MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 2, pp. 384–400 (2025)

Hybrid least squares support vector machine for water level forecasting

Someetheram V.[1], Marsani M. F.[1], Kasihmuddin M. S. M.[1], Zamri N. E.[2]

[1] *School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia*
[2] *Department of Mathematics and Statistics, Faculty od Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia*

(Received 16 December 2024; Revised 17 April 2025; Accepted 18 April 2025)

Previous studies have highlighted the significant role of historical water level data in flood forecasting. In this study, we compare two standalone models, Support Vector Machine (SVM) and Least Squares Support Vector Machine (LSSVM), with hybrid models that integrate Ensemble Empirical Mode Decomposition (EEMD) with SVM and LSSVM, aiming to develop a more effective forecasting approach for hydrological data. Particle Swarm Optimization (PSO) is incorporated into these hybrid models to optimize the parameters of SVM and LSSVM, resulting in four models: SVM-PSO, LSSVM-PSO, EEMD-SVM-PSO, and EEMD-LSSVM-PSO. This study focuses on forecasting water levels in Sungai Gombak, Malaysia. The performance of the proposed models is evaluated and compared using several metrics, including RMSE, MSE, MAPE, and the squared correlation coefficient. Results indicate that the EEMD-LSSVM-PSO model outperforms the other models, demonstrating the highest forecasting accuracy for Sungai Gombak, Malaysia, with the lowest RMSE, MSE, and MAPE values and the squared correlation coefficient value close to 1 for the testing data.

Keywords: *machine learning; predictive model; statistical method; water level prediction; flood forecasting.*

2010 MSC: 49M27, 49M30, 62-07, 62P12, 65-04 **DOI:** 10.23939/mmc2025.02.384

1. Introduction

Floods are one of the most frequent natural disasters in Malaysia, causing significant economic damage, particularly in areas such as Sri Muda, Sungai Gombak, and Mentakab. These events are primarily driven by rising water levels during heavy rainfall, which exceed the capacity of catchment areas. Developing reliable early warning systems is crucial to protecting lives and minimizing damage in flood-prone regions [1–4].

Accurate prediction of hydrological data, such as water levels [5–7], is essential for effective flood forecasting. Traditional statistical methods, such as the autoregressive integrated moving average (ARIMA) [8] model, have been widely used, but they often struggle with the complexities of nonlinear and nonstationary data. In recent years, machine learning techniques [9, 10] have gained prominence for their ability to process such complex datasets. Among these, the Support Vector Machine (SVM) has proven to be a powerful tool due to its strong predictive capabilities and its ability to handle both linear and nonlinear data [11–13]. However, SVM has limitations, including slow convergence and high computational costs, which can hinder its performance in large-scale applications.

To address these issues, the Least Squares Support Vector Machine (LSSVM) [14] was developed as an extension of SVM, offering improved computational efficiency by transforming inequality constraints into equality constraints, using a Radial Basis Function (RBF) kernel [15, 16]. While LSSVM has shown promise in hydrological predictions, its performance highly depends on the optimization of its hyperparameters [17].

This work was supported by Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2022/STG06/USM/03/1

Particle Swarm Optimization (PSO), a global optimization algorithm inspired by swarm intelligence, has emerged as an effective method for optimizing the parameters of machine learning models like SVM and LSSVM. PSO's ability to quickly converge to optimal solutions makes it a suitable choice for enhancing the predictive accuracy of these models [18–24].

In addition to optimization, data decomposition techniques, such as Empirical Mode Decomposition (EMD) [25, 26] and its advanced form, Ensemble Empirical Mode Decomposition (EEMD) [27], have been successfully applied to improve the prediction of nonlinear and nonstationary time series data. EEMD addresses the mode mixing problem inherent in EMD, making it particularly useful for preprocessing hydrological data to enhance forecasting accuracy. Prior research has consistently demonstrated the superior performance of hybrid models that incorporate empirical mode decomposition (EMD) over single-model approaches in various field research in predicting a range of nonlinear phenomena, including runoff [28–30], wind speed [31], wave height [32], streamflow [33, 34], and vegetation dynamics [35]. Table 1 below shows the list of related research.

This study aims to combine the strengths of these approaches by developing hybrid models that integrate EEMD with SVM and LSSVM, which are optimized by PSO. The proposed models, EEMD-SVM-PSO and EEMD-LSSVM-PSO, are designed to enhance the accuracy of water level forecasting, thereby improving flood early warning systems in Malaysia. The objectives of this research are as follows:

1. To develop and evaluate the performance of hybrid models that combine EEMD with SVM and LSSVM, which are optimized by PSO, for water level prediction.
2. To optimize the hyperparameters of SVM and LSSVM using PSO to enhance their predictive accuracy.
3. To compare the performance of these hybrid models with standalone SVM and LSSVM models in forecasting water levels using nonlinear and stationary hydrological data.

The remainder of this paper is organized as follows: Section 2 provides a brief introduction to SVM, LSSVM, PSO, and EEMD methodologies. Section 3 details the performance metrics used in this study. Section 4 presents the study area, the results, and a detailed analysis of the proposed models. Finally, conclusions are drawn in Section 5.

2. Methodology

2.1. Ensemble empirical mode decomposition (EEMD)

EMD is an effective technique often used for the analysis of nonlinear data sets [40, 41]. It works by breaking down the original data into distinct components, generating numerous sets of Intrinsic Mode Functions (IMFs). The IMFs act as basis functions, which are determined by the signal itself, rather than by pre-determined kernels. The IMFs produced must satisfy two conditions [42]. Firstly, based on the original data set, the number of extremes and zeros must either be equal or differ by at most one. Secondly, the mean of the upper and lower envelopes must be zero at any point. The generated IMFs encompass a range of frequency bands, from high to low. Let $s(t) = (t = 1, 2, \ldots, l)$ denote an initial time series from the dataset. The steps for the EMD methods are as follows [43]:

Step 1: Find both local maxima and minima within the time series.

Step 2: Create upper and lower envelopes for the time series by connecting all local maxima and minima using cubic spline interpolation.

Step 3: Find the average envelope using the upper and lower envelopes according to equation (1), provided below,

$$m(t) = \frac{e_{\text{up}}(t) + e_{\text{low}}(t)}{2}. \tag{1}$$

Step 4: The difference between the original time series and the average envelope computed in Step 3 using equation (2) provided below:

$$h(t) = s(t) - m(t). \tag{2}$$

Table 1. Summary of related studies.

Author(s)	Data Source	Location	Detail of the Studies	Findings	Limitations
[14]	Survey	NA	The authors made a comparison between SVM and LSSVM	The authors concluded that LSSVM is preferred for large-scale problems	This research only utilizes single method models
[8]	Official data	Segamat River, Malaysia	This study forecast flood with ARIMA model by using rainfall and streamflow data set	The best ARIMA model is concluded in this study	This study does not compare ARIMA model with other single benchmark models
[2]	Official data	Red River, Vietnam	This study proposed a LSTM model to forecast flowrate for one, two and three days in advance	The study shows the forecast length one day gives better prediction than two days and three days prediction	This research does not compare with other machine learning models
[3]	Official data	Silchar and Dholai, India	This study proposed a SVM model combined with Firefly Algorithm (SVM-FA) to predict monthly river flow	The research concluded both SVM-FA and SVM models outperformed Radial Basis Function (RBF) models	This research does not imply developed SVM which is LSSVM for nonlinear data
[13]	Official data	Wadi Ouahrane basin, Algeria	The authors proposed various types of ML models for prediction of hydrological drought	The results showed SVM model has outperformed other ML models in predicting hydrological drought	This paper only utilizes linear SVM to deal with nonlinear data
[15]	Official data	Yellow River, China	The authors proposed Complete ensemble empirical mode decomposition (CEEMDAN) with LSSVM and grey model in predicting rainfall and runoff	The performance of proposed model has outperformed single LSSVM model with high accuracy	This study does not apply PSO to optimize the parameters of SVM
[36]	Official data	Kelantan River, Malaysia	The authors proposed SVM using radial basis function (RBF) to forecast flood	SVM has outperformed Neural Network models	This study does not use any optimize techniques
[37]	Official data	Yangtze River, China	The authors proposed PSO with SVM to predict river water level	PSO-SVM outperformed SVM in forecasting river water level	This study does not utilize decomposition method
[38]	Official data	Mahanadi basin, India	The study proposed FFA with PSO to forecast ground water level	SVM-FFAPSO performs the best to predict ground water level	This study does not utilize decomposition method
[39]	Official data	Clackamas River	The study proposed EMD to improve forecasting model	EMD based model achieved superiority among other models	This study does not utilize SVM as forecasting tool

Step 5: Verify whether $h(t)$ satisfies the characteristics of IMFs. If yes, $s(t)$ is the first IMF and $s(t)$ is replaced by the residuals $r(t) = s(t) - h(t)$. If not $r = s(t)$, replace with $h(t)$.
Step 6: Repeat Steps 1 to 5, continuing until the termination condition is met.

