International Journal of Engineering & Technology, 7 (4.30) (2018) 573-579

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Effect of Dimensionality Reductions Technique in Modelling and Forecasting River Flow

Shuhaida Ismail[1]*, Ani Shabri[2], Aida Mustapha[3], Siraj Mohammed Pandhiani[4]

[1]Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia,84000 Muar, Johor, Malaysia.
[2]Faculty of Science, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
[3] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia.
[4]Jubail University College, Saudi Arabia
**Corresponding author E-mail: shuhaida@uthm.edu.my*

Abstract

The ability of obtain accurate information on future river flow is a fundamental key for water resources planning, and management. Traditionally, single models have been introduced to predict the future value of river flow. This paper investigates the ability of Principal Component Analysis as dimensionality reduction technique and combined with single Support Vector Machine and Least Square Support Vector Machine, referred to as PCA-SVM and PCA-LSSVM. This study also presents comparison between the proposed models with single models of SVM and LSSVM. These models are ranked based on four statistical measures namely Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Correlation Coefficient (r), and Correlation of Efficiency (CE). The results shows that PCA combined with LSSVM has better performance compared to other models. The best ranked models are then measured using Mean of Forecasting Error (MFE) to determine its forecast rate. PCA-LSSVM proven to be better model as it also indicates a small percentage of under-predicted values compared to the observed river flow values of 0.89% for Tualang river while over-predicted by 2. 08% for Bernam river. The study concludes by recommending the PCA as dimension reduction approach combined with LSSVM for river flow forecasting due to better prediction results and stability than those achieved from single models

Keywords: *Dimensionality Reduction; Forecasting, River Flow; Least Square Support Vector Machine; Principal Component Analysis,*

1. Introduction

The prediction of river flow values is undeniably crucial especially for a sound planning and smooth operation of water resource management system. It is important as the information obtained from the prediction process will be beneficial in water management by optimizing the management of water resources and help in prevent natural disaster such as flood control [21].

There are variety of statistical modelling approaches that have been developed for a reliable prediction of river flow such as the physically based distribution model and empirical models. Physically based distribution model known as knowledge-driven modelling requires various information pertaining the catchment area whilst empirical models, known as data-driven modelling requires only set of historical data and mathematically identifies the connection between the inputs and output data. In river flow forecasting, data-driven model, which uses only historical river flow data has becomes increasingly popular as it offers fast computational time, requires minimum information without losing its accuracies [19, 20, 38, 48].

In the past, several models such as Regression analysis and Box-Jenkins have been used to forecast the future river flow values. However, with the development of advanced Artificial Intelligent (AI) methods which based on the machine learning technique, various other forecasting models have emerge every day. The statistical learning framework proposed by Vapnik [42, 46] has led to the introduction of the kernel-based Support Vector Machine (SVM). SVM has been introduced as the new statistical learning technique due to its strong theoretical statistical framework and has gained popularity due to various promising features such as better empirical performance, robustness, resistant to over-fitting problem and etc. With that, SVM has been successfully applied to solve various problems such as data mining, classification, regression, bioinformatics, feature recognition, and time series forecasting [18, 24, 29, 30, 39, 44, 51, 53].

Previous researchers suggested that SVM could be applied to hydrological time series with a nonlinear nature to achieve a better prediction performance than linear modelling approaches. Therefore, SVM received wide attention in modelling and forecasting hydrological processes such as rainfall-runoff forecasting [7, 8], flood-stage forecasting [51], stream flow forecasting [9, 24, 28], sediment yield [29, 30] and reservoir inflow forecasting [23, 24, 25].

In 2005, Suykens and Vandewalle introduces a least square version of the original Support Vector Machine known as LSSVM. LSSVM was modified from the existed SVM and encompasses similar properties as SVM. LSSVM has been successfully applied to diverse fields such as pattern recognition and regression problems [9, 12,17]. In water resource management, LSSVM method started to receive wide attention. Several studies have been carried out using LSSVM in the modelling of environmental and ecological systems, such as water quality prediction [52], lake water pollution, hydrological time series, and river flow forecasting [4, 38, 40, 54], and rainfall-runoff [33].

Hydrological time series may be contaminated by various noises, hence data pre-processing are crucial to eliminate the noises and smooth out the data before further action. There are plenty data pre-processing technique. One of the commonly used technique is dimensionality reduction and Principal Component Analysis (PCA) is a well-known method for that purpose. PCA was first introduced by Karl Pearson in 1901 and was further developed by Hotelling in 1930 [14]. The central idea of PCA is to find the right projection of the data and this can be achieved by projecting the large interrelated variables into a smaller set of derived variables, while retaining as much as possible of the variation present in the data set. Over the years, PCA has been successfully employed in various areas such as in pattern recognitions [45], time series forecasting [35], process monitoring [2], and etc.

