INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

High-Resolution Downscaling with Interpretable Relevant Vector Machine: Rainfall Prediction for Case Study in Selangor

Raghdah Rasyidah Abdul Rashid [a], Shazlyn Milleana Shaharudin [a,b,*], Nurul Ainina Filza Sulaiman [c], Nurul Hila Zainuddin [a], Hairulnizam Mahdin [d], Summayah Aimi Mohd Najib [e], Rahmat Hidayat [f]

[a] *Department of Mathematics, Faculty of Science and Mathematics Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia*
[b] *Department of Statistics, Columbia University, Amsterdam Ave, New York, 10027, United States*
[c] *Department of Mechanical &Manufacturing, Kolej Vokasional Besut, Besut, Terengganu, Malaysia*
[d] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia*
[e] *Department of Geography and Environment, Faculty of Human Sciences, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia*
[f] *Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia*
*Corresponding author: *shazlyn@fsmt.upsi.edu.my*

***Abstract*—Due to the discrepancy in resolution between existing global climate model output and the resolution required by decision-makers, there is a persistent need for climate downscaling. We conducted a study to determine the effectiveness of Relevant Vector Machine (RVM), one of the machine learning approaches, in outperforming existing statistical methods in downscaling historical rainfall data in the complex terrain of Selangor, Malaysia. While machine learning eliminates the requirement for manual feature selection when extracting significant information from predictor fields, considering multiple pivotal factors is essential. These factors include identifying relevant atmospheric features contributing to rainfall, addressing missing data, and developing a significant model to predict daily rainfall intensity using appropriate machine-learning techniques. The Principal Component Analysis (PCA) technique was employed to choose relevant environmental variables as input for the machine learning model, and various imputation methods were utilized to manage missing data, such as mean imputation and the KNN algorithm. To assess the performance of the RVM-based rainfall model, we collected a dataset from the Department of Irrigation and Drainage Malaysia. We used Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) as evaluation metrics. This study concluded that Relevance Vector Machine (RVM) models are suitable for forecasting future rainfall since they can support large rainfall extremes and generate reliable daily rainfall estimates based on rainfall extremes. In this study, the RVM model was employed to determine a predictive association between predictand variables and predictors.**

***Keywords*— Statistical downscaling; missing value; PCA; RVM; forecasting; missing data.**

Manuscript received 20 Nov. 2023; revised 14 Jan. 2024; accepted 20 Mar. 2024. Date of publication 31 May 2024.

I. INTRODUCTION

Providing trustworthy and practical climate change projections at a high spatial resolution has been a significant and persistent obstacle in climate science. The availability of such detailed projections is crucial in guiding region-specific adaptation strategies and preparedness plans to tackle potential extreme weather occurrences in the future. Almost all the research aims to improve the quality of climate information and hazard data when it affects the surrounding conditions, such as floods [1]. Climate change's persistent trajectory has led to rising temperatures, shifting precipitation patterns, and heightened climatic variability, resulting in significance to crop preference and agricultural systems [24]. One of the leading causes of drastic climate change is the modification of the distribution of atmospheric statistical patterns that can affect a large part of the earth's population. This is because the increase in temperature and rainfall affects the air supply as well as the living plants in the future. Therefore, it is crucial to know the current global climate and what the future holds to determine the weaknesses and what can be done to adapt to climate change [4]. Precipitation downscaling enhances the inadequate resolution and precipitation representation in global climate models, enabling end users to evaluate the potential hydrological effects of climate change [9].

The task of modeling precipitation modeling poses significant challenges in climate modeling, often yielding reduced accuracy through numerical approach alone. An alternative strategy for forecasting rainfall involves utilizing statistical downscaling models [20]. Statistical downscaling is an approach that combines both statistical and dynamic methods. Using machine learning, previously unsolvable problems could be solved by identifying trends in data. Several Machine Learning (ML) techniques can be used to classify data based on research [16]. For an artificially intelligent machine to make informed decisions, it must identify objects in its surroundings and forecast environmental behavior [11]. Hence, machine learning techniques tend to prioritize prediction overestimation.

In this study, the relevance vector machine represents a machine learning technique to forecast precipitation levels in climate change scenarios. Compared to SVM, RVM's most significant advantage is its ability to reduce the computational complexity of the kernel function and overcome its limitation in keeping the chosen kernel function by the Mercer condition [17]. To the contrary of its predecessor, SVM [26]. RVM incorporates a probabilistic regression algorithm. Using the theory of Bayesian inference, RVM is a compelling machine learning method. In previous studies, the RVM has been implemented successfully in various fields. However, its application to modeling rainfall-induced shallow landslides still needs to be improved [25].

As we all know, rain is Malaysia's primary agriculture source. The landscape of the Earth is experiencing numerous changes because of present-day development initiatives. Hence, the environment has transformed to accommodate the changed conditions created by human-induced alterations. Consequently, a reciprocal dynamic exists where every human action prompts a response from the environment, usually appearing in the form of interdependent processes frequently initiated by natural events. Recently, the drastically increased rainfall due to sudden climate change has also caused significant floods in Selangor.

From the findings, we can safely deduce that the main reason for the unexpected flood in December 2021 was the high amount of precipitation in the area. Major flooding caused extensive damage to many houses and even areas beyond repair. Selangor has never experienced this, and most people still need to prepare. Due to the unusual amount of rainfall distribution equal to the average amount of rainfall distribution for one month on 18 December 2021 in Selangor, the flood was exacerbated by the tidal event that occurred [2]. For example, Taman Sri Muda in Shah Alam has been severely flooded. Flooding is expected in the area, but the rapid ascent of water caught numerous residents unprepared [2].

