Water Flow Prediction in Perak River using Thin Plate Spline Basis Function Neural Network

By

Ahmad Fakharuden Yahya Bin Abd Rahim

15190

Dissertation submitted in partial fulfilment of

The requirements for the

Bachelor of Engineering (Hons)

(Civil)

JANUARY 2014

Universiti Teknologi PETRONAS

Bandar Seri Iskandar

31750 Tronoh

Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

Water Flow Prediction in Perak River using Thin Plate Spline Basis Function Neural Network

By

Ahmad Fakharuden Yahya Bin Abd Rahim

15190

A project dissertation submitted to the

Civil Engineering Programme

Universiti Teknologi PETRONAS

in partial fulfilment of the requirement for the

BACHELOR OF ENGINEERING (Hons)

(CIVIL)

Approved By,

_____

(Dr.Muhammad Raza Ul Mustafa)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

August, 2014

ABSTRACT

Radial Basis Function Neural Network (ANN) technique has been found to be one of the most powerful tool use to predict the values of water discharge in Perak River. This technique has been proven to be the best alternatives to replace the previous forecasting technique such as Linear Regression Analysis and Flow Rating Curve which are less suitable to be applied to predict the non-linear stage and discharge data. The specific discharge data analysed from the developed Thin Plate Spline Basis function were important and crucial for the operational of river water management such as flood control system and construction of hydraulic structures, hence contribute towards the relevancy of this research paper. The data of the water level which were used as the input and discharge as the output were equally important for the training and testing purpose and those are taken for the three most recent years of 2011, 2012 and 2013. 780 data was used for the training whereas the remaining of 190 data was used for the testing purpose before run the analysis using the MATLAB software. At an optimal number of spread at 1.6607 and 30 hidden number the model architecture of using thin plate spline basis function showed a higher predictive performance than the normal Gaussian method at 0.986 for testing which is slightly lower than the training and Root Mean Square (RMS) of 2.310 which lower than the training due to the marginal difference in the minimum and maximum value of data. The comparison between the result obtained with the common kernel function used such as Gaussian shows that Thin Plate Spline Basis Function produce a more satisfactory result. Hence, the application of the thin plate spline basis function is recommended for the application in the other hydrology or non-hydrological field in future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Perak River is the second longest river in the Peninsular of Malaysia. The river which records a flow for over 400 km and covers a catchment area of 15,000 km^2 bring significance contribution towards the advancement and development of the country and nation in Malaysia. For example, it serves to provide water irrigation to the nearby paddy field, preserve and conserve the ecological system, strategic business location, electricity generation (construction of Temenggor dam) and for river operational and management system. Therefore, the study of the Perak River flow is crucial to ensure the river could maintain for its function and also able to overcome the flooding issues in this country due to the inconsistencies of river water levels and tides fluctuation.

Conventionally, river flow is measured using tedious and complex conceptual model such as curve fitting and regression model. However, it was identified that these methods tend to produce less accurate and inconsistent prediction result. Therefore, it has come towards concern to develop a technique which can produce a high accuracy, reliable and efficient result compared to the previous technique.

Recent studies show that there are several methods which can be applied and used in forecasting the river water flow. However, one of the simplest and practical techniques is through the use of neurocomputing or numerical modelling technique known as Artificial Neural Networks (ANN). In fact, this method has been applied in many study areas and activities worldwide such as in flow forecasting, pollution simulation and parameter identification (Jain & Chalinsgaonkar, 2000).

In general, Artificial Neural Network applies the same concept as human brain, where it consists of billions of neurons or nodes which are arranged in the form of layers. The signals received from each pre-determine layer will process the information from the supplied input to produce the favour output result. The unique

part of this technique is the ability of the system to generalize and learn from the examples and the input sourced in order to predict or forecast and modified the out coming result as close as possible to the targeted outcome (Jain & Chalinsgaonkar, 2000).

As a matter of fact, the accuracy of ANN increases with the increase of the input data. Since ANN system covers a large application and areas. In this research paper, the ANN technique which will be used for hydrological measurement will focus much onto the application of the Radial Basis Function (RBF) using thin plate spline algorithm.

1.2 Problem Statement

Numbers of research and study has showed that there are several forms of technique which have been used to predict the water discharge in the river. However, in previous days, researchers tend to apply the conventional flow rating curve method such as linear regression analysis to predict the stage and discharge in the river. This linear method normally require the application of linear formula such as $y = mx + c$ to form a linear relationship between the stage and discharge before the data could be further analysed. However, this such of forecasting techniques is found to be in appropriate and tend to produce less satisfactory result since the stage and discharge is in a non-linear form in their nature due to the variation of time and river basin capacity volume. Therefore, it has come towards concern to shift to the application of non-linear technique such as radial basis function neural network using thin plate spline (TPS) in order to forecast the non-linear form of gathered data. Besides, it has been identify that the used of thin plate spline basis function has never been performed previously in any other places around the world including Malaysia. In fact, the modification of this technique using the thin plate spline radial basis function (RBF) is necessary as an alternative to obtain higher predictive performance which could produce similar or better result than the common Gaussian algorithm or any other conventional approach (Refer to Appendix 2)

1.3 Objectives of Study

The objectives of this research study are listed as follows:

1) To develop Thin Plate Spline Basis function Neural Network model for the prediction of water discharge at Perak River.
2) To evaluate the performance of the Thin Plate Spline basis function model using different statistical performances measures.

1.4 Scope of study

The scope of study area is limited towards forecasting the flow in the Perak River by using Radial Basis Function neural network using Thin Plate Spline (TPS). Feed forward network is one of the model architecture inside the artificial neural network which comprised of more than one types of architectures such as multilayer perceptron (MLP), Support Vector Machine (SVM), Generalised Regression Neural Network, Radial Basis Function, Neurofuzzy and others. However, the scope area in this study will focus only into the application of Radial Basis Function Network. Few types of function listed inside the RBF namely Multiquadric (MQ), Gaussian (EXP), and Thin Plate Spline (TPS) and logarithmic have been known to perform their own specific algorithm and function. However, among of these functions, thin plate spline is chosen in this research study to be applied in developing the selected basis function model using MATLAB computing software.

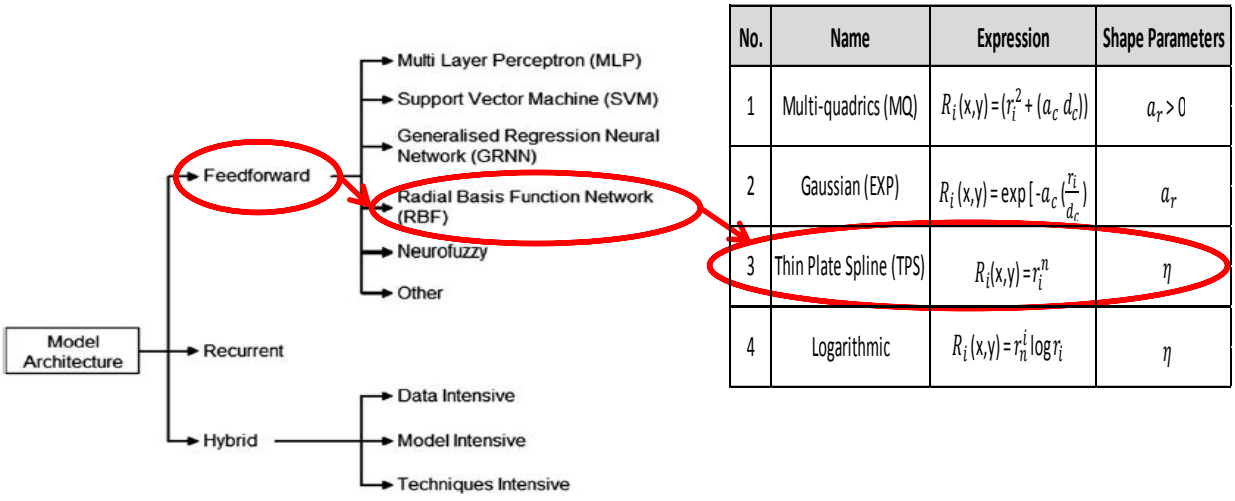


Fig. 11. Taxonomy of model architectures.

No.	Name	Expression	Shape Parameters
1	Multi-quadrics (MQ)	$R_i(x,y)=(r_i^2+(a_c\,d_c))$	$a_r>0$
2	Gaussian (EXP)	$R_i(x,y)=\exp[-a_c(\frac{r_i}{d_c})$	a_r
3	Thin Plate Spline (TPS)	$R_i(x,y)=r_i^n$	η
4	Logarithmic	$R_i(x,y)=r_n^i \log r_i$	η

Figure 1: Taxonomy of model architecture

1.5 Relevancy of the project

It is very important to determine the predicted discharge of the river flow in order to address the related issues such as flood impact and Hydraulic Structures construction project. However, the conventional method applied to gather and monitor the forecasted data was time consuming and less effective. In addition, less reliable and accurate result would affect the reliability of the data gathered. Hence, the application of this technique using soft computing measure such as MATLAB to predict the river flow using radial basis function is relevant to overcome the lack in the previous technique as it is lower in cost and less time consuming because it use less number of manpower to conduct the process. Furthermore, the process are much involved with the application of software such as MATLAB and Microsoft EXCEL which could be handled by few people or even one or two person only.

1.6 Feasibility of the project within the scope and time frame

The measurement and the collection of the data such as rainfall, water level and discharge could be time consuming since it require one to measure the flow using specific equipment for certain duration of time. Fortunately, these data for the Perak River flow were directly obtained from the Department of Irrigation and Drainage (DID) of Malaysia, which has turned the progress of the fieldwork to be much easier, faster and low in cost especially for the purpose of the data analysis and soft computing model. Perak River was chose as the study area because of its location which is near to the University (Research Centre) (Refer to Figure 6). Hence, any works which require the transportation and mobilization to the site for the purpose of pictures collection (Refer to Appendix 1) will be much easier and efficient. In addition, the fluctuating of river water level due to raining and draught season has great implication on the river basin capacity volume which has turn the flow data to be non-linear thus make the research using the radial basis function is relevant to be conducted at the river. Being the main river that runs through several major towns of Perak such as Batu Gajah, Pusing, Ipoh and Pasir Putih, the contribution of Perak River is significant towards the construction of infrastructures such as bridges and culverts thus require necessary effort to get the reliable and accurate river flow data to serve for these purpose. In fact, a proper study on Perak River would improve much in managing the hydrological structures.

CHAPTER 2

THEORY/ LITERATURES REVIEW

2.1 Literature review

There are numbers of research which proposed for several application of forecasting technique to predict the water flow in the sea and river. In fact, an accurate and reliable forecasting technique is vital to maintain the operational river management as well as to prevent or minimize the flooding impact onto the people who stay nearby to the river area. Therefore, in order to address and overcome the issues, it was found that Artificial Neural Network (ANN) could be one of the best solution and powerful tool designed to achieve the objective of the study conducted. Review and analysis of the result from the previous researches paper found that this system brought lots of positive impact and advantages from the aspects of reliability and accuracy. However, there were also several minor arguments recorded in those literatures which could influence one perception towards the reliability of those techniques in solving the issues arise.

In general, Artificial Neural Network (ANN) is a technique which comprised of a very complex networking system. According to Jain and Chalisgaonkar (2000), ANN was designed in a very special way to imitate the function of human brain which consists of billions of interconnected neuron that promotes a unique interconnection between the layers. For instance, the input data which denoted by x is transferred through the input and hidden layer of i and j respectively before reach the output layer of k as a vector, z (Refer to Figure 2) (Supharatid, 2003). This finding was also supported by Lippman (1987) which briefly stated that ANN is a complex network which consist of a large set of simple neural cells. In fact, this network system had demonstrates a general topological structure which could map the input and output vectors through a combination of nonlinear function through a non-linear and linear transformation of information through the network (Chat and Abdullah, 2002).

However, there were a limit in the work scope of ANN, a comprehensive review made by ASCE, 2000 found that even though there were extensive application of ANN in the hydrological engineering, ANN cannot be treated as a replacement for the other hydrological modelling technique due to the reason that the physics of the basic or foundation process in the system was confidentially stored in the optimal weight and threshold value and never been exposed to the user even after the end of the training stage. Therefore, thorough studies regarding the application of ANN must be done in order to ensure that this system will able to meet the designed objective.

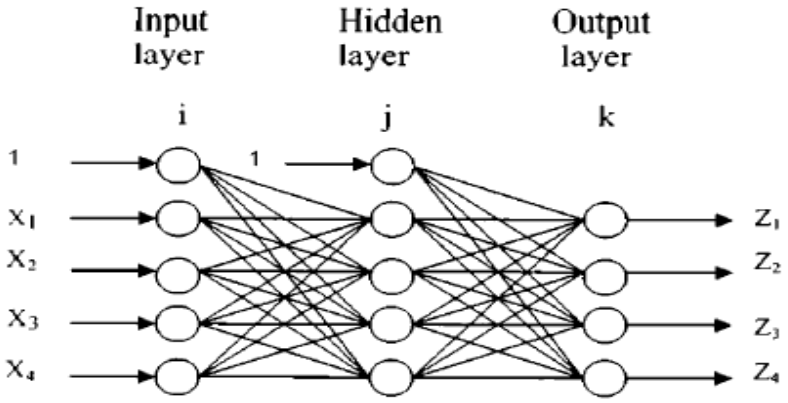


Figure 2: The interconnection between nodes (Source taken from Supharatid, 2003)

Despite of its complexity structures, ANN is designed to meet several purposes and target to solve few hydrological flow forecasting problems. According to Zhou and Han (1993) the principle of the existence of ANN is to address the issues of flooding event. The process could be implemented through applying the algorithm of the neural networks such as conjugate gradient descent, back propagation, thin plate spline, etc. by inserting the load of past input data, neural cells and noise without necessary to design any mathematical models (Brion & Lingireddy, 2003). However, it is difficult to describe these few variables using the others network such as linear regression analysis function during flood condition due to the non-linear form of this data variables. Therefore, ANN is found to be the best alternative to solve the problems since artificial neural network was able to complete the information in the network through parallel interaction between the neural cells and non-linear information transfer. Furthermore, the system also requires special learning process to enable the process of mapping the variables to be possible in order to produce accurate result (Feng and Lu, 2010).

