Rainfall-Runoff Modelling Using Artificial Neural Networks (ANNs)

Kalteh, Aman Mohammad

2007

Rainfall-Runoff Modelling

Using Artificial Neural Networks (ANNs)

Aman Mohammad Kalteh

LUND
UNIVERSITY

Akademisk avhandling som för avläggande av teknisk doktorsexamen vid tekniska fakulteten vid Lunds Universitet kommer att offentligen försvaras vid Institutionen för Bygg och Miljöteknologi, Avdelningen för Teknisk Vattenresurslära, John Ericssons väg 1, Lund, Hörsal V:C, Fredagen den 11 Maj 2007, kl. 13.

Academic thesis submitted to Lund University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy (Ph.D. Engineering), will be publicly defended on May 11, 2007 at 1 pm in Lecture Hall V:C, Department of Water Resources Engineering, John Ericssons väg 1, Lund.
Faculty opponent: Professor Robin Clarke, Institute for Hydraulic Research, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

Organization	Document name
LUND UNIVERSITY Lund Institute of Technology Department of Water Resources Engineering Box 118 SE-221 00 Lund, Sweden	DOCTORAL THESIS Date of issue: **May, 2007** **CODEN** LUTVDG / (TVVR-1037) / 2007
Author(s) Aman Mohammad Kalteh	**Sponsoring organization** Ministry of Science, Research & Technology (MSRT) of Iran

Title and subtitle

Rainfall-runoff modelling: Using artificial neural networks (ANNs)

Abstract: This thesis considers two types of artificial neural networks (ANNs), namely, self-organizing map (SOM) and feed-forward multilayer perceptron (MLP).

The thesis starts with the issue of understanding of a trained ANN model by using neural interpretation diagram (NID), Garson's algorithm and a randomization approach. Then the applicability of the SOM algorithm within water resources applications is reviewed and compared to the well-known feed-forward MLP. Moreover, the thesis deals with the problem of missing values in the context of a monthly precipitation database. This part deals with the problem of missing values by using SOM and feed-forward MLP models along with inclusion of regionalization properties obtained from the SOM. The problem of filling in of missing data in a daily precipitation-runoff database is also considered. This study deals with the filling in of missing values using SOM and feed-forward MLP along with multivariate nearest neighbour (MNN), regularized expectation-maximization algorithm (REGEM) and multiple imputation (MI). Finally, once a complete database was obtained, SOM and feed-forward MLP models were developed in order to forecast one-month ahead runoff. Some issues such as the applicability of the SOM algorithm for modularization and the effect of the number of modules in modelling performance were investigated.

It was found that it is indeed possible to make an ANN reveal some information about the mechanisms governing rainfall-runoff processes. The literature review showed that SOMs are becoming increasingly popular but that there are hardly any reviews of SOM applications. In the case of imputation of missing values in the monthly precipitation, the results indicated the importance of the inclusion of regionalization properties of SOM prior to the application of SOM and feed-forward MLP models. In the case of gap-filling of the daily precipitation-runoff database, the results showed that most of the methods yield similar results. However, the SOM and MNN tended to give the most robust results. REGEM and MI hold the assumption of multivariate normality, which does not seem to fit the data at hand. The feed-forward MLP is sensitive to the location of missing values in the database and did not perform very well. Based on the one-month ahead forecasting, it was found that although the idea of modularization based on SOM is highly persuasive, the results indicated a need for more principled procedures to modularize the processes. Moreover, the modelling results indicated that a supervised SOM model can be considered as a viable alternative approach to the well-known feed-forward MLP model.

Key words: Artificial neural networks, Estimation, Feed-forward multilayer perceptron, Forecasting, Hydrological modelling, Missing values, Rainfall-runoff modelling, , Self-organizing map

Classification system and/or index terms (if any):

Supplementary bibliographical information:		Language English
ISSN and key title	1101-9824	ISBN 978-91-628-7138-3
Recipient's notes	Number of pages 146	Price
	Security classification	

Distribution by: Department of Water Resources Engineering, Lund Institute of Technology, Box 118, SE-221 00 Lund, Sweden

Signature Aman Mohammad Kalteh **Date:** May 11, 2007

DEPARTMENT OF WATER RESOURCES ENGINEERING
LUND INSTITUTE OF TECHNOLOGY, LUND UNIVERSITY
CODEN: LUTVDG/(TVVR-1037) /2007

Doctoral Thesis

Rainfall-Runoff Modelling

Using Artificial Neural Networks (ANNs)

by

Aman Mohammad Kalteh

May 2007

Rainfall-Runoff Modelling:
Using Artificial Neural Networks (ANNs)

Doktorsavhandling
Institutionen för Teknisk Vattenresurslära
Lunds Tekniska Högskola, Lunds Universitet

Doctoral Thesis
Department of Water Resources Engineering
Lund Institute of Technology, Lund University
Box 118
SE-221 00 Lund
Sweden

http://aqua.tvrl.lth.se/

Cover:
Sefid Rud Dam – Iran, Source: Water News Network: www.wnn.ir (accessed 27.10.2006).

CODEN: LUTVDG/(TVVR-1037) /2007
ISBN
978-91-628-7138-3
ISSN
1101-9824

بسم الله الرحمن الرحيم

وجعلنا من الماء كل شيء حي

To my dear parents,
sisters
and
brothers

ACKNOWLEDGEMENTS

First of all, I would like to express my profound gratitude to the Almighty Allah for all good abilities and opportunities he granted me in my life and career.

I would like to offer my sincerest thanks to my mother, Taghan Bibi, and to my father, Ashour, who supported me in my whole life with their indescribable kindness and sacrifice as well as with other matters. Also, special thanks go to my sisters and brothers for encouraging and supporting me in my study.

