Middle-East Journal of Scientific Research 21 (11): 2182-2188, 2014
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2014.21.11.21844

Dissolved Oxygen Prediction Using Support Vector Machine in Terengganu River

[1]Afifah Tarmizi, [1]Ali N. Ahmed and [2]A. El-Shafie

[1]School of Ocean Engineering, University Malaysia Terengganu, UMT, Malaysia
[2]Department of Civil & Structural Eng, University Kebangsaan Malaysia, UKM, Malaysia

Abstract: This study was conducted using Support Vector Machine to predict dissolved oxygen for water quality in Terengganu River and see the difference between the two scenarios, using the five parameters (Scenario 1) and also using data from the previous station to predict dissolved oxygen the next stations (Scenario 2). This model has the ability to simulate water quality parameters to accurately prediction error are relatively small. The correlation coefficient indicates the highest value is 0.99, 0.97 and 0.96 after applying Scenario 2. The Type 1 of SVM regression and 10-fold cross-validation produce the accurate result. Found that, the MSE value is in range 0.009 to 0.714. Scenario 2 shows the better result than Scenario 1, with a significant improved from 1% to 2%.

Key words: Department of Environmental (DOE) · Chemical oxygen demand (COD) · Dissolved oxygen (DO) · pH and total suspended solid (TSS)

INTRODUCTION

All living thing especially humans use water for personal hygiene, transportation, agricultural production, industrial and manufacturing processes, hydroelectric power generation, recreation, navigation and a variation of other purposes.Water quality is a condition of water including chemical, physical and biological characteristics. Water quality can identify with do the sampling and water sample will be tested to know the level parameter of water. In ascertaining water pollution levels, Department of Environmental (DOE) has delineation the several parameters of water quality. Commonly, has six parameter to evaluate the surface water quality is consists of ammonia nitrogen (NH_3-N), biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), pH and total suspended solid (TSS) (Suhaimi *et al*., 2009) [1].

Dissolved oxygen is very important in determining the quality of water. DO is a type of gas dissolved in the water on the second sequence after nitrogen.DO content in water can be reduced by the process of respiration. Reductions of oxygen can cause obstruction of diffusion by salinity stratification.Low DO content can affect aquatic life and can cause a foul odor.Among these factors that influencing DO is temperature, flow speeds and turbidity. Water temperature is a main function in the water for aquatic life. It can handle the physiological functions of the organism with the components that affect the water quality of aquatic life. Water temperature plays an important role in the process of spawning and hatching, meanwhile water temperature can also cause death to aquatic life if the temperature is hot or cold dramatically.

Flow is an important for ecological factor to influence in the spread of dissolved salts, salt and food to organisms in the water. Flow is an important for ecological factor because influence in the spread of dissolved salts, salt and food to organisms in the water. Flow speeds are depending on the depth and width of the river bed. Typically, DO content in the high flow speeds is highest.Turbidity is the water clearness level, how far the sunlit can penetrate the water into river base. In the slow flows of water, turbidity was caused by the deposited materials at river base. If the turbidity is highest, then DO contained in the water is lower.

Nowadays, water quality level gradually dwindles because of urbanization. To find solution of pollution, a number of monitoring programs have been commenced. According to Singh *et al*. (2011) [2], water quality monitoring programs are very expensive, need time and manpower intensive and hence difficult to sustain over longer period. Najah *et al*. (2011) [3] stated,water quality modeling is the basis of water pollution control project

Corresponding Author: Afifah Tarmizi, School of Ocean Engineering, University Malaysia Terengganu, UMT, Malaysia.

