CHAPTER III

METHODOLOGY

Applications of the artificial neural networks (ANNs) in hydrology and water resources planning emerged in the early 1990's. In this thesis, ANN is applied with an attempt to develop nonlinear rainfall-runoff relations for the river discharge forecast in the designated watershed area. Implementation of ANN is procedurally based on the following five components: (1) Data adjustments, (2) Input determination, (3) Data preprocessing, (4) Network construction and training, and (5) Statistical criteria for optimal network selection. Figure 3.1 depicts the diagram of ANN implementation in sequential order.

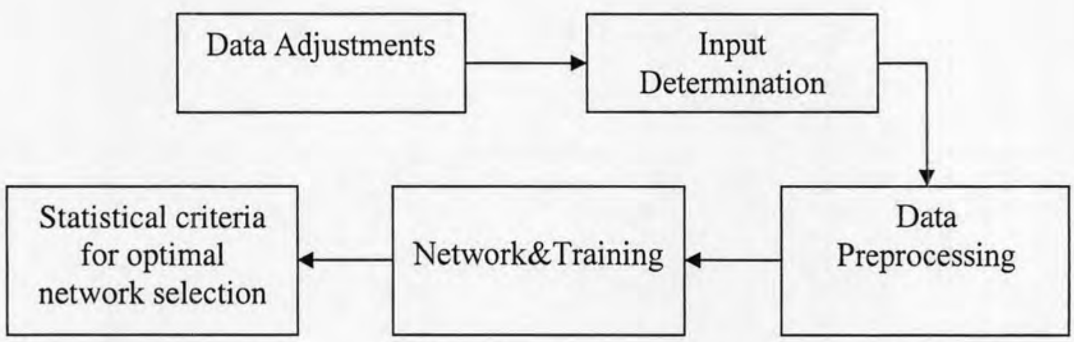

Figure 3.1: Five components of ANN implementation.

Due to missing records in the original rainfall and runoff data, the collected information must be adjusted to properly fit the appropriate choice of network input to the network, Once the input-target pairs are selected from the previous step, the data shall be preprocessed prior to ANN's training procedure. These are described in the data preprocessing, and network construction and training sections in this chapter. Eventually, the trained network is selected by statistical criteria to achieve the optimal network.

3.1 Data Adjustments

The original format of rainfall and runoff data used for ANN development is obtained from the Royal Irrigation Department of Thailand (see Tables 3.1 and 3.3). These data were recorded, in the spreadsheet format, between April 1996 - December 2005. This is not suitable for statistical analysis and numerical computation. It is necessary to convert these original data to a more proper format (excel format) for the ANN implementation. These adjustments are described in details as follows.

3.1.1 Rainfall Data

Hourly rainfall data was measured in *mm/hr*. A sample of the original rainfall data in September, 1998 is shown in Table 3.1. From this rainfall data sheet, there are many time slots of which data are missing. These missing records were due to failure of the automatic rain gauge and probably human factors. To adjust the rainfall data, the values at the time slot with no recorded data are replaced by 0 and those at the period in which the automatic rain gauge failed are replaced by -1 (see Table 3.2).

Table 3.1: Original rainfall data in September of the year 1998

ปริมาณน้ำฝนรายชั่วโมง - มิลลิเมตร

สถานี P.64 อ. อมก๋อย จ. เชียงใหม่ ปีน้ำ 2541

กันยายน

	เวลา - ชั่วโมง																								ฝนอัตโนมัติ	ฝน
วันที่	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	24:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	รวม	ธรรมดา
1	เครื่องขัดข้อง																								-	6.2
2										5.0	10.2	0.1								0.2	2.0	2.6			20.1	20.7
3							0.5	8.3	5.4	0.3	0.1														14.6	14.9
4				0.1				0.2																	0.3	0.3
5																	0.5	0.3			0.2				1.0	1.0
6				0.6						0.9			0.1		0.7	0.6	2.6	3.0	1.6	1.6	1.0	0.5		0.1	13.3	13.6
7			3.3	1.3	0.1	7.9	0.1						1.9												14.6	15.1
8													0.1	0.1											0.2	0.3
9										0.1										0.2	0.1			0.1	0.5	0.5
10						0.1	1.5	10.7	0.8	12.5															25.6	26.7
11																										
12																										
13																										