A full-time scholarship from the Ministry of Science, Research, and Technology (MSRT) of Iran to pursue my doctoral studies in abroad is gratefully acknowledged. The data used from Iran had been published by the Ministry of Energy of Iran. Thus, I would like to thank my brother, Anneh Mohammad, doctoral student at the Department of Mechanical Engineering, Amir Kabir University of Technology, Tehran, Iran for provision of the data to me.

I would also like to thank Associate Professor Peder Hjorth who collaborated with me throughout the work. I wish to extend my sincerest gratitude for his brilliant thoughts and inspiring guidance. I am also thankful to Professor Ronny Berndtsson for accepting me at the Department of Water Resources Engineering at Lund Institute of Technology.

ABSTRACT

Over the last decades or so, artificial neural networks (ANNs) have become one of the most promising tools for modelling hydrological processes such as rainfall-runoff processes. In most studies, ANNs have been demonstrated to show superior result compared to the traditional modelling approaches. They are able to map underlying relationships between input and output data without detailed knowledge of the processes under investigation, by finding an optimum set of network parameters through the learning or training process. This thesis considers two types of ANNs, namely, self-organizing map (SOM) and feed-forward multilayer perceptron (MLP).

The thesis starts with the issue of understanding of a trained ANN model by using neural interpretation diagram (NID), Garson's algorithm and a randomization approach. Then the applicability of the SOM algorithm within water resources applications is reviewed and compared to the well-known feed-forward MLP. Moreover, the thesis deals with the problem of missing values in the context of a monthly precipitation database. This part deals with the problem of missing values by using SOM and feed-forward MLP models along with inclusion of regionalization properties obtained from the SOM. The problem of filling in of missing data in a daily precipitation-runoff database is also considered. This study deals with the filling in of missing values using SOM and feed-forward MLP along with multivariate nearest neighbour (MNN), regularized expectation-maximization algorithm (REGEM) and multiple imputation (MI). Finally, once a complete database was obtained, SOM and feed-forward MLP models were developed in order to forecast one-month ahead runoff. Some issues such as the applicability of the SOM algorithm for modularization and the effect of the number of modules in modelling performance were investigated.

It was found that it is indeed possible to make an ANN reveal some information about the mechanisms governing rainfall-runoff processes. The literature review showed that SOMs are becoming increasingly popular but that there are hardly any reviews of SOM applications. In the case of imputation of missing values in the monthly precipitation, the results indicated the importance of the inclusion of regionalization properties of SOM prior to the application of SOM and feed-forward MLP models. In the case of gap-filling of the daily precipitation-runoff database, the results showed that most of the methods yield similar results. However, the SOM and MNN tended to give the most robust results. REGEM and MI hold the assumption of multivariate normality, which does not seem to fit the data at hand. The feed-forward MLP is sensitive to the location of missing values in the database and did not perform very well. Based on the one-month ahead forecasting, it was found that although the idea of modularization based on SOM is highly persuasive, the results indicated a need for more principled procedures to modularize the processes. Moreover, the modelling results indicated that a supervised SOM model can be considered as a viable alternative approach to the well-known feed-forward MLP model.

TABLE OF CONTENTS

LIST OF APPENDED PAPERS

The following papers are included in the thesis:

I. Kalteh, Aman Mohammad; (2007). **Rainfall-runoff modelling using artificial neural networks (ANNs): modelling and understanding.** (manuscript).

II. Kalteh, Aman Mohammad; Hjorth, Peder; Berndtsson, Ronny; (2007). **Review of self-organizing map (SOM) in water resources: analysis, modelling, and application.** *Environmental Modelling & Software* (submitted)*.

III. Kalteh, Aman Mohammad; Berndtsson, Ronny; (2007). **Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP).** *Hydrological Sciences Journal*, 52(2), 305-317.

IV. Kalteh, Aman Mohammad; Hjorth, Peder; (2007). **Imputation of missing values in a precipitation-runoff process database.** *Nordic Hydrology* (submitted)**.

V. Kalteh, Aman Mohammad; Hjorth, Peder; (2007). **Monthly runoff forecasting by means of artificial neural networks (ANNs).** *Hydrological Sciences Journal* (submitted).

Chapter 1. INTRODUCTION

1.1. Background

Relating rainfall (or more generally speaking, precipitation) to runoff, i.e. rainfall-runoff modelling, is, arguably, a fundamental activity of hydrological modelling. This relationship is believed to be highly non-linear and complex, since generating runoff from a watershed not only depends on various meteorological and watershed characteristic variables but also changes in both spatial and temporal scales. Generally three types of models, including deterministic (physical) models, conceptual models and empirical/systems theoretic/black-box models, are being used by hydrologists in order to model this relationship. The deterministic (physical) models describe the relationship using physical laws of mass and energy transfer (Dawson & Wilby, 2001). In contrast, in conceptual models instead of using physical laws of mass and energy transfer, a simplified, but a plausible or reliable conceptual representation of the underlying physics is adopted (Jain & Srinivasulu, 2006).

An alternative modelling approach for hydrological processes such as rainfall-runoff process is the empirical/systems theoretic/black-box models, which try to find a relationship between historical inputs and outputs (ASCE Task Committee, 2000a) without detailed understanding of the physics involved in the process under investigation, such as artificial neural networks (ANNs). According to ASCE Task Committee (2000a; b), an ANN is a massively parallel-distributed information processing system resembling biological neural networks of the human brain and capable of solving large-scale complex problems such as pattern recognition, non-linear modelling, classification, association and control. Due to the superiority of their performance compared to the alternative counterparts, ANN models have been widely used by hydrologists particularly in modelling of the rainfall-runoff process (e.g. Hsu *et al.*, 1995; Lorrai & Sechi, 1995; Minns & Hall, 1996; Dawson & Wilby, 1998; Tokar & Johnson, 1999; Rajurkar *et al.*, 2002; Wilby *et al.*, 2003; Giustolisi & Laucelli, 2005; Jain & Srinivasulu, 2006). This thesis adopts two types of ANN models, namely, feed-forward multi-layer perceptron (MLP) which is the most commonly used ANN in hydrological applications and self-organizing map (SOM) (Kohonen, 2001). A comprehensive review of ANNs along with their applications in hydrology can be found in Maier & Dandy (2000), ASCE Task Committee (2000a; b), Dawson & Wilby (2001) and Kalteh *et al.* (2007).

