Research Journal of Earth Sciences 1(1): 15-21, 2009
ISSN 1995-9044

Predicting for Discharge Characteristics in Langat River, Malaysia Using Neural Network Application Model

[1]M. Ekhwan Toriman, [1]Hafizan Juahir, [2]Mazlin Mokhtar, [3]Muhammad Barzani Gazim, [4]Sharifah Mastura Syed Abdullah and [5]Othman Jaafar

[1,4]School of Social, Development and Environmental Studies, National University of Malaysia
[2] Institute of Environment and Development (LESTARI), National University of Malaysia
[3]School of Environmental Sciences and Natural Resources, National University of Malaysia
[5]Department of Civil Engineering, National University of Malaysia

Abstract: Artificial neural networks have been shown to be able to approximate any continuous non-linear functions and have been used to build data base empirical models for non-linear processes. In this study the model are applied to estimate the daily water discharge with two input variables (from two rainfall stations in the same catchments area at Langat River, Malaysia. These model are highly non-linear, therefore two possible ways of analysis is carry out analysis and using the training and testing of sum square error (SSE) we can select the most significant model suit to the case study.

Key words: Artificial Neural networks % non-linear function % river discharge % Langat River

INTRODUCTION

Artificial neural networks have been used in developing non-linear models in industry for such a long time and robustness of the model is one of the main criteria that need to be considered. Robustness of the model can be defined as one of the baseline to judge the performance of the neural network models and it is really related to the learning or training classes [1]. Even though neural networks have a significant capability in predicting a non-linear function, inconsistency of accuracy still seem became a problem where neural networks model seems cannot cope or performed well when it is applied to a new unseen data. Furthermore in this day, that advance process control and supervision of industrial processes require an accurate prediction of the process models and at the same time promoted an interest in robustness of neural networks models. Lack of robustness in neural network models is basically due to the overfitting and poor generalisation of the models (e.g. [2]. Therefore, a lots of researchers was interested and concentrate on how overfitting can be avoided by improved the learning algorithm or by combining the neural networks (e.g. [3-6]. In view of the robustness of neural network a lot of techniques have been developed like regularisation and the early stopping method (e.g. [7-9]. Reference [10] implemented the universal learning rule and second order derivatives to increased the robustness in neural network models.

Case Studies

Study Area: The Langat River is one of the most important river of Malaysia. Many town having very large population like Cheras, Kajang, Sepang and Dengkil are situated along the river bank. In this study the NN model has been applied as a model of short-term water discharge prediction allocation at the selected water quality monitoring station by Department of Irrigation and Drainage in the Langat River, at Selangor, Malaysia. The Langat River Basin occupies the south and south-eastern parts of the state of Selangor Darul Ehsan. It is about 78 km long and ranges from 20 km to 51.5 km wide. It has a total catchment 1,987.8 km^2. The source of the Langat River is at the Pahang-Selangor border where hilly terrain reaching up to 1,500 m above mean sea level can be found. It finally drains into the Melaka Strait on the mangrove coastline of southwestern Selangor. The major tributaries of the Langat River are Semenyih River, the labu River and the Mantin River. The general flow of the Langat river is north and north-east towards the south and south-west in the eastern half of the basin and westward on the western part.

Corresponding Address: Dr. Mohd Ekhwan Toriman, School od Social, Development and Environmental Studies, FSSK. University Kebangsaan Malaysia, 43600 Bangi Selangor Malaysia

Table 1: Summary of Major Land Use Types in the Langat River Basin

Major Land Use Type	Area (km^2)	Percentage of Land Use
Agriculture	1,335.57	55.13
Forest	467.80	19.31
Wetland and Swamps	308.36	12.73
Urban and Built-up Areas	150.12	6.20
Mining	38.91	1.61
Others	121.79	5.02
TOTAL	2,422.55	100.00%

In general, agriculture and forest are the dominant types of land use in the Langat River Basin. The classification of land use types (in Sq. km.) in the Langat River Basin is shown in Table 1.0. Agriculture is the main land use type (55.13%), followed by forest (19.31%) and wetland (12.73%). Urban and built-up areas only occupy 6.20% of the total land use. Mining (1.61%) is a relatively minor land use type.