Improving the forecasting accuracy is fundamental but it is a difficult task faced by decision-makers in many areas. Various studies showed that the prediction's accuracy can be improved by using hybrid or combination models and it has become a common practice to improve the forecasting accuracy [53]. Several studies have proven that hybrid models can be an effective way to improve the predictions achieved by any models that are used separately [31]. In recent years, more hybrid forecasting models have successfully solved many prediction problems.

The aim of this research is to study the effect of forecasting performance using dimensionality reduction technique and combined it with forecasting model of SVM and LSSVM. The performance of developed PCA-SVM and PCA-LSSVM are then evaluated against single models and it is shown that the proposed model yielded more accurate results with less error.

2. Methodology

2.1. Principal Component Analysis (PCA) Model

PCA is a well-known technique for noise or dimensionality reduction and feature extraction without jeopardizing the accuracy of the model. With that capability, PCA has become to most used technique in data analysis. In PCA, the dataset are first transform a correlated variables into smaller numbers on uncorrelated variables called principal components (PC). The obtained eigenvalues are then sorted in descending order of the corresponding eigenvectors, which later makes the number of PC in the datasets can be reduced. The first few PC contained the highest variance of the dataset and reduces from step to step [16].As the first PC contain the highest variance, therefore it is the most informative PCs among others.

Suppose that n original observation are taken from $X_{1,}X_2,..., X_k$ with the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0.$ therefore, the first principal component of Y_1 is a linear combination for X's original variables is defined as:

$$Y_1 = \sum_{i=1}^{k} a_{1i} X_i \qquad \ldots(1)$$
$$= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1i}X_i$$

where $a_{11}, a_{12}, \ldots, a_{1k}$ is the first weights vector that can maximize the variance of Y_1 subject to the given constraint:

$$\sum_{i=1}^{n} a_{1i}^2 = 1 \qquad \ldots.(2)$$

The second principal component Y_2 which has the second largest variance is defined as:

$$Y_2 = \sum_{i=1}^{k} a'_{2i} X_i$$
$$= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2i}X_i \qquad (3)$$

where $a_{21}, a_{22}, \ldots, a_{2i}$ is the second weight vector that the variance of Y_2 is maximizes, subjected to the given constraint:

$$\sum_{i=1}^{n} a_{2i}^2 = 1 \qquad (4)$$

Following PCA concept, the second principal component which is Y_2 is linearly independent with the first principal component, Y_1. Thus, the independent condition is specified by the constraint that is:

$$\sum_{i=1}^{2} a_{1i} a_{2i} = 0 \qquad (5)$$

It is similar for the third principal component and so on.

2.2. Least Square Support Vector Machine

As oppose to SVM, LSSVM adopt least square linear systems as its loss function and equality constraints compared to convex quadratic programming and inequality constraints in SVM. With this feature, computational time for LSSM are lesser than SVM and able to converge the problem quickly without losing its accuracy.

LSSVM has been used to estimate the nonlinear $y(x)$ of the form:

$$y(x) = w^T \varphi(x) + b \qquad (6)$$

When LSSVM is used for function estimation, the optimization problem is formulated by minimizing the regular function [43] as:

$$\min R(w,e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} e_i^2 \qquad (7)$$

subject to

$$y(x) = w^T \phi(x_i) + b + e_i, \qquad i = 1,2,..., n, \qquad (8)$$

To solve this optimization problem, Lagrange function is constructed as:

$$L(w,b,e,\alpha) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i \{w^T \phi(x_i) + b + e_i - y_i\} \qquad (9)$$

where α_i is Lagrange multipliers. The solution of (9) can be obtained by partially differentiating with respect to w, b, e_i and α_i accordingly:

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i \phi(x_i),$$
$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} \alpha_i = 0,$$
$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i,$$
$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^T \phi(x_i) + b + e_i - y_i = 0$$

After elimination of e_i and w as the solution is given by the following set of linear equations:

$$\begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \phi(x_i)^T\phi(x_j)+\gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{10}$$

where $y=[y_1,\ldots,y_n]$, and $\mathbf{1}=[1;\ldots;1]$, $\alpha=[\alpha_1,\ldots,\alpha_n]$. This finally leads to the following LSSVM model for function estimation:

$$y(x)=\sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i,\mathbf{x})+b \tag{11}$$

where α_i and b are the solution to the linear system. In LSSVM, Radial Basis Function (RBF) is popular kernel function as has superior efficiency and has better performance compare to other kernels. The RBF is defined as $K(x_i,x_j)=\exp\left(-\left\|x_i-x_j\right\|^2/2\sigma^2\right)$.

2.3. Proposed PCA-SVM and PCA-LSSVM Models

In this section, the proposed PCA-SVM and PCA-LSSVM are explained. The data will first has to undergo data pre-processing which is dimensionality reduction technique using PCA. At this stage, numbers of principal components are obtained from the original input variables. Next, the newly data set are constructed based on the selected PCs. Once finish with the construction, the new data are then inputted into SVM and LSSVM for forecasting purposes.

