Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Hourly Flood Forecasting Using Hybrid Wavelet-SVM

Baheerah Shada[1*], N.R. Chithra[2], Santosh G. Thampi[3]

1. Post Graduate Student, National Institute of Technology, Calicut, India
2. Assistant Professor, Department of Civil Engineering, National Institute of Technology Calicut, India
3. Professor, Department of Civil Engineering, National Institute of Technology Calicut, India

Corresponding author: *baheerashada@gmail.com*

https://doi.org/10.22115/SCCE.2022.317761.1383

ARTICLE INFO

Article history:
Received: 01 December 2021
Revised: 02 March 2022
Accepted: 04 April 2022

Keywords:
Artificial neural networks;
Denoising;
Peak discharge;
Performance rating criteria;
Support vector machine.

ABSTRACT

The floods of 2018 and 2019 have underlined the urgent need for development and implementation of efficient and robust flood forecasting models for the major rivers in the State of Kerala, India. In this paper, the development and application of two hourly flood forecasting models are presented – one using Support Vector Machine (SVM) and the other based on hybrid wavelet-support vector machine (WSVM). The study was performed on the Achankovil River in Kerala. Wavelet technique was used to denoise the input signal (rainfall and water level) and the effective components of the input signal obtained after denoising were input to the SVM/ WSVM models for forecasting. These models' performance was assessed using standard performance rating criteria. Further, the performance of these models was compared with that of a flood forecasting model based on hybrid wavelet-artificial neural network (WANN) developed for this river in a previous study. Results of this study demonstrated the ability of the WSVM model to predict floods reasonably well. It was observed that the WSVM model performed better when compared to the WANN model. The WSVM model was able to accurately estimate peak discharge magnitude and time to peak, both of which are critical inputs in many water resource design and management applications.

How to cite this article: Shada B, Chithra NR, Thampi SG. Hourly flood forecasting using hybrid wavelet-SVM. J Soft Comput Civ Eng 2022;6(2):01–20. https://doi.org/10.22115/scce.2022.317761.1383

1. Introduction

The critical contribution of flood forecasting to reducing risk to life and property and minimising economic losses make studies for development of flood forecasting techniques/ models extremely important and relevant. Water-related disasters happen very frequently and are a major threat to human life and socio-economic development [1]. Complete control of floods is not possible due to several reasons such as topographical constraints, uncertainties associated with the timing, magnitude and place of occurrence of floods etc. So, rather than trying to completely control floods, appropriate measures to prevent damages caused by floods can be initiated if reliable and timely flood forecasts are available. Structural protection measures like dams and levees were traditionally employed for flood management. These structures reduce damages caused by floods by modifying its characteristics, say, by reducing the peak flood discharge and the corresponding river stage as well as the spatial extent of the area inundated. However, these cannot completely avoid floods. Hourly flood forecasting with adequate lead time is very effective and useful to minimise loss to life and property and damages caused by floods.

Many models have been proposed and used for flood forecasting in rivers all over the world [1]. Physical and conceptual models are, in general, data intensive in nature, making them difficult to implement in developing countries [2]. Computational simulations are becoming increasingly complex and time expensive in a variety of engineering challenges [3]. Conventional time series models were used for investigating the rainfall-runoff process in the previous two decades. The rainfall-runoff process being highly nonlinear and non-stationary in behaviour, these models find it difficult to capture the transformation satisfactorily. In many rainfall-runoff modelling studies, soft computing models like support vector machine and artificial neural networks have been used due to its ability to capture the nonlinear behaviour and flexibility in data [2]. Unlike the artificial neural networks (ANN) which reduce only the empirical risk associated, SVM helps to reduce the structural risk also [4]. SVMs are called “kernel machines” because it uses a kernel function for mapping the nonlinear function to a linear function. In SVM, training data is used to directly determine the decision boundaries. SVM is based on statistical learning theory and can minimise classification errors of the training data and the testing data [5]. Researchers have shown that the SVM approach helps in faster training when compared to ANN and ANFIS (Adaptive Neural-Fuzzy Inference System) models [5]. Also, the results obtained from the SVM models are reported to be more accurate when compared to those obtained using the ANN models [5]. A study used five surrogate models, namely, multiple regression, random forest, extreme gradient boosting, SVM and k-nearest neighbours to predict seismic vulnerability and environmental impacts of a class of buildings. The SVM was found to be the most accurate among these with respect to prediction of the total annual loss [6].

Even though SVM exhibits high flexibility in modelling hydrologic time series such as runoff, it does not handle non-stationary data very well. This problem can be overcome if the data is pre-processed [5]. Wavelet transform is a very efficient and popular technique for data pre-processing and is capable of dealing with non-stationary signals [7]. Signal denoising can be effectively performed with wavelet transforms. Many researchers reported that the capability of simple ANN and SVM models with regard to flood prediction can be considerably improved

when these are combined with wavelets [5], [8], [9]. In view of the above, it was felt that a model combining SVM with the wavelet technique would be promising for hourly flood forecasting applications.

The main objective of this study is to develop a flood forecasting model for the Achankovil river basin using SVM and hybrid wavelet-SVM, and to compare the performance of both these models as well as that of a hybrid wavelet-ANN model which had already been developed [10].

2. Research significance

Conceptual models based on physical laws provide a comprehensive description of hydrological processes. However, these models are computationally intensive and complicated and require a lot of data. A single data set like that of water level at a gauge site is not sufficient for calibration and testing of these models. Relative ease of developing and using models based on soft computing techniques and their satisfactory performance has resulted in the development and application of such models for diverse problems. Among these models, the SVM exhibits high flexibility in modelling hydrologic time series and with wavelet transforms, signal denoising can be effectively performed. Hence, it was felt that development of a model combining SVM with the wavelet technique would be quite promising for flood forecasting.

3. Theory

3.1. Support vector regression

Support Vector Machines (SVMs) which are used for classification as well as regression was introduced by Vapnik, a Russian mathematician in the early 1960s. SVMs are based on the Structural Risk Minimisation (SRM) principle which is an inductive principle that is commonly used in machine learning. As SVMs are based on SRM, it reduces the structural risk associated with the model and thus improves its generalization capability. SVM has been extensively used by researchers in various engineering fields including civil engineering, electronics and electrical engineering, mechanical engineering, financial, medical etc [11].

The basic idea of SVM classification is to divide the data points of different groups by a clear gap that is as wide as possible (maximum margin) using an optimal hyperplane for linearly separable patterns. For patterns that are not linearly separable, kernel mapping is used to change the data representation in the input space to a linearly separable form in a higher-dimensional space (feature space) and to fix the optimal hyperplane. Support vector machines can also be used in regression problems. Overall, support vector regression and support vector classification use the same principle but in regression, a margin of tolerance (epsilon) is set for approximation. Figure 1 shows the flow chart for basic SVM based regression.

The objective of SVM based regression is to estimate a functional relationship between a set of sampled values $\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots \dots . \boldsymbol{x_n}\}$ and desired values $\boldsymbol{y} = \{\boldsymbol{y_1}, \boldsymbol{y_2}, \dots \dots . \boldsymbol{y_n}\}$. The regression function is formulated as follows [12](Vapnik, 1995):

$$f(\boldsymbol{x}) = \big(\boldsymbol{w}.\Phi(\boldsymbol{x})\big) + \boldsymbol{b} \tag{1}$$

where $\boldsymbol{w}$ and $\boldsymbol{b}$ are the weight vector and bias terms which are the coefficients in this regression function and $\Phi(\boldsymbol{x})$ is a non-linear mapping function.

