Prediction of River Discharge by Using Gaussian Basis Function

By

Nur Farahain Bt Mohd Idrus

14390

Dissertation submitted in partial fulfilment of
the requirements for the

Bachelor of Engineering (Hons)

(Civil)

MAY 2014

Universiti Teknologi PETRONAS,
Bandar Seri Iskandar,
31750 Tronoh,
Perak Darul Ridzuan.

Table of Contents

List of Figures

List of Tables

ABSTRACT

For design of water resources engineering related project such as hydraulic structures like dam, barrage and weirs river discharge data is vital. However, prediction of river discharge is complicated by variations in geometry and boundary roughness. The conventional method of estimation of river discharge tends to be inaccurate because river discharge is nonlinear but the method is linear. Therefore, an alternative method to overcome problem to predict river discharge is required. Soft computing technique such as artificial neural network (ANN) was able to predict nonlinear parameter such as river discharge. In this study, prediction of river discharge in Pari River is predicted using soft computing technique, specifically gaussian basis function. Water level raw data from year 2011 to 2012 is used as input. The data divided into two section, training dataset and testing dataset. From 314 data, 200 are allocated as training data and the remaining 100 are used as testing data. After that, the data will be run by using Matlab software. Three input variables used in this study were current water level, 1-antecendent water level, and 2-antecendent water level. 19 numbers of hidden neurons with spread value of 0.69106 was the best choice which creates the best result for model architecture after numbers of trial. The output variable was river discharge. Performance evaluation measures such as root mean square error, mean absolute error, correlation of efficiency (CE) and coefficient of determination (R^2) was used to indicate the overall performance of the selected network. R^2 for training dataset was 0.983 which showed predicted discharge is highly correlated with observed discharge value. However, testing stage performance is decline from training stage as R^2 obtained was 0.775 consequently presence of outliers have affect scattering of whole data of testing and resulted in less accuracy as the R^2 obtained much lower compared to training dataset. This happened because less number of input loaded into testing than training. RMSE and MSE recorded for training much lower than testing indicated that the better the performance of the model since the error is lesser. The comparison of with other types of neural network showed that Gaussian basis function is recommended to be used for river discharge prediction in Pari river.

CHAPTER 1: INTRODUCTION

1.1 Background

River discharge is the volume of water which flows through it in a given time. It is usually measured in cubic meters per second. The volume of the discharge will be determined by factors such as climate, vegetation, soil type, drainage basin relief and the activities of man. Major river water uses such as sources of drinking water supply, irrigation of agricultural lands, industrial and municipal water supplies, navigation, fishing, boating and body-contact recreation and aesthetic value. River discharge data is required in different water resources engineering related project such as design of hydraulic structures like dam, barrages, weirs. River discharge prediction is an essential tool to ensure proper management of water resources and the optimal distribution of water to consumers. One of most complex hydrologic phenomena to comprehend is prediction of river water discharge. This is because of the tremendous spatial and temporal variability of watershed characteristics and precipitation patterns, and the number of variables involved in the modelling of the physical processes.

Fekete and Vorosmarty (2002) defined one important way to authorize components of hydrological models is to compare predicted and observed runoff, the final computed as river discharge at gauging station divided by upstream contributing catchment area. Discharges can be measured more accurately than other constituents of the land-based energy and water cycles except temperature. Accurate estimation of runoff from a given rainfall event and an accurate hydraulic model for a given discharge is needed for river discharge prediction via conventional method. Basically, there is two techniques for river flow forecasting which is conventional method and soft computing technique. Conventional method tends to be inaccurate because it is linear. When compared to soft computing technique such as artificial neural network (ANN), Gene Expression Programming (GEP) and Support vector machine (SVM), they are able to predict nonlinear function such as river discharge. Thus, Pari River, Perak which are situated at Kinta River catchment is chosen as study area to predict river discharge using Gaussian radial basis function artificial neural network in this study.

Bustami et al., (2006) claimed ANN be able to generalize result from unseen data and well-suited in modelling dynamic systems on a real-basis. The artificial network approach have rigorously used in the water resources literature. The ANN have been extensively used in hydrology for simulating rainfall-runoff and other hydrological process (Demiral et al.,2008). This method is carried out by signals are passed between nodes through connection links.

Hydrol (2008) claimed that ANNs are robust tools for modeling many of the nonlinear hydrologic processes such as rainfall-runoff, stream flow, ground-water management, water quality simulation, and precipitation. He stated that artificial neural networks tend to be very data severe, and there seems to be no recognized methodology for design and effective implementation. Among the three ANN models which are a feed forward back propagation (BP), a radial basis function (RBF) and an adaptive neural network-based fuzzy inference system (ANFIS) employed, RBF was used because RBFs allow for a straightforward interpretation of the internal representation produced by the hidden layer. The main features of RBF are they are two-layer feed-forward networks, the hidden nodes implement a set of radial basis functions, the output nodes implement linear summation functions. The neurons in the hidden layer contain Gaussian transfer functions whose outputs are inversely proportional to the distance from the centre of the neuron. The training is very fast and to conclude the networks are very good at interpolation.

1.2 Problem Statement

River discharge measurements is resource intensive and time consuming exercise. The common method of estimation is by establishing conventional ground-based methods. Nevertheless, the method limited only for relatively large rivers and that, even for these, measurement accuracy will be considerably lower than is possible with ground-based measurements (Dingman and Bjerklie, 2006). Consequently, an alternative method is required to approximate nonlinear functions, and therefore become useful tools for handling water resources problems such as rainfall forecasting, stream flow forecasting, water level forecasting, and applications in urban drainage systems. In this study, the prediction of river discharge would be based on Gaussian radial basis function neural network modelling since it has not been used yet in previous studies.

1.3 Objectives

The main objective of this study is to predict river discharge in Pari River with the following specific objectives:

1) To predict river discharge of Pari River, Perak using Gaussian function
2) Performance evaluation of Gaussian function by using various statistical measures

1.4 Scope Of Study

In the proposed study author will described about prediction of river discharge by using Gaussian radial basis function (RBF). Prediction of river discharge in Pari River was performed using hydrological data such as discharge, rainfall, and river water level. The scope of study can be described as below:

1) The study is limited to predict river discharge in Pari River by developing radial basis function neural network using Gaussian function.
2) To evaluate the performance of the prediction model using statistical measures such as root mean square error (MSE), mean absolute error and correlation coefficient.

CHAPTER 2: LITERATURE REVIEW

2.1 Prediction of River Discharge using Conventional Method

From ancient times, soft computing techniques have been developed for the solution of hydrological field problems. The motive is that many existent world hydrological problems can barely be solved by means of conventional techniques because needed information is not available or the systems under consideration are not well defined. Conventional methods are used to assess discharge capacity at low depths, when the flow is only in the main channel. Though, the classical formulae for discharge capacity estimation do not yield reliable solutions when overbank flow occurs. It may lead either to overestimation of discharge capacity, which is dangerous, or to underestimation of capacity, which may cause waste of resources (Ozbek et al., 2004). Divided channel method (DCM), is one of conventional methods to predict river discharge. The compound cross section is divided into relatively large homogeneous sub-areas. Vertical (V), horizontal (H) and diagonal (D) imaginary interface planes were considered between the main channel and the floodplain subsections.

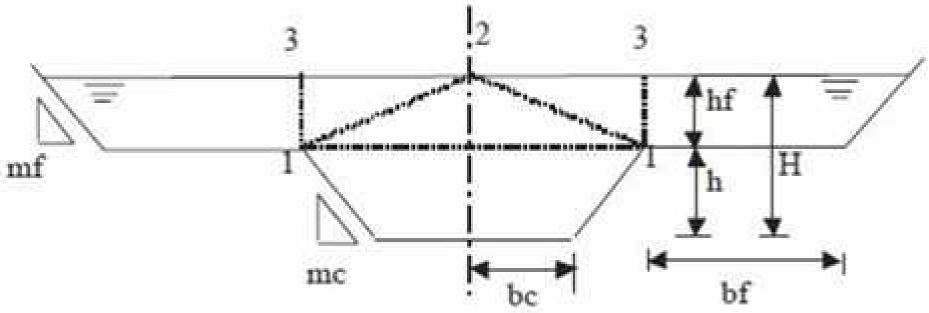


FIGURE 1: Compound channel cross-section with horizontal (H), diagonal (D) or vertical (V) planes shown as 1-1, 1-2 and 1-3, respectively (Wormleaton and Merrett, 1990

Another conventional method available is regression model that relate observed peak discharge to some measure of the impounded water volume: depth, volume, or some combination (Walder and O'Connor, 1997). Computer implementation such as computational models commonly account for hydraulic limits not reflected by regression relations but may be cumbersome to practice and usually need information that may be poorly known.

Belayneh et al.,(2014) had compared autoregressive integrated moving average models (ARIMA) to artificial neural network. It have been the most extensively used stochastic models for hydrologic drought forecasting. Stochastic models are linear models and are limited in their ability to forecast non-linear data. Both regression models and autoregressive integrated moving average (ARIMA) models are typical models for statistical time series methods for forecasting. Nevertheless, they are basically linear models assuming that data are static, and have a limited ability to capture non-statics and nonlinearities in the hydrologic data.

According to Campolo et al., (2009), river discharge prediction is useful in order to achieve objective to flood control and its mitigation, water supply for municipal and industrial uses, water quality control and power production optimization. For flood control and mitigation, prediction of river discharge is crucial in order to prevent flood as it can be such devastating disasters that anyone can be affected at almost any time. Hidayat et al., (2014), defined mitigation of disasters such as floods and droughts and water allocations for example drinking water and irrigation are among water-related issues that require reliable streamflow data and predictions. Similar to many other hydrological processes that exhibit a high degree of nonlinearity, conflicting spatial and temporal scales, and uncertainty in parameter estimates, estimation of streamflow and forecasting is still a great challenge.

In order to deal with the issue, artificial neural network (ANN) is proposed as one of the computational tools suitable. Based on his study, to represent catchment processes, a physically based model is superior to use particularly even though the performance is quite low because available data are limited. Furthermore, ANN can produce a somewhat precise prediction that can be used as a tool for working watershed management by using past data only.

2.2 Artificial Neural Networks (ANNs)

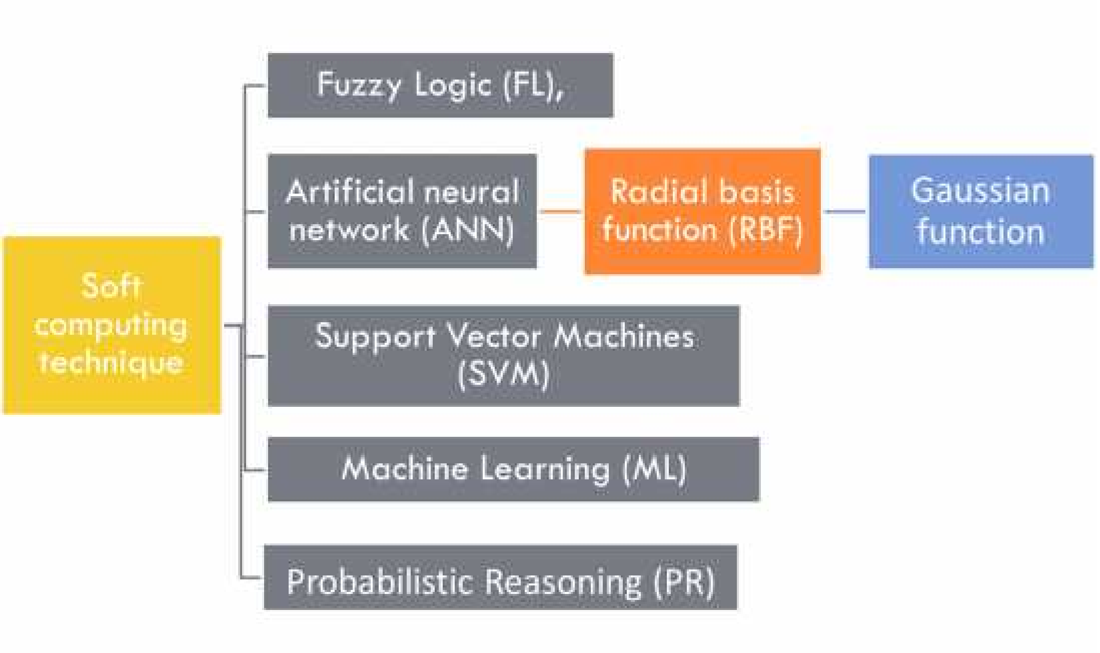

Soft computing technique
Fuzzy Logic (FL),
Artificial neural network (ANN)
Radial basis function (RBF)
Gaussian function
Support Vector Machines (SVM)
Machine Learning (ML)
Probabilistic Reasoning (PR)

FIGURE 2: Soft computing technique

:

There are numerous soft computing techniques such as multiple regression, aritificial neural network, genetic programming, and adaptive neuro-fizzy interence system (ANFIS). In this study, artificial neural network will be used. According to Kalagirou et al.,(2014), a neural network is a enormously equivalent distributed processor. It are able to store experiential knowledge and making it available for use since it contains a natural propensity.