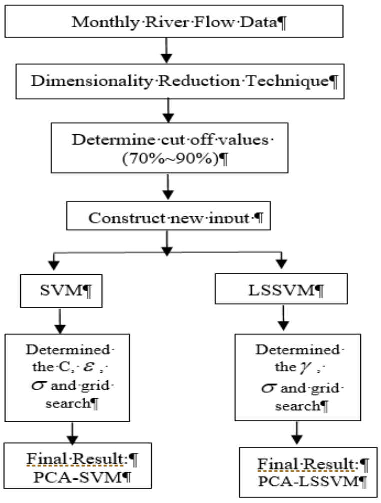
Monthly·River·Flow·Data¶
Dimensionality·Reduction·Technique¶
Determine·cut·off·values·(70%~90%)¶
Construct·new·input·¶
SVM¶
LSSVM¶
Determined·the·C,·ε,·σ and·grid·search¶
Determined·the·γ,·σ and·grid·search¶
Final·Result:¶ PCA-SVM¶
Final·Result:¶ PCA-LSSVM¶

Fig 1: The Proposed PCA-SVM and PCA-LSSVM

2.4 Performance Criteria

In this study, the forecasted values are evaluated using several widely used criteria in evaluating the time series forecasting results. The performance criteria used in this study are Mean Absolute Error (MAE) , Root Mean Square Error (RMSE), Correlation Coefficient (*r*) and Nash–Sutcliffe Coefficient Efficiency (CE).

In order to perfectly evalute the performance of the model, one of the required criteria is a measure of absolute error such as MAE. Meanwhile RMSE is a good evaluator as its highly sensitive to even small errors. While r measured on the relationship between the predicted flows, $\hat{y}_t$ againts the observed flows, y_t. CE on the other hand, is commonly used by the hydrologist to assess the hydrological power prediction model [33]. The criterion of judge is that, the model with the smallest MAE, RMSE and the highest r and CE values will be selected as the best model. The performance evaluation are defined as follows:

$$\text{MAE}=\frac{1}{n}\sum_{t=1}^{n}\left|y_t-\hat{y}_t\right| \tag{12}$$

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2} \tag{13}$$

$$r=\frac{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\bar{y}\right)\left(\hat{y}_t-\bar{\hat{y}}\right)}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\bar{y}\right)^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\bar{\hat{y}}\right)^2}} \tag{14}$$

$$\text{CE}=1\text{-}\frac{\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2}{\sum_{t=1}^{n}\left(y_t-\bar{y}\right)^2} \tag{15}$$

where y_t and $\hat{y}_t$ are the actual and the forecasted values at the time *t*. Meanwhile, $\bar{y}$ and $\bar{\hat{y}}$ denotes for mean of actual and forecasted values, respectively.

3. Experiments and Datasets

This section explained in detail about datasets used in the study as well as the parameters setting for each of the experiments.

3.1. Datasets

Figure 2 shows time series plotting for Bernam and Tualang Rivers in Selangor and Perak respectively. The 1st case study was collected from Bernam River from January 1966 to December 2008, consists of 516 monthly data. The training set are from January 1966 until December 2003 while the test set consists data from January 2004 until December 2008.

Meanwhile, the 2nd case study was gathered from monthly river flow from Tualang River station located in Kinta, Perak. The data was collected from January 1976 to March 2007. The first 315 monthly data were used for training and another 60 monthly data recorded from April 2002 to March 2007 were used for testing.

