12

Using Principal Component Scores and Artificial Neural Networks in Predicting Water Quality Index

Rashid Atta Khan[2], Sharifuddin M. Zain[2], Hafizan Juahir[1],
Mohd Kamil Yusoff[1] and Tg Hanidza T.I.[1]
[1]*Department of Environmental Science, Faculty of Environmental Study, University Putra Malaysia, Serdang*
[2]*Chemistry Department, Faculty of Science, University of Malaya, Kuala Lumpur*
Malaysia

1. Introduction

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. Industrial and municipal wastewater discharges can be considered as constant polluting sources, unlike surface water runoff which is seasonal and highly affected by climate. According to Aiken et al. (1982), 42 tributaries in Peninsular Malaysia are categorized as very polluted including the Langat River. Until 1999, there were about 13 polluted tributaries and 36 polluted rivers due to human activities such as, industry, construction and agriculture (Department of Environment, Malaysia (DOE), 1999). In 1990, there were 48 clean rivers classified as clean but the number is reduced to 32 rivers in 1999 (Rosnani Ibrahim, 2001).

Surface water pollution is identified as the major problem affecting the Langat River Basin in Malaysia. Increase in developing areas within the river basin has in turn increased pollution loading into the Langat River. To avoid further degradation, the DOE have installed telemetric stations along the river basin to continuously monitor the water quality. As a result, abundant data were collected since 1988. There are 927 monitoring stations located within 120 river basins throughout Malaysia. Water quality data were used to determine the water quality status and to classify the rivers based on water quality index (WQI) and Interim National Water Quality Standards for Malaysia (INWQS). WQI provides a useful way to predict changes and trends in the water quality by considering multiple parameters. WQI is calculated from six selected water quality variables, namely dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solid (SS), ammonical nitrogen (AN) and pH (DOE, 1997). It is a well-known phenomenon that the contribution of pollution loading into river systems from the environment involves a complex interaction of many factors (e.g. chemical, physical and meteorological interaction). These primary pollutants are emitted from land use activities surrounding the river basin (e.g. agriculture, forest, urban, industrial and others) Rapid urbanization along the Langat River plays an important role in the increase of point source

(PS) and non-point source (NPS). In view of this complex interaction, use of modelling techniques to solve this problem, is needed. However, the problem of obtaining models that adequately represent the dynamic behaviour of field data is not easy. Lack of good understanding and description of the phenomena involved, the availability of reliable and complete field data set and the estimation of the numerous parameters involved are the major factors contributing to this problem. Beck (1986) noted that, increase in model complexity will undoubtedly increase the number of parameters, leading to the problems of identification.

Applications of ANN (Artificial Neural Networks) to environmental problems are becoming more common (Silverman and Dracup, 2000; Scardi, 2001; Recknagel et al., 2002; Bowden et al., 2005; Muttil and Chau, 2007). The applications of ANN, which are computing systems that were originally designed to simulate the structure and function of the brain (Rumelhart et al, 1986) is a relatively new concept in environmental modeling. If trained properly, a neural network model is capable of 'learning' linear as well as the nonlinear features in the data (Elsner and Tronis, 1992).

ANN consists of a set of simple processing units (neurons) arranged in a defined architecture and connected by weighted channels which act to transform remotely-sensed data into a classification. The classification techniques of ANN are unlike the conventional ones. It is distribution-free, may sometimes use small training sets (Hepner et al., 1990) and, once trained; it is rapid computationally, which will be of value in processing large data sets (Gershon and Miller, 1993). Furthermore, ANNs have been shown to be able to map land cover more accurately compared to many widely used statistical classification techniques (Benediktsson et al., 1990; Foody et al., 1995) and alternatives such as evidential reasoning (Peddle et al., 1994).

It has been proposed that the best tool to model non-linear environmental relationship is ANN (Zhang and Stanley, 1997; Jain and Indurthy, 2003). Research have been undertaken at Imperial College, London which attempts to investigate the capability of ANN approach in modelling spatial and temporal variations in river water quality (Clarici, 1995). ANNs were used as a predictive model to predict cyanobacteria Anabaena spp. in the River Murray, South Australia (Maier et al., 1998). DeSilets et al. (1992), have also used ANN to predict salinity. Ha and Stenstrom (2003), proposed a neural network approach to examine the relationship between storm water quality and various types of land use.

ANN has been successfully applied on the study of river water quality in Malaysia (Zarita Zainudin, 2001; Mohd Ekhwan Toriman and Hafizan Juahir, 2003; Hafizan Juahir et al., 2003a,b; Hafizan et al, 2004a,b; 2005; Ruslan Rainis et al., 2004). An approach for identifying possibilities of water quality improvement could be developed by using this concept. Such information could provide opportunities for better river basin management to control river water pollution in Malaysia. In the Malaysian context, Hafizan Juahir et al. (2003a) 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 Toriman and Hafizan Juahir, 2003) and the application of the second order back propagation method (Hafizan Juahir et al., 2004a) on water quality of the Langat River have also been demonstrated.

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. Water quality is very difficult to model because of the different interactions between pollutants and meteorological variables. The principal component analysis (PCA) is one of the approaches to avoid this problem and has received increasing attention as an accepted method in environmental pattern recognition (Simeonov et al., 2003; Wunderline et al., 2001; Helena et al., 2000; Loska and Wiechula, 2003)

The objective of this study is to use the PCA method to classify predictor variables according to their interrelation, and to obtain parsimonious prediction model (i.e., model that depend on as few variables as necessary) for WQI with other physico-chemical and biological data as predictor variables to model the water quality of the Langat river. For this purpose, principal component scores of 23 physico-chemical and biological water quality parameters were generated and selected appropriately as input variables in ANN models for predicting WQI.

2. Methodology

2.1 The data and monitoring sites

The water quality data in this study were obtained from seven stations along the main Langat River (Fig. 1).