ASCE (2000) indicated that ANNs have initiate a number of applications in the area of water quality modeling factors such as flow rate, contaminant load, medium of transport, water levels, initial conditions and other site-specific parameters influenced the water quality. They claimed that ANN application is suitable because the estimation of such variables such as river discharge is complex to carry out. When compared to previous studies, ANN also was used for different types of modelling factors. In recent years, Artificial Neural Network (ANN) is well known for prediction and forecasting in various fields. The Artificial Neural Network concept was first proposed by McCulloch and Pitts in 1943 (Ruslan et.al,2014). Besides that, ANN also acknowledged as one of the most popular black-box models. Fundamentally, black-box model is a computer program into which users enter information and the system utilizes pre-programmed logic to return output to the user. From the previous study relating flood prediction, ANN is proven to be a great tool in producing pleasingly results in hydrological fields.

Gazzaz et.al.,(2012) proposes artificial neural network (ANN) modeling has the potential to reduce the computation time and effort and the possibility of errors in the calculation. ANN can be used to model the nonlinear system which it relates the inputs and outputs of a system. The data is usually flow forward, where the input layer will receive the input vector and transmit the values to the next layers. ANNs is able to extract the relation between the inputs and outputs of a process, without the physics being explicitly provided to them. They are able to provide a mapping from one multivariate space to another, given a set of data representing that mapping. Even if the data is noisy and contaminated with errors, ANNs have been known to identify the underlying rule. These properties suggest that ANNs may be well-suited to the problems of estimation and prediction in hydrology.

2.2.1 Advantages of Artificial Neural Network

In complex systems, the ANNs is a representative to an innovative and attractive solution to the problem of relating output variables to input ones (Gazzaz et al., 2012). Moreover, prediction is a common intention for employment of the neural network technology. The major steps for development of ANN models comprise of defining the proper model inputs, specifying network type, pre-processing and partitioning of the available data, defining network architecture, describing model performance criteria, and validating the model.

The advantages of ANN models over conventional methods have been discussed in detail by (Daliakopoulos et al.,2005). One of the most essential features of ANN models is their ability to perform tasks that a nonlinear program. It also able to adjust to recurrent changes and identify patterns in a complex natural system, it can continue without any problem by their parallel nature even an element of the neural network fails.

There are basically four different types of neural networks, including feedforward neural network, radial basis function (RBF), Kohonen self-organizing network and the recurrent neural network. However, in the present study, radial basis function (RBF) were employed to predict river discharge.

2.3 River Discharge Prediction Using Radial Basis Function (RBF)

The main features of RBF are as they build up of two-layer feed-forward networks, the hidden nodes and output nodes play their own roles. Basically, the network training is separated into two stages. The first stage is the weights from the input to hidden layer are determined, and second stage is determination of the weights from the hidden to output layer. In a nutshell, the training is absolutely fast and to conclude the networks are very good at interpolation. RBF has an input layer, a hidden layer and an output layer (as shown in Figure 2.2).

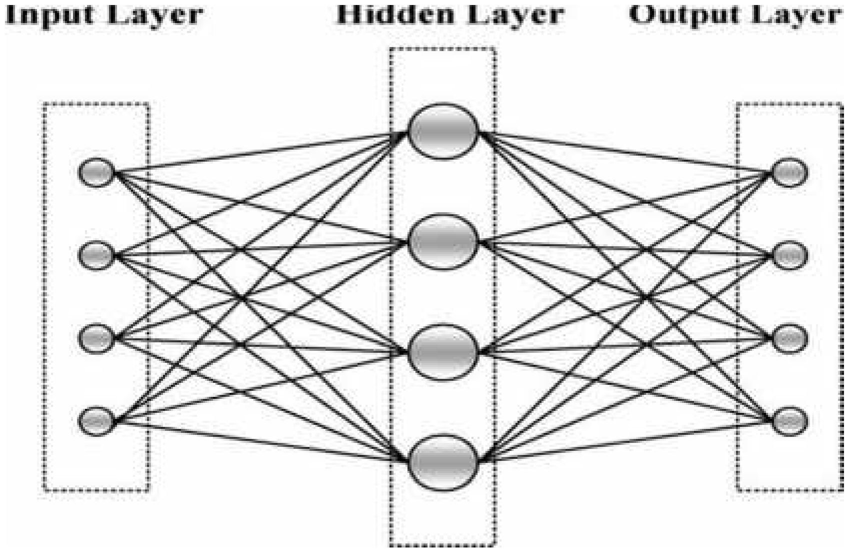


FIGURE 3 *:*Typical radial basis function

Orr (1996) simplifies that radial basis function is simply a class of function. In principle, they could be employed in any sort of model either linear or nonlinear and any sort of network (single-layer or multi-layer. An RBF network is nonlinear if the basis functions can move or change size or if there is more than one hidden layer. For single-layer network functions, it is fixed in position and size so nonlinear optimization but only for the regularization parameters in ridge regression and the optimal basis functions in forward selection.

2.3.1 Gaussian Function

Some of the most commonly used radial basis function are gaussian functions, multi-quadric functions, generalized multi-quadric functions, inverse multi-quadric functions, generalized inverse multi-quadric functions, thin plate spline function, Cubic Function, and linear function. Gaussian methods is the most favourable and commonly be used and they are characterized by identifying the specific centre and spread value. The spread which consist in the hidden function of RBFNN is the key components of the effectiveness of the outcome model. When applying the Gaussian method, the transfer function in the hidden nodes is responsible in transforming the information received from the input layer into the output response. In predicting river discharge for Pari river, Gaussian function will be used as the hidden unit function. Gaussian function equation is as shown below:

$$\phi(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right)$$

$$\text{width parameter} \quad \sigma > 0$$

Where,

ø(r) = radial basis function

σ = width parameter

2.4 River Discharge Prediction Using Feedforward Neural Network (FNN)

Since the introduction of the error back propagation learning algorithm, feedforward neural networks have been applied productively in many different problems. This network architecture and the corresponding learning algorithm can be viewed as a generalization of the popular least-mean-square (LMS) algorithm (Daliakopoulosa, 2005). A multilayer perceptron network comprises of an input layer, one or more hidden layers of computation nodes, and an output layer. From figure 4, FNN with one hidden layer consists of four input neurons, with three nodes and one output. The input signal propagates through the network in a forward direction, layer by layer. When compared RBF and FNN, radial basis function networks tend to learn much faster than a FNN.

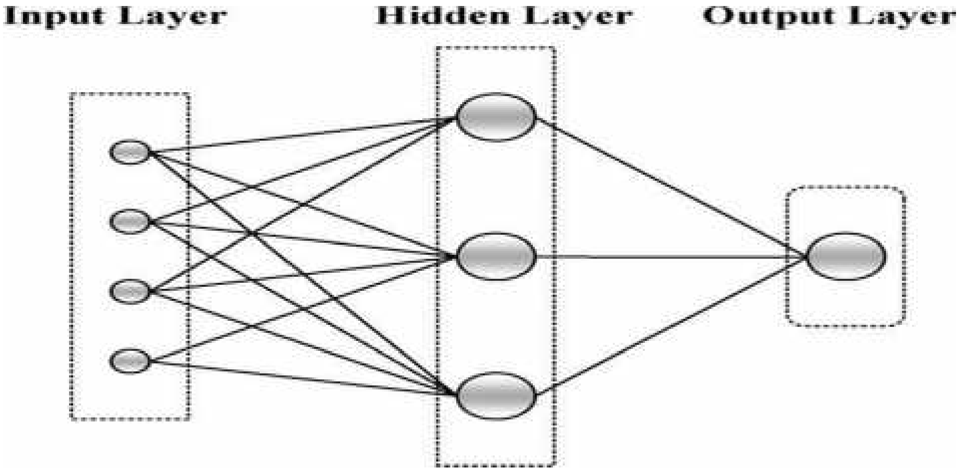


FIGURE 4: Feedforward Neural Network (FNN)

2.5 River Discharge Prediction Using Generalized Regression Neural Network (GRNN)

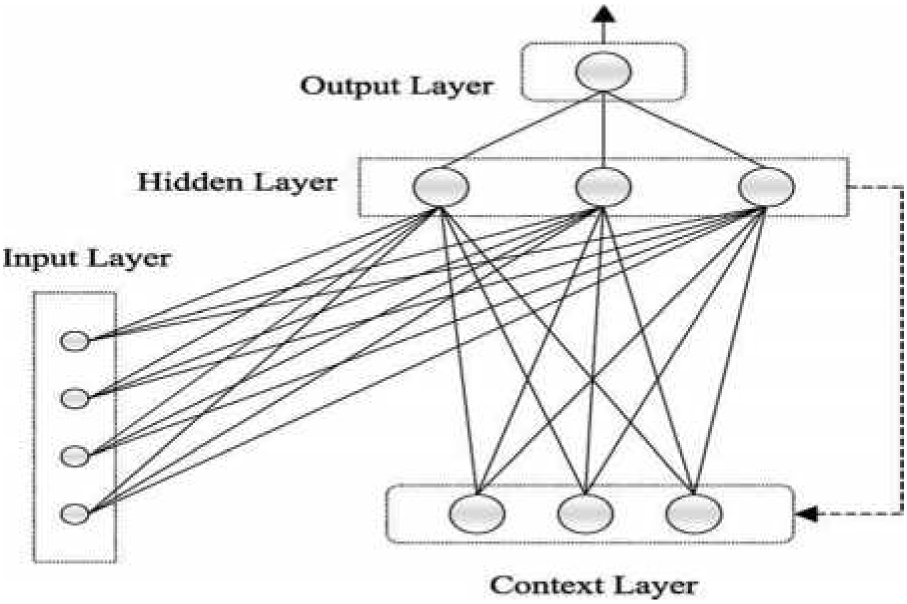


FIGURE 5: Typical Generalized Regression Neural Network

Generalized regression neural network feeds the outputs of the hidden layer back to itself. Fig. 5 shows a typical recurrent network consist of four layers: input layer, content layer, hidden layer and output layer. A context layer is intersected with the hidden layer and plays the role of the network memory. Taormina et al.,(2012) specifies these generalized regression networks can have an infinite memory depth and thus find relationships through time as well as through the instantaneous input space. Most real-world data contains information in its time structure. However, when compared the three neural networks, RBF is proven the most commonly used as it produced the best result in hydrology field. Thus, RBF is chosen to be used in this study.

2.6 Artificial Neural Network For Different Prediction Models

Tidal River

Hidayat et al.,(2014) proved that the inclusion of data from tide predictions at sea leads to a better model performance. Furthermore, the optimized ANN-based hind cast model yields a good discharge estimation, as shown by a consistent performance during both the training and validation periods. By using this model, discharge can be predicted from astronomical tidal predictions at sea plus water level measurements from a single station at an upstream location.

Urban Flood Control

Compared to Chang et al., (2014), they studied about how static artificial neural network (ANN) and two dynamic ANNs which are Elman neural network (Elman NN) and nonlinear autoregressive network with exogenous inputs-(NARX network) are used to construct multi-step-ahead floodwater storage pond (FSP) water level forecast models through two scenarios, in which scenario I adopts rainfall and FSP water level data as model inputs while scenario II adopts only rainfall data as model inputs.

Drought Forecasting

During last decade neural networks have shown boundless ability in modeling and forecasting nonlinear and non-stationary time sequences. In this study linear stochastic models (ARIMA/SARIMA), recursive multistep neural network (RMSNN) and direct multi-step neural network (DMSNN) was compared for drought forecasting which conducted at Kansabati River Basin, which lies in the Purulia district of West Bengal, India. Basically, drought forecasting shows an important part in the mitigation of effects of drought on water resources systems (Mishra and Desai,2006).

Sea Level Forecasting

In the present study conducted at Darwin Harbor, Australia, its hourly sea levels were anticipated using two different method, which are data driven techniques; adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN). To select the optimal input combination of hourly sea level, multi linear regression (MLR) technique was used (Karimi et al., 2012). For coastal engineering, in land drainage and reclamation studies, an accurate approximation of sea level disparities in estuaries where contributing rivers discharge into the sea is very vital elements.

CHAPTER 3: METHODOLOGY

3.1 Study Area and Data Source

The research data are attained from department of irrigation and drainage (DID) daerah Kinta, Perak for data of Pari River. Pari River is located in the southern city of Ipoh, Perak, Malaysia. Figure shows location of Pari River and its flow.

Pari River is a subcatchment of Kinta River and a drainage area of roughly 284 km² above Kinta River confluence and received an average mean annual rainfall of 2250 mm. based on information provided by DID, the main stream length is 39.78 km with a time of concentration (Tc) value of 14.4 minutes. Pari River drainage area is about 45% completely developed and the remaining 55% is forest and agricultures areas.

