International Journal of Civil Engineering and Technology (IJCIET)
Volume 10, Issue 06, June 2019, pp. 314-323, Article ID: IJCIET_10_06_031
Available online at http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=10&IType=6
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

DAILY FORECASTING OF DAM WATER LEVELS USING MACHINE LEARNING

Wong Jee Khai
Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional, 43000, Kajang, Selangor, Malaysia
Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional, 43000, Kajang, Selangor, Malaysia

Moath Alraih
Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional, 43000, Kajang, Selangor, Malaysia

Ali Najah Ahmed*
Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional, 43000, Kajang, Selangor, Malaysia

Chow Ming Fai
Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional, 43000, Kajang, Selangor, Malaysia
Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional, 43000, Kajang, Selangor, Malaysia

Ahmed EL-SHAFIE
Department of Civil Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia

Amr EL-SHAFIE
Civil Engineering Department, Giza High Institute for Engineering and Technology, Giza, Egypt
*Corresponding Author

ABSTRACT

The design and management of reservoirs are crucial towards the improvement of hydrological fields subsequently leading to better Integrated Water Resources Management (IWRM). Different forecasting models used in designing and managing dams have been developed recently. This report paper proposes a time-series forecasting model formed on the basis of assessing the missing values. This is followed by different variable selection to determination to gauge the reservoir's water level. The investigation gathered data from the Klang Gates Dam Reservoir as well as daily rainfall data. The two sets of data are consolidated into a coordinated set formed on *the basis of directing it as a research dataset. Furthermore, the proposed model applies a Time Series (TS) Regression Model to develop the forecasting model of the reservoir's water level. The tried results demonstrate that when the Time Series Regression forecasting model is used to select variables with complete variables, it gives a better forecast result than the SVM model.*

Key words: Model, Forecasting. Time Series Regression, Support Vector Machine

Cite this Article: Wong Jee Khai, Moath Alraih, Ali Najah Ahmed, Chow Ming Fai, Ahmed EL-SHAFIE3 and Amr EL-SHAFIE, Daily Forecasting of Dam Water Levels Using Machine Learning, *International Journal of Civil Engineering and Technology* 10(6), 2019, pp. 314-323.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=10&IType=6

1. INTRODUCTION

The Klang Gate Dam Reservoir is located in Gombak District, Selangor, Malaysia. It is also known as the Bukit Tabur Dam. It provides water for irrigation, hydroelectricity, domestic consumption, helps in controlling floods, water sports et cetera [1, 2, 3, 4, 5]. This makes it an important asset to the area. Consequently, it is equally important for the local administrators to plan and manage all their water resources completely through precise forecasting.

Past investigations of reservoir water levels have distinguished three essential issues:

- Very few studies on reservoir water levels exist. In fact, parallel examinations in the hydrological field apply machine learning procedures in forecasting water levels [1-4]. They concentrated on forecasting water levels during the flood stages in basins, reservoirs, lakes, pumping stations, et cetera. Majority of these water level forecasts on the flood stages gathered data on hurricanes, particular climate or seasonal rainfall.
- Very few variables have been applied in reservoir water level forecasting. Moreover, previous literature indicates the existence of few parallel studies on forecasting [5-9]. The water level was used as the dependent variable, while rainfall was used as the independent variable. Therefore, a couple of autonomous factors were chosen. It is extremely hard to know the main variable set in the reservoir water level.
- There were no ascription techniques used in any of the data sets of the reservoir's water level. Previous studies of water level forecasts in the hydrological field have demonstrated that the collected data are constant and long-term. However, majority of them fail to disclose ways of dealing with the missing values such as human error and mechanical failure.

Numerous reservoir-related anticipating models have been created with the primary aim of controlling and predicting the reservoir's state. While this is a commendable target, the outcomes acquired to date have frequently earnestly repudiated the genuine values because of the intricacy of hydrological forms [5-14]. Hence, different advancement strategies have been adjusted, including straight programming, dynamic programming, stochastic programming and hereditary calculations [8, 15]. More seasoned models usually fuse less intricate recipes in their forecast, delivering fewer specific outcomes. More up to date models have joined more hydrological data to accomplish reliable estimates [15-18]. This study gathered data on the Klang Gate Dam alongside the related data on daily atmospheric conditions to improve these issues. The two sets of data were combined into single dataset dependent on the date. Afterwards, the study credited missing values and chose a superior ascription technique to develop better forecasting models. Finally, we assessed the factors dependent on various models.