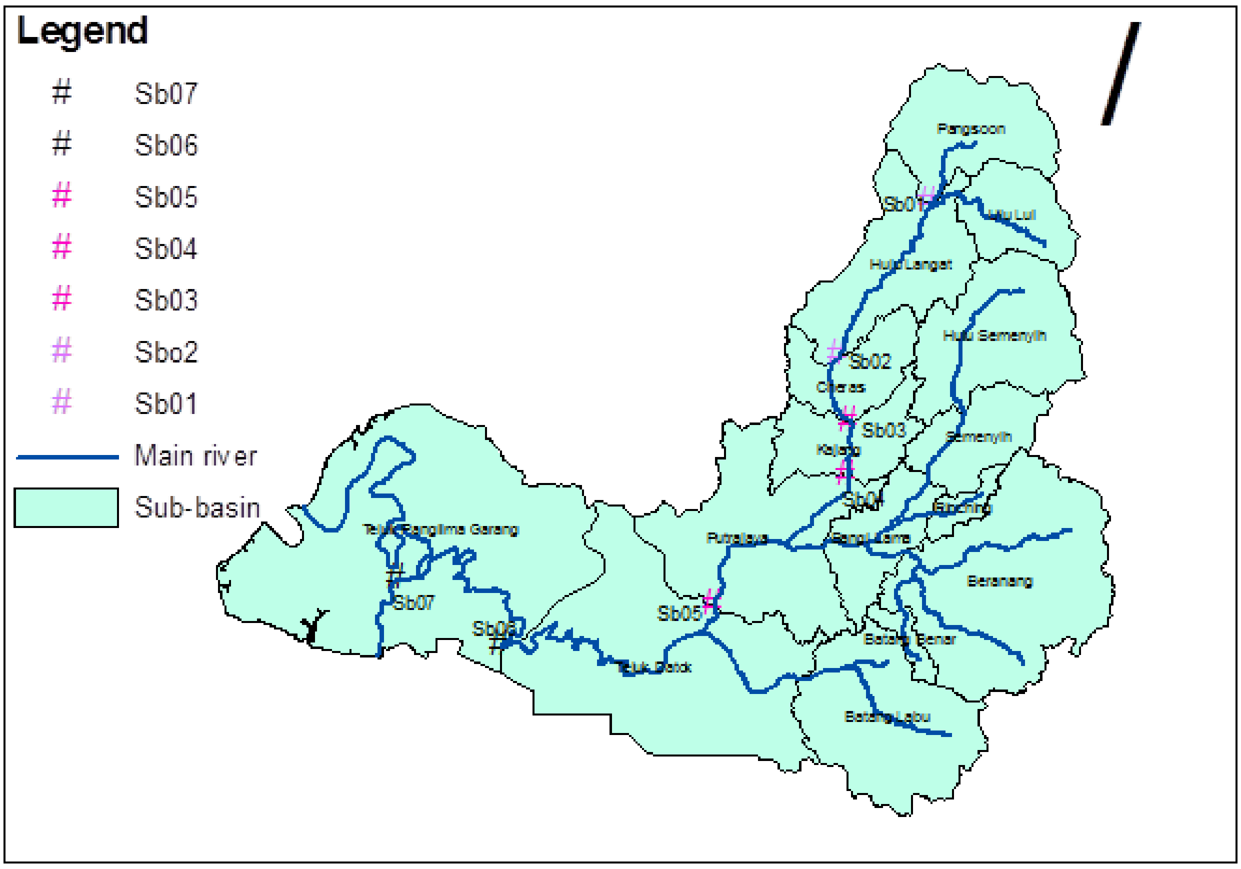


Fig. 1. Data from seven water quality stations (Sb) were selected in this study along the main river.

The water quality monitoring stations are manned by the DOE and Ministry of Natural Resource and Environment of Malaysia. The selected stations are illustrated in Table 1. The data used in the study is from September 1995 to May 2002. Seven sites were chosen, namely, Teluk Panglima Garang (site 7), Teluk Datok (site 6), Putrajaya (site 5), Kajang (site 4), Cheras (site 3), Hulu Langat (site 2), Pangsoon and Ulu Lui (site 1). Sites 3 to site 7 are located in the region of high pollution load as there are a several wastewater drains situated in the middle and downstream of the Langat River basin. Site 2 is partly situated in the middle stream region, designated as moderately polluted. Site 1 and a part of site 2 are located upstream of the Langat River, in an area of relatively low river pollution. It is worth mentioning here that some stations have missing data and not all stations were consistently sampled.

Although there are 30 water quality parameters available, only 23 completely monitored parameters were selected. A total of 254 samples were used for the analysis. The 23 water quality parameters were dissolved oxygen (DO), biological oxygen demand (BOD), electrical conductivity (EC), chemical oxygen demand (COD), ammoniacal nitrogen (AN), pH, suspended solids (SS), temperature (T), salinity (Sal), turbidity (Tur), dissolved solid (DS), total solid (TS), nitrate (NO), chlorine (Cl), phosphate (PO), zinc (Zn), calcium (Ca), iron (Fe), potassium (K), magnesium (Mg), sodium (Na), E.coli and coliform.

DOE Station No.	**Study Code**	**Distance From Estuary (km)**	**Grid Reference**	**Location**
2814602	Sb07	4.19	2°. 52.027′N 101° 26.241′E	Kampung Air Tawar (penghujung jalan)
2815603	Sb06	33.49	2° 48.952′N 101° 30.780′E	Telok Datuk, near Banting Town
2817641	Sb05	63.43	2° 51.311′N 101° 40.882′E	Bridge at Kampung Dengkil
2918606	Sb04	81.14	2° 57.835′N 101° 47.030′E	Near West Country Estate
2917642	Sb03	86.94	2° 59.533′N 101° 47.219′E	Kajang bridge
3017612	Sb02	93.38	3° 02.459′N 101° 46.387′E	Junction to Serdang, Cheras at Batu 11
3118647	Sb01	113.99	3° 09.953′N 101° 50.926′E	Bridge at Batu 18

Table 1. DOE sampling station at study area.

2.2 Principal component analysis

In this work, PCA was performed on the above mentioned water quality parameters to rank their relative significance and to describe their interrelation patterns. Chosen PC scores of the 23 water quality parameters were used as input variables in ANN model to predict the WQI. The principal components (PCs) can be expressed as

$$z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + ... + a_{im}x_{mj} \tag{1}$$

Where *z* is the component score, *a* is the component loading, *x* the measured value of variable, *i* is the component number, *j* is the sample number and *m* is the total number of variables.

The PCs generated by PCA are sometimes not readily interpreted; therefore, it is advisable to rotate the PCs by varimax rotation. Varimax rotation ensures that each variable is maximally correlated with only one PC and a near zero association with the other components (Abdul-Wahab et al., 2005; Sousa *et al.*, 2007). Varimax rotations applied on the PCs with eigenvalues more than 1 are considered significant (Kim and Mueller, 1987) where the typical criteria are 75-95% of total variance (Chen and Mynett, 2003). The rotations were carried out, in order to obtain new groups of variables. Variables with communality greater than 0.7 are considered, having significant factor loadings (Stevens, 1986).

2.3 Artificial neural networks for WQI prediction

In this work, the back propagation (BP) ANN was used in the development of all the prediction models. The Activation Transfer Function of a back-propagation network is usually a differentiable Sigmoid (S-shape) function, which helps to apply the non-linear mapping from inputs to outputs. A three layer back-propagation ANN is used in this study. The number of input and output neurons is determined by the nature of the problem under study. In this study, the networks were trained, tested and validated with one hidden layer and 1 to 10 hidden neurons. This choice was based on the work of Jiang et al. (2004), who found that the results with one hidden layer was better than that of two hidden layers, and the best performance was obtained using a structure with 3 to 6 neurons in the hidden layer. The output neuron (layer) gives the predicted WQI value.

Two different types of ANN models were developed. In the first type, prediction was performed based on the original PCs. In the second type of ANNs developed, scores of rotated (varimax rotation) PCs (ANN-RPCs) with eigenvalues greater than 1 were selected as input. For this model, prediction of WQI was performed using two to six rotated principal components separately.

The original PCs and rotated PCs (RPCs) data sets consist of 305 observations (305 rows) and are divided into training, testing and validating phases for WQI prediction. The ANN predicted WQI values are compared to the WQI values calculated using the DOE-WQI formula which is based on 6 water quality parameters, namely the DO, COD, BOD, AN, SS and pH (DOE, 1997). The input data matrix consists of 23 water quality variables (column) and 305 observations (rows) [23×305]. The observed data for each station is arranged according to time of observation from September 13, 1995 to June 7, 2002. Table 2 describes the data structure. The validation data is at least 10% of the whole data set, with 75% training set and 25% testing set data (Kuo et al., 2007).

No. of Observations	**Input parameters**								**Output**
	Input$_1$	Input$_2$	Input$_3$	.	.	.	.	Input$_{23}$	Output$_1$
1	Obs$_{1,1}$	Obs$_{1,2}$	Obs$_{1,3}$	.	.	.	.	Obs$_{1,23}$	O$_{1,1}$
2	Obs$_{2,1}$	Obs$_{2,2}$	Obs$_{2,3}$	.	.	.	.	Obs$_{2,23}$	O$_{2,1}$
. . . …	. . . ….	. . . ….	. . . …	. . . ….	. . . ….	. . . ….	. . . ….	. . . …	. . . . …
305	*Obs$_{305,1}$*	*Obs$_{305,2}$*	*1*					*Obs$_{305,23}$*	*O$_{305,1}$*

Table 2. The data structure for ANN prediction model.

