Application of artificial neural network models for predicting water quality index

5 authors, including:

Hafizan Juahir
Universiti Sultan Zainal Abidin
318 PUBLICATIONS 5,827 CITATIONS

SEE PROFILE

Sharifuddin Zain
University of Malaya
202 PUBLICATIONS 3,541 CITATIONS

SEE PROFILE

Mohd. Ikhwan Toriman
Universiti Sultan Zainal Abidin | UniSZA
202 PUBLICATIONS 3,648 CITATIONS

SEE PROFILE

Mazlin Mokhtar
National University of Malaysia (UNIVERSITI KEBANGSAAN MALAYSIA) ; UKM
308 PUBLICATIONS 4,124 CITATIONS

SEE PROFILE

APPLICATION OF ARTIFICIAL NEURAL NETWORK MODELS FOR PREDICTING WATER QUALITY INDEX

Hafizan Juahir[1], Sharifuddin M. Zain[1], Mohd. Ekhwan Toriman[2], Mazlin Mokhtar[3], Hasfalina Che Man[4]

[1] *Department of Chemistry, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia*

[2] *School of Social, Development and Environment, FSSK Universiti Kebangsaan Malaysia, Bangi 43600, Selangor.*

[3] *Institute of Environment and Development (LESTARI) Universiti Kebangsaan Malaysia, Bangi 43600, Selangor*

[4] *Department of Biology and Agriculture Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Selangor.*

Abstract: This study discusses the development and validation of an Artificial Neural Network (ANN) model in estimating water quality index (WQI) in the Langat River Basin, Malaysia. The ANN model has been developed and tested using data from 30 monitoring stations. The modeling data was divided into two sets. For the first set, ANNs were trained, tested and validated using six independent water quality variables as input parameters. Consequently, Multiple Linear Regression (MLR) was applied to eliminate independent variables that exhibit the lowest contribution in variance. Independent variables that accounted for approximately 71% of the variance in WQI are Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Suspended Solids (SS) and Ammoniacal-Nitrate (AN). The Chemical Oxygen Demand (COD) and pH contributed only 8% and 2% to the variance, respectively. Thus, in the second data set, only four independent variables were used to train, test and validate the ANNs. We found that the correlation coefficient given by six independent variables (0.92) is only slightly better in estimating WQI compared to four independent variables (0.91) which demonstrates that ANN is capable of estimating WQI with acceptable accuracy when it is trained by eliminating COD and pH as independent variables.

Key words: *Artificial Neural Network; Multiple Linear Regression; Water Quality Index (WQI).*

Abstrak: Kajian ini membincangkan pembangunan dan pengabsahan model Rangkaian Neural Buatan (ANN) dalam menganggar indeks kualiti air (WQI) di Lembangan Sungai Langat, Malaysia. Model ANN telah dibina dan diuji menggunakan data dari 30 stesen cerapan. Permodelan ini telah dibahagikan kepada dua set. Set pertama, enam pembolehubah kualiti air telah diuji dan disahkan dalam ANN sebagai parameter masukan. Oleh itu, Regresi Linear Berganda (MLR) telah digunakan bagi menyisihkan

pembolehubah-pembolehubah bersandar yang memberikan nilai yang kecil di dalam variasi. Pembolehubah-pembolehubah tak bersandar yang menerangkan kira-kira 71% dalam variasi indeks kualiti air ialah Oksigen terlarut (DO), Pemintaan Oksigen Biokimia (BOD), Pepejal Terampai (SS) dan Amoniacal Nitrogen (AN). Permintaan Oksigen Kimia (COD) dan pH masing-masing menyumbangkan hanya 8% dan 2% kepada variasi. Maka pada set yang kedua, hanya empat pembolehubah tak bersandar telah diguna dalam ANN bagi tujuan aturan, mengujian dan pengabsahan. Didapati pekali korelasi yang diberikan oleh enam pemboleh tak bersandar (0.92) memberikan nilai yang kurang baik dalam menganggarkan WQI berbanding hanya empat pembolehubah tak bersandar (0.91). Ini menunjukkan ANN berupaya menganggarkan WQI dengan ketepatan yang diterima apabila diuji tanpa COD dan pH sebagai pembolehubah tak bersandar.

