COPYRIGHT NOTICE

This is the peer-reviewed version of the following article:

Choudhury, T., Wei, J., Barton, A., Kandra, H., Aziz, A. (2018) Exploring the application of artificial neural network in rural streamflow prediction – A feasibility study. 27th IEEE International Symposium on Industrial Electronics, ISIE 2018; Cairns, Australia; 13th-15th June 2018 Vol. 2018-June, p. 753-758.

Which has been published in final form at:
https://doi.org/10.1109/ISIE.2018.8433644

Exploring the Application of Artificial Neural Network in Rural Streamflow Prediction – A Feasibility Study

T. A. Choudhury [1*], Jackie Wei [2], Andrew Barton [3], Harpreet Kandra [4], Abdul Aziz [5]

School of Engineering and Information Technology [1,2,3,5]
Federation University Australia
Victoria, Australia
t.choudhury@federation.edu.au [1*],
jackiewei@students.federation.edu.au [2],
a.barton@federation.edu.au [3]
h.kandra@federation.edu.au [4]

GWM Water [5]
Horsham, Victoria, Australia
abdul.aziz@gwmwater.org.au [5]

***Abstract*— — Streams and rivers play a critical role in the hydrologic cycle with their management being essential to maintaining a balance across social, economic and environmental outcomes. Accurate streamflow predictions can provide benefits in many different ways such as water allocation decision making, flood forecasting and environmental watering regimes. This is particularly important in regional areas of Australia where rivers can play a critical role in irrigated agriculture, recreation and social wellbeing, major floods and sustainable environments. There are several hydrological parameters that effect stream flows in rivers and a major challenge with any prediction methodology, is to understand these parameter interdependencies, correlations and their individual effects. A robust methodology is, thus, required for accurate prediction of streamflow under usually unique, waterway-specific conditions using available data. This research employs an approach based on Artificial Neural Network (ANN) to provide this robust methodology. Data from readily available sources has been selected to provide appropriate input and output parameters to train, validate and optimise the neural network. The optimisation steps of the methodology are discussed and the predicted outputs are compared and analysed with respect to the actual collected values.**

Keywords—artificial neural network, intelligent multi-variable control, real time monitoring, streamflow prediction, hydrological circle, back propagation

I. Introduction

Streamflow is a critical hydrological parameter for the planning and management of water resources. Indeed, an understanding of streamflow characteristics is central to water security. For much of Australia, the hydrological variability of rural streamflow is a key challenge with the statistical properties of streamflow data typically highly skewed, with historical data sets sometimes short and discontinuous [1]. The accurate prediction of streamflow can therefore be very difficult with limited data sets available from which to learn from. Furthermore, forecasts of streamflow are closely related to forecasts of weather which are notoriously difficult especially in the light of climate change. However, the potential benefits from accurate streamflow prediction is very significant with improved economic, social and environmental outcomes very clear. Examples include more precise irrigation agricultural practices, avoidance of water restrictions for towns and communities, greater certainty around environmental watering and waterway health, and more accurate information around the magnitude of floods and the movement of floods through riverine systems. .

In hydrology, there are known to be several factors, ranging from both natural and human induced, that affect streamflow behaviour [2]. These parameters include, but not limited to, (i) precipitation from rainfall and snowmelt, (ii) evaporation, (iii) temperature, (iv) interaction between surface water and groundwater, (v) soil moisture content, (vi) humidity, and (vii) fire. While these factors have different levels of non-linear effects to streamflow, they are unpredictable in nature across the year. The task of streamflow prediction therefore becomes very difficult due to non-linearity as well as variability of the hydrologic and associated weather processes.

In this preliminary study, Artificial Neural Network (ANN) is proposed to model the hydrological process and predict the streamflow. ANN is a non-linear predictive data modelling tool that works by learning the process dynamics to establish correlations between outputs and selected input parameters. The generalisation ability of ANN allows the developed model to predict values for any input conditions within limits of the training boundaries. ANN further provides a flexible approach to streamflow prediction. Conventional regression and statistical methods requires accurate geometric series of data for the models to predict – which might not be available in rural conditions.