2.4 Determination of model performance

The model's behaviour in both learning (training and testing) and validating phase, is evaluated using the following statistical methods; the correlation coefficient (R) at 95% confidence limit, given by equations;

$$\text{Coefficient of correlation (R), } r = \frac{\left[\sum_{i=1}^{n} x_i \hat{x}_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum \hat{x}_i\right)\right]^2}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum \hat{x}_i - \frac{1}{n}\left(\sum \hat{x}_i\right)^2\right]}} \qquad (2)$$

and the mean bias error or residual error given by;

$$\text{Mean bias error (MBE), } MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i) \qquad (3)$$

Where $\hat{x}_i$ and x_i represent observed values and the corresponding forecast values for i =1,2,…..,n.

The prediction performance evaluated using these two methods are used to evaluate the accuracy of the forecast and for comparing the forecasting ability of each approach.

The 95% confidence limit is used to determine that the predicted output lie within the confidence range. It is assumed that a predicted value fall into an interval within which there is an associated uncertainty. According to Wackerly et al. (1996), this uncertainty is derived from the residual errors that have already been calculated within that range of values. If the residual errors are randomly distributed, there is a general rule of thumb which states that they will lie within two standard deviations of their mean with a probability of 0.95. This method was used in the measurements of the ANN prediction performance conducted by some researchers (Bishop, 1995; Tibshirani, 1996; Shao et al., 1997; Zhang et al., 1998; Lowe and Zapart, 1999; Townsend and Tarassenko, 1999)

ANN models and statistical analyses were carried out using MATLAB 7.0 and XLSTAT2008 (Excel2003 add-in) for Windows.

3. Results and discussion

Post PCA, out of the 23 principal components generated, only six PCs with eigenvalues higher than 1 (Table 3) were selected for the ANN input parameters. Selected PCs explained 75.1% of the total variation. Furthermore, communality values were high for the selected PCs, for example, the values are 93% for Cond., 95% for Sal, 98% for DS and TS (Table 4). These results further confirm the choice of the selected number of PCs (Stevens, 1986).

For the first six rotated PCs (RPCs), the loadings from PCA are given in Table 4. The highest correlations between variables are noted in bold. For instance, Cond., Sal, DS, TS, Cl, Ca, K, Mg and Na, have high correlations with RPC1. Eighteen variables with strong loadings were included in the six selected RPCs. Significant variables in RPC1 are Cond., Sal., DS, TS, Cl, Ca, K, Mg, and Na; in RPC2 they are DO, BOD and AN; in RPC3 they are SS and Tur and in RPC4, NO_3^- and PO_4^{3-}. The only meaningful loads in RPC5 and RPC6 are pH and Zn.

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	9.074	2.387	2.067	1.492	1.225	1.026
Variability (%)	39.451	10.380	8.987	6.488	5.326	4.459
Cumulative %	39.451	49.830	58.817	65.305	70.631	75.091

Table 3. Descriptive statistics of selected original PCs with eigenvalues more than 1.

Variables	RPC1	RPC2	RPC3	RPC4	RPC5	RPC6	Communalities
DO	-0.205	**-0.722**	-0.121	0.046	0.485	-0.066	0.82
BOD	0.035	**0.740**	0.071	0.110	0.110	0.022	0.58
COD	0.340	0.103	0.081	-0.166	0.268	0.326	0.34
SS	-0.042	-0.009	**0.920**	0.010	-0.025	0.017	0.85
pH	0.189	-0.109	-0.204	0.020	**0.792**	-0.083	0.72
AN	-0.092	**0.797**	-0.151	0.161	0.023	-0.032	0.69
T	0.337	0.368	-0.242	-0.298	-0.317	0.208	0.54
Cond.	**0.963**	0.022	-0.043	0.035	0.013	-0.022	0.93
Sal.	**0.974**	0.023	-0.038	0.030	0.008	-0.004	0.95
Tur.	-0.031	-0.007	**0.863**	0.011	-0.140	-0.035	0.77
DS	**0.988**	0.017	-0.034	0.013	0.009	-0.005	0.98
TS	**0.985**	0.017	0.069	0.014	0.007	-0.003	0.98
NO_3^-	0.018	0.033	0.107	**0.688**	-0.126	0.300	0.59
Cl	**0.986**	0.010	-0.029	-0.004	0.020	0.005	0.97
PO_4^{3-}	0.023	0.312	-0.106	**0.700**	0.112	-0.073	0.62
Zn	-0.019	0.044	-0.011	0.186	-0.128	**0.767**	0.64
Ca	**0.980**	0.028	-0.026	-0.043	-0.024	0.039	0.97
Fe	-0.080	0.043	0.475	0.540	0.066	0.192	0.57
K	**0.984**	0.004	-0.031	-0.004	0.004	0.010	0.97
Mg	**0.974**	0.000	-0.022	-0.028	-0.002	0.037	0.95
Na	**0.986**	0.002	-0.025	-0.020	0.005	0.017	0.97
COLI	-0.254	0.361	0.097	-0.424	0.457	0.056	0.60
COLIFORM	-0.032	0.049	-0.025	0.042	-0.077	-0.517	0.28

Table 4. Rotated factor loadings using six PCs.

Using the original principal component scores as inputs, the best architecture consist of a three layer network with 23 input neurons, 10 neurons in the hidden layer and one neuron in the output layer. Considering RPC scores as inputs, the best architectures were achieved with almost the same number of hidden neurons. The hidden neurons consist of 9 and 10 neurons respectively. Training was carried out for a maximum 10000 iterations. Selection of the network was performed at maximum correlation coefficient (R) and 95% confidence limit.