Katakunc*i:Rangkaian Neural Buatan; Regresi Linear Berganda; Indeks Kualiti Air.*

1. Introduction

River systems have been substantially altered world wide, with massive ecological significance arising from the burgeoning global human population (Calder, 1999). River basin generally constitute areas with a high population density owing to favourable living conditions such as the availability of fertile lands, water for irrigation, industrial or drinking purposes and efficient means of transportation. Langat River Basins is the most rapid urban area in Malaysia. The Langat River catchment straddles the main urban conurbation in the Klang Valley forming parts of the growing urban complex in Selangor. Since 1970, Selangor has been experiencing extensive urban expansion from agriculture to urban-industrial-commercial landscape. Urbanization has taken place in the form of new industrial zones, housing estates and increase in population. According to Mohd. Ekhwan (2002), transformation from rural to urban landscapes also contributes to water quality changes based on rectification of topographical maps dated from the year 1970 to 1996 where a number of residential and industrial areas within 500 m from Langat River were computed.

The management of river water quality is a major environmental challenge. One of the major challenges is in determining point and non-point sources of pollutants. The discharge of industrial and municipal wastewater can be considered a constant polluting source, but not so for the surface runoff which is seasonal and highly affected by climate. In this regard, Malaysian government bodies that are responsible for the management of rivers in Malaysia are the Department of Environment (DOE) and the Department of Irrigation and Drainage (DID) with the cooperation of private sectors such as Alam Sekitar Malaysia Sdn. Bhd. (ASMA). One of the major works that had been carried out

in determining river water quality of the Langat river is the joint collaboration work of DOE and the Consultancy Unit of University Malaya (UPUM).

This analysis is based on the reports and data obtained from the above study. In this work, the development and optimization of different neural networks based on their predictive performance are discussed. To complement, Multiple Linear Regression (MLR) is also applied, which allows us to reduce the dimensionality of a highly dimensioned data set by explaining the correction amongst a large number of variables in terms of a smaller number of underlying factors without losing much information.

In natural environment, water quality is a multivariate phenomenon, at least as reflected in the multitude of constituents which are used to characterize the quality of water body. In an earlier study by Horton (1965), the various water quality data is integrated into an overall index, forming the water quality index. The study of WQI and its application as a successful tool for evaluating of water quality has increased exponentially since then (Stambuk-Giljanovic, 1999).

One of the challenges of today's water quality research is to reduce the cost and develop more intelligent computer aided tools in evaluating water quality. This study investigate the ability of Intelligent Predictive Tools (IPT) to estimate the WQI at Langat River basin. Approaches based on IPT are highly desirable in estimating the non-linear behaviour of urban water quality under historical and future scenarios. The most popular predictive model usually applied to model non-linear environmental relationship is the Artificial Neural Network (ANN) (Zhang and Stanley, 1997; Jain and Indurthy, 2003). The use of ANN in this study was justified from the availability of abundant historical data for the Langat River Basin. Successful application of ANN as a water quality forecaster at the Langat River Basin can be seen in several recent publications. Among the papers that have been published are; Hafizan et al. (2003) showed that the ANN model gives a better performance compared to the autoregressive integrated moving average (ARIMA) model in forecasting DO; the use of ANN for river regulation (Mohd. Ekhwan and Hafizan, 2003) and the application of second order backpropagation method (Hafizan et al., 2004).