Indeed, ANN is very flexible and unique due to its ability to learn and adjust the computation by its own. It might be true that ANN could learn by itself, however, as a matter of fact ANN cannot learn without the prior knowledge inserted into the input layer before data normalization process took place (Jain and Chalinsgokar, 2000). In the other words, ANN undergoes its learning process by using a set of input and load of output targeted vector which were important in the training set upon selected at the beginning of the process. In the first stage of training, the weight which contained inside the nodes will go into the process of normalization through initialization of the network weights by using some previous data or input (Jain and Chalinsgokar, 2000). The learning function will then modify or adjust the weight in the network based on the difference between the computed output and the targeted output value which fall within the permeable value limit (e.g 0.0 – 1.0). Upon the completion of the optimization stage, the set of weight will then be considered as the learning set which represents the knowledge regarding the specific problem. In fact, the subject has been strongly supported in the thesis paper wrote by Li Hua Feng and Jia Lu (2010). Due to its ability in self-learning, self-organization and self-adaptation ANN has been successfully adapted for pattern recognition. In fact, the functional relationship between the input and the output could easily be obtained using the sets taken from the final training set (Feng & Lu, 2010).

Normally, the weights inside the node are adjusted through the function known as back propagation method (BPN). This function is known as back propagation since the learning process take place in both forward and backward direction through the network (Feng & Lu, 2010). In fact, this algorithmic function has been widely used during the training to adjust the interrelation between the weights (Rumelhart. et al, 1986). Previously, it was mentioned that a set of input and output were selected for the training and from there network will compute the output based on the input inserted and result obtained will be subtracted from the targeted outcome to determine the output layer error, refer to Equation (1). This error will then be used to adjust the weight inside the nodes and the propagation of calculation will took place in numbers of iteration until the target value is achieved (Jain and Chalisgankar, 2000).

On the contrary, it was determined that Back Propagation method is subjected to certain limitation which in turns contributes to the weakness of this method. The BPN tend to undergo slow convergence along the network, therefore an effective ways to overcome this problem is through the application of Levenberg-Marquardt algorithm method.

$$E_p = \sum_{k=1}^{N} (t_{pk} - z_{pk})^2 = \sum_{k=1}^{N} v_{pk}^2 \quad (1)$$

E_p = Total Error
t_{pk} = Targetted output
z_{pk} = Output predicted
v_{pk} = Error of Output unit k for p data pattern

Equation 1

One of the important parts in the ANN system is the determination of the hidden layer numbers (Refer to Figure 2). Previously, it was mentioned that there are three sets of layer which consist of the input, hidden and output layer respectively in which each layer consist of billion number of neurons. Normally, the number of nodes in the hidden layer was determined using the application of Kolmogorov's theorem whereby the least number of nodes should follow the formula of 2n+1 (where n represent the number of nodes in the input layer) (Feng and Lu, 2010). In a different view, according to Mustafa. et.al, 2014 the application of trial and error method to determine the number of neurons in the hidden layer has been found to produce a better result as compared to the existing conventional regression analysis method. In fact this method is vital to ensure that during the training, the configuration set which gave the maximum Efficiency Index (EI) and minimum Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD) is selected and this must be done with reference to the minimum allowable number of hidden nodes (Shamseldin, 2010).

Hence, it is proven that ANN did promote lots of advantages compared to the conventional regression approach. As a matter of fact, the more number of input information is added into the system, the more coefficient of correlation improves and more errors could be minimized. According to Jain and Chalinsgokar (2010) the error calculated based on the equation (1) constructed in theirs paper found that the error recorded for the training and testing data was enormously smaller compared to the conventional method (Jain and Chalisgaonkar, 2000). In the table 1 presented in

their paper (Refer to Appendix 3), it shows that the error calculated become less as more water level, H and discharge, Q data were inserted as the input hence proved that ANN could be used as the perfect tool to map the loop rating curve or hysteresis effect.

Narrowing into the ANN structure system, Radial Basis Function Neural Network (RBFNN) can be considered as one of the important component which lying inside the feed forward structure. It is much similar to the mother neural network structure components since it retained the same engineering layer concept which comprise of three type of layer known as the input, hidden and output layer and both of the input and output layer play major roles in assigning the input data and transforming the response of the network into output result (Fernando and Sahmseldin, 2009).

However, in the RBFNN the main uniqueness lying in the structure of the hidden layer and the output layer. The hidden layer consisted of non-linear function which has its own specific function shape. According to Kasiviswanathan and Agarwal (2012), in their research paper, they mentioned that the function node in the RBFNN is different compared to the one applied in the Back Propagation Neural Network. It does not implement the same mechanism of multiply and add of the weighted summation, instead, it computes a respective field from the individual function overlaps. In addition, the function nodes is not a problem dependent function since it rely heavily on the network designer on how to set up the function based on the model performances (Kasiviswanathan & Agarwal, 2012).

Whereas, the output layer is normally consist of only one node. As a matter of fact the numbers of nodes in the output layer in RBFNN depend solely on the variables fixed by the designer. On the other hand, it was known that RBFNN has a higher reliability, faster convergence and interpretation and produce very small error compared to the conventional multilayer perceptron. However, among the types of RBFNN, Gaussian methods is the most favourable and commonly be used and they are characterized by identifying the specific centre and spread value (Ruslan et al., 2013).

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 denoted by the symbol Φ(x) and this function is responsible in transforming the information received from the input layer into the output response (Kasiviswanathan & Agarwal, 2012). In fact, in order to complete the transformation at the output response, linear transformation will took place between the hidden and the output layer and the weight is linearly sum up before projected the output value in the output layer node by the sigma and spread. According to Kasiviswanathan and Argawal (2010), the performances of RBFNN and the activation function are critically rely on the centre position and spread which indicate the radial distance of the RBF centre. Since the spread value have much influence on the activation function, it is best to know that the higher spread value will produce larger and scattered data point from the centre which will reduce the maximum function response. An example of performance comparison recorded for larger and smaller spread value is shown in the table in Figure 3. Based on the table, for both model of network of 4-4-1 and 2-24-1 the best performance is recorded by the optimal spread value of 1.0 and the lowest is shown by the higher spread value at 2.5. Therefore it shows that the value of spread must be properly determined for better performance of the network. According to Ruslan et al. (2013), in order to calculate the radial basis function, RBF Kernel Function and spread will be applied onto the value of the Euclidean distance measure by the hidden neuron from the neuron's centre point.

Figure 3: The comparison of spread performances

Spread $2\sigma^2$	Model performance at 1000 iteration			
	Calibration		Verification	
	RMSE (m^3/s)	CC (%)	RMSE (m^3/s)	CC (%)
	Network 4-4-1			
0.5	145.6	69.6	127.0	72.4
1.0	138.4	72.9	115.9	76.6
1.5	145.4	69.2	119.6	70.9
2.0	148.7	66.3	121.6	68.4
2.5	155.3	64.0	125.2	66.4
	Network 4-24-1			
0.5	151.8	68.2	119.6	71.5
1.0	132.5	75.0	113.2	78.5
1.5	139.0	72.5	124.8	73.6
2.0	146.4	70.4	130.5	70.1
2.5	154.6	69.1	138.4	68.4

Highest Performance at optimal spread value of 1.0

Lowest Performance at optimal spread value of 2.5

In RBFNN, the important stages such as the training, testing and validating play a significant role in determining the predictive performance of the model architecture. According to Maier et al. (2010), in order to develop an effective ANN model the training set is normally used to determine the unknown weight connection, whereas testing is used to determine the stopping characteristic of the model in order to avoid over fitting while validation is used to evaluate the reliability of the model developed.

In training, there are two mechanisms applied for the non-linear transformation between the input and hidden layer. First, the weight in between those two layers is monitored using the unsupervised training and second is the transformation of information from the function layer to the output layer where the process will be monitored using supervised training. In fact, the training process involved the calculation for the centres, widths and weights. According to Maier et al. (2010) there are several methods and ways in determining the centres. One of the most famous and less tedious ways is by using random selection, other methods such as mathematical algorithm (genetic algorithm or least square learning algorithm) could also been applied as an alternatives. In addition, least squares learning algorithm can also be performed to adjust the connection between the hidden layer and the output layer after the determination of the centre and weight is complete.

In any function including radial basis function it is necessary to measure the performance of the selected network by using specific statistical performance measures. According to Mustafa et al. (2012), the model architecture's performance were measured by using error basis measurement such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of efficiency (E), Mean Squared Relative Error (MSRE) and coefficient of determination (R^2) to indicate the overall performance of the selected network. However, conflicts exists when the squared error metrics were dominated by the errors of high magnitude which then lead to over fitting problem in the model especially during the high flow and negligible value during the low flow condition (Maier et al., 2010). Eventhough the error measured using absolute error is based on the absolute difference between the actual and the modelled output data, the application did not provide the performance information of the model selected in term of overall under or over prediction.

Therefore in order to compare the outputs of different magnitude more easily, it is recommended to consider the application of relative errors metrics such as Average Absolute Relative Error (AARE) and Normalized Root Mean Square Error (NRMSE) (Maier et al., 2010). At the end, the best selected model is indicated by the minimum time of training and for the one which could produce the least error in total (Refer to Figure 4).

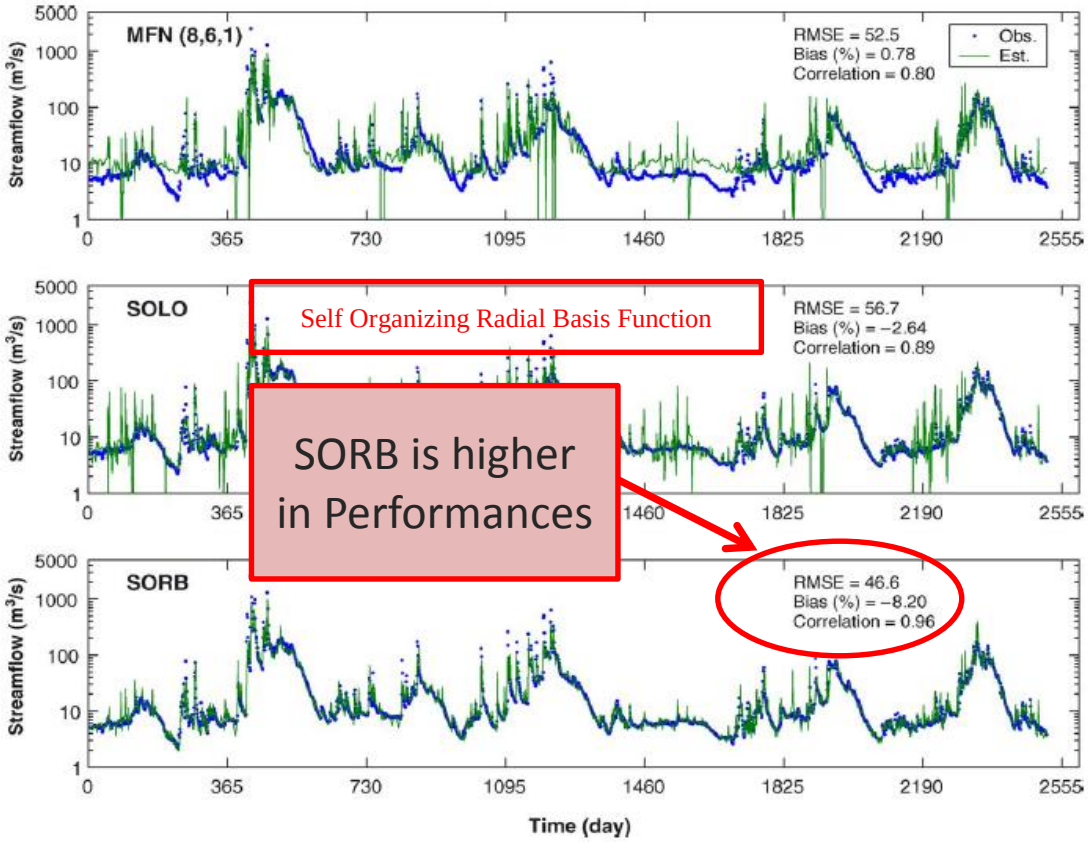


Figure 4 : Comparison between model architecture

Based on the literatures reviews above, ANN technique used has been proven to show improvement in the water flow forecasting techniques in comparison to the conventional method. This technology is very important in river flow calculation process since the stage, discharge and other non-linear hydrological variables play significant roles in determining the correct discharge value from the inserted stage data. The application and the development of radial basis function seem to bring more advantages in producing the accurate outcome result for the betterment of hydrological research study.

CHAPTER 3

METHODOLOGY / PROJECT WORK

Since the development of radial basis function neural network (ANN) structure involved much in modelling and simulation of the data obtained, the methods which are used throughout the research will revolve around the application of software such as Microsoft Excel and MATLAB using stage data as the input and discharge as output obtained from the Perak River.