1.2. The objectives of study

The general objective of the present thesis is to contribute to *modelling rainfall-runoff processes using feed-forward MLP and SOM ANN types for efficient planning and management of water resources such as flood control and management*. The thesis is based upon the following research questions, as prior to each question (s) an introduction is given:

i. ANNs have been known as black-box models due to their problem in providing insight into the modelled relationship. Is it possible to get any explanation about the built ANN model?

ii. The feed-forward MLP is the most commonly used ANN type in hydrological applications. Did SOM get enough attention among hydrologists?

iii. In general, the problem of missing values is a common obstacle in time series analysis and specifically in the context of a precipitation-runoff process modelling where it is essential to have serially complete data. How can the problem of missing values be dealt with using the feed-forward MLP and SOM?

iv. Over the last decades or so, ANNs have become one of the most promising tools for modelling hydrological processes such as rainfall runoff processes. Does a single model seem to be an appropriate approach for modelling such a complex, non-linear, and discontinuous process that varies in space and time? Is the SOM suitable for decomposition (segmentation)? What is the effect of the number of clusters (segments) based on SOM in modelling performance? Can SOM models be applied for building functional relationships between input and output data of the system under investigation?

1.3. The appended papers

The five appended papers in the thesis attempt to provide answers to the above research questions. Figure 1.1 demonstrates the main topics and interactions among the papers. To clarify, a brief summary for each paper is given in the following paragraphs.

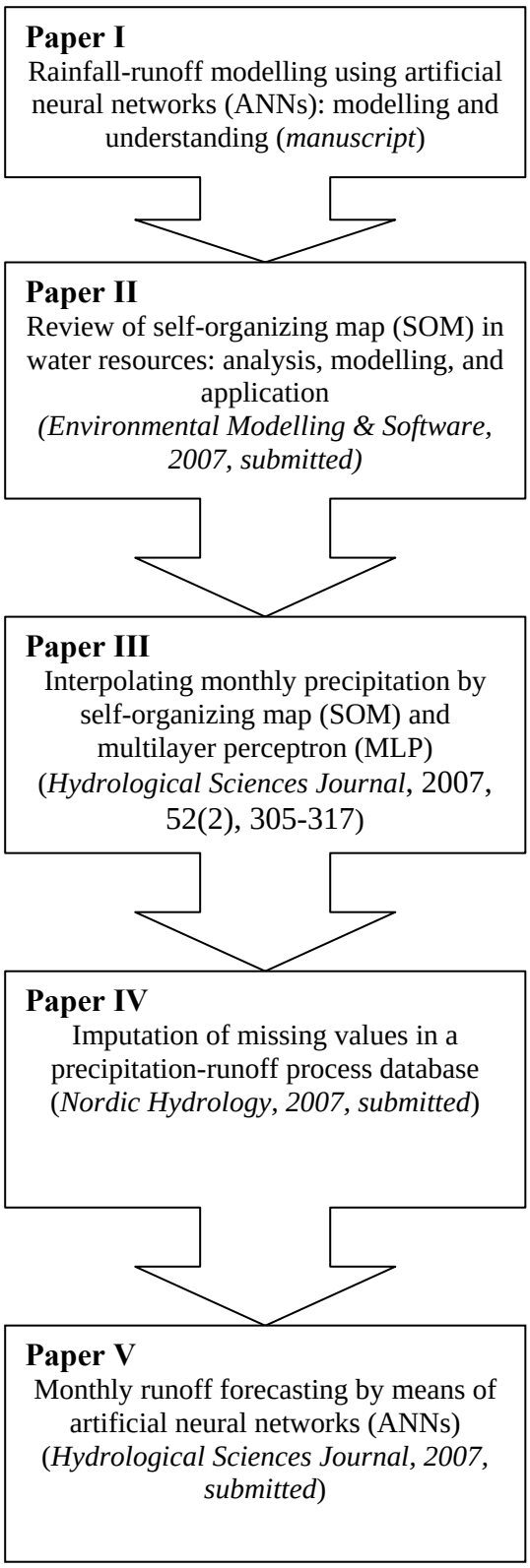
Regarding the first question, ***Paper I*** – this will be referred to as ***P1*** in this thesis – addresses three approaches for understanding of ANN models based on an ANN rainfall-runoff model example. It describes three different approaches, including Neural Interpretation Diagram, Garson's algorithm, and randomization approach, for understanding of the relationship learned by the ANN model. The results indicate that ANNs are promising tools not only in accurate modelling of the complex process but also in providing insight from the learned relationship which would assist the modeller in understanding the process under investigation as well as in evaluation of the model.

Addressing the second question, ***Paper II*** – this will be referred to as ***P2*** in this thesis – the SOM algorithm is reviewed in water resources. The paper attempts firstly to explain the algorithm and secondly, to review published applications with main emphasis on water resources problems in order to get a good picture of how well SOM can be used to solve a particular problem. It was concluded that SOM is a promising technique suitable to apply in many types of water resources processes.

In response to the third question, ***Paper III*** – this will be referred to as ***P3*** in this thesis – deals with the missing values problem in a precipitation database using feed-forward MLP and SOM models in northern Iran. The paper uses SOM both for regionalization (using component plane visualization technique) and for estimating monthly precipitation for stations with missing data for 1-, 2-, 5-, and 10-year periods using a jack-knife procedure to obtain objective results. According to the results, the SOM is able both to find regions with similar precipitation mechanisms and to estimate with accuracy. The results indicate that precipitation estimation will improve considerably by taking into account the regionalization properties in the SOM modelling. Furthermore, the SOM results were compared with those from a well-known feed-forward MLP. Our findings suggest that without regionalization feed-forward MLP is generally better than SOM. However, when regionalization is included SOM performs better than MLP.