14																										
15																										
16													0.4												0.4	0.4
17		0.8	0.8	0.2																				0.1	1.9	1.8
18	0.2	2.9	2.6	0.1																					5.8	6.6
19																		4.0	1.7	0.4	0.1				6.2	6.1
20									5.2																5.2	5.2
21																										
22																										
23																										
24	กราฟไม่ชัดเจน																								-	0.6
25							5.3																		5.3	5.3
26																										
27																										
28																										
29																										
30																										
31																										
																								รวม	115.0	119.1

Table 3.2: Adjusted rainfall data in September of the year 1998

ปริมาณน้ำฝนรายชั่วโมง - มิลลิเมตร

สถานี P.64 อ. อมก๋อย จ. เชียงใหม่ ปีน้ำ 2541

กันยายน

	เวลา - ชั่วโมง																								ฝนอัตโนมัติ	ฝน
วันที่	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	24:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	รวม	ธรรมดา
1	-1.0	-1.0	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	6.2
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	10.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	2.0	2.6	0.0	0.0	20.1	20.7
3	0.0	0.0	0.0	0.0	0.0	0.0	0.5	8.3	5.4	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.6	14.9
4	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.3	0.0	0.0	0.2	0.0	0.0	0.0	1.0	1.0
6	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.1	0.0	0.7	0.6	2.6	3.0	1.6	1.6	1.0	0.5	0.0	0.1	13.3	13.6
7	0.0	0.0	3.3	1.3	0.1	7.9	0.1	0.0	0.0	0.0	0.0	0.0	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.6	15.1
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.1	0.5	0.5
10	0.0	0.0	0.0	0.0	0.0	0.1	1.5	10.7	0.8	12.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.6	26.7
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		

14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4
17	0.0	0.8	0.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.9	1.8
18	0.2	2.9	2.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.8	6.6
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	1.7	0.4	0.1	0.0	0.0	0.0	6.2	6.1
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.2	5.2
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
24	-1.0	-1.0	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.6
25	0.0	0.0	0.0	0.0	0.0	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.3	5.3
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
29	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
31																										
																								รวม	115.0	119.1

3.1.2 Runoff Data

The three hourly runoff data was recorded as water level in meters (m). A sample of the original runoff data in April, 2001 is shown in Table 3.3. In this thesis, the ANN model is developed to predict the three hourly "discharge". Hence, water level in meters (m) is transformed into discharge with the unit of CMS or m^3/s. This transformation is based on the relationship of the discharge-water level curve given in Figure 3.2. Similar to rainfall data, the unavailable discharge value at any period is set to -1.