ANN has been successfully applied in similarly complex problems and across a wide range of water engineering applications in the past. Some of these applications include: (i) flood and river flow forecasting [3, 4], (ii) modelling rating curves [5], (iii) predicting sewerage capture efficiency [6]. Several authors, previously, have studied the use of

ANN in streamflow forecasting [7-10]. However, there has been no study available on the use of ANN to predict streamflow in Australian rural river systems – the geography of which makes the modelling challenging.

To test the feasibility of ANN in this study as an appropriate methodology for streamflow prediction, input variables have been limited to just two parameters being daily total rainfall and current river height (or daily maximum water level depth). These two parameters have been chosen on the a-priori assumption that they are dominant predictors and contributors to estimating future streamflow for rural catchments [2, 11]. Future work will refine this approach and look to broaden the selection of input parameters.

To study the viability of the application of ANN in streamflow forecasting under such conditions, the research deliberately: (i) selects a highly variable streamflow site and (ii) simplifies the method in only using the minimum number of input variables to produce reliable and useful results.

II. Simulation Model

A. Database Collection and Processing

Data collection is a critical step in the development of an ANN model as the reliability and robustness of the model depends directly on the quality and accuracy of the collected data. For the purpose of training, optimising and testing the proposed ANN model, a database (DB) that consisted of the daily total rainfall and daily maximum water level depth as the input parameters (or predictors) was prepared. The daily total discharge was used as the output parameter (or predictand) in DB. The daily total discharge is a direct indicator of the streamflow.

The streamflow site selected for data collection was the Genoa River, and for reasons of having a reasonable length and continuous set of historical data. The Genoa River starts in southern New South Wales (NSW) and flows southwards into Victoria where it joins the Wallagaraugh River (which also starts in NSW). Both rivers flow into Mallacoota Inlet before entering Bass Strait at Mallacoota [12]. The catchment has an area of 126 square kilometres with adjacent land used by the local council, cattle grazing as well as different environment activities. The Genoa River is unregulated without any major water storage or major diversions, with local consumption relying on natural flows or small structures such as weirs for their water supply [13].

The streamflow and related data were collected from the Department of Environment, Land, Water and Planning (DWELP) under the State Government of Victoria [14]. The data collected also consists of daily recorded data obtained from the Bureau of Meteorology (BoM) [15]. The data range lies from 1993 to 2016, with a total data set of 7,652.

Before further use the database (DB) was normalised using (1) [16]. This linear transformation ensures that all process parameters are treated equally by the ANN and, thus, avoids calculation error relating to different parameter magnitudes.

$$X_{NORM} = \frac{X - X_{MIN}}{X_{MAX} - X_{MIN}} \tag{1}$$

In (1), X_{NORM} represents the normalised input parameter value and X the actual parameter value. X_{MAX} and X_{MIN} are the maximum and minimum values contained within the dataset.

The maximum and minimum of each input and output variables are given in Table I. These value limits also represent the extent to which the network will be trained and can be tested for performance.

TABLE I. PARAMETER PHYSICAL LIMITS

Variable	Lower Limit	Higher Limit
Daily Total Rainfall (mm)	0.00	185.80
Daily Maximum Water Level Depth (m)	0.06	2.78
Daily Total Discharge (ML/d)	0.00	19,899.62

B. Defining ANN Model

The ANN consists of a mathematical model of a group of interconnected artificial neurons and is used as a non-linear statistical data modelling tool. ANN can be implemented to model complex and non-linear process relationships between the input and output parameters, particularly where there are process variability, fluctuations and with little known prior information.

ANN takes a connectionist approach to computation where the strength of each connection between the neurons is represented by the term ‘weight’ [17]. These weights form the basis to process generalisation and evaluation of input and output parameter relationships. This can be achieved by proper optimisation of the weight matrix, and is achieved through a training process. The process of tuning the weight matrix is called a paradigm. The most powerful paradigm is the back-propagation paradigm, which is widely used and also employed for this study [18].