Additionally, this also means that the shifting process of EMD will stop when the residual becomes a monotonic function, where IMF extraction is no longer possible. The outcome of the EMD decomposition includes a series of IMFs and a residual component derived from the original data, as indicated in equation (3),

$$s(t) = \sum_{i=1}^{n} C_i(t) + r_n(t), \tag{3}$$

where n is the number of IMFs, $r_n(t)$ represents the final residuals, which depict a trend and act as the central tendency of the signal $s(t) = (t = 1, 2, \ldots, l)$. The term $c(t) = (t = 1, 2, \ldots, l)$ represents the Intrinsic Mode Functions (IMFs), which exhibit periodicity and are nearly orthogonal to each other. Each IMF independently characterizes the local properties of the original signal when describing them. The frequency of each IMFs varies high to low. The versatility of the EMD has been demonstrated across various applications for signal extraction from noisy and non-linear data [40]. However, a significant limitation of EMD is the frequent occurrence of mode mixing, which happens when a single IMF contains signals of greatly different scales or when a signal of the same scale appears in multiple IMFs [27].

To address this issue, Wu [27] introduced the Ensemble Empirical Mode Decomposition (EEMD). EEMD mitigates mode mixing by incorporating a finite amount of Gaussian white noise into the data series before computing the overall average, which effectively reduces mode mixing [27]. In summary, EEMD, an extension of the EMD method, aims to eliminate mode mixing by introducing white noise into the data prior to analysis [44]. The process of EEMD is briefly explained below.
Step 1: Initialize the ensemble number, M, and the noise amplitude and let $m = 1$.
Step 2: Introduce a white noise series $n_m(t)$ into the original dataset $s(t)$ and obtain the below equation $s_m(t) = s(t) + n_m(t)$.
Step 3: Perform data decomposition on $s_m(t)$ using EMD to obtain Intrinsic Mode Functions (IMFs), considering the added white noise.
Step 4: Repeat these two steps iteratively until the residue $r(t)$ either becomes a monotonic function or contains at most one local extreme point, indicating that no more IMFs can be extracted. Crucially, use $m = m + 1$ white noise series for each repetition if $m < M$; otherwise proceed with Step 5.
Step 5: Compute the ensemble mean, $y_n(t)$ of the corresponding IMFs from all decompositions to obtain the final IMFs and residual,

$$y_n = \frac{1}{2} \sum_{m=1}^{M} c_{n,m}. \tag{4}$$

2.2. Support vector machine

Vapnik [12] originally proposed SVM to address problems in both regression and classification domains. This increased interest in SVM can be attributed to its robust mathematical foundation, rooted in the principles of Structural Risk Minimization (SRM) and Empirical Risk Minimization (ERM). In this SVM model, let the training sets be $S = \{(x_i, y_i) \mid i = 1, 2, 3, \ldots, N\}$, $x_i = \mathbb{R}^n$, $y_i = \mathbb{R}$. Next, the high dimensional feature space will contain the optimal decision function. Equation (5) shows the decision function used in this paper,

$$f(x) = \langle w, \phi(x) \rangle + b, \tag{5}$$

where $\phi(x)$ represents the high-dimensional feature space that derives a nonlinear mapping from the input space; w is the weight, and b is the bias. The parameters w and b in the equation (5) is derived from solving the constrained minimization problem originally introduced by Vapnik [12] shown in equations (6) and (7),

$$\text{minimize } \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} (\xi_i^+ + \xi_i^-), \tag{6}$$

$$\begin{gathered} y_i - \langle w, \phi(x) \rangle - b \leqslant \varepsilon + \xi_i^+, \\ \langle w, \phi(x) \rangle + b - y_i \leqslant \varepsilon + \xi_i^-, \\ \xi_i^+, \xi_i^- \geqslant 0. \end{gathered} \tag{7}$$

In the above expressions, the constant $C > 0$ serves as a parameter regulating the penalty level for instances surpassing the error threshold and represents the error tolerance. Furthermore, ξ_i^+ and ξ_i^- are positive variables, where ξ_i^+ denotes the upper excess deviation and ξ_i^- signifies the lower excess deviation. Incorporating Lagrange multipliers, the problem expressed in equation (6) is subsequently converted into a dual space in equation (8) and equation (9) below:

$$W(a_i - a_j^*) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_i - a_i^*) \left(a_j - a_j^*\right) \langle \phi(x_i), \phi(x_j) \rangle - \varepsilon \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} (a_i - a_i^*) \tag{8}$$

such that

$$\varepsilon \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0, \quad \alpha_i^* \in [0, C], \tag{9}$$

where a_i and a_i^* are Lagrange multipliers. Lagrange multipliers, subject to the imposed constraints, must adhere to the conditions. The resultant solution is presented in equation (10):

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle \phi(x_i), \phi(x_j) \rangle + b. \tag{10}$$

The inner product $\langle \phi(x_i), \phi(x_j) \rangle$ is defined by the kernel function $K(x_i, x)$. Therefore, the equation can be define as in equation (11):

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b. \tag{11}$$

Equation (12) specifies the radial basis function (RBF) as the kernel function selected for the SVM model in this paper,

$$K_{\text{svm}}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_{\text{svm}}^2}\right), \tag{12}$$

where σ_{svm} is the width of the kernel function. Hence, the cost parameter y and kernel parameter σ_{svm} of SVM need to be optimized.

2.3. Least squared support vector machine

LSSVM is a developed method from support vector machine (SVM) [4,14]. Previously, SVM was used to deal with small sample of the problem which were said to be difficult to forecast. Therefore, LSSVM was introduced to solve nonlinear problems in high dimensional problem and also minimize the squared error [15]. In order to reduce the complexity of the computational process, the inequality constraints are replaced with equality constraints that transforms the quadratic programming problem to a system of linear equations. In the function estimation LSSVM model, equation (13) and equation (14) present the optimization problem as follows:

$$\text{minimize } Z(w, b, e) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \tag{13}$$

such that

$$y_i = \langle w, \phi(x_i) \rangle + b + e_i, \quad i = 1, 2, \ldots, N, \tag{14}$$

where Z is the loss function, e_i is the error, $\gamma > 0$ is the regularization constant. Equation (13) can be addressed through the Lagrange function and Karush–Kuhn–Tucker (KKT) conditions, while equation (15) demonstrates the structure of this Lagrange function,

$$L(w, b, \xi, \lambda) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \lambda_i \left(\langle w, \phi(x_i) \rangle + b + e_i - y_i\right). \tag{15}$$

Next, determine the partial derivatives of using the expression provided in equation (16) and following the principles of the Karush–Kuhn–Tucker (KKT) conditions,

$$\begin{aligned}
&\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \lambda_i \phi(x_i), \\
&\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i = 0, \\
&\frac{\partial L}{\partial e_i} = 0 \Rightarrow \lambda_i - \gamma e_i = 0, \\
&\frac{\partial L}{\partial \lambda_i} = 0 \Rightarrow \langle w, \phi(x_i) \rangle + b + e_i - y_i = 0.
\end{aligned} \tag{16}$$

Equation (16) facilitates transforming the optimization problem into solving a series of linear equations, detailed in equation (17):

$$\begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & J + \frac{1}{\gamma}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{17}$$

where I is dimensional column vector, $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_n]^T$, $y = [y_1, y_2, \ldots, y_n]^T$, $J_i j = \phi(x_i)^T \phi(x_j) = K_{\text{lssvm}}(x_i, x_j)$. The kernel function satisfying the Mercer condition, forms the core of the LSSVM model, as depicted in equation (18),

$$f(x) = \sum_{i=1}^{N} \lambda_i * K_{\text{lssvm}}(x_i, x_j) + b. \tag{18}$$

Equation (19) presents the kernel function chosen in this paper for the LSSVM model, which is the radial basis function (RBF) [45],

$$K_{\text{lssvm}}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_{\text{lssvm}}^2}\right), \tag{19}$$

where σ_{lssvm} is the width of the kernel function. Hence, the cost parameter y and kernel parameter σ_{lssvm} of LSSVM need to be optimized.