Although downscaling literature contains several details that examine different modeling techniques, there currently needs to be a single study that examines the effectiveness of downscaling large-scale atmospheric information to catchment scale precipitation and the effectiveness of RVM models using machine learning techniques [11]. A Relevance Vector Machine (RVM) applies Bayesian treatment principles to a generalized linear model with the same functional form as the SVM RVM, which proves to be an excellent time series prediction model, even with some limitations [14]. The purpose of this study was to predict the distribution of precipitation for the state of Selangor so that the floods that occurred in December 2021, which were a nightmare for the people there, would not happen again. Hence, methods combining Machine Learning algorithms with time-series data are considered a solution to mitigate these limitations [8]. In this study, data is only available for Selangor, which comprises several districts and a few stations. The flood crisis that recently hit Selangor has limited this study to the state of Selangor alone.

A limitation of the study is that it emphasizes stations solely in Selangor in peninsular Malaysia, with no consideration of other states. In reviewing available atmospheric data, it is essential to recognize that more variables could be attributed to the server's utilization for processing high-dimensional data, particularly precipitation data. Machine learning techniques use a variety of algorithms for analysis, all of which require a lot of storage space and long iterations for large data sets. For accurate analysis of time series data, studies on machine learning require enormous computational demands. The study also stated that higher rainfall contributed to the major floods in Selangor at the end of last year. The city center of Selangor is not prepared to deal with significant flooding, unlike other states in the east that have been accustomed to dealing with such flooding. Further, this study aims to determine how rainfall distributions in Selangor will change.

II. Materials and Methods

This paper uses machine learning techniques to describe a statistical downscaling method for daily rainfall that combines classifications and regressions. A prediction model is proposed, which consists of data collected from predictors (atmospherics) and predictands (rainfalls). Using the imputation method, predictors were standardized and replenished as a pre-processing step. After that, PCA was employed to select predictor variables and reduce data dimensionality. In the next step, new data matrixes were generated between the selected predictors and predictand.

A. Missing Data

Prediction becomes challenging with the presence of missing data within a time series. Data analysis with time series is different from other forms of data analysis because of the temporal significance. An analysis of a time series without those missing values destroys its continuity. Assessing the effects of forecasting models or imputation methods on predicted outcomes proves difficult when missing values are substituted with imputation methods. Imputation methods alter the original time series and possibly affect prediction performance significantly [28].

First, addressing a multivariate time series containing missing values is inherently problematic due to numerous factors, including potential errors during collection, which may compromise subsequent analytical applications. Various techniques for imputing missing values have been introduced to minimize their influence on multivariate time series analysis. Furthermore, imputed missing values from a multivariate time series are helpful for many types of data analysis applications [23]. Coping with missing data through imputation is a critical concern in learning from incomplete datasets, with various techniques developed and successfully handling missing values [32]. Imputation as the primary technique for managing missing data entails replacing missing observations with possible values through single or multiple imputation approaches. Single imputation entails finding a probable value for each missing data point to impute, with specific contexts yielding unbiased estimates through these strategies [21].

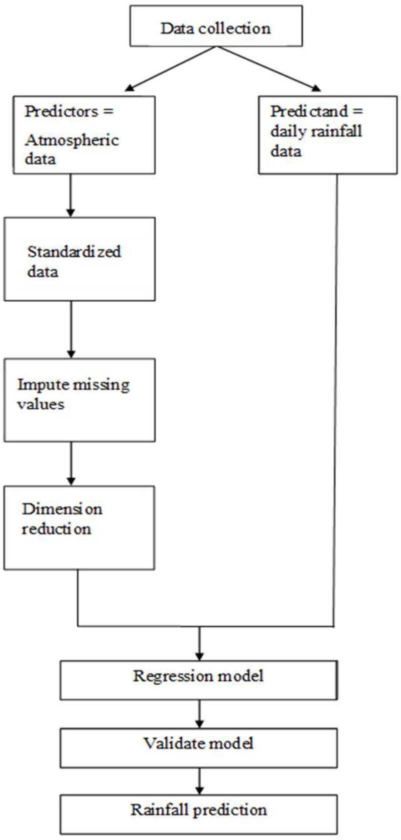

Fig. 1 A flowchart of the statistical downscaling process

Frequently, hydrologists need more data issues when processing time series data sets, which can make interpretation challenging. It can cause data distributions to be skewed, interval precisions to be compromised, and model efficiency to be significantly reduced by such issues. The presence of missing values, caused by sensor malfunctions or human errors, prompts the employment of various ad hoc techniques like deletion or single-value substitution for their management [28]. The removal and imputation of missing values are currently the most well-studied methods for handling missing values in time series data. Ensuring the accuracy of model outcomes requires identifying appropriate methods of handling missing data, such as deleting or imputing. It is not feasible to uniformly apply multiple imputations when encountering missing data. A comparison of numerous imputations and listwise deletion must consider the application-specific nature of the missing data when determining which method to use [12]. We compared several imputation methods to select the most suitable method for obtaining the entire data set for statistical downscaling for all stations in Selangor, as shown in Fig. 1.