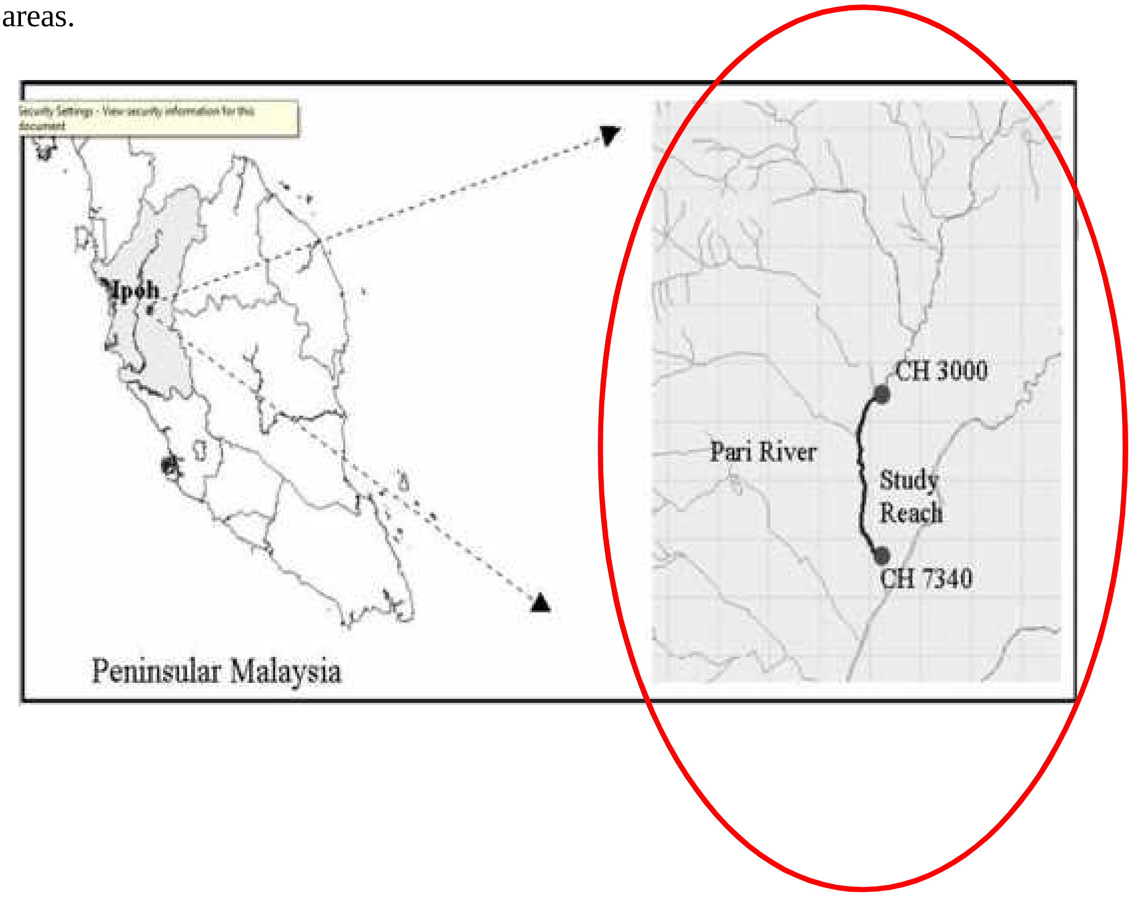

Security Settings - View security information for this document
Ipoh
Peninsular Malaysia
CH 3000
Pari River
Study Reach
CH 7340

FIGURE 6: Pari River Catchment Area

3.2 Development of RBF model

RBF network model is motivated by the nearby tuned response observed in biological neurons. Interpolation of multivariate functions is the theoretical basis of the RBF approach. The solution of the exact interpolating RBF mapping passes through every data point by differentiating number of hidden layer neurons and spread constants is used in the study.

3.2.1 Selection of Data

Data of water level of Pari River from year 1990 until 2013 were collected. Data of year 2011 and 2012 were chosen because of their recentness and the most completed data available from the DID. In order to get accurate estimation, the data must be adequate and specific for the modelling task. Other than that, input data must be selected according to their relevance for the modelling. Lastly, input data must be as limited as possible to reduce the training time and possibility of over fitting.

3.2.2 Partitioning of Data

Steps In Partitioning The Data:

1. Arrange all data according to date (river discharge and water level data)
2. Plot graph using full data
3. From the graph, divide the data into training and testing
4. Plot graph for training data only
5. Plot graph for testing data only
6. Do statistical analysis

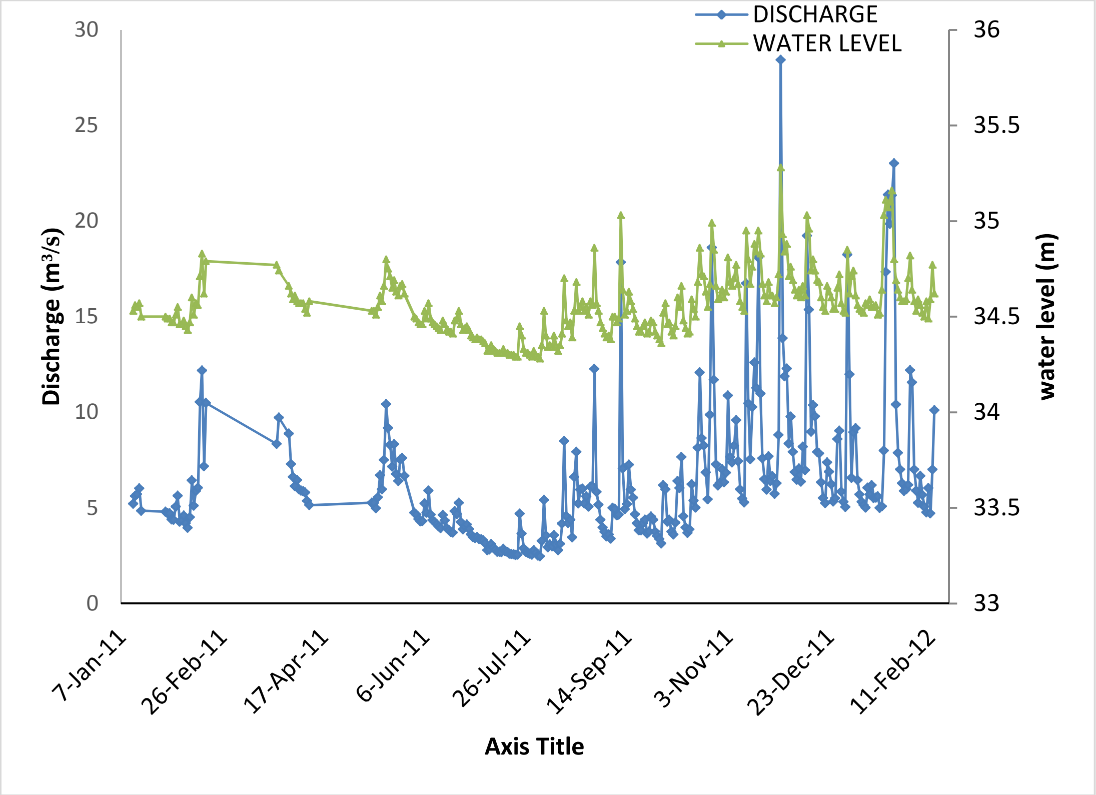


FIGURE 7: Time Series of Whole Data

The available datasets were divided into two training datasets and testing dataset. The partitioning of data for training and testing were based on the data trend. For each one of the input variables, the time series was divided in two different subsets. One subset for training the neural network (13 Jan 2011-05 November 2011) and one for model testing (6 November 2011-13 Feb 2012). Total available data are 314 and 214 of them was used for training purpose meanwhile 100 used for testing purpose. The time series of daily river discharge and water level for training and testing is display in figure 8 and figure 9. Data is divided into training and testing by follows condition that all data must be available and consistent and data for training is more than data for testing. Training set is used to adjust the weights on the neural network meanwhile testing set is used only for testing the final solution in order confirm the actual predictive power of the network. Training and testing data later will run by Matlab software.

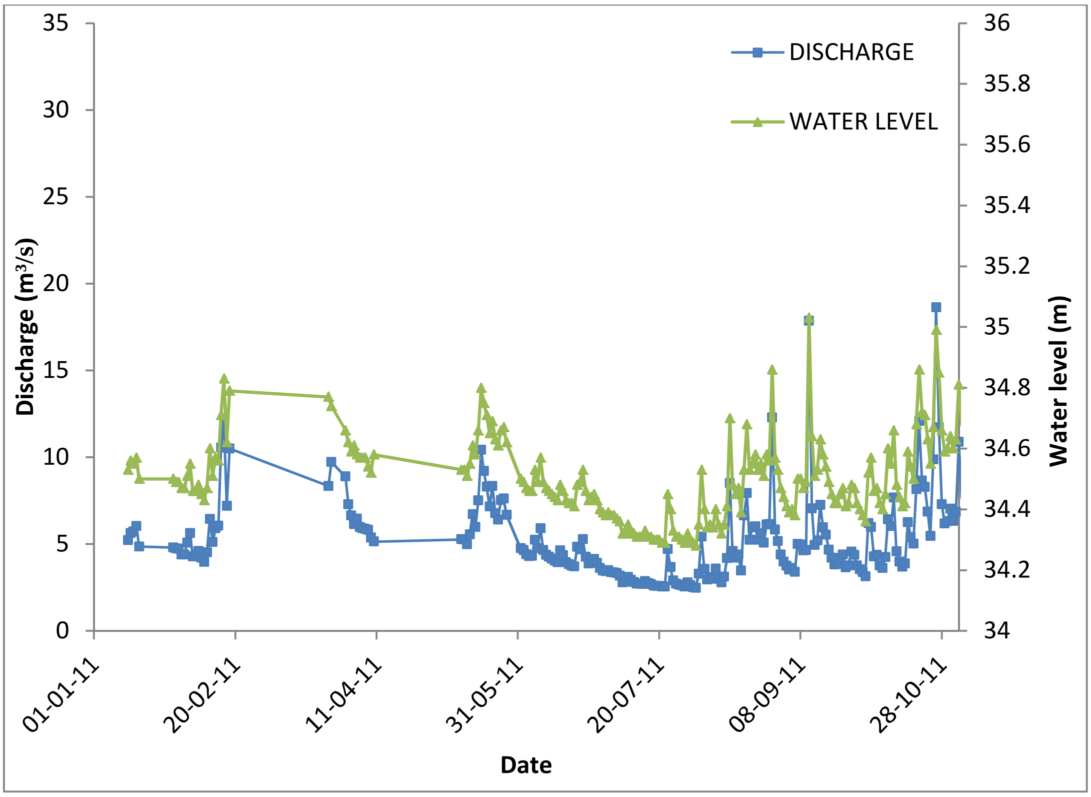


FIGURE 8: Time Series of Training Data

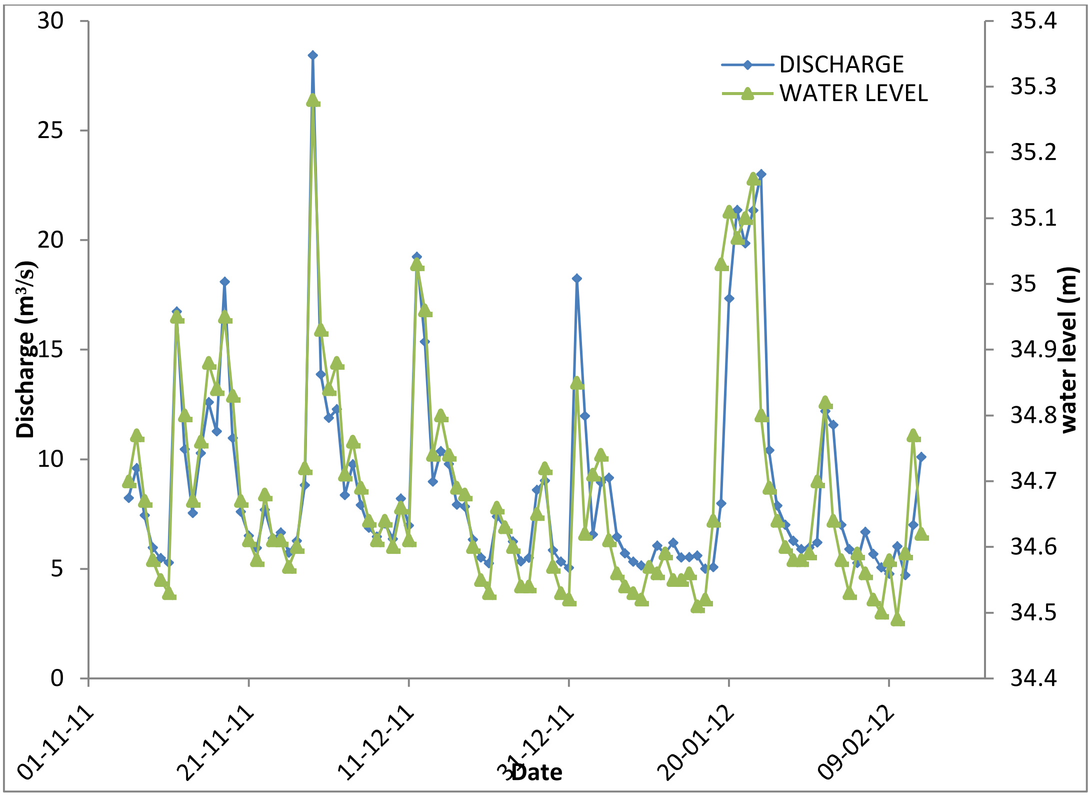


Figure 9: Time series of testing data

3.2.2 Statistical Analysis

TABLE 1: Statistical Parameters Of The Applied Data Set

	Training			Testing		
	Discharge	Rainfall	Water Level	Discharge	Rainfall	Water Level
Mean	5.241	4.493	34.503	8.777	11.440	34.684
Standard Deviation	2.381	10.597	0.136	4.583	15.239	0.163
Variance	5.670	112.306	0.019	21.004	232.229	0.027
Minimum	2.470	0.000	34.280	4.720	0.000	34.490
Maximum	18.620	67.500	34.503	28.420	65.000	35.280
Skew coefficient	2.208	3.497	0.817	2.033	1.582	1.472

Statistical analysis is done to identify complexity of the data, determine maximum, minimum, range and variation in data, to examine the correlated elements between the flow and suspended sediment variable, and lastly to compare testing and training data (Mustafa et al., 2012). Statistical parameters included in the data analysis are mean, standard deviation, variance, maximum and minimum value and coefficient of skewness.