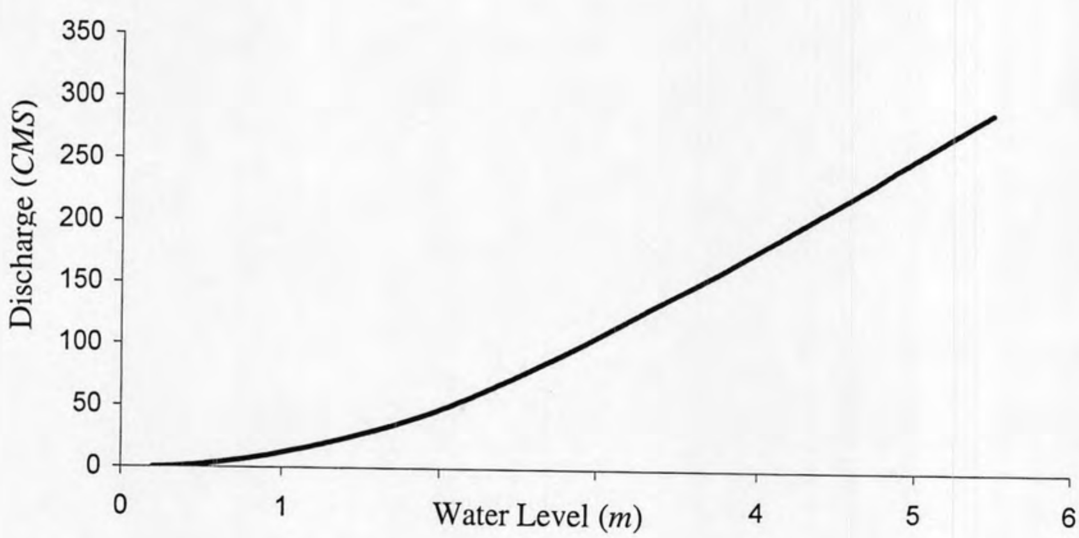

Figure 3.2: Relationship between discharge (CMS) and water level (m). Source: the Royal Irrigation Department of Thailand.

Table 3.3: Original runoff data in April of the year 2001

RID Data Processing Division

PC Version H.1 (P.64)

Station : Ban Luang, Om Koi, Chiang Mai,(P.64)
Stream : NamMae Tuen
River : Ping
River System : Ping River

RoyalIrrigation Department
Thailand
Hydrology Division
Type of Gage: Staff Gage

Hourly Gauge Height
April 2001 (Water Year)

Time and Gage Height in Meters (A.D.)

Date	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	24:00
1						0.35			0.35			0.35			0.35			0.35						
2						0.35			0.35			0.35			0.35			0.35						
3						0.39			0.38			0.38			0.38			0.37						
4						0.37			0.37			0.37			0.37			0.36						
5						0.36			0.36			0.35			0.35			0.34						
6						0.34			0.34			0.34			0.33			0.33						
7						0.33			0.33			0.33			0.33			0.32						
8						0.33			0.33			0.33			0.32			0.32						
9						0.33			0.33			0.33			0.33			0.33						
10						0.33			0.33			0.33			0.33			0.33						
11						0.32			0.32			0.32			0.32			0.32						
12						0.32			0.32			0.32			0.32			0.32						
13						0.32			0.32			0.32			0.32			0.32						
14						0.32			0.32			0.32			0.32			0.32						
15						0.33			0.33			0.33			0.33			0.3						
16						0.32			0.32			0.32			0.31			0.31						

	Time and Gage Height in Meters (A.D.)																							
Date	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	24:00
17						0.3			0.3			0.3			0.3			0.3						
18						0.3			0.3			0.3			0.3			0.3						
19						0.3			0.3			0.3			0.3			0.3						
20						0.3			0.3			0.3			0.3			0.3						
21						0.31			0.31			0.31			0.3			0.3						
22						0.3			0.3			0.3			0.3			0.3						
23						0.3			0.3			0.3			0.3			0.3						
24						0.3			0.33			0.28			0.3			0.31						
25						0.29			0.29	0.3		0.29			0.29			0.28						
26						0.28			0.29			0.3			0.3			0.3						
27						0.28			0.28			0.3			0.3			0.3						
28						0.29			0.29			0.29			0.29			0.29						
29						0.28		0.27	0.27			0.27			0.28			0.28						
30						0.27		0.27	0.27			0.28			0.28			0.28						

