Application of MLP Neural Network and M5P Model Tree in Predicting Streamflow: A Case Study of Luvuvhu Catchment, South Africa

E. K. Onyari and F. M. Ilunga

***Abstract*—Reliable estimation of discharge is important in water resource planning and management, as well as in systems operation. This paper presents a rainfall runoff modelling approach using data mining techniques namely multi layer perceptron neural network and M5P-Model tree. Both models were developed, trained and verified for the discharge at Luvuvhu River, Mhinga gauging station. The relevant inputs into the models were selected by minimum Redundancy maximum relevance (mRMR) algorithm. The M5P Model Tree developed with 66% training set was realized to be the best model that predicted flow with a RMSE of 2.666, and a correlation coefficient of the observed and the predicted flow of 0.89. A MLP-ANN with 4 hidden nodes performed satisfactorily with RMSE ranging from 3.42 to 5.22. It is concluded that Model tree M5 predicts better than ANN-MLP, although it is quite sensitive to data splitting.**

***Index Terms*—Streamflow prediction, MLP-ANN, M5P-model tree, mRMR.**

I. Introduction

In water resource planning, design and infrastructure development reliable quantitative estimation of discharge is crucial. Predicting flow presents many advantages as decision makers can anticipate extreme events i.e. both high and low flows, so as to plan and manage them well. Discharge can be measured reasonably well, however due to the spatial and temporal variability of rainfall; the forcing function that causes the discharges is not easy to characterize [1], thus this has necessitated researchers to use different methodologies in modelling streamflow. Modelling has proved to be an essential tool in predicting flows, with the techniques used ranging from physically-based models to data mining models.

Previous studies have been supportive of the latter especially artificial neural networks (ANN) for flow prediction. ANNs are effective in pattern recognition and function approximation [2] which are the main characteristics of water resources problems.

The advantage of ANN is that no prior knowledge of the catchment characteristics is required, because even if the exact relationship between the input and output is unknown but is acknowledged the network can be trained to learn the relationship [3]. ANNs can be taken as black box models since they neither learn based on assumptions relating to the input-output transfer function nor the physical interaction of the parameters.

ANNs have found increasing applications in water resources and environmental systems for instance in rainfall runoff modelling [4], short term flow prediction [5]; stage-discharge (rating curve) modelling [6] and flood forecasting [7].

Multi layer perceptron neural network (MLP- ANN) which uses back propagation algorithm has been used widely in water resources successfully. Back propagation is a supervised learning method where the algorithm works towards minimising the error between its output and the target. It is explained literature that it appears in practice that the back-propagation method leads to solutions in almost every case, although, the error back-propagation method does not guarantee convergence to an optimal solution since local minima may exist.

In addition, [8] concluded that standard multi-layer; feed-forward networks are capable of approximating any measurable function to any desired degree of accuracy and called them universal approximators. However, the study pointed out that errors in approximation may arise from inadequate learning, having insufficient number of hidden units or the relationship between the input and the output being insufficiently deterministic. In this regard, and following the successful application of MLP-ANN in various aspects of water resources the method was used in this study.

In [9] M5 are described as tree based models with their leaves having multivariate linear models, these model trees are thus analogous to piecewise linear functions. M5 model tree is relatively new in water resources but in the events it has been used it has proved to be quite robust. For instance it was used in the water level-discharge relationship in [6], [10] and it was found that M5 had the same predictive accuracy as an ANN built with the same data. M5 learns efficiently and tackles tasks with very high dimensionality.

In this study a MLP-ANN with four hidden nodes and an M5-Pruned (M5P) model tree were applied to the Luvuvhu catchment to predict streamflow at Mhinga gauging station (A9H012) while accounting for rainfall data from Thohoyandou station (07236646) and streamflow from Mutshindudi River (A9H025) and Nandoni dam outflows (A9H030).

II. Machine Learning Approaches

A. Artificial Neural Network

ANNs were developed with the intention of mimicking the functioning of the human brain [11]. They contain several simple units each having a small amount of memory, which is

Manuscript received September 15, 2012; revised November 30, 2012.

The authors are with the Department of Civil Engineering, University of South Africa, Florida campus, South Africa (e-mail: onyarek@unisa.ac.za, ilungm@unisa.ac.za).