1) EM Algorithm: Expectation-Maximization (EM) is the maximum likelihood method for creating a new data set. All missing values are imputed with values estimated by the maximum likelihood methods. This approach begins with the expectation step, during which the parameters (e.g., variances, covariance, and means) are estimated, perhaps using listwise deletion. Those estimates then create a regression equation to predict the missing data. It has long been a widespread practice to impute continuous data using the expectation-maximization algorithm (EM). Using the method assumes that the full data are multivariate normal, which is not the case for categorical data [21]. The multivariate normal distribution is a crude approximation to the true data distribution, but it can still adequately reproduce this distribution [21]. Imputing categorical data with continuous-based methods may result in biased results, so caution should be exercised. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to "fill in" the missing data. The expectation and maximization steps are performed repeatedly until the stabilization of the system is established [6].

2) Non-linear Interactive Partial Least Square (NIPALS): An alternative imputation method suitable for principal component analysis (PCA) problems involving missing values is the Non-linear Interactive Partial Least Squares (NIPALS) approach. Commercial chemometric software implementations of this approach demonstrate varying levels of correctness when handling the common PCA problem with missing values [18]. A better strategy entails substituting the sample mean into the missing value. Provided an $n \times p$ rectangular dataset, the NIPALS algorithm operates according to the following steps:

Firstly, a matrix for the i th observed value in the j th variables was defined:

$$x = \{x_{ij}\}, 1 \leq i \leq n, 1 \leq j \leq p \qquad (1)$$

Assume X has a rank of a, then decompose X as follows:

$$X = \sum_{h=1}^{a} t_h p'_h \qquad (2)$$

$$t_h = (t_{h1}, \dots, t_{hi}, \dots, t_{hn})' \; and \; p_h = (p_h, \dots, p_{hi}, \dots, p_{hn})'$$

as the principal factor and principal component, respectively. Lastly, approximate the missing value using NIPALS to the cell *(i)* as:

$$x_{ij} = \sum_{l=1}^{k} t_{ij} P_{ij} \qquad (3)$$

3) Mean Imputation: The approach to single imputation entails identifying a sole probable value for each missing data point for imputation purposes. Under specific circumstances, these strategies can yield unbiased estimates [21]. Instead of the missing data value, the mean value of a variable is substituted, ensuring that the sample mean for that variable remains unchanged. The underlying theory supporting mean substitution posits that the mean is a reasonable estimate for observations randomly chosen from a normal distribution.

According to [18], there are several advantages to using this technique, including the fact that it is not too complicated and is easy to integrate into most statistical packages. Based on the complete data, each variable's mean and standard deviation were computed. Next, they were compared with their corresponding counterparts after substituting the missing values. The absolute difference between means and standard deviations was employed to evaluate the effectiveness of both procedures. For a method to be considered efficient, it must reproduce means and standard deviations accurately. Statistics must also consider the accuracy of estimating parameters [18]. The formula can be written as:

$$P_x = \frac{\sum_{i=1}^{n} P_i}{n} \tag{4}$$

where P_x is the observed rainfall data, $P_i$is the initial rainfall data, and n denotes the number of rainfall days.

4) K-Nearest Neighbor: KNN imputation relies on identifying the nearest neighbors based on a distance measure to fill in missing attribute values. After determining the KNN, the mean/median or mode of identified attribute values of the missing attribute imputed the missing value. According to [7], through the application of KNN imputation, the handling of missing hydrological data with the KNN method has shown to be reliable and practical. The initial stage of nearest-neighbor imputation involves assuming n observations gathered on p covariates. The corresponding $n \times p$ data matrix is given by **X**=(x_{is}), in which x_{is} represents the i^{th} observation of the s^{th} variable. Let **O**=(O_{is}) signifies the corresponding $n \times p$ matrix dummies with entries as follows:

$$0_{is}\{1\ , if\ x_{is}\ was\ observed\ 0\ , for\ missing\ values$$

The calculation of the distance between two observations of x_i and x_j was made using Lq-metric from the observed data,

$$d_q(x_{i,}x_j) = [\frac{1}{m_{ij}}\sum_{s=1}^{p} \square = 1|x_{is} - x_{js}\ I(o_{is}=1)I(o_{js}=1)]^{\frac{1}{q}} \tag{5}$$

where $m_{ij} = \sum_{s=1}^{p} \square\ I(o_{is}=1)I(o_{js}=1)]$ signifies the number of valid components in the distance computation.

5) Markov Chain Monte Carlo (MCMC): Imputation was accomplished through Monte Carlo simulation or MCMC method. According to Little and Rubin (2019), the expectation-maximization (EM) calculates the maximum probable valuations for the MCMC method to substitute missing data The choice of using the MCMC method in the multi-imputation procedure was driven by the assumption of multivariate normality. According to [22] in his research, the MCMC method, following Bayesian inference principles, includes several steps, such as imputation for handling missing data. The missing values for each observation were simulated by estimating the mean and covariance matrix. After that, for the posterior step (P-step), the mean vector and covariance matrix from the imputed step were simulated.

B. Principal Component Analysis

PCA decreases the dimensionality of the constructed feature matrix, thereby minimizing linear feature correlation among data and eliminating redundant attributes. This results in a low-dimensional feature matrix that maintains significant characteristics for the classification model [32]. The step in the PCA algorithm was mainly to acquire the input data matrix. Next, the data matrix was centered on all observations' mean subtraction. Then, generate a correlated database. The correlated database is in the form of a matrix. After that, the eigenvalues and eigenvectors of the correlation matrix in PCA are calculated. Select the eigenvectors associated with the largest (more than 1) eigenvalues, calculate the contribution of factor loadings corresponding to the selected eigenvectors, and form a new data matrix for further analysis.