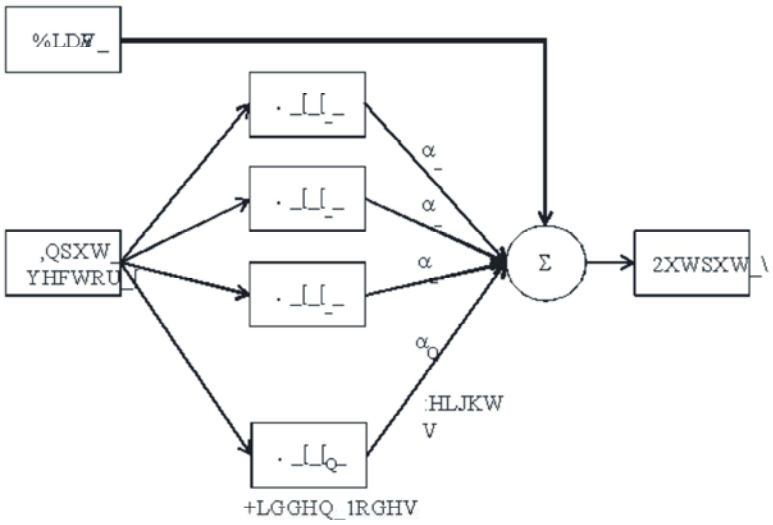


Fig. 1: Structure of SVM (Chen and Yu, 2007)

where with water quality modeling can predict the water quality tendency of varieties according to the current water environment quality condition, transfer and transformation rule of the pollutants in the river basin. Water quality modeling also yielded data without compromising with the quality and interpretability.

Support vector machine is a one of model used to predict water quality other than ANN. Tan *et al*. (2012) [4] stated, that the SVM and ANN have a large number of studies, nevertheless SVM have a lot advantage than ANN. It is because SVM can solve the small sample, nonlinear, high dimension and local minimum points and other practical issues. Hipni *et al*. (2013) [5] also stated,SVM is a modular design allows one to be implemented independently where their components can be designed. A local minimum does not have any effect on these models and they certainly do not face dimensionality problem. According to Liao *et al*. (2011) [6], support vector machine is an emerging machine learning technology that has already been used for water quality assessment in the field of environment. Figure 1 shows the structure of SVM.

According to Zhang (2001) [7], SVM is an output functional relationship is to assume that the output variable can be expressed approximately as a linear combination of its input vector components. These linear models include the linear least squares method for regression and the logistic regression method for classification. Because a linear model has limited prediction power by itself, there has been extensive research in nonlinear models such as neural networks.

History of SVM beginning in 1960s, where SVM was developed at AT&T Bell Laboratories by Vapnik. Due to this industrial context, SV research has up to date had a sound orientation towards real world application. SV classifiers became competitive with the best available system. Nevertheless, in regression and time series prediction applications obtained the excellent performances (Smola & Scholkoff, 2004) [8].

MATERIALS AND METHODS

Study Area Data Analysis: Terengganu is the state in the Malaysia Peninsular with an area of 12,955 km^2. The Terengganu River is important source of water supply for the state of Terengganu. It originated from Kenyir Lake and flows through the Kuala Terengganu and finally into the South China Sea. Figure 2 shows the map of the study area. This station includes four locationsthe main stream of the river. The river water quality parameters were monitored at four different stations by the Department of Environment (DOE) overa five year period from 2007 to 2011.

In this study provided two scenario, first scenario is predict DO using five parameter and second scenario is predict DO the next station from the previous stations. Five water quality parameters were selected for the SVM modeling in this study that is temperature (Temp), water pH, electrical conductivity (COND), nitrate (NO_3) and ammonia nitrogen (NH_3-NL).

These parameters were chosen because it representing the land use into the model. pH is important in water quality is to determine whether the water is acid

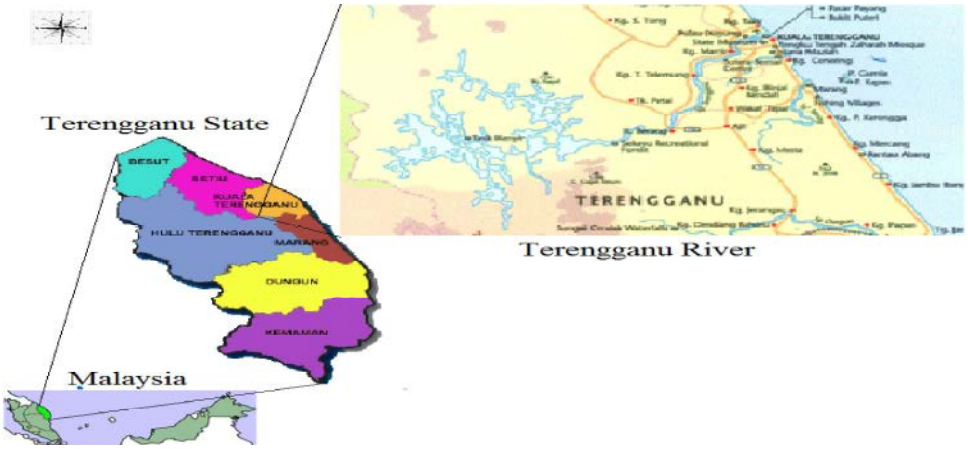


Fig. 2: Map showing the geographical setting of the study area

Table 1: Basic statistics of the measured water quality parameter in Terengganu River