2.4. Particle swarm optimization

The Particle Swarm Optimization (PSO) algorithm was initially introduced by Kennedy and Eberhart [46]. PSO is inspired by the coordinated movement observed in natural phenomena like bird flocking, bee swarming, and fish schooling. Renowned for its simplicity in coding, cost-effectiveness, and consistent performance, PSO has established itself as a powerful algorithm for solving optimization problems [47]. In PSO, individuals within the population are referred to as particles, collectively forming a swarm. These particles commence their optimization journey with random initial positions and velocities. Throughout the optimization process, particles adapt their positions and velocities as they navigate the search space. Additionally, each particle retains memory of the best position it has encountered in the search space. The parameters of SVM and LSSVM will be optimized using the PSO algorithm. Initially, upper and lower bounds are defined for the SVM parameters and LSSVM parameters. Subsequently, random values within these bounds are generated for each particle, which are then employed as inputs for the SVM and LSSVM model. Following this, the fitness function is applied, with this study utilizing Mean Absolute Percentage Error (MAPE) as the fitness criterion to determine suitable SVM and LSSVM model parameters. The MAPE value for each particle is calculated using the fitness function in equation (20),

$$\eta_{\text{MAPE}} = \frac{1}{\omega} \sum_{i=1}^{\omega} \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \tag{20}$$

where ω represents the number of subsets, y_i denotes the actual value, $\hat{y}_i$ denotes the predicted value, and D is the dimension defining the length of each particle. Each member of the swarm, referred to as a particle, is represented as a vector X_i encompassing the parameters targeted for optimization within the objective function. In the multidimensional search space, denoted as m, the position $X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iD})$ and velocity $V_i = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{iD})$ of the i-th particle are initialized randomly within the range of possible solutions. To enhance the optimization process, the algorithm

computes the objective function value for each particle, subsequently updating both their velocities and positions in accordance with specific equations as in equation (21) and equation (22),

$$v_{id}^{t+1} = \omega v_{id}^{t} + c_1 \cdot r_1 \cdot \left(p_{id} - x_{id}^{t}\right) + c_2 \cdot r_2 \cdot \left(p_{gd} - x_{id}^{t}\right), \tag{21}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1}. \tag{22}$$

The ideal location of the particle represents as $P_i = (P_{i1}, P_{i2}, P_{i3}, \ldots, P_{iD})$. Optimal swarm location is $P_g = (P_{1g}, P_{2g}, P_{3g}, \ldots, P_{Dg})$. Under i-th particle condition at t-th iteration, x_{id}^{t} and v_{id}^{t} are d-th position and constituent of speed. Positive coefficient c_1, c_2, r_1 and r_2, where r_1 and r_2 are distributed evenly within the range of 0 to 1, c_1 and c_2 represent constants. The procedure of PSO optimizing the SVM and LSSVM parameters described as follows:
Step 1: Initialise the parameters of PSO.
Step 2: The collective of particles begins its journey with randomly assigned individual velocities and positions.
Step 3: Fitness evaluation: The various initialized parameters are fed into LSSVM, and then the fitness value of each particle is evaluated using the fitness function of PSO by using equation (20).
Step 4: Update both the global and individual best values based on the outcomes of the fitness value.
Step 5: Velocity computation: The particle moves towards a new position by updating its velocity. The velocity for each particle is derived using equation (21).
Step 6: Position Update: Each particle transitions to its subsequent position following the guidelines outlined in equation (22).
Step 7: Termination: Continue iterating through Steps 3 to 7 until the specified termination criteria are met.

2.5. EEMD-LSSVM-PSO

As previously mentioned, water level data exhibits high nonlinearity and nonstationarity, making accurate water level forecasting challenging. To address this, the proposed EEMD-LSSVM-PSO model is utilized based on the principle of decomposition and ensemble learning. The detailed steps are outlined below.
Step 1: Data Preparation by using initial water level data.
Step 2: Using Ensemble Empirical Mode Decomposition (EEMD), the original water level data is broken down into a finite number of Intrinsic Mode Functions (IMFs) and a residue.
Step 3: For constructing the EEMD-LSSVM-PSO forecasting model, employ LSSVM to develop a forecasting model for each obtained IMF and residue. Consequently, obtain the forecasting values for all IMF and residue components from the model.
Step 4: PSO is used within the proposed EEMD-LSSVM model to optimize the parameter selection. Determine the ranges of penalty coefficient C and kernel parameter of the LSSVM. Where penalty parameter ranges $[1, 100]$, kernel parameters of radial basis function ranges $[0.01, 10]$.
Step 5: Utilize LSSVM methods to train the initialized particles, calculate their individual fitness values, and update both the global optimal value and the optimal value for each individual particle.
Step 6: Define the stopping criterion: stop iterations upon reaching the maximum number; otherwise, generate a new group based on the velocity equation and return to Step 5, continuing until termination conditions are met. Ultimately, identify the particle with the lowest fitness value in the group as the optimal solution sought.
Step 7: Confirm whether the maximum iteration is achieved, and if so, end the iteration and retrieve the optimized parameters of LSSVM.
Step 8: Obtain the optimal parameter of each IMF and residual of EEMD-LSSVM-PSO and then use EEMD-LSSVM-PSO model to test the sample water level data set. These procedures allow EEMD to identify distinct information scales in the original load data. Furthermore, since each IMF shares similar frequency characteristics, the hybrid model can reduce the complexity of the LSSVM model and improve its forecasting efficiency and accuracy.