2. Description of Study Area

The Langat River has a total catchment area of approximately 1815 km^2. It lies within latitudes 2° 40' 152" N to 3° 16' 15" and longitudes 101° 19' 20"E to 102° 01' 10"B. The catchment is illustrated in (Figure 1). The main river length is 141 km mostly situated about 40 km east of Kuala Lumpur. The Langat River has several tributaries with the principal ones being the Semenyih River, Lui River and Beranang River. There are two reservoirs, the Langat Reservoir and the

Semenyih Reservoir, located in the up-stream of the Langat River and Semenyih River catchments, respectively. The Langat Reservoir, built in 1981 has a catchment area of 54 km^2 while the Semenyih Reservoir, built in 1982 with the purpose to supply domestic and industrial water is 41 km$^{2.}$ The Langat Reservoir is also used to generate power supply at moderate capacity.

The climate is characterised by high average and uniform annual temperatures, high rainfall and high humidity. This climate has a dominant impact on the hydrology and geomorphology of the study area. Generally the study area experiences two types of seasons: the wet season starts in April until the end of November, and a relatively drier period from January to March. According to Dale (1959), the study area is located in the western region of Peninsular Malaysia.

The rainfall region for the west is differentiated on the basis of a double maximum occurring during the two inter- monsoon seasons in April and October to September, but the double minimum is prevalent during the two monsoon seasons. The highest intensity storms normally occur during the months of October, November, December and April. The driest months in the west are July and February.

Recently, water quality in the Langat River basin has become considerably poor due to rapid urbanization along the river. Several parts of the river are seriously polluted by untreated urban wastewater and by diffuse pollution originating from agricultural activities. The rapid growth of urban land use has been at the expense of agricultural land and reclamation of peat and swampy areas. Peat and swampy areas have been reduced from 298 km^2 in 1984 to about 155 km^2 in 1997. Agricultural related land use recorded slight reduction from about 1362 km^2 in 1984 to 1282 km^2 in 2001. Nevertheless, agricultural land uses remain a dominant feature of the entire basin. About 55% of the total land area within the basin is still devoted to agricultural use (Table 1). Most of these agricultural activities are permanent cropping (i.e. rubber, palm oil plantations and other permanent cropping).

Figure 1: The Langat River Basin

Table 1: Land use change within the Langat River Basin, 1984-2001

Land use/ cover (km^2)	1984	1990	1995	1997	2001
Agriculture	1362.98	1438.59	1300.76	1252.62	1283.07
Forest	831.35	689.49	718.58	735.52	560.05
Urban	38.41	123.52	194.23	223.08	434.13
Waterbody	-	5.11	9.67	13.53	36.96
Other	27.89	12.90	42.68	45.46	-

3. Methodology

Based on the WPCP report, the WQI for Langat river was evaluated based on six physico-chemical and biological determinants, namely Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Suspended Solids (SS), Ammoniacal-Nitrogen (AN) and pH (Table 2). This traditional approach needs high monitoring cost involving labour, equipment, chemicals etc.

Generally the WQI is calculated by summing up individual quality index (I_i) and weighting these parameters in a total quality evaluation. The final WQI is obtained using the following equation:

$$WQI = \sum_{i}^{n} w_i I_i \qquad (1)$$

where,

$$\sum_{i=1}^{n} w_i I_i = weighted \quad sum \qquad (2)$$

and,

I_i = water quality index of parameter i
w_i = weighting factor of parameter i
n = number of parameters

Table 2: Weighting factor of each parameter applied in calculating the WQI at Langat River   Basin.

Parameter	w_i
DO	0.22
BOD	0.19
COD	0.16
AN	0.15
SS	0.16
pH	0.12

We used, data from 30 sampling points monitored by DOE in collaboration with ASMA from which water quality parameters were observed (Figure 2). The historical water quality parameters (1995-2002) were applied to develop the ANN model consisting of 300 sets of observation.