Two dams are located in the study area. The Langat Dam in the upstream of the Langat River has an active storage capacity of 30 million m^3. Semenyih Dam located in the Semenyih River has an active reservoirs and release water whenever water level of the Langat River is low.

Eight water intakes (water treatment plant) are located in the study area and produce more than 200 MGD (million gallons per day) of treated water. The Langat plant supplies 85 MGD of treated water to areas in Cheras, Pandan and Hulu Langat, while the Cheras Mile 11 plant supplies 6 MGD of treated water to areas in Balakong, part of Cheras and Kajang. The Bukit Tampoi treatment plant supplies 6.9 MGD of treated water to areas of Dengkil. One water treatment plant is located on Semenyih River with Semenyih dam regulation flow to the Semenyih treatment plant. The output capacity of this plant is 120 MGD and supplies treated water to areas in Semenyih, Petaling Jaya South, Bukit Gasing, Shah Alam, Klang and Subang Jaya. Salak Tinggi water treatment plant is located at Salak Tinggi and draws raw water from Labu River. The operator of these plants, Puncak Niaga, constantly monitors the quality of the raw water at intake points.

Neural Network for Time Series and Non-linear Dynamic Relationship: The optimisation method used is the Levenberg-Marquardt algorithm [11, 12]. Regularisation and 'early stopping' method has been used which is consist of 100 epoch and the regularisation parameter used is 0.001. Regularisation is achieved by modifying the networks training objective in equation (1) to include a term to penalise unnecessary large network weights as follows:

$$J = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}(i) - y(i))^2 + r \parallel W \parallel^2 \quad (1)$$

where N is the number of data points, $\hat{y}$ is the networks prediction, y is the target value, W is a vector of networks weights and δ is the regularisation parameters. Therefore training will be automatically stop when either one of this criteria is achieved. All weights and biases have been selected randomly in the range of –0.1 to 0.1.

Time Series Prediction: In this case, we carried out 3 different dynamics of the model where the output is the discharge from the catchments area. The time series model can be shown as follows;

a. $\hat{y}(t) = f[y(t-1), y(t-2)]$ (2)

b. $\hat{y}(t) = f[y(t-1), y(t-2), y(t-3)]$ (3)

c. $\hat{y}(t) = f[y(t-1), y(t-2), y(t-3), y(t-4)]$ (4)

By trial and error we carry out analysis which is by varying the hidden neuron from 1 to 10 and plot the training, testing and validation Sum square error (SSE) for all time series model. By using the training and testing SSE we can select the most significant model that suit to the case study.

Non-linear Dynamic Relationship: The dynamic model structure selected for this approach is:

$$\hat{y}(t) = f[y(t-1), y(t-2), u1(t-1), u1(t-2), u2(t-1), u2(t-2)] \quad (5)$$

where $\hat{y}(t)$ is the models prediction water discharge form the catchments area at time t, $y(t\text{-}1)$ and $y(t\text{-}2)$ are the water discharge at time *(t-1)* and *(t-2)* and $u1(t)$ and $u2(t)$ are the raining rates at time t at 2 different station.

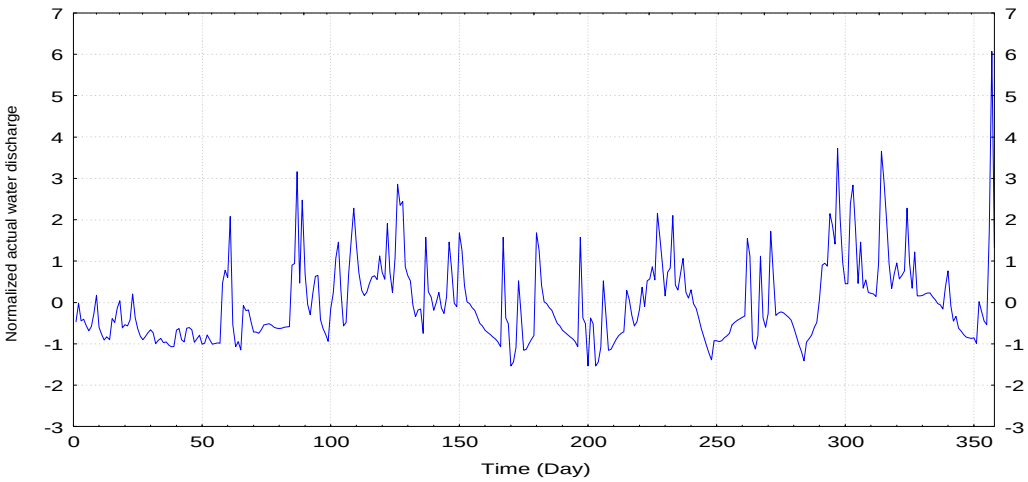

Fig. 1: Data used in parameter estimation on neural network training and validation