3.1 General Process Flow of ANN Development

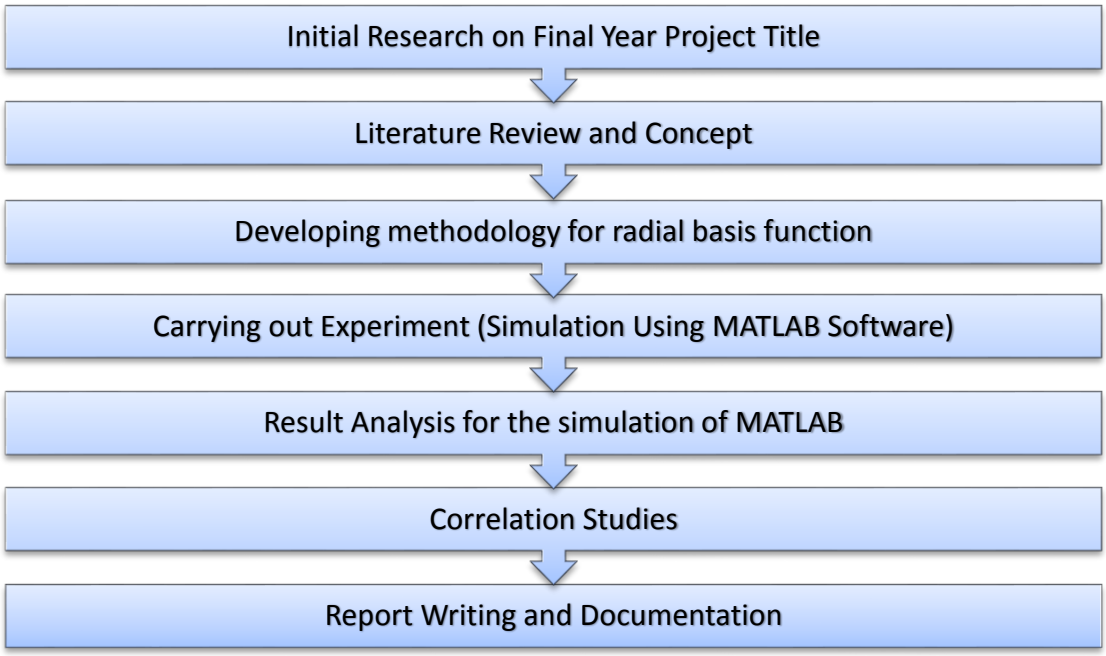


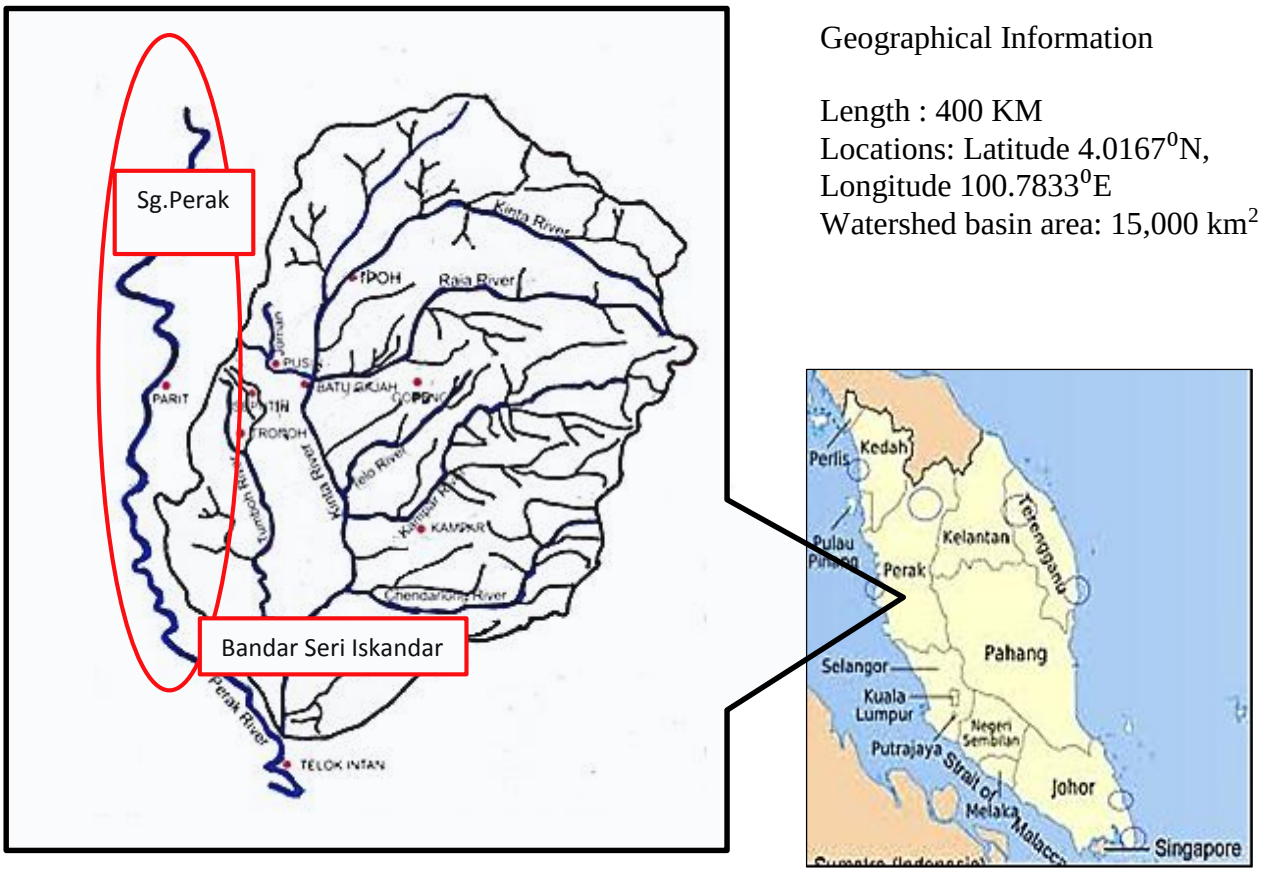
Figure 5: General Flow of Methodological process

3.2 Data Source and Study Area

In this case study, groups of hydrological measurement data were obtained from the records provided by the Department of Irrigation and Drainage (DID), Ministry of Natural Resources and Environment, Kuala Lumpur. In actual, the data were measured at several specified locations and stations at Sg.Perak. The study area for this research can be clearly identify from Figure 6.

Based on the data given, the records consist of two variables of hydrological resources which are Water Level (WL) and Discharge (DC) respectively. Each of these data comprises of their own specific value and unit (m and m^3/s respectively) which were tabulated into group form according to subsequent years onwards starting from year 1990 until the recent year of 2013. The daily data were tabulated according to the months from January until December for each and every years using the software of Microsoft Excel. However out of these 23 years historical data only the three most recent years of data of 2011, 2012 and 2013 were chose to be presented into graph and table form due to the recentness and relevancy factors.

The water level and discharge data were measured in daily basis throughout the 12 months in a year. Based on the observation made, it was found that the minimum discharge value is recorded at 125.8 m^3/s and maximum at 394.7 m^3/s for the year of 2011 to 2013. Whereas for the water level the minimum value is recorded at 31.67m and maximum at 33.5m.



Geographical Information

Length : 400 KM
Locations: Latitude 4.0167^0N,
Longitude 100.7833^0E
Watershed basin area: 15,000 km^2

Figure 6: Location Map of Study Area at Sg.Perak River

3.3 Development of Radial Basis Function Model

3.3.1 Input Data Selection

The input and output data is very important for the training and testing stages (Mustafa et al., 2012). This method is primarily done based on the priori knowledge and or availability of the stage and discharge data recorded which is the water level and discharge respectively.

Thousands of hydrological data were received from years of 1990 to 2013 from the DID department for each of the hydrological variables (Water Level and Discharge). The data were then transferred into a more organize form in Excel sheet format for the specific calculation process. Out of these numbers, data from years 2011 to 2013 were chose for each variables due to its recentness and completion of data set structure. In fact, it is vital to ensure for the selected data to have a complete and consistent data set since it will affect much the accuracy of the result obtained. Moreover, loopholes or missing dataset are probable to produce skew and scattered data which eventually increase the complexity of the learning process of the RBF model.

3.3.2 Partitioning of Data

The partitioning of the data for training and testing were particularly done according to the data trend (Figure 8). Based on the plotted graph in Figure 8, the minimum data set number for water level is recorded at 31.67m on 23 and 24 February 2013 and the maximum water level recorded is 33.5m on 15 April 2012. On the other side, the minimum number of discharge was recorded at 125.8m^3/s on 24 February 2013 and the maximum at 394.7 m^3/s on 15 April 2012. Hence, in order to select the data for the input load, it is important to select the data range which included for this both criteria of minimum and maximum value for the next data analysis stage for both training and testing. Hence, 780 data were chose for the training from the total of 970 data of water level and discharge and the remaining of 190 data were then used for the testing purpose throughout the development process.

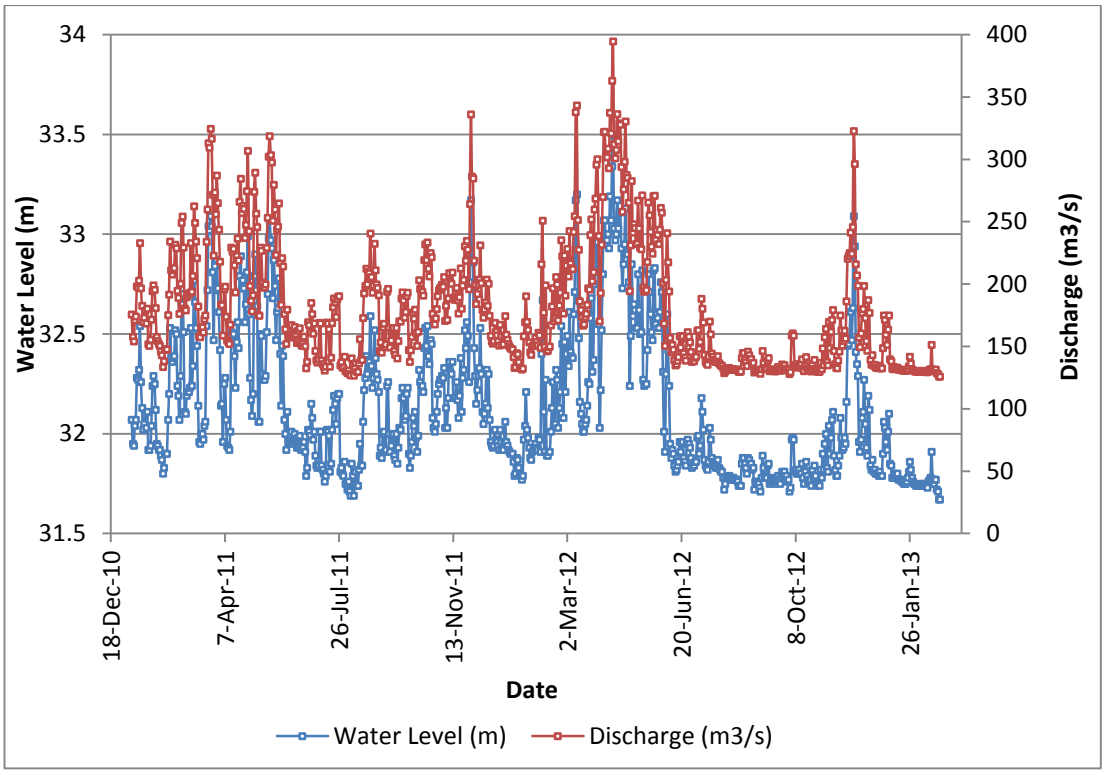


Figure 7: Time Series for water level and discharge

3.3.3 Statistical Data Analysis

From all of these number of data set, 780 data which were chose for training and the remaining of 190 data for testing will undergo statistical data analysis. In fact, the statistical data analysis of the input and observed output data are performed for both training and testing data to examine the complexity involved in the data set. The statistical data is analysed by using parameters such as Mean, Variance, Standard Deviation (SD), Minimum and Maximum value before transferred it into the table form for interpretation purpose

Parameters	Training (Jan 1, 2011 - Feb 24, 2013)		Testing (2013) (Feb 25, 2013 - Sept 3, 2013)	
	Water Level (m)	Discharge (m^3/s)	Water Level (m)	Discharge (m^3/s)
Mean	31.5	181.87	23.9	96.51
Variance	601.09	1111794.2	962.42	479634.9203
Standard Deviation	24.52	1054.42	31.02	692.56
Minimum	31.67	125.8	31.69	126.8
Maximum	33.5	394.7	32.33	207.9

Table 1: Summary of the Statistical Data Analysis

Table one above shows the summary of the statistical data analysis performed after each of the specific parameters has been determined. The determination of these statistical data analysis is vital in order to construct the graphs of relationship between predicted and observed data in the next stages and also to foresee any changes or factors that could affect the predictive performance of the developed model of basis function.