3.1 Multiple Linear Regression

Multiple linear regression (MLR) was applied as well in this work to justify the relationship between the water quality parameters and their impact on WQI. An MLR model takes the form :

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \varepsilon \qquad (3)$$

where Y is the response variable, and there are $p-1$ explanatory variable x_1, $x_2,\ldots, x_{p-1}$, with p parameters (regression coefficients) β_o, β_1, $\beta_2,\ldots, \beta_{p-1}$.
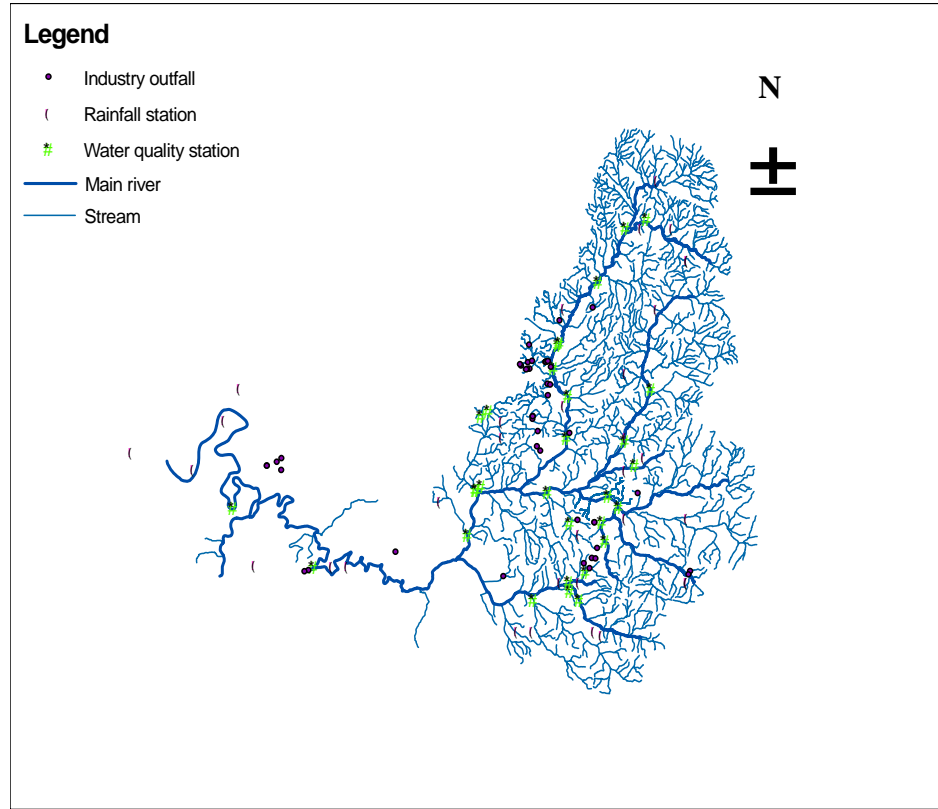

Figure 2: Selected sampling points in the Langat River Basin

In this study the adjusted R^2 value was calculated. This adjusted value indicates the loss of predictive power or shrinkage. While R^2 explains how much of the variance in Y is accounted for by the regression model, the adjusted value tells us how much variance in Y would be accounted for if the model is derived from the population from which the sample was taken. One version of R^2 that explains how well the model cross-validates is the Stein's formula below :

$$adjusted\ R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right]\left(1 - R^2\right) \tag{4}$$

In Stein's equation, R^2 is the adjusted value, n is the number of subjects and k is the number of predictors in the model. It is useful to use the equation to cross-validate a regression model.

3.2 Artificial Neural Network as WQI Prediction Model

In general, ANN models are specified by network topology, node characteristics and training or learning rules. It is an interconnected set of weights that contains the knowledge generated by the model. An ANN is composed of a large number of simple processing units, each interacting with others via excitatory or inhibitory connections (Figure 3). Distributed representation over a large number of units together with interconnectedness among processing units, provide a fault tolerance. Three different layers can be distinguished:

(i) An input layer - connecting the input information to the network. In this study two to six input nodes were applied, which are DO, BOD, COD, SS, AN and pH.

(ii) Hidden layer (one or more hidden layer) - acting as the intermediate computational layer. Multi-layer feed forward networks formed by only one hidden layer.