Assume that a dataset. $x^{(1)}, x^{(2)}, \ldots\ldots x^{m}$ with n dimension inputs, n-dimension data has to be reduced to k-dimension $(k << n)$ using PCA. Firstly, the standardization of the raw data ensures that it has unit variance and zero mean.

$$x_j^i = \frac{x_j^i - x_j}{\sigma_j}\ \forall j \tag{6}$$

Then, calculate the data co-variance matrix as follows:

$$\Sigma = \frac{1}{m}\sum_i^m \square\ (x_i)(x_i)^T, \Sigma_\square^\square \in R^{n*n}$$

Next, compute the eigenvector and eigenvalue of the co-variance matrix using the following equation:

$$U^T\Sigma_\square^\square = \lambda\mu \tag{7}$$

$$U = [|\ |\ |\ u_1\ u_2 \ldots\ldots. u_n\ |\ |\ |\ ], u_i \in R^n$$

Projection of raw data into a k-dimensional subspace involves selecting the top K eigenvectors of the covariance matrix, establishing them as the new original basis for the data [15]. The following equation outlines the calculation of the corresponding vector:

$$x_i^{new} = [u_1^T x^i\ u_2^T x^i\ \ldots\ \cdots\ u_K^T x^i\ ] \in R^K \tag{8}$$

As a result, n-dimensional data will be reduced to k-dimensional data if the raw data is of n dimensions [15]. Figure 2 demonstrates the flowchart of the PCA algorithm:

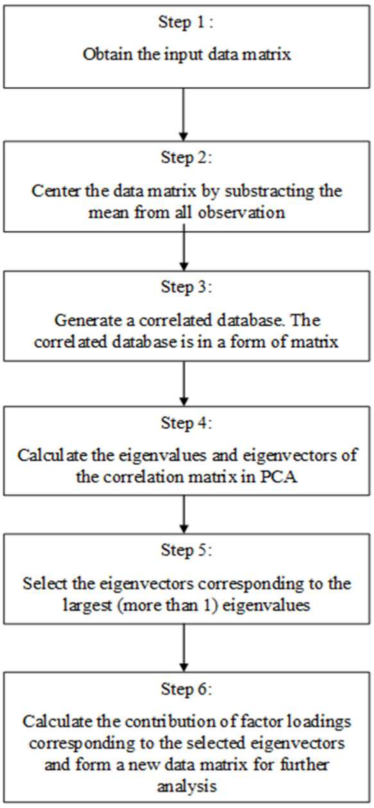
Step 1 :
Obtain the input data matrix
Step 2:
Center the data matrix by substracting the mean from all observation
Step 3:
Generate a correlated database. The correlated database is in a form of matrix
Step 4:
Calculate the eigenvalues and eigenvectors of the correlation matrix in PCA
Step 5:
Select the eigenvectors corresponding to the largest (more than 1) eigenvalues
Step 6:
Calculate the contribution of factor loadings corresponding to the selected eigenvectors and form a new data matrix for further analysis

Fig. 2 A flowchart of the PCA algorithm

C. Relevance Vector Machine

The relevance vector machine (RVM) is a new classifier based on support vector machines (SVMs) and Bayesian approaches. RVM operates as a learning mechanism within the Bayesian framework. In this classifier, the core functions were created without excessive parameterization [29]. As a result of RVM's ability to reduce the computational complexity and overcome its shortcoming in keeping the kernel function in line with the Mercer condition, incorporating it into a hybrid prediction system can achieve higher accuracy and stability [13]. A key advantage of RVM is its ability to produce decision functions that are sparser than SVM while maintaining classification accuracy. In this way, computational complexity can be significantly reduced, making the decision function more suitable for real-time decision-making applications [19]. RVM is regarded as particularly well-suited for modeling specific tasks like the SVM approach; RVM initially transforms training data samples from the original input space into a higher-dimensional space known as the feature space. Consequently, a hyperplane can be established within this feature space to discriminate the data [25]. Following is a description of the RVM model's principle.

Suppose a data set $\{x_{n,}t_n\}_{n=1}^{N}$, where $x_n \in R^n, t_n \in R$. The following equation shows the relationship between x_n and t_n:

$$t_n = y(x_n;\omega) + \xi_n = \sum_{i=1}^{N} \square\, \omega_i \varphi_i(x) + \omega_0 + \xi \qquad (9)$$

where $\omega = (\omega_0, \omega_1, \ldots\ldots \omega_{n)}$ is the weight vector, ξ_n is the independent additive term subject to $\xi_n \sim N(0, \sigma^2)$, $\varphi_i(x) = K(x, x_i)$ is the nonlinear essential function and $K(x, x_i)$ is the kernel function. Hence, $p(x) = N(y(x_n), \sigma^2)$ signifies the normal distribution of t_n with mean $y(x_n)$and variance, σ^2. Subsequently, t_n is assumed to be independent of each other, in which the likelihood of the complete data set can be written as follows:

$$p(\omega, \sigma^2) = (2\pi\sigma^2)^{-N/2} exp\, \{-\frac{1}{2\sigma^2} \left\|t - \phi\omega\right\|^2 \qquad (10)$$

where $t = (t_1, t_2, \ldots, t_N)^T$and $\phi = [\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_N)]^T$ are the $N \times (N-1)$ kernel function matrix in which

$$\varphi(x_N) = [1, K(x_n, x_1), K(x_n, x_2), \ldots, K(x_n, x_N)]^T$$

By utilizing training data, we can acquire the distribution density of target values. $y\, *$. For the new input vector $x\, *$, we get:

$$p(y) = \int_{\square}^{\square} \square\, p(y^* | \omega, \alpha, \sigma^2) p(\omega, \alpha, \sigma^2 | y) d\omega d\alpha d\omega\sigma^2 \quad (11)$$