1.1. Objectives of Study

- To develop models using historical inflows and water level data to predict reservoir water level based on different methods or models.
- To investigate the potential of forecasting method that can be used to accurately and efficiently to predict dam water level.
- To determine, investigate, optimize and compare the best method based on the simplicity of the model, the effectiveness of the model development procedure, and the accuracy level achieved.

1.2. Support Vector Machine (SVM)

Analysts in the hydrological field have settled on SVM as the better forecasting technique when contrasted with past more models such as Artificial Neural Networks (ANNs) and Adaptive Neuro-Fuzzy Inference System (ANFIS) [20, 21, 22]. SVM have improved speculation capacity, novel, and comprehensively ideal structures and the capacity to be quickly prepared. Values make SVMs progressively robust, effective and stable [9, 13, 17, 23]. The precision of figures is uplifted with better speculation capacity, which is exceptionally critical in hydrology forecast. Exact execution is one element that is featured in SVMs that originates from the unrivaled Structural Risk Minimisation (SRM) [9, 14, 18, 24]. It guarantees the minimization of an upper bound on the natural hazard which takes out the issue of poor speculation. SVMs started their valuable need in problems that require characterization strategies however have advanced to be utilized in regression problems [7, 10].

SVMs defeat every one of the shortcomings displayed by ANNs while keeping up the values of ANNs. ANNS' model design is subjective. They do not stress speculation execution, unlike SVMs, which settle this issue in a thorough hypothetical setting [11, 21, 24]. The learning calculations of SVMs uncover various concealed units that solidly set up the model's design, disposing of any subjectivity. Customary neural system models frequently actualize the standard of observational hazard minimization to limit preparing mistakes, while the hidden idea of the SVMs is the minimization of the upper bound of the speculation error by striking a harmony between the mistake itself and the machine's ability. SVMs ensure a common ideal arrangement, dissimilar to conventional neural system models, which yield locally common mechanisms [24, 22].

One of the features of SVMs is adaptability, which is supplemented by the utilization of Kernel works that map the information to a higher and infinite dimensional space in a certain way. This mapping takes into consideration the linear arrangement in the more upper dimensional space to compare to a non-straight method in the underlying lower dimensional info space [5, 9, 13]. This is one of the fundamental reasons that SVM is the best decision for discovering arrangements, particularly in hydrology, where the issues are frequently non-straight. In distinguishing synthetic procedures, SVM is more precise than ANFIS at fitting the ideal elements at a somewhat higher computational expense [6, 7].

The design and management of reservoirs are crucial towards the improvement of hydrological fields subsequently leading to better Integrated Water Resources Management (IWRM). As mentioned earlier, various forecasting models used in designing and managing dams have been developed. This paper proposes a different approach of applying SVM to forecast daily water levels with a case study of the Klang Gate Dam, located in Malaysia. The proposed model uses both dam water levels L(t-i) and rain gauge R(t-i). Furthermore, the proposed SVM model uses the Kernel functions that have high flexibility in the forecasting computation to calculate the common optimal solution. This enables the mapping of data at a higher and infinite-dimensional space in an implicit way.

2. RESEARCH METHOD

2.1. Proposed Model

Reservoirs are not only important for domestic purposes but also for national purposes that contribute to economic development. Hence, it is immensely crucial to forecast long-tem water levels for reservoirs. Equally, it is important to plan and manage water resources comprehensively to ensure minimum costs. This paper proposes two models, a time-series forecasting model and an SVM model that are based on the attribution of missing values and chosing variables. The proposed model initially applied five attribution methods that include the median of close points, series mean, mean of close points, linear, and regression. The results are then compared with a delete approach to forecast the missing values. After determining the main variables that affects the daily water levels, the proposed strategy positioned the significance of the environmental factors by means of factor investigation. It then successively isolates the immaterial factors.