SVM based regression model uses a loss function known as ε-insensitivity loss function (L_ε) defined as:

$$L_\varepsilon(f(\boldsymbol{x}), \boldsymbol{y}) = \begin{cases} |f(\boldsymbol{x}) - \boldsymbol{y}| - \varepsilon & \text{for} \quad |f(\boldsymbol{x}) - \boldsymbol{y}| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\boldsymbol{y}$ is the desired output and ε defines the region of insensitivity [12].

In SVM regression, the problem is to find $f(x)$ that minimizes regularized risk function,

$$R_{reg} = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(f(\boldsymbol{x}), \boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{3}$$

where $\frac{1}{2}\|w\|^2$ is the regularization term and C is the regularization constant[12].

The non-linear regression function is a function that minimizes the regularized risk function subject to the loss function as [12]

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(\propto_i - \propto_i^*)K(\boldsymbol{x}, \boldsymbol{x_i}) + \boldsymbol{b} \tag{4}$$

where $\propto_i$, $\propto_i^*$ are Lagrangian multipliers and $K(\boldsymbol{x}, \boldsymbol{x_i})$ is the kernel function. The dimensionality of the input space can be changed using kernel functions to achieve a good regression model. Linear, polynomial, sigmoid and radial basis are some of the commonly used kernel functions [11]. Kernel functions act as a bridge from linearity to non-linearity for algorithms which can be expressed in terms of dot product. Linear kernel function is the simplest kernel function. If all the training data is normalized, polynomial kernel function would be suitable. Parameter σ plays a major role in the case of the Gaussian radial basis kernel function. The performance of this kernel function is greatly affected by the selection of the parameter σ. It has to be carefully selected depending on the problem [13]. The constant C, the radius of the insensitive tube ε, and the kernel parameters are those which have an impact over the effectiveness of the nonlinear SVR. Because these values are mutually connected, changing the value of one has an impact on the other associated parameters. The smoothness/flatness of the approximation function is determined by the parameter C. Due to underfitting of the training data, a lower value of C causes the learning machine to make bad approximations. A high C value overfits the training data and focuses on minimising solely the empirical risk, allowing for more complicated learning. The parameter determines the breadth of the -insensitive zone used for fitting the training data and is connected to smoothing the complexity of the approximation function. The parameter influences the number of support vectors, and hence the complexity and generalisation capabilities of the approximation function are both controlled by its value. It also controls the approximation function's precision. Smaller values of ε result in a larger number of support vectors, resulting in a more complicated learning machine. Larger ε values result in more flat regression function estimations.

Table 1
Some commonly used kernel functions.

Kernel Function	Formula
Linear	$K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}.\boldsymbol{z}$
Polynomial	$K(\boldsymbol{x}, \boldsymbol{z}) = (1 + (\boldsymbol{x}.\boldsymbol{z}))^d$ where d is degree of polynomial
Gaussian	$K(\boldsymbol{x}, \boldsymbol{z}) = \exp[-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|^2}{2\sigma^2}]$ where σ is the width of kernel

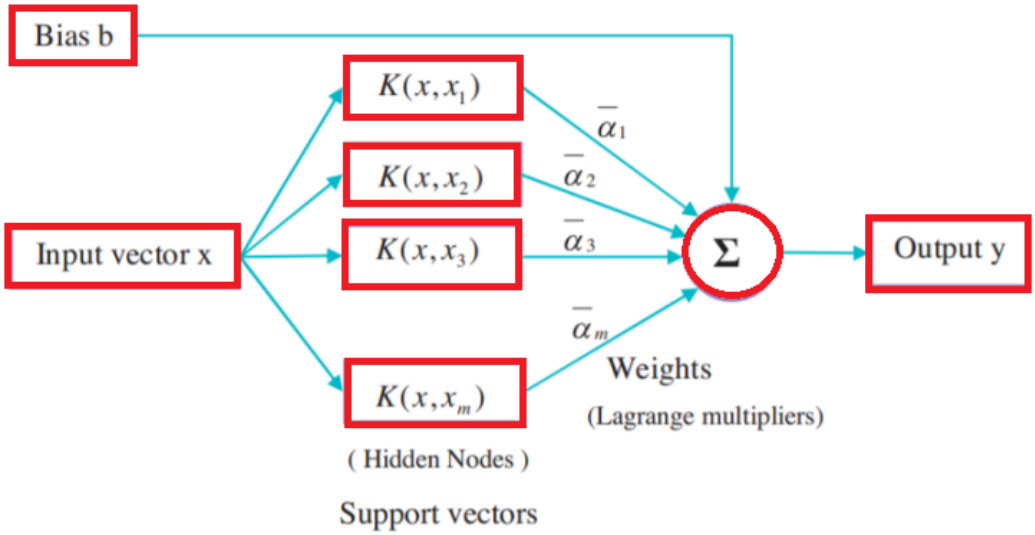

Fig. 1. Flow chart for SVM based regression.

3.2. K Fold cross validation

Although SVMs are good in generalization, overfitting may still occur because of data bias in training. K fold cross validation can be used to overcome this. In K fold cross validation, the original training data set will be divided into k equally sized subsets. From the k subsets, a single subset will be retained as a validation set, and the remaining k-1 subsets will be used as the training set. The cross validation process will then be repeated k times (the folds), with each of the k subsets. The final performance of a k fold model training will be the average of validation performances in k subsets. Usually the value of k is determined based on the availability of samples, generally from 2 to 10. The advantage of k fold cross validation is that in each round, the training sets and validation set are independent.

3.3. Hybrid wavelet-SVM technique

A wavelet is a zero-mean, quickly fading wave-like oscillation. The signal/time series is convolved against specific instances of a wavelet at various time scales and places in a wavelet transform. Hybrid wavelet-SVM approach is a combination of wavelet and support vector machine techniques. Due to its multi-resolution capability, wavelet analysis helps to obtain the time–frequency representations of the signal with different resolution[10]. The wavelet

transforms help to decompose the time series signal into various resolutions by controlling scaling and shifting [14]. Unlike other analyses, wavelet analysis has the potential to reveal trends, self-similarity, discontinuities in higher derivatives, breakdown points etc. [10]. Thus, the time-frequency localization of a signal can be efficiently achieved through wavelet transforms. The main difference between the wavelet and Fourier transforms is that the latter can deal with stationary data only but the former can very well deal with non-stationary data [7].

The signal is separated into shifted and scaled replicas of the original (mother) wavelet using wavelet analysis. The wavelet, which is chosen as the mother wavelet, should satisfy the following: (i) the mean of the function of the wavelet signal should be zero and (ii) wavelet signal has to be localized in both time and frequency domains. Wavelets can be classified into discrete and continuous types. Those which are strictly finite in the time domain are known as discrete wavelets, and others are called continuous wavelets. Selection of the type of wavelet transform (discrete or continuous), mother wavelet, and decomposition level are some of the important aspects to be studied before performing wavelet analysis for hydrological forecasting as these factors affect the results significantly.

The use of Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT) mainly depends on the purpose for which it is to be used. To understand non-stationary and complex localized variability of a time series, the CWT method can be used. For denoising and identification of true components, it would be better to use the DWT method [15]. Commonly used mother wavelets include Haar, Daubechies, Symn, Mexican Hat etc.

4. Study area

Kerala is located between 8.3º and 12.8º North latitudes and 74.9º and 77.9º East longitudes in the South Western part of peninsular India. Physiographically, the state can be divided into three zones, viz., highlands, midlands and the lowlands, all running almost parallel to each other along its length. The Western Ghats are located in the highlands which is spread over almost half the area of the state. It has large peaks like the Anaimudi with an elevation of 2694 m above the MSL [16]. The highlands are covered by forests as well as cardamom, coffee and tea plantations. The midlands are around 40% of the state and have an undulating topography of valleys and hills [16]. Most of the area under midlands are urban settlements and agricultural land. The lowlands comprise of the western coastal plains and houses beaches, backwaters, river deltas and lagoons.