Sampling Site		DO (mgl^{1})	COND (µs)	pH	NH_3-NL (mgl^{1})	TEMP (^{o}c)	NO_3 (mgl^{1})
DO-1	Mean	6.2507	22.0702	6.39561	0.14298	27.03	0.63788
	Min	2.91	3	5.49	0.01	24.08	0.01
	Max	7.34	56	7.83	1.07	30.33	2.93
	SD	0.6679	7.9292	0.49984	0.20246	0.93272	0.41124
	CV	10.6852	35.9272	7.81528	141.598	3.45067	64.4701
DO-2	Mean	6.14421	53.7526	6.21842	0.10074	27.22	0.66514
	Min	4.84	42	5.43	0.01	25.34	0.27
	Max	6.96	69.6	7.28	0.38	29.82	1.28
	SD	0.41378	8.30821	0.35534	0.09938	0.98139	0.22066
	CV	6.73447	15.4564	5.71438	98.6483	3.60539	33.1745
DO-3	Mean	5.79965	20.2807	6.36351	0.14596	27.3207	0.75398
	Min	4.43	1	5.67	0.01	23.35	0.01
	Max	6.78	45	8.41	2.46	31.93	4.7
	SD	0.68031	8.37656	0.47859	0.37206	1.28036	0.7154
	CV	11.7302	41.3031	7.52089	254.895	4.68642	94.8834
DO-4	Mean	5.54018	20	6.28807	0.14598	27.4347	0.64504
	Min	4.41	9	5.59	0.01	24.58	0.01
	Max	7.53	38	8.09	0.827	29.78	3.22
	SD	0.66269	5.17416	0.40917	0.19299	1.08466	0.40315
	CV	11.9616	25.8708	6.50706	132.201	3.95361	62.5005

or alkaline. Aquatic organisms and bacteria are sensitive to pH changes. Electrical conductivity is major water quality parameter due to the dilution effect of stream flow and can be used as general water quality indicator. The best value of conductivity for aquatic life is below 300µs/cm. Aquatic life can died if the value of conductivity exceeded 500µs/cm.

Nitrate usually came from human activity such as industrial, agricultural activities, human and animals waste. Nitrite and ammonium are more toxic than nitrate to aquatic life. Ammonia nitrogen (NH_3-NL) is used to measure the amount of ammonia that is a toxic pollutant often found in agriculture fertilizer and domestic sewage. NH_3-NL has been promoted as a tool to define the status of surface water quality in Malaysia (Najah *et al*., 2011)[9]. The high ammonia concentrations can stimulate excessive aquatic production and indicate pollution.

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The analyzed result of the study sites are given in Table 1.

$$cv(\%) = \frac{SD}{Mean} \times 100 \quad (1)$$

All parameter showed a coefficient of variation between 3.45067% and 254.895% for temperature and ammonia nitrogen. Temperature showed the lowers

Table 2: The correlation coefficient between DO and the input parameters

	COND	pH	NH_3-NL	TEMP	NO_3
DO-1	0.072705	0.442027	0.018208	0.271913	0.03234
DO-2	-0.10069	0.110975	-0.2035	0.251447	-0.17683
DO-3	0.094284	-0.25662	-0.02591	0.17145	-0.33867
DO-4	-0.06933	-0.15418	-0.26138	-0.10309	-0.22663

variation which might be due to the buffering capacity of the river (Najah *et al*., 2011). Geographical variations in the study area might influence the variable between samples. Table 2 shows the value of correlation coefficient between DO and the input parameter. The highest value is pH (0.442027), where pH value is approximately to 1.