DOI: 10.7763/IJIMT.2013.V4.347

interconnected by communication channels that carry numerical data. These units use their local data and the inputs they receive through connections to do computations.

The architecture of a MLP-ANN is made up of a number of interconnected nodes arranged into three types of layers: input, hidden and output. Fig. 1 shows a schematisation of the various components of a MLP with one hidden layer. The input layer simply sends the input values x_i to the units in the hidden layer, but it does not perform any operation upon the input signal. A hidden layer receives signals from the nodes of the input layer and transforms them into signals which are sent to all output nodes which, in turn, transform them into outputs. The weights at the connections, from the input to the hidden node and from the hidden layer to the output layer, are calibrated using the steepest descent algorithm. This algorithm is used for solving the non-linear problems.

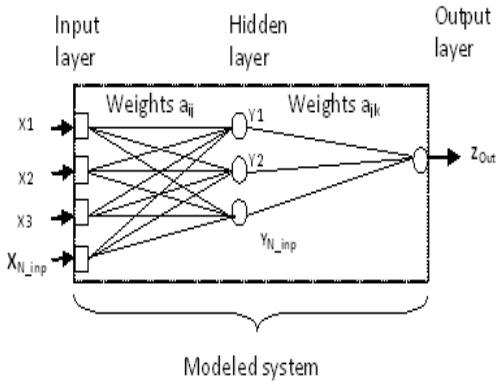

Fig. 1. A Multi-layer perceptron with one hidden layer.

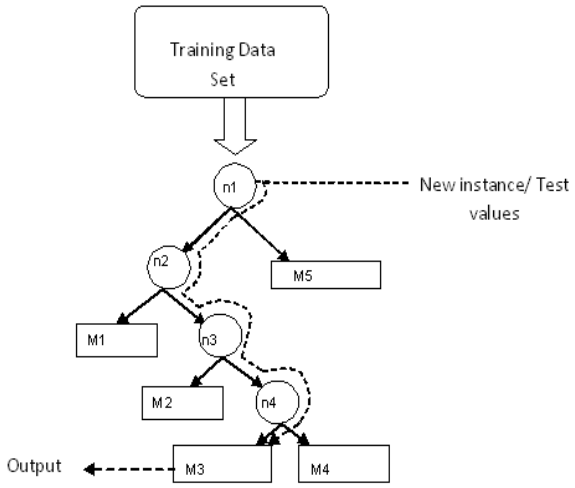

Fig. 2. A M5P model tree, n_i are split nodes and M_i are the models.

B. M5P Model Tree

A model tree is used for numeric prediction and at each leaf it stores a linear regression model that predicts the class value of instances that reach the leaf. In determining which attribute is the best to split the portion T of the training data that reaches a particular node the splitting criterion is used. The standard deviation of the class in T is treated as a measure of the error at that node and each attribute at that node is tested by calculating the expected reduction in error. The attribute that is chosen for splitting maximises the expected error reduction at that node. The standard deviation reduction (SDR) which is calculated by (1) is the expected error reduction.

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} \times sd(T_i) \quad (1)$$

where T_i corresponds to T_1, T_2, T_3 ... sets that result from splitting the node according to the chosen attribute. See Fig. 2.The linear regression models at the leaves predict continuous numeric attributes. They are similar to piecewise linear functions and when finally they are combined a non-linear function is formed [6]. The aim is to construct a model that relates a target value of the training cases to the values of their input attributes. The quality of the model will generally be measured by the accuracy with which it predicts the target values of the unseen cases.

The splitting process terminates when the standard deviation is only a small fraction less than the standard deviation of the original instance set or when a few instances remain. For detailed reading check [12] as M5 model trees are clearly explained.

III. Experimental Set Up

A. Study Area

The Luvuvhu catchment is located in the north-eastern part of South Africa in Limpopo Province, with an area of approximately 5941 km^2. It originates from Soutpansberg Mountains, flows through kruger National park and empties into the Limpopo river at the border with Mozambique and Zimbabwe. The Luvuvhu Catchment is one of the 18 Water Management Areas (WMA) that has been identified by the Department of Water Affairs and Forestry.