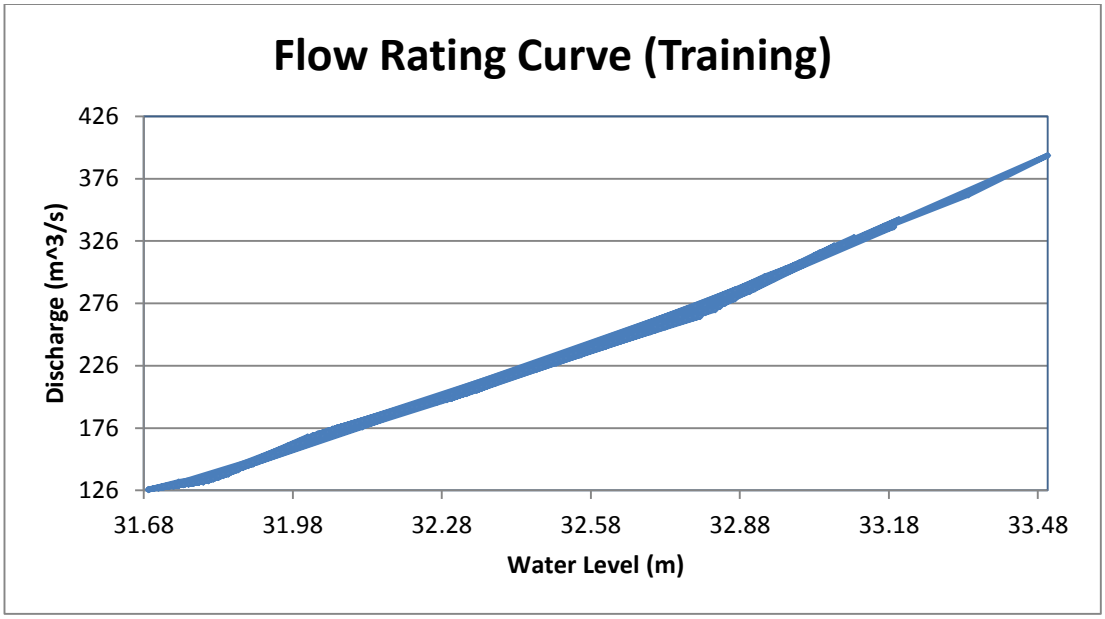
Based on the table 1 summary of statistical data analysis, the value of the maximum water level and discharge for both training and testing were found to be proportionally increased due to the weathervane factor such as high frequency of rainfall event. Subsequently, for the discharge value of training, the difference between the maximum and minimum value is comparatively higher at 268.9 m^3/s compared to the testing at 81.1 m^3/s. This is due to the effect of volume of water that pass through the basin area which increase as the result of the rainfall.

The higher variance of discharge value for training indicate that the values of the data were randomly separated from the mean value discharge data due to the distance of location and recorded time, whereas, the low difference of discharge in testing value implies that the capacity in term of the stream flow in which the area (Sg.Perak) can hold is larger during the testing compared to the training period. Thus, it showed that the testing could produce skewer tabulation graphs compared to the training. Hence, this river could experience bad flooding impact due to the heavy rain condition thus causing low variance of discharge data during the testing period.

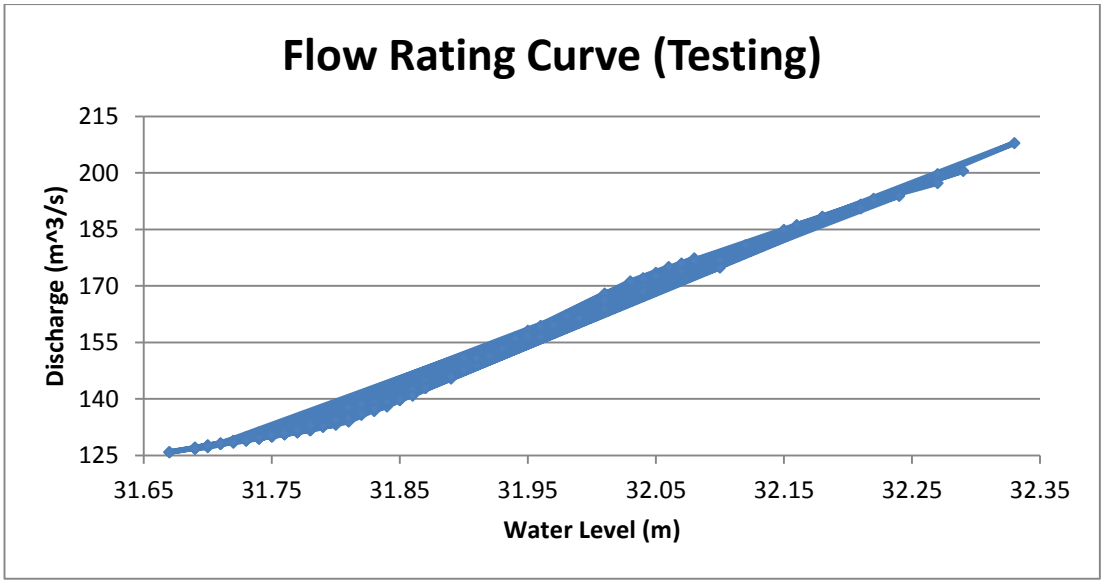
From the table, it was also found that the value of minimum and maximum value of water level and discharge for testing is inclusive in the range of minimum and maximum data set of water level and discharge for training. Therefore this shows that the training is much complex compare to the testing thus enable the system to perform better during the testing process.

Conventional Flow Rating Curve for Training and Testing

After the completion of statistical data analysis in methodological part, the result from the tabulated discharge and water level data in the Excel sheet for training and testing were used for the establishment of flow rating curve model. The development of this model is to show the demonstration of the calculation using the conventional method of linear relationship model. In fact, this model is developed using a linear formula of (y= mx +c). The flow rating curve graphs are shown as below.



Graph 1: Conventional Flow Rating curve of Discharge and Water Level for Training

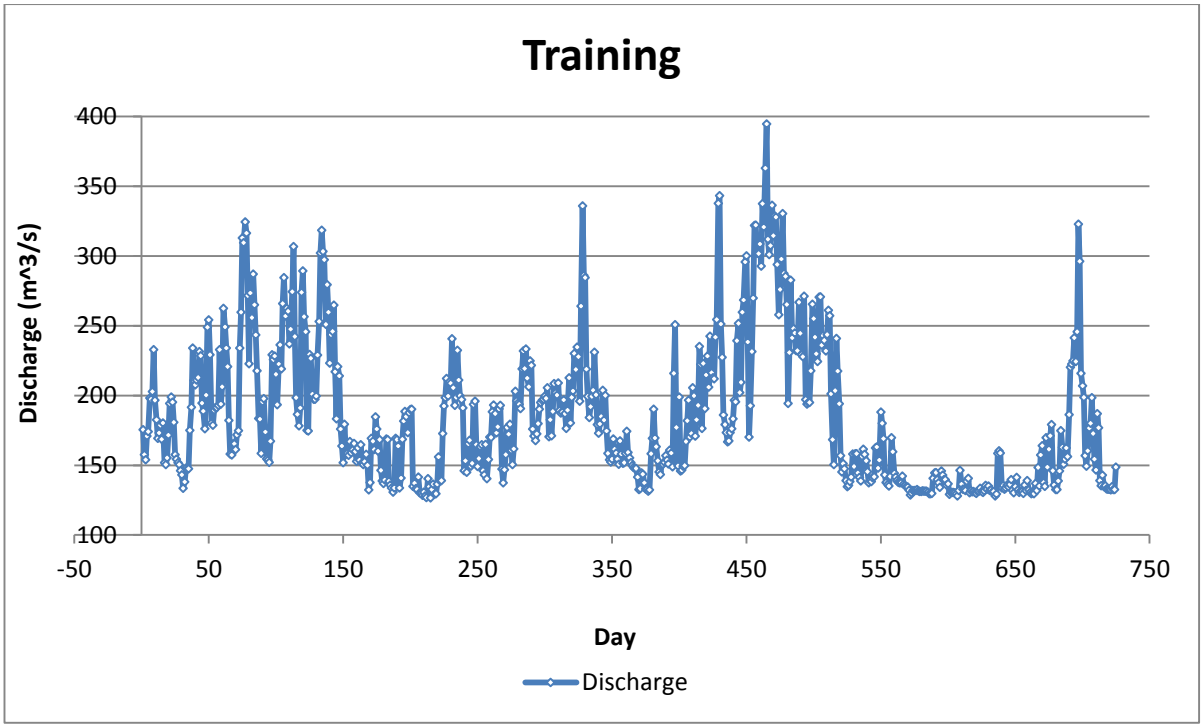


Graph 2: Conventional Flow Rating curve of Discharge and Water Level for Testing

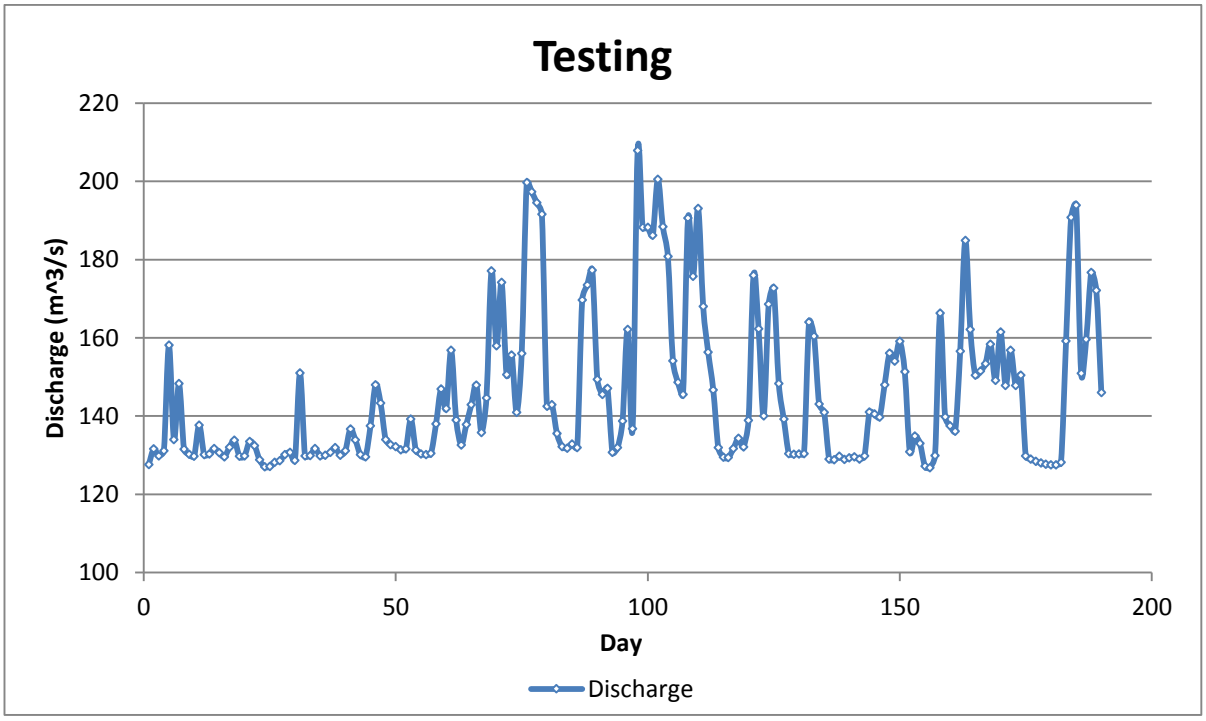
The designed graph above shows the relationship between the water level (m) and the discharge data (m^3/s) which was plotted for both training and testing on the Microsoft Excel sheet using the linear relationship formula Y = mX + C. The x-axis denotes for the water level data (the elevation of water surface from a specified datum) whereas, y-axis represented the discharge value (the amount of water that flow throughout the basin area), m is the gradient form between the two axes and C is the interception of y-axis at zero water levels.

As per discussed in the problem statement, the application of linear flow rating curve is less suitable to be applied with the non-linear data such as discharge and water level since it will produce a linear development between the two variables. Observation shown from the graphs show that the discharge value increase proportionally with the increase of the water level , instead in real situation due to variation in time and fluctuating value of water level the relationship shown should be in a non-linear form. Hence, the results demonstrated promote a higher tendency to produce less accurate data which then require the application of non-linear techniques such as radial basis function neural network at the Perak River. For comparison purpose, the development of radial basis function in the further stage will involve with the application of thin plate spline algorithm to compare with the linear flow rating curve data before assessed using specific statistical performance measure.

Daily Discharge Hydrograph



Graph 3: Daily Hydrograph of Discharge vs Day for Training



Graph 4: Daily Discharge Hydrograph of Discharge vs Day for Testing

Based on the two daily discharge (hydrograph), of the Perak River shown above for both training and testing period, it was found, from the days of 50 until 150, the discharge has recorded for a vigorous fluctuation trend which is mainly due to the inconsistencies of the rainfall event. However, the discharge value increase drastically from the day of 450 until 460 thus recorded for the maximum discharge value at 394.7 m^3/s on the day of 465 before fluctuating again. The value of discharge then record for a gradual decrease along the days after. Above those period, at the days of six hundred the discharge record for the lowest value which is at 128 m^3/s before increase again until 322.8 m^3/s on the day of 695. Whereas, for the training, the highest discharge is recorded at 207. 9 m^3/s on the days of 98, this value is much lower than the testing due to the numbers of included data. In fact, the lowest discharge value recorded during training is 127.2 m^3/s on the days of 155. The hypothesis show that, a smaller marginal difference between the maximum and minimum discharge value will tend to produce more consistent water flow prediction

3.3.4 Normalization of Data

The normalization of data is important to ensure for fast convergence and minimization of global error during the network training (Rojas, 1996). On the other words, it is a process in which the data set is scaled in order to optimize the accuracy of the numerical computation by reducing redundancy hence minimizes the simulation failure (Mustafa et al, 2012).