The back-propagation paradigm used in this research is the Levenberg-Marquardt (LM) algorithm [19]. Standard back-propagation algorithms are: (i) very slow, (ii) require a lot of off-line training, and (iii) suffer from temporal instability as they tend to get stuck to the local minima [20]. Given a network of not more than few hundred weights (which complies with the current model), the LM algorithm proves to be more efficient in comparison to the conjugate gradient and variable learning rate types algorithms [19].

Further to overcome the complex non-linear relationship between the output streamflow (daily total discharge) and input hydrological parameters, this research utilises a simple neural network model based on multi-layer perception (MLP).

The MLP ANN architecture consists of three main components: (i) input layer (daily total rainfall and daily maximum water level depth), (ii) output layer (daily total discharge), and (iii) hidden layer.

The number of neurons required in the ANN model to describe each physical parameter within the process it is trying to model, is dependent on the parameter nature. A real valued parameter can be represented by only one neuron, while a parameter representing classifications requires x neurons to describe 2^x categories [21]. All the hydrological input and output parameters (listed in sub-section A) are real valued and are, thus, described by one neuron each.

The hidden layer within ANN structure contributes most in establishing the process generalisation and parameter correlations. The number of hidden layers and number of neurons in each hidden layer is determined from the network optimisation process. The ANN architecture used in this research is presented in Fig.1.

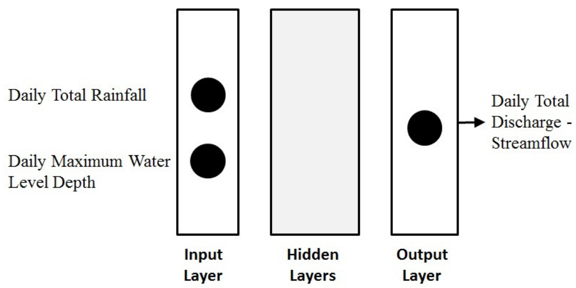

Fig. 1. The artificial neural network (ANN) architecture

C. ANN Model Development

ANN model development process involves training, testing and optimisation steps and uses a constructed database (DB). The process of network optimisation determines the number of hidden layers; number of neurons in each hidden layer in conjunction with the optimisation of the weight population to produce the lowest error performance function. All the ANN simulations in this study were performed with MATLAB [22].

To begin with, an assumption is made on the number of hidden layers, the number of neurons in each hidden layer and the initial weight population. The network response of the stream flow (total daily discharge) value was then computed and compared with the actual values to obtain the error performance function. Based on the comparisons, the network weight matrix is re-computed and optimised further. This process of network training is run for a number of iterations or *epochs*.

The model is then trained for a sufficient amount of time so as to optimise the network and generalise the process underpinning the input and output parameters. The network will fail to learn the function if the number of *epochs* is set too low; but if it is set too high, the network tends to memorise the training data instead of generalising the function (over-fitting). The maximum number of epoch set in this study is 1000 with the log-sigmoid as the transfer function in all layers.

For the network training, optimisation and testing, the database DB is divided into three sets: (i) 60% of the data for network training (DB-TRN), (ii) 20% of the data for network validation (DB-VAL), and (iii) 20% for network testing (DB-TST). The three basic steps to ANN model development are briefly described below.

(i) *Training step*: trains the network using the training set to obtain the optimised weight matrix producing minimum error between the predicted and actual output values.

(ii) *Validation step*: tests the trained network with the validation set to prevent network over-fitting.

(iii) *Test step*: simulate the trained network with a test data set which has previously been unseen to the network. If the residual error between the predicted and actual output parameter is computed to be equivalent to the training step, it can be concluded that proper generalisation is achieved and the networks are ready to predict intermediate conditions.

For the test and validation step, the weight population and other network parameters are kept static as the network is being tested and validated with the input data and checked for prediction and network performance.

Correlation coefficient (R) and Mean Absolute Error (MAE) are selected as the model's performance measuring functions. The correlation coefficient, R, values on the test set provides an understanding of how well the trained network's response, to the unseen inputs, fits with the respective collected daily total discharge (streamflow) value. Larger R values represents better fit indicating stronger model performance in generalising the process dynamics. The MAE generated by the network on the test set provides a measure of the "generalisation error" of the trained network models. The lower this error value, better is the network's ability to generalise process and predict streamflow with sufficient accuracy under unknown conditions.

