NEURAL NETWORK MODEL FOR PREDICTION OF DISCHARGED FROM THE CATCHMENTS OF LANGAT RIVER, MALAYSIA

ZAINAL AHMAD AND HAFIZAN JUAHIR[2]

School of Chemical Engineering, University Sains Malaysia, Engineering Campus, Seri Ampangan, 14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia.
[2]Chemistry Department, Faculty of Science, University Malaya
E-mail: chzahmad@eng.usm.my

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, neural networks models were used to model the daily river flows or discharged in Langat River, Malaysia. Two possible ways of modelling were implemented which is by time series prediction and by the dynamics function of the system which include the past value of the discharged and also the rainfall in the input. The sum square error (SSE), residue analysis and correlation coefficient based on the observed and prediction output is chosen as the criteria of selection of which models is appropriate. It was found that the developed neural networks models using dynamics function provided satisfactory model discharges.

Keywords: *Artificial neural networks, time series prediction, nonlinear process modelling, water discharged*

1. 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 [1, 2]. 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 [3,4]. 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. Lack of robustness in neural network models is basically due to the overfitting and poor generalisation of the models [5,6]. Therefore, many researchers was interested and concentrate on how overfitting can be avoided by improved the learning algorithm or by combining the neural networks [2, 7,8, 9, 10, 11, 12]. In view to the robustness of neural network, a lot of techniques have been developed like regularisation and the early stopping method [9,13,14]. Reference [14] implemented the universal learning rule and second order derivatives to increase the robustness in neural network models.

Multiple-regression analysis was also being used to develop potentially observable variables to estimate river discharge using remote sensing techniques [15,16]. In this study the multiple regressions are applied to estimate the water discharged using information from the satellite remote sensing. The result was quite good where the relative error in this case study is quite small. Nguyen *et al* [17] apply the neural networks model with back propagation training methods to forecast the river discharge in Thailand and the performance based on the two case studies is quite convincing eventhough the result based on the small basin river is less accurate but the overall performance is quite reasonable. Currently, several black-box models (having little or no physical considerations) have also been used. These models include the time-series approach using Box-Jenkins ARIMA models, multiple regression models [16,18]. Recently, back propagation neural networks (BPNNs), a particular type of neural network, have been developed and successfully used in many fields

In this study, neural networks modelling using Levenberg-Marquardt optimisation techniques were used to develop the models for forecasting daily discharges (with lead time equal to one day) in Langat River , Malaysia. These model actually a BPNN type of model but the different is the optimisation method to update the weight in the training algorithm is using Levenberg-Marquardt techniques [15]. Besides daily rainfall from two rainfall station, the inputs to these models also include the past values of the discharge itself. Two proposed model were develop in this study, one is by time series prediction using water discharged and the other using dynamics of the process including the past values of the rainfall in the input. The evaporation data was excluded in this study where for run-off water, the evaporation data is not really significant compare to rainfall [17].

This paper is organised as follows. Section 2 presents the concept of neural network modelling. Section 3 presents the application of the proposed technique; the results and discussion of the proposed modelling are presented in Section 4. Finally, the last section concludes this paper.

2. NEURAL NETWORK MODEL

The objectives of this paper are to utilise the neural network modelling techniques to model the discharged from the catchments of Langat River. The modelling techniques consist two different models which are by the time series prediction or by the dynamics function of the system.

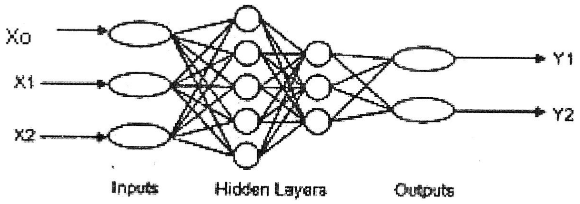

Fig. 1: Structure of Feedforward Neural Network

The neural networks were trained by the Levenberg-Marquardt optimisation algorithm with regularisation and "early stopping". All weights and biases were randomly initialised in the range from –0.1 to 0.1. In this case, the input layer has several nodes, each representing an input variable. The hidden layer also has several nodes and represents the non-linearity of the network system as the basic structure shown in Fig. 1. The output layer has only one node which represents the forecast value corresponding to each set of input values. In principle, the neural networks may have several hidden layers, but in practice, only one or two layers are used. The number of nodes in the hidden layer is determined mainly by trial and error as shown in Table 1 and Table 2.