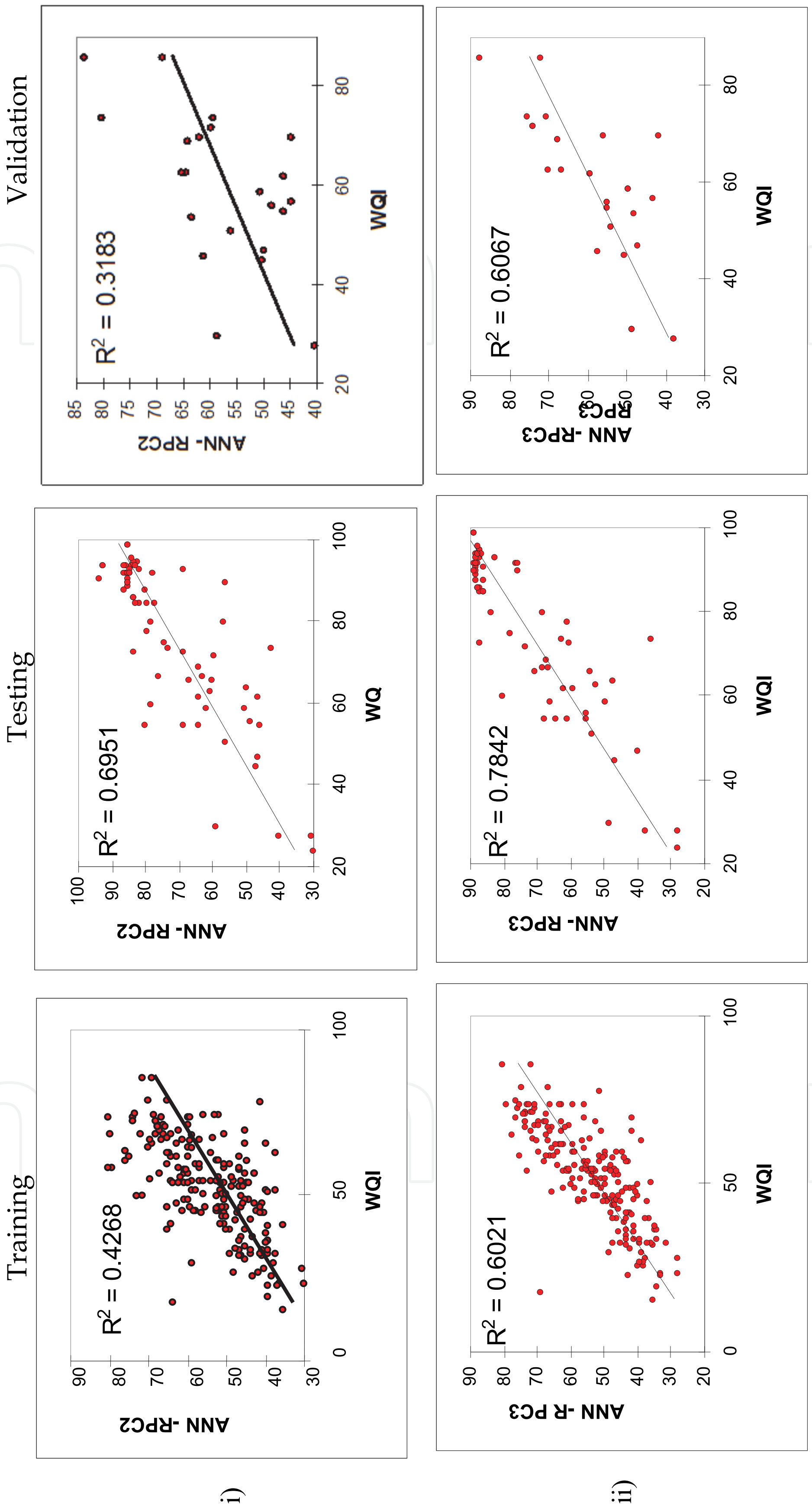
Table 5 and Figure 2 illustrate the prediction performances of ANN models using different combinations of PC scores as input variables. ANN using the first 2 PCs (PC1 and PC2) does not perform very well as far as accuracy is concerned for all the training, testing and validation phases. It is observed that the prediction performance of the validation phase is slightly worse compared to the training and testing phases. It is important to point out that for this model, the cumulative percentage in explaining the variance given by these two RPCs is only 49.8%. None of the strong loading variables contains the variables forming the WQI equation. DO, BOD and pH loadings in PC2 explain only 10.4% of the total variance.

Based on the results, it is apparent that the WQI prediction performance increases with the increase in number of input variables. The highest accuracy in predicting WQI is given by model ANN-RPC6, which contains six RPCs with 75.1% variation explained, giving an R^2 value of 0.64 (training), 0.87 (testing), and 0.72 (validation) respectively.

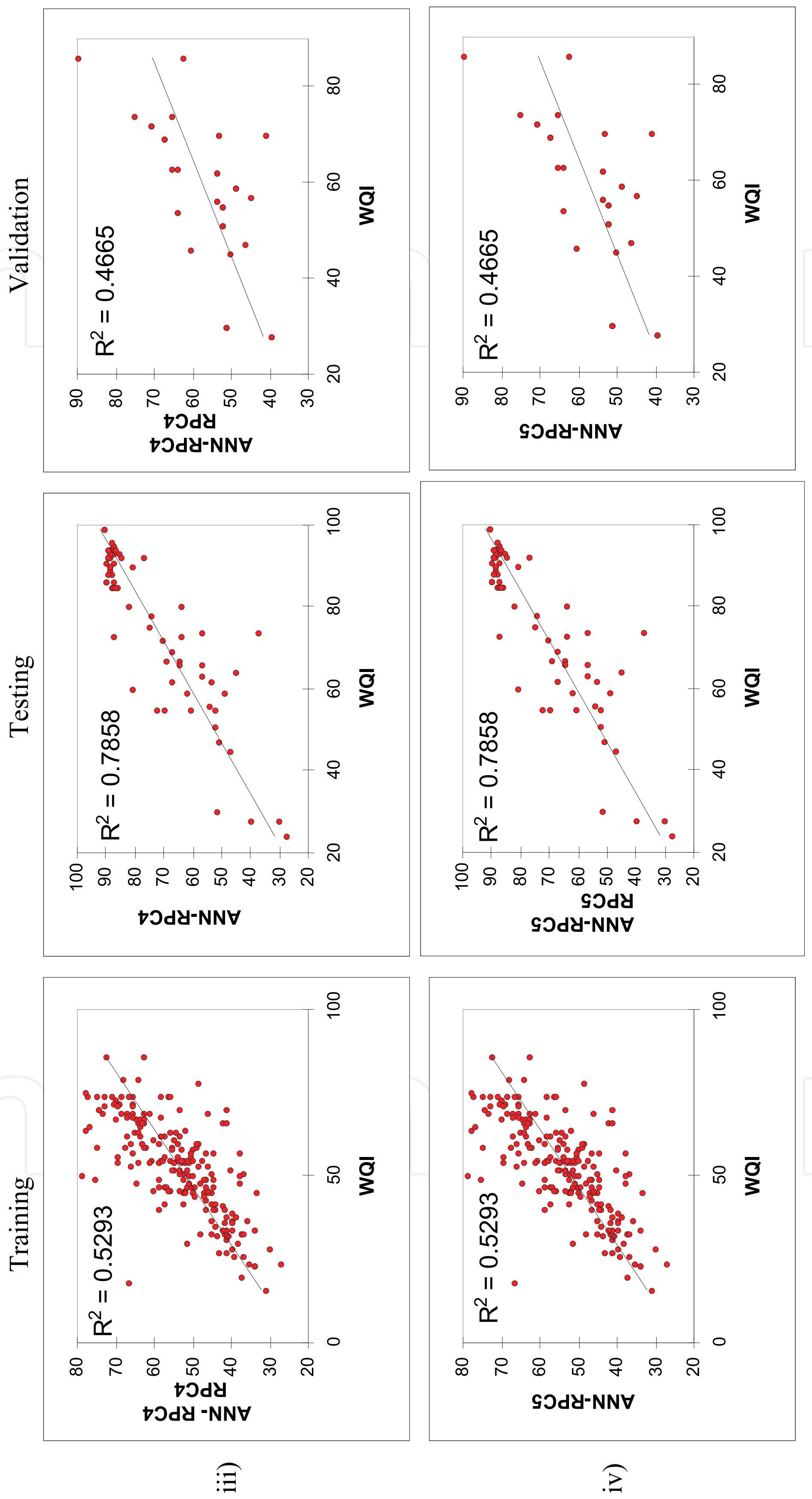
Model	No.of PC	R squared			MBE		
		Training	Testing	Validation	Training	Testing	Validation
ANN-RPC2 (2 inputs)	2	0.43	0.70	0.32	28.01	-167.90	-40.71
ANN-RPC3 (3 inputs)	3	0.60	0.78	0.61	64.95	-109.68	6.60
ANN-RPC4 (4 inputs)	4	0.53	0.79	0.47	0	-165.04	-89.78
ANN-RPC5 (5 inputs)	5	0.53	0.79	0.47	140.12	-143.75	-44.77
ANN-RPC6 (6 inputs)	6	0.64	0.87	0.72	67.93	-58.57	-44.61
ANN-PC23 (23 original PC inputs)	23	0.60	0.85	0.66	-18	-81.59	-49.83

Table 5. The prediction performances of the different ANN models.