Training and Validation Data: To prevent any discrepancy in the unit for the input and output data, all the measurements or the analysis data will be scale down to zero mean and with the same standard deviation. In addition, the comparison on individual SSE for 20 neural networks will be made compare to SSE for combination neural networks and all the simulation and programming work is carried out using Matlab™ 6.5.

Water Catchments Area in Langat River: This area basically a reserve for water supply in area of Sg Rasah in Malaysia. There is a lot of stations that monitor the water flow to the catchments area and in this case we choose two rainfall stations which the number is 2917001(R_1) and 3018101 (R_2) and one water discharge station, the number is 2917401 (Q) that are really related or proportional to the water discharge from the catchments area. Two possible ways of analysis is carried out which is by using time series prediction and 2nd order dynamics of the system. As we know that these models are highly non-linear, therefore the 2nd order system might give some significant improvement in the prediction models. This data consist one year compilation of discharged water (365 daily observation data) from the catchments area and rain falls. We divided the data to training, testing and validation where 100 for training and testing and the rest for validation. The data are normalized using its mean and standard deviation and is shown in Fig. 1.

RESULT AND DISCUSSION

We have trained the network with two input neuron, 1 to 10 hidden neurons and one output neuron. For each neural network performed 10 simulations. The initial weights were randomly chosen in the interval [-1,1]. The results obtained are shown in Table 3 to 6. The first column represents the number of hidden neurons used. The second until fourth column represents the training, testing and validation of SSE.

Table 2 shows the result was obtained for each time series model a, b and c. Model a with five hidden neurons give the best result is a training is 24.938, testing is 60.243 and validation is 114.36. Model b with four hidden neurons, the best result are obtained is a training is 21.73, testing is 63.603 and validation 118.86. While model c with two hidden neuron give the best result is a training is 25.389, testing is 58.443 and validation is 108.05.

Based on this analysis times series model c with hidden neuron of 2 is shows some significant performance to the model. Therefore in time series analysis, model with 2 hidden neuron were used as a basic model for analysis.

It is shown in Table 3 that 2 hidden neuron shows some significant result on the training and testing SSE, therefore 2nd order dynamics system with 2 hidden neuron were used as a basic model for analysis.

Comparison of Models: Fig. 2.0 and 3.0 shows the water discharge (Q) prediction as an example (for the four alternative neural network models investigated) compared to measure validation data set. It would be possible to compare the models predictions to the parameter estimation or training data, however one would expect a good fit in this case as the objective is to minimize the sums of squared errors. The validation set is of key significance, as it indicates the ability of the model generalize. The validation data and the corresponding predictions seem at first sight to be similar to each model with actual data set. It is difficult to assess the

Table 2: SSE of time series model a, b and c

	SSE of model								
	a			b			c		
Hidden	Train	Test	Valid	Train	Test	Valid	Train	Test	Valid
1	25.5402	59.2728	107.8102	25.7264	59.5412	108.1299	25.9671	58.1243	108.6586
2	25.747	60.0362	110.305	25.2596	60.2782	107.7173	25.389	58.443	108.05
3	24.2025	62.3941	109.2165	23.1246	67.6818	119.2022	28.1454	67.8047	143.4154
4	22.9107	68.5121	123.9763	21.73	63.603	118.86	22.1982	62.2297	132.4314
5	24.938	60.243	114.36	24.0468	61.3469	115.9002	24.0839	63.2657	127.4278
6	23.3872	63.2555	118.6575	22.4024	62.38	129.211	25.9651	65.9725	139.339
7	22.8761	63.409	118.7675	21.7003	64.4457	119.306	20.5021	74.7742	156.7594
8	22.3055	76.028	127.2101	21.8205	71.9189	122.845	17.1718	75.7905	158.9433
9	24.0683	66.5252	125.047	22.5453	72.8349	117.794	71.5871	94.0683	203.4477
10	24.552	70.0233	130.055	22.4597	66.5399	129.9118	15.8481	79.8287	169.3541