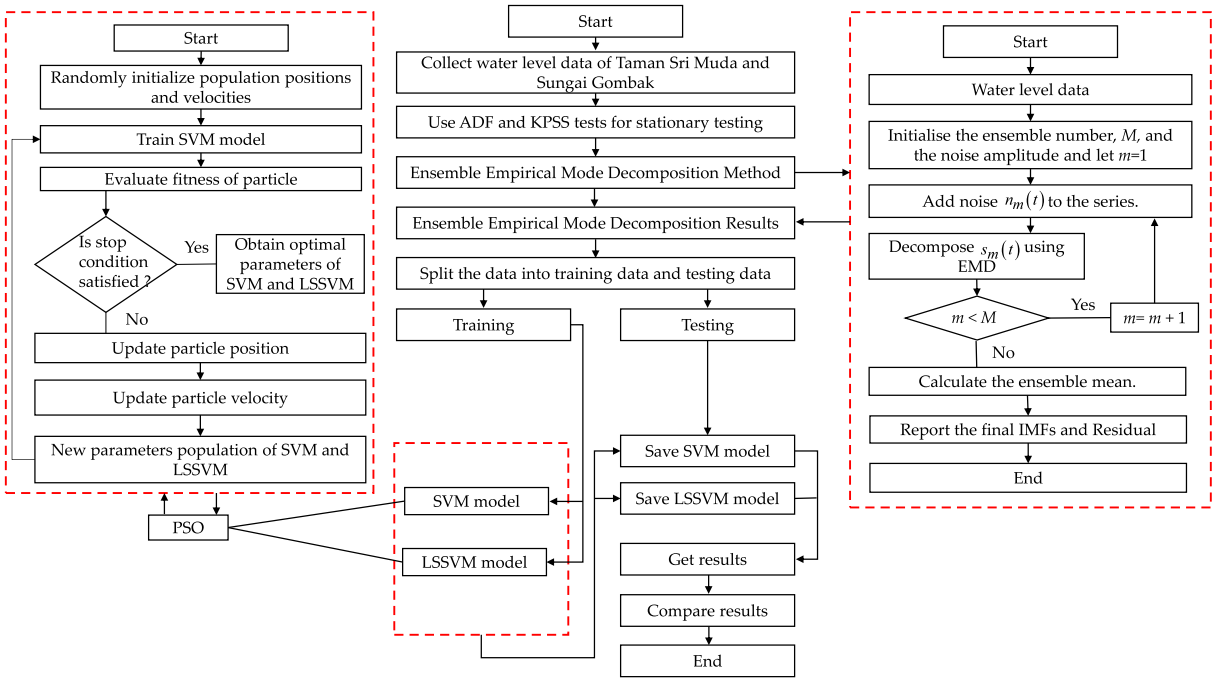

Fig. 1. The flowchart of proposed model EEMD-SVM-PSO and EEMD-LSSVM-PSO.

2.6. Stationarity tests

Autocorrelation is computed to determine how the observation are correlated in the time series. Autocorrelation function (ACF) used as a first attempt in determining the stationarity and to identify the existence of the seasonality [48]. For further confirmation, Augmented Dickey Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests are chosen to analyze the stationarity of the datasets. We implemented these methods in R software. ADF test is also called as unit root test. The null hypothesis of the test is that if the time series can be presented by a unit root, it is not stationary. Otherwise, the alternative hypothesis indicates that the data is stationary. The mathematical description of this test is explained in [49]. In KPSS test, the null hypothesis states that there is no unit root and the alternative hypothesis states that a unit root exists, meaning the data is non-stationary. The test confirms the non-stationarity of the data when the p-value is less than the significance level of 0.05. The equation of KPSS can be seen in detail in [50]. The flowchart in Figure 1 briefly illustrates EEMD-SVM-PSO and EEMD-LSSVM-PSO models.

3. Performance metrices

In this study, four different common indices are employed to assess the precision of SVM-PSO, LSSVM-PSO, EEMD-SVM-PSO and EEMD-LSSVM-PSO models. The error evaluation methods chosen are root mean square error (RMSE), mean squared error (MSE), mean absolute percentage error (MAPE) and squared correlation coefficient (R^2). The best prediction model will have the lowest value for RMSE, MSE and MAPE and R^2 value that almost reaches one.

3.1. Root Mean Squared Error (RMSE), Mean Squared Error (MAE), Mean Absolute Percentage Error (MAPE)

RMSE is widely utilized to gauge the discrepancy between model-predicted and observed actual water level data [51]. Essentially, RMSE, MSE and MAPE assesses the performance quality [36]. Equation (23), (24) and (25) shows the equation of RMSE, MSE and MAPE respectively. z_i is the observed value, $\hat{z}_i$ is the predicted value and n is number of data,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}, \tag{23}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2, \tag{24}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{z_i - \hat{z}_i}{z_i}\right|. \tag{25}$$

3.2. Squared correlation coefficient (R^2)

R^2 is computed to evaluate how well models explain variance, as depicted in equation (26). Recently, has proven effective in analyzing wind power forecasting, according to [52],

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \tag{26}$$

where $\bar{z}$ is mean value of water level data.

4. Results and discussion

4.1. Study area and data set

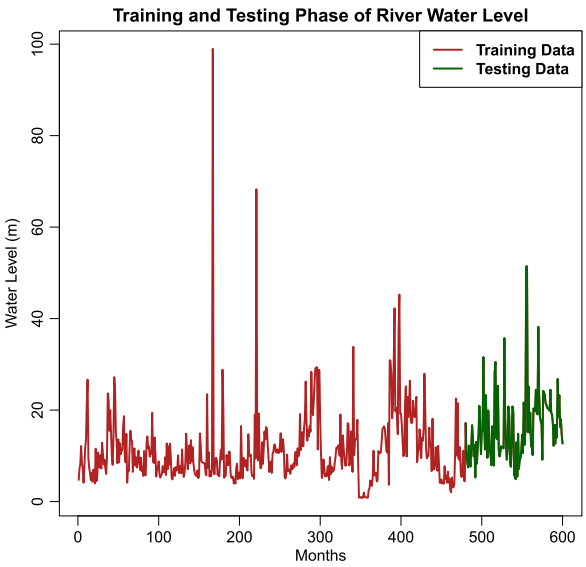

Fig. 2. The monthly water level time series at Sungai Gombak.

The study focuses on Sungai Gombak, spanning Selangor and Kuala Lumpur in Malaysia. Hydrological data provided by the Department of Irrigation and Drainage Malaysia covers the years 1960 to 2010, totaling 600 data points. The dataset was divided into monthly series to enhance prediction accuracy. For model development, 80 % of the data (from January 1960 to December 1990) was allocated to training SVM and LSSVM models, with the remaining 20 % (from January 2000 to January 2010) reserved for validation. The forecasting model uses 480 months of training data to predict water levels for 120 months of testing data. Monthly water level series for Sungai Gombak are depicted in Figure 2. In this region, flood alerts are categorized into normal, alert, warning, and danger levels, triggered by water levels exceeding 29.0 m, 30.0 m, 30.5 m, and 31.0 m respectively. Accurate water level prediction models are crucial for flood warning systems to alert authorities and protect residents. All the models will run a program by using R Software.

4.2. Descriptive statistic of the water level data

Descriptive statistics has been calculated to understand the water level data. The number of observations for this study is 600 months. The lowest water level recorded is 0.79, while the highest is 98.96. The average value of the water level data is determined for 600 months. The average water level data is 12.1721 and the standard deviation value 8.1766. The distribution skewness is 3.4638 which indicates the tail to the right and the Kurtosis of 25.3845 supports our conclusion that the distribution of water level is not normal and heavy-tailed.

4.3. Stationarity analysis

Figure 3 depicts the ACF plot for monthly water level data. An important aspect of these figures is that each sample variable exhibited both high and low volatility over time, indicating that the variability remained consistent throughout the period.

The suggestive linear trend lines in the water level plot indicate the sequence is stationary. For most of the cases, the ACF plot displayed a rapid decay for the lags. Hence, the ACF of water level data shows the presence of stationarity.

For further confirmation of presence or absence of stationarity of the water level data, ADF and KPSS tests have been chosen to test the existence of unit root in the data or not [53]. The null hypothesis of ADF test indicates that the data has unit root. The p-value of the water level data is 0.01 for ADF test. Since this value is less than 0.05, the null hypothesis is rejected. Next, the null hypothesis for KPSS test is that the water level data is stationary. Therefore, the p-value of the water level data is 0.1 for KPSS test. The null hypothesis is not rejected. Thus, the results of ADF and KPSS test showed that the observed water level data is stationary and has a predictable trend.

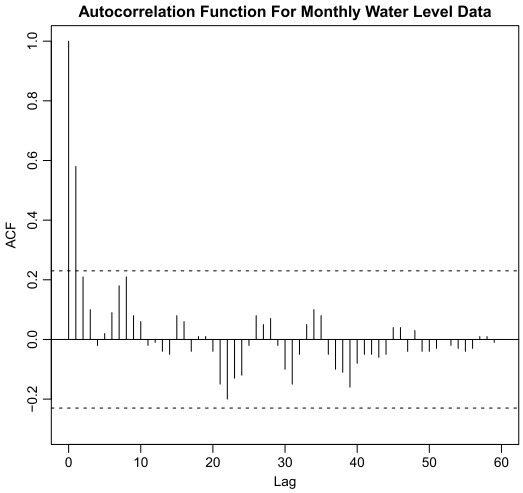

Fig. 3. The autocorrelation function plot.