(iii) Output layer - producing the desired output which is in this case the WQI.
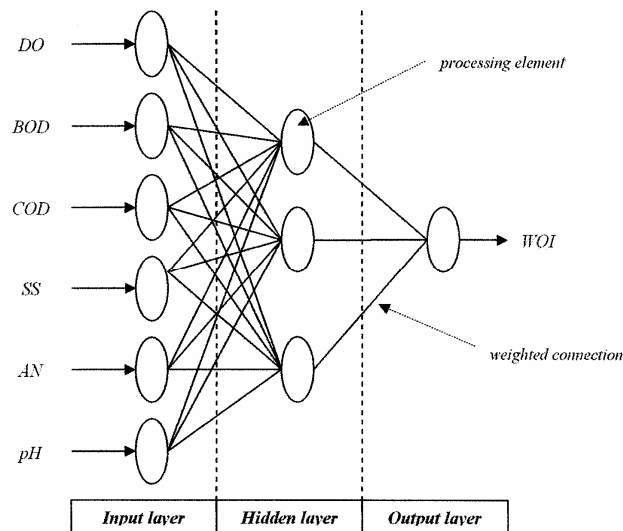
Figure 3: Example of ANN configuration

During training the error in the model's response to input examples is measured and the internal configuration is adjusted accordingly (weightings of connections between processing elements are changed) to reduce the error in the overall response (Sudheer *et al.*, 2003). Training is analogous to model calibration (Rojas, 1955). Levenberg-Marquardt Back Propagation algorithm was chosen as the training algorithm. The ANNs were trained starting from two inputs and 20 different initial networks, randomly initialized, from which the best performing network on the training data was chosen as the trained network. The training considered an early stopping approach and only the training set was used for determining weights and biases. Two activation functions were applied in these networks which are log-sigmoidal (logsig) and linear transfer function (purelin).

4. Results and Discussion

4.1 Multiple Linear Regression

One of the objectives of this work is to reduce the number of parameters needed to carry out water quality prediction without much loss of information. To meet this objective, MLR was employed to investigate the relationship of each water quality parameters to the dependent variable (WQI). In this study four variables meet the entry requirement to be included in the equation (DO, BOD, SS and AN)

which accounted for approximately 71% of the variance in WQI. Two variables did not meet the entry requirement, namely the COD and pH, which contributed only 8% and 2%, respectively. In relation to WQI, significant correlations were observed for the four chosen variables (r_{DO} = 0.735, r_{BOD} = -0.542, r_{SS} = -0.261 and r_{AN} = -0.418, p < 0.0001) while non-significant correlations were observed for COD and pH (r_{COD} = -0.138 and r_{pH} = 0.152, p > 0.0001).

The summary of MLR is shown in Table 3. The column labeled R contain values of the multiple correlation coefficient between the independent variables or predictors (DO, BOD, SS and AN) and the dependent variable or outcome (WQI). When only DO used, a simple correlation between DO and WQI (R=0.735) is observed. Using two predictors, DO and BOD, the correlation is 0.786. Using predictors DO, BOD and SS, the correlation is 0.809 and finally, upon adding AN as the fourth predictor, the correlation coefficient is 0.821.

The next column gives the values of R^2, which is a measure of how variability in the outcome is accounted for by the predictors. For the first model its value is 0.540 which means that DO accounts for 54% of the variation in WQI. However, when the other three predictors are included (model 4), this value increases to 0.673 or 67.3% of the variance in WQI. That means, if DO accounts for 54%, subsequent inclusion of BOD, SS and AN account for an additional 13.3%.