The performance of a trained network is sensitive to the size of the hidden layers. To find the optimal number of hidden layers to overcome the non-linearity associated with the hydrological process considered in this study, the number of hidden layers were varied from one to three. In each case, the number of neurons in each layer were also varied to find the optimal number of neurons under each condition.

For one hidden layer, the number of hidden layer neurons were varied from one to twenty (1 to 20) with increments of one neuron each time. For two hidden layers, the number of hidden layer neurons were varied from two and one to twenty and nineteen neurons, respectively, in each hidden layer (2-1 to 20-19) with increments of one neuron in each layer. For three hidden layers, the number of neurons were varied from 20-19-18 to 3-2-1 with increments of one neuron in each layer. The network trainings were repeated ten times. In each case, and the network generating maximum correlation coefficient, R, value on DB-TST, was stored and saved.

III. Results Analysis and Discussion

A summary of the network performance, in terms of R-value and generalisation error (MAE), for the networks with one, two and three hidden layers are presented in Table II. The standard deviation of the network performances, over different number of hidden layer neurons, are also presented in the table.

TABLE II. Performance Comparison of ANN Models With Different Hidden Layers

		Maximum Correlation Coefficient (R)	Standard Deviation - R	Minimum Mean Absolute Error (MAE)	Standard Deviation – MAE
1 Hidden Layer	***Training Set***	0.9952	0.0271	0.0033	0.0028
	Test Set	0.9968	0.0276	0.0034	0.0029
2 Hidden Layer	***Training Set***	0.9872	0.0147	0.0063	0.0021
	Test Set	0.9907	0.0109	0.0064	0.0022
3 Hidden Layer	***Training Set***	0.9971	0.0027	0.0021	0.0011
	Test Set	0.9981	0.0017	0.0021	0.0011

The maximum R-value and the minimum generalisation error on both the training and test dataset was generated by the ANN model with three hidden layers. On the test dataset (DB-TST), the standard deviations of R-values of all the networks with three hidden layers was found to be 84% lower in comparison to that of the network with two hidden layers and 94% lower in comparison to the network with one hidden layer. On the training dataset (DB-TRN), the standard deviations of R-values of all the networks with three hidden layers was found to be 82% lower to that of the network with two hidden layers and 90% lower to that of the network with one hidden layer. Similar trend was found with the standard deviation of the generalisation error produced by ANN models with three hidden layers in comparison to the ANN models with two and one hidden layers.

The lowest standard deviation of the performance parameters (R and MAE) of the ANN models with three hidden layers indicates greater stability and robustness of the developed ANN models. Although higher number of hidden layers increased the complexity of the network models, it was found to be necessary for ANN to generalise the process and learn the underlying relationship between the input and output hydrological parameter relationship. For further analysis, the ANN network models with three hidden layer were, thus, chosen.

The correlation coefficient and generalisation error variations of all the trained ANN models with three hidden layers on the DB-TRN are presented in Fig. 2 and Fig. 3, respectively. The R-value in Fig. 2 is seen to rise and stabilise with the increase in hidden layer neurons. The trend is repeated in Fig. 3, with the generalisation error dropping and keeping relatively steady with the increase in hidden layer neurons. The maximum R-value of 0.9981 and minimum MAE value of 0.0021 was observed for the ANN model with 7-6-5 hidden layer neurons. The average R-value of the all the networks was computed to be 0.9968 with the average generalisation error of 0.0028.

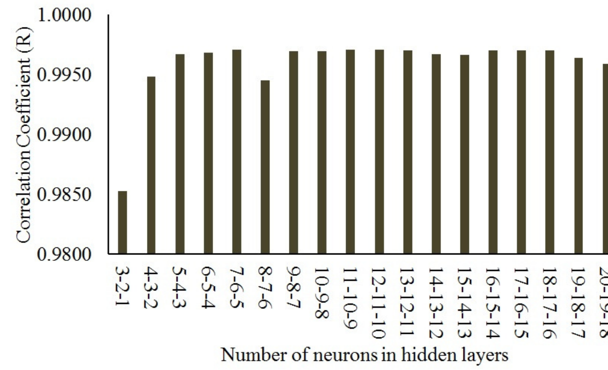

Fig. 2. Correlation coefficient (R) variations of different ANN models with three hidden layers on the training dataset (DB-TRN).