Based on the equation above, $y^*$is a normal distribution. The expected values $\underline{y^*}$ and variance, σ^2 are:

$$\underline{y^*} = \mu^T \phi(x^*)$$

$$\sigma_*^2 = \sigma_{MP}^2 + \phi^T(x^*) E \phi(x^*) \qquad (12)$$

From equation above, a further algorithm to determine the type of kernel in ϕ is presented as follows:

$$Hyperbolic\ tangent = tanh(ax^T y + c) \qquad (13)$$

$$Polynomial = \underline{(ax^T . y + c)^d} \qquad (14)$$

$$RBF = (exp\ exp\ \left(-\sigma\left\|x - y\right\|^2\right) \qquad (15)$$

$$Laplace = exp\ exp\ \left(-\frac{\|x-y\|}{\sigma}\right) \qquad (16)$$

$$ANOVA = \underline{\sum_{k=1}^{n} \square\, exp\, (-\sigma(x^k - y^k)^2)^d\quad} \qquad (17)$$

$$Bessel = (\frac{J_{v+1}(\sigma\|x-y\|)}{\|x-y\|^{\cancel{-n(v+1)}}} \qquad (18)$$

To implement the RVM model, the following steps must be followed:

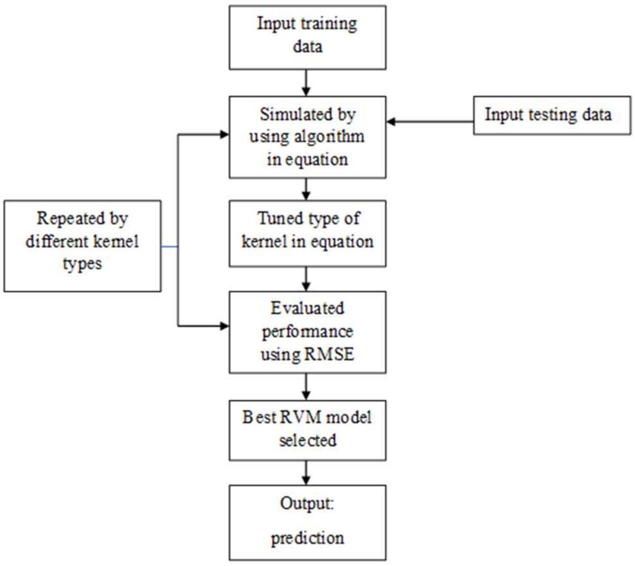
Input training data
Simulated by using algorithm in equation
Input testing data
Repeated by different kernel types
Tuned type of kernel in equation
Evaluated performance using RMSE
Best RVM model selected
Output: prediction

Fig 3 A flowchart of Relevance Vector Machine

B. Model Performance Assessment

It is essential to exhibit the statistical basis for the model's performance by comparing rainfall prediction values computed using different soft computing methods after the model was developed [3]. Models such as SVR, ANN, and RVM have been assessed based on Root Mean Square Error (RMSE) and Nash-Sutcliffe Efficiency (NSE). A comparison was also conducted between the three imputation methods, mean substitution, random forest, and K-nearest neighbor, according to two measures of performance: root mean square error and Nash-Sutcliffe Efficiency (NSE). The Nash–Sutcliffe efficiency index (E_f) is a widely utilized and potentially dependable statistic for evaluating the goodness of fit of hydrologic models. Nevertheless, no method has been devised for estimating the statistical significance of sample values [10]. The mathematical expression of NSE is represented below.

$$NSE = 1 - \frac{\sum_{i=1}^{n} \square (x_i - y_i)^2}{\sum_{i=1}^{n} \square (x_i - \underline{x_i})^2} \qquad (19)$$

In this equation, x_i is the observed value of rainfall, and y_i is the predicted value of rainfall. In this case, x_i is the average observation value (mean), and n is the number of observation sets. As described in the text, on average, the coefficient of efficiency is −∞ ≤ NSE ≤ 1, with NSE = 1 representing the perfect fit model and NSE≤ 0 representing the mean of the observed values being more predictive than the evaluated model.

The RMSE calculation was carried out through a series of three simple steps. As a result, a 'total square error' was obtained as the sum of each squared error, which means that each error has a more significant influence on the total in proportion to what it squares than its magnitude. Therefore, the influence of larger errors on the total square error outweighs that of smaller errors. A declining number of larger individual errors will lead to an increase in total square error. The mean-square error (MSE) was then calculated as the sum of the square errors divided by the number *n*. Finally, we take the square root of the MSE to find the RMSE [27]. RMSE can be expressed mathematically as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - y_t)^2} \quad (20)$$

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

III. Results and Discussion

A. Missing Values

The lowest RMSE values signify the best model performance, indicating minimal deviation between the estimated and observed values. In contrast, NSE values falling within the 0 to 1 range denote models exhibiting high predictive accuracy [23]. Therefore, the method that produced the lowest RMSE and highest NSE values was chosen as the most suitable approach for filling out the predicted data set with missing data.