The hidden neuron is varies from 1 to 10 and one output nodes for both cases. The purpose of this variation is to determine how many hidden neuron to be used in that particular model based on the sum square error (SSE) in testing data. The least SSE in the testing data which is corresponding to the number of hidden nodes in the model will selected as a final model. The results obtained are shown in Table 1 and 2 for model 1 and model 2 respectively. Both model used 2 hidden neurons in their final neural networks model. The architecture of the each model used in this study is shown in Table 3.

Table 1: SSE for different Number of hidden nodes in Model 1

No. of hidden nodes	SSE Training	SSE Testing	SSE validation
1	25.9671	58.1243	108.6586
2	**25.389**	**58.443**	**108.05**
3	28.1454	67.8047	143.4154
4	22.1982	62.2297	132.4314
5	24.0839	63.2657	127.4278
6	25.9651	65.9725	139.339
7	20.5021	74.7742	156.7594
8	17.1718	75.7905	158.9433
9	71.5871	94.0683	203.4477
10	15.8481	79.8287	169.3541

Table 2: SSE for different Number of hidden nodes in Model 2

No. of hidden nodes	SSE Training	SSE Testing	SSE 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

Table 3: Architecture of the models for neural networks modelling

Case Model	No. of nodes in input	No. of hidden neurons	No. of nodes in output
1	4	2	1
2	6	2	1

Hidden neurons use the logarithmic sigmoid activation function whereas output layer neurons use the linear activation function. To cope with the different magnitudes in the input and output data, all the data were scaled to zero mean and unit standard deviation. The data for neural network model building will be divided into 3 sections which are: 1). Training data (for network training); 2). Testing data (for cross-validation based network structure selection and early stopping); and 3). Unseen validation data (for evaluation of the final selected model). The two proposed model for modelling are shown below:

Model 1:Time series prediction In the formal term, time series are a sequence of number where in this case they are a sequence of water discharge from the Langat River in daily series or at time, *t* (day). The discharged water at time *t*, is a function of the past discharge water at time *t-1, t-2, t-3* and *t-4* as shown below:

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

where $\hat{y}(t)$ is the models prediction of water discharge form the catchments area at time t , $y(t-1)$, $y(t-2)$, $y(t-3)$ and $y(t-4)$

Model 2: Dynamic prediction

The discharged at time t, is a function of the past discharged water at time *t*-1 and *t*-1, the past rainfall from the station 1 at time *t-1* and *t*-2, and past rainfall from the station 2 at time *t-1*and *t-2.*The lag is still at one day. The dynamic model structure is shown as followed:

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

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

3. APPLICATION

3.1 Prediction of Discharged from Langat River, Malaysia

The Langat River Basin occupies the south and south-eastern parts of the state of Selangor Darul Ehsan in Malaysia. It is about 78 km long and ranges from 20 km to 51.5 km wide. It has total catchments of 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 the mean sea level can be found. It finally drains into the Melaka Strait on the mangrove coastline of south western Selangor. The major tributaries of the Langat River are the Semenyih River, the Labu River and the Mantin River. The purpose of this study is to predict the water discharged from checkpoint 1 (WO2917401) which is shown in Fig. 2. Therefore when the water discharge from checkpoint 1 can be estimated for a certain period of time, some preventive measures can be taken at the downstream of the river basin where flooding might happen. This is particularly important since checkpoint 1 is near to an industrial area and also near to a pig farm. Flood in this area can cause contamination to the basin of the Langat River.

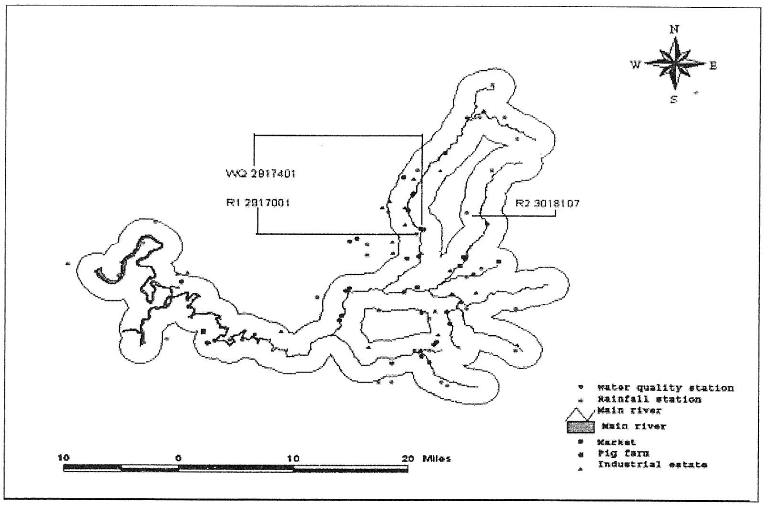

Fig. 2: River basin of the Langat River in Malaysia

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 4. 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.