For water resources management purposes, the catchment has been sub-divided into 14 quaternary catchments. The Luvuvhu catchment has a Mean annual precipitation (MAP) of over 608 mm/yr, potential evaporation is estimated at 1678 mm/ yr and natural mean annual runoff (MAR) estimated to be 520 $\times 10^6$ m^3/yr. There is a high variability of rainfall and evaporation throughout the catchment. The catchment has several rainfall and streamflow gauging stations. In this study only Thohoyandou- 07236646 rainfall station, and three flow gauging stations namely Mhinga (A9H012), Mutshindudi (A9H025), and Nandoni dam outflow (A9H030) were used since they fall in the sub quaternary catchment of interest.

B. Input Data Selection

Data input selection is an initial and necessary step of any modeling practice, and proper selection of data input variables dictates the modeling accuracy of the processes or systems in question [13]. Successful application of ANN model or model trees requires proper input data selection, as this enables a better understanding of the true driving forces of the modelled system. In addition, a firm understanding of the process being modelled is essential for proper selection of input variables. This fundamental understanding will not only help in choosing proper input data but also help in avoiding wrong input data which will confuse the training process.

In rainfall-runoff modeling, the inputs consist of rainfall data while the output is discharge at the outlet. However, it is noted that the travel time throughout the catchment can vary

thus the contributions from various parts of the catchments varying considerably. This requires that rainfall data of several days before the discharge of interest be available for the learning of the network. Again, of importance is the inclusion of the previous output variables i.e. a flow at one day before (*t*-1), two days before (*t*-2) as an input to determine the flow at time t, a method referred to as recurrent back propagation [5].

The data that was used to predict flow at Mhinga station(gauge A9H012) included flows at Mhinga, Mutshindudi, and Nandoni dam outflow for a period of 3 years. There was a major constrain on the available data since the Nandoni dam was built in 2006. The data period used ranged from 2007 July 26th to 2010 july 11th as this was the data that was common to all gauges, although gauge A9H012 had 23 years of data and A9H025 had 15 years of data. Fig. 3 shows a graph of measured rainfall and streamflow that were used in model training and verification.

The data was arranged with a lag of several days introduced, where (*QL*, *t*, *QL*, *t*-1, *QL*, *t*-2, *QL*, *t*-3, *QL*, *t*-4, *QL*, *t*-5) corresponded to flows in Luvuvhu river at Mhinga station,(*QM*, *t*, *QM*, *t*-1, *QM*, *t*-2, *QM*, *t*-3, *QM*, *t*-4, *QM*, *t*-5)are flows at Mutshindudi river,(*QN*, *t*, *QN*, *t*-1, *QN*, *t*-2, *QN*, *t*-3, *QN*, *t*-4, *QN*, *t*-5)are Nandoni dam outflows, and (*RT*, *t*, *RT*, *t*-1, *RT*, *t*-2) are Thohoyandou rainfall data. Dimension reduction was done to remove insignificant or non contributing inputs. In this case the minimum Redundancy Maximum Relevancy (mRMR) was used to select the input variables. mRMR algorithms were proposed by [14]; where maximum relevance drives the selection process in favour of the most relevant set of variables with no attention to minimum redundancy and vice versa. Therefore, mRMR selection is based on two selection processes favouring variables that bring high relevance and low redundancy on average.

Based on mRMR method five flow variables were selected, and they include: *Q*(*L*, *t*), *Q*(*L*, *t*-1), *Q*(*M*, *t*) , *Q*(*N*, *t*-1), and *Q*(*N*, *t*-2). The mRMR selected variables together with the rainfall variables with two days lag were used as inputs into the model. The function below shows the relationship between discharge at Mhinga station at time t and all relevant variables.

$$Q(L, t) = \text{f}\,(Q(L, t\text{-}1), Q(M, t), Q(N, t\text{-}1), Q(N, t\text{-}2), R(T, t), R(T, t\text{-}1), R(T, t\text{-}2)$$

where *Q*(*L*, *t*-1) is the discharge 1 day before at Mhinga station,*Q*(*M*, *t*) is the discharge at time t at Mutshindudi station,*Q*(*N*, *t*-1) is the discharge 1 day before at Nandoni dam station, *Q*(*N*, *t*-2) is the discharge 2 days before at Nandoni dam station, *R*(*T*, *t*) is the rainfall at time t before at Thahoyandou rainfall station, *R*(*T*, *t*-1) is the rainfall 1 day before at Thohoyandou rainfall station, and *R*(*T*, *t*-2) is the rainfall 2 days before at Thahoyandou rainfall station.