Table 1 shows the statistical parameters of training and testing data. Mean value for discharge of testing data is slightly higher than training data with difference of 3.536 m^3/s. It means that testing has higher river discharge. When compared mean value of water level of training and testing, water level for testing is higher than training data with difference of 0.181 m. Both of the values for mean differences relatively low means that both training and testing data have a relatively constant stream discharge with low fluctuation.

Standard deviation difference for discharge of testing data is higher than training data by 2.202 m^3/s means that the difference quite low. Low standard deviation signifies the distribution of data is converged. Apart from that, having large standard deviation is an indicator that the data may contain no outliers.

For river discharge, there is slightly huge difference between minimum value which is 2.25 m^3/s. For minimum data for water level for training and testing is 34.28m and 34.49m with difference of 0.21m. The river has no flow during dry season. Apart from that, the minimum discharge of pari river for testing data has increased. This could be deduced as Pari river having modification to allow for larger capacity. For river discharge, the maximum value of training and testing data slightly large by difference of 9.80m^3/s. Simultaneously, for water level, the value shows small significant change by different of 0.777m. This implies the maximum capacity, in term of streamflow, which Pari river can hold during wet seasons is around 28.42m^3/s.

Low value of skewness coefficient indicated that the discharge data are mostly below the average or mean value for both training and testing data. When skew coefficient higher, complexity higher. For both training and testing for discharge and water level, skew coefficient is low meaning that the complexity is low. In predicting river discharge using gaussian radial basis function, low skewness is important as high value may affect the performance of the radial basis function negatively. This due to the increased complexity of the model's learning process as no significant pattern can be identified.

3.3 Model Structure

Formerly studied in the context of artificial intelligence, the ANN approach has now become one of the primary technologies especially in machine learning. Furthermore, it is a mainstream technology for data-driven modelling. Fundamentally, a neural network is a universal function approximation method which includes a large number of unknown parameters. Then the parameters are identified by solving an optimization problem. The purpose of an ANN is to generalize a relationship of the following form. In this study, ANN radial basis function (ANN-RBF) will be used to estimate spatiotemporal value of monthly river discharge. The step involved is Radial Basis Function (RBF) is used to assess the value of river discharge in Pari River in specific month and a spatial point within the study region, considering the value of monthly river discharge in other rivers. For spatial estimation, in the study Gaussian function will be used.

3.4 Normalization of Data

Data normalization is important to characterize the data in their unique form or commonly form or commonly known as standard form. The formula that is used in this research to normalize the data is as equation below:

$$v_p = 2 \text{ x } \frac{(X_p - X_{min})}{(X_{max} - X_{min})} - 1$$

The current normalized data is denoted as is the current original data V_p . X_m denotes the minimum value of the whole data and X_m denotes the maximum value of the whole data. The data in the context of study are river discharge and water level value. It is important to normalize the data to ensure a fast learning process of RBF model, hence producing estimation in a short time. In this study, the data were normalized between -1 and 1. The advantage of using [-1,1] for runoff modelling is that low and high flows event happening out the array of the calibration data may be accommodated (Dawson and Wilby, 1999).

3.5 Selection of ANN Model Architecture

in this research paper, since the radial basis function is used as the design model therefore, there are three layers which are input, hidden and output layer. The layers consist of specific number of neurons that should to be determined in this stage (Fakharuden,2014).

3.5.1 Input layer

Input data selection of RBF models is crucial for training and testing stage. The determination of input layer based on the number of input and the type of input variables. In this study, there are three input variables and they are current water level,1-antecendent water level, and 2-antecedent water level. The notation for each type of variable is W_t for current water level,W_{t-1} for 1-antecendent water level and W_{t-2} for 2-antecedent water level. The method to carry out the selection is subjected to recommendation from previous research papers..

3.5.2 Kernel

For this study, Gaussian function has been chosen as the kernel of the model.

3.5.3 Spread Coeffiecient

The default equation in the MATLAB software defined the spread of RBF model. In this study, the calculated spread is 0.69106. The spread values were evaluated through numbers of trial. In this study, the number of hidden layer and spread which created the lowest mean square error (MSE) value was selected as the best or optimum criterion for the model architecture.

3.5.4 Hidden Layer

According to Mustafa et al., (2012), trial and error method is used to determine number of neuron in hidden layer and to obtain the fitting structure of the network. Details of trial and error is presented in Table 2. Hidden layer determination is computed by using Matlab 7.8.0 and Microsoft excel spreadsheet. Data of Pari river from partitioning data is loaded into Matlab to allow selection process. Simulation will run after loaded the data. Then, graph of index versus stream and discharge will appear. Data for testing, 100 was entered into the simulation respectively. Then, number of neurons in the hidden layer will be demanded. Number of hidden neuron can started with any number. In this study, the number of neuron in hidden neuron is started with value of 4 and consequently increased by 1 for next trial. The determination ended with 100 neurons in hidden. Value of 4 is the optimum minimum number of neuron in hidden layer can be identified the spread coefficient.

TABLE 2: Determination of neuron in hidden layer using trial and error method

no of trial	no of neuron in hidden layer	MSE	
		training	testing
1	4	0.893	16.344
2	5	0.425	9.274
3	6	0.357	5.584
4	7	0.389	13.470
5	8	0.250	8.064
6	9	0.155	9.695
7	10	0.170	11.032
8	11	0.127	6.316
9	12	0.140	8.771
10	13	0.112	7.771
11	14	0.176	11.988
12	15	0.142	7.385
13	16	0.112	6.433
14	17	0.110	8.223
15	18	0.105	6.830
16	19	0.098	4.767
17	20	0.100	6.315
18	50	0.080	8.8571
19	100	0.071	6.722

TABLE 3 : Analysis of trial and error method

	MSE	
	training	testing
lowest value	0.071	4.767
no.of layer in hidden neuron	100	19
highest value	0.893	16.344
no.of layer in hidden neuron	4	4

From determination of neuron in hidden layer using trial and error method, analysis of trial and error method is carried out to determine the lowest and highest value of MSE and number of layer in hidden neuron for both phase of training and testing. Neurons that produced the lowest MSE is the best solution to find hidden neurons (Sheela and Deepa, 2013). From trial and error method, number of layer in hidden neuron that gave the lowest MSE for training and testing were 100 and 19 correspondingly. This could be the best option to choose the number of hidden neuron. However, from 20 neurons until 100 neurons, the MSE values at testing phase were greatly higher than at training phase. This could lead to overfiting problem. Overfiting is a phenomena whereby a model is excessively complex. In this study, high number of neuron leads to the problems. Nevertheless, overfiting is a common issue in radial basis function (Safwan,2013).

Figure 10 shows target and output discharge after testing RBF graph for 100 neurons. Over fitting occurred as a result of too high number of neurons. The model architecture has over approximate the complexity of the target problem resulted in degradation of generalization capacity (Sheela and Deepa, 2013). Enlarging image shows the area that affected by over fitting and Matlab was used to trace the problem. The horizontal axis represents the targeted output discharged value indicated by blue colour line meanwhile the vertical axis represents the predicted output discharged indicated by green colour line. As shown in the figure, output value does not follow the trend of target value. In testing phase the model will produce poor performance differ from training phase which the model will perform well. Therefore, in order to determine the best choice for number of neuron in hidden layer, 4 neurons until 19 neurons were evaluated and it was determined that 19 neurons give the lowest value of MSE which are 0.098 for training and 4.767 for testing. Thus, 19 neurons were used in design of radial basis function architecture.

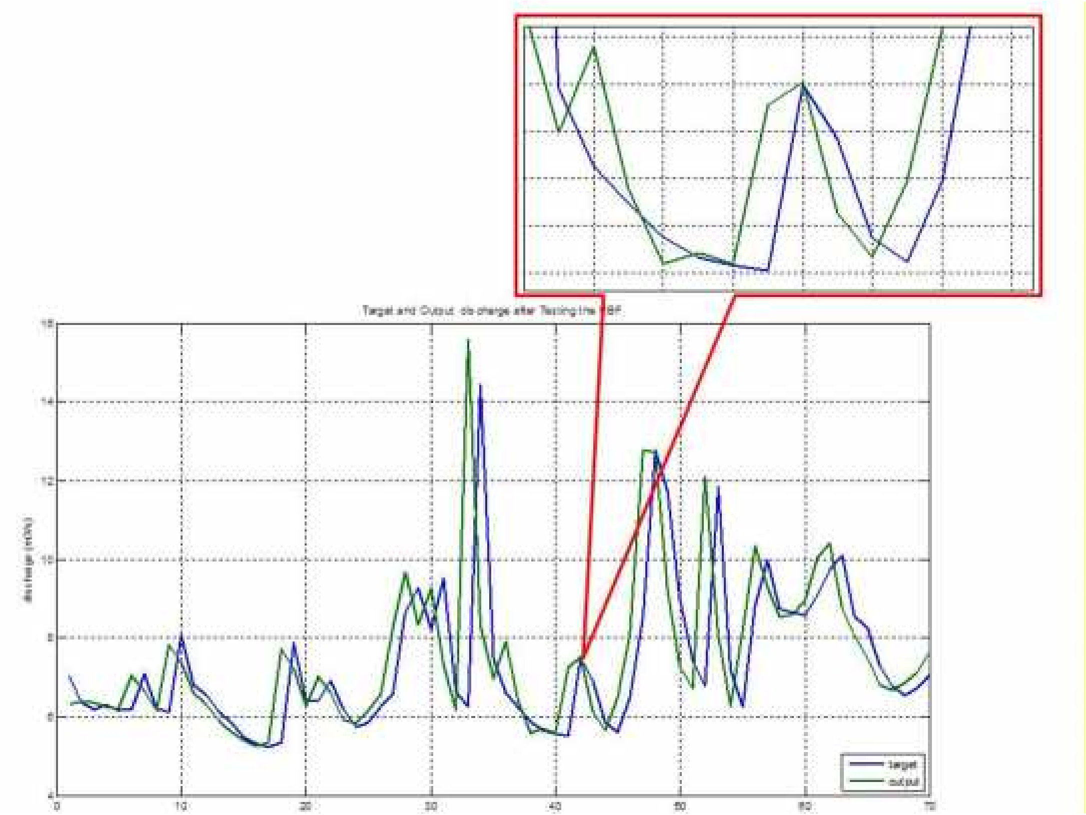


FIGURE 10: target and output discharge after testing the RBF

3.5.5 Performances Evaluation Measures

The most commonly employed error measured were the root mean square error (RMSE), the mean square relative error (MSRE), the coefficient of efficiency (CE) and the coefficient of determination (r^2) (Dawson and Wilby, 1999). They claimed that a reliable measure of goodness of fit at the high flows can be produced by square error despite the fact that relative errors are partial towards moderate flows. Based on formula shown below, z_n the observed discharged value and y_n the predicted value for discharged and $\bar{z}$ is the mean of the observed discharged value and N is the total number of observation for the computed error.

$$RMSE = \left[\frac{1}{N} \sum_{n=1}^{N} (z_n - y_n)^2 \right]^{1/2}$$

$$MAE = \frac{1}{N} \sum_{n=1}^{N} (z_n - y_n)$$

$$CE = 1 - \frac{\sum_{n=1}^{N} (z_n - y_n)^2}{\sum_{n=1}^{N} (z_n - \bar{z})^2}$$

3.5.6 Output Layer

There is only one output layer for this study of radial basis function using Gaussian function. The output is discharge value with respect to forecast water level. Summary of the RBF model is as follows:

Spread,σ	=	0.69106
Kernel function	=	Gaussian function
Input variables	=	3(W_t,W_{t-1},W_{t-2})
Hidden layer	=	19 neurons
Output neuron	=	1

The architecture of RBF model is as shown in figure 11.