In response to the third question, ***Paper IV*** – this will be referred to as ***P4*** in this thesis – also deals with missing values problem, but in a precipitation-runoff process database and by using SOM, feed-forward MLP, multivariate nearest neighbour (MNN), regularized expectation-maximization algorithm (REGEM) and multiple imputation (MI) in northern Iran. The paper also argues the pros and cons of these methods. It was found that the SOM and MNN tend to give similar and robust results. REGEM and MI build on the assumption of multivariate normal data, which we don't seem to have in one of our cases. MLP tends to produce inferior results because it fragments the data into 68 different models. Therefore, it was concluded that it makes most sense to use either the computationally simple MNN method or the more demanding SOM.

Finally, ***Paper V***– this will be referred to as ***P5*** in this thesis – aims to respond to the questions raised in the fourth research question. Generally, the paper is about one-month ahead runoff forecasting. It aims at decomposing the process into different clusters or segments based on SOM ANN approach, and thereafter modelling different clusters into corresponding outputs using separate feed-forward MLP and supervised self-organizing map (SSOM) ANN models. Specifically, three different SOM models have been employed in order to cluster the input patterns into two, three, and four clusters respectively so that each cluster in each model corresponds to certain physics of the process under investigation and thereafter modelling of the input patterns in each cluster into corresponding outputs using feed-forward MLP and SSOM ANN models. The employed models were developed on two different watersheds, one Iranian and one Canadian. It was found that although the idea of decomposition based on SOM is highly persuasive, our results indicate that there is a need for more principled procedure in order to decompose the process. Moreover, according to the modelling results the SSOM can be considered as an alternative approach to the feed-forward MLP.



- *An ANN rainfall-runoff model is developed in order to understand the mechanisms modelled by ANNs.*
- *The SOM algorithm is comprehensively reviewed in water resources in order to get a picture of the state of the art of its application.*
- *Missing values problem is addressed by using SOM and feed-forward MLP in a monthly precipitation data,*
- *The SOM application for regionalization (input determination) is introduced.*
- *Missing values problem is addressed by using SOM, feed-forward MLP, multivariate nearest neighbour (MNN), regularized expectation-maximization algorithm (REGEM) and multiple imputation (MI) in a daily precipitation-runoff data,*
- *The pros and cons of these methods are discussed.*
- *The SOM and feed-forward MLP are applied for modelling rainfall-runoff,*
- *The suitability of SOM for decomposition is investigated.*

Figure 1.1. Overview of the appended papers

Chapter 2. MISSING VALUES

In general, the problem of missing values is a common obstacle in time series analysis and specifically in the context of a precipitation-runoff process modelling where it is essential to have serially complete data.

There may be various reasons for missing values, for instance equipment failure, errors in measurements or faults in data acquisition and natural hazards such as landslides. Whatever the reasons, missing values produce a significant problem for water resources applications, which generally require continuous database (e.g. Zoppou *et al.*, 2000; Junninen *et al.*, 2004; Ramirez *et al.*, 2005). Consequently, finding efficient and principled methods to deal with the problem of missing values is an important issue in most hydrological analyses. However, hydrological modellers commonly discard the observations with missing values and only use the observations with complete information hence this means that a lot of the information contained in the data set is lost. Furthermore, the method is inadequate for analyses that require serially complete data. As an alternative to this listwise deletion procedure ("or complete-case analysis"), the modellers may impute ("or fill in") a value for the missing values by using e.g. the mean of the observed variables. Such a procedure may, however, seriously distort statistical properties like standard deviation, correlations, or percentiles.

During the last twenty years, statisticians have tried to introduce model-based methods such as maximum likelihood using the expectation-maximization (EM) algorithm and multiple imputation (MI), arguing that they provide sounder and more promising solutions for a wider range of situations. However it is worth to mention that each of the above methods for handling missing values carries assumptions about the missing data mechanisms. According to Little & Rubin (2002) missing data mechanisms can be divided into three classes that consist of missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR occurs when the probability of an observation having a missing value for a component does not depend on either the available values or the missing values. MAR occurs when the probability of an observation having a missing value for a component may depend on the available values, but not on the missing values itself. And finally, NMAR occurs when the probability of an observation having a missing value for a component could depend on the value of that component.

According to Little & Rubin (2002), there are different methods for dealing with missing values and they can be divided into three categories such that they will be shortly reviewed in the following subsections.

2.1. Listwise /pairwise deletion

In the listwise deletion strategy (or "complete-case analysis"), all of the cases (observations) with missing values are discarded from the analysis. This strategy is easy to carry out and it is the one most commonly practised and it is the default option in most statistical software packages. However, it should be mentioned that this strategy may be satisfactory when small fraction of database are missing and the missing data mechanism is MCAR (Little & Rubin, 2002). Pairwise deletion (or "available-case analysis") uses different sets of sample observations for each statistic. This strategy preserves more information compared to

listwise deletion. However, the reliability of its estimates depends not only on the MCAR assumption about missing data mechanism but also on existing correlations among the variables in the database (Pigott, 2001; Little & Rubin, 2002; Tsikriktsis, 2005).

2.2. Imputation-based procedures

In imputation-based procedures, observations with missing values are imputed with plausible values rather than deleting observations entirely. Imputation has several advantages such as its efficiency and precision if compared with complete-case analysis because no observations are discarded. However, it suffers from implementation difficulties, especially in a multivariate database. Moreover, some techniques can falsify data relationships and distributions (Schafer & Graham, 2002). Many different missing data imputation procedures have been applied and among them the mean, regression and hot deck methods are the most commonly used ones. In mean imputation procedure, means from sets of recorded values are substituted. The regression imputation procedure can be summarized as two-step approach as follows (Frane, 1976):

- First, the relationships among variables are estimated,
- And second the regression coefficients are used to estimate missing value. It requires MAR data.

The hot deck imputation procedure uses values from similar observations to impute.

