International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 1
ISSN: 2088-5334

Rainfall Prediction Using Statistical Downscaling Based on Support Vector Machine in Selangor

Nur Farah Amieera Mat Hussin [a], Shazlyn Milleana Shaharudin [a,*], Nurul Ainina Filza Sulaiman [b], Noor Hamizah Mohamad Sani [a], Sumayyah Aimi Mohd Najib [c], Hairulnizam Mahdin [d], Mohd Saiful Samsudin [e], Rasyidah [f]

[a] *Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Perak, Malaysia*
[b] *Department of Mechanical &Manufacturing, Kolej Vokasional Besut, Kampung Raja,Besut, Terengganu, Malaysia*
[c] *Faculty of Human Sciences, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia*
[d] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia*
[e] *Environmental Technology Division, School of Industrial Technology, Universiti Sains Malaysia, Gelugor, Penang, Malaysia*
[f] *Department of Information Technology, Politeknik Negeri Padang, Sumatera Barat, Indonesia*

Corresponding author: [*]shazlyn@fsmt.upsi.edu.my

Abstract*—Global climate change gains notoriety in literature discussions for potentially triggering extreme change intensity and regularity, like floods and droughts. In this study, the amount of daily rainfall in Selangor was predicted using a downscaling model based on the machine learning technique of the Support Vector Machine (SVM) approach. The collected atmospheric data (predictor) and daily rainfall data (predictand) between 2008 and 2018 used incorporate five imputation methods: mean imputation, K-nearest neighbor, Nonlinear Iterative Partial Least Squares (NIPALS) algorithm, Markov Chain Monte Carlo (MCMC) multiple imputation algorithm, and Expectation Maximization (EM) algorithm. The predictor selection was obtained using Principal Component Analysis (PCA). Primarily, gamma, cost, and epsilon were determined using K-fold cross-validation. Once the parameter value was identified, varying kernel types (linear, RBF, polynomial, and sigmoid) allowed the SVM performance as a regression model to be measured. The SVM model was developed by first handling missing data using imputation methods. The model generating the lowest RMSE value performs best because the difference of the estimated and observed value is minor. PCA efficiently reduced data dimension while retaining key variabilities. The SVM model with a Radial Basis Function (RBF) kernel outperformed others in predicting daily rainfall by displaying the lowest RMSE during calibration (13.95071) and validation (12.60423). The most fitting parameter set for the SVM model is C set to 4.00, γ set to 1.935, and ε set to 0.2. Based on the study, the SVM model performance is limited when applied to this dataset. For future studies, exploring advanced imputation techniques and broadening the methodology to other tropical climates for broader applicability are recommended.

***Keywords*—Support vector machine; principal component analysis; gamma; cost; K-fold cross-validation.**

Manuscript received 4 Mar. 2024; revised 26 Aug. 2024; accepted 18 Dec. 2024. Date of publication 28 Feb. 2025.

I. INTRODUCTION

Climate change globally acquires notoriety in literature discussions due to its potentially critical consequences on the Earth's environment. Generally, it has been acknowledged the surge in extreme climate intensity and regularity, like in the examples of floods and droughts, is potentially triggered by global climate change [1]. Climate prediction models and statistical analysis could approximate future rainfall data in this regard. These climate data, including rainfall, temperature, and wind speed, could be predicted via many climate prediction models [2], [3]. However, General Circulation Models (GCMs) and Support Vector Machines (SVMs) are extensively utilized in various studies to forecast climate properties, such as temperature and rainfall. For classification and regression problems, SVM is an effective machine-learning technique. Selangor, a river basin near Kuala Lumpur, the capital of Malaysia, produces about 70% of the city's water needs for domestic and industrial use. There are two water supply dams in the Selangor River basin: the Tinggi

Dam and the Selangor Dam. A water intake is located 21 and 42 km downstream of the Tinggi and Selangor dams. During the rainy season, when uncontrolled flow downstream of the dam is sufficient for abstraction, no discharge from the dam is required. However, releases are necessary during the dry season when downstream flows fall below normal. To achieve this goal, flow forecasting is needed so that future flows, based on specific time intervals, are estimated using current rainfall, and river flow can be known. In addition, drought monitoring, recognition, and forecasting play a vital in the planning and management of natural resources and the water resource system in this country, as well as avoiding any flash floods, water rationing problems, or the destruction of river water in the surrounding areas in Selangor.

This study was conducted to help and estimate changes in rainfall in the future to find appropriate measures to reduce the problem of this natural disaster, this is why this study was conducted so that every issue that arises as a result of an increase in the amount of rain can be overcome. Hence, this study seeks to predict the daily rainfall amount in Selangor via a downscaling model based on the Support Vector Machine (SVM) approach and to find which imputation method to handle missing predictand data in Selangor. In addition, this study also incorporated five types of appropriate imputation methods as part of its methodology. It used K-fold cross-validation before applying the following learning technique: the Support Vector Machine (SVM). Besides, this study also aims to investigate the precipitation forecast on the data set acquired from Malaysia's Department of Irrigation and Drainage (DID) using a support vector machine. This study was also conducted in Selangor, consisting of several stations on the map. Additionally, the data considered in this study were taken daily for 10 years, from 2008 to 2018.