4.4. Decomposition by ensemble empirical mode decomposition

The EEMD technique is utilized to decompose the original water level time series into multiple independent Intrinsic Mode Functions (IMFs) along with one residue component. The EEMD technique plays a pivotal role in preprocessing the monthly water level of Sungai Gombak, Malaysia from 1960 to 2010. Before applying EEMD, it is essential to configure two key parameters, in this study, 100 chosen for the ensemble size, and each ensemble member is enriched with white noise featuring a standard deviation of 0.2. It is worth noting that these parameter settings align with those previously employed by [27]. The paper analyses the impact of noise and selection ensemble parameter on decomposition. Therefore, we refrain from duplicating these details here. The outcomes of the decomposition process are visually represented in Figure 4 and Figure 5. The water level data was decomposed into eight (8) Intrinsic Mode Function (IMF) components and one residual component. These components are organized in a sequence based on their frequency, passing from high to low. Figures 4 and 5 illustrate that the periodicity of these eight IMFs gradually increases as their frequency decreases, while their amplitudes exhibit a corresponding decrease [40]. The graph displays a time series from 1960 to 2010 on the x-axis, representing monthly intervals. The y-axis shows the water level data, which has been decomposed. The outcome of EEMD which are the IMFs and residual becomes the input variables for forecasting water levels.

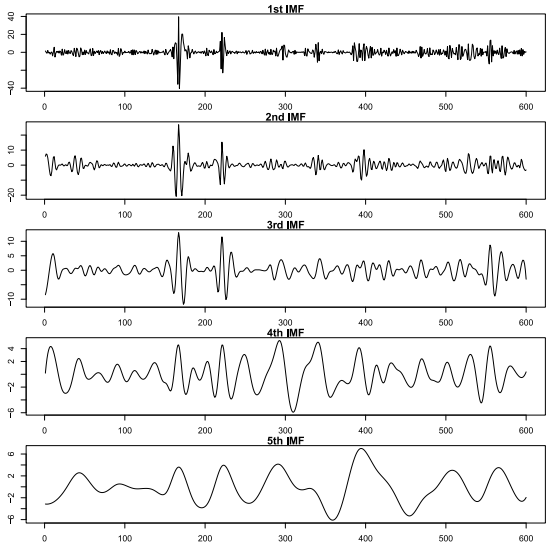

Fig. 4. 1st IMF to 5th IMF that obtained after EEMD decomposition.

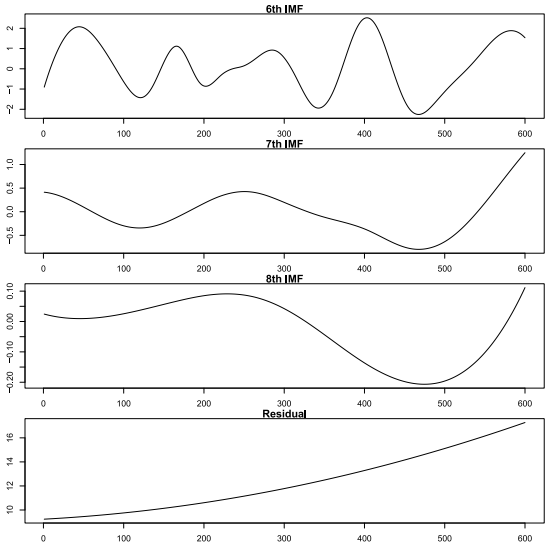

Fig. 5. 6th IMF to 8th IMF and the residual obtained after EEMD decomposition.

4.5. Particle swarm optimization analysis

The SVM-PSO and EEMD-SVM-PSO models employ the Radial Basis Function (RBF) as their kernel function, necessitating the selection of two crucial parameters: the regularization parameter and the kernel function parameter for SVM. The LSSVM-PSO and EEMD-LSSVM-PSO models employ the Radial Basis Function (RBF) as their kernel function, requiring the selection of two crucial parameters: the regularization parameter and the kernel function parameter for LSSVM. Notably, prior studies have suggested that varying kernel functions have a minimal impact on performance, underscoring the significance of the kernel parameter in SVM and LSSVM performance. Between these two parameters, kernel parameter plays a pivotal role in precisely defining the structure of high-dimensional space, thereby controlling the complexity of the ultimate solution. On the other hand, cost parameter governs the model's complexity and the degree of penalization for fitting deviations.

Table 2. Parameter settings of the methods.

Methods	Parameters	Value
PSO	Population Size	25
	Maximum Iteration	150
	Acceleration constants (c1,c2)	(1,6)
SVM	Cost	[1,100]
	Kernel	[0.01,1]
LSSVM	Cost	[1,100]
	Kernel	[0.01,1]

During the training phase, we begin by optimizing the SVM model's parameters and the LSSVM model's parameters for each IMF and the residual using PSO. The bounds (initial ranges) of the solution space used in the PSO technique, as well as the upper and lower limits of the proposed methods, are shown in Table;2. We evaluate the validation error using equation (20) and identify the parameters that yield the lowest validation error as the most suitable ones for EEMD-SVM-PSO and EEMD-LSSVM-PSO, as detailed in Table;3, and for SVM-PSO and LSSVM-PSO in Table;4. Subsequently, we employ these optimal parameters to train corresponding SVM-PSO, LSSVM-PSO, EEMD-SVM-PSO and EEMD-LSSVM-PSO models.

Table 3. The optimized value for hyperparameters of SVM and LSSVM models based EEMD.

EEMD outcomes	**EEMD-SVM-PSO parameters**		**EEMD-LSSVM-PSO parameters**	
	C	$\boldsymbol{\sigma}_{\text{svm}}$	$\boldsymbol{\gamma}$	$\boldsymbol{\sigma}_{\text{lssvm}}$
Imf 1	9.565	0.265	15.495	0.903
Imf 2	7.659	0.766	15.032	0.886
Imf 3	5.659	0.258	4.984	0.786
Imf 4	5.698	0.289	3.197	0.319
Imf 5	4.265	0.167	8.065	0.194
Imf 6	4.159	0.749	9.936	0.495
Imf 7	4.058	0.411	7.597	0.334
Imf 8	3.019	0.564	6.968	0.612
Residual	2.148	0.496	6.749	0.789

Table 4. The optimized value for hyperparameters of SVM and LSSVM models.

SVM-PSO parameters		**LSSVM-PSO parameters**	
C	$\boldsymbol{\sigma}_{\text{svm}}$	$\boldsymbol{\gamma}$	$\boldsymbol{\sigma}_{\text{lssvm}}$
15.549	0.964	18.659	0.580

4.6. Discussion of experimental results of different methods

The experimental results will be discussed in detail in this section. This paper claims that EEMD decomposition method has performed effectively to verify the superiority of the EEMD-SVM-PSO and EEMD-LSSVM-PSO for the model to predict the future water level. Four forecasting models will be compared, namely SVM-PSO, LSSVM-PSO, EEMD-SVM-PSO, and EEMD-LSSVM-PSO, to

determine the best model that improves forecasting accuracy. Figure 6 shows a radar plot of RMSE values of IMFs and residual of SVM and LSSVM. Figure 7 shows radar plot of RMSE, MAPE, MSE and R^2 values in testing phase of all proposed models.

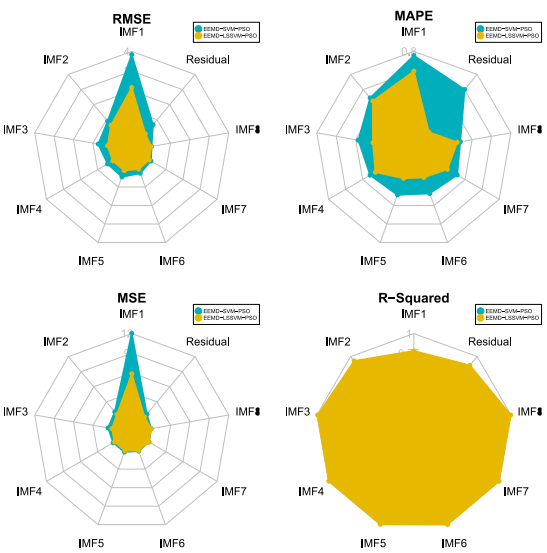

Fig. 6. Radar plot of RMSE, MAPE, MSE and R2 values of IMFs and residual of EEMD-SVM-PSO and EEMD-LSSVM-PSO.