Table 3: Summary of regression model

Model	R	R-square	Adjusted R-square	Std. Error of the Estimate
1	0.735[a]	0.540	0.539	13.12502
2	0.786[b]	0.617	0.615	11.99071
3	0.809[c]	0.655	0.652	11.40159
4	0.821[d]	0.673	0.669	11.11370

a. Predictors: (Constant), DO
b. Predictors: (Constant), DO, BOD
c. Predictors: (Constant), DO, BOD, SS
d. Predictors: (Constant), DO, BOD, SS, AN

The adjusted R^2 demonstrates how well the model generalizes and ideally the values should be very close to R^2. In this study the difference for the final model is small. The difference between the adjusted R^2 from the R^2 was only 4% (0.0673 – 0.669). This shrinkage means that if the models were derived from the population rather than a sample it would account for approximately 0.4% less variance in the outcome. To gain clearer evidence, Stein's formula was applied to the R^2 to get some idea of the likely values in different samples. The formula was given in equation (4) and can be applied by replacing *n* with the sample size (N=329) and *k* with the number of predictors. This gives an adjusted R^2 equal to 0.659 indicating that the cross-validity of this model is very good.

The β values (see Table 4) indicate the relative influence of the entered variables. DO has the greatest influence on WQI (β=0.541), followed by BOD (β=-0.295), SS (β=-0.208) and then AN (β=-0.147). The direction of influence for DO is positive and the other three is negative. The model can be written in the following form:

$$WQI = 46.822 + 4.483DO – 0.420BOD – 0.006SS – 1.156AN \qquad (5)$$

The β values explain the relationship between WQI and each predictor. A positive value indicates that there is a positive relationship between predictor and the outcome whereas a negative coefficient represents a negative relationship. In this study three out of four predictors have negative β values. In this case, the WQI shows a decrease when BOD, SS and AN increase. Meanwhile only DO exhibits positive β values indicating that the increase of DO will be followed by an increase in WQI.

Table 4: Coefficient of regression model

Predictors	Unstandardized Coefficients		Standardized Coefficients	t	Significant level (p)
	B	Std. Error	Beta		
(Constant)	46.822	1.905		24.576	<0.001
DO	4.483	0.301	0.541	14.877	<0.001
BOD	-0.420	0.050	-0.295	-8.392	<0.001
SS	-0.006	0.001	-0.208	-6.439	<0.001
AN	-1.156	0.272	-0.147	-4.249	<0.001

4.2 Artificial Neural Network

Based on the MLR method, only four variables (DO, BOD, SS and AN) meet the requirement as good predictors for WQI. However, to justify the best predictor combination for an ANN model, six and five variable models were also trained. Six different models were employed to evaluate their performance in predicting WQI. The ANNs were trained using two to six parameters. The forecasting performance of each model is shown in Table 5 and Figure 4.

Model NN [5,1,8,1] with five predictors, one hidden layer, 8 hidden neurons and one output neuron performed the best in the training phase with SSE of 1.86 while the model with six predictors performed the worst with an SSE of 19.06. In the testing phase, model NN [6,1,10,1] with six predictors performed the best result in predicting WQI with SSE of 10.0951 and the worst model is NN [3,1,12,1] with SSE of 18.542. In the validation phase, model NN [4,1,18,1]

performed the best prediction of WQI with the lowest SSE of 9.732 and CC of 0.9501. The model with three predictors (NN [3,1,12,1]) performed the worst.

Table 5: Performance of ANN model

Models	No. of hidden neurons	SSE (Sum of Squared Error)			Correlation Coefficient (CC)
		Training	Testing	Validation	
NN[4,1,18,1]	18	2.442	11.624	9.732	0.950
NN[5,1,8,1]	8	1.857	11.762	11.479	0.940
NN[2,1,6,1]	6	8.752	15.247	15.021	0.917
NN[6,1,10,1]	10	19.061	10.095	15.160	0.919
NN[3,1,12,1]	12	4.255	18.542	16.669	0.913

The observed and predicted SSE and CC from the validation phase for each model is graphically shown in Figure 4. The graphical results also indicate that the four predictors model (NN [4,1,18,1]) performs a lot better than the other ANN models. All figures clearly show that the performance increases with an increase in the number of hidden neurons. Increasing from one to 18 hidden neurons, results in an increase in the CC and decrease in the SSE value for the four predictor models. The figure values also exhibit that after the optimal numbers of neuron (number of neuron which gives the best performance) have been reached, the overall performance of each model begins to deteriorate.