Therefore, the drive of this study is to help estimate future rainfall revolutions to find suitable measures to reduce the problem of this natural disaster. This is why this study was carried out so that every issue arising from increased rainfall can be overcome. When identifying the most appropriate machine learning model, selecting predictors (atmospheric variables) is one of the issues to address. A good predictor must be informative, and the relationship between the predictor and the prediction (local climate factor) should be stationary. Dimensionality reduction methods like MCA, PCA, and independent component analysis can be used to identify informative predictors. Predictor selection is also based on interactive mode-fitting approaches. Selecting the right dimension reduction method saves time and effort in selecting and analyzing relevant features. Moreover, a dimensionality reduction approach can identify and extract relevant characteristics (predictors) to enable faster predictor analysis while requiring less information. At the same time, using a dimensionality reduction strategy helps extract a small set of significant features that express a considerable data set [4]. Therefore, researchers studying climate change impacts will find it helpful to use the appropriate statistical downscaling methods that integrate predictor selection mechanisms. This study aims to highlight climate change-related changes in rainfall events using Support Vector Machines. Missing data analysis, fitting predictors selection, historical data calibration, and pre-selected predictors validation were performed. As such, most Malaysian studies revolving around climate change focus on the effects of climate change instead of a specific parameter like rainfall and imminent trends.

II. Material and Methods

A. Missing Value

The imputation of missing values poses a challenge when dealing with machine learning and data mining [5]. The bias generated by missing values may affect the quality of the mining outcome. These imputation techniques seek to estimate population parameters accurately to ensure that the power of data mining and data analysis techniques is retained. This study also incorporated five types of appropriate imputation methods as parts of its methodology, which are the mean imputation method, K-nearest neighbor approach, NIPALS (Nonlinear Iterative Partial Least Squares) algorithm, MCMC (Markov Chain Monte Carlo) multiple imputation algorithm, and EM (Expectation Maximization) algorithm.

Mean imputation is commonly employed to replace data [25]. Depending on data distribution, the missing values are substituted with the sample mean, median, or mode. For large missing values, each value is imputed with an equal imputation value, the mean, resulting in a change in distribution shape. The more missing values, the smaller the standard deviation will be. Besides, the K-nearest neighbor (KNN) approach is a non-parametric classification algorithm that makes no assumptions about the elementary dataset [26]. It is known for its simplicity and effectiveness. The Expectation Maximization (EM) algorithm achieves maximum parameters' likelihood estimates when missing data occurs. However, the EM algorithm is also more frequently applied when there is unobserved latent, where the data's purpose was never for observation [27]. The NIPALS algorithm is applied to the dataset, while a PCA model predicts missing values [28]. MCMC method Monte Carlo simulation was used for imputation. In the MCMC method, the expectation-maximization (EM) technique attains maximum likelihood estimates for missing data substitution [29].

B. Principal Component Analysis

The principal component analysis (PCA) is significant in feature extraction and data compression. A large data matrix can also be reduced to a smaller dimension by maintaining mostly the original variability [6]. The PCA method is standard in data compression and feature selection. Regarding feature selection, a data space transforms into a reduced-dimensional feature space. In the consequent section, we will briefly discuss some basic PCA knowledge. According to the varimax rotation method, only PCs with eigenvectors with values bigger than one can be utilized to generate new variables known as varifactors (VF) or factor loadings. In contrast to factor loading, factor scores describe the transformations of observations according to the original variables. In contrast to factor loading, factor scores describe the transformations of observations according to the original variables. The factor loading correlation is computed with the following formula:

$$c_j = \sum_{p=1}^{m} |a_{pj}| \quad (1)$$

where a_p, as the eigenvector, denotes the j^{th} entry with $j = 1,2,\dots,N$ and $p = 1,2,\dots,m$. $|\boldsymbol{a_{pj}}|$ represents the absolute values of $\boldsymbol{a_{pj}}$. In the correlation of factor loading, c_j sorts in descending order and uses d_j to store the order. To conclude, the steps involved in the PCA algorithm are firstly to obtain the input data matrix. Secondly, it centers the data matrix by subtracting the mean from all the observations. Thirdly, it generates a correlated database in the form of matrix correlation. Fourth, it calculates the eigenvalues and eigenvectors of the correlation matrix in PCA, and lastly, it selects the eigenvectors corresponding to the largest eigenvalues (more than 1).