3.2 Input Determination

Determination of appropriate model inputs is very important in many model designs (Faraway and Chatfield, 1998) since they can directly or indirectly affect the output accuracies. There is no exception for the ANNs development process. Nonetheless, it should be noted that not all of the determined input variables can affect the output. Some input may be correlated, noisy or insignificantly related to the model output. These can cause difficulties in the learning process, particularly when a large number of inputs are given to the ANNs. There have been reports of shortcomings in these cases on (Maier and Dandy, 1997; Back and Trappenberg, 1999; Gavin *et al.*, 2005); (1) input dimensionality, computational complexity and memory requirements of the model also increase, (2) learning process becomes more difficult to achieve with irrelevant inputs, (3) poor model accuracy may originate from the inclusion of irrelevant inputs due to the increase of number of local minima in the error surface, and (4) some types of ANNs with many irrelevant inputs behave poorly since the network uses almost all its resources to represent irrelevant portions of the input-output mapping. On the other hand, the networks with irrelevant inputs that can efficiently provide reliable results generally require more data in order to concentrate on important regions of the input space. Therefore, it is necessary to adopt analytic procedures for appropriate ANN model input selection.

There are generally many parameters involved in the processes of finding the rainfall-runoff relationship. These parameters are concerned with physical, geographical and meteorological constraints. Essential watershed characteristics may have different magnitude of impact on the rainfall-runoff processes. A few of them can be listed as follows: Topography of the watershed; Geography relating the intensity and duration of the rainfall event, rainfall season and drought period etc.; geology influencing the streamflows and groundwater flow; and sociology resulting in variation of the water distribution. Occasionally, models that utilized too many variables are not very well accepted by modelers (Furundzic, 1998). Mainly, because it is impossible to use all parameters which affect the process, in order to develop a model, particularly, in the rural area with incomplete hydrologic measurements. The recorded data, practically used as input to ANN, is the past runoff and rainfall data. Both data are considered as representations of soil moisture state and the water storage capability of the watershed, respectively (Rujurkar *et at.*, 2004). Other data such as temperature (Furundzic, 1998) and evaporation (Nazemi *et at.*, 2003) can also be included, in addition to both rainfall and runoff data (Furundzic, 1998, Singh *et at.*, 2006). Once the input selection is made, the appropriate lag time of network input can then be determined. This is important for multilayer perceptrons structure, which is regarded as a static system. The lag time of each input is incorporated as the dynamic information for the network (Maier and Dandy, 2000).

In this study, the inputs to ANN are determined from the available hydrologic data: hourly rainfall data and three hourly discharge data. The appropriate lag time for rainfall and runoff data are considered separately as follows.

3.2.1 Lag time determination for rainfall data

For hourly rainfall data, the lag time can be derived from the analysis of correlation. Let $E(t_i)$ be a value of the discharge difference at predicted time t_i, defined by

$$E(t_i) = D(t_i) - D(t_i - t_D). \tag{3.1}$$

Here, $D(t_i)$ is the discharge at time t_i and t_D is the number of time step backward for the discharge. In this study, $t_D = 3$. Rainfall lags are determined from the correlation between $E(t_i)$ and the corresponding rainfall values $R_k(t_i)$ at lag time k, where $k = t_i, t_i - 1, t_i - 2, \cdots, t_i - 7$. The correlation C_k at lag time k can be computed by

$$C_k = \frac{\sum_{i=1}^{N}(E(t_i) - \overline{E})(R_k(t_i) - \overline{R_k})}{\sqrt{\sum_{i=1}^{N}(E(t_i) - \overline{E})^2 \sum_{t=1}^{N}(R_k(t_i) - \overline{R_k})^2}}. \tag{3.2}$$

Here N is the total number of time series. The computed correlations are given in Table 3.4.

Time	t	t-1	t-2	t-3	t-4	t-5	t-6	t-7
C_k	0.01	0.10	0.13	0.12	**0.23**	**0.34**	**0.24**	0.13

Table 3.4: Correlations between the discharge difference (E) and the corresponding rainfall values at lag time k (R_k).

It is evident from the computed correlations that the rainfall data of lag times $t-4, t-5$ and $t-6$ are suitable for the input vector to the ANN.