As a matter of fact, in radial basis function the input must undergo data normalization process using specific formula applied in the MATLAB software in order to achieve stable conversion within the activation function limit in the function nodes (Maier and Dandy, 2000). However, in this research paper the data normalization were carefully conducted, this is to ensure that the data scaled within the range of the activation function, so the size of the weight adjusted will be almost negligible (Mustafa et al., 2012). Therefore in this paper, the data will be normalized so as to fall between the limits range from 0 to 1 by using the common normalization using equation 2.

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

The formula that is commonly been used to normalize the subsequent data is shown in the equation above. The current v_p symbol denotes for the normalized or transformed data series whereas the x_p is the raw data series such that $1 \leq p \leq p$ in which p is the number of data and x_{min} and x_{max} are the minimum and the maximum value of the original data series respectively which is in this case the data referred to the water level and discharge data series (Mustafa et al., 2012)

3.4 Selection of ANN Model Architecture

Since the radial basis function is used as the design model in this research paper, therefore, there are three layers (input, hidden and output layer) which consist of specific number of neurons that should to be determined in this stage. In fact, the selection of appropriate number of neuron in the input, hidden and output has a great significance on the accuracy of the model structure developed (Maier et al., 2010).

3.4.1 Selection of Input Layer

One of the selective methods used to identify the number of neuron in the input nodes of Radial Basis Function model is by using trial and error method through the training phase. There are two stages involve in the training stage in which at the first stage, the transfer functions are determine at the hidden layer which include the determination of spread value while in the second stage involve with the determination of centres and weights in the hidden and the output layer using the application of thin plate spline algorithm (Maier et al., 2010). From there it will undergo testing and validating process before the exact architecture of the model is finally determined which will consist of the numbers of input data selection.

3.4.3 Spread Coefficient

The spread value used throughout the process has been predetermined using the default equation in the MATLAB soft computing at spread, σ = 1.6607 which had known to produce the best performance measure compared to the other group of trial and error using different set of spread and hidden value (Refer to figure 9). In fact,

through the numbers of iteration of the trial and error, the optimal spread value at 1.6607 showed that the model has reach its best activation function within its cluster with a minimal scattered data from the line of agreement or the mean value (Refer to Figure 15). In addition, since the value of spread is higher than 1.0, the model performance should has increased throughout the training and testing stage (Mustafa et.al, 2012).

3.4.4 Hidden layer Selection

In this research paper the process of determination of the hidden layer was conducted through *trial and error* method. Based on the numbers of literatures review, trial and error method did produce an effective and fast result to select for the optimum number of hidden layer. In subsequent, Figure 9 show the exact method on how the hidden neuron is determined.

The process of trial and error was conducted using two main software which are Microsoft Excel software and MATLAB version 7.8.0. The partitioned data were loaded into the Excel sheet as part of the process to enable the selection process. The simulation will run automatically until the basic load graph appear. By entering the fixed value of testing and training data at 190 and 780 data respectively, the desired value of the hidden neuron will be requested. In this research paper, the number of hidden neuron is started with 4 and will increase by one neuron for the subsequent trials. This is mainly because, the hidden value of 4 is the optimal minimum number of hidden neuron to be inserted before the spread value could be identified.

In this methodology, the spread value, σ were made up through numbers of trial. Starting from 0.1 until 2 the spread value were changed for each iteration took place after the sequence. The iteration would only stopped after analysing the error produced through Mean Square Error (MSE). The number of hidden layer and spread which produced the lowest MSE value was selected as the best or optimum criterion for the model architecture. In this research paper, the optimum number of spread is found to be 1.6607 at the hidden layer of 30 (Refer to Figure 9).

No.of hidden neuron	Performance of Training		Performance of Testing	
	SSE	MSE	SSE	MSE
4	3317186.627	4263.736	400130.8709	2105.952
5	4449543.376	5719.2074	445690.1815	2358.1491
6	1429691.009	1837.6491	179124.5878	942.761
7	1154584.303	1484.0415	99855.0929	525.5531
8	2945284.381	3785.7126	340854.4082	1793.9706
9	1661321.264	2135.3744	444439.7851	2339.1568
10	900779.1956	1157.8139	89534.3915	471.2336
11	2077929.758	2670.8609	252105.5098	1326.8711
12	2878320.306	3699.6405	277400.7205	1460.0038
13	188602.1538	242.4192	21106.3907	111.0863
14	335442.9689	431.1606	28571.3723	150.3756
15	828986.1447	1065.5349	268465.8414	1412.9781
16	155238.3223	199.5351	14720.0643	77.474
17	147049.058	189.0091	33121.6335	174.3244
18	207724.3154	266.9978	77331.6459	407.0087
19	229488.7757	294.9727	34698.0631	182.6214
20	17399.4418	22.3643	2783.681	14.651
21	41161.384	52.9067	6660.0795	35.053
22	17236.4045	22.1548	3721.9753	19.5893
23	29241.2576	37.5852	4759.9252	25.0522
24	41246.1107	53.0156	5797.8751	30.5151
25	81376.6348	104.5972	14426.7214	75.9301
26	47071.3627	60.503	16171.372	85.1125
27	12766.0906	98.68368	22505.4864	110.55145
28	21539.1815	113.05452	27692.23485	133.11104
29	6307.3506	8.1071	1364.284	7.1804
30	6680.8687	8.5872	1014.2007	5.3379
31	19644.8069	25.2504	4030.5935	21.2137
Minimum Value of Error	6307.3506	8.1071	1014.2007	5.3379

Figure 9: Determination of the optimal number of neurons in hidden layer using the trial and error approach

Parameter	MSE			
	Lowest Error Value		Highest Error Value	
Hidden Number	30		5	
Phase	Training	Testing	Training	Testing
Value	8.5872	5.3379	5719.21	2358.15

Table 2: Analysis of trial and error method

Table 2 above shows the simplified form of the trial and error made in Figure 9. From the table above it was found, the lowest value of Mean Square Error (MSE) produced during the training is 8.5872 and 5.3379 for testing. Since the number of MSE produced during the testing using 30 number of hidden layer are the lowest among the others, this layer was found to be the best layer for optimum hidden neuron selection to be used inside the radial basis function architecture. In addition, the trending graph produced using the 30 hidden neuron did not consist of any over

fitting problems and the factorised value are still lower than the standard deviation calculated at the first stage of data analysis. Larger number of hidden neurons number is normally associated with an overfitting problem which happened due to excessive noise generated during generalization of model. Figure 10 shows the example of overfitting which is due to overloaded of information using 100 numbers of hidden neurons.

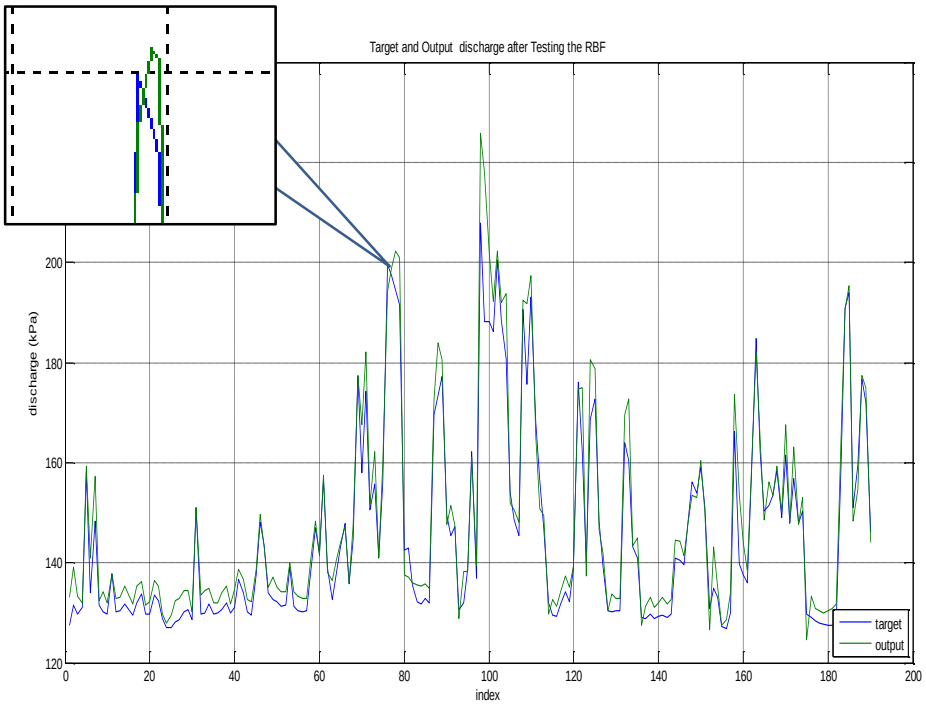


Figure 10: Overfitting problems due to excessive noise during generalization

The figure above displays the graphs of time series resulted from the insertion of 100 numbers of hidden neuron. The overlapped image is the enlargement of the area which was affected by overfitting problem and traced using the application of MATLAB software. In fact, the vertical axis represent the predicted output discharged which was patterned by the green colour graph whereas the horizontal axis is the targeted output discharged value, which represented by the blue graph. As shown in the figure, the RBF model designed failed to follow the sharp edges trend produce by the blue graph pattern. The model might perform well during the training but it has high tendency to perform poor during testing. Therefore, in between 4 to 31 number of neuron, 30 neuron shows the lowest MSE value for both training and testing phase. Hence the optimal number of hidden neuron was chose at 30.

3.4.5 Selection of Performances Evaluation Measures

The evaluation measure is a process in assessing or determining the performance of the calibrated data against one or more criteria (Maier et al., 2010). In actual, the performance of the model is assessed using a quantitative error metrics such as root means square error (RMSE), Sum of Squared Errors (SSE), Nash Sutcliffe efficiency (E) and Mean Absolute Error (MAE). However, in most of the time the difference between SSE and MAE are Squared Errors tend to be populated by the high magnitude error, therefore absolute error were used as an alternative based on the absolute difference between actual and modelled output (Maeir et al., 2010).

According to Mustafa et al. (2012) each of the evaluation criteria has its own function and formula as stated in Figure 11 where, $\hat{u}_k$ is the predicted value for discharged u_k , is the observed discharged value and $\bar{u}$ is the mean of the predicted target value and N is the total number of observation for the computed error.

$$\mathrm{RMSE} = \left[\frac{1}{N}\sum_{k=1}^{N}(\hat{u}_k - u_k)^2\right]^{1/2}$$

$$\mathrm{CE} = 1 - \frac{\sum_{s=1}^{N}(\hat{u}_s - u_s)^2}{\sum_{s=1}^{N}(\hat{u}_s - \bar{\hat{u}})^2}$$

$$R^2 = \frac{[\sum_{s=1}^{N}(u_r - \bar{u}_r)(z_r - \bar{z}_r)]^2}{\sum_{s=1}^{N}(u_r - \bar{u}_r)^2 \cdot \sum_{s=1}^{N}(z_r - \bar{z}_r)^2}$$

Figure 11: Examples of performance measurement

3.4.6 Selection of Output Layer

It is possible for the outcome layer to produce more than one output layer. However in this case study of forecasting of river flow using thin plate spline basis function, there was only one variable, which is to forecast the water level with respect to discharge value, therefore the output layers for the radial basis function using the thin plate spline was fixed at one. Besides, the selection of one output layer will enable for fast convergence of information through linear relationship from the hidden layer to the output neuron.

- Spread, σ = 1.6607
- Kernel Function = Thin Plate Spline Basis Function
- Input Variables = H_{n+1}
- Hidden Layer = 30
- Output Neuron = 1

As per summarized in the line above, the final result of the model architecture were construct based on the description listed above in order to get the full picture of the network :

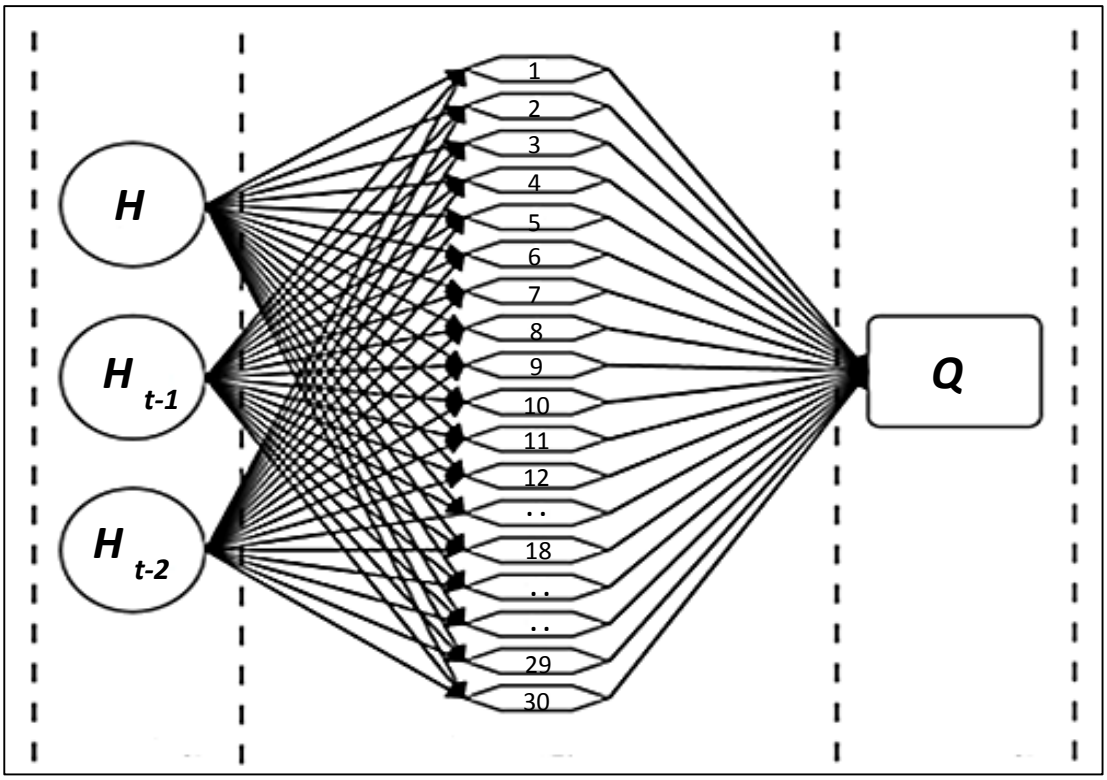


Figure 12: Final model of Thin Plate Spline Radial Basis Function

3.5 Project Activities Flow

Below are the process flow of the activities and stages taken throughout the FYP 1 and FYP 2 to complete the progress report.