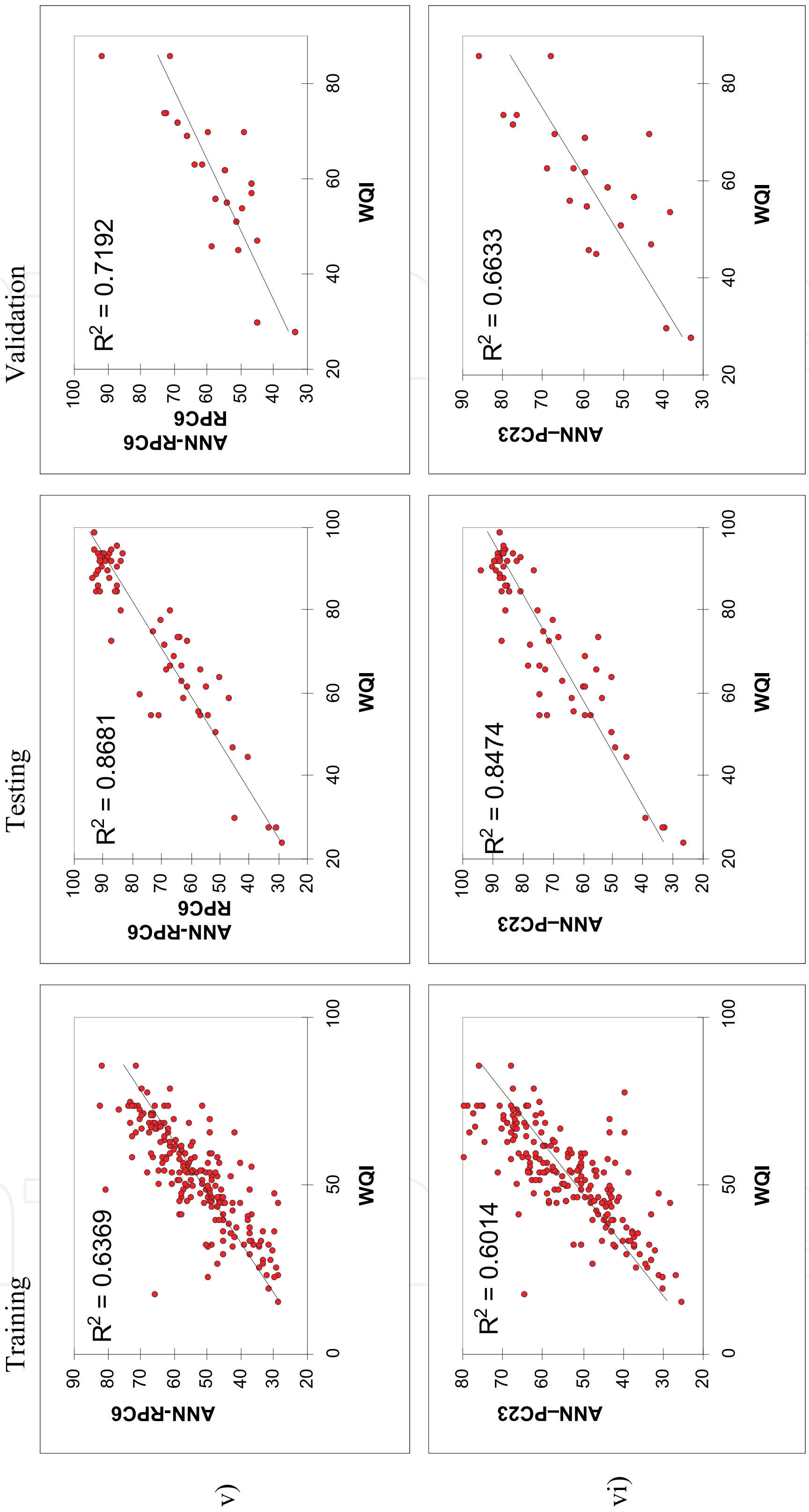
From table 5, it can be observed that the prediction performance of the ANN model using original PCs (23 input PC scores) is not significantly different from the RPC models. However, as RPC models use fewer variables and is far less complex, the advantage over the ANN-PC23 model is obvious. Comparing the MBE values, it is generally observed that the signs for the validation phases are negative for both the un-rotated and rotated PC models. This is an indication that the predicted WQI values are consistently underestimated in both approaches.


Training
Testing
Validation
R² = 0.4268
ANN -RPC2
WQI
R² = 0.6951
ANN- RPC2
WQ
R² = 0.3183
ANN- RPC2
WQI
R² = 0.6021
ANN -R PC3
WQI
R² = 0.7842
ANN- RPC3
WQI
R² = 0.6067
ANN -RPC3
WQI
i)
ii)

Part I


Training
Testing
Validation
iii)
iv)
R2 = 0.5293
R2 = 0.7858
R2 = 0.4665
ANN- RPC4
ANN-RPC4
ANN-RPC5
WQI
0
50
100
20
40
60
80
90
30
70