Table 3: SSE of second order dynamics model

	SSE of model		
Hidden	Training	Testing	Validation
1	25.6352	56.9722	108.2547
2	23.645	53.131	101.86
3	23.7987	58.3818	119.155
4	27.9955	56.0337	109.8377
5	26.6053	55.5742	105.5446
6	15.9779	68.3178	147.9366
7	14.6947	66.1117	150.6525
8	13.2706	74.6968	157.1994
9	18.3003	61.6179	119.7136
10	17.9822	56.2264	128.2235

performance of the model by looking at such plots and statistical analyses are normally needed.

However from the plotted is given (Fig. 2.0 and 3.0), it shows that the second order model give a good simulation where is the predicted and actual value is close to each other. While for the time series model, it cannot follow the actual value as good as a second order model. The predicted and actual value not too close to each other.

To consider of performance from both model which used in this study, the model comparison is common technique. Statistical test were based on correlation analysis is common in model identification. The correlation analysis as the primary measure of model performance in particular the coefficient of correlation, autocorrelation of the residuals (error between predictions and actual data) and the partial autocorrelation.

For a perfect predictor, the coefficient determination should be +1 or –1. In general the definition of r tells us that $100r^2$ is the percentage of the total variation of the predicted values which is explained by, or is due to their relationship with actual values. This is important measure of the relationship between two variables, beyond this scenario, it permits valid comparisons of the strength of several relationships [13, 14, 15].

From the Fig. 4, it can be seen that the scatter plot shows the highly positive correlation for the second order model compare to the time series model c. The coefficient of correlation for the second order model is 0.73 and 0.49 for the time series model.

The autocorrelation function should ideally resemble an impulse. This would indicate that the residuals “white”, i.e. no correlation exists between the residuals and any time shifted replica of the series [16]. There should also be no correlation between the residuals and any linear or

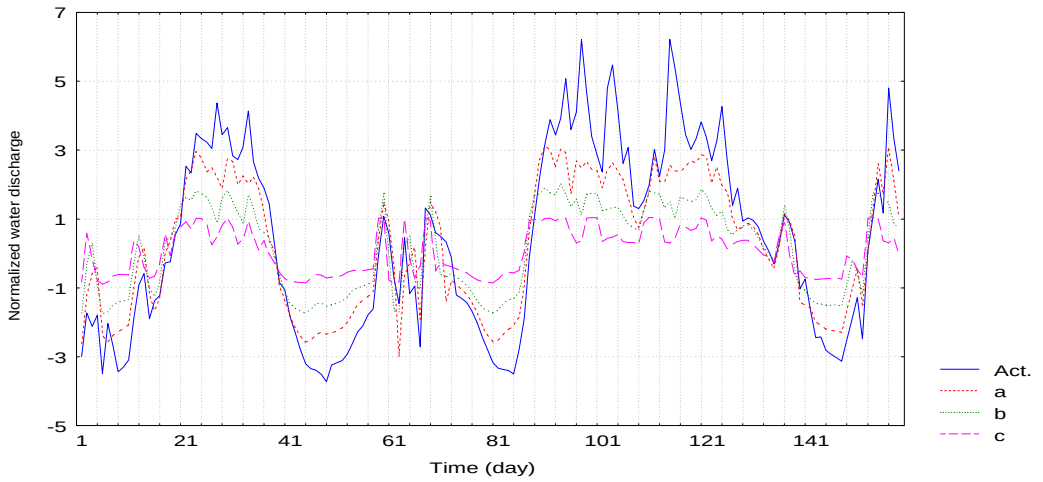

Fig. 2: Water discharge predictions for a time series model (validation data)