3.2.2 Lag time determination for discharge data

Unlike rainfall data, the three hourly runoff values were recorded. Since the discharges were recorded once every three hours only during the daytime from 6 a.m. to 6 p.m.. The discharges $D(t_i - t_D)$, $D(t_i - 2t_D)$ and $D(t_i - 3t_D)$ are used as another input to the ANN, where t_D is a three-hour period, for the target discharge $D(t_i)$. Therefore, in each day, the values at 6 a.m., 9 a.m. and 12 a.m. and the values at 9 a.m., 12 a.m. and 3 p.m. are used as previous discharge input to the ANN for the target discharge at discharge value at 3 p.m. and 6 p.m., respectively. Main reason for choosing at least three lag times of discharge is that it can represent the history of the changes in discharge at that location. Due to the limitation on three hourly discharges information, the choices of numbers of used previous discharge are 3 or 4. Based on the number of input pattern, three previous discharges are selected because the number of input patterns is more than the number of input patterns obtained from selecting four previous discharges.

From above, the input-target pair at the predicted time t_i is characterized by $\{\vec{I}(t_i), D(t_i)\}$. Here, the rainfall data at $t_i - 4, t_i - 5$ and $t_i - 6$, and the discharges at $t_i - t_D, t_i - 2t_D$ and $t_i - 3t_D$ constitute an input vector $\vec{I}(t_i)$ corresponding to the discharge $D(t_i)$. Therefore, the input vector $\vec{I}(t_i)$ can be written as

$$\vec{I}(t_i) = [R(t_i - 6)\ R(t_i - 5)\ R(t_i - 4)\ D(t_i - 3t_D)\ D(t_i - 2t_D)\ D(t_i - t_D)]^T, \qquad (3.3)$$

where D denotes the discharge, R denotes rainfall, t_i is the predicted time and, $t_D = 3$ representing the time step for runoff lags. The ANN model can be mathematically described by

$$\tilde{D}(t_i) = f(\vec{I}(t_i)), \qquad (3.4)$$

where $\tilde{D}(t_i)$ is the predicted discharge from the model. In the network development processes, one set of the input-target pairs $\{\vec{I}(t_i), D(t_i)\}$ is used for the network training to produce the output $\tilde{D}(t_i)$ provided that $\| \tilde{D}(t_i) - D(t_i) \| < \varepsilon$, for $\varepsilon > 0$. This is called the training set. Another set of input-target pairs is used for the model testing, which is called the testing set. In this studied area, the total numbers of distinctive input-target pairs derived from the original records of rainfall and discharge are 2,840 pairs or patterns.

3.3 Data Preprocessing

In this section, the data preprocessing for the ANN's training is described. The data herein is the set of input-target pairs derived from previous section. There are two consecutive steps for this process: First step concerns the filtration and data separation of total input-target pairs (the grouping of filtered input-target pairs as training and testing sets). Second step involves the transformation of data. Details of each step are presented in the following subsections.

3.3.1 Data Filtration and Data Separation

Available data are randomly separated into two subsets. One subset is used for calibrating the parameters of the model (training process) and the other is used for validating the model (testing process). Before separation, the derived input–target of 2840 pairs is examined for filtering. It is found that some input vectors are "meaningless" due to insufficient information. This is resulted from the only available rain gauge station in the studied area, which insufficiently represents the distribution of the rainfall event in the whole watershed. These "meaningless" input-target pairs appear to have strange information as follows: (i) there is no or few rainfall quantity but discharge significantly increases, and (ii) discharge decreases even with substantial rainfall intensity. Therefore, it is necessary to filter primitively the set of input-target pairs in order to remove these "meaningless" input-target pairs prior to training and testing the model. The conditions for removing these input-target pairs are described as follows.