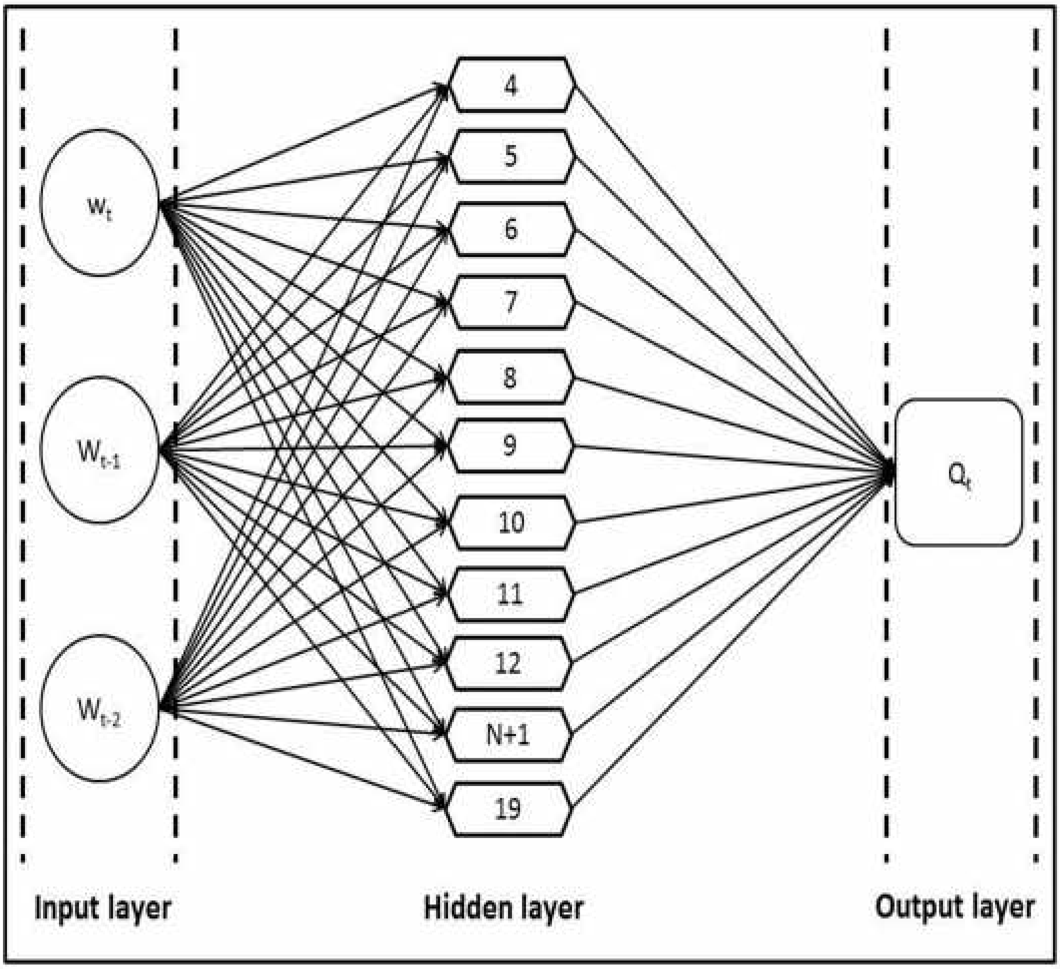
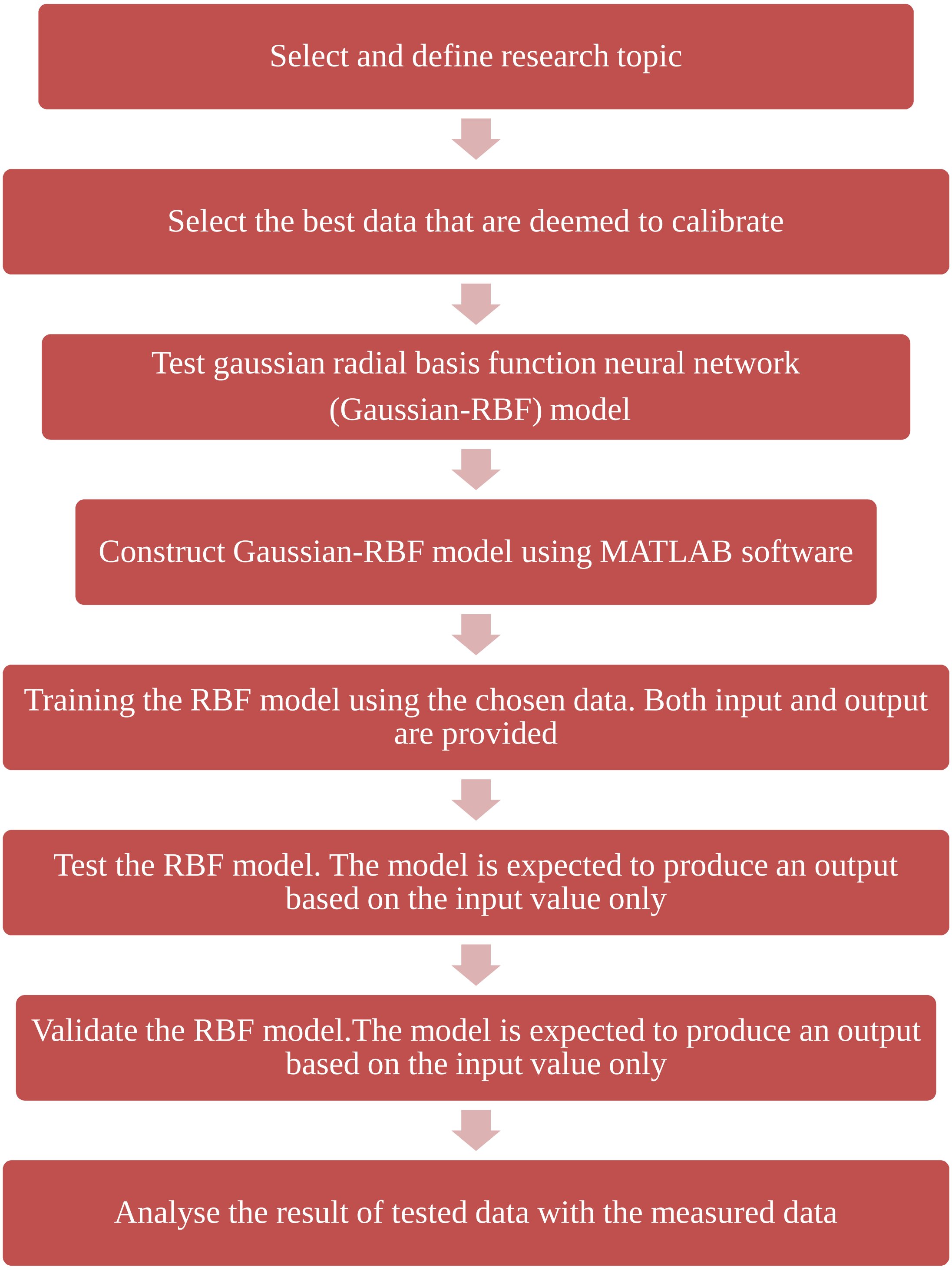


FIGURE 11: Final model of Gaussian Radial Basis Function

3.6 Project Flow Activities

Below are the steps for the project throughout the FYP1 and FYP2 until completion.


Select and define research topic
Select the best data that are deemed to calibrate
Test gaussian radial basis function neural network (Gaussian-RBF) model
Construct Gaussian-RBF model using MATLAB software
Training the RBF model using the chosen data. Both input and output are provided
Test the RBF model. The model is expected to produce an output based on the input value only
Validate the RBF model.The model is expected to produce an output based on the input value only
Analyse the result of tested data with the measured data

3.7 Project Key Milestones

Milestone planning is used to show the major steps that are needed to reach the goal on time. When several tasks have been completed the milestone is reached. Final year project (FYP) is divided into 2 sections, namely section FYP1 and FYP2. Below are the milestones for both FYP1 and FYP2.

- Semester 1 (FYP1)

TABLE 4: Key milestone for FYP1

Milestone	Week
Project proposal	Week 1
Extended proposal	Week 7
Proposal defence	Week 10
Interim report	Week 14

- Semester 2 (FYP2)

TABLE 5: Key milestone for FYP2

Milestone	Week
Progress report	Week 7
Pre-sedex	Week 10
Technical paper	Week 12
Viva presentation	Week 13
Project dissertation (hard bound)	Week 15

3.8 Gantt Chart

Below is the gantt chart for the whole course of FYP. For FYP1, the total time allocated is 14 weeks while 15 weeks for FYP2. FYP1 would focus on the execution of research and FYP2 would focus on complete documentation of the research.

TABLE 6: Gantt chart for FYP1

no	activities	duration (week)													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	selection project title	■													
	1. survey project title	■													
2	project comfirmation		■												
	1. tite comfirmed by supervisor		■												
	2. start preliminary research work		■												
3	data gathering		■	■	■										
	1. collect and identify relevant journals			■	■										
	2. proposal drafting process			■	■										
4	3. understand problem statement			■	■										
	4. identify objective			■	■										
	5. literature review review				■	■	■								
	6. working on extended proposal					■	■								
5	submission of extended proposal							■							
	1. amendment of extended proposal								■	■					
6	proposal defence										■				
	1. presentation preparation									■	■				
	2. consultation with supervisor									■	■				
7	interim report preparation											■	■	■	
	1. amendment from extended proposal												■	■	
8	2. consultation with supervisor												■	■	
9	submission of interim report														■

TABLE 7: Gantt chart for FYP2

Item No.	activities	Duration													
		Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14
1	project work continues	■	■												
	1. continuation from FYP 1	■	■	■	■										
	2. Working on Matlab software			■	■	■	■								
2	Submission of progress report							■							
	1. progress report draft amendment								■						
3	project work continues								■	■					
	1. result and discussion								■	■					
	2. Data Analysis								■	■					
4	Pre-SEDEX										■				
	1. pre-sedex preparation									■					
	2. Slide Presentation									■					
5	Submission of Draft Final Report											■			
	1. amendment from draft final report											■			
	2. Consult with supervisor for amendment											■			
	3. Meeting with supervisor regarding the report											■			
6	submission of dissertation (soft bound)											■			
7	Submission of Technical Paper												■		
8	Viva													■	
9	Submission of Project Dissertation (Hard Bound)														■

3.9 Project Key Milestones

Milestone planning is used to show the major steps that are needed to reach the goal on time. When several tasks have been completed the milestone is reached. Final year project (FYP) is divided into 2 sections, namely section FYP1 and FYP2. Below are the milestones for both FYP1 and FYP2.

- Semester 1 (FYP1)

TABLE 8: Key milestone for FYP1

Milestone	Week
Project proposal	Week 1
Extended proposal	Week 7
Proposal defence	Week 10
Interim report	Week 14

- Semester 2 (FYP2)

TABLE 9: Key milestone for FYP2

Milestone	Week
Progress report	Week 7
Pre-sedex	Week 10
Technical paper	Week 12
Viva presentation	Week 13
Project dissertation (hard bound)	Week 15

3.10 Tools and Software

There are two mains software that were used throughout the research which are MATLAB and Microsoft Excel. In reality, MATLAB is a very great programming tools that have wide application in many types of engineering and non-engineering related field for specific purpose such as math and computations, algorithm development, data acquisition, modelling, simulation and prototyping, data analysis, exploration and visualization, scientific and engineering graphics and application development, including graphical user interface building. In this study, this programming tools is used to help in define and develop the flow network model. Aside from Matlab and Microsoft Excel, others related software used is Notepad and Microsoft Word.

CHAPTER 4: RESULT

4.1 Statistical Model Analysis

Statistical model analysis was performed by plotting predicted discharge vs observed discharge for both dataset training and testing. The graphs plotted were shown in figure 12. From the figure, it clearly shown that predicted is highly correlated with observed value. To conclude that RBF model developed, which is Gaussian function has gained an adequate learning process as a result of the high numbers of loaded input data and sufficient learning time. However, the presence of outliers in the plotted graph is few significant but the predicted discharge still follow the same pattern as observed discharge. This can be explained by plotted graph of observed and predicted discharge versus time series as shown in figure 13.