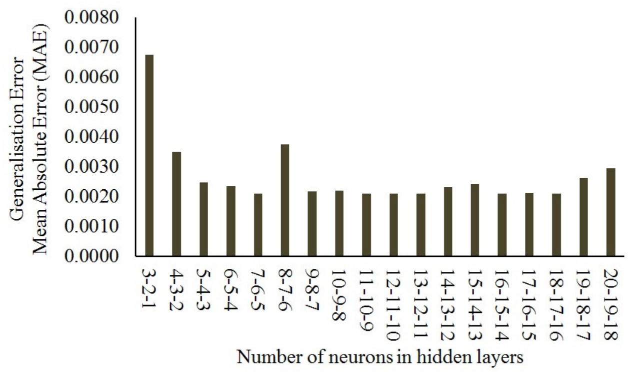

Fig. 3. Generalisation error variations of different ANN models with three hidden layers on the training dataset (DB-TRN).

The correlation coefficient and generalisation error variations of all the trained ANN models with three hidden layers on DB-TST are presented in Fig. 4 and Fig. 5, respectively. As evident, the trend in performance parameter variations of the ANN models on DB-TST was similar to that of the models on DB-TRN (Fig. 2 and Fig. 3). This presents robustness of successful network training. The maximum R-value of 0.9981 and minimum generalisation error value of 0.0021 was found out for the network with 7-6-5 hidden layer neurons. The ANN model with 7-6-5 was therefore selected for further analysis and is referred to as NET.

The selected ANN model (NET) is used to simulate the entire dataset (DB) to view the overall performance of the model in predicting the streamflow (daily total discharge) under varying input conditions. The plot of the predicted values with the corresponding actual values (from DB) and relative error is presented in Fig. 6. The trend presented by the predicted streamflow was in strong alignment with the actual streamflow variations from DB. The magnitude of the average relative error of the predicted values with respect to the actual values was 0.04%. The NET's performance on DB aligns with the training and testing performance presented in Fig. 2 to Fig. 5.

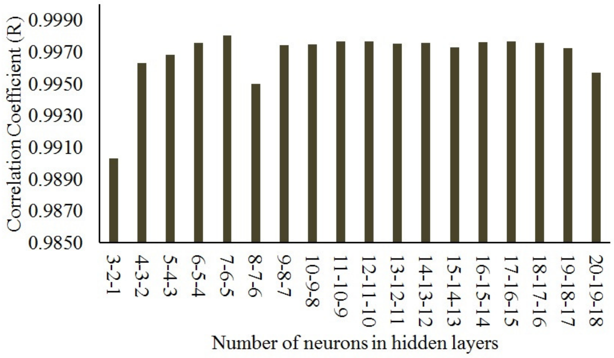

Fig. 4. Correlation coefficient (R) variations of different ANN models with three hidden layers on the test dataset (DB-TST).

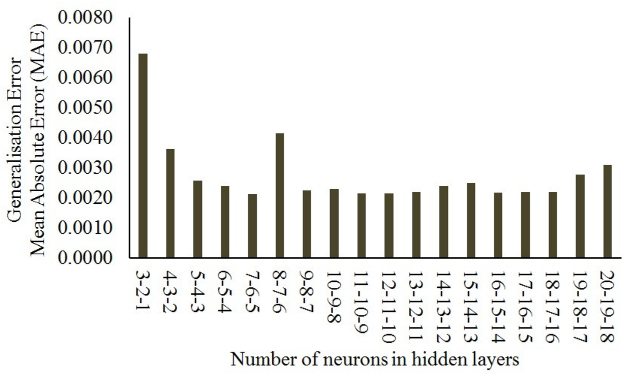

Fig. 5. Generalisation error variations of different ANN models with three hidden layers on the test dataset (DB-TST).