C. Turning Parameter

The methods of cross-validation and repetitive random subsampling are similar, but in the sampling for the cross-validation, no two steps overlap [30]. The KCV technique divides a dataset into k-independent subsets, with all but one of these subsets employed to train a classifier while the others evaluate the generalization error. The K-fold cross-validation was run multiple times to increase the number of estimates. In K-fold cross-validation, the sample data was classified into K disjoint subset $t_h (h = 1,2,3,\dots,K)$ of equivalent size. As a term, "fold" signifies the number of resulting subsets. The mean of the K value signifies the CV estimates of extra-sample error. Denoted by $c_h (h = 1,2,\dots,K)$, the training set was achieved by excluding the h^{th} subset, t_h.Let $m = \frac{n}{K}$ be the number of a subset's units. The CV-estimator is the average error of the K analyses, formulated as follows [7]:

$$err^{cv} = \frac{1}{K}\sum_{n=1}^{K}\frac{1}{m}\sum_{j\in t_h} L\left(y_j, \hat{f}c^h(x_j)\right) \qquad (2)$$

where err^{cv} is the CV-estimator, L indicates the loss function, and $\hat{f}c^h$ shows the estimated function of random covariates. K-fold CV is known to be a biased estimate of Err and the bias will be decreased by cumulating the number of folds [13]. If the training sets $c^1, c^2, \dots, c^k$ are the samples of size $(n-m)$ and the interval of the training set on a different sample is estimated, then err^{cv} becomes an unbiased estimator of Err for sample size $(n - m)$. Thus, the approximate estimate obtains of $E_c\left[Err\left(\hat{f}_c\right)\right] \approx E_c(n-m)\left[Err\left(\hat{f}_c(n-m)\right)\right]$.

D. Machine Learning

The study of computational science, machine learning (ML), involves analyzing and interpreting patterns and structures in data to improve outcomes and decision-making processes. It is an automated process, and the algorithm learns over time while gaining experience. The models created by MLs are not just able to analyze large volumes of complex data but are also capable of producing appropriate results [8]. The steps involved in constructing machine learning models can be divided into two phases: calibration and validation. A data set should be divided into two periods to minimize discrepancies and understand the enhanced model's features. Based on the study by Hadipour et al. [9], the dataset was divided into 70% for training and 30% for testing, following an appropriate percentage distribution for statistical scale reduction development.

The ML Model algorithm was primarily trained with a set of exercises via hyperparameter value substitution. Meanwhile, the model's performance was evaluated using the current calibration period of the statistical measurement. The most appropriate hyperparameters selected will be employed to test the data during the validation period. Then, test data were evaluated using a range of kernel types, followed by the selection of hyperparameters. After that, the model will produce a prediction. Fig. 1 shows the complete ML procedure used in this study.

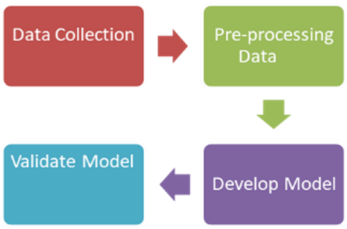

Fig. 1 Machine Learning procedure of this study

E. Support Vector Machine

SVM functions as a two-layer neural network and is effective for both linear and non-linear regression. The principle of SVM is based on statistical learning theory and the structural risk minimization approach. For Support Vector Regression, mapping the original data into a new feature space can be performed via the function of φ. The regression prediction function is formulated as follows [10]:

$$f(x) = w^T\varphi(x_i) + b \qquad (3)$$

where w and b are attained using the solution of the following optimizing problem:

Min

$$\frac{1}{2}\|w\|^2 + c\sum_{t=1}^{t}(\xi_i + \xi_i^*) \qquad (4)$$

Subject to

$$(\langle w^t\phi(x_i)\rangle + b) - y_i \le \varepsilon + \xi_i$$
$$(\langle w^t\phi(x_i)\rangle + b) - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0 \ (i = 1,2,\dots,n)$$

Regarding Equation 4, the first item simplifies the function, aiding in refining generalization. The C parameter, known as penalty terms, indicates the penalty for an experimental error. If the C value is small, the penalty for experimental error is similarly small. Hence, there is less fit and forecasting error. Also, if C shows a large value, the penalty for experimental error becomes extra-large, leading to over-learning. ε represents the non-sensitive loss function with a positive constant. The difference between the predictive value, $f(x)$, and the real value, yi is disregarded when less than the value of ε. However, the error becomes $|f(x_i) - y_i| - \varepsilon$ when an error difference is more than ε. The ε signifies the error expectation between the predictive and the real value. For the lower value of ε, the demanding error is also less but with a greater prediction precision. The non-linear data issues could be simplified by introducing the Lagrange and kernel functions. Equation 4 was translated into the following quadratic function maximizing problem.

Max

$$\sum_{l=1}^{l}(\alpha_i^* + \alpha_i)y_i - \varepsilon\sum_{l=1}^{l}(\alpha_i^* + \alpha_i) \\ \frac{1}{2}\sum_{l=1}^{l}(\alpha_i^* + \alpha_i)(\alpha_j^* + \alpha_{ji})K(x_i, x_j) \quad (5)$$

Subject to

$$0 \le \alpha_i, \alpha_i^* \le C \qquad (i = 1,2,\dots,n)$$

$$\sum_{l=1}^{l}(\alpha_i^* - \alpha_i^*) = 0$$

Therefore, the final prediction function will be as follows:

$$f(x, \alpha_i, \alpha_i^*) = \sum_{l=1}^{l}(\alpha_i - \alpha_i^*)K(x_i, x) + b \quad (6)$$

