flow</sub>) is vital in flood and water management, determining potential of river water flow, agricultural practices, hydropower generation, and environmental flow study. This research aims to explore capability of hybrid support vector machines (SVM) with Whale Optimisation Algorithm (WOA) model for forecasting streamflow at Badarpur Ghat gauging station of Barak river basin and evaluate its enactment with the conventional SVM model. Root mean squared error (RMSE), mean absolute error (MAE), and Nash-Sutcliffe efficiency (NSE) statistical measures are considered as evaluating standards. Assessment of outcomes indicates that the optimization algorithm could enhance the accurateness of standalone SVM model in monthly streamflow forecasting. Compared to conventional artificial intelligence methods without a data pre-processing system, the comparatively good performance of applied hybrid model gives an effective alternate to achieve better precision in streamflow forecasting. Results confirm that enhanced SVM model can better process a multifaceted hydrogeological data set, have higher prediction accuracy, and possess better generalisation capability.

Keywords: Barak river basin, Streamflow, SVM, SVM-WOA

1. Introduction

Precise streamflow forecasting is crucial for decision-makers related to water supplies, navigation, hydropower generation, flood control, and drought mitigation [1, 2, 3]. Nonetheless, reliable and accurate streamflow modeling is still challenging to attain because of the complex behavior of streamflow itself driven by its frenzied features like stochasticity, non-stationarity, and non-linearity [4, 5, 6]. Incorporating only appropriate input variables into a prediction model is crucial for ensuring high precision of predicted streamflow data. Yet, intrinsic non-linear relationship amid input and output variables in Q<sub>flow</sub> forecast is proving to be a continuing scientific test as traditional regression-based methods cannot accurately model streamflow data.

Models forecasting streamflow can be categorised into data-driven (DD) and physical-based models. Physical-based models give an idealised demonstration of hydrogeological developments by simulating complex relations amid meteorological, subsurface, and land surface constituents of water cycle [7].
contrast, DD models mathematically capture non-linear or linear interactions amid $Q_{\text{flow}}$ and its descriptive variables [8]. DD methods like ANN (artificial neural network), SVM, and ANFIS (adaptive neuro-fuzzy inference system), have been successfully applied by many investigators in hydrological field of study; for example, sediment transport modelling [9, 10, 11, 12], runoff modeling [13, 14, 15], water table depth prediction [16, 17, 18], streamflow forecasting [19, 20, 21]; pan evaporation estimation [22, 23]; flood prediction [24, 25, 26, 27, 28].

Zakaria and Shabri [29] explored ability of SVM in forecasting streamflow at ungauged stations and assessed its performance against MLR (multiple linear regression) models. Results revealed that SVM outperformed prediction capability of conventional MLR. Shabri and Suhartono [30] investigated potential of least-squares SVM model for improving accuracy of streamflow forecasting using recorded data from two stations of Kinta River in Perak, Malaysia. Kisi et al. [31] evaluated accurateness of ANN, SVM, and ANFIS models for daily intermittent streamflow forecasting considering data from Uzunkopru and Babaeski stations located in Thrace region of Turkey. Their findings revealed that ANN performed better in Uzunkopru station whereas ANFIS performed better in Babaeski station. Alizadeh et al. [32] assessed performance of feed-forward NN, recurrent NN, radial basis function NN, time delay NN, K-nearest neighbors (KNN), and SVM models to predict monthly flow of Alavian and Dez river basins located in Iran. Interest in developing hybrid streamflow forecasting models has grown in last two decades. Hybrid models combine different fitting techniques and decomposition approaches. Guo et al. [33] proposed an advanced SVM model using adaptive insensitive features to predict monthly streamflow and achieved outcomes were compared with ANN results and conventional SVM models. They found that improved SVM model processed complex hydrological data series better with high prediction accuracy. Yaseen et al. [34] treated non-stationary characteristics and uncertainty of streamflow data (Kelantan and Johor Rivers) utilising grey models (GM) and rolling mechanism (RM) pre-processing techniques before considering them as input. Again, Yaseen et al. [35] applied a hybrid ANFIS-FFA model to predict streamflow using recorded monthly $Q_{\text{flow}}$ data of River Pahang, Malaysia. Comparative analysis of results revealed that FFA was capable of improving forecasting accurateness of conventional ANFIS model. Zuo et al. [36] projected a hybrid VMD-LSTM (variational mode decomposition with long short-term memory) to predict 1–7 days ahead of daily streamflow for Han and Jing Rivers, China. Results indicated that hybrid VMD-LSTM model was more efficient and robust than other applied models. Pan et al. [37] proposed a new oil layer recognition model utilising semi-supervised SVM and improved WOA. Results showed that improved SVM-WOA had better recognition precision in oil layer recognition.