2.3. Model-based procedures

Two model-based procedures, including maximum likelihood using the expectation-maximization (EM) algorithm and multiple imputation (MI), are briefly illustrated in this subsection. Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. A general method for maximum likelihood in missing data problems was described by Dempster *et al.* (1977) on their seminal paper on the EM algorithm. Maximum likelihood procedures for missing multivariate normal data can be parameterized based on the mean vector and the covariance matrix. In every iteration of the EM algorithm, estimates of the mean and the covariance matrix are adjusted in three stages. First, based on estimated parameters (the mean and the covariance matrix), parameters of the model such as regression coefficients for the variables with missing values are computed based on the variables with available values. Second, missing values are filled in with their conditional expectation values which are the product of the available values and the estimated model parameters such as regression coefficients. Third, the parameters of the mean and the covariance matrix are re-estimated. Then, the EM algorithm process cycles back and forth until the imputed values and the estimates of the mean and the covariance matrix parameters do not change substantially.

In MI, a researcher specifies an appropriate imputation model, imputes several complete databases (usually 3-5 times), performs desired statistical analysis on each database separately by using standard complete-data methods, and thereafter combines results

(Allison, 1998; Patrician, 2002). The MI procedure conducted for this thesis (***P4***) will be discussed in the section corresponding to missing values in Chapter 3. Both maximum likelihood using the EM algorithm and MI require the assumptions of multivariate normality and MAR.

Chapter 3. METHODOLOGY

3.1. Understanding of ANNs

ANNs are often criticized for their black-box nature, as they are unable to reveal any explanation regarding how the model was built. As the ASCE Task Committee (2000b) stated: "*For ANNs to gain wider acceptability, it is increasingly important that they have some explanation capability after training has been completed. Most ANN applications have been unable to explain in a comprehensibly meaningful way the basic process by which ANNs arrive at a decision. It is highly desirable that ANNs be capable of imparting an explanation, even if only a partial one, as an integral part of its function.*" In this thesis, in order to obtain some explanation from a trained feed-forward MLP ANN, the following approaches have been used:

3.1.1. Neural interpretation diagram (NID)

The NID was proposed by Özesmi & Özesmi (1999) and was used for interpretation of the connection weights obtained from an ANN model based on their magnitude and direction in order to visually depict the relationship between input output data as well as interactions among the input variables. Positive effect of an input variable on the output variable is depicted if the connection weight from a neuron in the input layer to a neuron in the hidden layer and the connection weight from the neuron in the hidden layer to a neuron in the output layer are both positive or both negative, whereas negative effect of an input variable on the output variable is depicted if these connection weights are of opposite signs. And finally, input variables can be identified as interacting if the connection weights entering the same hidden neuron are of opposite signs.

3.1.2. Garson's algorithm

Garson (1991) introduced a method which determines the relative contribution of each input variable used in modelling the output based on the products of the input-hidden and hidden output connection weights. Although this algorithm provides the contribution of each input variable on the output quantitatively, as it will be explained later it uses absolute values of connection weights (i.e. without considering the direction of the relationship) which would lead to misinterpretation of the importance of input variables in modelling the output of the system. However, the algorithm can be summarized , referring to Figure 3.1, as follows.

i. Preparation of the input-hidden-output neuron connection weights obtained from the ANN model.

ii. Calculation of contribution of each input neuron to the output via each hidden neuron. For example, the contribution of the input neuron 1 via hidden neuron A is computed as follows:

$$C_{A1} = W_{A1} \times W_{OA} \tag{1}$$

iii. Calculation of the relative contribution of each input neuron to the outgoing signal of each hidden neuron, for example:

$$r_{A1} = \frac{|C_{A1}|}{|C_{A1}| + |C_{A2}| + |C_{A3}|} \tag{2}$$

And sum of input neuron contributions, for example:

$$S_1 = r_{A1} + r_{B1} \tag{3}$$

The same must be done for S_2 and S_3.

iv. And finally, calculation of relative importance of each input variable, for example:

$$RI_1 = \frac{S_1}{(S_1 + S_2 + S_3)} \times 100 \tag{4}$$

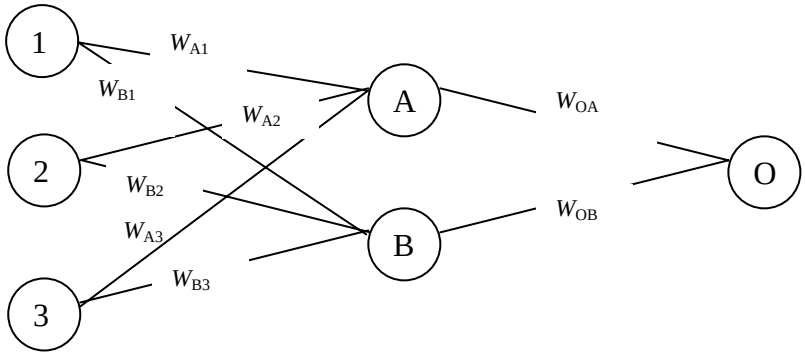


Figure 3.1. An ANN structure for the illustration of Garson's algorithm

3.1.3. Randomization approach

This approach was proposed by Olden & Jakson (2002) who used an additional measure to investigate the importance of the input variables, i.e. the overall connection weight as defined below. In this thesis, this approach was carried out as follows:

i. Construct a number of ANN models using the original input output data with randomly generated initial weights;
ii. Select the ANN model with the best performance and calculate and record:

a) input-hidden-output connection weights: the product of input-hidden and hidden-output connection weights, for instance C_{A1} : see Garson's algorithm, step 2;
b) overall connection weight: the sum of the products of the input-hidden and hidden-output connection weight for each input variable, for instance $C_1 = C_{A1} + C_{B1}$;

iii. Randomly permute the original output data;
iv. Construct an ANN using randomized outputs;

v. Repeat steps 2,3 and 4 a large number of times.