In Equation 5 and Equation 6, the function of $K(x_i, x)$ represents the kernel function. The kernel function, $K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j)$ was presented for computing the inner products. Some kernels are possibly used in the SVM; however, the standard kernels employed are polynomial, linear, radial basis function, and sigmoid. The most prevalent and competent kernel is the radial basis function with parameter γ, which is the kernel parameter gamma. Multiple experiments have demonstrated that the gamma value significantly influences the performance of the SVM model [11]. A high gamma value may decrease the structural risk and result in a more slippery function curve (while maximizing the experimental error). Also, the over-small gamma will result in an over-fitting model. In the Support Vector Machine for regression, the parameters that will be chosen when the model set up are penalty terms C, non-sensitive loss, ε, and kernel parameter γ. Fig. 2 illustrates the essential steps for ensuring the accuracy of SVR models, which have been simplified as follows.

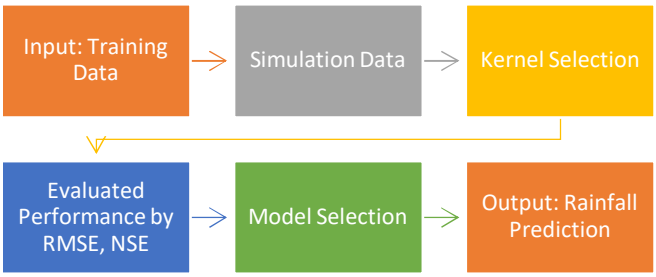

Fig. 2 Steps of SVR Model

F. Evaluation Model Performance Assessment

After the model is developed, its performance is rigorously evaluated by comparing rainfall forecast values using model performance metrics. SVM performance was evaluated using Root Mean Error (RMSE) and was compared among six imputation methods. Using RMSE, high-flow values can also be analyzed for suitability and relevance. The advantage of using RMSE is that the error is unbiased and follows a normal distribution. As a result, RMSE is generally better at illustrating model performance differences because it gives more weight to adverse conditions [12]. It is widely recognized that a higher RMSE value indicates a greater prediction error. The mathematical expression of RMSE is stated below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - y_t)^2} \quad (7)$$

where x_t is the measured value of rainfall, y_t is the predicted value of rainfall, and n is the number of data sets.

III. Results and Discussion

A. Handling Missing Predictors Data

There are six different imputation methods used to deal with the challenge of missing data, including the mean imputation method, nearest neighbor approach, NIPALS algorithm, MCMC (Markov Chain Monte Carlo), multiple imputation algorithm, and also EM (Expectation Maximization) algorithm. We then compared the error values using Root Mean Square Error (RMSE), and the results were generated according to lower RMSE values among the five methods used. As a result, the model generating the lowest RMSE value performs best because the estimated and observed value differences will be minor. Therefore, we chose the mean imputation method with the lowest RMSE value to substitute the missing data in the predictor dataset by calculating the RMSE value of each predictor variable.

TABLE I
RMSE VALUE OF PREDICTOR VARIABLES FOR EACH IMPUTATION METHODS

Missing Value	RMSE
Mean Imputation	0.751*
Nipals	1.334
MCMC	1.631
Em Algorithm	2.225
Nearest	1.686

*Indicates the lowest value of RMSE

B. Selection of Predictor Variables and Reduction of Dimensional Data

In the high-dimensional data reduction stage, PCA has been used as one of the methods in the dimensionality reduction approach for predictors' amount reduction via the extraction of the number of principal components (PC) with no critical information or data loss. Table II demonstrates the analysis results, including eigenvalues, variances, and cumulative percentages of variance.

Six components were extracted from Table II, considering the eigenvalues and total variation. Variation percentages and eigenvalues are shown in descending order. Nonetheless, according to the Kaiser criteria, eigenvalues beyond 1.00 will be opted to interpret the components [13]. Next, the results display eigenvalues greater than 1.00 for Components 1, 2, and 3, which are 1.595, 1.077, and 1.008, respectively. Nevertheless, eigenvalues approaching 1.00 also need to be considered, such as components 4 and 5, which are 0.973 and 0.88, but other factors, such as cumulative percentage values, must be considered. The rule of thumb here is that the cumulative percentage explained by the components must be at least 70%. According to [14], Calinski and Harabasz, an index is most appropriate to determine the best number of clusters, and a PCA score between 65% and 70% produces the most reasonable number of clusters. Therefore, components 4

and 5 are also considered because they have a cumulative value of 77.55% and 92.32%.