In current study, potential effectiveness of SVM-WOA method was compared with standard SVM model for evaluating prediction accuracy. The novelty lies in the use of SVM-WOA for streamflow prediction in Barak river basin, as this river basin receives flood water almost every year. Hence streamflow prediction becomes of utmost importance for planners and decision-makers of this study location.

2. Study area
The Barak Valley is collectively formed by Karimganj, Cachar, and Hailakandi districts falling amid 24°8’ and 25°8’ N, and 92°15’ and 93°15’ E, and covers an area of 6,922 sq. km (Figure 1). A warm and humid climate is experienced in the valley with an average annual precipitation of 2,440 - 4,100
mm, most of which is during May to September (southwest monsoon season), and average monthly minimum and maximum temperature of 9.2°C and 33.9°C. Barak is the region's main river, with its distributaries and tributaries. The Barak river traverses the valley in a westerly direction up to Karimganj supporting agricultural activities and giving food source to inhabitants. The valley has a rising and falling topography contributed by grasslands, floodplains, wetlands, hillocks, mountains, etc. [38].

![Figure 1. Location of Badarpur Ghat gauging station in Barak valley, Assam](image)

3. Material and Methodology

3.1. SVM

Vapnik and Chervonenkis [39] first introduced SVM, on basis of statistical learning theory. SVM is a group of supervised learning techniques utilised for regression and classification works governed by binary classification in random property space and hence, an appropriate technique [40]. A proficient learning system based on efficient optimisation theory, SVM applies structural risk minimization principle for attaining a general optimum solution.

A training dataset is used for determining $f(x)$ by computing $W$ (coefficient vector) and $b$ (bias term). After that, coefficients obtained are applied for minimizing error function as given in following equations:
\[
\begin{align*}
\text{Min} & \quad \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \\
\text{Subject to:} & \\
W \phi(x_i) + b - y_i & \leq \varepsilon + \xi_i^* \\
y_i - W \phi(x_i) + b & \leq \varepsilon + \xi_i \\
\xi_i, \xi_i^* & \geq 0, \quad i = 1, 2, \ldots, N
\end{align*}
\]

Existence of training error is unavoidable in training process. Hence, there must be a penalty term that is shown by C. \( N \) and \( \phi \) - number of samples, and kernel function, respectively. Furthermore, \( \xi_i, \xi_i^* \) - two slack variables indices. \( \varepsilon \) - loss function associated with training process accuracy.

### 3.2. WOA

An alternate swarm algorithm established on basis of natural behaviour of humpback whales is known as WOA [41]. In WOA, whales travel in a dimensional search space where number of parameters is represented by \( n \). Every whale course signifies solution of a candidate. Preliminary step for determining a global solution is by modification of every solution’s location. Following equation is used for updating location:

\[
\bar{X}_{t+1} = \bar{X}^*(t) - \bar{A}[\bar{C}, \bar{X}^*(t) - \bar{X}(t)]
\]