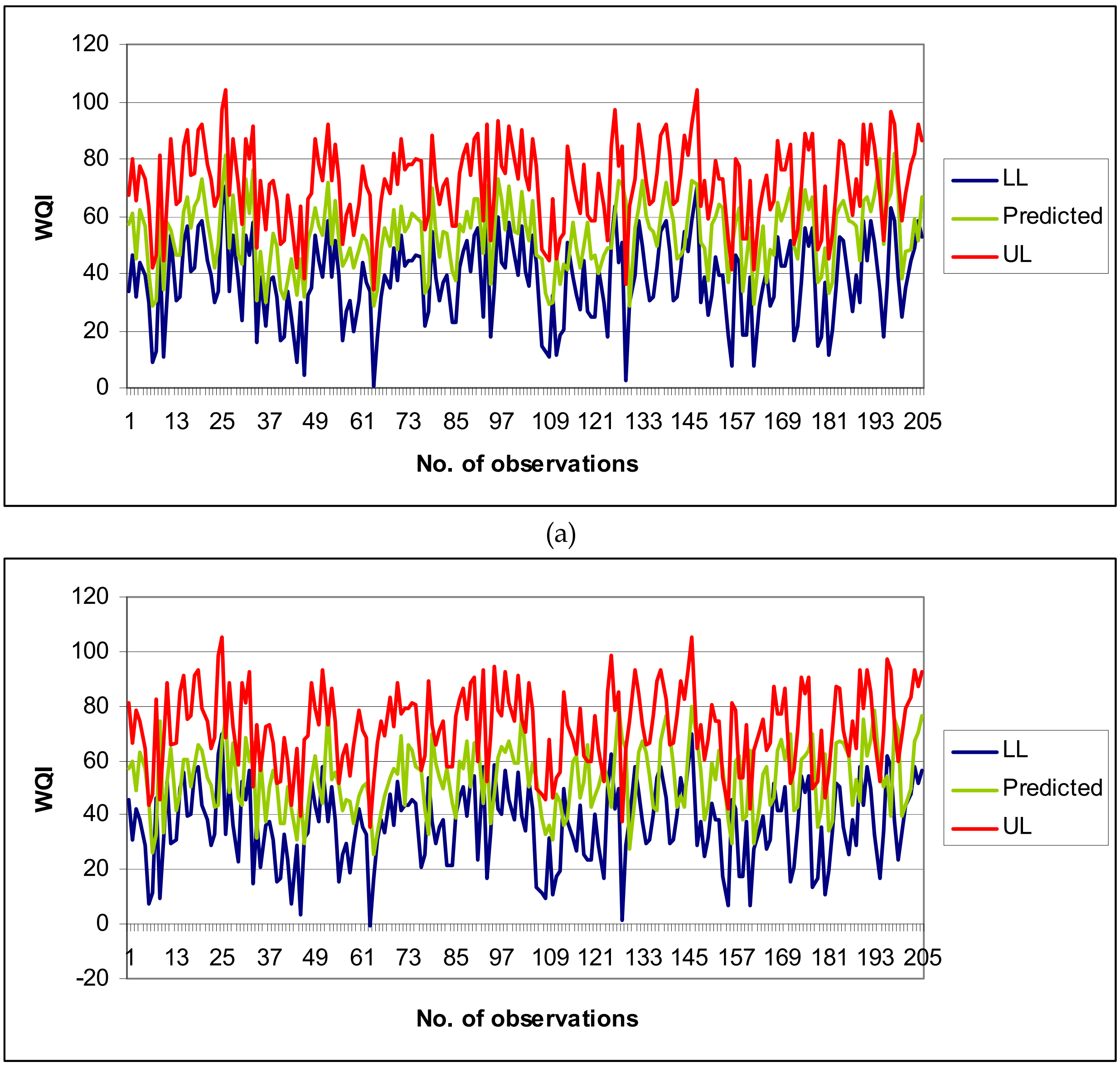
Part II



Part III

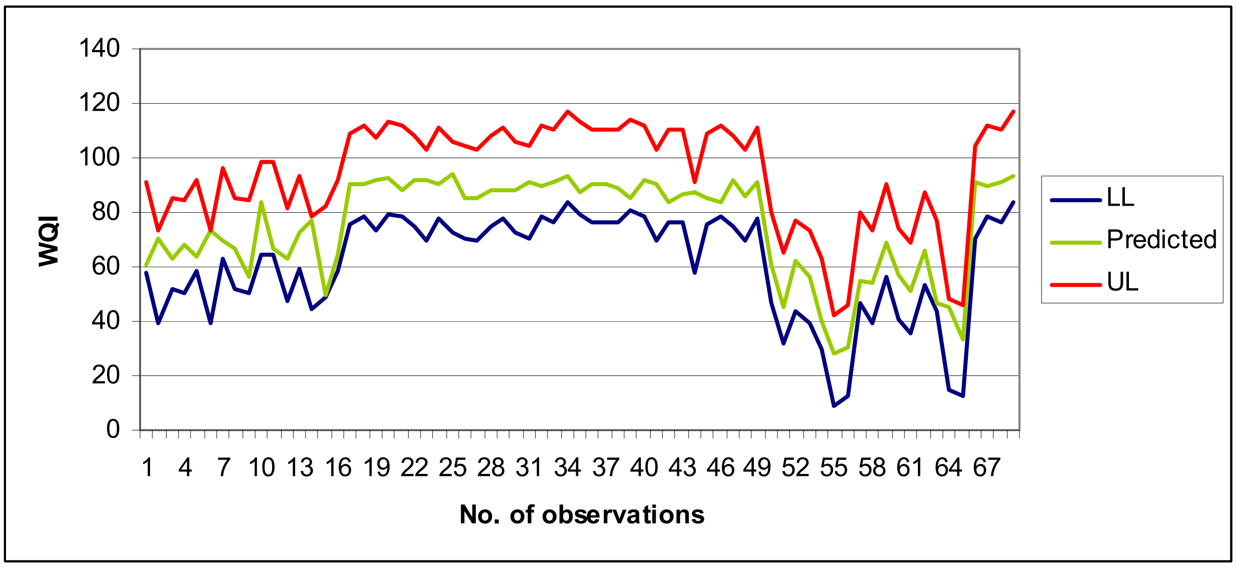
Fig. 2. The prediction performances for different combination of PC scores during training, testing and validation phases : (i) 2 RPCs, (ii) 3 RPCs, (iii) 4 RPCs, (iv) 5 RPCs, (v) 6 RPCs and, (vi) 23 original PCs.

This study also attempts to allocate 95% confidence interval on the WQI prediction produced by the best ANN model. Figure 3, 4 and 5 show the comparison between predicted values and the upper (UL) and lower limits (LL) lying within 95% confidence interval. This was carried out for ANN-RPC6 and ANN-PC23 models. It can be seen that only 4.3% out of the 305 predicted values were identified beyond the 95% confidence limit (1% fall below the LL and 3.3% fall beyond the UL) for ANN-RPC6. For ANN-PC23, 25% of the 305 observations fall beyond the upper and lower of 95% confidence interval limit (14% fall below the LL and 11.8% fall beyond the UL). This basically shows that by using reduced rotated PC scores as input, better results can be obtained without losing information. It is thus apparent that ANN prediction using scores of varimax rotated PCs result in a more accurate WQI prediction.

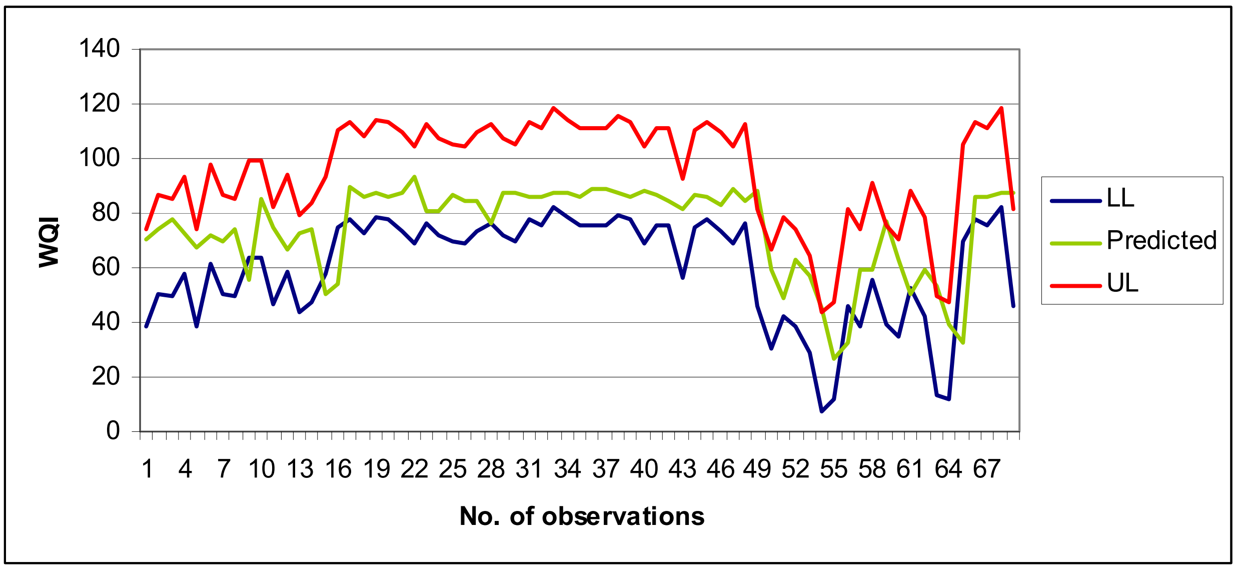


(b)

Fig. 3. Predicted WQI within the 95% confidence interval during training phase using (a) six rotated PCs, and (b) 23 original PCs.