Table 4: 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%**

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.

In this case study, the data was consisting of 1 year compilation of discharged water (365 daily observation data) from the catchments area and rain falls. Then the data were divided into training, testing and validation where 100 for training and testing and the rest are for the validation. The data are normalized with zero mean and standard deviation and it is shown in Fig. 3.

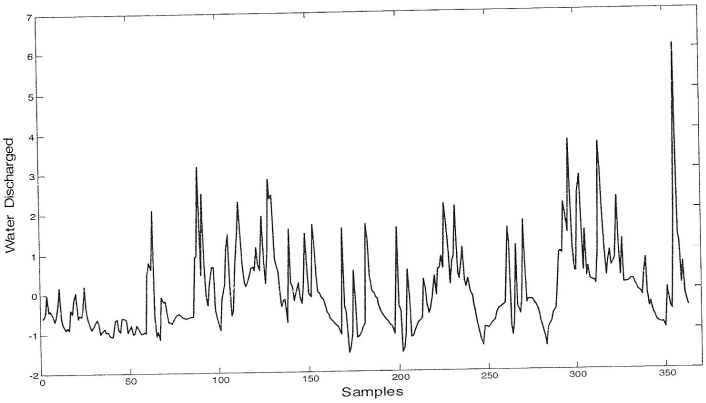

Fig.3: Data used in parameter estimation on neural network training, testing and Validation

4. RESULTS AND DISCUSSION

The neural networks model were trained with 4 input nodes, for model 1 and 6 input nodes for model 2. The data was divided into three set which is training and testing data with 101 samples and validation data with 158 samples. The calculated performance statistics are shown in Table 5.

Table 5: Performance statistic for model 1 and model 2.

	Model 1	Model 2
SSE	144.53	116.25
Correlation coefficient	0.51	0.75
Coefficient determination, r^2	0.19	0.61

As seen in Table 3, the coefficient determination, r^2 for model 1 is about 0.19 and for model 2 is 0.61. For a perfect predictor, the coefficient determination should be +1 or –1. In general the definition of r tells us that 100 r^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 an important measure of the relationship between two variables, beyond this; it permits valid comparisons of the strength of several relationships [5].

Based on this value, indicating that the model 2 is better than model 1 even though the r^2 value is not really reaching 1.

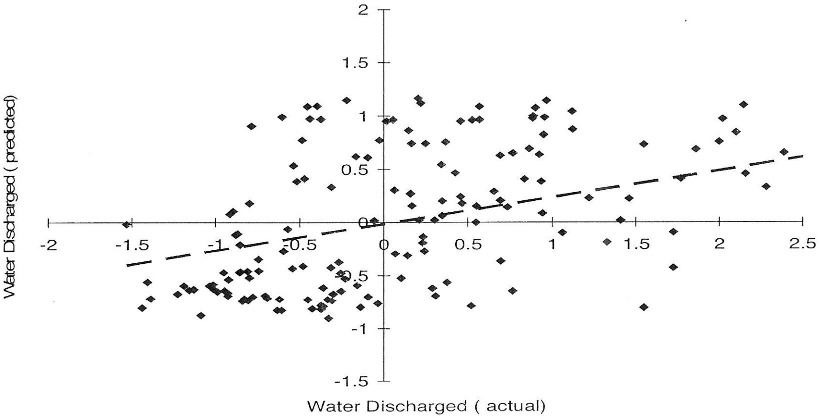

Fig 4: Scatter plot for actual and predicted value for Model 1.