Kerala is bounded by the Western Ghats to its east and Arabian Sea to its west. There are 44 rainfed rivers in the State, 41 of which flow towards the west and empty into the Arabian Sea. Also, the State is home to 34 lakes [16], and 61 dams [17]. The rivers are of relatively short length with steep bed slopes and hence the lead time available is very short. In short, Kerala is highly vulnerable to floods and hence there is an urgent need for developing and implementing a robust and efficient flood forecasting model for at least the major rivers in the State, if not for all the rivers.

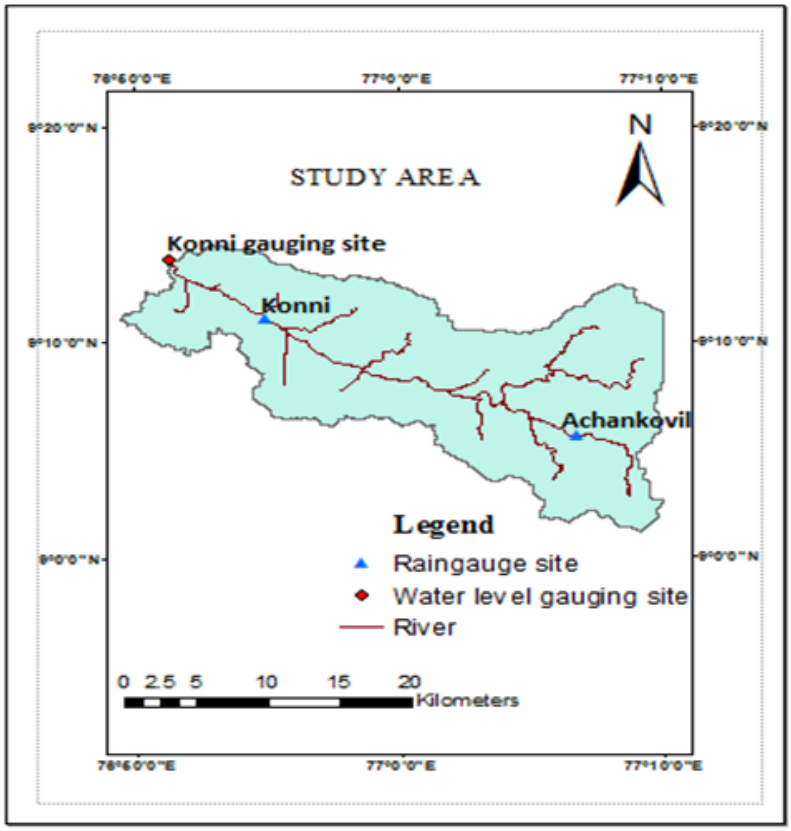

Fig. 2. Study area showing the gauging sites (Source: Alexander et al., 2018 [10]).

The river chosen for this study is the Achankovil River up to the river gauging station at Konni (Figure 2). Two rain gauge sites namely, Konni estate and Achankovil station, located in the study area are shown in Figure 2. The average rainfall in the river basin is about 2700 mm [10]. The river flows westwards through Pathanamthitta, Alappuzha, and Kollam, districts of Kerala before meeting the Arabian Sea at Thottapally. The geographical coordinates of the Achankovil river basin extends from 8^0 75' 0" to 9^0 5' 0" N latitudes and 76^0 25' 0" to 76^0 75' 0" E longitudes. The Pamba river basin is located on its northern side whereas the Pallikkal and Kallada river basins are located on its southern side. The Western Ghats define the basin's eastern border, while the Arabian Sea forms its western border. The Achankovil river basin covers a total area of 1484 km^2. The length of the river is 128 km. Like all the river basins in Kerala, this river basin can also be divided into three physiographic zones based on elevation, namely the low lands, mid lands and high lands. The study area is the upstream part of the river basin and is mostly prone to flash floods during the monsoons. The catchment area contributing to the Konni River gauging station is 449.4 km^2.

5. Methods

The overall methodology adopted in this study is presented in Figure 3.

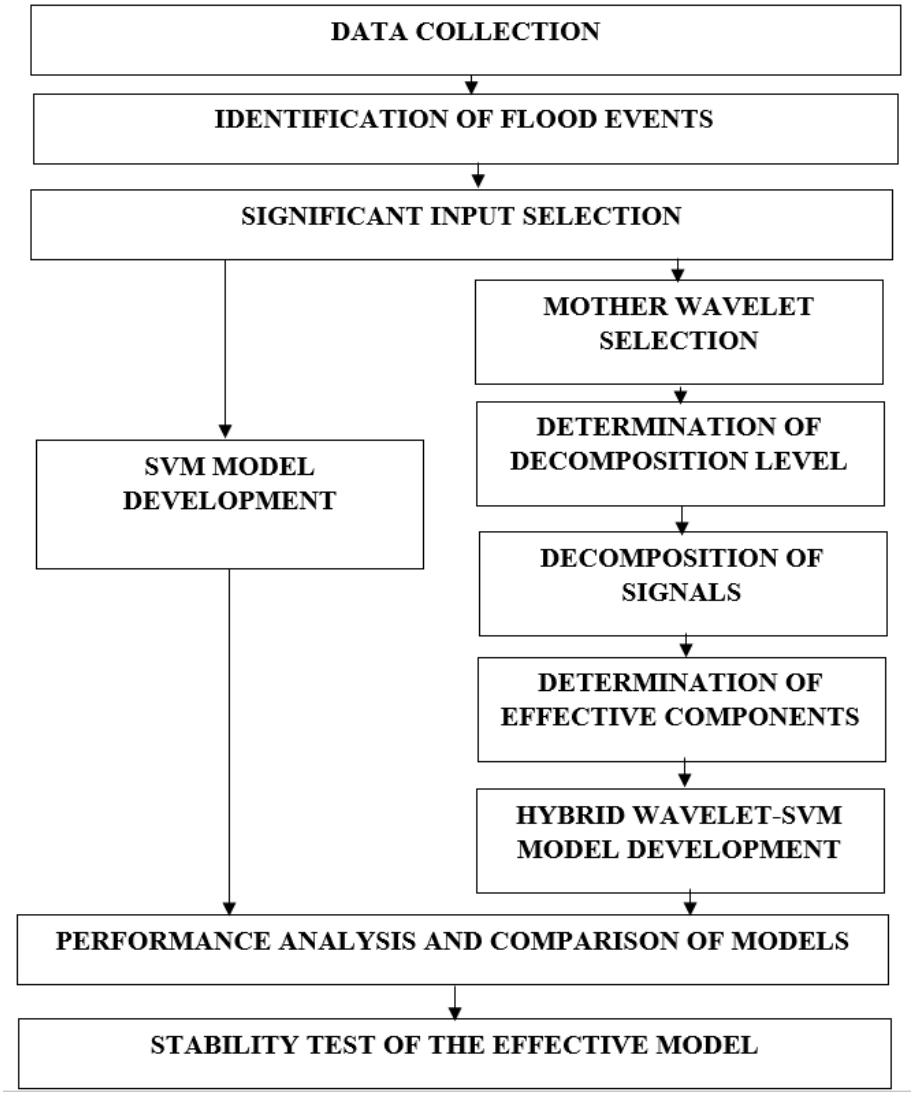

Fig. 3. Methodology adopted in the study.

5.1. Data collection

This study used five-year time series data of hourly water level at the Konni gauging site and hourly rainfall at the Konni estate and Achankovil stations for the years 2011–2015.

5.2. Identification of flood events

Flood events during the period 2011-2015 were identified from the river stage data at the Konni river gauging station by setting a threshold value of 2m for the stage. A flood event was identified from the pattern of increase in water level reaching a peak followed by a decrease in water level. The corresponding hourly rainfall values were also identified.