These results indicate that the ANN agrees well with the MLR model. ANN model with predictors giving the highest contribution of variance is seen as a good combination set of input.

5. Conclusion

This study clearly shows that it is possible to reduce the number of parameters in predicting water quality without much loss of information. Using MLR, COD and pH contributed little to the variance of the WQI and consequently, these parameters were left out in the new MLR model for WQI prediction. From the ANN analysis, it is also observed that the model with 4 input parameters exhibit the best predictive performance. This model exhibits its optimum performance when 18 hidden neurons were employed. The ANN performance observed in this study agrees well with the decision based on MLR to remove COD and pH as predictors.
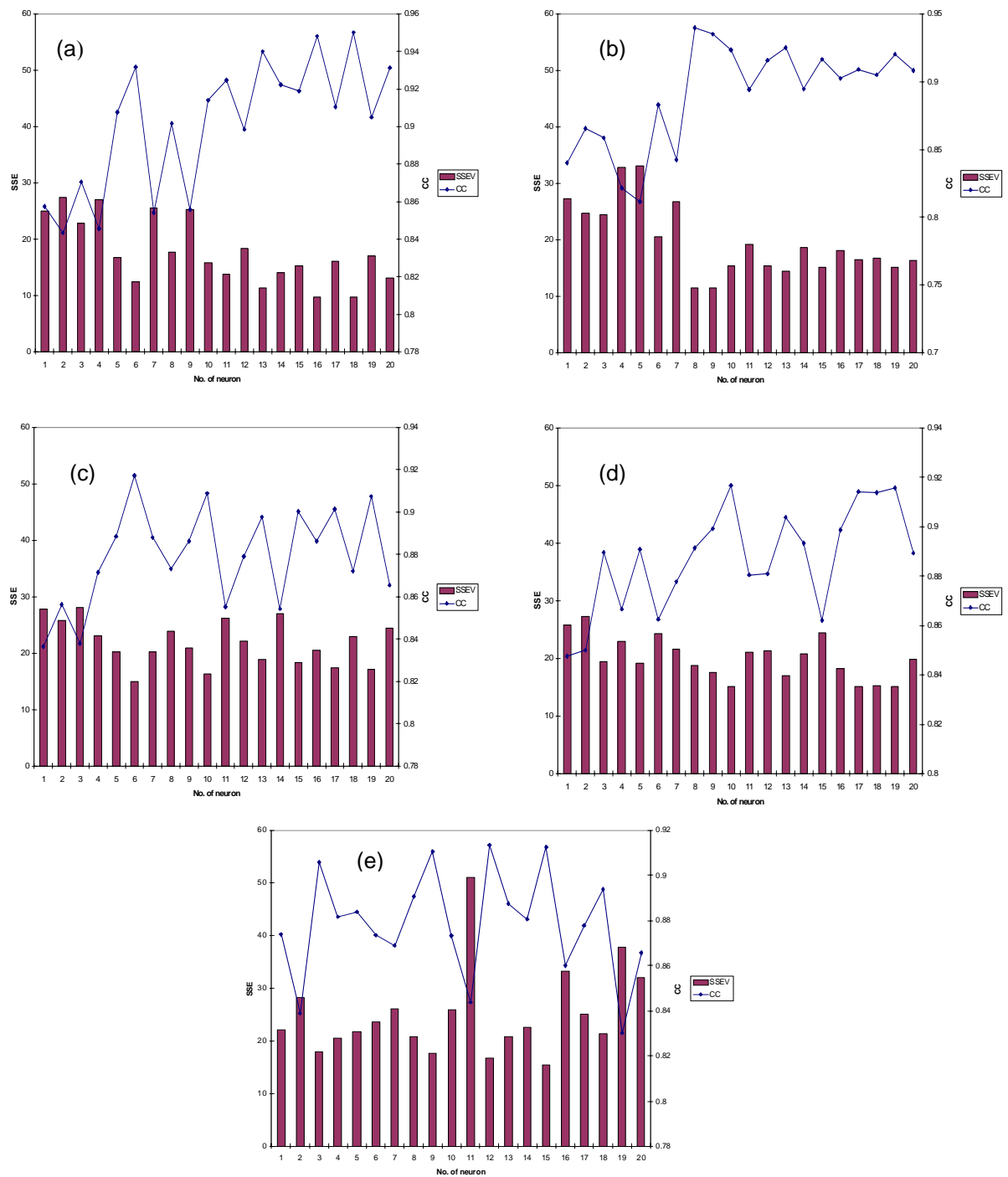
Figure 4: The performance of ANN model with different combination of network (a) [4,1,18,1], (b) NN[5,1,8,1], (c) NN[2,1,6,1], (d) NN[6,1,10,1] and (e) NN[3,1,12,1].