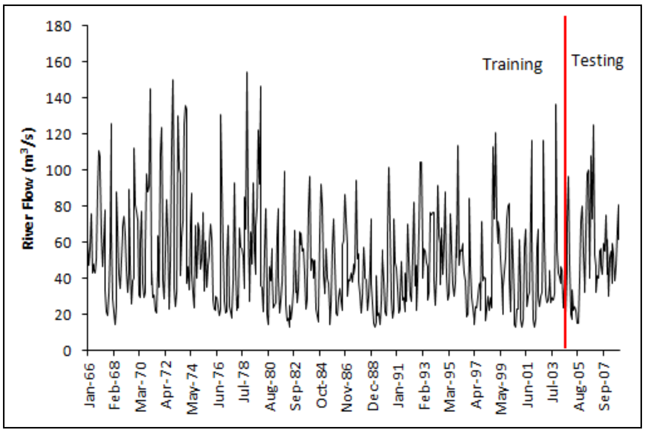

Fig 2: Monthly River Flow for Bernam Rivers

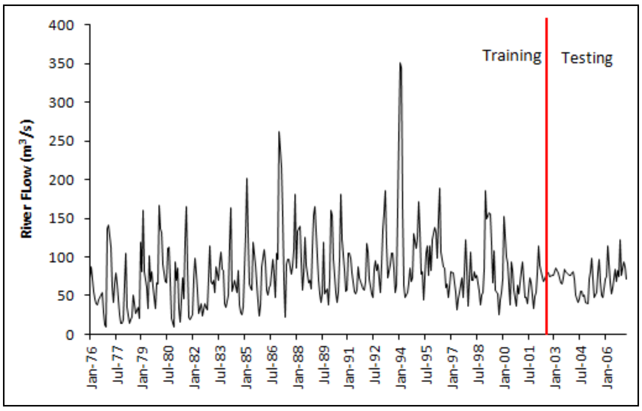

Fig 3: Monthly Flow of Tualang River

Based on visual inspection, the graph in Fig 2 and 3 shows a hydrological time series has an up and down pattern, and seasonal variation with the highest monthly flow usually occurs between April to May and October to January each year.

3.2. Experiment using Single SVM

In this study, the parameters of C and ε for SVM was set in the range [1, 10] with an increment of 1.0, and [0.1, 0.5] at increment of 0.1, respectively. Meanwhile, RBF and σ^2 used as kernel function and the value of RBF kernel bandwidth are fixed at 0.5. Furthermore, 10-fold cross validation for hyper-parameter pair of C and ε are applied. The purpose of cross validation on the training set is increase the reliability of the results as the iteration of the cycle will be repeated ten times and the optimum prediction error are obtained.

Table 1: Single SVM Result for Bernam and Tualang Rivers

Rivers	Input	MAE	RMSE	r	CE
Bernam River	Input 2	17.202	21.385	0.542	0.176
	Input 4	16.266	20.380	0.551	0.252
	Input 6	15.413	20.276	0.516	0.260
	Input 8	15.148	20.026	0.533	0.278
	Input 10	**15.076**	19.695	0.551	0.301
	Input 12	15.190	**18.796**	**0.604**	**0.364**
Tualang River	Input 2	12.229	15.427	**0.676**	0.281
	Input 4	10.420	14.252	0.516	0.387
	Input 6	10.172	13.833	0.559	0.422
	Input 8	10.440	14.428	0.576	0.371
	Input 10	9.728	13.769	0.589	0.428
	Input 12	**9.637**	**13.617**	0.585	**0.440**

Table 1 shows the performance results for Single SVM approach for both catchment areas. Based on the obtained results, it can be seen that the accuracies of SVM model for both catchment have increased gradually. Despite the fact that input 10 for Bernam River has the lowest MAE, while input 2 for Tualang River has the highest r, input 12 was selected as the best input variables for Bernam and Tualang River as the variances among other input variables are small and relatively insignificant.

3.3. Experiments using Single LSSVM

As for LSSVM, the parameter of γ are set the range of [10, 1000] and σ^2 in the range of [0.01, 10]. Same as SVM, 10-fold cross validation were also applied here.

Table 2: Single LSSVM Result for Bernam and Tualang Rivers

Rivers	Input	MAE	RMSE	r	CE
Bernam River	Input 2	15.707	20.075	0.525	0.274
	Input 4	15.948	20.116	0.521	0.271
	Input 6	15.192	19.634	0.555	0.306
	Input 8	15.501	**19.537**	**0.572**	**0.313**
	Input 10	15.253	19.748	0.548	0.298
	Input 12	**14.766**	19.776	0.547	0.296
Tualang River	Input 2	10.672	14.603	0.510	0.356
	Input 4	10.072	13.844	0.581	0.421
	Input 6	10.094	13.963	0.580	0.411
	Input 8	10.540	14.400	**0.595**	0.374
	Input 10	10.063	14.056	0.581	0.403
	Input 12	**9.992**	**13.765**	0.574	**0.428**

Table 2 shows the performance results obtained in the training and testing period of the Single LSSVM approach for Bernam and Tualang Rivers. The results showed that input eight has the best performance for Bernam River, while input 12 is for Tualang River. By considering these results, input 12 was selected as the best input variables for both catchment areas as the variances among other input variables are relatively insignificant. As forecasting model with 12 input variables has shown the superior results for both of Single SVM and LSSVM, therefore, input 12 will be employed for PCA algorithm in PCA-SVM and PCA-LSSVM models.

3.4. Experiments using PCA-SVM and PCA-LSSVM

This section will explain in details about the experimentation using dimensionality reduction technique combined with SVM and LSSVM. To extract PCs and reduce the dimension of the input variables, the monthly river flow for Bernam dan Tualang Rivers using twelve input variables are use respectively. Input 12 are choose to proceed with data pre-processing technique as it has the best forecasting results of MAE, RMSE, MAPE, r and CE for both single SVM and LSSVM models. Furthermore, input 12 correctly representing the whole dataset as it's denotes for twelve months in a year.