The Weka software was used for M5P model tree and MLP-ANN calibration and verification, the most optimal MLP-ANN that was built had 4 hidden nodes. The training time was less than 1 minute in all cases.

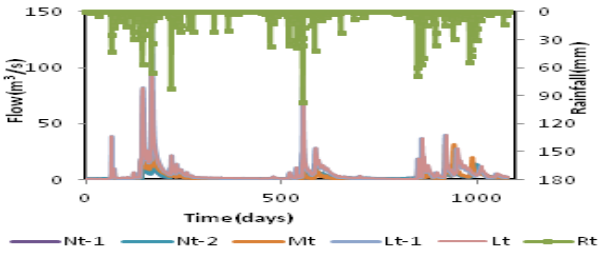

Fig. 3. A representation of corresponding measured rainfall and streamflow data at Mhinga station.

IV. Results and Discussions

Machine learning techniques whether it is artificial neural network or a model tree will have learnt well if they have good generalization ability. A model performs well when it has learnt the main characteristics in a training set, and correctly classifies new information. Model performance was assessed on the basis of the values of the root mean square error (RMSE), Mean absolute error (MAE) and the correlation coefficient.

A. Training and Verification

The available data was split into two sets, one for training and another for verification. The percentage of data that was used was varied so as to check for the sensitivity of the two models to data splitting.

TABLE I: RMSE, MAE and Correlatin Coefficient.

Model (% Training set)	RMSE	MAE	Correlation Coefficient
MLP 66%	3.4286	1.9558	0.8248
MLP 75%	4.9377	3.2063	0.7368
MLP 80%	5.2154	3.6164	0.6971
M5P 66%	2.6659	1.2302	0.8936
M5P 75%	3.109	1.6581	0.869
M5P 80%	7.696	4.8665	0.5037

The data percentages included 66%, 75%, and 80% of the total, these data was used for training and the rest for verification. The M5P Model Tree developed with 66% training set was realized to be the best model that predicted flow with a RMSE of 2.666, and a correlation coefficient of the measured and the predicted flow of 0.89. This is explained by the fact that this method is a dynamic committee machine with leaf models that are specialized in particular areas of the input space, this was also realised in [6] in rating curve estimation. M5P MT with 80% training set had the worst performance, this shows that M5P MT is sensitive to data splitting. The predictive accuracy of the M5P model was observed to be better than that of an ANN model built with the same data. Fig. 4a-4c show results for the verification models with different training sets when the M5P model tree is used while Fig. 5a-5c show results for the verification

models with different training sets when the MLP-ANN was employed. The different performance measures for numeric prediction are given in Table I.

A MLP-ANN with 4 hidden nodes predicted flows with a RMSE ranging from 3.42 to 5.22. This agrees with [8] who concluded that standard multi-layer; feed-forward networks are capable of approximating any measurable function to any desired degree of accuracy, although errors in approximation can arise due to inadequate learning, insufficient number of hidden units or the input output relationship being insufficiently deterministic. In this case the challenges of overestimation and underestimation of peak flows, as well as low flows, and shifting of the hydrographs, either the peaks arriving early like in the M5P MT 80% training set could be attributed to limited data for training, thus inadequate learning.