5.3. Selection of significant inputs (water level and precipitation)

Partial autocorrelation analysis was performed for the hourly water level time series with confidence band of 95% for different lags (in hours) to recognize the effect of previous flow

values on the subsequent flow values. This is one of the best methods for identifying significant lags.

Autocorrelation Function (ACF) is used to express the correlation between the observations at a time and the observations at previous times. Autocorrelation coefficient is a measure of the correlation between the observations at different times. The relationship between an observation in a time series and the observations at previous time steps with the relationships of intervening observations eliminated is known as partial autocorrelation. After removing the effect of any correlations attributable to terms at lower lags, the partial autocorrelation at lag k is the correlation that remains.

Data driven approaches, have the ability to select the critical model inputs [18]. But in various flood prediction studies based on data driven methods, the lag for input precipitation was selected based on the time of concentration [10,19]. Alexander et al. (2018) reported that the time of concentration of this catchment is 4 hours.

5.4. SVM model development

The flood events identified were grouped as training and testing events. The model was developed using fifteen training events and four testing events. The data was arranged by appending flood events one after the other and was used as input for SVM training. A total of 20608 data points were input to the model. Training and testing were performed using the Regression Learner App in MATLAB R2019b. Linear, quadratic, coarse Gaussian, medium Gaussian and fine Gaussian kernel functions were used for training. Also, 5-fold cross validation was utilized to reduce overfitting problems

5.5. Selection of mother wavelet

A suitable mother wavelet has to be used as the type of wavelet used affects the results of time series analysis. A large number of wavelets are used in time series analysis. The choice of a suitable mother wavelet for a problem is a great challenge. The choice is governed largely by the purpose and by the wavelet function's usual features like its number of vanishing moments and the region of support. The vanishing moment of a wavelet reflects its ability to ability to represent the polynomial behaviour of the data, while the support region indicates its capacity to localize [15].

Generally, mother wavelets are of two types, namely, orthogonal and non-orthogonal. Orthogonality refers to the property by which the information captured by one wavelet is completely independent of the information captured by another. Orthogonal wavelets are found to be ideal for hydrological variables because these are efficient in wavelet decomposition, denoising, multi resolution analyses etc [10]. Meyer, Daubechies (db), and Haar wavelets are some of the major orthogonal wavelets. In the case of hydrologic time series, wavelets under the Daubechies family yield better results [20]. These are a family of wavelets with orthogonal properties and are compactly supported with extreme phase. For a given support width, these wavelets have the maximum number of vanishing moments. A number of wavelets come under

the Daubechies family - designated as db1, db2 etc. The index numbers 1, 2 etc. represent the number of vanishing moments.

5.6. Selection of decomposition level

The accuracy of features identified in a time series depends on the decomposition level selected and hence the choice of an appropriate level of decomposition or temporal scale is very important. In earlier studies, trial and error procedure was used for this purpose. A formula to determine the minimum level of decomposition L_{min} [21,22] is

$$L_{min} = int[logN] \tag{5}$$

where N is number of data points.

The maximum level of decomposition L_{max} for a DWT [23] is:

$$L_{max} = int[log_2 N] \tag{6}$$

5.7. WSVM model development

It is required to forecast the water level (Q_{t+i}) at time $t+i$, where i is the lead time of the river flow time series. Values in the time series up to time t form the input to the SVM model. The output will be the water level at time $t+i$.

$$Q_{t+i} = f\,(\,Q_t, Q_{t-1} \ldots\ldots Q_{t-j}\,,\, P(A)_t\,,\, P(A)_{t-1}, \ldots\ldots P(A)_{t-k}\,,\, P(K)_t, P(K)_{t-1} \ldots\ldots P(K)_{t-k}\,) \tag{7}$$

where f is the unknown function, the value i represents hourly lead time, while the indices j & k denote time steps for y (water level at Konni) and $P(A)$ and $P(K)$ are the precipitation values at Achankovil and Konni estate respectively. A flow chart of the hybrid wavelet SVM method adopted in this study is presented (Fig. 4).

The magnitude of peak discharge and the time to peak are the two most important parameters of the flood hydrograph and hence these have to be predicted accurately for good flood forecasts. Some of the components of the input data may contain noise. Such components have to be identified and the signal has to be reconstructed without these components. The sharpest features of the original signal may be lost due to the removal of high frequency information completely and this would affect the peak value prediction during floods. In order to reduce such errors and enhance the accuracy of prediction of the peak values and for more efficient de-noising, an approach called thresholding can be adopted. In this approach, an optimal threshold value is found and the portion of the components which exceed this limit is discarded. The optimal threshold value has to be carefully determined as it can greatly affect denoising. A very small value of threshold can result in considerable amount of noise remaining in the input. A very large value of threshold also affects the analysis as some of the relevant features of the signal may be filtered out. Many methods are available for determining the optimal threshold value. Because of its simplicity and effectiveness, universal threshold method is the most widely used [24] and it is expressed as follows:

$$\lambda = \sigma\sqrt{2ln(N)} \tag{8}$$

where, σ is the average variance of the noise and N is the signal length. σ is calculated using the median estimate method [24].

$$\sigma = \frac{Median(|W_{j,k}|)}{0.6745} \tag{9}$$

where, $W_{j,k}$ represent all the detail wavelet coefficients[1]. After computing the threshold, effective components are selected using soft thresholding function.

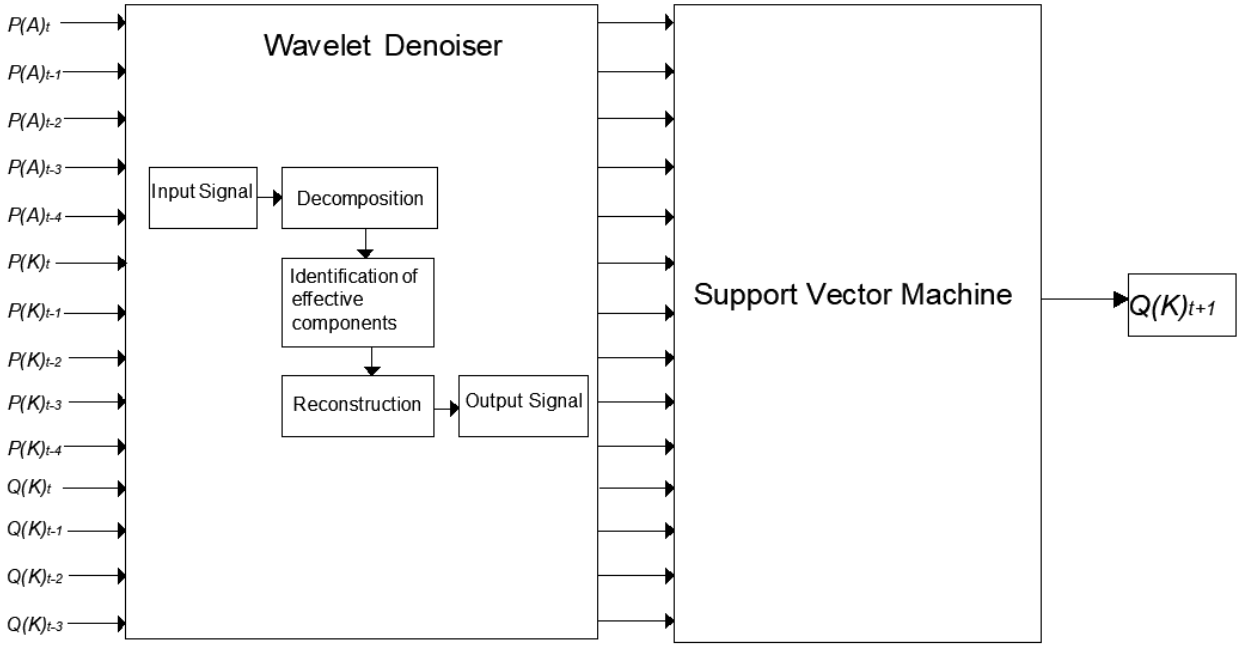

Fig. 4. Flow chart of the hybrid wavelet SVM method.