3.4.1. Data Pre-Processing Technique using PCA

By using PCA as a data pre-processing technique, the input variables are first, changed into numbers of principal components (PCs). After selecting the appropriate number of PCs, it will be changed again into input variables and known as new dataset and used it to forecast using SVM and LSSVM model. By using this method, the information of input variables will present with a minimum loss and adequate summary of the data. In this study, there are two values for total variation were selected in the range of 70% to 90%. Table 3 summarizes the descriptive statistics of PCA technique for Bernam dan Tualang Rivers respectively. Based from Table 3, the first cut-off values for Bernam river was observed from the fifth PCs with total variance of 79.72% and the second cut off value was selected from the eighth PCs with a total variance of 89.27%. In total, the new dataset are constructed for Bernam River ranging from five to eight input variables.

Table 3: Descriptive Statistics of PCs for Bernam and Tualang Rivers

Rivers	PCs	Eigen Value	% of Variance	Cumulative of %	Cut-off
Bernam River	1	1730.24	19.13	19.13	
	2	1674.33	18.51	37.64	
	3	1563.21	17.28	54.92	
	4	1209.83	13.38	68.30	
	5	**1032.99**	**11.42**	**79.72**	**1st**
	6	332.27	3.67	83.39	
	7	269.20	2.98	86.37	
	8	**262.70**	**2.90**	**89.27**	**2nd**
	9	250.73	2.77	92.04	
	10	250.07	2.76	94.81	
	11	236.94	2.62	97.43	
	12	232.62	2.57	100.00	
	1	5118.80	21.99	21.99	
	2	3954.71	16.99	38.98	

Tualang River	3	3785.18	16.26	55.24	
	4	**3701.94**	**15.90**	**71.15**	1st
	5	3285.96	14.12	85.27	
	6	900.93	3.87	89.14	
	7	**540.17**	**2.32**	**91.46**	2nd
	8	431.67	1.85	93.31	
	9	425.47	1.83	95.14	
	10	412.38	1.77	96.91	
	11	390.35	1.68	98.59	
	12	328.58	1.41	100.00	

Meanwhile, the first cut-off value for Tualang river was observed from the fourth PCs with a total variance of 71.15% and the second cut off value was selected from the seventh PCs with a total variance of 91.46%. In total, the new input variables for Tualang river were ranging from four to seven input variables.

3.4.2. Forecasting using SVM Model

Table 4 shows the performance results for SVM with PCA approach for Bernam and Tualang Rivers. It is evident that the accuracies for each PCs have increased gradually for Bernam River. However, the accuracy for Tualang River has increased exponential. By considering these results, the lowest MAE for Bernam river was collected from SVM with 5PCs. On the contrary, the lowest RMSE and the highest values of r and CE were recorded from SVM with 6PCs. For comparisons purpose SVM with 6PCs was selected to represent PCA-SVM model for Bernam river as SVM with 6Pcs giving the best results for most of the statistical performances.

Table 4: PCA-SVM Results for Bernam and Tualang Rivers

Rivers	PCs	MAE	RMSE	r	CE
Bernam River	5PCs	**14.095**	19.237	0.581	0.862
	6PCs	14.430	**19.202**	**0.583**	**0.863**
	7PCs	14.797	20.075	0.545	0.850
	8PCs	14.957	20.091	0.548	0.850
Tualang River	4PCs	24.014	27.030	0.354	0.841
	5PCs	10.798	14.752	0.503	0.953
	6PCs	10.477	14.201	0.567	0.956
	7PCs	**10.215**	**13.929**	**0.592**	**0.958**

Whereas for the second case study, the best results for SVM with PCA for Tualang River were obtained from 7PCs with the lowest MAE, RMSE, with the highest r and CE statistics. Therefore, SVM with 7PCs were selected as the best result representing Tualang River.

3.4.3. Forecasting using LSSVM Model

The PCA-LSSVM model uses the same inputs structures of the dataset as PCA-SVM which are PCA5 to PCA8 for Bernam River, PCA4 to PCA7 for Tualang River. For PCA-LSSVM model, the same process was employed as PCA-SVM model.

Table 5: PCA-LSSVM Results for Bernam and Tualang Rivers

Rivers	PCs	MAE	RMSE	r	CE
Bernam River	5PCs	14.956	19.752	0.546	0.855
	6PCs	14.862	19.702	0.550	0.856
	7PCs	14.815	19.739	0.546	0.855
	8PCs	**13.214**	**17.895**	**0.654**	**0.881**
Tualang River	4PCs	12.871	17.118	0.331	0.936
	5PCs	10.822	14.985	0.473	0.951
	6PCs	9.922	13.748	0.575	0.959
	7PCs	**9.899**	**13.660**	**0.580**	**0.959**

Table 5 shows the performance results for LSSVM with PCA approach for both selected cases studies. Following the results in Table 5, the lowest MAE, and RMSE with the largest r and CE for Bernam river were extracted from LSSVM with 8PCs. Meanwhile, LSSVM with 7PCs was choose to represent PCA-LSSVM model.