TABLE I
RMSE values of predictand variables for each imputation method

Imputation Method	RMSE
EM algorithm	2.2252
Mean Imputation	0.7518
Nearest Neighbor	1.6858
MCMC	1.6307
NIPALS	1.3341

According to the findings, mean imputation is the most effective method for imputation of predictand for Selangor states. Meanwhile, the EM algorithm is regarded as the least fitting imputation method for this study because its results revealed the greatest RMSE.

B. Principal Component Analysis (PCA)

During the reduction of high-dimensional data, PCA was utilized to identify the principal components (PCs) necessary to reduce the number of predictors without losing significant information. Presented in Table II below is a summary of the analysis, which includes eigenvalues, variance, and cumulative percentage variance:

TABLE II
Result for principal component analysis

Dimension	Eigenvalue	Percentage variation (%)	Cumulative Percentage (%)
Component 1	1.595	26.575	26.575
Component 2	1.077	17.592	44.527
Component 3	1.088	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	100.00

According to Table 2, six components were identified from the eigenvalues and total variation. PCA typically yielded a number of components equivalent to the selected variables. The variation percentage and eigenvalues were arranged in decreasing sequence. With reference to the Kaiser criterion [30], eigenvalues exceeding 1.00 will be selected to interpret the component. Consequently, the results reveal that Components 1, 2, and 3 have eigenvalues exceeding 1.00, namely, 1.595,1.077 and 1.088, respectively. Components 4 and 5 are also accepted since their eigenvalues are nearest to 1.00, which are 0.973 and 0.886, respectively.

TABLE III
Results for correlation of PC loading

Dimension	F1	F2	F3	F4	F5
Precipitations	-0.284	0.647	-0.008	0.235	0.666
Max temperature	0.840	0.192	-0.137	0.087	0.070
Min temperature	0.628	0.508	-0.062	-0.430	-0.161
Wind	0.267	0.368	0.576	-0.450	0.509
Relative humidity	0.182	0.216	0.767	0.496	-0.288
Solar	0.557	-0.426	-0.255	0.526	0.266

The correlation of PC loading of each variable of predictor and factor is shown in Table 3. The selected eigenvalues determined the number of factors generated [23]. Here, a selection of predictors was based on the highest or strongest PC loading for every predictor variable, according to [5]. According to this argument, the higher the PC loading of a variable, the more significant its contribution to the factors accounting for a particular PC [23]. Maximum temperature demonstrates a significant positive loading of 0.840 on Factor 1 (0.840). For the second factor, precipitations show the highest number (0.647), but loadings don't appear significant. Next, Factor 3 shows a strong positive loading on Relative Humidity (0.767). Then, Factor 4 and Factor 5 do not have any strong loading values. Such factors are expected to account for as much variance as possible since they were extracted sequentially [23]. Thus, the new dataset was structured in a matrix format. 132594 rows represent the number of days by station, while the extracted factors are represented by five columns designated as Factor 1, Factor 2, Factor 2, Factor 3, Factor 4, and Factor 5.

C. Relevance Vector Machine

Determining the kernel's superior performance entailed analyzing the RMSE and NSE values throughout the calibration and validation periods, as indicated in Table 3. RMSE offers a strong measure of the model's predictive accuracy and is considered the primary criterion for evaluating fit when the model's primary objective is prediction.

TABLE IV
Performance of RVM varying kernel function

Type of Kernel	Calibration			Validation	
	Number of Relevance Vector	RMSE	NSE	RMSE	NSE
RBF	50	14.09	-0.20	13.73	-0.40
Polynomial	4	13.91	-0.11	12.90	-0.24
Laplace	302	18.32	-0.93	15.67	-0.83
Hyperbolic tangent	102	14.49	-0.21	13.84	-0.42
Bessel	14	14.09	-0.14	13.14	-0.28
ANOVA	13	14.09	-0.16	12.94	-0.24

The polynomial kernel was chosen for the calibration and validation period as it generated the lowest RMSE value while increasing NSE. During calibration, the Polynomial kernel produced the fewest relevant values.

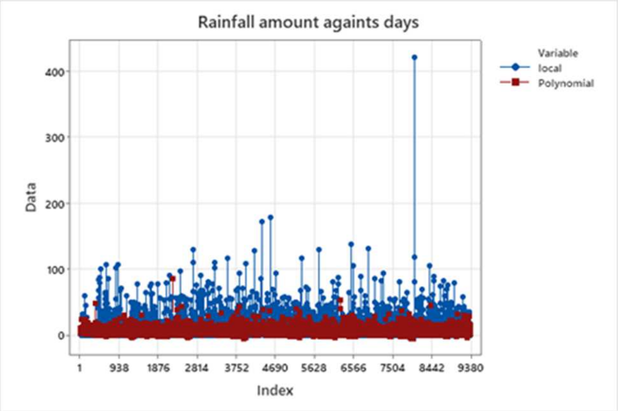

Fig 4 Performance of RVM in predicting daily rainfall amount in calibration period

Figure 4 shows the trend of the RVM model's prediction value, which clearly shows that it could follow the trend of the predictand because it can predict extreme values as predictand extreme values.