The soft thresholding function is defined as:

$$W_{st} = \begin{cases} sgn(W_{j,k})(|W_{j,k}| - \lambda); & |W_{j,k}| \geq \lambda \\ 0; & |W_{j,k}| \leq \lambda \end{cases} \tag{10}$$

5.8. Performance evaluation

The performance of the models can be evaluated by using a number of statistical techniques that can assess the predictive ability of the models. The performance measures used are the percentage deviation in peak stage *(Dev*), time difference to peak stage (*Dep*), Nash–Sutcliffe Coefficient (*NSC*), coefficient of determination (R^2) and root mean square error (*RMSE*). These are defined as follows [10]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}} \tag{11}$$

[1] In Discrete Wavelet Transform, the signal will be decomposed into high scale, low frequency coefficients called approximation and low scale, high frequency coefficients called detail.

where, N is the total number of observations, O_i is the observed value at the i^{th} time, P_i is the computed value at the i^{th} time.

$$R^2 = [\frac{\sum_{i=1}^{N}(O_i - O_{avg})(P_I - P_{avg})}{\sqrt{\sum_{i=1}^{N}(O_i - O_{avg})^2}\sqrt{\sum_{i=1}^{N}(P_i - P_{avg})^2}}]^2 \quad (12)$$

where, O_{avg} is the mean of observed values, and P_{avg} is the mean of computed values.

$$NSC = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{(O_i - O_{avg})^2} \quad (13)$$

$$Dev = \frac{(P_p - O_p)}{O_p} \times 100 \quad (14)$$

where, O_p is the peak of observed values and P_p

is the peak of computed values.

$$Dep = \left(T_{P_p} - T_{O_p}\right) \quad (15)$$

where, T_{P_p} is the time to peak for computed values and T_{O_p} is the time to peak for observed values.

6. Results and discussions

6.1. Selection of inputs

Nineteen flood events (E1 – E19) and the corresponding hourly rainfall values during the period 2011-2015 were identified. Details pertaining to the flood events identified are presented in Table 2. Figures 5 and 6 show partial autocorrelation function (PACF) graphs for two events, E3 and E6, respectively. From the partial autocorrelation statistics of the flood events identified, presented in Table 3, it was observed that the 3 h antecedent water level values (the average of all the PACF values) were the most significant for making forecasts. Also, the time of concentration of this catchment reported by Alexander et al. [10] is 4 hours. Therefore, one-, two- and three-hour antecedent water levels and one-, two-, three- and four- hour antecedent rainfall along with the present water level and rainfall were fixed as input to the flood forecasting model.

6.2. SVM model development

Three SVM models were developed to forecast the water level at one-, three-, and six-hour lead times. The RMSE and R-squared values during training with the SVM models are presented in Table 4. It can be seen that the least RMSE value was obtained for the linear SVM model whereas the R-squared values were the highest with this model for all the lead times. Hence this model was selected for prediction.

Table 2
Details of flood events identified.

Event	Beginning		End		Rainfall (mm)		Water level (m)	
	Date	Time (h)	Date	Time (h)	Konni	Achankovil	Mean	Max
E1	27/6/2015	10	28/6/2015	23	70.70	45.00	2.31	2.71
E2	29/6/2015	9	30/6/2015	1	42.00	21.70	2.00	2.13
E3	29/10/2015	21	4/11/2015	7	20.90	129.90	2.56	4.34
E4	4/11/2015	8	11/11/2015	24	120.90	298.10	2.64	3.32
E5	12/11/2015	1	17/11/2015	15	36.40	75.00	2.21	2.75
E6	3/6/2011	1	7/6/2011	5	114.10	96.40	2.68	3.30
E7	14/6/2011	5	15/6/2011	9	34.80	42.20	2.10	2.40
E8	17/6/2011	12	18/6/2011	18	21.40	8.80	2.22	2.57
E9	17/8/2012	2	18/8/2012	11	19.40	39.80	2.62	3.40
E10	22/6/2013	16	27/6/2013	21	193.30	175.80	2.74	3.37
E11	9/7/2013	7	12/7/2013	4	57.30	25.80	2.23	2.49
E12	22/7/2013	15	27/7/2013	16	115.50	71.80	2.42	2.80
E13	4/8/2013	11	7/8/2013	18	76.80	49.40	3.17	4.36
E14	16/9/2013	18	22/9/2013	24	121.70	123.80	2.83	3.45
E15	19/10/2013	9	21/10/2013	5	41.80	57.80	2.68	3.71
E16	14/7/2014	19	17/7/2014	7	37.10	37.50	2.40	2.93
E17	1/8/2014	4	8/8/2014	7	131.10	93.20	2.43	2.87
E18	20/8/2014	15	25/8/2014	24	277.00	137.20	3.60	6.72
E19	30/8/2014	4	6/9/2014	24	183.20	183.40	2.65	3.70

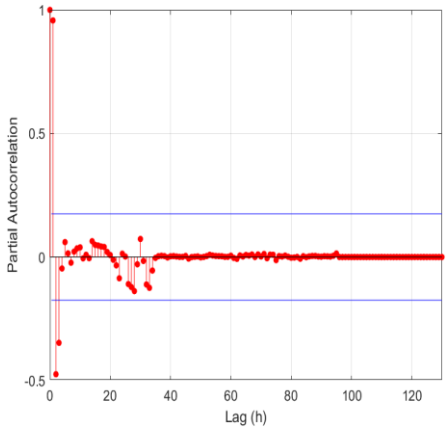

Fig. 5. PACF plot for E3.

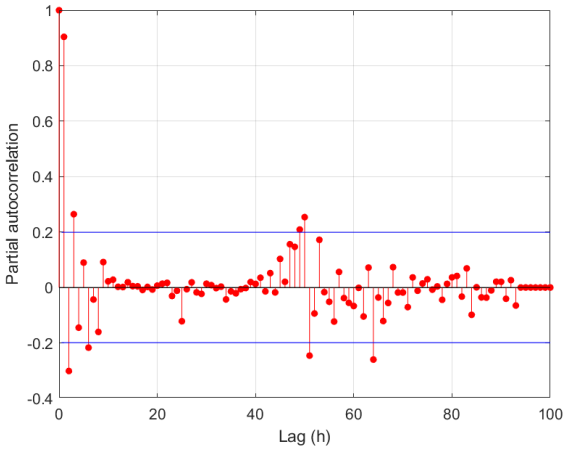

Fig. 6. PACF plot for E6.

Table 3
Significant lags from PACF plots.

Event	Significant lag obtained
E1	2
E2	2
E3	3
E4	4
E5	3
E6	3
E7	2
E8	3
E9	1
E10	4
E11	2
E12	2
E13	1
E14	4
E15	2
E16	1
E17	1
E18	2
E19	4

Table 4
Performance of various SVM models during the calibration period.