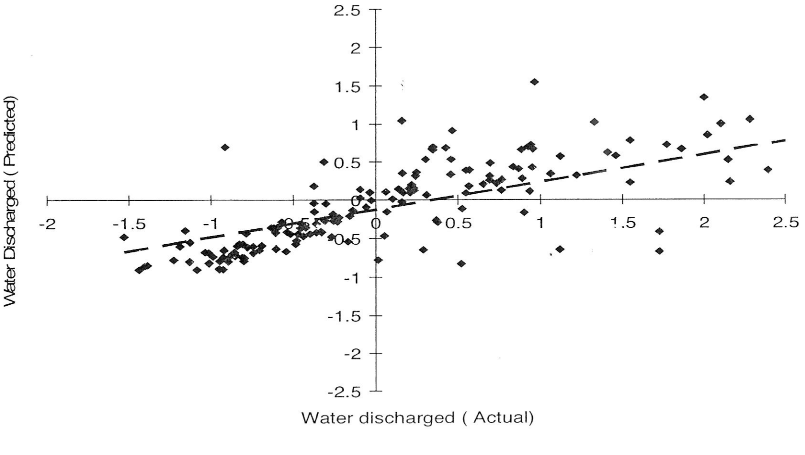

Fig. 5: Scatter plot for actual and predicted value for Model 2

The correlation coefficient was also high on model 2 which is 0.61 compared to model 1 which is 0.51. Correlation coefficient is basically a numerical index of relationship between 2 variables. A positive (measure of direction) correlation or direct relationship indicates that a high score on one variable is associated with a high score on the second variable. A negative correlation or inverse relationship indicates that a high score on one variable is associated with a low score on the second variable. The magnitude of the correlation coefficient indicates the strength of the relationship between the two variables. This magnitude can vary from 0.00 to 1.00. The closer the correlation coefficient is to either -1.00 or +1.00 the stronger the relationship. The more strongly two variables are related, the better the prediction. This phenomenon can be seen in Fig. 4 and Fig. 5 for model 1 and model 2 respectively. Model 2 gives a clear view that the predicted and the actual process output is strongly related which mean that the prediction output is quite near to the actual data. Even though the correlation coefficient value for model 2 isstill low but it can differentiate which model is better than the others.

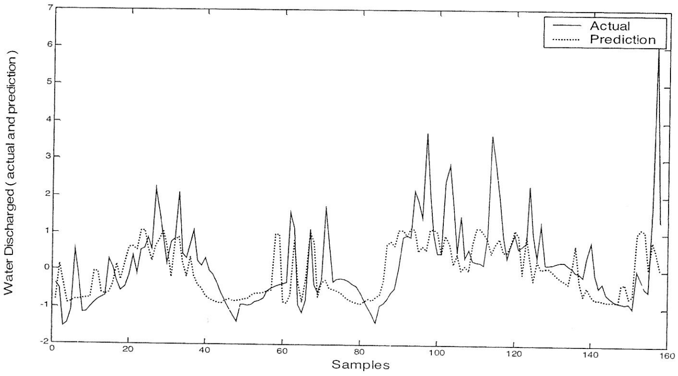

Fig. 6: Water discharge predictions for a time series model (validation data).

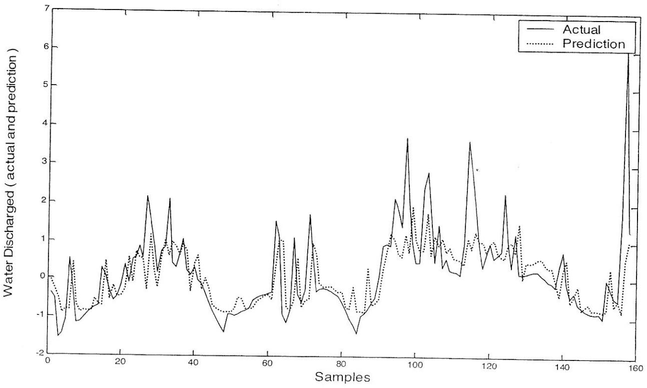

Fig. 7: Water discharge predictions for a second order dynamics model (validation data)

The result obtained for SSE for model 2 is 116.25 compared to model 1 which is 144.53. From these results as well as from the r^2 and the correlation coefficient, it appears that model 2 always outperformed the first model. As shown in Fig. 6 and Fig. 7, that the actual and predicted output for model 2 is reasonably good compare to first model. From both figures, it was also found that the water discharged was always under estimated during the high flow period but in the low flow period the prediction model is reasonably good. It might be due to the quality of the data in the testing and training that not really covers the range of the high flow rate while develop the model.

The residue analysis was also carried out in this study. The residual analysis is a very useful analysis to help us to identify the performance of both models. The analysis for the residual is presented in Fig. 8. It can be seen that the residue for model 2 is more consistent where the output is near to zero even though in some samples the residue is high. It shows that second model output or prediction always near to the actual data.