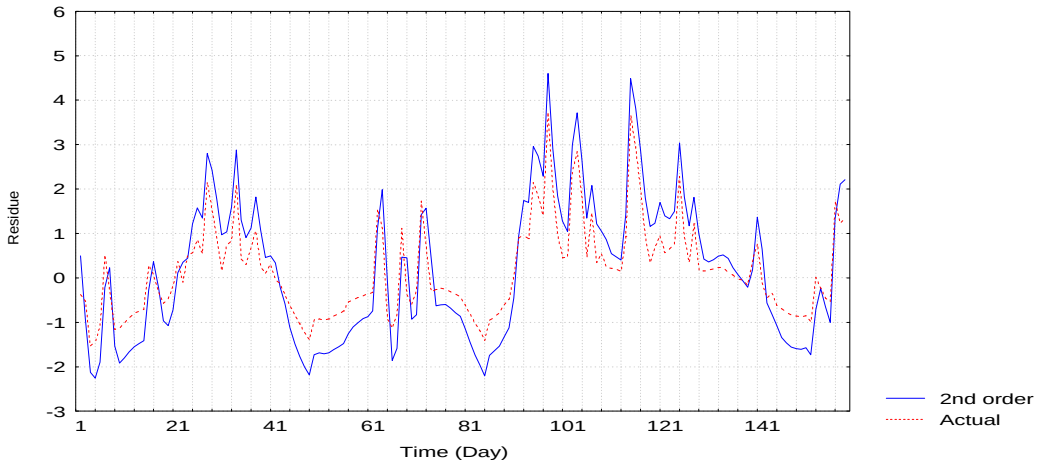

Fig. 3: Water discharge predictions for a second order dynamics model (validation data)

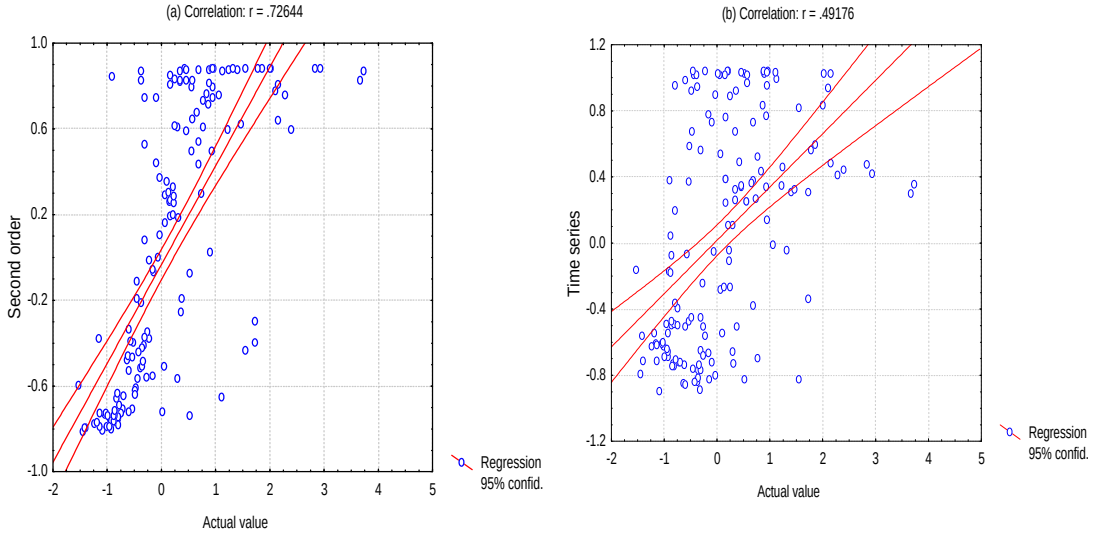

Fig. 4: Scatter plot between actual and predicted value of second order model (a) and time series model (b)