TABLE II
RESULTS OF PRINCIPLE COMPONENTS (PC'S)

Dimension	Eigenvalue	Percentage of Variability (%)	Cumulative Percentage (%)
Component 1	1.595	26.575	26.575
Component 2	1.077	17.952	44.527
Component 3	1.008	16.799	61.327
Component 4	0.973	16.219	77.546
Component 5	0.886	14.771	92.317
Component 6	0.461	7.683	100.000

Additionally, PCA can identify factors that significantly influence each variable. According to [15], loading refers to the forecast of the original variable onto the PC subspace with alternating coefficients between the PC and the variable. As part of Principal Component Analysis (PCA), principal components (PCs) should be rotated using the varimax method to facilitate the interpretation of the relationship between PCs and the original variables [16]. Furthermore, the varimax rotation of the axes outlined by PCA produced a new factor set. Every factor mainly includes a subset of the original variable and is distributed into groups of independent variables. A substantial PC loading coefficient was deemed 'weak' when the correlation was between 0.49 and 0.30, 'moderate' if the correlation was between 0.74-0.50, and 'strong' if it was greater than 0.75 (>0.75). As a result, this study considered a positive and negative PC loading of 0.74.

TABLE III
RESULT OF EACH CORRELATION PC LOADING BETWEEN VARIABLES AND FACTORS

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Rainfall	-0.284	0.647	-0.008	0.235	0.666
Maximum temperature	0.840*	0.192	-0.137	0.087	0.070
Minimum temperature	0.628	0.508	-0.062	-0.430	-0.161
Wind	0.267	-0.368	0.576	-0.450	0.509
RH	0.182	0.216	0.767*	0.496	-0.288
SR	0.557	-0.426	-0.255	0.526	0.266

*Indicates strong loading values >0.74

Table III above presents the correlations between the principal component loadings and each predictor variable. A measure of the number of factors produced by the chosen eigenvalues is provided in Table II. Here, the highest or strongest PC loading for each predictor variable was introduced to predictor selection based on the study by [17]. A variable with a higher PC loading contributes more to a PC if it contributes more to the PC that belongs to that variable. According to Table III, Factor 1 displays a strong positive loading at maximum temperature (0.84), while the third factor has a strong positive loading at RH (0.77). Due to the sequential extraction of factors, each factor accounts for the majority of the remaining variance as possible. Therefore, a new data set was produced in matrix format: 132564 rows represent the number of days respective of the stations, and six columns noted as Factor 1, Factor 2, Factor 3, Factor 4, and Factor 5 represent the extracted factor.

C. Statistical Downscaling Model Based on Regression Approaches

This data set is also divided into 70% of the calibration and 30% of the validation phase based on the study by [18]. A demanding task in Support Vector Machine (SVM) is determining the kernel type and accurately choosing the parameters gamma (γ), penalty period (C), and epsilon (ε). The feature selection process is crucial for identifying relevant variables within the search space, while the penalty parameter (C) and kernel function settings significantly influence the performance of SVM [19]. Furthermore, [20] noted that SVM requires an optimization process to avoid overfitting the model. The convolution method has been applied to SVM to optimize the selection parameter k for multiple cross-validation (KCV). In machine learning, the statistical technique of cross-validation could estimate the skills of the model. Moreover, k-fold cross-validation is often used to evaluate SVM with hyperparameter sets. Therefore, Table IV below demonstrates that the parameter C shows the highest performance at 0.5, where the error and dispersion values show the smallest values among the other 30 iterations while γ is constant at 1.935. This finding is so because the dispersion value is the average MSE of 30 test error estimates. The model with the smallest RMSE value was selected and will be employed in SVM in this study. Upon identifying C, the optimal value of ε is attained, as presented in Table IV. According to the results, the range used values can produce a good combination of 30 parameters by considering all-time series features while maintaining the generalization performance.

Fig. 3 shows the visualization of the rotation parameters between ε and C. A darker color indicates a smaller error on the blue scale. SVM model effectiveness is calculated by averaging the error estimates over all k trials. Despite a large imbalance in the response variable, stratified variation in the error estimates throughout the range is quite similar. Hence, the most fitting parameter set for the SVM model is C=4.00, γ=1.935, and ε=0.2. Once the parameter values have been identified, varying the kernel type (linear, Radial Basis Function (RBF), polynomial, and sigmoid) will allow the performance of SVM as a regression model to be measured. Due to the importance of determining the best kernel function in SVM applications, selecting the best one is an important step. Many previous studies have investigated the aptness of distinctive kernel functions used in downscaling [21],[22].

SVM model performance is measured by the RMSE for the calibration and validation periods, as presented in Table V. Since regularization is used to improve validation accuracy during the calibration period, SVM's calibration performance is generally much better than its validation performance. Thus, the SVM models with RBF kernels had the lowest RMSE values in calibration and validation, which were 13.9507 and 12.60423. According to a study by [23], applying the RBF kernel in SVM can also reduce the computational complexity more than the Polynomial kernel in downscaling monthly precipitation. A sigmoid kernel, for example, without comparing its capabilities, will produce poor predictions without identifying the kernel type.