Dissolved Oxygen Prediction: The SVM with the its non-linear and stochastic modeling proficiencies was consumed a prediction model that copycatted the DO pattern at the Terengganu River based on the five input parameter and can be expressed as follows:

$$DO_N = f_{MLP-NN}(pH_N, Temp_N, NO3_N, NH3_N, COND_N) \quad (2)$$

where DO_N is the dissolved oxygen at station N and $f_{MLP\text{-}NN}(.)$ is the non-linear function predictor built. The predicted DO at the previous station can be expressed follows:

$$DO_{N+1} = f_{MLP-NN}(pH_{N+1}, Temp_{N+1}, NO3_{N+1}, NH3_{N+1}, COND_{N+1}, DO_{pN}) \quad (3)$$

The performances of the models were evaluated according to three statistical indexes namely Coefficient of Efficiency (CE), Mean Square Error (MSE) and Coefficient of Correlation (CC).CE is often used to evaluate the performance, MSE can be used to determine how well the network output fits the desired output and CC is often used to evaluate the linear relationship between the predicted and measured dissolved oxygen. The three statistical indexes aredefined as follows:

$$CE = 1 - \frac{\sum_{i=1}^{n}(DOm - DOp)^2}{\sum_{i=1}^{n} DOm - \overline{DO}m)^2} \quad (4)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(DOm - DOp)^2 \quad (5)$$

$$cc = \frac{\sum_{i=1}^{n}(DOm - \overline{DO}m)(DOp - \overline{DOp})}{\sqrt{\sum_{i=1}^{n} DOm - \overline{DO}m)^2 \sum_{i=1}^{n} DOm - \overline{DO}p)^2}} \quad (6)$$

where n is the number of observations, Do_p and DO_m are the predicted and measured dissolved oxygen, respectively and DO_m is the average of measured dissolved oxygen. The accuracy improvement (AI) is to measure the significance of the proposed Scenario 2 over Scenario 1, is define as follow:

$$AI(\%) = \left(\frac{CCScen2 - CCScen1}{CCScen2}\right) x100 \quad (7)$$

RESULT AND DICUSSION

Table 3 and Table 4 show the value of correlation coefficient between all parameter and station (Scenario 1) and between all previous station and next station (Scenario 2). These shows the relationships between both scenarios are interrelated. Model 1 shows thehighest values for both scenarios.

The best model for Scenario 1 is Model 1, meanwhile for Scenario 2 is Model 1, Model 4 and Model 11. From this model, that shows the highest value for correlation coefficient in range 0.99 to 0.96 for both scenarios. According to Najah *et al*. (2011), [9] a value of R^2 should be approximately to 1, R^2 more than 0.9 indicates a very satisfactory model performance, a value between 0.6-0.9 indicates a fairly good performance and value below 0.5 indicates unsatisfactory performance.

For Scenario 1, DO-1 shows the highest value for all models, meanwhile for the other stations shows the equivalents value. MSE value for Model 1 shows the lower value compares the other models, the value is 0.032, 0.009, 0.036 and 0.041. In this study, Model 1 is the better model compares the other models because values are approximately to 0.

For Scenario 2, Model 7 shows the highest value is 0.311, meanwhile for the Model 11 shows the lower value is 0.009. At the station DO-3 shows the equivalent value within 0.077, 0.082 and 0.032. The value of MSE is in range 0.009 to 0.714. Najah *et al*. (2011) stated, the MSE value should approaching to 0 to get the right results. The smaller values of MSE ensure better performance [10].

DO predicted reached the best result when C equal to 10, ϵ equal to 0.1 and γ equal to 19. These results depict that this model can be used to deal with nonlinear

Table 3: Correlationcoefficient for Scenario 1

		CC			
Model	Input Parameters	DO-1	DO-2	DO-3	DO-4
1	Cond, pH, Ammonia, TEMP, Nitrate	0.97	0.97	0.96	0.94
2	pH, Ammonia, TEMP, Nitrate	0.91	0.96	0.89	0.92
3	COND, Ammonia, TEMP, Nitrate	0.89	0.94	0.88	0.91
4	COND, pH, Ammonia, Nitrate	0.85	0.92	0.87	0.87
5	COND, pH, TEMP, Nitrate	0.81	0.85	0.86	0.84
6	COND, pH, Ammonia, TEMP	0.80	0.82	0.82	0.81