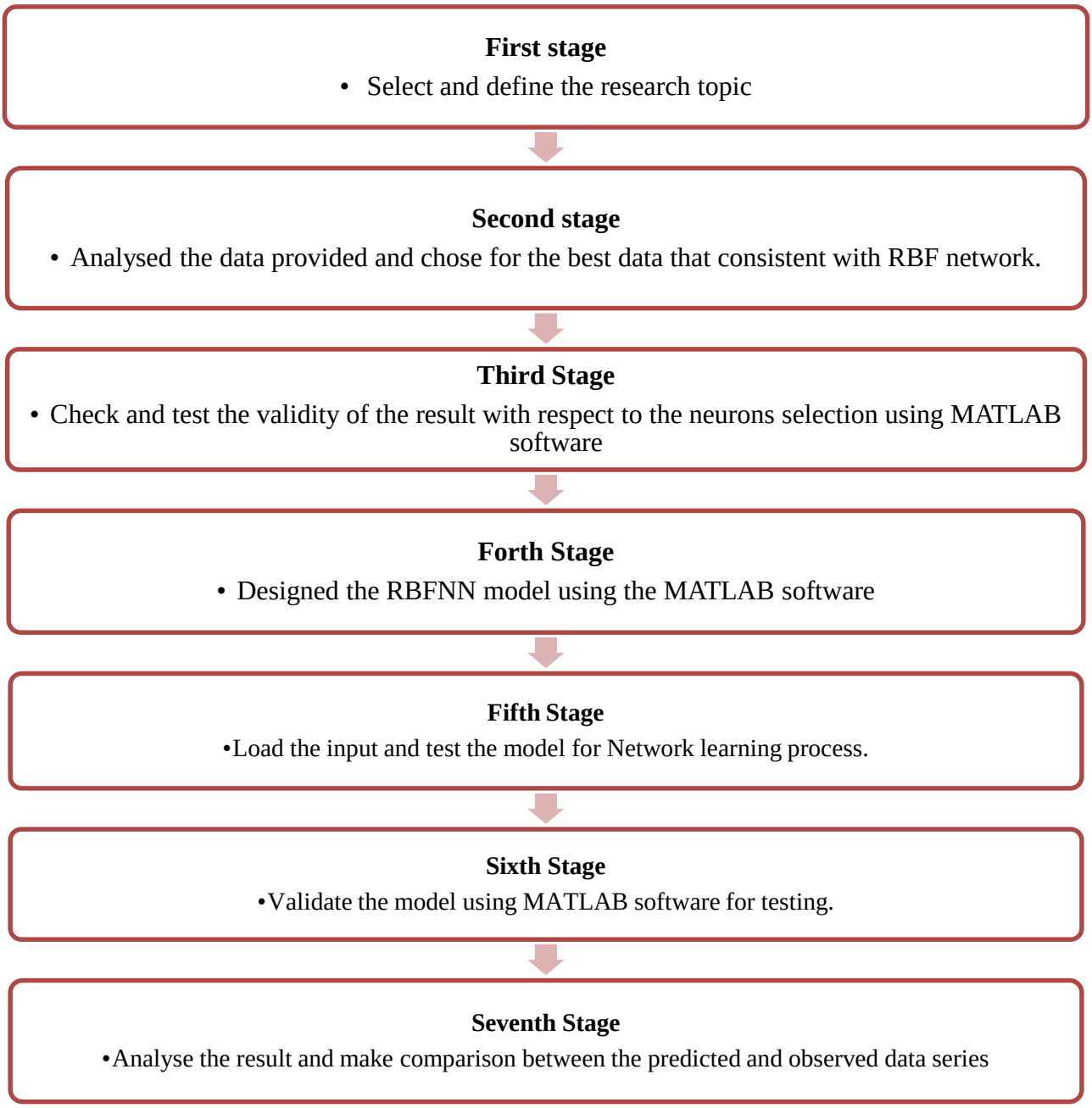


Figure 13: Process flow of development of model architecture of Radial Basis Function using Thin Plate Spline.

3.6 Project Key Milestone

In order to ensure for a proper progress flow and general overview on these Final Year Project, it is best to construct the key milestone which cover for both FYP 1 and FYP 2

Item No.	Task Detail	Start Date	Finish Date	Duration	January			February				March				April		
					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
Selection of Project Topic						○												
1.0	Surveying of FYP titles proposed by the supervisors	13/1/2014	17/1/2014	1 Week	→													
2.0	Confirmation of FYP title and supervisor	17/1/2014	22/1/2014	5 Days	→													
Preliminary Research Work									○									
3.0	Briefing by the supervisor	22/1/2014	24/1/2014	3 Days		→												
4.0	Understanding and familirazation of the background of FYP title	24/1/2014	2/2/2014	10 Days		→	→											
5.0	Collection of relevant journals, archives and articles	25/1/2014	11/2/2014	18 Days		→	→	→										
6.0	Reading and analysis of the previous studies from journals	30/1/2014	16/2/2014	18 Days			→	→		→								
Preparation of Extended Proposal										○								
7.0	Understanding and identification of problem statement	20/1/2014	31/1/2014	12 Days		→	→											
8.0	Extended Proposal drafting process	31/1/2014	16/2/2014	23 Days			→	→		→								
9.0	Consultation with supervisor	10/2/2014	16/2/2014	14 Days				→		→								
Submission of the extended proposal to the supervisor										○								
Preparation of Proposal Defence														○				
10.0	Amendment and correction of Extended Proposal	24/2/2014	3/3/2014	10 Days							→	→						
11.0	Additional readings of journals and articles	6/3/2014	17/3/2014	12 Days							→	→	→					
12.0	Preparation of presentation slides and practice	18/3/2014	21/3/2014	4 Days									→					
13.0	Consultation with supervisor for any ammendement	22/3/2014	24/3/2014	2 Days									→					
Proposal Defence															○			
Preparation of Interim Report																		
14.0	Meeting with the supervisor for the data collection	3/4/2014	5/4/2014	3 Days												→		
15.0	Performing data analysis with statistical measures	7/4/2014	11/4/2014	5 Days													→	
15.0	Consultation with the supervisor for any ammendement	11/4/2014	14/4/2014	3 Days													→	
Submission of interim report																		○

Table 3: Key Milestone for FYP 1

Item No.	Task Detail	Start Date	Finish Date	Duration	May-14			Jun-14				Jul-14				Aug-14		
					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
Parametric study							○											
1.0	Study on the parameters required in FYP 2 research	22/5/2014	24/5/2014	3 Days														
2.0	Revised the literatures review	27/5/2014	30/5/2014	4 Days														
3.0	Collect and specify the data for computation	3/6/2014	7/6/2014	5 Days														
Data simulation analysis										○								
3.0	Brief explanation by Supervisor	11/6/2014	12/62014	2 Days														
4.0	Thorough study on MATLAB software	6/6/2014	18/6/2014	10 Days														
Development of RBF using MATLAB										○								
3.0	Construct the code for Thin Plate Spline function	17/6/2014	20/6/2014	5 Days														
4.0	Run the program and counter problems raised	19/6/2014	26/6/2014	5 Days														
Trial and Errors											○							
3.0	Classified the trial and error using Excel	24/6/2014	27/6/2014	5 Days														
4.0	Analyse the graph produced for each trials	26/6/2014	3/7/2014	5 Days														
Preparation of Progress Report												○						
3.0	Construct and improve methodology section	4/7/2014	9/7/2014	3 Days														
4.0	Analyse and discuss the result	3/7/2014	10/7/2014	5 Days														
Submission of Progress Report												○						
Collecting material for poster exhibition																		
Pre SEDEX																○		
Preparation and submission of technical paper and draft report																		
Submission of interim report / Dessertation (soft bound) (FYP 1 and II)																	○	
Submission of interim report / Dessertation (hard bound) (FYP 1 and II)	FYP 2																	

Table 4: Key Milestone for FYP 2

3.7 Gantt Chart

No.	Detail / Week	Week No. / Date													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Selection of the project topic														
2	Preliminary Research Work / Project drafting														
3	Identify and understanding the problem statement														
4	Familiarize to the existing ANN technique and framework														
5	Submission of the extended proposal to supervisor														
6	Proposal Defence														
7	Developed the framework and model using MATLAB software														
8	Submission of interim Draft Report														
9	Submission of Interim Report														
10	Collect data	FYP 2													
11	Validate data														
12	Developed structured framework for troubleshooting														

Table 5: Gantt chart for FYP 1

No.	Detail / Week	Week No. / Date													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Continuation of Project Work														
2	Submission of Progress Report														
3	Continuation and Improvement on Project Work														
4	Pre - SEDEX														
5	Submission of Draft Report														
6	Submission of dissertation (Soft Bound)														
7	Submission of Technical Paper														
8	Oral Presentation														
9	Submission of Project Dissertation (Hard Bound)														

Mid Sem Break

Table 6: Gantt chart for FYP 2

3.8 Tools and Software

In order to conduct the study on this radial basis function neural network using the Thin Plate Spline, it is necessary to have a well understanding and practice on the related software that help to run the program. There are two main software that were used throughout the research which are MATLAB and Microsoft Excel. In fact, MATLAB (Refer to Appendix 5) is a very powerful programming tools that have broad 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 project this programming tools is used to help in determine and develop the flow network model. Apart from this two software, others related software used is Notepad and Microsoft Word.