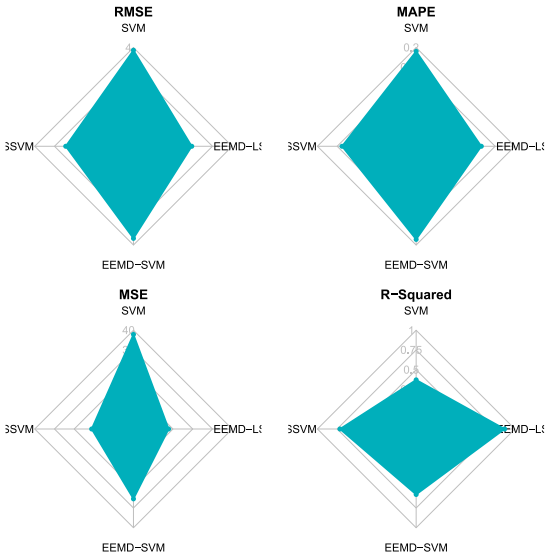

Fig. 7. Radar plot by using RMSE, MAPE, MSE and R-squared values in testing phase of proposed models.

A radar plot, also known as a spider plot or star plot, is an effective visualization tool for comparing multiple variables across different categories. The radar plot in Figure 6 illustrates the RMSE, MAPE, and MSE values on each subplot, representing the performance of two models across various Intrinsic Mode Functions (IMFs). Each subplot corresponds to one of the error metrics (RMSE, MAPE, MSE, R_2), while the lines represent the performance of the two decomposition models (EEMD-SVM-PSO and EEMD-LSSVM-PSO). The RMSE, MAPE, MSE values for the EEMD-SVM-PSO and EEMD-LSSVM-PSO models decrease from IMF1 to IMF8. These values represent the error in predicting each IMF. The higher RMSE in IMF1 suggests that the model struggles more with predicting high-frequency components, which are typically noisier and more volatile. Conversely, the lower RMSE, MAPE, MSE values for the subsequent IMFs indicate better prediction accuracy for lower frequency components. The residual RMSE is relatively low after applying the SVM and LSSVM models to each. This indicates a highly effective decomposition and prediction process where most of the variability in the data is captured by the model. Both models show higher RMSE, MAPE, MSE values for high-frequency IMFs and lower RMSE, MAPE, MSE values for low-frequency IMFs. This is expected, as high-frequency components are more challenging to predict due to their noise. The EEMD-LSSVM-PSO model, however, demonstrates a superior ability to handle these complex, high-frequency components more effectively than the EEMD-SVM-PSO model. Each model's R_2 value represents the proportion of variance in the water level data explained by the model. The distance from the center of the plot to each data point on the axis represents the R_2 value for that model. A longer distance from the center indicates a higher R_2 value, indicating better performance in explaining variance in the data. The R_2 radar graph clearly demonstrates that the EEMD-LSSVM-PSO model provides the best performance in predicting water levels in Sungai Gombak. Its high R_2 value indicates a robust capability to explain a significant proportion of the variance in the water level data, making it a reliable tool for accurate forecasting. The integration of EEMD enhances the model by decomposing the time series into more manageable components, while LSSVM improves the model's predictive accuracy, and PSO optimization further refines the model's performance. This study highlights the importance of combining advanced decomposition techniques with powerful predictive models and optimization strategies to achieve high accuracy in water level predictions.

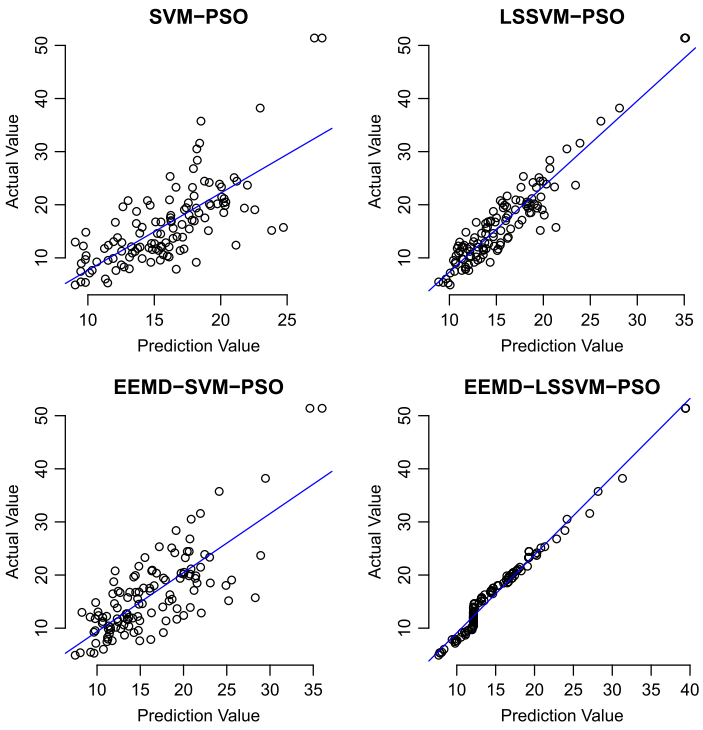

Fig. 8. The scatterplots for every models.

The radar plot in Figure 7 illustrates the RMSE, MAPE, and MSE and R^2 values on each subplot, representing the performance of four models. Each subplot corresponds to one of the error metrics (RMSE, MAPE, MSE, R^2) while the lines represent the performance of four different models SVM-PSO, LSSVM-PSO, EEMD-SVM-PSO, and EEMD-LSSVM-PSO based on testing data. The comparative analysis of RMSE values highlights the superior performance of the EEMD-LSSVM-PSO model in predicting water levels in Sungai Gombak. This model's ability to achieve the lowest RMSE,MSE and MAPE underscores its robustness in handling complex and nonlinear time series data. The significant improvement over the SVM-PSO and EEMD-SVM-PSO models can be attributed to the effective decomposition of the time series data by EEMD and the accurate prediction capabilities of LSSVM, further refined by PSO optimization.

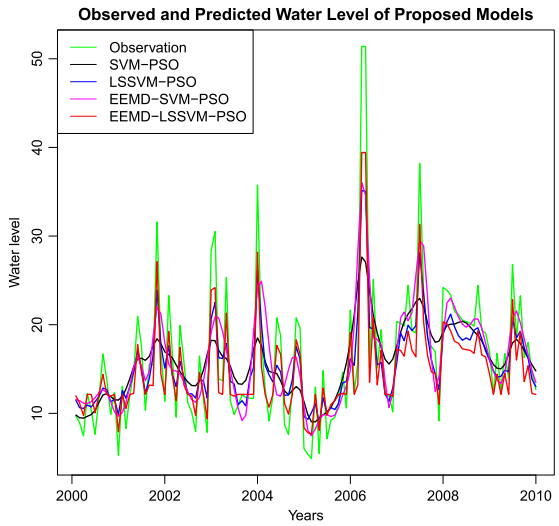

Fig. 9. Comparison of observed and predicted water level of proposed models.

Figure 8 shows the scatterplots of actual values and predicted values. By comparing the single models, namely SVM-PSO and LSSVM-PSO, the R^2 of the testing data shows that LSSVM-PSO has performed more effectively than SVM-PSO. Therefore, this proves that single LSSVM-PSO model can result in better forecasting model compared to single SVM-PSO model [14]. This is due to LSSVM-PSO high efficiency for large scale problems with parameter optimization compared to SVM-PSO and also with the help of PSO to optimize the parameters of LSSVM. To focus on the SVM models, the combined decomposition method with SVM which is called as EEMD-SVM-PSO has improved the R^2 for testing data by 65%. This indicates a good prediction by proposing EEMD-SVM-PSO that outperformed single SVM-PSO model. Next, according to the LSSVM models, after the data were decomposed by EEMD-LSSVM-PSO has better results of R^2 for testing phase enhanced by 64%. This indicates a good prediction by proposing EEMD-LSSVM-PSO that outperformed single LSSVM model [15]. Hence, the combined models (EEMD-SVM-PSO and EEMD-LSSVM-PSO) work effectively in decomposing the data by EEMD and optimize the SVM and LSSVM parameters by PSO enhances the forecasting model. The EEMD-LSSVM-PSO model exhibits the highest R-squared value among all models, indicating superior performance. Approximately 87.05% of the variance in the water level data is explained by this model. The combination of EEMD with LSSVM, further optimized by PSO, results in a model that captures the underlying patterns in the data exceptionally well.