where \( \bar{X}(t) \) - solution vector in \( t \)th iteration; \( \bar{X}_{t+1} \) - solution vector in \( t+1 \)th iteration; \( \bar{X}^*(t) \) - probable location of prey in \( i \)th iteration; \( \bar{A} \) and \( \bar{C} \) - coefficient vectors. In every iteration, \( \bar{A} \) and \( \bar{C} \) are updated:

\[
\bar{A} = 2\bar{a}, \bar{r}_1 - \bar{a} \\
\bar{C} = 2, r_2
\]

where \( \bar{a} \) - reduces from 2 to 0; \( r_1 \) and \( r_2 \) - arbitrary vectors. By decreasing \( \bar{a} \) value, shrinking Encircling Mechanism (SEM) is executed. New location of a search tool can be described anywhere between range of instrument’s actual location and the present location of best agent by adjusting random values for \( \bar{A} \). Following equation is applied for defining spiral-shaped movement of whales:

\[
\bar{X}(t+1) = D', e^{bl}, \cos(2\pi l) + \bar{W}^*(t)
\]

where \( D' = \bar{W}^*(t) - \bar{X}(t) \), \( b \) - a constant value to identify logarithmic spiral shape, and \( k \) - an arbitrary number. The working procedure of hybrid SVM-WOA model is showcased in Figure 2.

### 3.3. Evaluating Constraint

The authors collected 20 years (2001-2020) of discharge data (\( Q_t \)) from CWC, Shillong whereas precipitation (\( P_t \)); average temperature (\( T_{avg}\)), and relative humidity (\( R_{h(t)} \)) are collected from IMD, Pune. 70% of collected data, i.e., 2001-2014, and rest 30%, i.e., 2015-2020 are considered to train and test proposed models. In present work, forecasting performance of established models (SVM and SVM-WOA) is
assessed utilising different quantitative statistical evaluation indices. Statistical indices utilised here are RMSE, NSE, and MAE. In mathematical terms, these indicators are expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} [T_{i,o} - T_{i,f}]^2}$$

$$\text{NSE} = \frac{\sum_{i=1}^{N} [T_{i,o} - T_{i,f}]^2}{\sum_{i=1}^{N} [T_{i,o} - \bar{T}_{i,o}]^2}$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |T_{i,o} - T_{i,f}|$$

where $T_{i,f}$ and $T_{i,o}$ = Predicted and collected discharge data; $\bar{T}_{i,f}$ and $\bar{T}_{i,o}$ = Average of predicted and collected discharge data; $N$ = data length.

Table 1. Modeling input combination structures

| Input combination | Model description | SVM          | SVM-WOA     |
|-------------------|-------------------|--------------|-------------|
| $Q_t$             | SVM1              | SVM-WOA 1    |             |
| $Q_t$, $P_t$      | SVM 2             | SVM-WOA 2    |             |
| $Q_t$, $P_t$, $T_{avg(t)}$ | SVM 3             | SVM-WOA 3    |             |
| $Q_t$, $P_t$, $T_{avg(t)}$, $R_{hi(t)}$ | SVM 4             | SVM-WOA 4    |             |

Figure 2. Flowchart of SVM-WOA algorithm
4. Result and Discussion
In present work, to forecast monthly $Q_{\text{flow}}$ values, two models, a hybrid SVM-WOA, and standalone SVM, were used in Badarpur Ghat gauging station located in Barak valley. Classification of input data was done with four parameters as shown in Table 1 and in this way four input combinations were created. For easier comparison and briefness, outcomes of applied models are given in Table 2 and Figures 3, 4 and 5. As stated in Table 2, NSE coefficients using SVM-WOA range between 0.9764-0.9863 in training phase, and 0.961-0.9748 in testing phase. Similarly for SVM, NSE values range between 0.9472-0.9536 in training and 0.9207-0.9302 in testing. Among all models, the best value is found for SVM4 (RMSE- 19.2213, NSE-0.9536, MAE-9.973) and SVM-WOA4 (RMSE-10.534, NSE-0.9863, MAE-5.028). From Figures 3 and 4 it is clear that the prediction values by the hybrid model are closer to the observed values, whereas for the basic model there is a considerable difference between the forecasted and observed time series, especially in peak values.