Four scenarios are used in this study: In the first scenario, only the daily precipitation data between time (t) and (t-7) was used, as shown in equation (2.1):

$$L = fR(t-i) \qquad i = \{0,1,2,3,4,5,6,7\} \tag{1}$$

where:

L = daily water level of the dam

R = daily rainfall

Second scenario, two inputs which are daily rainfall data for time (t) to (t-1) and daily dam water level for time (t-1) as shown in Equation 2.2:

$$L = f(R(t-i)\; L(t-1) \qquad i = \{0,1\} \tag{2}$$

The third scenario, using two inputs variable, the daily rainfall data for time (t) to (t-7) and daily dam level for time (t-1) to (t-7) as shown in Equation (2.3):

$$L = f(R(t-i)\; L(t-j) \qquad i = \{0,1,2,3,4,5,6,7\}\,; j = \{1,2,3,4,5,6,7\} \tag{3}$$

Finally, the fourth scenario is using the same input variables as in scenario three, but the data were normalized to be less or equal to one by dividing all data set by the maximum value as shown in Equation (2.4):

$$L = f(\hat{R}(t-i)\; \hat{L}(t-j) \qquad i = \{0,1,2,3,4,5,6,7\}\,; j = \{1,2,3,4,5,6,7\} \tag{4}$$

where:

$\hat{L}$ = normalized data set of (L) daily water level of the dam

$\hat{R}$ = normalized data set of (R) daily rainfall

The best scenario will then be determined based on data prediction accuracy.

The four statistical formulas listed below were also used to evaluate the forecasting efficiency in this study:

Root Mean Square Error (RMSE)

Mean Absolute Error (MAE)

Mean Absolute Percentage Error (MAPE)

Correlation Coefficient (R)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

$$R = \frac{\sum_{i=1}^{n}[(y_i - \bar{y})(\hat{y}_i - \tilde{y})]}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2(\hat{y}_i - \tilde{y})^2}}$$

3. RESULTS AND ANALYSIS

This section provides the results we get from the two models, the TS Regression and the SVM Models. The examination of the predicting limit of these models depended on the precision of their conjectures. We grouped the results into TS Regression data and SVM data. Afterwards, we used the two models to create data forecasts applying four scenarios. Both models were utilized to forecast the dam water level in a year. The best estimation of each model is considered the most appropriate and exact predicting scenario. We compared the estimation power of both models to determine their superiority. Four sorts of measurable assessment are utilized to evaluate the execution of the proposed models in various time delays in preparing, checking and testing information. The results are shown in table 1 below.

Table 1 Statistical evaluation of Time Series Regression Model and Support Vector Machine Model

Model	Evaluation Criteria	Scenario			
		1	2	3	4
Time Series Regression	RMSE	0.397411	0.384693	0.089438	0.000780
	MAPE	0.337437	0.328921	0.070163	0.059477
	MAE	0.322678	0.314541	0.067137	0.000591
	R	0.242556	0.333864	0.977641	0.983869
Support Vector Machine	RMSE	0.398440	0.409439	0.210117	0.001929
	MAPE	0.333257	0.347281	0.172963	0.152606
	MAE	0.318597	0.331865	0.165543	0.001516
	R	0.216641	0.125873	0.858351	0.893410

Model simulation was done to estimate the dam’s water level. Next, to that, we computed the RMSE as well as the correlation coefficient between the forecasted and recorded values. See table 1 above for specific results. From the table, it is evident that scenario 4 applies systemic old data of daily rainfall and the dam’s simulated water level with the lowest RMSE, MAPE and MAE values and a maximum R value among all the scenarios. Moreover, it can be seen from the table that Scenario 4 offers the top precision of data forecasting, while Scenarios 1 and 2 provide the second lowest. Besides, scenarios 1 and 2 also have erratic forecasts compared to scenarios 3 and 4. Scenario 4 gave the lowest values for validation errors RMSE, MAPE and MAE. However, scenario 3 in the TS regression model had a slightly higher correlation coefficient value. Scenario 4 topped for higher accuracy values, while the TS regression model bet the SVM model in prediction accuracy.