Through downscaling, the issues of projection rainfall in Selangor states were discovered, prompting the application of correlated procedures to achieve the most accurate results in predicting rainfall data. Imputation methods are explored in this study, focusing on their efficiency in dealing with missing ground data. To accomplish this, we will identify the most effective method to deal with missing ground data in Selangor state. For Selangor state, Value 1 imputation proves to be the most appropriate method for handling missing atmospheric data. Therefore, Value 1 imputation can also analyze missing data issues for Peninsular Malaysia and other tropical countries with similar data sets. For an advanced analysis to be effective, preprocessing steps must be well managed.

After Principal Component Analysis (PCA) was used to reduce the extensive dimensional data and choose the predictor variables, it is a constant challenge for hydrologists to deal with high-dimensional time-series data that requires highly complex computations. Machine learning models can have difficulties recognizing patterns due to the high dimensionality of data. By employing the PCA method, the dimensionality of the data was reduced through the identification of uncorrelated features (PCs). We extracted five PCs based on the number of variables as predictors. We cut the information when the eigenvalue is more significant than one, and the cumulative percentage is greater than 70%. Variables' contributions to that chosen PC are reflected in rotation discovery as a loading of factors. According to the results, the high loading indicates that reasonable predictor variables were selected, and five variables were selected for further analysis. PCA can assist statistical downscaling approaches in selecting predictor variables while reducing the dimensionality of time-series data during preprocessing.

Finally, the RVM-based Polynomial kernel was discovered to be the most appropriate model for analyzing rainfall prediction in Selangor—lastly, the performance of RVM by using indicator statistical measures such as RMSE and NSE. The RVM model is potentially beneficial for hydrologists or climatologists, particularly for downscaling studies, in determining rainfall projections for long-trend analysis. This study has proven that it manages to cater to all the issues regarding missing data records, high dimensional data, and the statistical downscaling approach processes.

IV. Conclusion

Economic growth and population growth will result in increased man-made greenhouse gas emissions, which will impact the hydrological cycle globally and regionally. Flooding and tsunamis are likely to occur because of the influence of climate change on hydrological cycles. Malaysia is experiencing the most devastating effects of climate change as a result of flooding. The disruption caused by floods to people's daily lives is significant, often lasting for weeks. The situation is likely to worsen in the coming years due to climate change. In December 2021, the latest flooding occurrence in Selangor caused the number of Malaysians to flee to flood evacuation centers to increase due to continuous heavy rainfall. Meteorologists should analyze rainfall projections to prevent more tragic floods in Selangor state.

Machine learning techniques have been proposed as a statistical downscaling approach for the projection of daily rainfall amounts in Selangor states. For missing data, mean imputation is the most effective method for imputation of predictand for Selangor states. In this study, we explained statistical downscaling based on the RVM model. In particular, the polynomial kernel for RVM exhibited significant enhancement in predicting daily rainfall amounts compared to other kernels, as validated by statistical measures. RVM has proven its ability to predict extreme values, which could be helpful to hydrologists or climatologists in analyzing environmental models and improving the assessment of climate change.

Acknowledgment

The support and funding for this research are courtesy of Universiti Pendidikan Sultan Idris and the Ministry of Higher Education Malaysia as part of the Post-Doctoral Sponsored Scholar.

References

[1] K. Abbass, M. Z. Qasim, H. Song, M. Murshed, H. Mahmood, and I. Younis, “A review of the global climate change impacts, adaptation, and sustainable mitigation measures,” Environmental Science and Pollution Research, vol. 29, no. 28, pp. 42539–42559, Apr. 2022, doi:10.1007/s11356-022-19718-6.

[2] Bernama (2021, Dec 19). Once in 100 years: One month average rainfall poured down in one day. Th star. https://www.thestar.com.my/news/nation/2021/12/19/floods-heavy-rain-lasting-over-24-hours-equals-to-average-monthly-rainfall-occurring-once-in-100-years-says-environs-ministry

[3] 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,” Stochastic Environmental Research and Risk Assessment, vol. 30, no. 6, pp. 1769–1784, Sep. 2015, doi: 10.1007/s00477-015-1153-y.

[4] Daniel, F. (2020). What is Machine Learning? Emerj The AI Research and Advisory Company. https://emerj.com/ai-glossary-terms/what-is-machine-learning/

[5] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[6] H. Kang, “The prevention and handling of the missing data,” Korean Journal of Anesthesiology, vol. 64, no. 5, p. 402, 2013, doi:10.4097/kjae.2013.64.5.402.

[7] H. Lee and K. Kang, “Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling,” Advances in Meteorology, vol. 2015, pp. 1–12, 2015, doi: 10.1155/2015/935868.

[8] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, “Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting,” Machine Learning with Applications, vol. 7, p. 100204, Mar. 2022, doi: 10.1016/j.mlwa.2021.100204.

[9] D. Maraun et al., “Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user,” Reviews of Geophysics, vol. 48, no. 3, Sep. 2010, doi:10.1029/2009rg000314.

[10] R. H. McCuen, Z. Knight, and A. G. Cutter, “Evaluation of the Nash–Sutcliffe Efficiency Index,” Journal of Hydrologic Engineering, vol. 11, no. 6, pp. 597–602, Nov. 2006, doi: 10.1061/(asce)1084-0699(2006)11:6(597).

[11] P. Mehta et al., “A high-bias, low-variance introduction to Machine Learning for physicists,” Physics Reports, vol. 810, pp. 1–124, May 2019, doi: 10.1016/j.physrep.2019.03.001.