Model	$RMSE$			R^2		
	Lead Time 1	Lead Time 2	Lead Time 3	Lead Time 1	Lead Time 2	Lead Time 3
Linear SVM	0.13	0.24	0.35	0.95	0.78	0.63
Quadratic SVM	0.19	0.46	0.54	0.89	0.37	0.10
Fine Gaussian SVM	0.45	0.44	0.45	0.43	0.42	0.37
Medium Gaussian SVM	0.33	0.35	0.39	0.69	0.64	0.54
Coarse Gaussian SVM	0.16	0.25	0.36	0.92	0.81	0.63

6.3. WSVM model development

Daubechies wavelets were used as the mother wavelet in this study based on the findings of Maheswaran and Khosa [20]. Being orthogonal in nature, these wavelets are more suitable for de-noising purposes. In this study, db4 (Daubechies 4) wavelet of the Daubechies family was chosen as the mother wavelet, considering the differentiability of the input signals. In this study, wavelet analysis was performed using the Discrete Wavelet Transform (DWT). As the DWT method uses orthogonal wavelets, it helps to overcome the data redundancy problem in the Continuous Wavelet Transform (CWT) method. From equations (5) and (6), the minimum level of decomposition was 3 and the maximum level was 10. For all decomposition levels between 3 and 10, the input data was decomposed into approximations and detailed components using the db4 mother wavelet.and the effective components were identified using the universal threshold method. The signal was then reconstructed back using the soft threshold method. For each decomposition level, the correlation of the reconstructed signal was compared with that of the observed water level time series. Decomposition level 5 (db4_5) yielded better correlation and hence this was selected for performing further analysis.

6.4. Performance analysis and comparison of models developed

With the original input values, three simple SVM models with 1, 3, and 6 h lead times were built, and three WSVM models with 1, 3, and 6 h lead times were developed with the de-noised inputs, and these were tested for four unknown flood occurrences (E5, E12, E13, E15). For a lead time of 3 hours, Figure 7 shows a comparison of the hydrographs predicted by the SVM and WSVM models with the observed hydrograph. The observed stage hydrographs and the ones computed with the SVM and WSVM models for the testing events E5 and E12 are presented in Figures 8 and 9 respectively. The performance measures of the SVM and WSVM models were compared with those for the WANN model already developed [10] for the study area (Table 5). The performance measures reveal that the SVM, WSVM and WANN models perform very well in terms of its forecasts; but the overall performance of the WSVM model is slightly better than that of both SVM and WANN models. From Figures 8 and 9, it can be seen that all of the models' predicted hydrographs follow the same pattern as the observed hydrograph. However, the performance of all the models declines with increase in lead time in terms of values of R^2, *RMSE, NSC* and departure to peak. Satisfactory results are obtained up to a lead time of 3 h. But the prediction error increases for 6-hour lead time. This may be because the time of concentration of the catchment is only 4h.

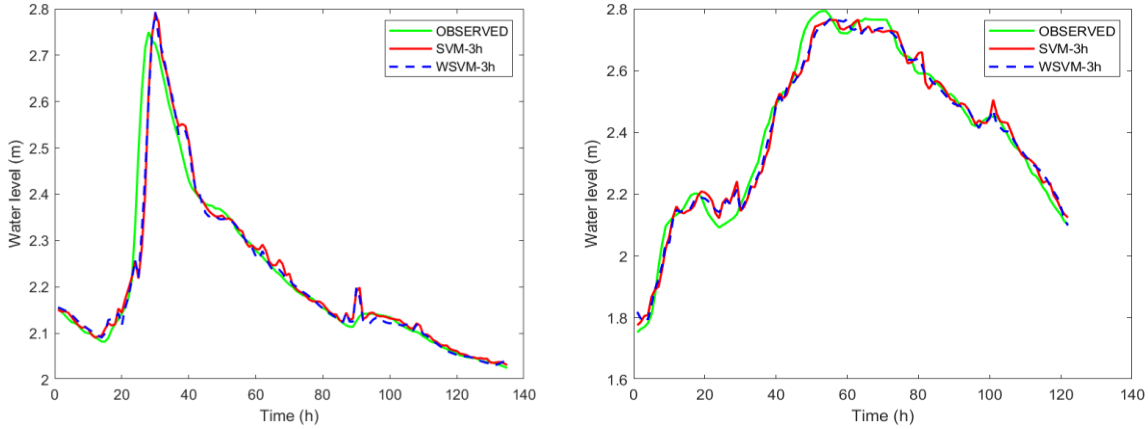

Fig. 7. Comparison of the observed and computed stage hydrographs for 3 h lead time for the testing events E5 and E12.

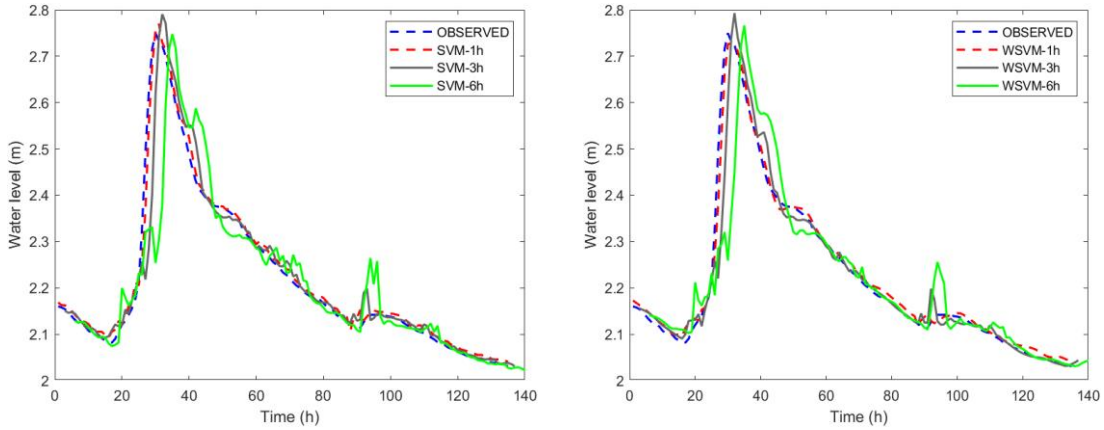

Fig. 8. Comparison of the observed and computed stage hydrographs for 1, 3 and 6 h lead times using a) SVM and b) WSVM models for the testing event E5.

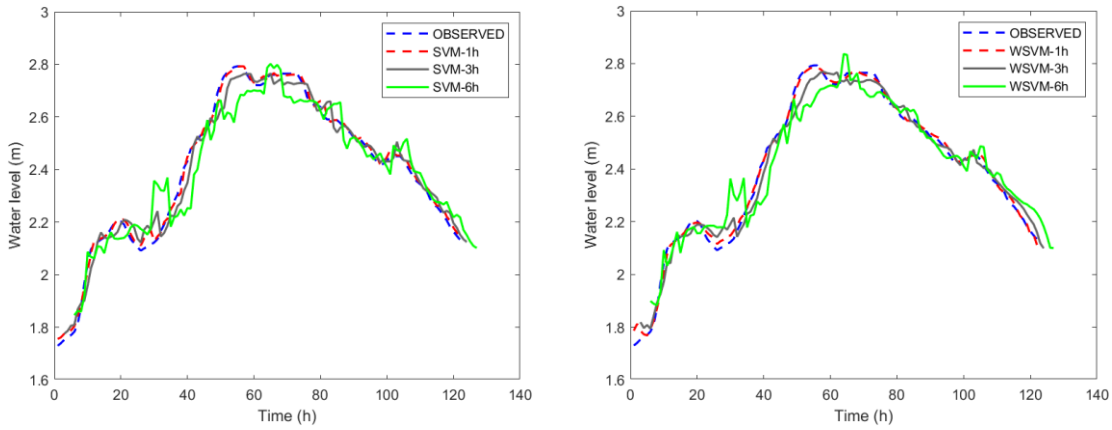

Fig. 9. Comparison of the observed and computed stage hydrographs for 1, 3 and 6 h lead times using a) SVM and b) WSVM models for the testing event E12.

7. Stability test

The stability of any machine learning model has to be validated to indicate how well the model is able to generalize the unseen data set. The stability of the WSVM model was analysed by changing the training data set and testing data set [25]. Events E3, E8, E14, and E16 were selected as testing events and all other events as training events. WSVM model was trained and tested using these events. The performance measures of this modified model were analysed to determine how a change in the input will affect the output of the model. The observed and computed hydrographs for the testing events E3, E8, E14, and E16 for 1-, 3-, and 6-h lead times are shown in Figure 10.