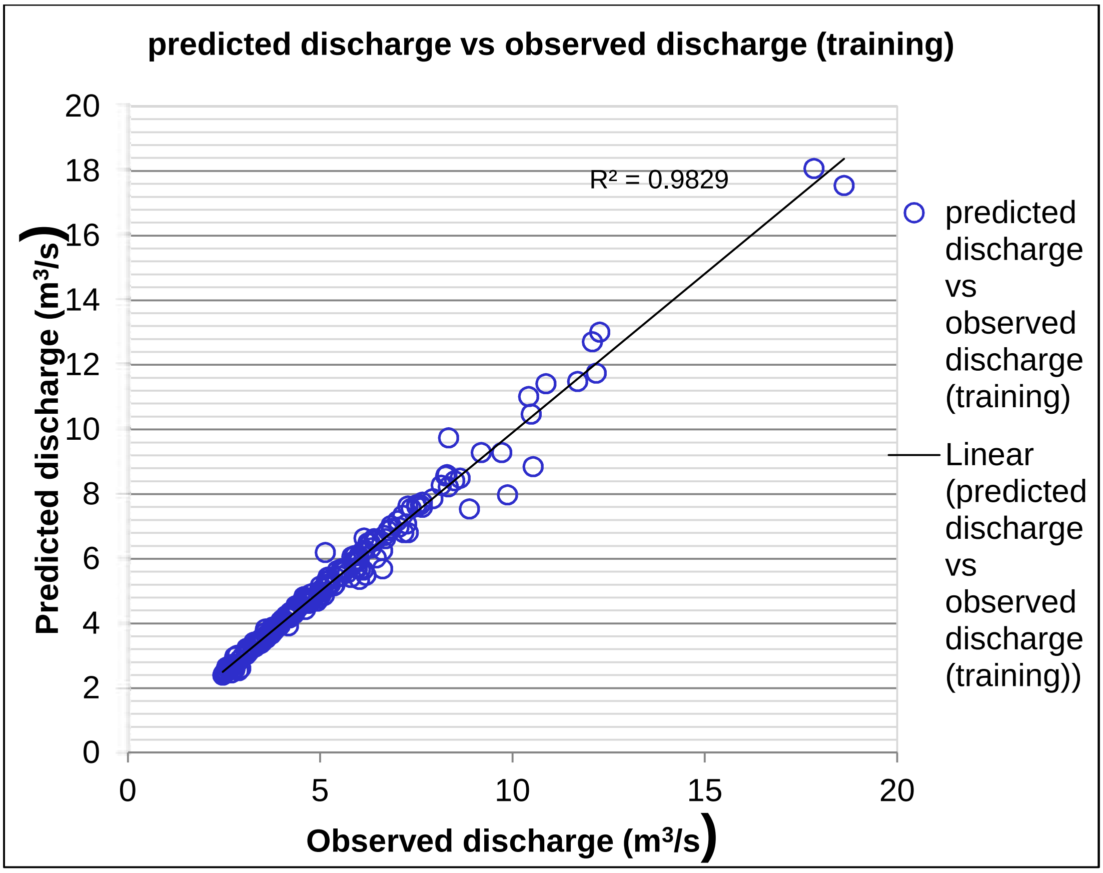


Figure 12: plotting of predicted discharge and observed discharge for training data

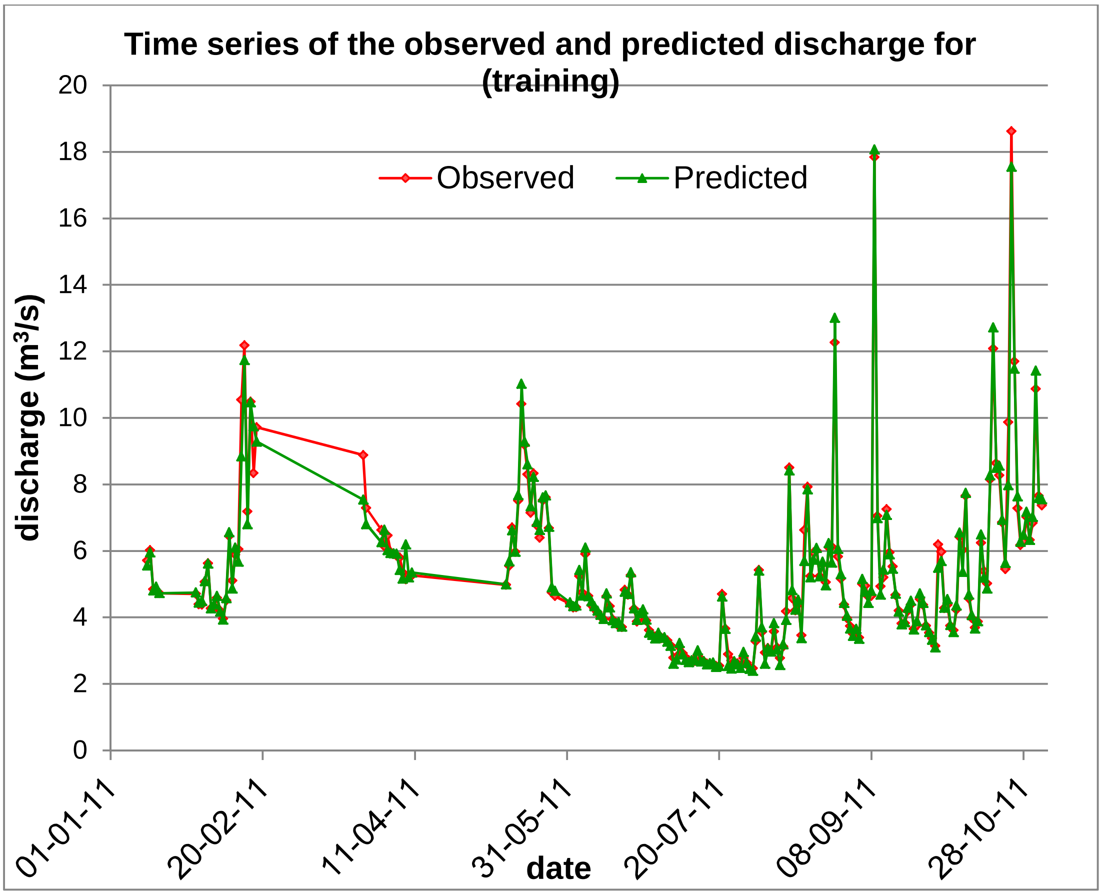


Figure 13: Time series of the observed and predicted discharge for training data

From time series of the observed and predicted discharge for training graph, the comparison between observed discharge and predicted discharge can be clearly recognized. The same pattern shown in the graph described that training stage has high correlation as it allows the predicted discharge to be made close to observed discharge.

However, testing stage performance is decline from training stage. From predicted discharge vs observed discharge testing graph shown in figure 14, there is huge presence of outliers can be seen. Outliers can affect the accuracy of prediction as more outliers, the less the accuracy. From the figure, the outliers have negatively affect scattering of whole data of testing. Thus produce poorly correlation for predicted and observed discharge and lead to lower accuracy of predictive performance. Outliers is exist due to the high marginal difference between the predicted and observed value at those particular points.

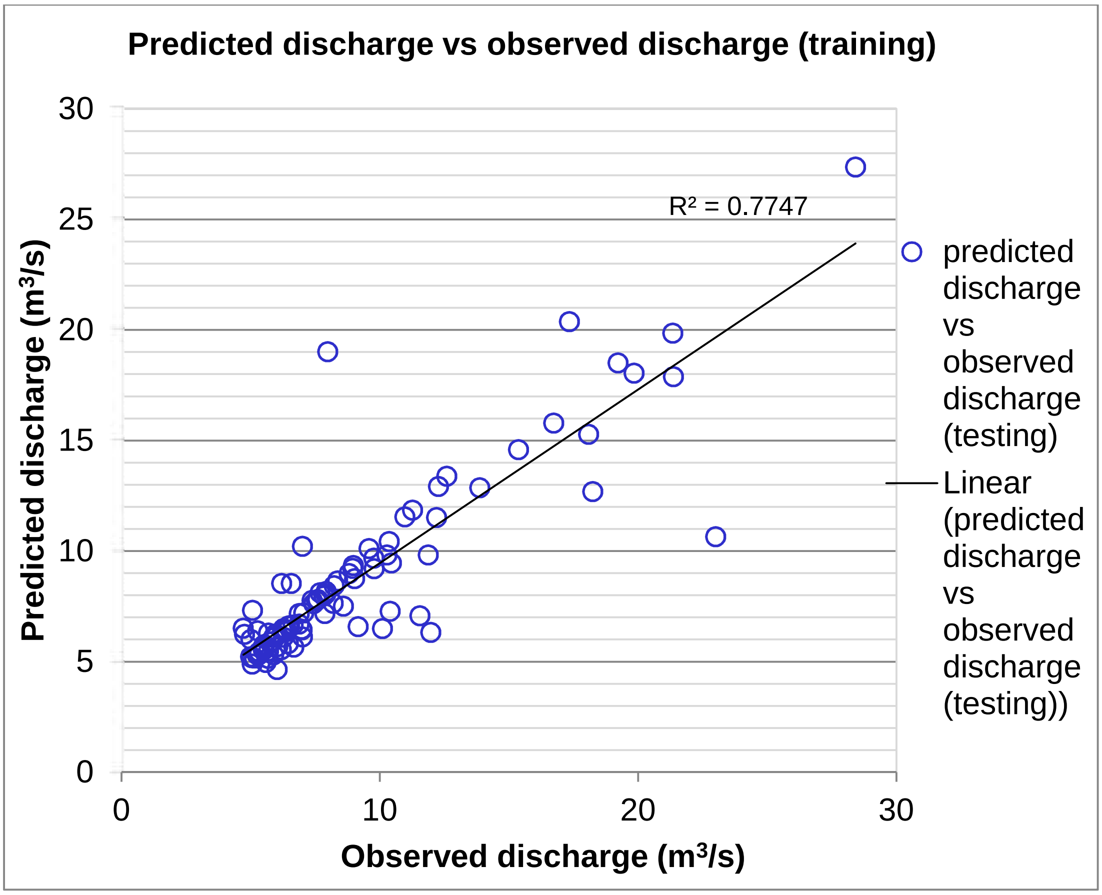


FIGURE 14: Plotting of predicted discharge vs observed discharge for testing data

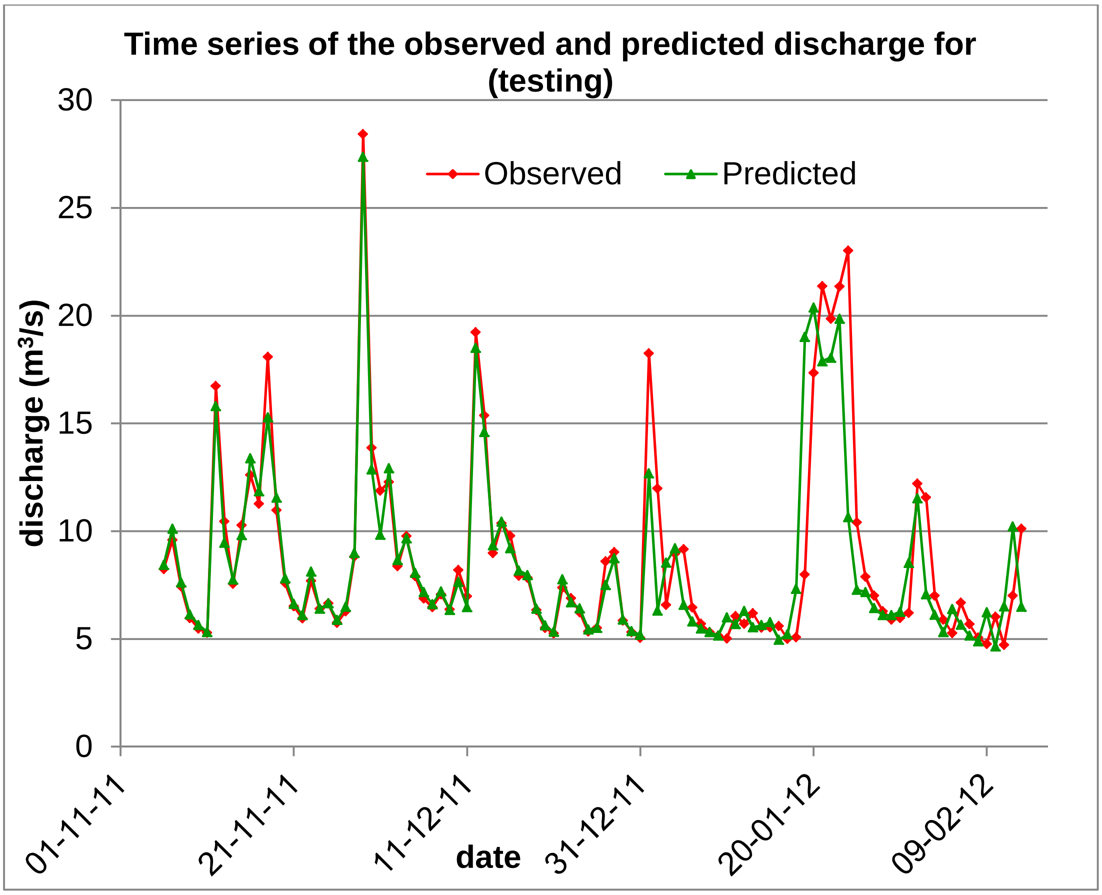


Figure 15: Time series of the observed and predicted discharge for testing data

From figure 15, at time series of 27 August 2012, the predicted and observed discharge recorded were 12.68 m^3/s and 18.25 m^3/s respectively. This is quite huge in difference compared to other values. Bear in mind that if observed discharge is high, same goes to the margin of prediction error, it also will be high as well. Thus, it was found that Gaussian radial basis function encounter with a problem to adapt with value of large magnitude and as a consequence effect in discrepancy of the data along the line of agreement.