[12] T. B. Pepinsky, “A Note on Listwise Deletion versus Multiple Imputation,” Political Analysis, vol. 26, no. 4, pp. 480–488, Aug. 2018, doi: 10.1017/pan.2018.18.

[13] W. Qiao, K. Huang, M. Azimi, and S. Han, “A Novel Hybrid Prediction Model for Hourly Gas Consumption in Supply Side Based on Improved Whale Optimization Algorithm and Relevance Vector Machine,” IEEE Access, vol. 7, pp. 88218–88230, 2019, doi:10.1109/access.2019.2918156.

[14] J. Quinonero-Candela and L. K. Hansen, “Time series prediction based on the Relevance Vector Machine with adaptive kernels,” IEEE International Conference on Acoustics Speech and Signal Processing, May 2002, doi: 10.1109/icassp.2002.5743959.

[15] G. T. Reddy et al., “Analysis of Dimensionality Reduction Techniques on Big Data,” IEEE Access, vol. 8, pp. 54776–54788, 2020, doi:10.1109/access.2020.2980942.

[16] Rogers I, Kirkham C. JikesNODE and PearColator: A Jikes RVM operating system and legacy code execution environment. In2nd ECOOP Workshop on Programm Languages and Operating Systems (ECOOP-PLOS’05) 2005 Jul 26.

[17] D. A. Sachindra, K. Ahmed, Md. M. Rashid, S. Shahid, and B. J. C. Perera, “Statistical downscaling of precipitation using machine learning techniques,” Atmospheric Research, vol. 212, pp. 240–258, Nov. 2018, doi: 10.1016/j.atmosres.2018.05.022.

[18] S. M. Shaharudin, “Imputation methods for addressing missing data of monthly rainfall in Yogyakarta, Indonesia,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1.4, pp. 646–651, Sep. 2020, doi: 10.30534/ijatcse/2020/9091.42020.

[19] B. Ribeiro and C. Silva, “RVM Ensemble for Text Classification,” International Journal of Computational Intelligence Research, vol. 3, no. 1, 2007, doi: 10.5019/j.ijcir.2007.81.

[20] L. Song, W. Duan, Y. Li, and J. Mao, “A timescale decomposed threshold regression downscaling approach to forecasting South China early summer rainfall,” Advances in Atmospheric Sciences, vol. 33, no. 9, pp. 1071–1084, Jul. 2016, doi: 10.1007/s00376-016-5251-7.

[21] M. R. Stavseth, T. Clausen, and J. Røislien, “How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data,” SAGE Open Medicine, vol. 7, p. 205031211882291, Jan. 2019, doi:10.1177/2050312118822912.

[22] Suhaimi, N., Ghazali, N. A., Nasir, M. Y., Mokhtar, M. I. Z., & Ramli, N. A. (2017). Markov Chain Monte Carlo method for handling missing data in air quality datasets. Malaysian Journal of Analytical Science, vol. 21, no. 3, Jun. 2017, doi: 10.17576/mjas-2017-2103-05.

[23] Sulaiman, Nurul Ainina Filza. Statistical Downscaling of Projecting Rainfall Amount Based On SVC-RVM Model. Tanjong Malim, 2022.

[24] N. ur Rehman and H. Aftab, “Multivariate Variational Mode Decomposition,” IEEE Transactions on Signal Processing, vol. 67, no. 23, pp. 6039–6052, Dec. 2019, doi: 10.1109/tsp.2019.2951223.

[25] D. Tien Bui et al., “A Novel Integrated Approach of Relevance Vector Machine Optimized by Imperialist Competitive Algorithm for Spatial Modeling of Shallow Landslides,” Remote Sensing, vol. 10, no. 10, p. 1538, Sep. 2018, doi: 10.3390/rs10101538.

[26] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine", J. Mach. Learn. Res., vol. 1, pp. 211-244, Jun. 2001.

[27] Willmott, Cort J., and Kenji Matsuura. “Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance.” *Climate Research* 30, no. 1 (2005): 79–82. http://www.jstor.org/stable/24869236.

[28] S.-F. Wu, C.-Y. Chang, and S.-J. Lee, “Time Series Forecasting with Missing Values,” Proceedings of the 1st International Conference on Industrial Networks and Intelligent Systems, 2015, doi:10.4108/icst.iniscom.2015.258269.

[29] J. Xingmeng, W. Li, P. Liwu, G. Mingtao, and H. Daidi, “Rolling Bearing Fault Diagnosis Based on ELCD Permutation Entropy and RVM,” Journal of Engineering, vol. 2016, pp. 1–7, 2016, doi:10.1155/2016/1308108.

[30] Y. Zhang, B. Zhou, X. Cai, W. Guo, X. Ding, and X. Yuan, “Missing value imputation in multivariate time series with end-to-end generative adversarial networks,” Information Sciences, vol. 551, pp. 67–82, Apr. 2021, doi: 10.1016/j.ins.2020.11.035.

[31] H. Zhao, J. Zheng, J. Xu, and W. Deng, “Fault Diagnosis Method Based on Principal Component Analysis and Broad Learning System,” IEEE Access, vol. 7, pp. 99263–99272, 2019, doi:10.1109/access.2019.2929094.

[32] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, “Missing Value Estimation for Mixed-Attribute Data Sets,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 1, pp. 110–121, Jan. 2011, doi: 10.1109/tkde.2010.99.