CHAPTER 4

RESULTS AND DISCUSSION

4.2 Statistical Model Analysis using Thin Plate Spline Basis Function

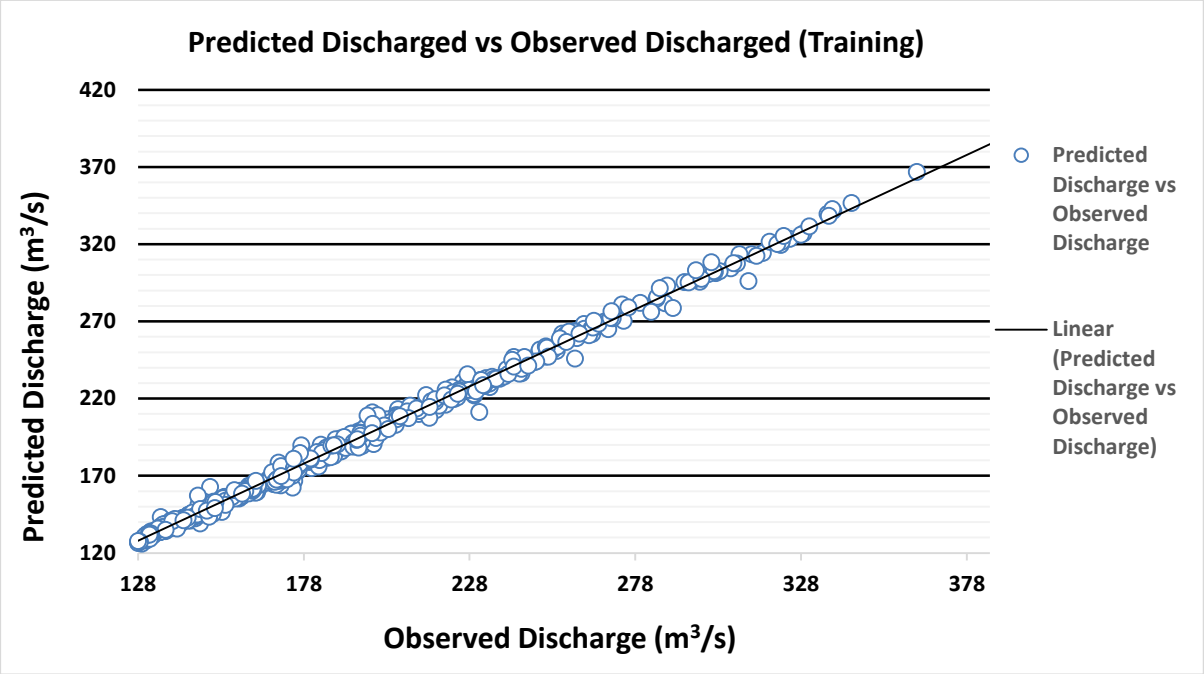


Figure 14: Graphs of predicted versus observed discharged value for training

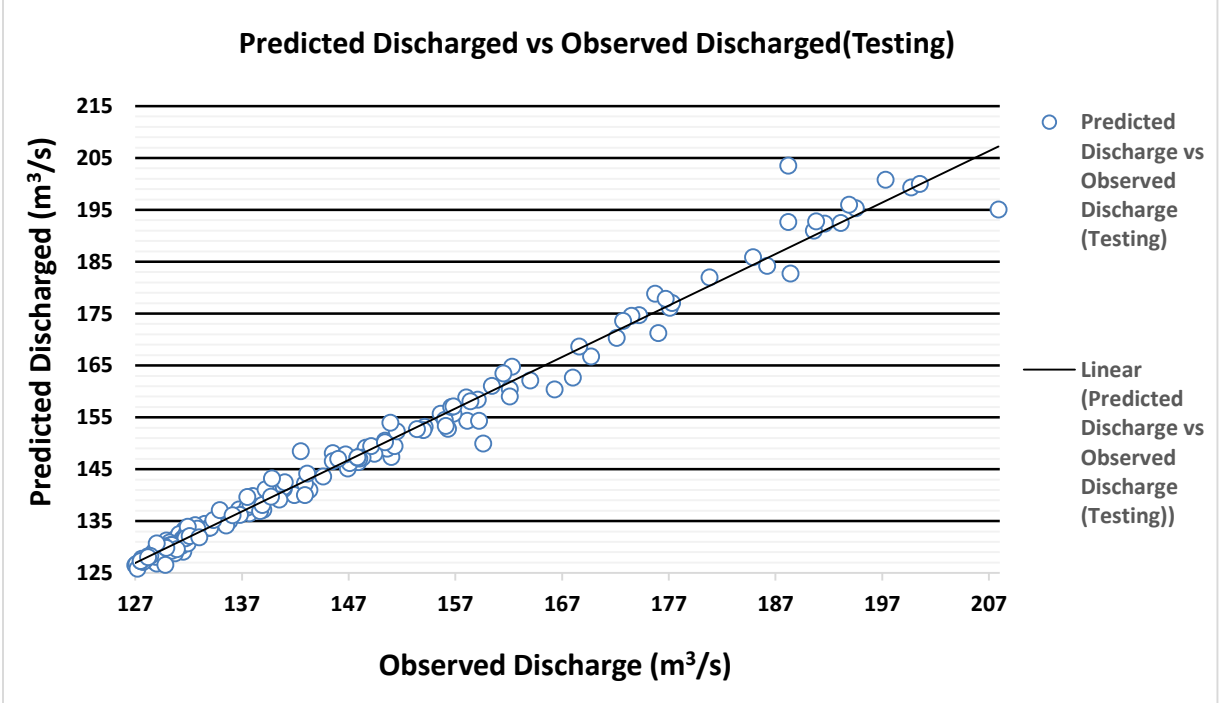


Figure 15: Graphs of predicted versus observed discharged value for testing

The graphs above show the comparison of the predictive and the observed data between the training and testing using the same basis function of thin plate spline (TPS). Through the use of an Excel program, the coefficient of determination, R^2 was determine by using specified formula as per stated in Figure 11. From the result, it was found that the value of coefficient determination, R^2 of the training data set is higher than the testing at 0.9966 compared to the testing, 0.9859. Thus it shows that during training, the model basis function performed with a higher precision to the targeted outcome value since there is less variation to the existing perfect line of agreement. This is mainly because of the system has gained an adequate learning process due to the high numbers of loaded input data and sufficient learning time.

The detail analysis on the figure found that, there are few points which stray far from the best fit line thus resulted in lower accuracy of predictive performance for the testing and training model. This is mainly due to the high marginal difference between the predicted and observed value at those particular points. In fact, from the figure it shows that both training and testing did produce slight unsatisfactory result for high discharged value. In figure 15, the particular point in testing recorded a value of 368.03 m^3/s and 394.7 m^3/s for predicted discharged, which is quite high compared to the other values in the data set. Therefore, the Thin Plate Spline (TPS) is found to encounter with a problem to learn with a large magnitude value and thus result in discrepancy of the data along the line of agreement. The same condition also happen to one particular point picked at the observed discharged value at 203.52 m^3/s and 188.2 m^3/s for testing data set. This situation might happened due to large marginal difference between the observed and predicted value and also inconsistency in maximum and minimum value in data set which attributed to low accuracy of the model predictive performance later.

However, with the slight difference between the coefficient determination, R^2 at 0.0107 (less than 0.1) between the training and testing model it can be concluded that the RBF model architecture using the thin plate spline algorithm has shown a good agreement with line of perfect agreement and able to forecast the data as close as possible to the observed data.

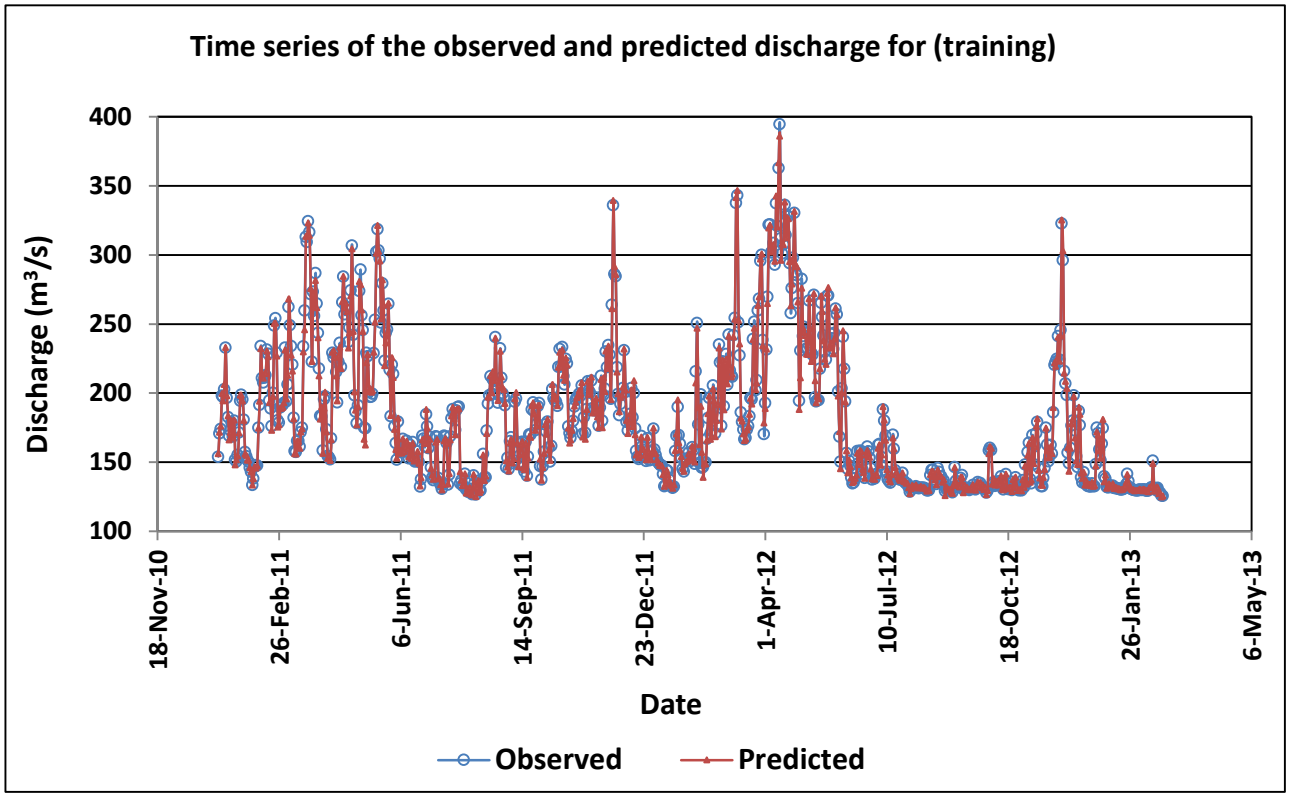


Figure 16: Time Series of Observed and Predicted Discharge for Training

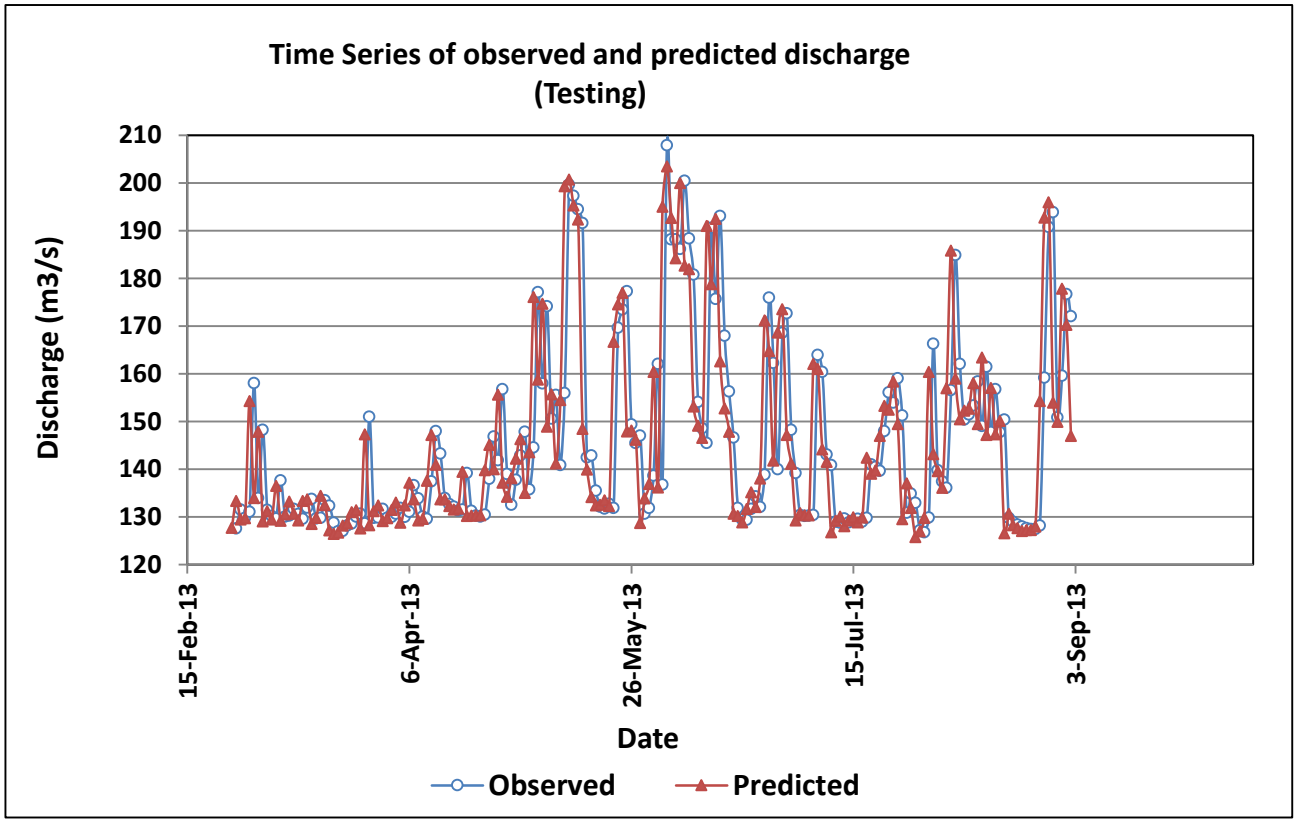


Figure 17: Time Series of Observed and Predicted Discharge for Testing

As illustrated in graphs above, the training data set graph shows a crowded and lengthy data set compared to the testing and this is due to the high amount of loaded input data which has been selected for the learning process at 970 data instead of 190 for testing. This is purposely been done in order to promote an adequate learning process for the algorithm before the testing could be executed.

During training, eventhough there are lots of data loaded in the network system, however, the trend shows a very systematic increment and decrement of linear line shape by closely follow the shape of the line in the observed discharged data. Hence, this suggest that the network system has learned the pattern of water level variation in response to discharged very well during the training process. In another part, the application of thin plate spline algorithm during testing did performed well which indeed showed a good correlation between the observed and predicted value pattern. Hence, it show that the network of Radial Basis Function using the Thin Plate Spline basis function could generalize at its optimum function when subjected to different regime and environment.

4.3 Statistical Performance Measure Analysis

Usually the analysis of the model performance is made to the basis of error measurement. Some of the most common statistical performance measures involved the Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) and Coefficient of Efficiency (CE). In fact, each of these parameters is a very powerful indicator towards the predictive 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 (Refer to Figure 11) before the result could be interpreted. The table below illustrated the simplified form of the result obtained for each parameters involved for both testing and training data set.

Data Set	RMSE	R^2
Training	2.927	0.997
Testing	2.310	0.986

Table 7: Statistical analysis of the model performance

As shown in the table above, it was found that the values of error in each parameter for both training and testing did not vary much from each other. In general, the model of RBF architecture designed shows a very close criteria to be a perfect model in predicting the discharge since overall error obtained from testing recorded for lower value in comparison to the training value. This might be due to well-trained learning process undergo during the training. However, a detail analysis regarding the result should be first determined before a conclusion could be made.

From the excel program, it was found that the value for MSE, RMSE and MAE for testing did produce a very well and satisfactory result in predicting the flow discharge of the Perak River. However for the simplification purpose, only RMSE and Coefficient of Determination, R^2 were chose to be presented in the result part. In fact, RMSE which stand for Root Mean Square Error is the Root factor to the actual MSE and this two type of parameters is sufficient to indicate and analyse the performance of the model developed.

From the Excel, the high value of MSE for the training data is much higher compared to the testing data set due to the size of the error which correlate the predicted with the observed discharged value in the system. Therefore as a result, the squared error basis such as RMSE shows a higher tendency of being dominated by the high magnitude error during the training process. From the table, the value of RMSE recorded for training is much higher at 2.927 compared to the testing at 2.310 Therefore, the training show that the cluster of input inserted into the system is far from the actual mean value obtained thus result of high error magnitude.

On the contrary, for the testing it showed that the model were easily predict the observed data set with a high predictive accuracy due to good correlation between the water level and discharge data used in testing thus promote for a lower magnitude of error value as compared to the training. As an alternative, the significant of low error measurement for MAE value in testing would proved that there is less absolute error of difference between the predictive and targeted output.