Table 4: Correlationcoefficient for Scenario 2

		CC		
Model	Input Parameters	DO-2	DO-3	DO-4
1	DO-1	0.99		
2	DO-1		0.90	
3	DO-2		0.91	
4	DO-1,DO-2		0.97	
5	DO-1			0.65
6	DO-2			0.66
7	DO-3			0.70
8	DO-1, DO-2			0.74
9	DO-1, DO-3			0.89
10	DO-2, DO-3			0.90
11	DO-1, DO-2, DO-3			0.96

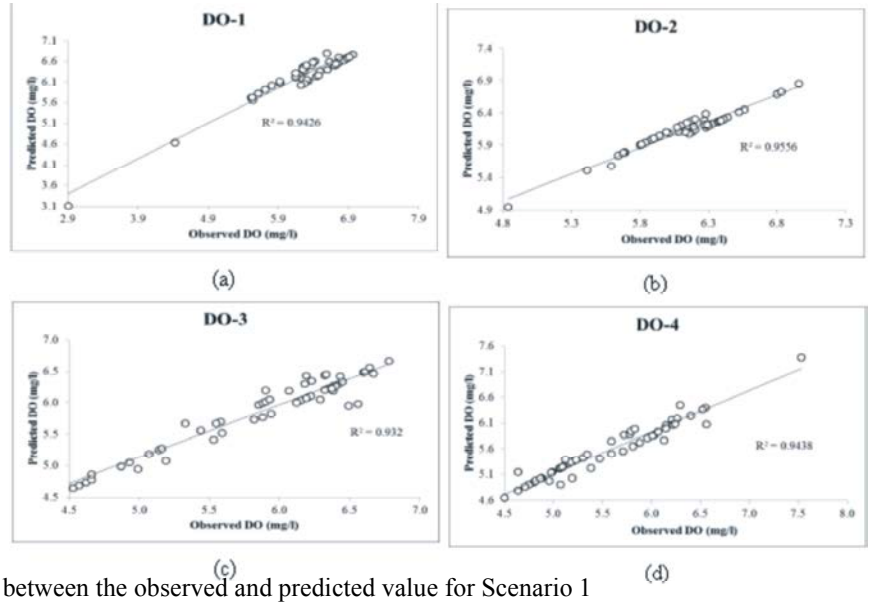
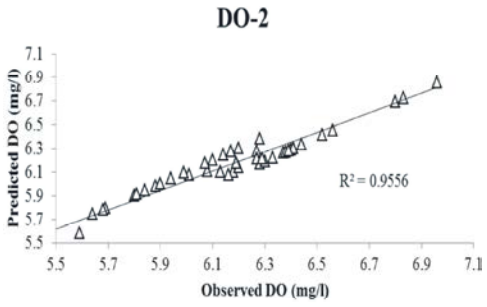


Fig. 3: Scatter plots between the observed and predicted value for Scenario 1



(e)

Fig. 3: Scatter plots between the observed and predicted value for Scenario 2

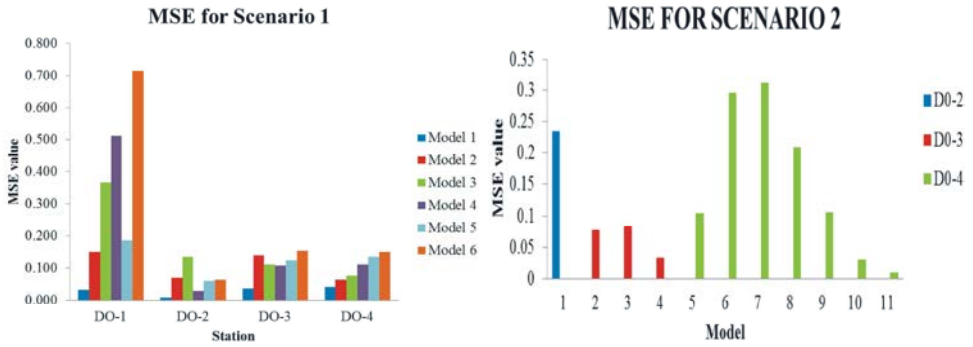


Fig. 4: MSE for Scenario 1 and Scenario 2

Table 5: The optimal SVM parameter

	SVM parameter				
Parameter	C	ϵ	γ	Kernel Type	CC
DO-1	7	0.1	10	RBF	0.97
DO-2	10	0.1	19	RBF	0.99
DO-3	10	0.1	19	RBF	0.97
DO-4	10	0.1	20	RBF	0.96