In this thesis, the overall connection weights were used in order to assess the importance of the input variables on the output in such a way that if the overall connection weight of an input variable is greater than 95% of the randomized overall connection weights for the same input variable, then the input variable can be considered to be significant with a 95% confidence level.

3.2. Missing values

In ***P4***, the following methods have been applied to impute missing values in a daily precipitation-runoff database in a watershed located in northern Iran as follows.

3.2.1. Self-organizing map (SOM)

In ***P4***, a batch learning algorithm was used for constructing the SOM, because the batch SOM is faster and eliminates the specification of learning rate; hence avoids convergence problems (Kohonen, 2001). The batch training of the SOM is similar to the sequential training algorithm, which is commonly used, in the sense that both are iterative, but in the batch algorithm the whole database is presented to the map before any updates are made, that is, in each training iteration, the batch algorithm lists input vectors one by one under the best matching units and updates the neurons according to the whole database. The procedure can be summarized as follows. At the outset of the training, a weight vector w_j must be initialized to each neuron and thanks to an unsupervised or self-organization procedure the input vectors (x_i) are compared with the SOM neurons to find the closest match which is called best matching units (BMUs). The most commonly used criterion for comparison is the Euclidean distance. Thus, the new weight vectors are calculated using the following rule:

$$w_j(t+1) = \frac{\sum_{i=1}^{n} H_{j^*}(t) x_i}{\sum_{i=1}^{n} H_{j^*}(t)} \tag{5}$$

where $H_{j^*}(t)$ is the neighbourhood function (Gaussian in this study) of the best matching neurons j^* at iteration t. Finally, n is the total number of the input vectors (x_i). The Gaussian neighbourhood function, which is commonly used, is as follows.

$$H_{j^*}(t) = \exp\left(-\frac{\left\| r_{j^*} - r_j \right\|^2}{2\delta^2(t)}\right) \tag{6}$$

where $\left\| r_{j^*} - r_j \right\|$ is the distance between neurons j^* and j and $\delta(t)$ is the neighbourhood radius at iteration t.

These procedures must be iterated several times until the results can be considered as steady.

In ***P4***, missing values were ignored during training. Once the SOM was trained, the BMUs for the incomplete data vectors were found and the missing values were filled in by copying the corresponding values of the BMUs (weight vectors).

3.2.2. Feed-forward MLP

The most widely used ANN in water resources and hydrological applications is the feed-forward MLP. The structure and training of feed-forward MLP model will be explained later in the rainfall-runoff modelling section.

In ***P4***, several MLP models were separately trained, one per missing variables combination. The number of input neurons was equal to the number of available values and the number of output neurons was equal to the number of missing values at each missing variables combination. The number of hidden neurons was determined by using the following formula which was adopted by Junninen *et al.* (2004) in their application on air quality data sets.

$$\begin{aligned} &H_{in} = (2 \times N_{in}) + 1, \\ &H_{out} = (2 \times N_{out}) + 1, \\ &\text{if } H_{out} < H_{in}, \text{ then } H_{MLP} = H_{in}, \text{ else } H_{MLP} = H_{out}, \end{aligned} \tag{7}$$

where H_{in} and H_{out} : the number of hidden neurons determined by the number of input and output neurons (N_{in} and N_{out}), respectively. In ***P4***, the above formula was used to determine the number of neurons in the hidden layer of each MLP model per missing variables combination.

3.2.3. Multivariate nearest-neighbour (MNN)

As described in Dixon (1979), this approach means that the missing values of a vector are filled in by determining what vector that is most similar to the vector of interest and replacing the missing values by corresponding values from the that vector. This similarity was computed by a distance function which was Euclidean in ***P4***. The procedure can be summarized as follows:

i. Divide the data set into two parts: one that contains vectors in which at least one of the components is missing while the remaining part will form the data in which the complete vectors, i.e. without any missing values in any components, are located.
ii. For each vector in the data matrix with missing values find the nearest neighbour from the complete data matrix. In distance calculation, use only components that are not missing.

3.2.4. Regularized expectation-maximization (REGEM) algorithm

In ***P4***, missing values are also filled in with the regularized EM algorithm. In an iteration, estimates of the mean and covariance matrix are revised in three steps. First, for each incomplete observation, the regression parameters of the variables are computed from the estimates of the mean and the covariance matrix. Second, the missing values are filled in

with conditional expectation values given the available values and mean and covariance matrix estimates, the conditional expectation values being the product of the available values and the estimated regression coefficients. Third, re-estimate the mean and the covariance matrix, the mean as the sample mean of the completed database and the covariance matrix as the sum of the sample covariance matrix of the completed database and an estimate of the conditional covariance matrix of the error of the imputed values (Schneider, 2001).

3.2.5. Multiple imputation (MI)

In ***P4***, Schafer's (1999) NORM program was used in order to generate 5 possible values for each missing observation using a data augmentation algorithm. NORM refers to the multivariate normal distribution, the model used to generate the imputations. The data augmentation algorithm treats parameters and missing data as random variables and simulates random values of parameters and missing data from their conditional distribution. The procedure of simulating parameters and missing data generates a chain that for a sufficient number of iterations converges to the Bayesian posterior distribution. Prior to the run of the data augmentation algorithm the EM algorithm was run, in order to obtain initial values to the data augmentation algorithm and to assess the number of iterations needed to create statistically independent imputations. By specifying 50 iterations as sufficient to ensure statistically independent values, the augmented databases at iterations 50; 100; 150; 200 and 250 were saved, hence generating 5 completed databases. As stated in Chapter 2, each completed database can be analyzed by using standard statistical methods separately and thereafter the obtained estimates can be combined by averaging over estimates, in order to obtain a single estimate. As stated by Rubin (1996), the main objective of MI is to get valid statistical inference about the database with missing values rather than optimal point prediction.