References

Calder, I.R., (1999). *The Blue Revolution: Land Use and Integrated Water Resources Management.* Earthscan Publications Ltd: London

Dale, W.L. (1959) The rainfall of Malaya, Part 1. *Journal Tropical Geography*, 13: 23-37.

Hafizan J., Sharifuddin M. Z., M. Nazari J. and Zainal A. (2004) Water discharge modeling by neural networks. Application of second order backpropagation method to improve river management at Langat River, Malaysia. Proceedings of the Water Environmental Planning: Towards Integrated Planning and Management of Water Resources for Environmental Risks. Bureau of Consultancy & Entrepreneurship, International Islamic University Kuala Lumpur, p. 307-324

Hafizan J., Sharifudin M. Z., Mohd Ekhwan T., M. Nazari J. and Klaewtanong, W. (2003) Performance of autoregressive integrated moving average and neural network approaches for forecasting dissolved oxygen at Langat River. *In* Noorazuan M.H. and Ruslan R. (eds.) Urban ecosystem studies in Malaysia: A study of change**.** Florida: Universal Publishers, p. 145-165.

Horton, R.K. (1965) An index number system for rating water quality. *Journal of Water Pollution Control Fed.,* 37(3): 300-305

Jain, A. and Prasad Indurthy, S. K. V. (2003) Comparative analysis of event-based rainfall-runoff modeling techniques-deterministic, statistical, and artificial neural networks, *Journal of Hydrologic Engineering*, 8: 93-98.

Mohd Ekhwan T. (2002) *Investigations into channel instability and river morphological change: The Langat River, Peninsular Malaysia.* Ph.D Thesis. Faculty of Science, Agriculture and Engineering, University of Newcastle Upon Tyne, 255 pp.

Mohd Ekhwan T. and Hafizan J. (2003) Artificial neural network modeling for Langat River discharge: Implication for river restoration. *In* Ajib C.K., Awang Hasmadi A.M., Baszley Bee, B.B. Idris, Fuziah Kartini A., H.B. Rozmi I. and Mohd Safar H (eds*.).* Abstract of Inovation and Research Expo, Universiti Kebangsaan Malaysia (Gold medal), B17.

Rojas, R. (1955). *Neural networks: A Systematic introduction*, Springer-Verlag, Berlin, 27-448.

Stambuk-Giljanovic, N (1999). Water quality evaluation by index in Dalmatia. *Water Resources*, 33:3423-3440.

Sudheer, K.P., Nayak, P.C and Ramasastri, K.S. (2003) Improving peak flow estimates in Artificial Neural Network river flow models. *Hydrological Processes,* 17: 677-687.

UPUM (2002) *Draf Laporan Akhir: Program Pencegahan Pencemaran dan Peningkatan Kualiti Air Sungai Langat.* Unit Perundingan Universiti Malaya.

Zhang, Q. and Stanley, S. J. (1997) Forecasting raw-water quality parameters for North Saskatchewan River by neural network modeling. *Water Resources*, 31: 2340-2350.