Table 6: Accuracy improvement for Scenario 2 over Scenario 1

	Scenario 1	Scenario 2	
Model	CC	CC	AI CC (%)
DO-2	0.97	0.99	2
DO-3	0.96	0.97	1
DO-4	0.94	0.96	2

issue to improve the precision of the water quality parameters prediction. According to Smola & Scholkoff (2004), commonly, four of kernel functions being used are Linear: $K(x,xj)=x^{T}xi$, Polynomial: $K(xi,xj)=(\gamma x^{T}xi+r)^{d}$, $\gamma>0$, RBF: $K(x,xj)=exp(-\gamma \|x-xi\|^{2})$, $\gamma>0$ and Sigmoid: $K(xi,xj)=\tanh(\gamma x^{T}xi+r)$. In this study, RBF was use because RBF can solve non-linear problems. The Type 1 of SVM regression and 10-fold cross-validation produce the accurate result.

In SVM case, it is important to determine approximate value of optimal hyper parameter C, ϵand γ. Parameter C, ϵ and γ were adopted in different range. The optimum value of C was determined through grid search over a space of 0.01-50000 with step of 10-1 and for γ is 0.001-20. ε was optimized in the range of 0.001 and 0.2 (Singh *et al*., 2011).

The range of C was set to [1-10] at increment of 1.0 and [0.1 - 0.5] at increment of 0.1 for ε and γ. The optimal values of hyper parameters are selected based on 10-fold cross validation repeated ten times until it reached the optimal result (Najah *et al*., 2011). Table 5 shows the best result for this study.

Table 6 shows that an accuracy improvement between two scenarios. After presenting Scenario 2 for all stations, the prediction accuracy was significantly enhanced. Scenario 2 was better than Scenario 1, with a significant improvement for all stations ranging from 1% to 2%.

CONCLUSION

Conclusion for this study is SVM can give a precise and robust result and able to give a fairly accurate prediction of water quality parameters.The model has ability to mimic the water quality parameters accurately with relatively small prediction error. The SVM can help in the optimization of water quality monitoring programs by reducing the number of sampling sites, frequency and water quality parameters. As a consequence, studies like this should be continued so that more water quality parameters willstudy and variety of possible scenarios. Henceforth, with this study the data's collector can take the initiative to develop a prediction model using SVM.

REFFERENCES

1. Suhaimi, S., A. Mohamad, A.L. Loh and M.T. Norhayati, 2009. Kajianindekskualiti air di Lembangan Sungai Paka, Terengganu (Water quality index study in Paka Basin, Terengganu), Journal of Sains Malaysiana, 38(2): 125-131.
2. Singh, K.P., N. Basant and S. Gupta, 2011. Support vector machine in water quality management, Analytica Chimica Acta, 703: 152-162.
3. Najah, A., A. El-Shafie, O.A. Karim, O. Jaafar and A.H. El-Shafie, 2011. An application of different artificial intelligences techniques for water quality prediction, International Journal of the Physical Sciences, 6(22): 5298-5308.

4. Tan, G., J. Yan, C. Gao and S. Yang, 2012. Prediction of water quality time series data based on least squares support vector machine, Procedia Engineering, 31: 1194-1199.
5. Hipni, A., A. El-shafie, A. Najah, A.O. Karim, A. Hussain and M. Mukhlisin, 2013. Daily forecasting of dam water level: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS), Water Recourse Management, 27: 3803-3823.
6. Liao, Y., J. Xu and W. Wang, 2011. A method of water quality assessment based on biomonitoring and multiclass support vector machine, Procedia Environmental Sciences, 10: 451-457.
7. Zhang, T. 2001. An introduction to support vector machine and other kernel-based learning methods (a review), Al Magazine, 22: 2.
8. Smola, A.J. and B. Scholkopf, 2004.A tutorial on support vector regression, Statistics and Computing, 14: 199-222.
9. Najah, A., A. El-Shafie, O.A. Karimm and O. Jaafar, 2011. Integrated versus isolated scenario for prediction dissolved oxygen at progression of water quality monitoring stations. Hydrology and Earth System Sciences, 15: 2693-2708.
10. Chen, S. and P. Yu, 2007. Pruning of support vector network on flood forecasting, Journal of Hydrology, 347: 67-78.