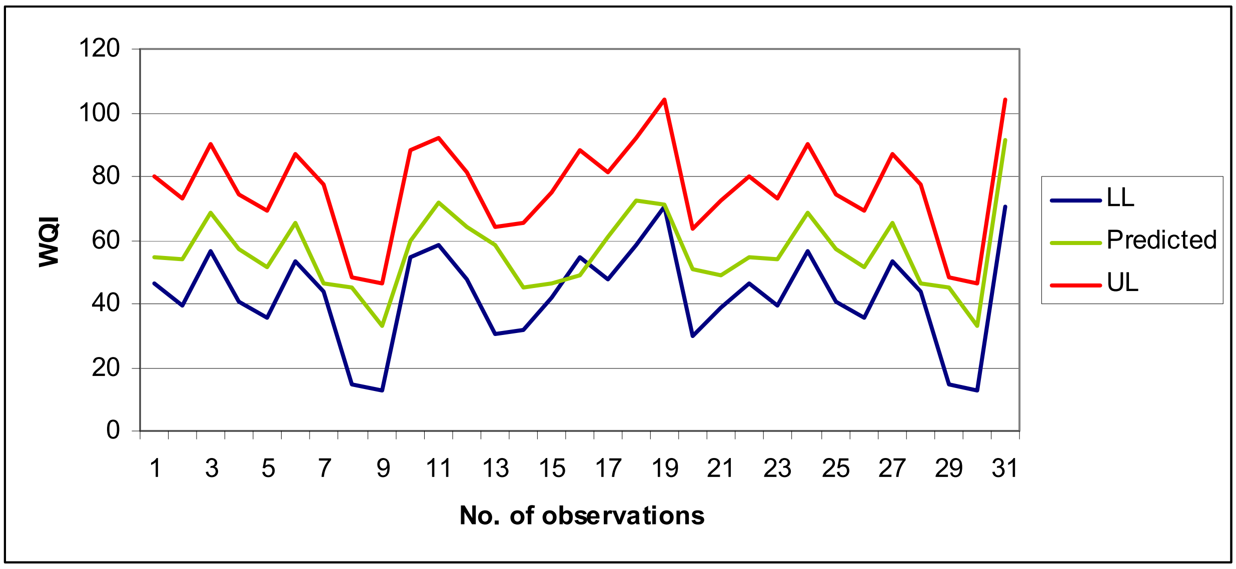


(a)

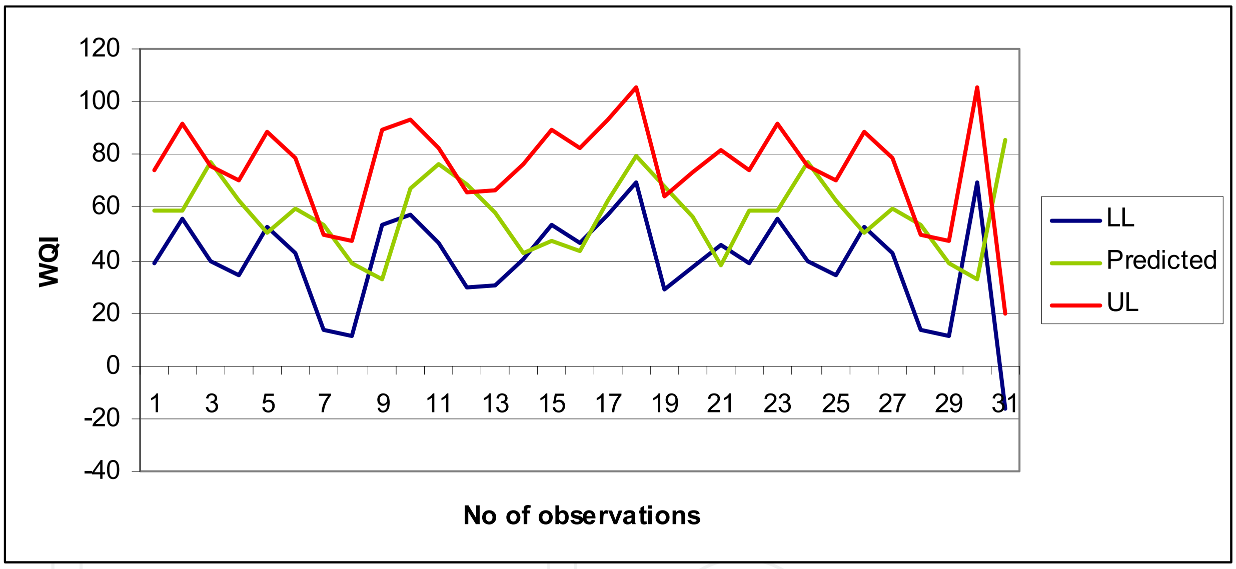


(b)

Fig. 4. Predicted WQI within the 95% confidence interval during testing phase using (a) six rotated PCs, and (b) 23 original PCs.



(a)



(b)

Fig. 5. Predicted WQI within the 95% confidence interval during validation phase using (a) six rotated PCs, and (b) 23 original PCs.

4. Conclusion

In this work, a combination of PCA and ANN is used to predict WQI based on 23 historical water quality parameters. The original predictors were selected based on the available Malaysian DOE data. To obtain the latent variables as inputs into the ANN, two different approaches were used; one based on un-rotated original PCs and the other based on varimax rotated PCs.

Using six PCs, significant loadings are observed for Cond, Sal, DS, TS, Cl, Ca, K, Mg and Na in PC1, DO, BOD and AN in PC2, SS and Tur in PC3, NO3- and PO43- in PC4, pH in PC5 and Zn in PC6. ANN models based on these 6 PC scores can predict WQI with acceptable accuracy (within 95% confidence limit). Moreover, the ANN model using the 23 original PCs as input, do not render the prediction more accurate, even with a complex network structure. The use of rotated PC scores based models is clearly more effective and efficient due to the elimination of collinearity and reduction of predictor variables without losing important information.

5. Acknowledgment

The authors acknowledge the financial and technical support for this project provided by the Ministry of Science, Technology and Innovation and Universiti Putra Malaysia under the Science Fund Project no. 01-01-04-SF0733. The authors wish to thank,the Department of Environment, and Department of Irrigation and Drainage, Ministry of Natural Resources and Environment of Malaysia, Institute for Development and Environment (LESTARI), Universiti Kebangsaan Malaysia, Universiti Malaya Consultancy Unit (UPUM) and Chemistry Department of Universiti Malaya, who have provided us with secondary data and valuable advice.

6. References

Abdul-Wahab, S.A., Bakheit, C.S. and Al-Alawi, S.M., 2005. Principal component and multiple regression analysis in modeling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software* 20, p.1263-1271.

Aiken, R.S., Leigh, C.H., Leinbach, T.R., and Moss, M.R., 1982. Development and Environment in Peninsular Malaysia. McGraw-Hill International Book Company: Singapore.

Beck, M.B., 1986. Identification, estimation and control of biological waste-water treatment processes. *IEE Proceeding* 133, p.254-264

Benediktsson, J.A., Swain, P.H., and Ersoy, O.K., 1990. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *I.E.E.E. Transactions on Geoscience and Remote Sensing*, 28, 540-551

Bishop, C. M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford Bowden, G.J., Dandy, G.C. and Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1-background and methodology. *Journal of Hydrology* 301, p.75-92.

Chen, Q. and Mynett, A.E., 2003. Integration of data mining techniques and heuristic knowledge in fuzzy logic modeling of eutrophication in Taihu Lake. *Ecological Modelling* 162, p.55-67.

Clarici, E., 1995. Environmental Modelling Using Neural Networks, PhD Thesis, Imperial College.

Department of Environment Malaysia, DOE 1997. Malaysia environmental quality reports 1999. Kuala Lumpur: Ministry of Science, Technology and Environment.

Department of Environment Malaysia, DOE 1999. Malaysia environmental quality reports 1999. Kuala Lumpur: Ministry of Science, Technology and Environment.