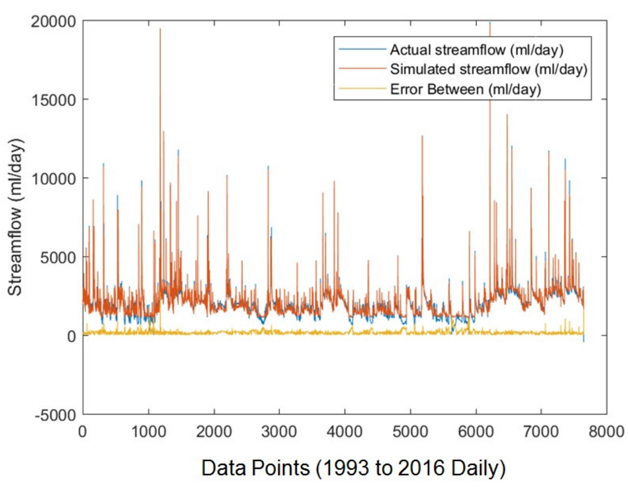

Fig. 6. Performance of selected ANN model (NET) on DB.

The accumulated streamflow (daily total discharge) value over the entire range of DB is presented in Fig. 7. The Figure presents a comparison of the actual and predicted cumulative data over time indicating total storage. The deviations between the plots are useful in determining any systemic under or over prediction by the ANN model. The plot of predicted streamflow (simulated output) is shown to be very close the actual streamflow data. The result further demonstrates the feasibility of application of ANN as a prediction tool.

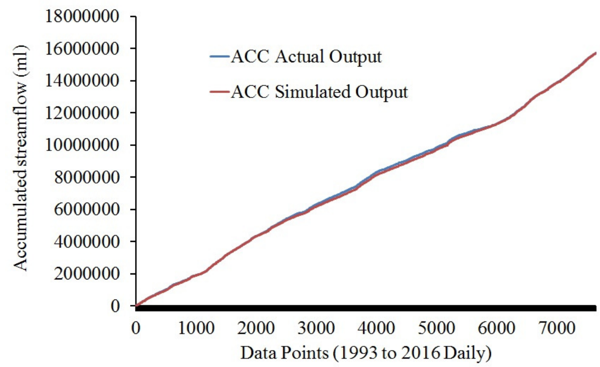

Fig. 7. A comparison of actual and predicted cumulative streamflow.

An application of this work, for example, would be during times of flood where a key monitoring station has lost communications or been damaged, and the ANN model is able to reliably predict expected peak flow rates. This offers redundancy during times of emergency and provides assurances to decision makers.

IV. Conclusion

Accurate streamflow predictions would provide many benefits such as improved water allocation decision making, more timely flood forecasting and environmental watering regimes. This paper presents the preliminary study of an ANN model developed to predict streamflow in rural Australian water catchments.

An Artificial Neural Network (ANN) model has been developed to predict streamflow using the Levenberg-Marquardt back-propagation algorithm and using steps of training, testing, validation and optimisation. After comparison of network performance with different hidden layers and neuron numbers, the ANN model, NET, with the most favourable correlation coefficient and mean absolute error was selected. The selected network, NET, showed strong performance in predicting the streamflow (the daily total discharge) under varying input conditions by learning the underlying relationships between the hydrological parameters presented. Results further provide applicability of ANN technique under availability of limited number of input parameters – rainfall and water level.

Future work will explore the use of additional input parameters such as soil moisture and relative humidity, and focus on particular forecast windows. Seasonal to sub-seasonal (S2S; approximately two months to two weeks) will be a particular focus for this work with applications being extended towards flood forecasting, improved water allocation and water resource system operations.

REFERENCES

[1] R. Grayson, R. Argent, R. Nathan, T. McMahon, and R. Mein, "Hydrological recipes. Estimation techniques in Australian hydrology. CRC for Catchment Hydrology, Dept," *Civil Engineering, Monash University,* 1996.

[2] R. C. Ward and M. Robinson, *Principles of Hydrology*: McGraw-Hill, 2000.