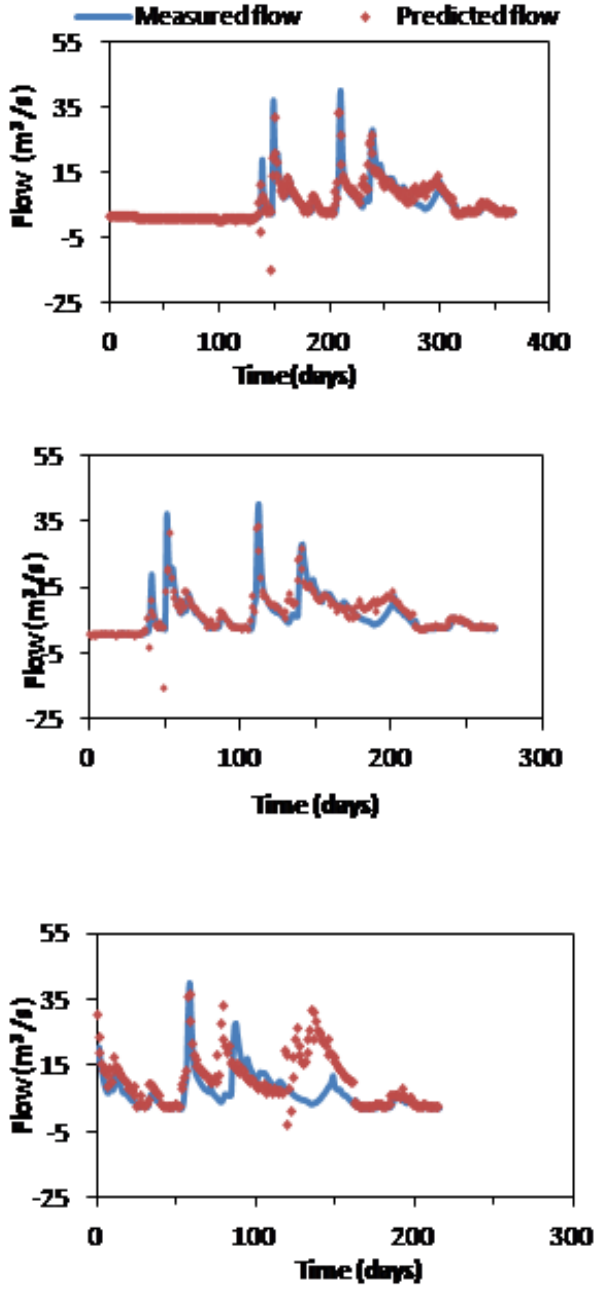

Fig. 4. Verification results for M5P model tree (a) 66% (b) 75% and (c) 80% of training data at Mhinga station. The continuous line represents measured flow and the diamond represents predicted flow.

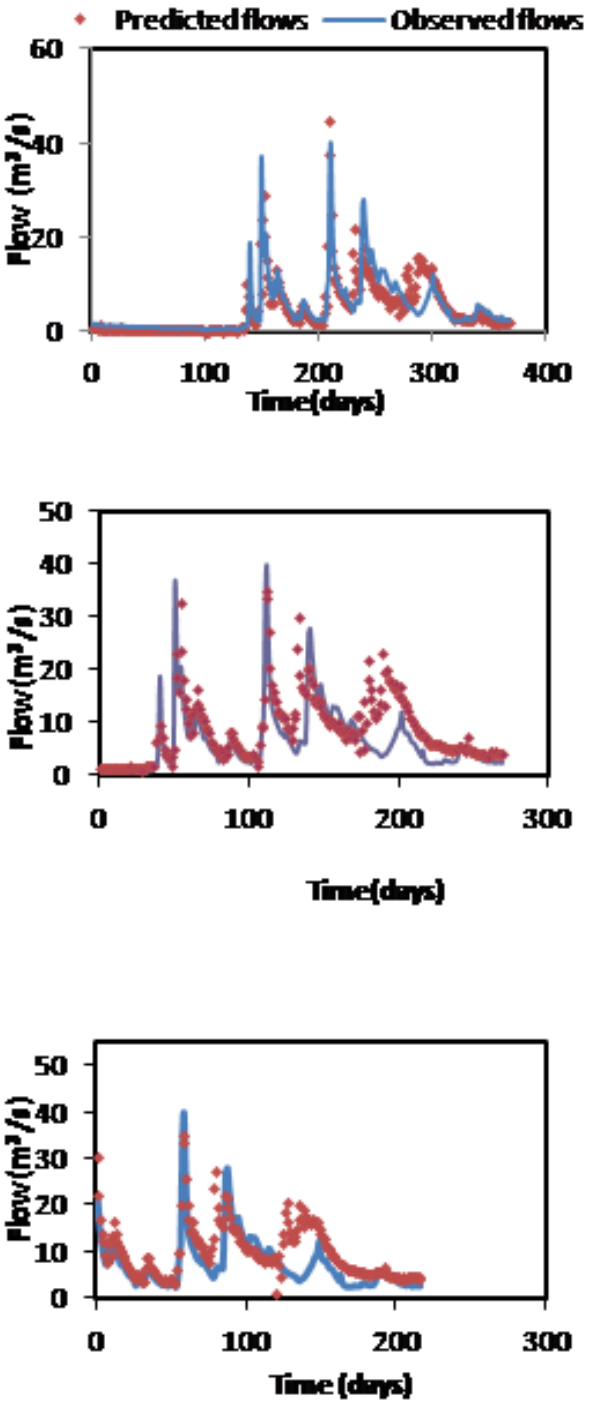

Fig. 5. Verification results for MLP –ANN with (a) 66% (b) 75% and (c) 80% of training data at Mhinga station. The continuous line represents measured flow and the diamond represents predicted flow.