Scatterplots comparing actual and predicted values for each model (SVM-PSO, LSSVM-PSO, EEMD-SVM-PSO, and EEMD-LSSVM-PSO) are crucial for visually assessing the performance of these predictive models. Each scatterplot typically displays the actual water level values and the predicted values. A well-performing model would have points lying close to the 45-degree line ($y = x$), indicating that the predicted values match the actual values closely. The scatterplots for the four models reveal varying levels of prediction accuracy, with the EEMD-LSSVM-PSO model demonstrating superior performance. This model's ability to produce predicted values closely aligned with actual values highlights its effectiveness in dealing with complex, nonlinear time series data. The visual analysis complements the RMSE values, providing a comprehensive understanding of each model's strengths and limitations in predicting water levels in Sungai Gombak. Figure 9 illustrates the comparison of water level prediction by the proposed models with the observed water level.

5. Conclusion

Flooding has increasingly impacted various regions over the past few decades, highlighting the urgent need for accurate forecasting models as part of an effective early warning system. In this study, we developed and evaluated a hybrid forecasting model integrating EEMD, LSSVM, and PSO to predict monthly water levels. This model harnesses the power of EEMD to decompose the data, reducing noise and capturing nonlinear variations, which are then processed by an LSSVM model optimized through PSO.

Our findings demonstrate that the EEMD-LSSVM-PSO model significantly enhances forecasting accuracy compared to traditional single models. Specifically, the EEMD-LSSVM-PSO model achieved the lowest values in RMSE, MAE, and MAPE, and an R-squared value nearing 1, outperforming the SVM-PSO and LSSVM-PSO models. This suggests that the integration of EEMD for data decomposition, coupled with LSSVM and PSO, provides a robust approach for accurate water level prediction. However, while the EEMD-LSSVM-PSO model shows promising results, it is essential to recognize its limitations and explore its performance across different contexts. Future work should address the following aspects:

(i) Advanced Decomposition Techniques: Further research could investigate the application of Complete Ensemble Empirical Mode Decomposition (CEEMD) or other advanced decomposition structures.

(ii) Alternative Optimization Algorithms: Exploring different metaheuristic algorithms such as Grey Wolf Optimization, Bat Algorithm, and Artificial Bee Colony could offer new insights into optimizing SVM and LSSVM parameters, potentially improving model performance further.

(iii) Comparative Analysis with Other Models: Comparative studies involving other machine learning models, such as Radial Basis Neural Networks and Artificial Neural Networks, could provide a broader perspective on the effectiveness and limitations of different forecasting approaches.

(iv) Practical Implementation Challenges: Future research should also consider practical aspects of deploying these models in real-world scenarios, including computational efficiency, scalability, and integration with existing flood management systems.

By addressing these areas, future research can build on the current findings to develop more sophisticated, adaptable, and practical forecasting solutions, ultimately contributing to more effective flood management and disaster preparedness.

[1] Sumi S. M., Zaman M. F., Hirose H. A rainfall forecasting method using machine learning models and its application to the Fukuoka city case. International Journal of Applied Mathematics and Computer Sciencel. **22**, 841–854 (2012).

[2] Le X.-H, Ho H. V., Lee G. River streamflow prediction using a deep neural network: a case study on the Red River, Vietnam. Korean Journal of Agricultural Science. **46** (4), 843–856 (2019).

[3] Sahoo A., Samantaray S., Ghose D. K. Prediction of flood in Barak River using hybrid machine learning approaches: A case study. Journal of Geological Society of India. **97** (2), 186–198 (2021).

[4] Shabri A., Suhartono. Streamflow forecasting using least-squares support vector machines. Hydrological Sciences Journal. **57** (7), 1275–1293 (2012).

[5] He S., Sang X., Yin J., Zheng Y., Chen H. Short-term Runoff Prediction Optimization Method Based on BGRU-BP and BLSTM-BP Neural Networks. Water Resources Management. **37**, 747—768 (2022).

[6] Belyakova P. A., Moreido V. M., Tsyplenkov A. S., Amerbaev A. N., Grechishnikova D. A., Kurochkina L. S., Filippov V. A., Makeev M. S. Forecasting Water Levels in Krasnodar Krai Rivers with the Use of Machine Learning. Water Resources. **49**, 10–22 (2022).

[7] Faruq A., Marto A., Abdullah S. S. Flood forecasting of Malaysia Kelantan river using support vector regression technique. Computer Systems Science and Engineering. **39** (3), 297–306 (2021).

[8] Ab Razak N. H., Aris A. Z., Ramli A. F., Looi L. J., Juahir H. Temporal flood incidence forecasting for Segamat River (Malaysia) using auto-regressive integrated moving average modelling. Journal of Flood Risk Management. **11**, S794–S804 (2016).

[9] Darabi H., Torabi Haghighi A., Rahmati O., Jalali Shahrood A., Rouzbeh S., Pradhan B., Tien Bui D. A hybridized model based on neural network and swarm intelligence-grey wolf algorithm for spatial prediction of urban flood-inundation. Journal of Hydrology. **603**, 126854 (2021).

[10] Shaaban N. N, Hassan N., Mustapha A., Mostafa S. A. Comparative Performance of Supervised Learning Algorithms for Flood Prediction in Kemaman, Terengganu. Journal of Computer Science. **17** (5), 451–458 (2021).

[11] Sukanya K., Vijayakumar P. Frequency Control Approach and Load Forecasting Assessment for Wind Systems. Intelligent Automation and Soft Computing. **35** (1), 971–982 (2023).

[12] Cortes C., Vapnik V. Support-vector networks. Machine Learning. **297**, 273–297 (1995).

[13] Achite M., Jehanzaib M., Elshaboury N., Kim T. W. Evaluation of Machine Learning Techniques for Hydrological Drought Modeling: A Case Study of the Wadi Ouahrane Basin in Algeria. Water (Switzerland). **14**, (2022).

[14] Wang H. F., Hu D. Comparison of SVM and LS-SVM for regression. 2005 International Conference on Neural Networks and Brain. 279–283 (2005).

[15] Guo S., Wen Y., Zhang X., Chen H. Runoff prediction of lower Yellow River based on CEEMDAN-LSSVM-GM(1,1) model. Scientific Reports. **13**, 1511 (2023).

[16] Lin S.-W., Ying K.-C., Chen S.-C., Lee Z.-J. Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Systems with Applications. **35** (4), 1817–1824 (2008).

[17] Someetheram V., Marsani M. F., Kasihmuddin M. S. M., Jamaludin S. Z. M., Mansor M. A. Double Decomposition with Enhanced Least-Squares Support Vector Machine to Predict Water Level. Journal of Water & Climate Change. **15** (6), 2582–2594 (2024).

[18] Zeng J., Tan Z.-H., Matsunaga T., Shirai T. Generalization of Parameter Selection of SVM and LS-SVM for Regression. Machine Learning and Knowledge Extraction. **1** (2), 745–755 (2019).

[19] Zhou C., Yin K., Cao Y., Intrieri E., Ahmed B., Catani F. Displacement prediction of step-like landslide by applying a novel kernel extreme learning machine method. Landslides. **15**, 2211–2225 (2018).

[20] Deng W., Yao R., Zhao H., Yang X., Li G. A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. Soft Computing. **23**, 2445–2462 (2019).

[21] Li X., Yang S.-d., Qi J.-x. A new support vector machine optimized by improved particle swarm optimization and its application. Journal of Central South University of Technology. **13**, 568–572 (2006).