[3] D. Nagesh Kumar, K. Srinivasa Raju, and T. Sathish, "River Flow Forecasting using Recurrent Neural Networks," *Water Resources Management,* vol. 18, pp. 143-161, April 01 2004.

[4] R. R. Shrestha, S. Theobald, and F. Nestmann, "Simulation of flood flow in a river system using artificial neural networks," *Hydrology and Earth System Sciences Discussions,* vol. 9, pp. 313-321, 2005.

[5] A. Zeroual, M. Meddi, and A. A. Assani, "Artificial neural network rainfall-discharge model assessment under rating curve uncertainty and monthly discharge volume predictions," *Water Resources Management,* vol. 30, pp. 3191-3205, 2016.

[6] M. A. Aziz, M. Imteaz, T. A. Choudhury, and D. Phillips, "Applicability of Artificial Neural Network in Hydraulic Experiments Using a New Sewer Overflow Screening Device," *Australian Journal of Water Resources,* vol. 17, pp. 77-86, 2013/01/01 2013.

[7] K. Anusree and K. O. Varghese, "Streamflow Prediction of Karuvannur River Basin Using ANFIS, ANN and MNLR Models," *Procedia Technology,* vol. 24, pp. 101-108, 2016/01/01/ 2016.

[8] D.-H. Lee and D.-S. Kang, "The Application of the Artificial Neural Network Ensemble Model for Simulating Streamflow," *Procedia Engineering,* vol. 154, pp. 1217-1224, 2016/01/01/ 2016.

[9] G. Uysal, A. Şensoy, and A. A. Şorman, "Improving daily streamflow forecasts in mountainous Upper Euphrates basin by multi-layer perceptron model with satellite snow products," *Journal of Hydrology,* vol. 543, pp. 630-650, 2016/12/01/ 2016.

[10] G. Uysal, A. A. ſorman, and A. ſensoy, "Streamflow Forecasting Using Different Neural Network Models with Satellite Data for a Snow Dominated Region in Turkey," *Procedia Engineering,* vol. 154, pp. 1185-1192, 2016/01/01/ 2016.

[11] V. Chow, D. Maidment, and L. Mays, *Applied Hydrology*: McGraw Hill, 1988.

[12] (24 January 2018). *Genoa River: East Gippsland Catchment Management Authority*.
Available: http://www.egcma.com.au/rivers/160/

[13] (24 January 2018). *Water NSW*.
Available: https://www.waternsw.com.au/

[14] L. Department of Environment, Water & Planning. (30 November 2017).
Available: http://data.water.vic.gov.au/monitoring.htm

[15] B. o. M. A. Government. (30 November 2017).
Available: http://www.bom.gov.au/

[16] T. A. Choudhury, N. Hosseinzadeh, and C. C. Berndt, "Using artificial neural network to predict the particle characteristics of an atmospheric plasma spray process," in *IEEE International Conference on Electrical and Computer Engineering (IEEE-ICECE)*, Dhaka, 2010, pp. 726-729.

[17] M. M. Nelson and W. T. Illingworth, *A practical guide to neural nets*: Addison-Wesley Publishing Company Inc., 1991.

[18] S. E. Fahlman, "Faster-learning variations on back propagation: an emperical study," *Proceedings of the 1988 Connectionist Models Summer School,* pp. 38-51, 1988.

[19] M. T. Hagan and M. B. Mehnaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks,* vol. 5, pp. 989-993, November 1994.

[20] A. J. Adeloye and A. De Munari, "Artificial neural network based generalized storage-yield-reliability models using the Levenberg-Marquardt algorithm," *Journal of Hydrology,* vol. 326, pp. 215-230, 2006.

[21] S. Guessasma, Z. Salhi, G. Montavon, P. Gougeon, and C. Coddet, "Artificial intelligence implementation in the APS process diagnostic," *Materials Science and Engineering B: Solid-State Materials for Advanced Technology,* vol. 110, pp. 285-295, 2004.

[22] "MATLAB," R2017b ed: The Math Works Inc., Natick, MA, USA