The calibration and validation results presented in Fig. 4 and Fig. 5 illustrate the comparison between observed daily rainfall (Predictand) and predicted rainfall using the Support Vector

Machine (SVM) model. In the calibration phase, the observed rainfall exhibits significant variability, with notable extreme peaks exceeding 2000 mm. However, the SVM-predicted values remain relatively stable and fail to capture these high-intensity rainfall events, suggesting that the model struggles with extreme precipitation while maintaining reasonable accuracy for moderate rainfall conditions. Similarly, in the validation phase, the observed rainfall maintains high variability, with multiple extreme values exceeding 500 mm. The SVM model continues to underestimate these peaks, producing more stable predictions that align with general rainfall patterns but not with extreme events. The consistency between calibration and validation results indicates that while the SVM model effectively captures overall rainfall trends, it has limitations in predicting extreme precipitation. This limitation suggests that further improvements, such as integrating ensemble learning techniques or hybrid models, may be necessary to enhance the predictive accuracy of extreme rainfall events in future studies.

It clearly shows that the daily rainfall forecast values produced by the SVM model are more or less the same and do not respond to extreme values. As you can see from the plot below, the SVM model looks different from the predicted pattern and data. It tends to flatten out as the calibration and validation processes proceed. As a result of the study by [24] that unbalanced and noisy data, SVM's performance is limited when applied to this dataset.

TABLE IV
THE RESULT OF OPTIMIZATION PARAMETER C AND ε

Parameter		Error	Dispersion
C	ε		
4.00*	0.2*	131.0784*	59.2384*
8.00	0.2	131.8495	59.5833
16.00	0.2	134.2943	59.6395
32.00	0.2	136.5345	60.5076
64.00	0.2	139.9322	60.5079
128.00	0.2	144.5168	61.1629
4.00	0.4	132.1314	54.5875
8.00	0.4	132.6679	55.2893
16.00	0.4	135.0581	55.0397
32.00	0.4	136.4924	55.1146
64.00	0.4	140.0597	55.3870
128.00	0.4	146.9883	56.8581
4.00	0.6	142.5411	49.7784
8.00	0.6	143.5305	51.0533
16.00	0.6	146.5203	51.3432
32.00	0.6	147.4821	50.4266
64.00	0.6	150.3398	51.6368
128.00	0.6	158.3768	52.5456
4.00	0.8	163.4084	44.7456
8.00	0.8	163.8017	46.0207
16.00	0.8	165.5818	45.9124
32.00	0.8	166.1530	46.1291
64.00	0.8	169.2882	49.0604
128.00	0.8	175.7528	51.0012
4.00	1.0	194.9591	39.6295
8.00	1.0	195.0145	39.9239
16.00	1.0	195.7446	40.0286
32.00	1.0	196.2499	40.1332
64.00	1.0	197.8295	43.6621
128.00	1.0	200.7167	48.3563

*Indicates the selected value

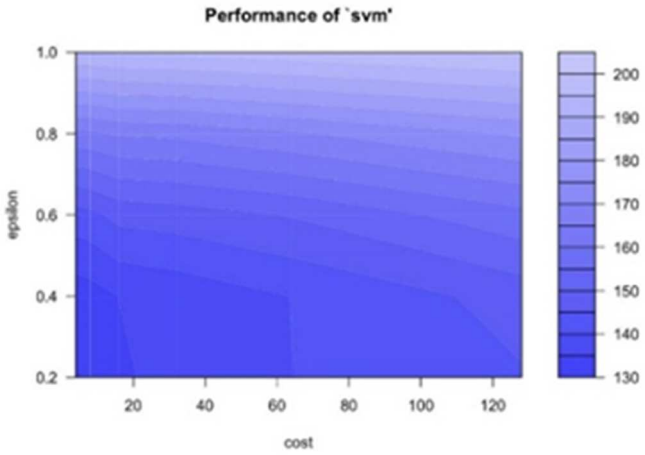

Fig. 3 Turning parameter ε by 30th -fold cross-validation

TABLE V
PERFORMANCE SVM BY VARYING THE KERNEL FUNCTION

Type of Kernel	Parameter			RMSE	
	C	γ	ε	*Calibration*	*Validation*
Linear	4	1.935	0.2	14.75032	13.72201
Polynomial	4	1.935	0.2	142.0788	263.2877
RBF	4*	1.935*	0.2*	13.95071*	12.60423*
Sigmoid	4	1.935	0.2	356360.9	168596.1

*Indicates the best kernel function

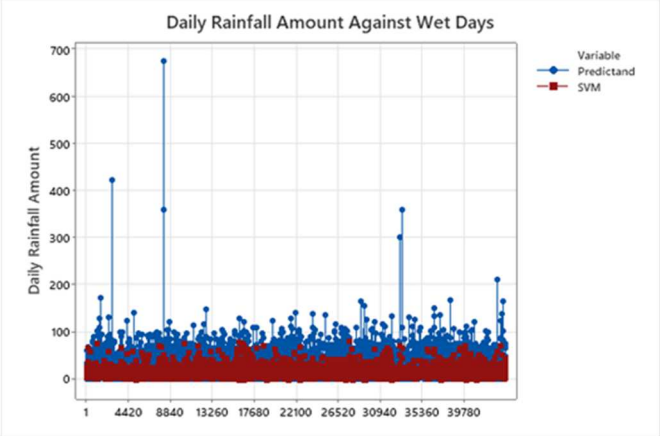

Fig. 4 Performance of SVM in predicting daily rainfall data amount in the validation period.