3.3. Rainfall-runoff modelling

In ***P5***, SOM and feed-forward MLP ANN models were used in order to forecast monthly runoff in one Iranian and one Canadian watershed. However, it may be mentioned that in ***P1*** a feed-forward MLP model was developed in order to model the rainfall-runoff relationship as well as understanding the mechanisms modelled by ANNs. The SOM and feed-forward MLP models will be discussed as follows.

3.3.1. Self-organizing Map (SOM)

The self-organizing map (SOM) is an unsupervised learning algorithm within the family of artificial neural networks (ANNs), which was initially proposed by Kohonen (1982a; b). The SOM is a fascinating neural network method that has found increasing interest in water resources applications such as classification of satellite imagery data and rainfall estimation (e.g. Murao *et al.,* 1993) and rainfall-runoff modelling and analysis (e.g. Hsu *et al.,* 2002). Typically SOM networks learn to cluster groups of similar input patterns from high dimensional input space in a non-linear fashion onto a low dimensional (most commonly two dimensional) discrete lattice of neurons in a Kohonen layer or output layer (Kohonen, 2001). This is done in such a way that neurons physically located close to each other in the output layer of the SOM respond to similar input patterns (combining clustering and

ordering processes in SOM). Output layers of higher dimensions are also possible, but they will not be so convenient for visualization purposes and consequently they are not so common (Vesanto, 1999). Discrete lattice of neurons can be arranged to be either hexagonal or rectangular but hexagonal is preferred because of being effective for visualization and/or more convenient to the eye (Vesanto, 1999; Kohonen, 2001). The main advantages of this algorithm are that it is non-linear and has an ability to preserve the topological structure of the data (Corne *et al.,* 1999; ASCE Task Committee, 2000a). In general, a SOM clusters the samples or patterns into predefined (i.e. the number of neurons is selected by the modeller) classes (clustering property) and also orders the classes into meaningful maps (topology preservation or ordering property). The typical structure of a SOM consists of two layers: an input layer and a Kohonen layer or output layer (Figure 3.2). The input layer contains one neuron for each variable (e.g. precipitation, temperature, etc.) in the data set. In the Kohonen layer all neurons, most commonly in a two-dimensional array, are connected to every neuron in the input layer through adjustable weights or network parameters or connection weights. The weight vectors in the Kohonen layer give a representation of the distribution of the input vectors in an ordered fashion.

To illustrate how the algorithm performs, a conceptual framework which summarizes the successive procedures needed in a typical water resources application, e.g. analysis and modelling, is shown in Figure 3.3. The conceptual framework can be divided into three categories, namely:

i. Data gathering and normalization

This category considers the issues of the data matrix preparation and normalization procedures which are required before the application of the SOM algorithm. In order to use the SOM Toolbox (available at: http://www.cis.hut.fi/projects/somtoolbox), the data matrix must be organized as shown in Table 3.1, in which m and p is the number of variables and patterns, respectively. After preparation of the data matrix, the next step is to normalize the data. As it will be explained in the training category, the distance measure that is mostly used in SOM applications is the Euclidean metric. Thereafter, the most important part in normalizing is to prevent the formation of a map with variables that have higher impact as compared to other variables. Consequently, normalization ensures that all variables have equal importance in the formation of the SOM. In this thesis, the normalization is carried out either by subtracting the mean and dividing by the standard deviation or by subtracting the minimum and dividing by the difference between maximum and minimum observation.

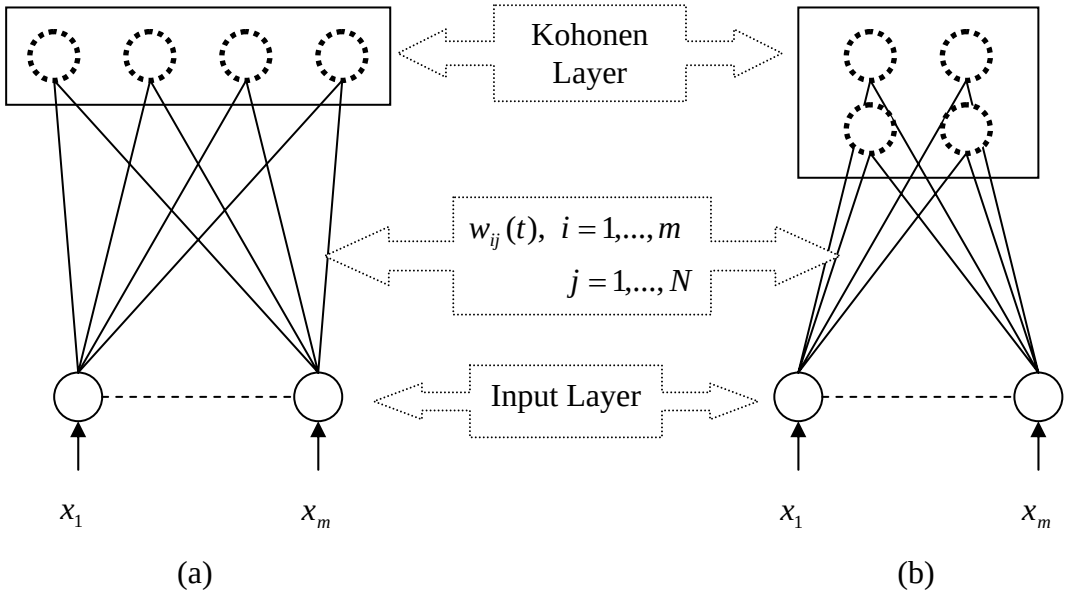


Figure 3.2. (a) A one-dimensional self-organizing map (SOM) with 4 neurons in the Kohonen layer, (b) a two-dimensional self-organizing map (SOM) with 2× 2 neurons in the Kohonen layer