4. Results and Discussions

For further analysis, the error statistics of SVM, LSSVM, PCA-SVM and PCA-LSSVM models were compared to each other to find the best model for river flow forecasting. Table 6 compares the results among the four approaches based on four statistical measurements, which are MAE, RMSE, r and CE.

The result obtained for Bernam River shows that the best approach was obtained from PCA-LSSVM with the lowest MAE, RMSE, and the highest r and CE statistics. Therefore, PCA-LSSVM is declared as the best model presenting Bernam River, followed by PCA-SVM. These models can be ranked as first, second best approaches for Bernam river, whereas single SVM is ranked the least good approached.

Table 6: Comparative Performances of All Models for Both Catchments

Rivers	Models	MAE	RMSE	r	CE
Bernam River	SVM	15.190	18.796	0.604	0.364
	LSSVM	14.766	19.776	0.547	0.296
	PCA-SVM	14.430	19.202	0.583	**0.863**
	PCA-LSSVM	**13.214**	**17.895**	**0.654**	0.881
Tualang River	SVM	**9.637**	**13.617**	**0.585**	0.440
	LSSVM	9.992	13.765	0.574	0.428
	PCA-SVM	10.215	13.929	0.592	0.958
	PCA-LSSVM	9.899	13.660	0.580	**0.959**

Based on percentage comparisons, the PCA-SVM produced some improvement over the SVM model for Bernam River, with about 137.08% improvements in CE and reductions in MAE with the value of 5.003%. Simirily, PCA-LSSVM were also reported have better forecasting results than LSSVM model. The PCA-LSSVM also shows improvements over LSSVM with the reductions in MAE and RMSE with the values of 10.51% and 9.511% respectively. The PCA-LSSVM also improved over LSSVM for both of CE and r with the values of 197.63% and 19.56% for Bernam River. Furthermore, Mean Forecasting Error (MFE) has been constructed to determine the capability of the forecasted models. Based on MFE, LSSVM with PCA has under-predicted the future monthly Bernam river flow data by 0.89% compare to 2.14% for PCA-SVM.

While SVM is ranked last for Bernam river, the model has proven it capability in predicting the future river flow value for Tualang river. The results showed that SVM is found having the best performance for Tualang river and it can be ranked as the first forecasting model while PCA-LSSVM is in the second place. Despite the satisfactory results achieved by SVM model, however, PCA-LSSVM is chosen to represent Tualang river. This is due to the fact that the variances between these two models for MAE, RMSE and R values are small and up to two-decimal place. Moreover CE statistics is shown having significant variance between both models with the total variation of 0.519. With that, PCA-LSSVM is chosen as the best model to represent the prediction capability for Tualang river.

Meanwhile, MAE and RMSE are also proved to have some reductions with the values of 5.99% and 2.29% respectively for Tualang river. Other than that, it can be observed that, there some significant improvement Tualang river with total percentage of 1.19% and 117.72% for both r and CE values. Aside from that, PCA-LSSVM also shows improvements over LSSVM for Tualang River. Based on the observations, PCA-LSSVM is better over LSSVM with reductions of 0.76%, 0.93% in RMSE and MAE with improvements of 124.06% and 1.04% for CE and r, respectively. These improvements attributable to the fact that the ability of the data pre-processing technique to eliminate the noises existed in the data.

Based on MFE comparisons for Tualang river, PCA-LSSVM has over-predicted the monthly river flow data by 2.08% compare to 3.28% for PCA-SVM.

5. Conclusions

This study explores the potential of using PCA as dimensionality reduction technique and it effect with combinations of LSSVM and SVM in river flow forecasting. Due to the complexity of the hydrologic system itself, an improvement in river flow forecasting model has always been an important task for researchers and hydrologist.

Previous researchers believed that PCA is useful tools in forecasting and it is a great tool in identifying components that can be used as potential variables in forecasting. They also stated that the use of PCA combined with forecasting model, had been proven having better representation than those without PCA. The findings in this study are consistent with other researches, as PCA-SVM and PCA-LSSVM have outperformed the single SVM and LSSVM, respectively and prove the claim made by those researchers that using PCA makes it easier to increase the prediction accuracy that if it were not used.

In the proposed approach, past monthly river flow data were analyzed using PCA, then single SVM and LSSVM model were setup and parameters were set accordingly. The results showed that the proposed PCA-SVM and PCA-LSSVM performed better and have good performance than single models. Other than that, there are two main conclusions successfully derived from the objectives.

First, although the results suggest that single SVM model performed better than single LSSVM model. However, the variances between these two models are small and almost insignificant. The results also proven that LSSVM has shown its capability in forecasting and undeniably, it can be used as an alternatif forecasting model. Based on author’s knowledge, there are very limited number of published research which comparing the performances between SVM and LSSVM. Most of the published research were comparing the performances either SVM or LSSVM and ANN.