IV. CONCLUSION

The impact of climate change on the hydrological cycle is evident through extreme weather events, such as floods and tsunamis. In Malaysia, flooding is one of the most devastating consequences, significantly disrupting daily life, particularly in rapidly urbanizing regions. Accurate rainfall forecasting is essential to mitigate flood risks, especially in Selangor, where meteorological analysis is crucial in disaster prevention. This study addressed rainfall forecasting challenges using low-scaling techniques and correlation analysis to improve predictive accuracy.

A key focus of this study was efficiently handling missing local data. Various imputation techniques were explored, and single-value imputation was identified as the most suitable method due to the minimal occurrence of missing data in Selangor. Additionally, Principal Component Analysis (PCA)

was employed for dimensionality reduction, facilitating the identification of significant predictor variables and mitigating computational complexity associated with high-dimensional hydrological data.

Furthermore, Support Vector Machine (SVM) was applied as a statistical downscaling approach to classify wet and dry days, enabling a more precise daily rainfall forecast. The Radial Basis Function (RBF) kernel-based SVM model was selected as the optimal predictive model, following k-fold cross-validation for parameter optimization. Through this approach, the study effectively tackled key challenges related to missing data, high-dimensional datasets, and statistical downscaling, contributing to an improved understanding of rainfall patterns in Selangor.

Future research could explore more advanced imputation techniques beyond NIPALS, Expectation-Maximization (EM) Algorithm, Markov Chain Monte Carlo (MCMC), and Mean Imputation to enhance data handling efficiency. Additionally, while this study focused on Selangor, the proposed methodology is adaptable to other tropical regions in Malaysia. Given the similarities in rainfall patterns between Malaysia and other tropical or even seasonal climates, researchers are encouraged to extend this framework to broader geographical contexts.

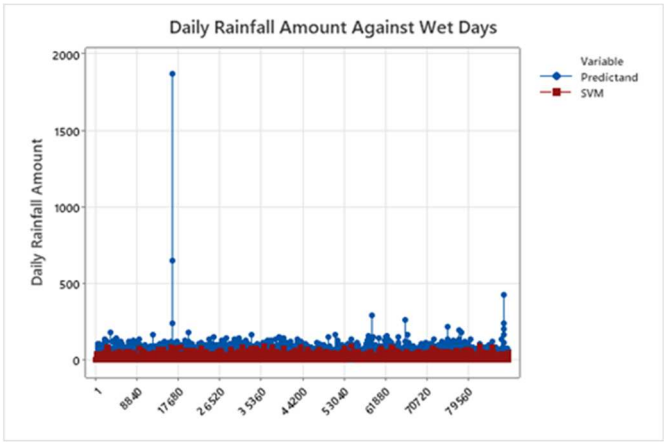

Fig. 5 Performance of SVM in predicting daily rainfall data amount in calibration period.

ACKNOWLEDGMENT

This research was supported by the Kurita Water and Environment Foundation, Japan, through the Kurita Overseas Research grant (23Pmy201) 2023-0202-101-11 and Fundamental Research Grants Scheme Vote (FRGS/1/2020/SS0/UPSI/02/11) offered by the Malaysian Ministry of Education.

REFERENCES

[1] M. H. I. Dore, "Climate change and changes in global precipitation patterns: What do we know?," *Environ. Int.*, vol. 31, no. 8, pp. 1167–1181, Oct. 2005. doi: 10.1016/j.envint.2005.03.004.

[2] K. Thorpe, R. Greenwood, A. Eivers, and M. Rutter, "Prevalence and developmental course of 'secret language,'" *Int. J. Lang. Commun. Disord.*, vol. 36, no. 1, pp. 43–62, Jan. 2001. doi: 10.1080/13682820150217563.

[3] R. Hock and B. Holmgren, "A distributed surface energy-balance model for complex topography and its application to Storglaciären, Sweden," *J. Glaciol.*, vol. 51, no. 172, pp. 25–36, 2005. doi: 10.3189/172756505781829566.

[4] O. Saini and S. Sharma, "A review on dimension reduction techniques in data mining," *Comput. Eng. Intell. Syst.*, vol. 9, no. 1, pp. 7–14, 2018.

[5] D. Zhang et al., "Comparison of NCEP-CFSR and CMADS for hydrological modelling using SWAT in the Muda River Basin, Malaysia," *Water*, vol. 12, no. 11, p. 3288, Nov. 2020. doi: 10.3390/w12113288.

[6] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Netw.*, vol. 5, no. 6, pp. 927–935, Nov. 1992. doi: 10.1016/s0893-6080(05)80089-9.

[7] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 2nd ed., vol. 1–3, Elsevier, 2024, pp. 542–545.

[8] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015. doi: 10.1126/science.aaa8415.

[9] N. L. Hadipour, M. R. Delavar, and A. M. Malekmohammadi, "A statistical scale reduction approach for geospatial data generalization," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 12, p. 221, 2016. doi: 10.3390/ijgi5120221.