4.3 Statistical Performance Measure Analysis

Statistical parameters were done for the analysis of model performance. In this study, the statistical parameters involved are The statistical analysis of the result were made by calculating each of the parameters using specific formula root mean square error (RMSE), mean absolute error (MAE), coefficient of efficiency (CE) and coefficient of determination (R^2). In fact, each of these parameters is a very dominant indicator towards the analytical of the overall performance of the developed model. The statistical analysis of the result were made by calculating each of the parameters using specific formula stated in statistical measures part.

TABLE 10: Statistical analysis of model's performance

Data set	RMSE	MAE	CE	R^2
Training	**0.312**	**0.012**	**0.997**	**0.983**
Testing	**2.183**	**0.284**	**0.951**	**0.775**

From calculation by using excel program, it was established that the value for MSE, RMSE and CE for testing did produce a very good and pleasing result in predicting the river discharge of the Pari River. Nevertheless, for the simplification purpose, only RMSE, CE and R^2 were chose to be presented in the result part. Basically, RMSE is the root factor to the actual MSE. Therefore, indication and analysis of the performance of the model developed is sufficient by using these two type of parameters.

From table 10, the simplified statistical analysis of model's performance in predicting river discharge in Pari River is shown. MSE value for testing is higher than training shows that the size of the error which correlate the predicted with the observed discharged value in the system. As a result, root mean square error also affected by the high magnitude error during testing resulting in high error in magnitude in testing process. This can be explained by the cluster of input inserted into the system is far from the actual mean value obtained.

From the table 10, RMSE value is quite low. This indicates that training and testing data set have small range of data which the maximum is 28.420 m^3/s and the minimum is 18.620m^3/s. RMSE value for both data set has increasing in value from 0.312 for training to 2.183 for testing.

However CE shows opposite pattern as shown by MSE and RMSE. The decreasing in value for both from training to testing is 0.977 to 0.951. Coefficient of efficiency (CE) is one of the most important parameter. This is because CE can gauge predictive performance of hydrological models. To prove that the RBF model developed be able to achieve prediction with high efficiency, thus CE is used as indicator. Ideally, the value of CE should be one. From the calculation, both CE value for training and testing considerably have high value of efficiency, which close to 1. From summary of CE value obtained, it can be assumed training and testing data for prediction of river discharge are sufficient.

Another parameter need to take into consideration is coefficient of determination (R^2). For R^2, the best option is the value range from 0-1. From table 10, the value of training data is 0.983 compared to testing data is 0.775. it clearly shows that value of training is much higher than testing. This is happens because the number of load input value which is higher compared to testing. As more input is loaded, the higher the performance of R^2 as a result of adequate learning process. In other word, training stage displays high correlation between predicted discharge and observed discharge compare to testing stage. Training stage has the highest correlation because the developed RBF model is able to predict river discharge near to observed discharge value. In contrast, testing data is difficult to forecast higher range of data because of high value of discharge data hence resulted in low R^2. To improve the correlation in testing data, training the RBF model by using larger range of data is needed.

.

CONCLUSION

Gaussian basis function neural network was successfully applied for prediction of river discharge at Pari River.in the end, it was applicable and suitable for river discharge prediction. After number of trials, 19 numbers of hidden neurons with spread value 0.69106 produced the best developed model and thus come out with satisfactory result. Thus, it was proven that three numbers of input variables which were water level produced the best outcome for output, which in this study was river discharge. The performance of Gaussian basis function is evaluated by using various statistical measures such root mean square error (RMSE), mean absolute error (MAE), coefficient of efficiency (CE) and coefficient of determination (R^2). Both values for RMSE and MSE of training and testing dataset recorded was low and it meant that less error produced by the model. In addition, coefficient of efficiency, CE obtained essential closed to 1 indicated high accuracy of the performance. Moreover, Gaussian basis function with coefficient of determination, R^2 of 0.983 and 0.775 show high accuracy and good performance of the developed radial basis function (RBF) model. Huge presence of outliers resulted in lower R^2 value obtained for testing compared to training. This happened due to less number of input data loaded. Despite of low value of R^2 obtained, Gaussian basis function still appropriate and suitable to predict river discharge. Thus, the objectives of this study which was prediction of river discharge by using Gaussian basis function were achieved. The application of developed model can be used in the future for predicting hydrological data because it can produce accurate and dependable data sources. For recommendation, in future, it is recommended to include a wide range of data for training of RBF model to produce prediction with high accuracy. Thus, increasing number of loaded data will resulted in prediction with high accuracy

References

A. Belayneh a, J. A. (2014). Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression. 418–429.

A.K. Mishra, V. D. (2006). Drought forecasting using feed-forward recursive neural network. 127–138.

A.R, F. (2013). Water Flow Prediction in Perak River using Thin Plate Spline Basis Function Neural Network.

Alp, H. K. (2004). Rainfall-Runoff Modelling Using Three Neural. 1-6.

Altunkaynak, A. (2014). Predicting Water Level Fluctuations in Lake Michigan-Huron Using Wavelet Expert System Methods. 2293–2314.

Bakar, A. S. (2013). Prediction of Suspended Sediment Concentration in Kinta River Using Soft Computing Technique.

Bijari, M. K. (2014). Fuzzy artificial neural network (p, d, q) model for incomplete financial time series forecasting. 831–845.

Committee, A. T. (2000). Artificial Neural Network in Hydrology.II Hydrologic Application. 124-137.

Fazlina Ahmat Ruslan, A. M. (2014). River, Flood Water Level Modeling and Prediction Using NARX Neural Network: Case Study at Kelang.

Fi John Chang, P. A. (2014). Real-time multi-step-ahead water level forecasting by recurrent neural networks.

Hidayat, H., Hoitink, A. J., Sassi, M. G., & Torfs, a. P. (2014). Prediction of Discharge in a Tidal River Using Artificial.

J.S, Y. (2014). River Flow Prediction Using Multi-Quadric Basis Function Neural Network For Perak River.

Jeofry, M. &. (2013). General Observations about Rising Sea Levels in Peninsular Malaysia. 363-370.

Jose Maria P. Menezes Jr., G. A. (2008). Long-term time series prediction with the NARX network:An empirical evaluation. 3335–3343.

K.Gnana Sheela, S. (2013). A New algorithm to find number of hidden neurons in Radial Basis Function Networks for Wind Speed Prediction in Renewable Energy Systems . 30-37.

L.Wilby, C. W. (1999). A Comparison of Aritificial Neural Network Used For River Flow Forecasting. 529-540.

Lu, M. L. (2014). Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?

M. R. Mustafa, M. H. (2012). Artificial Neural Networks Modeling in Water Resources Engineering: Infrastructure and Applications. Vol 6.

Majid Heydari, E. O. (2013). Development of a Neural Network Technique for Prediction of Water Quality Parameters in the Delaware River, Pennsylvania. 1367-1376.

Mehmet C. Demirel a, 1. A. (12 November 2008). Flow forecast by SWAT model and ANN in Pracana basin, Portugal.

rosmina bustami, n. b. (2006). artificial neural network for precipitation and water level prediction for Bedup River.

S.A. Kalogirou, E. M. (2014). Artificial neural networks for the performance prediction of large solar. 90-97.

Sepideh Karimi, O. J. (2013). Neuro-fuzzy and neural network techniques for forecasting sea level in Darwin Harbor,Australia. 50–59.

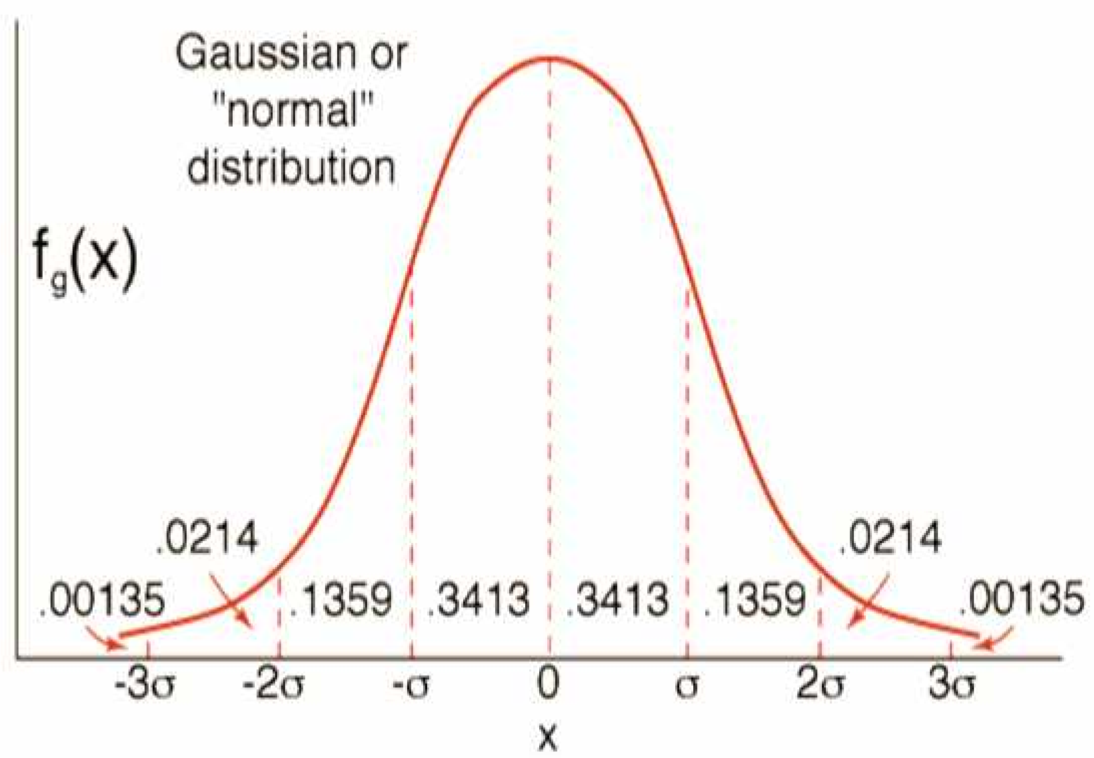
Srivastava, R. R. (2014). Predicting Monsoon Floods in Rivers Embedding Wavelet Transform, Genetic Algorithm and Neural Network. 301–317.

Sztobryn, M. (2013). Application of Artificial Neural Network into the Water Level Modeling and Forecast. 7(2).