Based on the table above, it clearly shows that the coefficient of determination, R^2 for the training is much higher compared to testing. This is mainly due to the number of load input value which is higher compared to testing. In fact, as more and more input is loaded the higher the improvement and the performance of R^2 value due to adequate learning process which took place. In the other words the discrepancy is sourced from the variability of the inherent data in training the training and testing data set. In other parts, the slight underperformance of R^2 at 0.986 which is lower than training, 0.997 during testing was attributed by the uncertainty associated such as larger variability in water level data set or insufficient length of training to predict for high values during testing.

However, despite of slight difference in the error measurement and coefficient of determination, the results below have shown a good performance of comparison when compared to the previous kernel function such as Gaussian and it produce a very satisfactory result as well as Multi-Quadric Function (Refer to Table 8). This analysis using the performance measurement has led to the significance of applying the Thin Plate Spline Basis Function training algorithm in modelling the non-linear complex behaviour of River Flow at Sg.Perak.

Author	Kernel Function	(R^2)
Ruslan et.al (2013)	Gaussian	0.911
Jain and Chalisgaonkar (2013)	Gaussian	0.928
Yangus, J.S. (2014)	Multi-Quadric (MQ)	0.995
Fakharuden, A.R. (2014)	**Thin Plate Spline (TPS)**	**0.986**

Table 8: Comparison of RBF performance using different kernel function

CHAPTER 5

CONCLUSION AND RECOMMENDATION

All in all, throughout this study paper, it was found that the application of thin plate spline basis function tend to produce a very satisfactory result in predicting the discharge of the Perak River. The model of radial basis function architecture using the thin plate spline were developed by applying the trial and error approach and the final model architecture were found to perform best at three input neurons consist of the water level data as the input, 30 number of neuron in the hidden layer and spread of 1.6607 with one output neuron of discharge value. The result obtained from the two stages of training and testing showed a very impressive and significant accuracy of predictive performance for testing at 0.986 and 0.997 for training. In spite of minimal discrepancy in the marginal difference between the ranges of data set, the model tend to produce a very good correlation between the predicted and the observed discharge value. Therefore it can be concluded that the study of water flow prediction using the thin plate spline basis function has achieved the designed objective. This such of developing prediction technique using the thin plate spline basis function is recommended to be used in future to predict for the other hydrological data in the related hydrological field thus provide an accurate and reliable data sources for the application in the industry.

REFERENCES LIST

[1] Feng, L. H., & Lu, J. (2010). The practical research on flood forecasting based on Artificial neural networks. *Science Direct*, 37, 2974-2977.

[2] Jain, S. K., & Chalinsgaonkar, D. (2000). Setting Up Stage-Discharge Relations Using ANN. *Journal of Hydrologic Engineering*, 5(4), 428-433

[3] Supharatid, S. (2003). Application of neural network model in establishing a stage-discharge relationship for a tidal river. Hydrological Process, 17, 3085-3099.

[4] Moradkhani, H., Hsu. K. L., Gupta, H. V., & Sorooshian, S. (2003). Improved streamflow forecasting using self organizing radial basis function artificial neural network. *Journal of Hydrology*, 295, 246-262.

[5] Dawson, C. W., Harpham, C., Wilby, R. L., & Chen, Y. (2002). Evaluation of artificial neural network techniques for flow forecasting in the river yangtze, china. *Hydrology & Earth System Sciences*, *6*(4), 619-626.

[6] Bors, A. G. (n.d.). Introduction of radial basis function network. *Online Symposium for Electronis Engineers*,

[7] Shamseldin, A. Y. (2010). Artificial neural network model for river flow forecasting in developing country. *Journal of Hydrinformatics*,

[8] Orr, M. J. (1996). Introduction to radial basis function networks.

[9] Zounemat-Kermani, M., Kisi, O., & Rajaee, T. (2013). Performance of radial basis and LM-feed forward artificial neural networks for predicting daily watershed runoff. *Applied Soft Computing*, *13*(12), 4633-4644

[10] Ajmera, T. K., & Goyal, M. K. (2012). Development of stage–discharge rating curve using model tree and neural networks: An application to Peachtree Creek in Atlanta. *Expert Systems with Applications*, *39*(5), 5702-5710.

[11] Gadkar, K. G., Mehra, S., & Gomes, J. (2005). On-line adaptation of neural networks for bioprocess control. *Computers & chemical engineering*, *29*(5), 1047-1057.

[12] Fernando, D. A., & Shamseldin, A. Y. **(2009).** Investigation of internal functioning of the radial-basis-function neural network river flow forecasting models. *Journal of Hydrologic Engineering*, *14*(3), 286-292

[13] Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. **(2010).** Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, *25*(8), 891-909.

[14] Kagoda, P. A., Ndiritu, J., Ntuli, C., & Mwaka, B. (2010**).** Application of radial basis function neural networks to short-term streamflow forecasting. *Physics and Chemistry of the Earth, Parts A/B/C*, *35*(13), 571-581.

[15] Mustafa, M. R., Rezaur, R. B., Saiedi, S., & Isa, M. H. **(**2012). River suspended sediment prediction using various multilayer perceptron neural network training algorithms—a case study in Malaysia. *Water resources management*, *26*(7), 1879-1897.

[16] Kasiviswanathan, K. S., & Agarwal, A. **(2012).** Radial Basis Function Artificial Neural Network: Spread Selection. *International Journal of Advanced Computer Science*, *2*(11).

[17] Moharrampour, M., Eskandari, M. R., Rahimi, H., Ghafouri, S. R., & Abad, M. R. A. A. (2012). Predicted Daily Runoff Using Radial Basic Function Neural Network RBF. *Advances in Environmental Biology*, *6*(2), 722-725.

[18] Yin, J. C., Zou, Z. J., & Xu, F. (2013). Sequential learning radial basis function network for real-time tidal level predictions. *Ocean Engineering*, *57*, 49-55.

[19] Ruslan, F. A., Samad, A. M., Zain, Z. M., & Adnan, R. (2013, November). Modelling flood prediction using Radial Basis Function Neural Network (RBFNN) and inverse model: A comparative study. In *Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on* (pp. 577-581). IEEE.

[20] Mustafa, M. R., Isa, M. H., & Rezaur, R. B. (2012). Artificial Neural Networks Modeling in Water Resources Engineering: Infrastructure and Applications.*World Academy of Science, Engineering and Technology*, *62*, 341-349.

[21] Isa, M. M. M. (2014). Comparative Study of MLP and RBF Neural Networks for Estimation of Suspended Sediments in Pari River, Perak.

APPENDICES

Appendix 1

Perak River

Perak River Current Flow

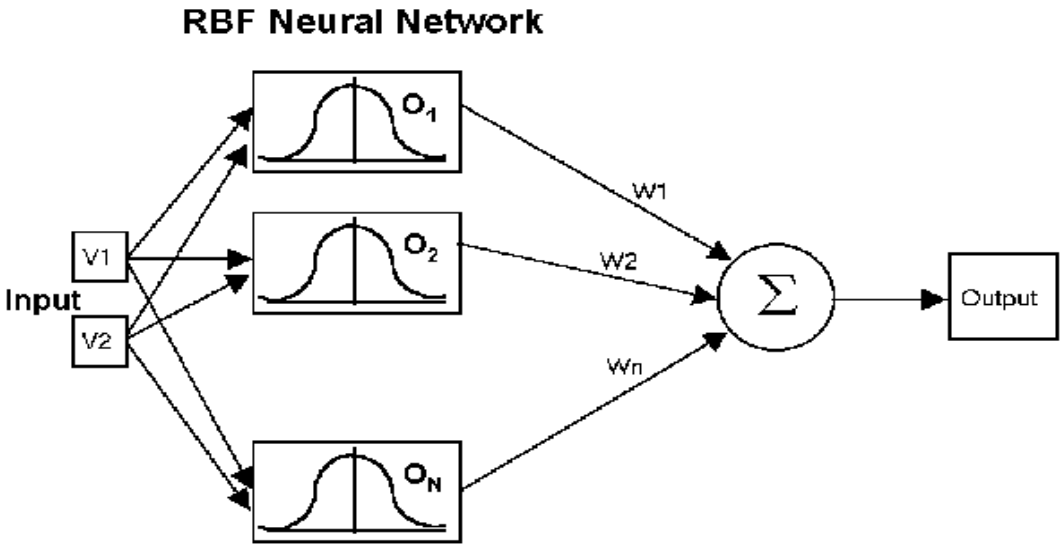
Appendix 2

TABLE 1. Sum of Squares of Errors and Coefficient of Correlation for ANN Models and Conventional Procedure—Training and Testing Data of Satrana Site

ANN model inputs (1)	Nodes in hidden layer (2)	Training Data		Testing Data	
		Sum of squares of errors (3)	Coefficient of correlation (4)	Sum of squares of errors (5)	Coefficient of correlation (6)
H_t	5	557,881	0.991	18,535	0.989
H_t and H_{t-1}	8	473,547	0.992	23,292	0.989
H_t and Q_{t-1}	5	547,474	0.991	22,728	0.988
H_t, H_{t-1}, and Q_{t-1}	6	373,559	0.994	13,574	0.990
H_t, H_{t-1}, and H_{t-2}	6	443,272	0.993	22,119	0.989
H_t, H_{t-1}, H_{t-2}, and Q_{t-1}	8	312,234	0.995	19,812	0.985
H_t, H_{t-1}, H_{t-2}, Q_{t-1}, and Q_{t-2}	10	266,000	0.996	18,235	0.988
Curve fitting	—	949,349	0.955	123,814	0.978

Table 1: Comparison between ANN and Conventional Approach for Sum of Squares of Error and Coefficient of Correlation

Appendix 3



Configuration of Radial Basis Function's Neural Network

Appendix 4

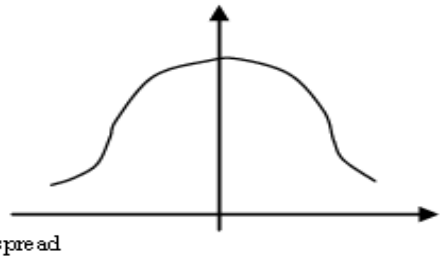


Fig. 2 Larger spread

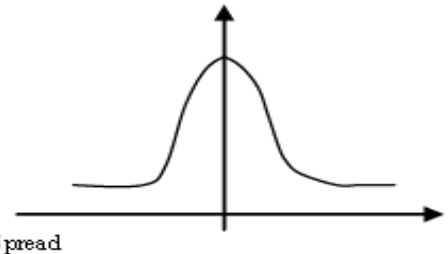


Fig. 3 Small Spread

Appendix 5

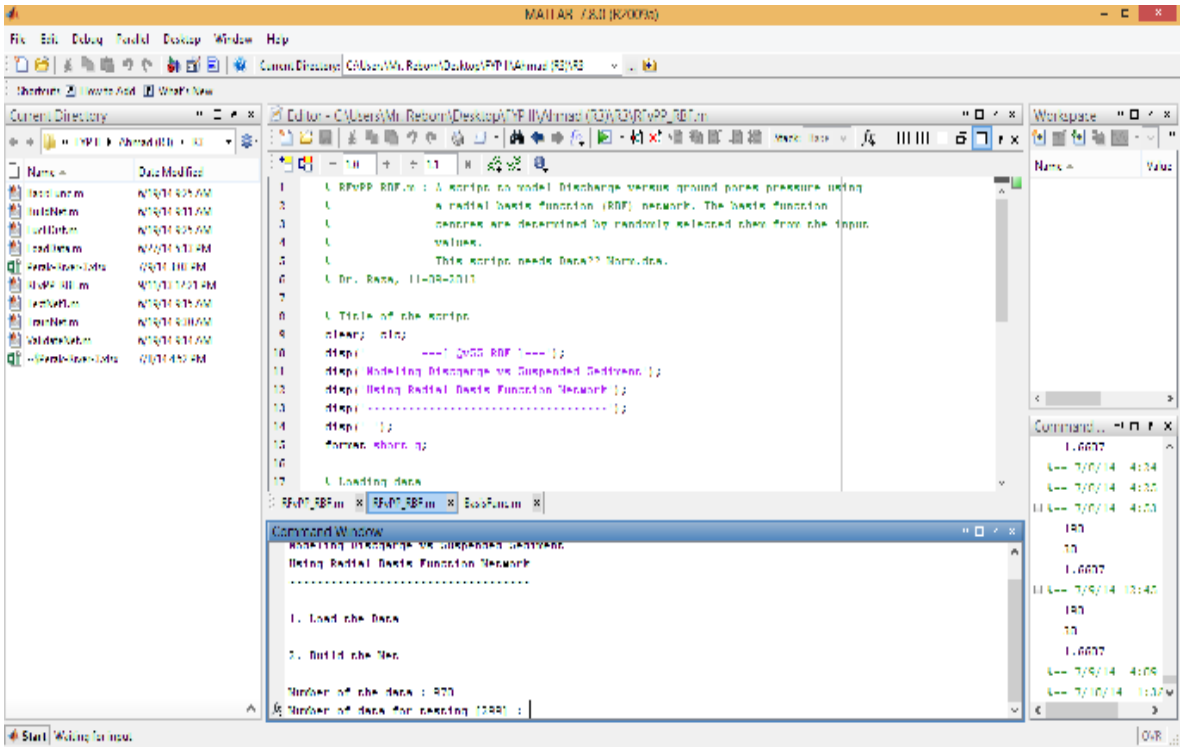


Figure 15: MATLAB Soft Computing Tools