| Station Name | Model Name | RMSE Training | MAE | NSE | RMSE Testing | MAE | NSE |
|--------------|------------|---------------|-----|-----|--------------|-----|-----|
| Badarpur Ghat | SVM1       | 22.367        | 15.835 | 0.9472 | 27.9436 | 18.829 | 0.9207 |
|               | SVM2       | 21.5218       | 14.949 | 0.9451 | 26.754 | 16.747 | 0.9235 |
|               | SVM3       | 21.0574       | 12.054 | 0.9494 | 25.862 | 15.442 | 0.9284 |
|               | SVM4       | 19.2213       | 9.973 | 0.9536 | 24.615 | 12.204 | 0.9302 |
|               | SVM-WOA1   | 13.205        | 9.429 | 0.9764 | 18.261 | 11.521 | 0.961  |
|               | SVM-WOA2   | 11.5349       | 8.362 | 0.9805 | 17.395 | 10.67 | 0.9682 |
|               | SVM-WOA3   | 11.087        | 6.175 | 0.9831 | 16.3489 | 8.136 | 0.9706 |
|               | SVM-WOA4   | 10.534        | 5.028 | 0.9863 | 15.573 | 7.384 | 0.9748 |

Figures 3 and 4 illustrate the scatter plot and time series plot between forecasted and observed $Q_{\text{flow}}$ by SVM and SVM-WOA models for test period. The solid violet line in scatter plot exhibit regression line fitted to forecasted and observed $Q_{\text{flow}}$. The dotted black line is the 1:1 line. It is observed from Figure 3 that $R^2$ values generated by hybrid SVM-WOA4 model were better than values specified by standard SVM4. Considering time series plot (Fig. 4), it is evident that forecasted $Q_{\text{flow}}$ values by SVM-WOA4 model give a good match with observed time series. An important point to note here is that extreme and low events are of significant concern to water resources planners and decision-makers. This is because decision-makers require accurate predictions of these two events as it is very crucial for avoiding drought and flooding conditions.
Figure 3. Scatter plot for observed vs predicted streamflow by proposed models during test phase

Figure 4. Time series showing comparison of observed predicted monthly streamflow using the (a) SVM and (b) SVM-WOA models
Box-plots with corresponding distributions were generated for demonstrating how closely SVM-WOA model forecasts against original data series and SVM approach. In Figure 4, boxplots are utilised for indicating degree of overall extent in observed and forecasted data according to corresponding quartile whiskers and values. In present study, application of proposed soft computing models is not restricted to $Q_{\text{flow}}$ forecasting and can also be utilised for forecasting other hydrological variables such as temperature, precipitation, groundwater, etc.

5. Conclusion
Streamflow forecasting is significant for integrated water resources management, planning, and building river structures with other industrial processes. In this study, precision of classical SVM technique was enhanced by integrating SVM with WOA algorithm. Hybrid SVM model performance comparison with classical SVM model revealed the evolutionary algorithms’ ability to optimize SVM membership function to minimize prediction error. Outputs suggested that the hybrid SVM-WOA model had more precision in predicting streamflow than the SVM-based model. Forecasting $Q_{\text{flow}}$ achieved a comparatively good agreement with observed $Q_{\text{flow}}$, particularly the peak $Q_{\text{flow}}$. In future, other integrated intelligent systems may be explored for $Q_{\text{flow}}$ forecasting in the Barak river basin that may increase forecasting performance.