V. Conclusion and Recommendations

The main aim of this study was to show the ability of a multilayer perceptron artificial neural network and a model tree M5P in predicting streamflow. The application of these two techniques to the Luvuvhu River at Mhinga station in South Africa has shown the possibility of using available data in a given catchment to predict streamflow, keeping in mind that data intensive models may not be successfully used especially in the developing countries where data is inadequate, as concluded in [15]. Both MLP-ANN and M5P model tree were found to predict flows significantly well,

despite insufficient training data. M5P MT has a better predictability when compared to MLP-ANN built with the same data. However, M5P MT seems to be more sensitive to data splitting. Machine learning techniques are said to be data dependent, and perform satisfactorily when long data series are available. A further study on the application of the same techniques to other catchments with relatively long data series should be carried out to reasonably compare the performance of the models in water resources.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of the South African Weather Service (SAWS) for providing the rainfall data and the Department of Water Affairs DWA of South Africa for making available the streamflow measurements that were used in this study.

REFERENCES

[1] T. Wagener, H. S. Wheater, and H. V. Gupta, *Rainfall-Runoff Modelling in Gauged and Ungauged Catchment*, London: Imperial College Press,2008.

[2] S. Lingireddy, D. Ramalingam, and J. Pavoni, "ANNs as function approximation tools- a case study," in *Artificial Neural Networks in Water Supply Engineering*, B. G. Lingireddy, Virginia: American Society of Civil Engineers, 2005, pp. 160-170.

[3] A. W. Minns and M. J. Hall, "Artificial neural networks as rainfall runoff models," *Hydrological Sciences Journal* , vol. 3, pp. 399-417, June 1996.

[4] C. W. Dawson and R. Wilby, "An artificial neural network approach to rainfall runoff modelling," *Hydrological Sciences*, vol. 43, no. 1, pp. 47-66, 1998.

[5] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Redwood city, California, USA.: Addison Wesley.1991.

[6] B. Bhattacharya and D. P. Solomatine, "Neural networks and M5P model trees in modelling water level- discharge relationship," *Neurocomputing* ,vol. 63, pp. 381-396, 2005.

[7] D. F. Lekkas, C. Onof, M. J. Lee, and E. A. Baltas, "Application of artificial neural for flood forecasting," *Global Nest: The International Journal* ,vol. 6, no. 3, pp. 205-211, 2004.

[8] K. Hornik, M. Stinhcombe, and H. White, "Multi layer feed forward networks are universal approximators," *Neural Networks* ,vol. 2, pp. 359-366, 1989.

[9] J. R. Quilan, "Learning with continous classes," in *Proc 5 th of the Australian Joint Conference on Artificial Intelligence*, Singapore, 1992, pp. 343-348.

[10] D. P. Solomatine, "Optimisation of hierarchical modular models and M5 trees," in *Proc. of International Joint Conference on Neural Networks.* Budapest, Hungary, 2004.

[11] S. Lingireddy and G. M. Brion, "Artificial neural networks in water supply engineering," in *Artificial Neural Networks in Water Supply Engineering*, G. Brion, S. Lingireddy, Virginia, USA: American society of Civil Engineers, 2005, pp. 1-9.

[12] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Technique*, 2nd ed., San Francisco, USA: Morgan Kaufmann Publishers, 2005.

[13] H. I. Mohamad and X. Cai, "Input variable selection for water resources systems using modified minimum redundancy maximum relevance (mMRMR) algorithm," *Advances in water resources*, vol. 32 , no. 4, pp. 582-593, April 2009.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max dependecy, max relevance and min-redundancy," *IEEE Transactions on pattern analysis and machine Intelligence* , vol. 27, no. 8, pp. 1226-1238, 2005.

[15] P. A. Kagoda, J. Ndiritu, C. Ntuli, and B. Mwaka, "Application of radial basis function neural networks to short term streamflow forecasting," *Journal of Physics and Chemistry of the Earth*, in press, corrected proof, 2010.