[10] N. A. F. Sulaiman, S. M. Shaharudin, S. Ismail, N. H. Zainuddin, M. L. Tan, and Y. A. Jalil, "Predictive modelling of statistical downscaling based on hybrid machine learning model for daily rainfall in east-coast Peninsular Malaysia," *Symmetry*, vol. 14, no. 5, p. 927, May 2022. doi: 10.3390/sym14050927.

[11] J. E. Wang and J. Z. Qiao, "Parameter selection of SVR based on improved K-fold cross validation," *Appl. Mech. Mater.*, vol. 462–463, pp. 182–186, Nov. 2013. doi: 10.4028/www.scientific.net/amm.462-463.182.

[12] R. C. Deo, P. Samui, and D. Kim, "Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine, and multivariate adaptive regression spline models," *Stoch. Environ. Res. Risk Assess.*, vol. 30, no. 6, pp. 1769–1784, Sep. 2015. doi: 10.1007/s00477-015-1153-y.

[13] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 141–151, 1960. doi: 10.1177/001316446002000116.

[14] S. M. Shaharudin, N. Ahmad, N. H. Zainuddin, and N. S. Mohamed, "Identification of rainfall patterns on hydrological simulation using robust principal component analysis," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 11, no. 3, pp. 1162–1167, Sep. 2018. doi: 10.11591/ijeecs.v11.i3.pp1162-1167.

[15] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, 2016. doi: 10.1098/rsta.2015.0202.

[16] A. Azid et al., "Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia," *Water Air Soil Pollut.*, vol. 225, no. 8, Jul. 2014. doi: 10.1007/s11270-014-2063-1.

[17] C. W. Liu, K. H. Lin, and Y. M. Kuo, "Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan," *Sci. Total Environ.*, vol. 313, no. 1–3, pp. 77–89, Sep. 2003. doi: 10.1016/s0048-9697(02)00683-6.

[18] K. K. Golnaraghi, *Artificial Neural Networks in Hydrology*. Springer, 2014. doi: 10.1007/978-3-642-38716-1.

[19] M. Y. Cho and T. T. Hoang, "Feature selection and parameters optimization of SVM using particle swarm optimization for fault classification in power distribution systems," *Comput. Intell. Neurosci.*, vol. 2017, no. 1, pp. 1–9, 2017. doi: 10.1155/2017/4135465.

[20] A. H. Ali and M. Z. Abdullah, "An efficient model for data classification based on SVM grid parameter optimization and PSO feature weight selection," *Int. J. Integr. Eng.*, vol. 12, no. 1, pp. 1–12, Jan. 2020. doi: 10.30880/ijie.2020.12.01.001.

[21] H. S. Wheater, S. Mathur, and A. K. Gupta, "Application of statistical downscaling methods for climate change impact assessment in hydrology," *J. Hydrol.*, vol. 391, no. 1–2, pp. 1–18, 2010. doi: 10.1016/j.jhydrol.2010.07.004.

[22] M. P. Goyal and R. S. Ojha, "Downscaling of precipitation using support vector machine (SVM)," *Hydrol. Sci. J.*, vol. 57, no. 2, pp. 227–238, 2012. doi: 10.1080/02626667.2011.637042.

[23] R. K. Mishra and R. K. Desai, "Downscaling of precipitation using support vector machine with radial basis function kernel," *Theor. Appl. Climatol.*, vol. 137, pp. 1769–1784, 2019. doi: 10.1007/s00704-018-2707-6.

[24] G. Halik, N. Anwar, B. Santosa, and Edijatno, "Reservoir inflow prediction under GCM scenario downscaled by wavelet transform and

support vector machine hybrid models," *Adv. Civ. Eng.*, vol. 2015, no. 1, pp. 1–9, 2015. doi: 10.1155/2015/515376.
[25] X. Wu, H. A. Khorshidi, U. Aickelin, Z. Edib, and M. Peate, "Imputation techniques on missing values in breast cancer treatment and fertility data," *Health Inf. Sci. Syst.*, vol. 7, no. 1, Oct. 2019. doi: 10.1007/s13755-019-0082-4.
[26] S. Deng, L. Wang, S. Guan, M. Li, and L. Wang, "Non-parametric nearest neighbor classification based on global variance difference," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, Mar. 2023. doi: 10.1007/s44196-023-00200-1.
[27] M. P. Becker, I. Yang, and K. Lange, "EM algorithms without missing data," *Stat. Methods Med. Res.*, vol. 6, no. 1, pp. 38–54, Jan. 1997. doi: 10.1191/096228097677258219.
[28] A. F. Ochoa Muñoz, V. M. Gonzalez Rojas, and C. E. Pardo Turriago, "Missing data in multiple correspondence analysis under the available data principle of the NIPALS algorithm," *DYNA*, vol. 86, no. 211, pp. 249–257, Oct. 2019. doi: 10.15446/dyna.v86n211.80261.
[29] W. Ruth, "A review of Monte Carlo-based versions of the EM algorithm," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2401.00945. doi: 10.48550/arxiv.2401.00945.
[30] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, Jun. 2022. doi: 10.1016/j.icte.2021.05.001.