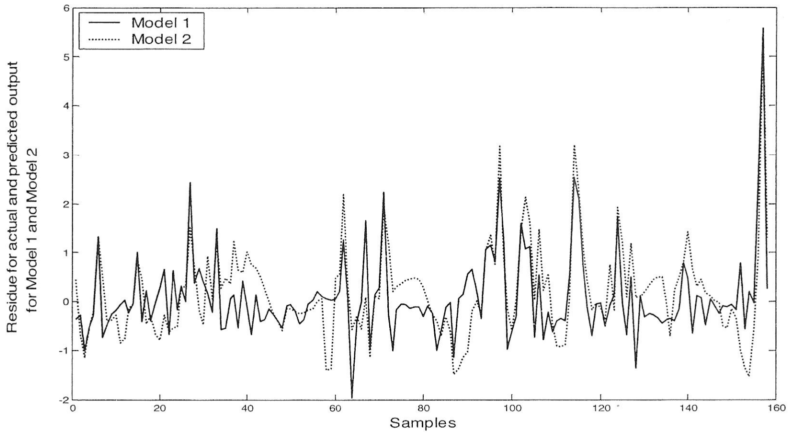

Fig 8: Residuals water discharge for Model 1 and model 2

5. CONCLUSION

The results obtained in this paper indicate the capability of neural networks models in modelling of daily river flows using daily rainfall data as inputs as well as the past value of the water discharged. The proposed two modelling techniques show that both models can performe well, which is measured by correlation coefficient analysis, coefficient determination and SSE but it was found that the second model had better performance based on that statistical analysis. It was noted that the contribution of past values of discharge was very important in neural networks modelling as well as the past value of the rainfall. This was not observed in the time series modelling techniques where it only used past value of the water discharged. This was the main reason why the time series modelling techniques was not performed in this study. Therefore the appropriate techniques to model the water discharged in this study is by using the past value of the rainfall as well as the past value of the water discharged.

ACKNOWLEDGEMENT

This work was supported by University Science Malaysia (USM) (for Z. Ahmad) and Ministry of Science (for Hafizan Juahir).

REFERENCES

[1] S, Chen, S.A Billings and P.M Grant, "Nonlinear system identification using neural networks". Int. Journal Control 51, pp.1191-1214, 1999.

[2] J. Zhang,, E.B.Martin, A.J.Morris and C.Kiparissides, " Inferential Estimation of Polymer Quality Using Stacked Neural Networks," Computer and Chemical Engineering, Vol, 21, pp. 1025-1030, 1997.

[3] C, Bishop. "Neural Networks for Pattern Recognition," Clarendon Press, Oxford 1995.

[4] J.A Heartz, A Krogh and R.G Palmer, "Introduction to the Theory of Neural Computation," (Addison-Wesley, Redwood City, CA), 1991.

[5] R, Caruana, S.Lawrence and C, Lee Giles, "Overfitting in Neural Networks: Backpropagation, Conjugate Gradient and Early Stopping," Advances in Neural Information Processing System, Vol. 13, pp 402-408, 2001.

[6] N. Morgan and H. Bourlard, "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, 1990.

[7] S. Hashem, "Optimal Linear Combination, "Neural Networks, Vol. 10: 4, pp. 599-614, 1997.

[8] A.J.C Sharkey, "Multi Nets System," Combining Artificial Neural Nets Ensemble and Modular, Amanda J.C Sharkey (Ed), Springer Publication London, 1999.

[9] D.V Sridhar, E.B.Bartlett and R.C.Seagrave, " An Information Theoretic Approach for Combining Neural Network Process Models," Neural Networks, Vol. 12, pp. 915-926, 1999.

[10] D.V Sridhar, E.B.Bartlett and R.C.Seagrave, " Process Modelling Using Stacked Neural Networks," AIChe Journal, Vol. 42:9, pp. 2529-2539, 1996.

[11] D.H Wolpert, " Stacked Generalisation," Neural networks, Vol. 5, pp. 241-259, 1992.

[12] J. Zhang, " Developing Robust Neural Network Models by Using Both Dynamic and Static Process Operating Data," Ind.Eng.Chem.Res, Vol. 40, pp. 234-241, 2001.

[13] K, Hagiwara and K. Kuno,"Regularisation Learning and Early Stopping in Linear Networks," International Joint Conference on Neural Networks (IJCNN 2000), pp 511 – 516, 2000.

[14] M. Ohbayashi, K. Hirasawa, K. Toshimitsu, J. Murata and J. Hu, " 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, Vol 2, pp. 968-973, 1998.

[15] M, Campolo, P, Andreussi and A. Soldati, "River Flood Forecasting with a Neural Network Model", Water Resources Research, 35(4), 1191-1197, 1999.

[16] H.N Phien, B.K Huong and P.D Loi, "Daily Flow Forecasting with Regression Analysis", Water SA 16 (3),pp179-184,1990

[17] T.D Nguyen, H.N Phien and A.D Gupta, "Neural Network Models for River Flow Forecasting", Water SA Vol. 25 (1) pp 33-39, 1999

[18] G.C Premier, R. Dinsdale, A.J Guwy, D.L Hawkes and S.J Wilcox, "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, pp.1027-1037, 1999

[19] J.E Freund and G.A Simon, "Modern Elementary Statistics", Prentice-Hall Inc., Upper Saddle River, New Jersey, 1992.