The results display a gradual improvement in model execution from scenario 1-2 in TS Regression i.e., R in scenario (1) 0.243 versus R in scenario (2) 0.334. A big inaccuracy from scenario 1 to 2 of the SVM was recorded where the values of RMSE, MAPE, and MAE somewhat expanded while the correlation coefficient decreased from 0.217 to 0.126. Nonetheless, However, it is important to note an improvement in inaccuracies in the third and fourth scenarios, using the suggested models, with a reduced percentage error.

A critical increment in the precision from scenario 2 to 3 and a high decrease in errors was accomplished. The R-value for TS Regression increased from 0.334 to 0.978 while for SVM increased from 0.126 to 0.858. Moreover, there was a significant drop in errors by 75% and 35% for both models respectively. This was possible through the use of accurate rainfall and water level inputs. We also notice that the precision of the predicted values somewhat improved in scenario 4 subsequent to normalizing the data providing a high value of correlation coefficient. It is demonstrated that the predicted values obtained by applying scenario 4 achieve the best results among the four scenarios. TS Regression gives better results and more precise data forecasts than the SVM. It demonstrates insignificant errors that are closer to 0 and high R-value which is practically adjacent to 1.

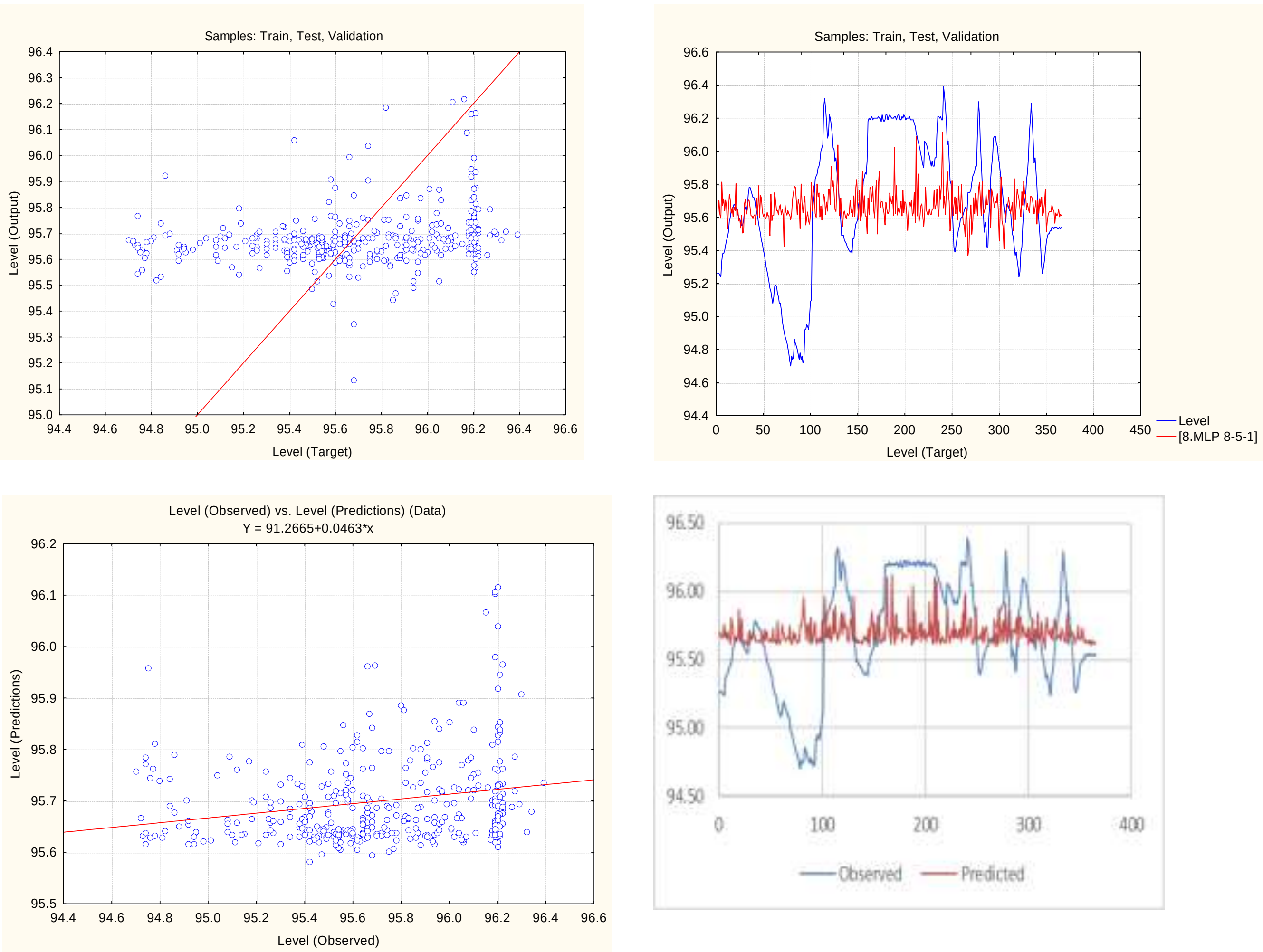

Figure 1 Correlation of observed dam level versus predicted dam level for TS and SVM scenarios one