Define $SR(t_i) = R(t_i - 6) + R(t_i - 5) + R(t_i - 4)$ and $DD(t_i) = D(t_i) - D(t_i - t_D)$. A given input-target pair is said to be "meaningless" and shall be removed if any one of the following two conditions: 1. $SR(t_i) < 1$ and $DD(t_i) \geq 1$ or 2. $SR(t_i) > 1$ and $DD(t_i) \leq 1$ holds. After applying the data filtering criteria to all input-target pairs, there remain input patterns of 2,403 or approximately 85% of the original data. From these remaining data, the input-target patterns are divided into (1) training set consisting of 1,443 input-target pairs and (2) testing set consisting of 960 input-target pairs, which are approximately 60% and 40% of the total input-target pairs, respectively.

3.3.2 Data Transformation

Data Transformation is another important step that can help improving the learning processes. This is carried out after the data is separated into their respective subsets (training and testing sets). General objective of this transformation is to ensure that all elements of each input vector and the corresponding target receive equal attention during the training process. This has to be done in such a way that the transformed data is consistent with the range of the nonlinear activation function in the output layer. For example, if the sigmoid function is used as the activation function, the transformed data must be between 0 and 1, such as [0.1, 0.9] or [0.2, 0.8]. In this thesis, the general linear transformation to [a, b] (Sajikumar and Thandaveswara, 1999) is used to scale down the input-target pairs. This linear transformation is described by

$$x_n = a + \frac{(x_o - x_{\min})}{x_{\max} - x_{\min}}(b - a). \tag{3.5}$$

Here, x_n and x_0 represent the transformed and the original data, respectively, whereas $x_{\min}$ and $x_{\max}$ are the minimum and the maximum values of the original data. The transformation is performed on both the input vectors and the targets. The advantage of using linear transformation (3.5) is quite obvious since it is simple and the sigmoid function is used as the activation function of neurons in this network.

3.4 Network and Training

In the training process, network geometry (i.e. the types of connection: full or partial connection, the number of hidden layers and the number of neurons or nodes in each layer) must be specified in advance. Learning process and ANN performance can be affected by the network geometry. First, we define the descriptions of network capability to learn from a training set using the concepts of generalization of network and overtrained network. Generalization of network is the network that can perform well even when the unseen data is fed into the network. Overtrained network is the network that overfits the training set. In other words, it has learnt the idiosyncrasies in the training set and unable to generalize. Clearly, the size of network is proportional to the number of neurons in the network. Smaller networks usually have better generalization ability, require fewer physical resources (e.g. less storage space), and with better processing speed during the training. However, the error surface of smaller networks is more complicated, and might fail to learn the complicated problems. On the other hand, the larger networks tend to learn quickly (in terms of number of training cycles) and more suitable for complicated problem provided that the networks are large enough. Unfortunately, their computational costs are high due to the size of network and the networks generally require a large number of training set to achieve good generalization ability (Maier, H.R., and Dandy, G.C., 2000). This section composes of two subsections namely (1) network construction and (2) network training, which are carried out by utilizing the MATLAB 7.0 Neural Network Toolbox.

3.4.1 Network construction

For a given geometric specification, a fully connected neural network (see Figure 2.6) is chosen. Number of nodes (or neurons) in the input and output layers are determined by the number of elements of input vector and the number of corresponding targets, respectively. There is no theory concerning the determination of numbers of hidden layers and hidden neurons. In fact these are problem-dependent. For this work, two hidden layers are used and the number of neurons in each hidden layer is varied from 12 to 30 neurons. The feedforward backpropagation network is created by using the function *newff* in the MATLAB Neural Network Toolbox, by

$$net = newff(PR,[N_{H_1}, N_{H_2},\ldots,N_{H_k}, N_O],\{f_{H_1},f_{H_2},\ldots,f_{H_k},f_O\},BTF,PF), \qquad (3.6)$$

where net is the created feedforward backpropagation network,

PR is the $R \times 2$ matrix of min and max values for R input elements of an input vector,

N_{H_i} is the number of neurons in the i^{th} hidden layer (H_i),

f_{H_i} is the activation function in the i^{th} hidden layer (H_i),

k is the number of hidden layers,

N_O is the number of neurons in the output layer,

f_O is activation function in the i^{th} hidden layer,

BTF is backpropagation network training function,

and PF is performance function (mean square error, *MSE*).