References
[1] Fang W, Huang Q, Huang S, Yang J, Meng E and Li Y 2017 Optimal sizing of utility-scale photovoltaic power generation complementarily operating with hydropower: A case study of the world’s largest hydro-photovoltaic plant Energy Convers. Manag. 136 161–72
[2] Yu L and Li Y P 2019 A flexible-possibilistic stochastic programming method for planning municipal-scale energy system through introducing renewable energies and electric vehicles J. Clean. Prod. 207 772–87
[3] Yaseen Z M, Kisi O and Demir V 2016 Enhancing long-term streamflow forecasting and predicting using periodicity data component: application of artificial intelligence Water Resour. Manag. 30 4125–51
[4] Narsimlu B, Gosain A K, Chahar B R, Singh S K and Srivastava P K 2015 SWAT model calibration and uncertainty analysis for streamflow prediction in the Kunwari River Basin, India, using sequential uncertainty fitting Environ. Process. 2 79–95
[5] Bayazit M 2015 Nonstationarity of hydrological records and recent trends in trend analysis: a state-of-the-art review Environ. Process. 2 527–42
[6] Ismail S, Samsudin R and Shabri A (2010) River flow forecasting: a hybrid model of self organizing maps and least square support vector machine. Hydrol Earth Syst Sci Discuss 7 8179–8212
[7] Rosenberg E A, Wood A W and Steinemann A C 2011 Statistical applications of physically based hydrologic models to seasonal streamflow forecasts Water Resour. Res. 47
[8] Hadi S J and Tombul M 2018 Streamflow forecasting using four wavelet transformation combinations approaches with data-driven models: a comparative study Water Resour. Manag. 32 4661–79
[9] Samantaray S and Sahoo A 2021 Prediction of suspended sediment concentration using hybrid SVM-WOA approaches Geocarto Int. 1–27
[10] Samantaray S, Sahoo A and Ghose D K 2020 Assessment of Sediment Load Concentration Using SVM, SVM-FFA and PSR-SVM-FFA in Arid Watershed, India: A Case Study KSCE J. Civ. Eng. 24 1944–1957
[11] Afan H A, Keshtegar B, Mohtar W H M W and El-Shafie A 2017 Harmonize input selection for sediment transport prediction. Journal of Hydrology, 552 366–375
[12] Rahman S A and Chakrabarty D 2020 Sediment transport modelling in an alluvial river with artificial neural network J. Hydrol. 588 125056
[13] Chang F-J, Chen P-A, Lu Y-R, Huang E and Chang K-Y 2014 Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control J. Hydrol. 517 836–46
[14] Samantaray S and Sahoo A 2020 Prediction of runoff using BPNN, FFBPNN, CFBBPNN algorithm in arid watershed: A case study Int. J. Knowledge-Based Intell. Eng. Syst. 24 243-251
[15] Van S P, Le H M, Thanh D V, Dang T D, Loc H H and Anh D T 2020 Deep learning convolutional neural network in rainfall–runoff modelling J. Hydroinformatics 22 541–61
[16] Luo Q, Wu J, Yang Y, Qian J, and Wu J, 2016 Multi-objective optimization of long-term groundwater monitoring network design using a probabilistic Pareto genetic algorithm under uncertainty. J. Hydrol. 534 352–363
[17] Samantaray S and Sahoo A 2021 Modelling response of infiltration loss toward water table depth using RBFN, RNN, ANFIS techniques. International Journal of Knowledge-based and Intelligent Engineering Systems, 25 227–234
[18] Yadav B, Gupta P K, Patidar N and Himanshu S K 2020 Ensemble modelling framework for groundwater level prediction in urban areas of India Sci. Total Environ. 712 135539
[19] Sahoo A, Samantaray S and Ghose D K 2019 Stream flow forecasting in mahanadi river basin using artificial neural networks Procedia Computer Science vol 157