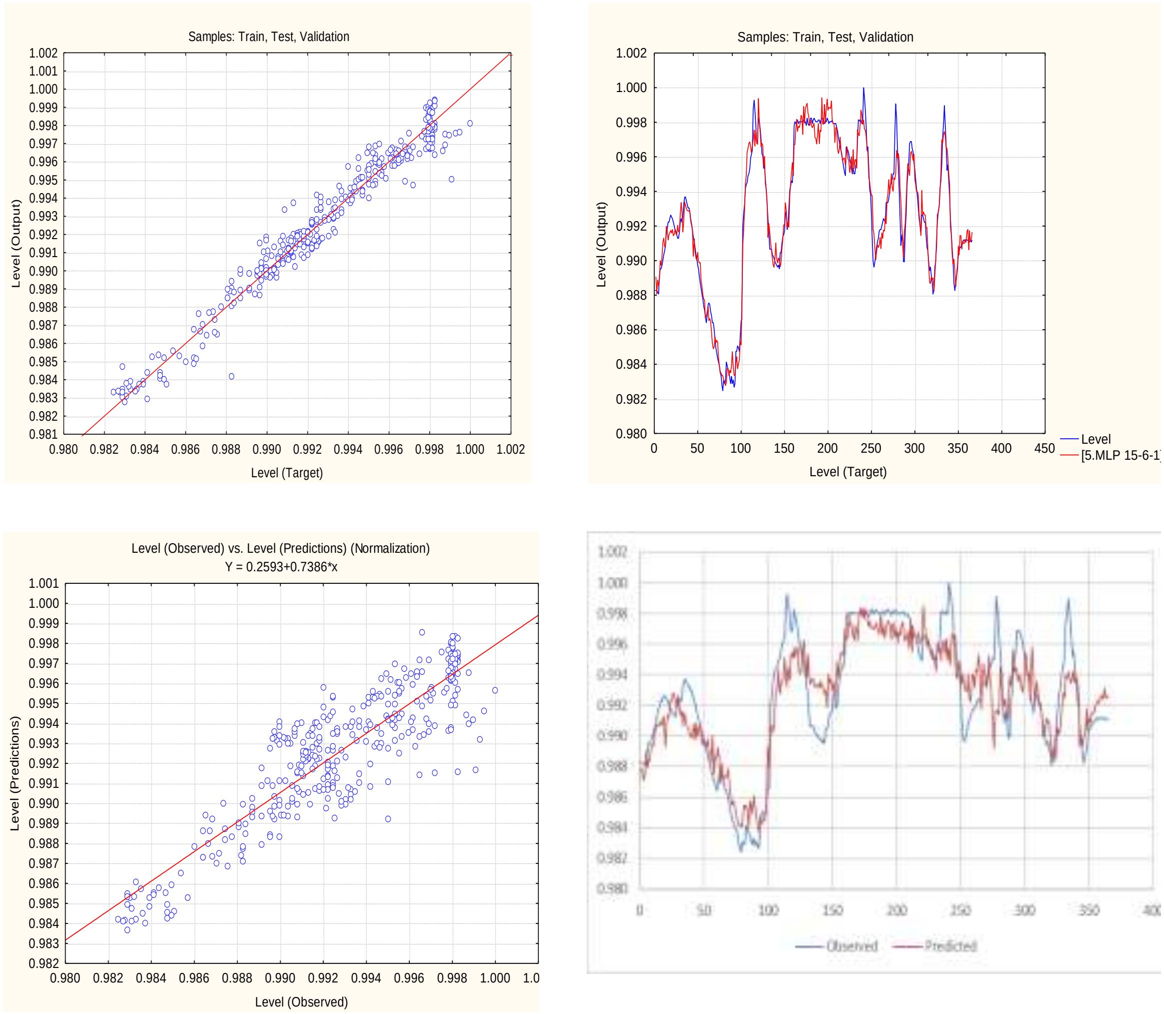

Figure 2 Correlation of observed dam level versus predicted dam level for TS and SVM scenarios four

Plotted graphs of the observed dam water levels compared to the forecasted dam water level in figure 1 indicates an inaccurate data distribution given the farness of both datasets from the line of best fit. Looking at figure 1, you notice a very big difference from the observed and predicted lines. This amounts to unreliable data prediction. Figure 1 present scatter graphs for both TS and SVM models for scenario 1. This scenario seems to offer minimal correlation and precision with inaccurate results for both AI models based on the values of the correlation between the recorded and predicted data. The line of best fit in figure 1 indicates poor correlation between the recorded and forecasted values evident from the sparse distribution of data away from the line of best fit. Moreover, it is shown in figure 1 that the forecasting precision is low and feeble where it is clearly found in the line of observed and estimated level. Subsequently, we conclude that scenario 1 is invalid and ought not be utilized in estimating models.

It is clear that figure 2 indicate the line of best fit where nearly all the data points lie or are very close to it. Moreover, the forecast line in figure 2 is almost at the base of the original line of the dam water level. Thusly, utilizing two input variables forms the most precise data predicting approach for dam water level. This data must be and normalized as applied in scenario 4. The TS regression model is the best for this work. Undoubtedly, the TS model outperforms the SVM model indicating its power and efficiency in predicting daily water level.

4. CONCLUSION

This paper proposes forecasting models using the Time Series Regression model alongside the Support Vector Machine model for the estimating water levels at the Klang Gate dam. In terms of efficiency, the paper puts the time series regression approach ahead of the SVM model. The analysis results indicate that when both forecasting models are used in choosing variables with full variables, they possess better forecasting prowess