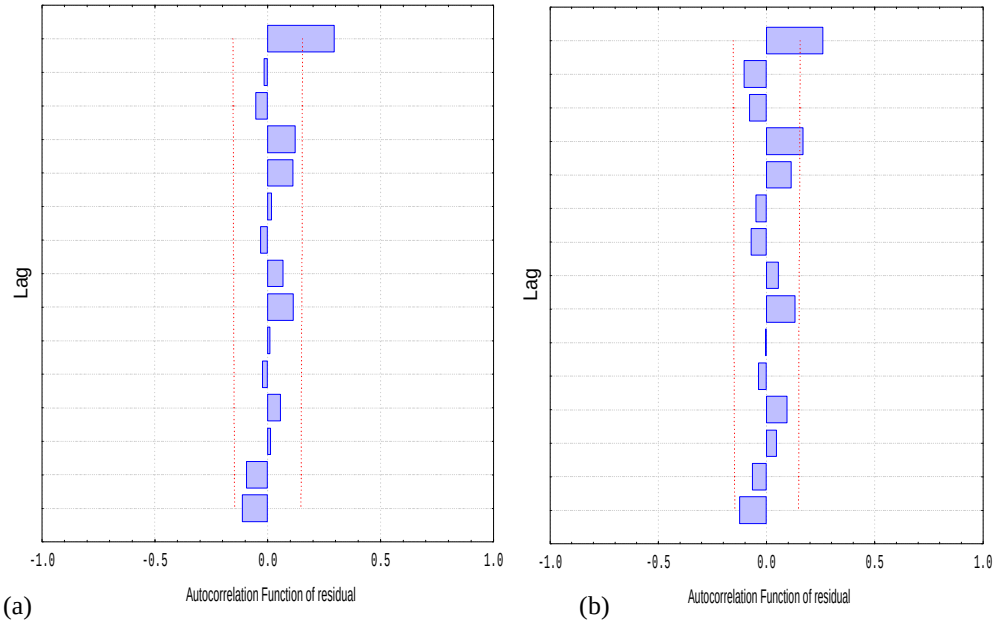

Fig. 5: Autocorrelation function of residuals water discharge for second order (a) and time series models (b)

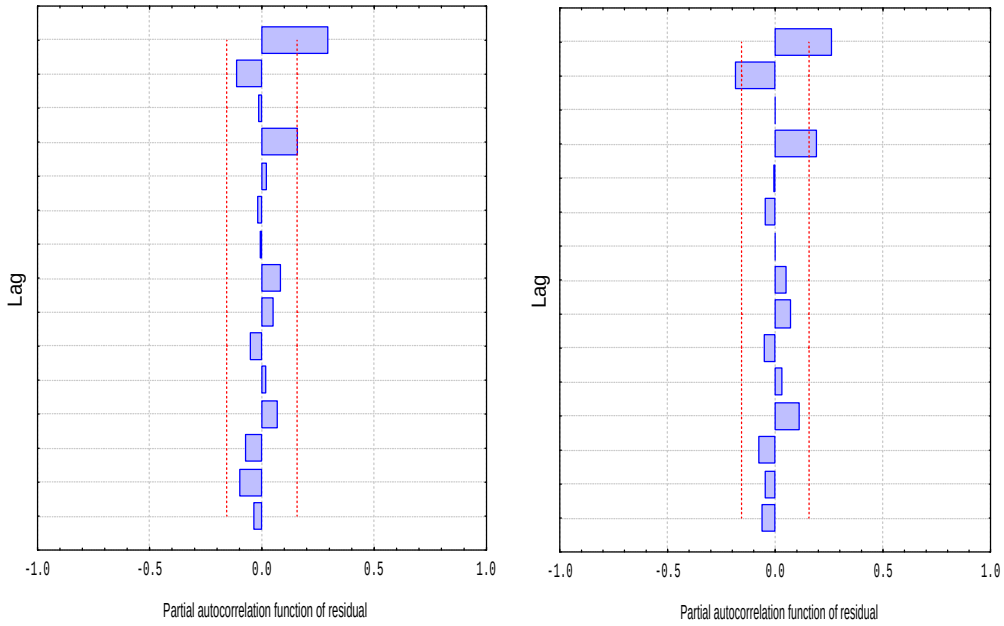

Fig. 6: Partial autocorrelation function of residuals water discharge for second order (a) and time series models (b)

non-linear combinations of past inputs and outputs. The autocorrelation and partial autocorrelation of the residuals should exhibit correlation only when the time series are not time shifted (one to the other). If the correlation functions are within a confidence interval of 95% of the conditions mentioned, then it is reasonable to accept the model as a fair representation to simulate of the real data.