Table 5

Performance measures of SVM, WSVM and WANN models for different lead times for four testing events.

Performance measures	Lead time								
	1h			3h			6h		
	SVM	WSVM	WANN	SVM	WSVM	WANN	SVM	WSVM	WANN
				E5					
RMSE (m)	0.02	0.02	0.03	0.05	0.05	0.04	0.09	0.09	0.12
R^2	0.99	0.99	0.98	0.91	0.92	0.97	0.75	0.76	0.60
NSC	0.99	0.99	0.97	0.91	0.92	0.96	0.75	0.76	0.54
Dev (%)	0.72	-0.81	1.82	1.48	1.57	1.81	-0.07	0.6	-0.38
Dep (h)	1	1	0	2	2	0	5	5	5
				E12					
RMSE (m)	0.02	0.02	0.04	0.04	0.04	0.06	0.08	0.07	0.09
R^2	1.00	1.00	0.97	0.98	0.98	0.96	0.91	0.93	0.86
NSC	1.00	1.00	0.97	0.98	0.98	0.94	0.91	0.92	0.86
Dev (%)	-0.01	-0.22	0	-0.99	-0.81	0.73	0.33	1.55	1.87
Dep (h)	0	0	0	9	2	0	9	8	0
				E13					
RMSE (m)	0.25	0.24	0.12	0.45	0.45	0.18	0.59	0.59	0.50
R^2	0.93	0.93	0.98	0.76	0.76	0.90	0.55	0.55	0.60
NSC	0.93	0.93	0.97	0.73	0.73	0.94	0.46	0.47	0.46
Dev (%)	-0.35	0.17	-4.05	4.6	5.35	-6.97	7.40	8.27	-13.80
Dep (h)	0	0	1	-2	-2	1	1	0	0
				E15					
RMSE (m)	0.08	0.07	0.15	0.23	0.22	0.15	0.37	0.38	0.37
R^2	0.98	0.99	0.98	0.85	0.87	0.93	0.60	0.59	0.59
NSC	0.98	0.99	0.97	0.85	0.86	0.93	0.58	0.56	0.59
Dev (%)	1.33	0.95	-1.76	2.78	2.69	-4.85	-0.04	0.66	6.46
Dep (h)	0	0	0	1	1	0	5	4	2

Table 6
Performance measures of WSVM models developed for stability test.

Performance Measures	Lead time		
	1h	3h	6h
	E3		
RMSE	0.11	0.25	0.45
R^2	0.96	0.79	0.36
NSC	0.96	0.77	0.22
Dev (%)	0.77	3.08	4.58
Dep (h)	1	1	1
	E8		
RMSE	0.04	0.12	0.22
R^2	0.98	0.82	0.32
NSC	0.98	0.79	0.11
Dev (%)	-0.14	3.25	4.64
Dep (h)	1	1	1
	E14		
RMSE	0.02	0.06	0.16
R^2	1.00	0.98	0.86
NSC	1.00	0.97	0.86
Dev (%)	-0.37	-1.01	0
Dep (h)	1	2	-5
	E16		
RMSE	0.06	0.09	0.16
R^2	0.97	0.93	0.74
NSC	0.97	0.92	0.70
Dev (%)	-0.25	-0.92	-3.17
Dep (h)	1	2	5

The corresponding performance measures are presented in Table 6. The performance measures presented in this Table indicate that the results obtained from the WSVM model are satisfactory. The computed stage hydrograph follows the same trend as that of the observed stage hydrograph. *RMSE* values are in the range of 0.02 to 0.06 and R^2 values in the range 0.96 to 1.00 for 1-h lead time. Peak values are also predicted satisfactorily. From this, it can be concluded that the performance of the proposed WSVM model continues to be good even with a different training sample. This is because of the solid mathematical processes in the hybrid model, viz., data pre-processing, cross validation and SVM generalization.

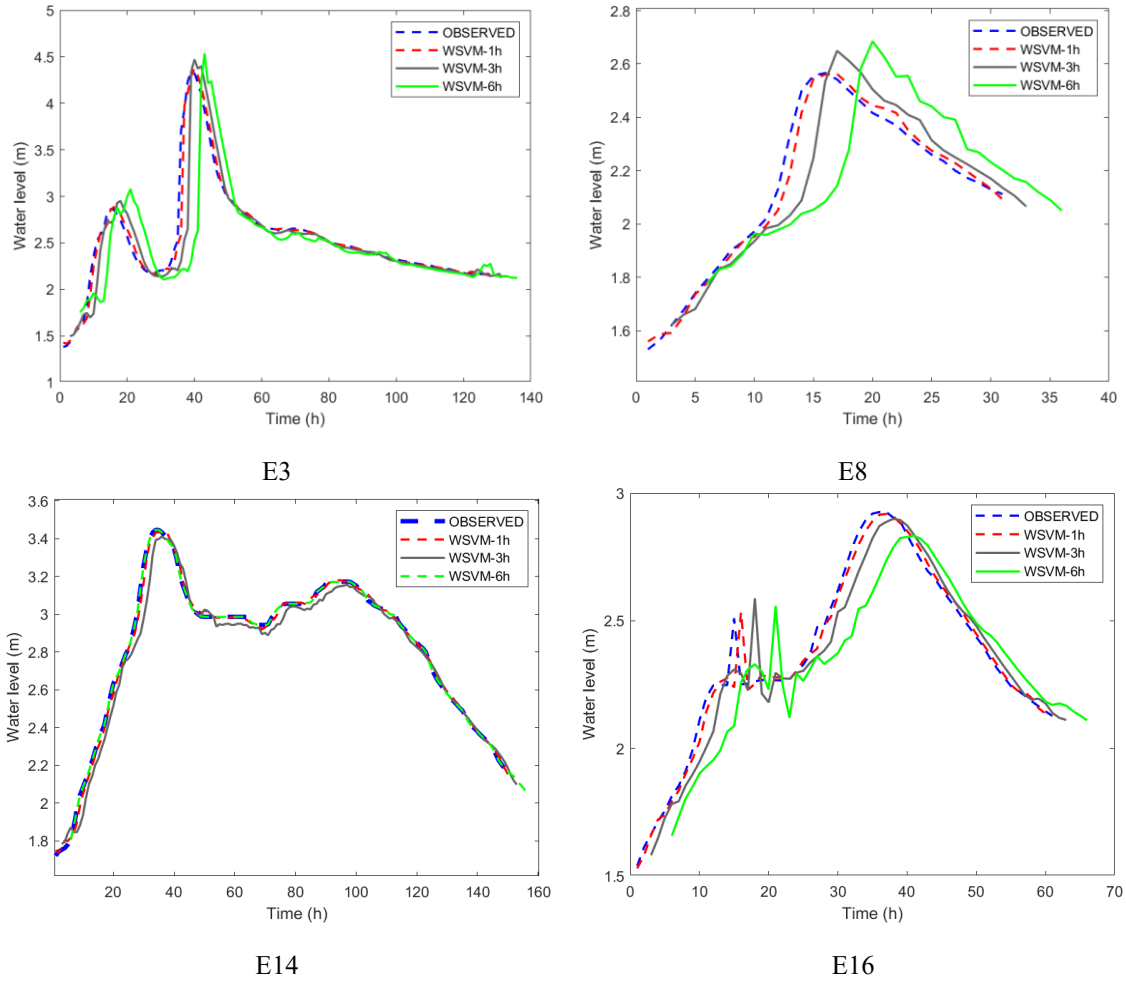

Fig. 10. Comparison of the observed and WSVM computed stage hydrographs for the testing events E3, E8, E14 and E16.