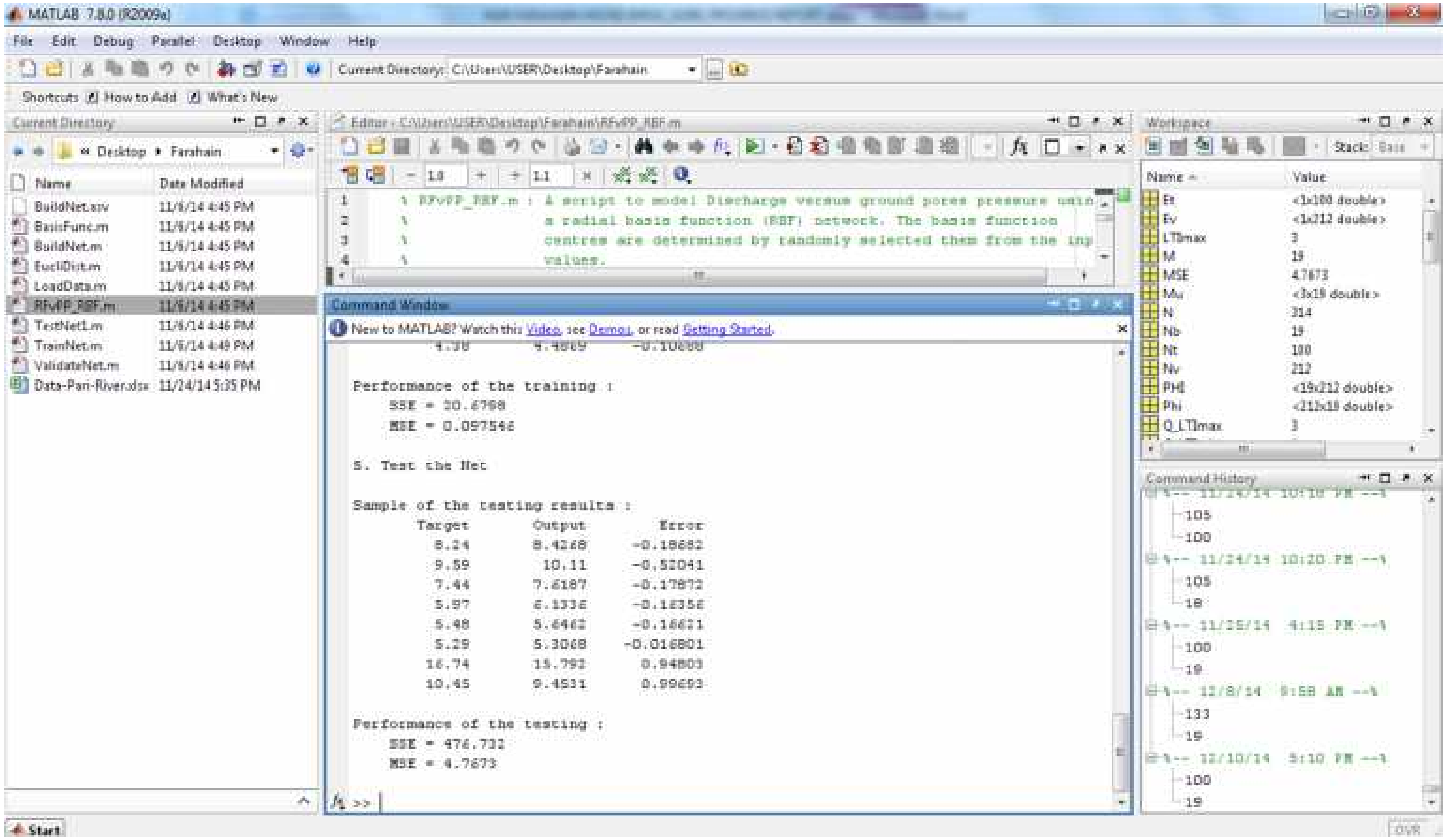
Vesna Rankovic, A. N. (2013). Predicting piezometric water level in dams via artificial neural networks. 1115–1121.

APPENDICES

Pari River Flow



Gaussian Distribution



MATLAB 7.8.0 (R2009a)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\USER\Desktop\Farahain
Shortcuts How to Add What's New
Current Directory
Desktop ▸ Farahain
Editor - C:\Users\USER\Desktop\Farahain\RFvPP_RBF.m
Workspace
Stack: Base
Command Window
Command History
Start
OVR

Name	Date Modified
BuildNet.asv	11/6/14 4:45 PM
BasisFunc.m	11/6/14 4:45 PM
BuildNet.m	11/6/14 4:45 PM
EucliDist.m	11/6/14 4:45 PM
LoadData.m	11/6/14 4:45 PM
RFvPP_RBF.m	11/6/14 4:45 PM
TestNet1.m	11/6/14 4:46 PM
TrainNet.m	11/6/14 4:49 PM
ValidateNet.m	11/6/14 4:46 PM
Data-Pari-River.xlsx	11/24/14 5:35 PM

```
% RFvPP_RBF.m : A script to model Discharge versus ground pores pressure usin
%                a radial basis function (RBF) network. The basis function
%                centres are determined by randomly selected them from the inp
%                values.
```

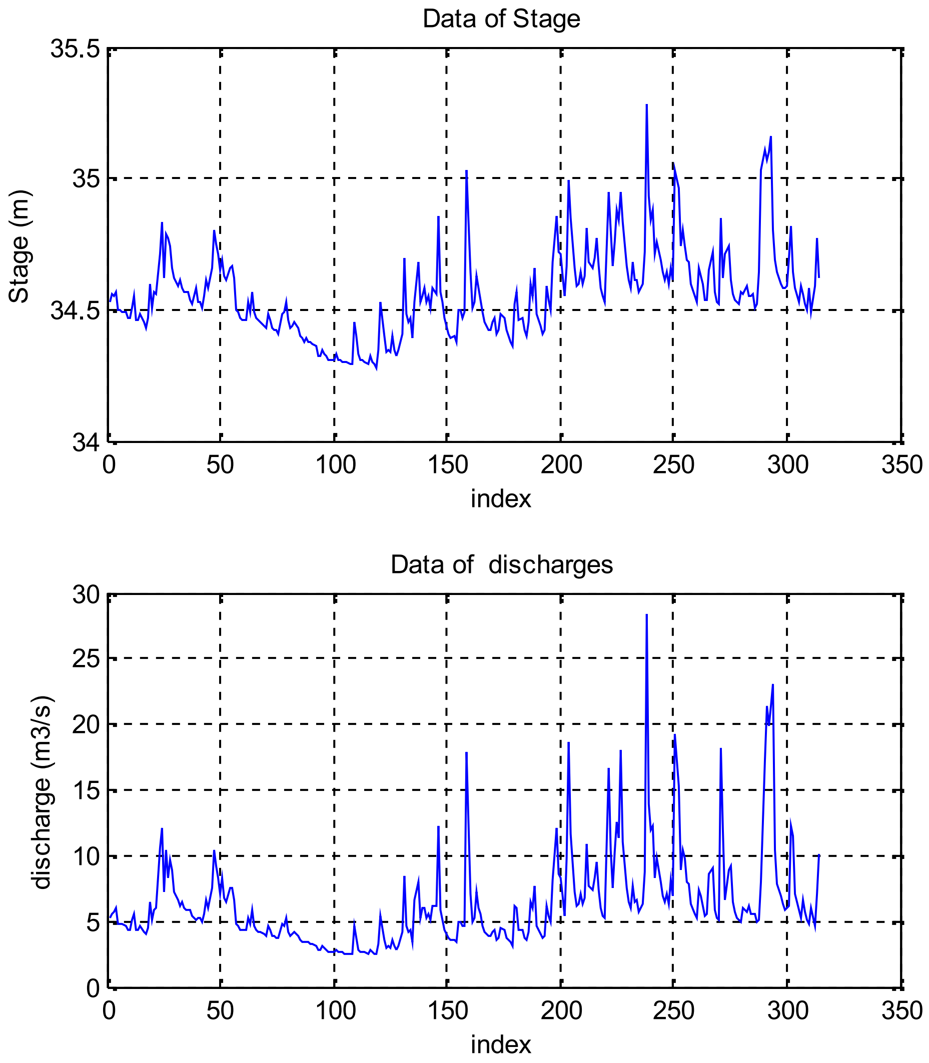
New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
         4.38         4.4869      -0.10888

Performance of the training :
    SSE = 20.6798
    MSE = 0.097546

5. Test the Net

Sample of the testing results :
         Target        Output         Error
           8.24        8.4268      -0.18682
           9.59         10.11      -0.52041
           7.44        7.6187      -0.17872
           5.97        6.1336      -0.16356
           5.48        5.6462      -0.16621
           5.29        5.3068     -0.016801
          16.74        15.792       0.94803
          10.45        9.4531       0.99693

Performance of the testing :
    SSE = 476.732
    MSE = 4.7673

>>
```

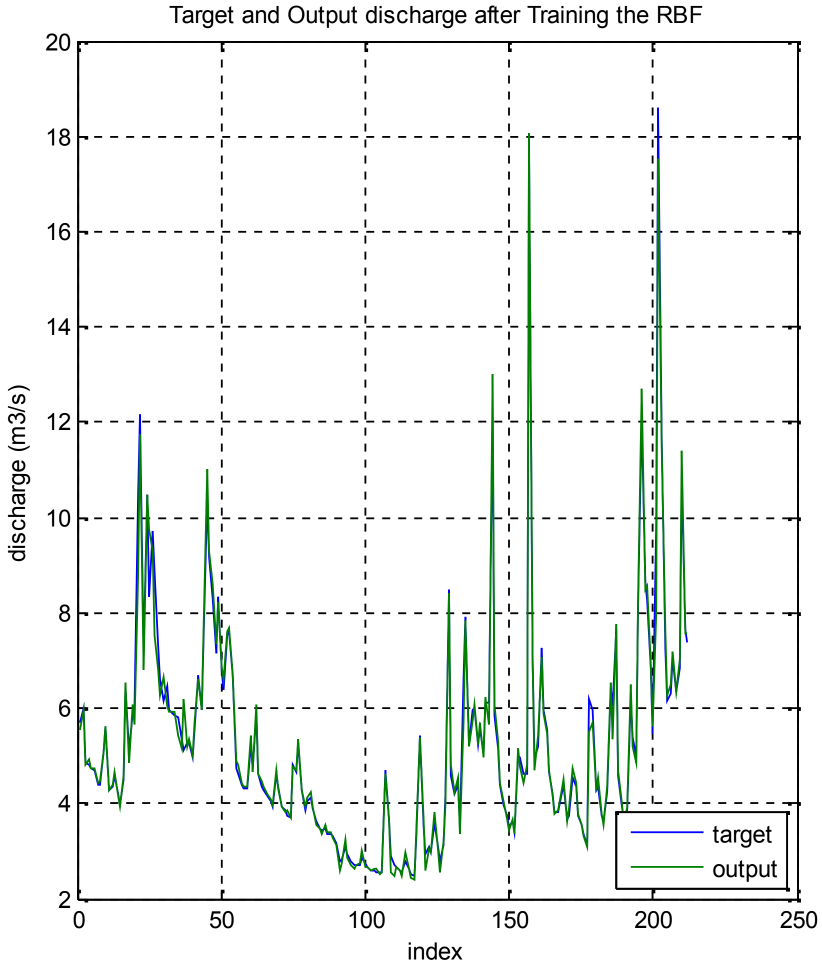
Name	Value
Et	<1x100 double>
Ev	<1x212 double>
LTImax	3
M	19
MSE	4.7673
Mu	<3x19 double>
N	314
Nb	19
Nt	100
Nv	212
PHI	<19x212 double>
Phi	<212x19 double>
Q_LTImax	3

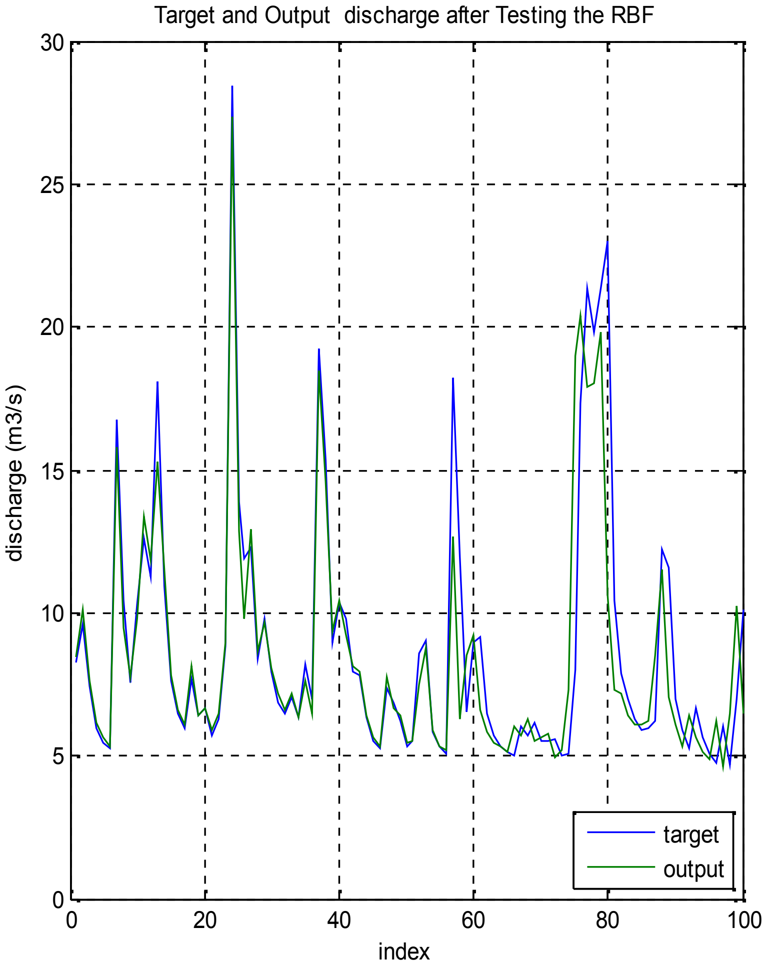
```
%-- 11/24/14 10:10 PM --%
   105
   100
%-- 11/24/14 10:20 PM --%
   105
   18
%-- 11/25/14  4:15 PM --%
   100
   19
%-- 12/8/14  9:58 AM --%
   133
   19
%-- 12/10/14  5:10 PM --%
   100
   19
```

Matlab coding



Graph generated from Matlab


Target and Output discharge after Training the RBF


Target and Output discharge after Testing the RBF