Second, the accuracy of forecasting models which used dimensionality reduction technique has improved significantly over single forecasting model. With that, it can be noted that the application of PCA as dimensionality reduction tools has proven its superiority and it can be applied as an alternative way in order to increase the prediction accuracy.

Acknowledgement

This research is funded by Universiti Tun Hussein Onn Malaysia grant scheme under vote number H073 and Universiti Teknologi Malaysia.

References

[1] Afshin M, Sadeghian A & Raahemifar K (2007), On efficient tuning of LS-SVM hyper-parameters in short-term load forecasting: A comparative study. *Proc. of the 2007 IEEE Power Engineering Society General Meeting (IEEE-PES).*

[2] Astel A, Mazerski, J, Polkowska Z, Namiesnik J (2004), Application of PCA and time series analysis in studies of precipitation in Tricity (Poland). *Advances in Environmental Research.* 8(3-4): 337-349.

[3] Bhagwat PP & Maity R (2012), Multistep-Ahead River Flow Prediction Using LS-SVR at Daily Scale. *Journal of Water Resource and Protection.* 4: 528-539.

[4] Bhagwat PP & Maity R (2013), Hydroclimatic streamflow prediction using Least Square-Support Vector Regression. *Journal of Hydraulic Engineering.* 19(3): 320-328.

[5] Cao LJ, Chua KS, Chong WK, Lee HP, Gu QM (2003), A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing*. 55(1-2): 321-336

[6] Chau KW, Wang WC, Cheng CT, Qiu L (2009), A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology.* 374(3-4): 294-306.

[7] Dibike YB, Slavco V, Solomatine DP, Abbott MB (2001), Model Induction with Support Vector Machines: Introduction and Applications. *Journal of Computing in Civil Engineering*. 15(3): 208-216.

[8] Elshorbagy A, Corzo G, Srinivasalu S, Solomantine DP (2010), Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology. *Hydrology and Earth System Sciences*. 14(10): 1931-1941.

[9] Gestel TV, Suykens JAK, (2001), Financial time series prediction using least squares support vector machines within the evidence framework. *Neural Networks, IEEE Transactions.* 12(4): 809-821.

[10] Guo X, Sun X, Ma J (2011), Prediction of daily crop reference evapotranspiration (ET_0) values through a least-squares support vector machine model. *Hydrology Research*. 42(4): 268-274.

[11] Guhathakurta P, Rajeevan M, Thapliyan V (1999), Long Range Forecasting Indian Summer Monsoon Rainfall by a Hybrid Principal Component Neural Network Model. *Meteorology and ATM Ospheric Physics*. 71(3-4): 255-266.

[12] Hanbay, D. (2009). An expert system based on least square support vector machines for diagnosis of valvular heart disease. *Expert Systems with Applications.* 36(4): 8368-8374.

[13] Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez, L (2000), Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Research*. 34(3): 807–816.

[14] Hotelling H (1933), Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. (24): 417–441.

[15] Hu TS, Lam KC, N, ST (2007), Rainfall-Runoff Modelling using Principal Component Analysis and Neural Network. *Nordic Hydrology*. 38(2): 235-248.

[16] Jolliffe IT (2002), *Principal Components Analysis*. Second Edition. New York. Springer.

[17] Kang YW, Li J, Cao GY, Tu HY, Li J, Yang J (2008), Dynamic temperature modeling 10 of an SOFC using least square support vector machines. *Journal of Power Sources.* 179: 683-692.

[18] Khan MS & Coulibaly P (2006), Application of Support Vector Machine in Lake Water Level Prediction. *Journal of Hydrologic Engineering.* 11(3): 199-205.

[19] Kisi O (2004), River flow modeling using artificial neural networks. *Journal of Hydrologic Engineering.* 9(1): 60-63.

[20] Kisi O (2008), River flow forecasting and estimation using different artificial neural network technique. *Hydrology Research*. 39(1): 27-40.

[21] Knight DW & Shamseldin AY (2006), *River Basin Modelling for Flood Risk Mitigation*. London, UK. Taylor & Francis.

[22] Legates DR & McCabe Jr GJ (1999), Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. *Water Resources Research.* 35(1): 233–241.

[23] Lin GF, Chen GR, Huang PY & Chou YC (2009), Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods. *Journal of Hydrology*. 3(32): 17-29.

[24] Lin JY, Cheng CT, Chau KK (2006), Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*. 51(4): 599-612.

[25] Lin GF, Chen GR, Huang PY (2010), Effective typhoon characteristics and their effects on hourly reservoir inflow forecasting. *Advances in Water Resources*. 33: 887-898.

[26] Liong SY & Sivapragasam C (2002), Flood stage forecasting with support vector machines. *Journal of American Water Resources*. 38(1),173 -186.