When we apply the function *newff*, the initial weights and biases of the network are automatically determined.

3.4.2 Network training

This subsection describes the training process of the created network. Here the batch mode for training is set in a way that the weight adjustment is performed after receiving all of input-target pairs. This mode is characterized by *train* function in Neural Network Toolbox. The variable *net* is referred as the network created by *newff* . The process for training the network *net* is described as follows.

Step 1. Set initial parameters related to the created network *net* prior to training namely:

net.trainParam.epochs is the maximum values of training cycles,

and *net.trainParam.goal* is the goal of minimizing mean square error (*MSE*).

Step 2. Train the network *net* by using the function *train.* The syntax of this function is given by

$$[net, tr, OutputNet, neterror] = train(net, ptr, ttr). \tag{3.7}$$

Here *tr* is the training record,

OutputNet is the network outputs,

neterror is the network error,

ptr is the network inputs,

and *ttr* is the network targets.

The network *net* in the right hand side of (3.7) is the network before training and one in the left hand side is the network after training.

3.5 Statistical Criteria for Optimal Network Selection

The desired optimal network structures in this study are evaluated by using various standard statistical performance evaluation measures. The selection procedure is based on the following statistics: Average absolute relative error ($AARE$); Correlation (C); Coefficient of determination (R^2); Normalized root mean square error ($NRMSE$) and; Relative error of maximum discharge ($REMD$). These are defined by

$$AARE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Q_i - \hat{Q}_i}{Q_i}\right| \times 100, \tag{3.8}$$

$$C = \frac{\sum_{i=1}^{N}(Q_i - \bar{Q})(\hat{Q}_i - \bar{\hat{Q}})}{\sqrt{\sum_{i=1}^{N}(Q_i - \bar{Q})^2 \sum_{i=1}^{N}(\hat{Q}_i - \bar{\hat{Q}})^2}}, \tag{3.9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2}, \tag{3.10}$$

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - \hat{Q}_i)^2}}{\frac{1}{N}\sum_{i=1}^{N}Q_i}, \tag{3.11}$$

and

$$REMD = \frac{Q_{\max} - \hat{Q}_{\max}}{Q_{\max}} \times 100. \tag{3.12}$$

Here, N is the number of data, Q_i is the i^{th} observed value, Q_i is the i^{th} predicted value, $\bar{Q}$ and $\bar{\hat{Q}}$ are the average of the observed and predicted values respectively. $AARE$ represents an average quantity of the relative error of the prediction with respect to the actual value of the discharge. It is clear from its definition that the smaller $AARE$ value indicates the better model performance. The correlation C is a commonly used statistics, which provides information on the strength of linear relationship between the observed and the predicted values. Coefficient of determination R^2 proposed by Nash and Sutcliffe (1970) is one of the widely used statistics for the evaluation of model performance. The closer C and R^2 are to 1.0, the better the model performance is. $NRMSE$ indicates the model's ability to predict a value away from the mean. It should be noted that the square term in the $NRMSE$ places a bias on the high value of discharge. Therefore, the errors in estimating the low discharge are insufficiently represented by $NRMSE$. Relative error of maximum discharge ($REMD$) quantifies the error in predicting the maximum discharge that are useful in water management (Srinicasulu and Jain, 2006).

Since the predicted values obtained from the network lies in the range [a, b], it is necessary to rescale the predicted values before making network comparison. This can be achieved by means of the mentioned statistic criteria to obtain the optimal network. The rescaled values are evaluated from the inverse function of linear transformation (3.4) as follows:

$$x_r = x_{\min} + \frac{(x_p - a)}{(b-a)}(x_{\max} - x_{\min}). \tag{3.13}$$

Here, x_r and x_p are the rescaled and the predicted values, respectively. The experimental results of ANN implementation are presented in the chapter VI.