[22] Brits R., Engelbrecht A. P., van den Bergh F. Locating multiple optima using particle swarm optimization. Applied Mathematics and Computation. **189** (2), 1859–1883 (2007).

[23] Sarath K. N. V. D., Ravi V. Association rule mining using binary particle swarm optimization. Engineering Applications of Artificial Intelligence. **26** (8), 1832–1840 (2013).

[24] Ostadrahimi L, Mariño M. A., Afshar A. Multi-reservoir operation rules: multi-swarm PSO-based optimization approach. Water Resources Management. **26**, 407–427 (2012).

[25] Huang N. E., Shen Z., Long S. R., Wu M. C., Shih H. H., Zheng Q., Yen N.-C., Tung C. C., Liu H. H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences. **454** (1971), 903–995 (1998).

[26] Stajuda M., García Cava D., Liśkiewicz G. Aerodynamic instabilities detection via empirical mode decomposition in centrifugal compressors. Measurement. **199**, 111496 (2022).

[27] Wu Z., Huang N. E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. Advances in Adaptive Data Analysis. **1** (1), 1–41 (2009).

[28] Wang W.-c., Chau K.-w., Xu D.-m., Chen X.-Y. Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resources Management. **29**, 2655–2675 (2015).

[29] Wang W.-c., Chau K.-w., Qiu L., Chen Y.-b. Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. Environmental Research. **139**, 46–54 (2015).

[30] Wang W.-c., Xu D.-m., Chau K.-w., Chen S. Improved annual rainfall-runoff forecasting using PSO-SVM. Journal of Hydroinformatics. **15** (4), 1377–1390 (2013).

[31] Liu H., Tian H.-q., Li Y.-f. Comparison of new hybrid FEEMD-MLP, FEEMD-ANFIS, wavelet packet-MLP and wavelet packet ANFIS for wind speed predictions. Energy Conversion and Management. **89**, 1–11 (2015).

[32] Duan W. Y., Han Y., Huang L. M., Zhao B. B., Wang M. H. A hybrid EMD-SVR model for the short-term prediction of significant wave height. Ocean Engineering. **124**, 54–73 (2016).

[33] Napolitano G., Serinaldi F., See L. Impact of EMD decomposition and random initialization of weights in ANN hindcasting of daily streamflow series: An empirical examination. Journal of Hydrology. **406** (3–4), 199–214 (2011).

[34] Karthikeyan L., Nagesh Kumar D. Predictability of nonstationary time series using wavelet and EMD based ARMA models. Journal of Hydrology. **502**, 103–119 (2013).

[35] Hawinkel P., Swinnen E., Lhermitte S., Verbist B., Van Orshoven J., Muys B. A time series processing tool to extract climate-driven interannual vegetation dynamics using Ensemble Empirical Mode Decomposition (EEMD). Remote Sensing of Environment. **169**, 375–389 (2015).

[36] Faruq A., Marto A., Abdullah S. S. Flood forecasting of Malaysia Kelantan River using support vector regression technique. Computer Systems Science and Engineering. **39** (3), 297–306 (2021).

[37] Qin Y., Lei Y., Gong X., Ju W. A model involving meteorological factors for short-to medium-term water-level predictions of small- and medium-sized urban rivers. Natural Hazards. 1–15 (2021).

[38] Samantaray S., Sahoo A. Prediction of flow discharge in Mahanadi River Basin, India, based on novel hybrid SVM approaches. Environment, Development and Sustainability. **26** (7), 18699–18723 (2024).

[39] Heddam S., Vishwakarma D. K., Abed S. A., Sharma P., Al-Ansari N., Alataway A., Dewidar A. Z., Mattar M. A. Hybrid river stage forecasting based on machine learning with empirical mode decomposition. Applied Water Science. **14** (3), 46 (2024).

[40] Huang S., Chang J., Huang Q., Chen Y. Monthly streamflow prediction using modified EMD-based support vector machine. Journal of Hydrology. **511**, 764–775 (2014).

[41] Dehghan Y., Amini Zenooz S. M., Pour Z. F. Analysis of sea level fluctuations around the Australian coast with anomaly time series analysis approach. Marine Environmental Research. **181**, 105742 (2022).

[42] Guhathakurta K., Mukherjee I., Chowdhury A. R. Empirical mode decomposition analysis of two different financial time series and their comparison. Chaos, Solitons & Fractals. **37** (4), 1214–1227 (2008).

[43] Flandrin P., Rilling G., Goncalves P. Empirical mode decomposition as a filter bank. IEEE Signal Processing Letters. **11** (2), 112–114 (2003).

[44] Colominas M. A., Schlotthauer G., Torres M. E. An unconstrained optimization approach to empirical mode decomposition. Digital Signal Processing. **40**, 164–175 (2015).

[45] Anuwar F. H., Omar A. M. Future solar irradiance prediction using least square support vector machine. International Journal on Advanced Science, Engineering and Information Technology. **6** (4), 520–523 (2016).

[46] Kennedy J., Eberhart R. Particle swarm optimization. Proceedings of ICNN'95 – International Conference on Neural Networks. 1942–1948 (1995).

[47] Chau K. W. Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River. Journal of Hydrology. **329** (3–4), 363–367 (2006).

[48] Wong W. M., Lee M. Y., Azman A. S., Rose L. A. F. Development of short-term flood forecast using ARIMA. International Journal of Mathematical Models and Methods in Applied Sciences. **15**, 68–75 (2021).

[49] Dickey D. A., Fuller W. A. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association. **74** (366a), 427–431 (1979).

[50] Kwiatkowski D., Phillips P. C. B., Schmidt P., Shin Y. Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics. **54** (1–3), 159–178 (1992).

[51] Wang W.-C., Chau K.-W., Cheng C.-T., Qiu L. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. Journal of Hydrology. **374** (3–4), 294–306 (2009).

[52] Wang D., Cui X., Niu D. Wind power forecasting based on LSTM improved by EMD-PCA-RF. Sustainability. **14** (12), 7307 (2022).

[53] Marsani M. F., Shabri A. Non-stationary in extreme share return: World indices application. ASM Science Journal. **13**, 1–9 (2020).

Гібридна машина опорних векторів з найменшими квадратами для прогнозування рівня води

Сометерам В.[1], Марсані М. Ф.[1], Касімуддін М. С. М.[1], Замрі Н. Е.[2]

[1] *Школа математичних наук, Університет Сайнс Малайзії, 11800 USM, Пенанг, Малайзія*

[2] *Кафедра математики та статистики, факультет природничих наук, Університет Путра Малайзії, 43400 UPM, Серданг, Селангор, Малайзія*

Попередні дослідження підкреслили значну роль існуючих даних про рівень води в прогнозуванні повеней. У цьому дослідженні порівнюються дві автономні моделі, машину опорних векторів (SVM) і машину опорних векторів з найменшими квадратами (LSSVM) з гібридними моделями Ensemble Empirical Mode Decomposition з SVM і LSSVM, щоб розробити більш ефективну модель прогнозування для гідрологічних даних. Оптимізація роїв частинок (PSO) включена в ці гібридні моделі для оптимізації параметрів SVM і LSSVM, у результаті чого створюються моделі під назвами SVM-PSO, LSSVM-PSO, EEMD-SVM-PSO і EEMD-LSSVM-PSO. Це дослідження зосереджено на прогнозуванні рівня води в Сунгай Гомбак, Малайзія. Ефективність запропонованих моделей була оцінена та порівняна за допомогою кількох показників, включаючи RMSE, MSE, MAPE та квадратичний коефіцієнт кореляції. Результати показують, що модель EEMD-LSSVM-PSO перевершила інші моделі, продемонструвавши найвищу точність прогнозування для Сунгай Гомбак, Малайзія, з найнижчими значеннями RMSE, MSE, MAPE та квадратичним коефіцієнтом кореляції, близьким до 1 для тестових даних.

Ключові слова: *машинне навчання; прогнозні моделі; статистичний метод; прогноз рівня води; прогнозування повеней.*