[20] Samanataray S and Sahoo A 2021 A Comparative Study on Prediction of Monthly Streamflow Using Hybrid ANFIS-PSO Approaches KSCE J. Civ. Eng. 25 4032–43
[21] Dalkılıç H Y and Hashimi S A 2020 Prediction of daily streamflow using artificial neural networks (ANNs), wavelet neural networks (WNNs), and adaptive neuro-fuzzy inference system (ANFIS) models Water Supply 20 1396-1408
[22] Kisi O 2015 Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J. Hydrol. 528 312–320
[23] Wang L, Niu Z, Kisi O and Yu D 2017 Pan evaporation modeling using four different heuristic approaches. Computers and Electronics in Agriculture, 140 203-213
[24] Sahoo A, Samantaray S and Ghose D K 2021 Prediction of Flood in Barak River using Hybrid Machine Learning Approaches: A Case Study J. Geol. Soc. India 97 186–98
[25] Samantaray S, Sahoo A and Agnihotri A 2021 Assessment of Flood Frequency using Statistical and Hybrid Neural Network Method: Mahanadi River Basin, India J. Geol. Soc. India 97 867–80
[26] Agnihotri, A., Sahoo, A. and Diwakar, M.K., 2022. Flood Prediction Using Hybrid ANFIS-ACO Model: A Case Study. In Inventive Computation and Information Technologies (pp. 169-180). Springer, Singapore
[27] Chang F-J and Tsai M-J 2016 A nonlinear spatio-temporal lumping of radar rainfall for modeling multi-step-ahead inflow forecasts by data-driven techniques J. Hydrol. 535 256–69
[28] Xie S, Wu W, Mooser S, Wang Q J, Nathan R and Huang Y 2021 Artificial neural network based hybrid modeling approach for flood inundation modeling J. Hydrol. 592 125605
[29] Zakaria Z A and Shabri A 2012 Streamflow forecasting at ungaged sites using support vector machines Appl. Math. Sci. 6 3003–14
[30] Shabri A and Suhartono 2012 Streamflow forecasting using least-squares support vector machines Hydrol. Sci. J. 57 1275–93
[31] Kisi O, Nia A M, Gosheh M G, Tajabadi M R J and Ahmadi A 2012 Intermittent streamflow forecasting by using several data driven techniques Water Resour. Manag. 26 457–74
[32] Alizadeh Z, Yazdi J, Kim J H and Al-Shamiri A K 2018 Assessment of machine learning techniques for monthly flow prediction Water 10 1676
[33] Guo J, Zhou J, Qin H, Zou Q and Li Q 2011 Monthly streamflow forecasting based on improved support vector machine model Expert Syst. Appl. 38 13073–81
[34] Yaseen Z M, Fu M, Wang C, Mohtar W H M W, Deo R C and El-Shafie A 2018 Application of the hybrid artificial neural network coupled with rolling mechanism and grey model algorithms for streamflow forecasting over multiple time horizons Water Resour. Manag. 32 1883–99
[35] Yaseen Z M, Ebtehaj I, Bonakdari H, Deo R C, Mehr A D, Mohtar W H M W, Diop L, El-Shafie A and Singh V P 2017 Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model J. Hydrol. 554 263–76
[36] Zuo G, Luo J, Wang N, Lian Y and He X 2020 Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting J. Hydrol. 585 124776
[37] Pan Y, Xia K, Wang L and He Z 2021 A Novel Approach to Oil Layer Recognition Model Using Whale Optimization Algorithm and Semi-Supervised SVM Symmetry (Basel). 13 757
[38] Vapnik V 1995 The nature of statistical learning theory, Springer, New York.
[39] Pai P and Hong W 2007 A recurrent support vector regression model in rainfall forecasting Hydrol. Process. An Int. J. 21 819–27
[40] Mirjalili S and Lewis A 2016 The whale optimization algorithm Adv. Eng. Softw. 95 51–67
[41] Mazumder M K, Boro F, Barbhuiya B and Singha U 2014 A study of the winter congregation sites of the Gangetic River Dolphin in southern Assam, India, with reference to conservation Glob. Ecol. Conserv. 2 359–66