The residuals analysis is very useful analysis to help us to identify the performance of the both models. The analysis for the residual are presented in Fig. 5 and 6. It can be seen the autocorrelation and partial autocorrelation function for the second order model is fair enough to achieve the 95% confidence interval (within read line in the figure). While for the time series model c, it fail to

achieve the 95% of the confidence criteria. Where the time series model c shows the two temporal structure for autocorrelation and three temporal structure for partial autocorrelation function out of the 95% confidence criteria.

CONCLUSION

Based on SSE produced of each model, two significant models were selected as the best model to simulate water discharge. Time series model c and second order model were selected to comparison for further analysis.

The comparisons of the second order and time series models performance, measured by correlation analysis showed that the second order model had better performance. It can be give a good simulation in the case of water discharge prediction at station 2917401 using two input variables, namely rainfall stations 2917001 and 3018101. The autocorrelation function remained within 95% confidence limits for the most part, as did the partial autocorrelation function.

Based on the results, of the two models investigated, the second order model would be preferred as water discharge modelling at that station.

ACKNOWLEDGEMENT

Special Thank to Universiti Kebangsaan Malaysia (Centre for Research and Innovation Management- GUP-ASPL-07-06-002) and Drainage and Irrigation Malaysia for general assistant.

REFERENCES

1. Bishop, C., 1995. ”Neural Networks for Pattern Recognition,” Clarendon Press, Oxford.
2. Caruana, R., S. Lawrence and C. Lee Giles, 2001. Overfitting in Neural Networks: Backpropagation, Conjugate Gradient and Early Stopping, Advances in Neural Information Processing System, 13: 402-408.
3. Chen, S., S.A. Billings and P.M. Grant, 1990. Nonlinear system identification using neural networks. Int. J. Control., 51: 1191-1214.
4. Freund, J.E. and G.A. Simon, 1992. Modern Elementary Statistics, Prentice-Hall Inc., Upper Saddle River, New Jersey.
5. Hagiwara, K. and K. Kuno, 2000. Regularisation Learning and Early Stopping in Linear Networks, International Joint Conference on Neural Networks (IJN 2000), pp: 511-516.
6. Hashem, S., 1997. Optimal Linear Combination, Neural Networks, 10(4): 599-614.
7. Heartz, J.A., A. Krogh and R.G. Palmer, 1991. Introduction to the Theory of Neural Computation, (Addison-Wesley, Redwood City, CA).
8. Morgan, N. and H. Bourlard, 1990. Generalisation and Parameter Estimation in Feedforward Nets: Some Experiments, In Touretzkey, D.S (Ed.), Advances in Neural Information Processing System, Vol 2, San Mateo CA, pp: 630-637.
9. Ohbayashi, M., K. Hirasawa, K. Toshimitsu, J. Murata and J. Hu, 1998. Robust Control for Non-linear System by Universal Learning Networks Considering Fuzzy Criterion and Second Order Derivatives, IEEE World Congress on Computational Intelligence: IEEE International Conference Proceeding on Neural Networks, 2: 968-973.
10. Premier, G.C., R. Dinsdale, A.J. Guwy, D.L. Hawkes and S.J. Wilcox, 1999. A comparison of the ability of black box and neural network models of ARX structure to represent a fluidised bed anaerobic digestion process, Water Research 33: 1027-1037.
11. Sharkey, A.J.C., 1999. Multi Nets System, Combining Artificial Neural Nets Ensemble and Modular, Amanda J.C. Sharkey (Ed), Springer Publication London.
12. Sridhar, D.V., E.B. Bartlett and R.C. Seagrave, 1999. An Information Theoretic Approach for Combining Neural Network Process Models, Neural Networks, 12: 915-926.
13. Sridhar, D.V., E.B. Bartlett and R.C. Seagrave, 1996. Process Modelling Using Stacked Neural Networks, AIChe J., 42(9): 2529-2539.
14. Wolpert, D.H., 1992. Stacked Generalisation, Neural networks, 5: 241-259.
15. Zhang, J., 2001. Developing Robust Neural Network Models by Using Both Dynamic and Static Process Operating Data, Ind. Eng. Chem. Res., 40: 234-241.
16. Zhang, J., E.B. Martin, A.J. Morris and C. Kiparissides, 1997. Inferential Estimation of Polymer Quality Using Stacked Neural Networks, Computer Chemical Engineering, 21: 1025-1030.