[27] Liu Z, Wang X, Cui L, Lian ., Xu J (2009), Research on Water Bloom Prediction Based on Least Squares Support Vector Machine. *WRI World Congress on Computer Science and Information Engineering, 2009.*

[28] Maity R, Bhagwat PP, Bhatnagar A (2010), Potential of support vector regression for prediction of monthly streamflow using endogenous property. *Hydrological Processes*. 24: 917–923.

[29] Misra D, Oommen T, Agarwal A, Mishra SK, Thompson AM (2009), Application and analysis of support vector machine based simulation for runoff and sediment yield. *Biosystems Engineering*. 103: 527-535.

[30] Mishra S, Choubey V, Pandey SK, Shukla JP (2014), An Efficient Approach of Support Vector Machine for Runoff Forecasting. *International Journal of Scientific & Engineering Research*. 5(3): 158-166.

[31] Noori R, Abdoli MA, Ameri A, Jalili GM (2009), Prediction of municipal solid waste generation with combination of support vec-

tor machine and principal component analysis: A case study of Mashhad. *Environmental Progress and Sustainable Energy*. 28: (249-258).

[32] Noori R, Khakpour A, Omidva, B, Farokhni, A (2010), Comparison of ANN and Principal Component Analysis-Multivariate Linear Regression models for predicting the river flow based on developed discrepancy ratio statistic. *Expert Systems with Applications*. 37: 5850-5862.

[33] Okkan U & Serbes ZA (2012), Rainfall–runoff modeling using least squares support vector machines. *Environmetrics*. 23: 549-564.

[34] Ouyang, Y. (2005). Evaluation of river water quality monitoring stations by principal component analysis. *Water Research*. 39: 2621-2635.

[35] Page RM, Lischeid G, Epting J, Huggenberger P (2012), Principal component analysis of time series for identifying indicator variables for riverine groundwater extraction management. *Journal of Hydrology*. 432-433: 137-144.

[36] Parinet B, Lhote A, Legube B (2004), Principal component analysis: an appropriate tool for water quality evaluation and management—application to a tropical lake system. *Ecological Modelling*. 178: 295-311.

[37] Pearson K (1901), On lines and planes of closest fit to systems of points in space. Phil Mag. (6), 2, 559-572.

[38] Samsudi R, Saad P, Shabri A (2011), River flow time series using least squares support vector machines. *Hydrology and Earth System Sciences.* 15: 1835-1852.

[39] Sivapragasam C, Liong SY, Pasha MFK (2001), Rainfall and runoff forecasting with SSA–SVM approach. *Journal of Hydroinformatics*. 3. 141-152.

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

[41] Stathis D & Myronidis D (2009), Principal component analysis of precipitation in Thessaly Region (Central Greece). *Global NEST Journal*. 11(4): 467-476.

[42] Suykens JAK & Vandewalle J (1999), Least squares support vector machine classifiers. *Neural Processing Letter*. 9(3): 293-300.

[43] Suykens JAK, Gestel TV (2005), *Least Square Support Vector Machine.* New Jersey: World Scientific.

[44] Tay FEH, & Cao LJ (2001), Improved Financial Time Series Forecasting By Combining Support Vector Machines with Self-Organizing Feature Map. *Intelligent Data Analysis.* 5. 339–354.

[45] Twining CJ & Taylor CJ (2003), The use of kernel principal component analysis to model data distributions. *Pattern Recognition*. 36: 217-227.

[46] Vapnik V (1995), *The Nature of Statistical Learning Theory*. New York: Springer.

[47] Wang H & Hu D (2005), Comparison of SVM and LSSVM for Regression. *International Conference on Neural Networks And Brain, 2005.* 1: 279-283.

[48] Wang WC, Chau KW (2009), A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology* 374(3-4): 294-306.

[49] Wang S, Zhang X, Yu L, Lai KL (2009), Estimating the impact of extreme events on crude oil price: An EMD-based event analysis method. *Energy Economics*. 31: 768–778.

[50] Ye J & Xiong T (2007), SVM versus Least Squares SVM. The 11th International Conference on Artificial Intelligence and Statistics (AISTATS).640-647.

[51] Yu PS, Chen ST, Chang IF (2006), Support vector regression for real-time flood stage forecasting. *Journal of Hydrology.* 328(3-4): 704-716.

[52] Yunrong X & Liangzhong J (2009), Water Quality Prediction Using LS-SVM And Particle Swarm Optimization. *Second International Workshop on Knowledge Discovery And Data Mining, 900-904.*

[53] Zhang GP (2003), Time Series Forecasting Using A Hybrid ARIMA And Neural Network Model. *Neurocomputing,* 50: 159-175.

[54] Zhao Y, Dong Z, Li Q (2012), Application Study of Least Squares Support Vector Machines in Streamflow Forecast. *Applied Mechanics and Materials*. 212-213: 436-440.