8. Conclusions

An improved hybrid WSVM model was developed to forecast flood events in the Achankovil River in Kerala, India. Initially, SVM was used for modelling. The SVM model was trained using linear, quadratic, fine Gaussian, medium Gaussian, and coarse Gaussian kernel functions. Results of model calibration indicated that the linear SVM was the most efficient and so this model was employed in further investigations. The simple SVM model was thereafter improved with wavelet pre-processing using db4 wavelet with decomposition level 5. The performance of the models was evaluated based on performance criteria, namely the *RMSE*, R^2, *NSC*, *Dev (%)* and *Dep (h)*. From the studies performed, it is concluded that the performance of the SVM model and the hybrid wavelet-SVM model are reasonably good. The performance of the hybrid wavelet-SVM is slightly better when compared to that of the SVM model. The *RMSE* value of the WSVM model lies in the range 0.02 to 0.24, R^2 and *NSC* in the range 0.93 to 1.00, *Dev (%)* in the range -0.81 to 0.95 and *Dep (h)* in the range 0 to 1 for one-hour lead time. However, as the lead time increased, the model performance deteriorated. The use of multi-scale time series and denoising of precipitation and water level data can be the reasons for the relatively better performance of the WSVM models. Comparison of the WSVM model to the WANN model

developed by Alexander et al. [10] showed that the performance of the WSVM model was better. This could be due to the better generalization ability of SVM when compared to ANN, thereby reducing over fitting problems. Stability test of the WSVM model was performed to determine how well the model responds to variations in the input data and satisfactory results were obtained.

Acknowledgments

Authors would like to thank all the faculty of Water Resources Department of National Institute of Technology for their valuable suggestions and in-depth discussions throughout and for the successful completion of this work.

Funding

This research received no external funding.

Conflicts of interest

The authors declare no conflict of interest.

Authors contribution statement

BS, NRC: Conceptualization; NRC, SGT: Data collection; NRC: Formal analysis; BS, NRC: Investigation; BS, NRC: Methodology; BS: Project administration; SGT: Resources; BS: Software; BS: Supervision; BS, NRC: Validation; BS, NRC: Visualization; BS: Roles/Writing – original draft; BS, NRC, SGT: Writing – review & editing.

References

[1] Jain SK, Mani P, Jain SK, Prakash P, Singh VP, Tullos D, et al. A Brief review of flood forecasting techniques and their applications. Int J River Basin Manag 2018;16:329–44. https://doi.org/10.1080/15715124.2017.1411920.

[2] Adnan RM, Petroselli A, Heddam S, Santos CAG, Kisi O. Comparison of different methodologies for rainfall–runoff modeling: machine learning vs conceptual approach. Nat Hazards 2021;105:2987–3011. https://doi.org/10.1007/s11069-020-04438-2.

[3] Amini A, Abdollahi A, Hariri-Ardebili MA, Lall U. Copula-based reliability and sensitivity analysis of aging dams: Adaptive Kriging and polynomial chaos Kriging methods. Appl Soft Comput 2021;109:107524. https://doi.org/10.1016/j.asoc.2021.107524.

[4] Behzad M, Asghari K, Eazi M, Palhang M. Generalization performance of support vector machines and neural networks in runoff modeling. Expert Syst Appl 2009;36:7624–9. https://doi.org/10.1016/j.eswa.2008.09.053.

[5] Komasi M, Sharghi S. Hybrid wavelet-support vector machine approach for modelling rainfall-runoff process. Water Sci Technol 2016;73:1937–53. https://doi.org/10.2166/wst.2016.048.

[6] Zaker Esteghamati M, Flint MM. Developing data-driven surrogate models for holistic performance-based assessment of mid-rise RC frame buildings at early design. Eng Struct 2021;245:112971. https://doi.org/10.1016/j.engstruct.2021.112971.

[7] Rhif M, Abbes A Ben, Farah IR, Martínez B, Sang Y. Wavelet transform application for/in non-stationary time-series analysis: A review. Appl Sci 2019;9:1–22. https://doi.org/10.3390/app9071345.

[8] Wei CC. Wavelet support vector machines for forecasting precipitation in tropical cyclones: Comparisons with GSVM, regression, and MM5. Weather Forecast 2012;27:438–50. https://doi.org/10.1175/WAF-D-11-00004.1.

[9] Adamowski JF. Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis. J Hydrol 2008;353:247–66. https://doi.org/10.1016/j.jhydrol.2008.02.013.

[10] Alexander AA, Thampi SG, Chithra NR. Development of hybrid wavelet-ANN model for hourly flood stage forecasting. ISH J Hydraul Eng 2018;24:266–74. https://doi.org/10.1080/09715010.2017.1422192.

[11] Han D, Chan L, Zhu N. Flood forecasting using support vector machines. J Hydroinformatics 2007;9:267–76. https://doi.org/10.2166/hydro.2007.027.

[12] Vapnik VN. The nature of statistical learning theory. vol. 37. 1st ed. 1995.

[13] Liu Z, Zuo MJ, Zhao X, Xu H. An analytical approach to fast parameter selection of gaussian RBF kernel for support vector machine. J Inf Sci Eng 2015;31:691–710.

[14] Seo Y, Kim S, Singh VP. Multistep-ahead flood forecasting using wavelet and data-driven methods. KSCE J Civ Eng 2015;19:401–17. https://doi.org/10.1007/s12205-015-1483-9.

[15] Sang Y-F, Singh VP, Sun F, Chen Y, Liu Y, Yang M. Wavelet-Based Hydrological Time Series Forecasting. J Hydrol Eng 2016;21:06016001. https://doi.org/10.1061/(asce)he.1943-5584.0001347.

[16] Oommen T, Coffman R, Sajinkumar KS, Vishnu CL. GEOTECHNICAL IMPACTS OF AUGUST 2018 FLOODS OF KERALA, INDIA Event: August 2018 Geotechnical Extreme Events Reconnaissance Turning Disaster into Knowledge Sponsored by the National Science Foundation GEER Association Report NO-058 2018:10–7. https://doi.org/10.18118/G6ZH3K.

[17] Central Water Commission Government of India 2019. National Register of Large Dams -2019 2019:300.

[18] Maier HR, Dandy GC. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environ Model Softw 2000;15:101–24. https://doi.org/10.1016/S1364-8152(99)00007-9.

[19] Sarkar A, Kumar R. Artificial Neural Networks for Event Based Rainfall-Runoff Modeling. J Water Resour Prot 2012;04:891–7. https://doi.org/10.4236/jwarp.2012.410105.

[20] Maheswaran R, Khosa R. Comparative study of different wavelets for hydrologic forecasting. Comput Geosci 2012;46:284–95. https://doi.org/10.1016/j.cageo.2011.12.015.

[21] Nourani V, Komasi M, Mano A. A multivariate ANN-wavelet approach for rainfall-runoff modeling. Water Resour Manag 2009;23:2877–94. https://doi.org/10.1007/s11269-009-9414-5.

[22] Wang W, Ding J. Wavelet Network Model and Its Application to the Prediction of Hydrology. Nat Sci 2003;1:67–71.

[23] Lei L, Wang C, Liu X. Discrete Wavelet Transform Decomposition Level Determination Exploiting Sparseness Measurement. Int J Electr Comput Energ Electron Commun Eng 2013;7:691–4.

[24] He C, Xing J, Li J, Yang Q, Wang R. A New Wavelet Threshold Determination Method Considering Interscale Correlation in Signal Denoising. Math Probl Eng 2015;2015. https://doi.org/10.1155/2015/280251.

[25] Zhou T, Wang F, Yang Z. Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction. Water (Switzerland) 2017;9. https://doi.org/10.3390/w9100781.