compared to the collection model. Daily rainfall and the dam's water level form the crucial variables. The suggested time series predictive model provides a superior forecast than the SVM that utilizes the four rating indices regardless of the selection presence or absence of variables. This confirms the compatability of the suggested time series forecast model in estimating water levels in the Klang Gate dam. Using the four accuracy measures, scenario 4 emerged the best for the data forecast where as the Time Series Regression emerged the best model because of its superior accuracy and dependable data forecast.Undoubtedly, the Time Series Regression model utilizing precipitation data and the dam water level is the top model given its accuracy and easiness to use. The model can be utilized for dependable data forecasting given its minimal error and higher correlation coefficient closer to one. Reservoirs that have conflicting set conditions such as weather and hydrological conditions can utilize the model to forecast hydrological states and promote the relevance of models. More research should be possible to enhance the hybrid models connecting wavelet deterioration with related AI models and creative calculations to anticipate hydrological factors with non-stationary and nonlinear connections. Additionally, further examinations ought to be led to evaluate the viability on the model on different info arrangement built from successful or wavelet parts.

ACKNOWLEDGEMENT

This work is supported by Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional Internal Grant: J510050716 and J510050822.

REFERENCES

[1] Akrami SA, El-Shafie A, Jaafar O (2013) Improving rainfall forecasting efficiency using Modified Adaptive Neuro-Fuzzy Inference System (MANFIS). Water Resour Manag, pp 1573–1650

[2] Bacanli UG, Firat M, Dikbas F. Adaptive neuro-fuzzy inference system for drought forecasting. *Stochastic Environmental Research and Risk Assessment.* 2009 Dec 1;23(8):1143-54

[3] El-Shafie A, Noureldin A. Generalized versus non-generalized neural network model for multi-lead inflow forecasting at Aswan High Dam. *Hydrology and Earth System Sciences.* 2011 Mar 11.