DeSilets, L., Golden, B., Wang, Q., and Kumar, R., 1992. Predicting salinity in the Chesapeake Bay using backpropagation. *Computer and Operations Research*, 19, p.227-285

Elsner, J.B., and Tronis, A.A., 1992. Nonlinear prediction, chaos, and noise. Bull. *Am. Meterol. Soc.* 73(1), p.303-314.

Foody, G.M., McCulloch, M.B., and Yates, W.B., 1995. Classification of remotely sensed data by an artificial neural network: issues related to training data characteristics. *Photogrammetric Engineering and Remote Sensing.*

Ha, H. and Stenstrom, M. K., 2003. Identification of land use with water quality data in stormwater using a neural network. *Water Research*, 37, p.4222-4230

Hafizan Juahir, Sharifuddin M. Zain, Zainol Mustafa, and Azme Khamis, 2001. Dissolved oxygen forecasting due to landuse activities using time series analysis at Sungai Lang at, Hulu Lang at, Selangor. *Ecological Environmental Modelling, Proceeding of the National Workshop*, Universiti Sains Malaysia, 3-4 September, p.157-164.

Hafizan Juahir, Sharifuddin M. Zain, Mohd. Ekhwan Toriman, M. Nazari Jaafar and W. Klaewtanong, 2003a. Performance of autoregressive integrated moving average and neural network approaches for forecasting dissolved oxygen at Langat River Malaysia. *Urban Ecosystem Studies In Malaysia: A study of change.* Universal Publishers, p. 145-165.

Hafizan Juahir, Sharifuddin M. Zain, M. Nazari Jaafar, Zainal Majid and M. Ekhwan Toriman, 2003b. Land use temporal changes: application of GIS and statistical analysis on the impact of water quality at Langat River Basin, Malaysia. presented *in 2nd Annual Asian Conference of Map Asia 2003*, 17-19, Oct., PWTC Kuala Lumpur.

Hafizan Juahir, Sharifuddin M. Zain, M. Nazari Jaafar and Zainal Ahmad, 2004a. An Application of Second order backpropagation method in Modeling River Discharge at Sungai Langat, Malaysia. *Water Environmental Planning: Towards integrated planning and management of water resources for environmental risks*, IIUM, p.300-307.

Hafizan Juahir, Sharifuddin M. Zain, M. Ekhwan Toriman and Mazlin Mokhtar, 2004b. Application of Artificial Neural Network Model In the Predicting Water Quality Index. *Jurnal Kejuruteraan Awam*, 16 (2), p.42-55

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L., 2000. Temporal evaluation of grounwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Water Research* 34, 807-816.

Hepner, G.F., Logan, T., Ritter, N., and Bryant, N., 1990. Artificial Neural Network classification using a minimal training set: comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing*, 56, 469-473

Jain, A. & 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, p. 93-98.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J. And Shao, D., 2004. Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment* 38, p.7055-7064.

Kim, J.,-O., and Mueller, C.W., 1987. Introduction to factor analysis: what it is and how to do it. *Quantitative Applications in the Social Sciences Series.* Sage University Press, Newbury Park.

Kuo, J.-T., Hsieh, M.-H., Lung, W.-S. And She, N., 2007. Using artificial neural network for reservoir eutrophication prediction. *Ecological Modelling*, 200, p.171-177

Loska, K. and Wiechula, D., 2003. Application of principal component analysis for the estimation of source of heavy metal contamination in surface sediments from the Rybnik Reservoir. *Chemosphere* 51, p.723-733.

Lowe, D. and Zapart, C., 1999. Point-Wise Confidence Interval Estimation by Neural Networks: A Comparative Study based on Automotive Engine Calibration. *Neural Computing & Applications*, Vol. 8, p.77-85.

Mohd. Ekhwan Toriman and Hafizan Juahir, 2003. Artificial Neural Network Modelling For Langat River Discharge: Implication For River Restoration. Pertandingan Minggu Penyelidikan dan Inovasi UKM, Pusat Pengurusan Penyelidikan, 3-5 Julai.

Muttil, N. and Chau, K.,-W., 2007. Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Application of Artificial Intelligence* 20, p.735-744.

Peddle, D.R., Foody, G.M., Zhang, A., Franklin, S.E., and Ledrew, E.F., 1994.

Multisource image classification II: An empirical comparison of evidential reasoning and neural network approaches. *Canadian Journal of Remote Sensing*, 12, 277-302.

Recknagel, F., Bobbin, J., Whigham, P., and Wilson, H., 2002. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modeling of algal blooms in freshwater lakes. *Journal of Hydroinformatics* 4(2), p.125-134.

Rosnani Ibrahim, 2001. River Water quality Status In Malaysia. *Proceedings National Conference On Sustainable River Basin Management* In Malaysia, 13 & 14 November 2000, Kuala Lumpur, Malaysia.

Rumelhart, D. E., Hinton, E. and Williams, J., 1986. Learning internal representation by error propagation. *Parallel Distributed Processing*, 1, p. 318-362.

Ruslan Rainis, Kamarul Ismail and Hafizan Juahir, 2004. Modeling The Relationship Between River Water Quality Index (WQI) and Land Uses Using Artificial Neural Networks (ANN). Presented in JSPS Seminar, December 15-17, Kyoto, Japan.

Scardi, M., 2001. Advances in neural network modeling of phytoplankton primary production. *Ecological Modelling* 146, p.33-45.

Schalkoff, R., 1992. Pattern Recognition: Statistical, Structural and Neural Approaches. New York, Wiley

Shao, R., Martin, E.B., Zhang, J. and Morris, A.J., 1997. Confidence bounds for neural network representations. *Computers and Chemical Engineering*, 21, p.1173-1178.

Silverman, D., and Dracup, J.A., 2000. Artificial neural networks and long-range precipitation in California. *Journal of Applied Meteorology* 31(1), p.57-66.

Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M. and Kouimtzis, Th., 2003. Assessment of the surface water quality in Northern Greece. *Water Research* 37, p.4119-4124.

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M. and Pereira, M.C., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software* 22, p.97-103.

Stevens, J., 1986. Applied Multivariate Statistics for the Social Science. Hill Sdale: New Jersey, USA, p.515.

Tibshirani, R., 1999. A comparison of some error estimates for neural network models. *Neural Computation*, 8, p.152-163.

Townsend, N.W. and Tarassenko, L., 1999. Estimations of error bounds for neural-network function approximators. *IEEE Trans Neural Netwoks*, 10(2), p.217.

Wackerly, D.D., Mendenhall, W and Scheaffer, R.L., 1996. Mathematical Statistics with Applications. 5th. Ed., Duxbury Press: Belmont, USA.

Wunderlin, D.A., Diaz, M.P., Ame, M.V., Pesce, S.F., Hued, A.C., Bistoni, M.A., 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba, Argentina). *Water Research* 35, 2881-2894.

Zarita Zainuddin, 2004. Modelling Nonlinear Relationship in Ecology and Biology using Neural Networks. In Koh Hock Lye and Yahya Abu Hassan, Ecological Environmental Modelling (ECOMOD 2001): *Proceedings of the National Workshop,* 3-4 September, USM, p.88-95

Zhang, J., Morris, A.J., Martin, A.J. and Kiparissides, C., 1998. Prediction of polymer quality in batch polymerization using robust neural networks. *Chemical Engineering Journal*. 69, p.135-143.

Zhang, Q. & Stanley, S. J.,1997. Forecasting raw-water quality parameters for North Saskatchewan River by neural network modeling. *Water Resource*, 31, p. 2340-2350.