[4] Efe MO. A comparison of ANFIS, MLP, and SVM in the identification of chemical processes. In2009 *IEEE Control Applications,(CCA) & Intelligent Control,(ISIC)* 2009 Jul 8 (pp. 689-694). IEEE.

[5] El-Shafie AH, El-Shafie A, El Mazoghi HG, Shehata A, Taha MR. Artificial neural network technique for rainfall forecasting applied to Alexandria, Egypt. *International Journal of Physical Sciences.* 2011 Mar 18;6(6):1306-16.

[6] Hipni A, El-shafie A, Najah A, Karim OA, Hussain A, Mukhlisin M. Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro-fuzzy inference system (ANFIS). *Water resources management.* 2013 Aug 1;27(10):3803-23.

[7] An S, Liu W, Venkatesh S. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition.* 2007 Aug 1;40(8):2154-62.

[8] Li ZZ, Xie YB, Zhang J, Li XL. Sediment carrying capacity prediction based on chaos optimization support vector machines. In2010 *International Conference on Display and Photonics* 2010 Jul 21 (Vol. 7749, p. 774916). International Society for Optics and Photonics.

[9] Behzad M, Asghari K, Eazi M, Palhang M. Generalization performance of support vector machines and neural networks in runoff modeling. *Expert Systems with applications.* 2009 May 1;36(4):7624-9.

[10] Dastorani MT, Moghadamnia A, Piri J, Rico-Ramirez M. Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental monitoring and assessment.* 2010 Jul 1;166(1-4):421-34.

[11] Firat M, Turan ME, Yurdusev MA. Comparative analysis of fuzzy inference systems for water consumption time series prediction. *Journal of hydrology.* 2009 Aug 15;374(3-4):235-41.

[12] Guo J, Zhou J, Qin H, Zou Q, Li Q. Monthly streamflow forecasting based on improved support vector machine model. *Expert Systems with Applications.* 2011 Sep 15;38(10):13073-81.

[13] Lin HJ, Yeh JP. Optimal reduction of solutions for support vector machines. *Applied Mathematics and Computation.* 2009 Aug 15;214(2):329-35.

[14] Moghaddasi M, Araghinejad S, Morid S. Water management of irrigation dams considering climate variation: a case study of Zayandeh-rud reservoir, Iran. *Water resources management.* 2013 Apr 1;27(6):1651-60.

[15] El-Shafie A, Jaafer O, Akrami SA. Adaptive neuro-fuzzy inference system based model for rainfall forecasting in Klang River, Malaysia. *International Journal of Physical Sciences.* 2011 Jun 18;6(12):2875-88.

[16] Feng CX, Yu ZG, Kingi U, Baig MP. Threefold vs. fivefold cross-validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data. *Journal of manufacturing systems.* 2005 Jan 1;24(2):93-107.

[17] Tripathi S, Srinivas VV, Nanjundiah RS. Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of hydrology.* 2006 Nov 15;330(3-4):621-40.

[18] Hejazi MI, Cai X, Borah DK. Calibrating a watershed simulation model involving human interference: an application of multi-objective genetic algorithms. *Journal of Hydroinformatics.* 2008 Jan 1;10(1):97-111.

[19] Lu RS, Lo SL. Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. *Water Research.* 2002 May 1;36(9):2265-74.

[20] Lin JY, Cheng CT, Chau KW. Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*. 2006 Aug 1;51(4):599-612.

[21] Najah A, El-Shafie A, Karim OA, Jaafar O. Integrated versus isolated scenario for prediction dissolved oxygen at the progression of water quality monitoring stations. *Hydrology and Earth System Sciences*. 2011 Aug 29;15(8):2693-708.

[22] Kushwaha A, Jain MK. Hydrological simulation in a forest dominated watershed in the Himalayan region using the SWAT model. *Water resources management*. 2013 Jun 1;27(8):3005-23.

[23] Naji A, Ismail A, Ismail AR. Chemical speciation and contamination assessment of Zn and Cd by sequential extraction in surface sediment of Klang River, Malaysia. *Microchemical Journal*. 2010 Jul 1;95(2):285-92.

[24] Najah AA, El-Shafie A, Karim OA, Jaafar O. Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation. *Neural Computing and Applications*. 2012 Jul 1;21(5):833-41.