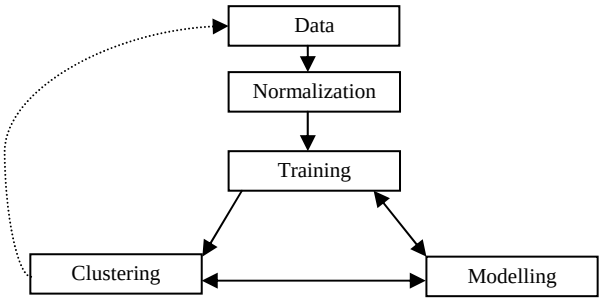


Figure 3.3. A framework of self-organizing map (SOM) application in water resources

Table 3.1. The data matrix for SOM's application in water resources

1						*m*
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
p						

ii. Training

After data preparation and normalization, an input vector from the data matrix is introduced to the iterative training procedure to form a SOM. The unsupervised or competitive training of connection weights in SOM can be summarized as follows (Haykin, 1999):

- Initialize randomly the weights for each SOM connection weight:

$$w_{ij},\ i = 1,...,m\ \ j = 1,...,N \tag{8}$$

where m denotes the number of input variables, N denotes the number of neurons in the Kohonen layer.

- Calculate the best matching unit (BMU) at each training step t, using the minimum-distance Euclidean criterion, of input pattern $x(t)$ from the training data as follows:

$$d_j = \left[\sum_{i=1}^{m}(x_i(t) - w_{ij})^2\right]^{0.5} \tag{9}$$

- Update the weights of the winner or BMU and its neighbours, using the following rule:

$$w_{ij}(t+1) = w_{ij}(t) + \varepsilon(t)(x_i(t) - w_{ij}(t)), \quad \text{if } j \in H_c(t)$$
$$w_{ij}(t+1) = w_{ij}(t) \qquad \text{if } j \notin H_c(t) \tag{10}$$

where $\varepsilon(t)$ is the learning rate and $H_c(t)$ is the size of a neighbourhood around the BMU c at iteration t. After the weights have been updated, the next input pattern from the training data is presented to the network and the process continues until convergence.

Although the SOM is usually used in applications with the above objective i.e. clustering or grouping of similar input patterns, we also used it in a supervised manner in order to build functional relationships between input and output data. However the supervised SOM (SSOM) is similar to SOM and the difference lies in a minor modification needed during training of the network so that finding the BMU in the SSOM is based on the input portion

of data presented to the network while updating applies to all input-output data. The supervised SOM was used in ***P3*** and ***P5***.

iii. Extracting information from the trained SOM

To summarize, after data gathering and normalization, a SOM is trained. Once the training of the SOM was accomplished the resulting map could be post processed based on visualization, clustering and local modelling purposes. This category has been comprehensively investigated by Vesanto (2002) hence the interested reader is referred to this reference. From the water resources engineer and hydrologist point of view, the trained SOM map is a valuable tool for visualization of relatively large amounts of data and for providing insight into the system under investigation such as precipitation processes (e.g. Kalteh & Berndtsson, 2007) rainfall-runoff processes (e.g. Hsu *et al.*, 2002). The former authors used a SOM component plane visualization technique (see ***P3***) for correlation hunting in order to regionalize precipitation stations in a large watershed, while in the latter application; the authors used the SOM output to characterize a rainfall-runoff process. The applications of SOM in water resources and hydrology have been comprehensively reviewed in ***P2***. As mentioned above, the SOM is an unsupervised clustering algorithm and in most applications reviewed in ***P2***, it is used for this purpose. In ***P5***, the SOM was used to cluster or group input patterns prior to modelling the input-output relationship. Thereafter, separate models were developed for each group of input patterns, corresponding to different physics of the processes in the watershed.

To illustrate the information which is available from a trained SOM, the data matrix of 34-year monthly precipitation data, the same as those in ***P3***, were used as input to the SOM. According to Table 3.1, the variables 1 through m (here m = 24) are precipitation data at 24 rain gauge stations and the observations or patterns from 1 through p (here p = 408) are the observational patterns of each variable corresponding to monthly precipitation. In the following paragraphs the results obtained by training a SOM with 11×10 neurons in the output layer using the above data are illustrated.

For instance, Figure 3.4 shows clustering results of the SOM in our illustration example. As mentioned previously, in most applications the SOM is used for this purpose. The number in each neuron indicates the number of projected patterns to that particular neuron. Thus, the SOM maps similar patterns onto the same neuron. In ***P5***, the SOM was used to cluster or group input patterns prior to modelling the input-output relationship. Thereafter, separate models were developed for each group of input patterns situated in a neuron, corresponding to different physics of the processes in the watershed. However, as mentioned previously, the SOM output preserves the topology so that the neurons that are physically close are more similar than those located further away from each other. Therefore, an interesting idea would be to cluster the output of a SOM by using other clustering algorithms such as the k-means clustering algorithm. This issue was clarified in ***P2***.

The SOM algorithm can, however, also be used for visualization of the contribution of each variable for all mapped neurons via the component plane visualization technique. In order to illustrate the applicability of this technique in hydrology, we applied to the correlation between the variables in our illustration example. To do so, we used the above technique of

SOM in which each component plane shows the values at a single precipitation station for all mapped neurons. This is done using color-coding as shown in Figure 3.5. In the component planes, each cell corresponds to each mapped neuron or data pattern. As seen from Figure 3.5, the colors of mapped neurons have been chosen so that the red color indicates high values. By looking at the component planes pattern similarity between the precipitation stations is visualized. Human eyes are impressive in this sense. For example, station 16009, 16025, and 16029 show similarity and consequently have high correlation in terms of precipitation. This finding can be used for regionalization ("input determination") as carried out in ***P3***.

An interesting feature of the SOM algorithm which can be used in modelling is that the SOM output array provides graphical visualization of the distribution function as depicted in Figure 3.6. Taking precipitation stations in our illustration example, i.e. the input vectors to the SOM are time series of precipitation, weight vectors in the SOM neurons approximate mean precipitation fields of the projected patterns to that particular neuron. This characteristic can be used for filling in the gaps in time series as in ***P3*** and ***P4*** as well as input-output mapping as in ***P5***.

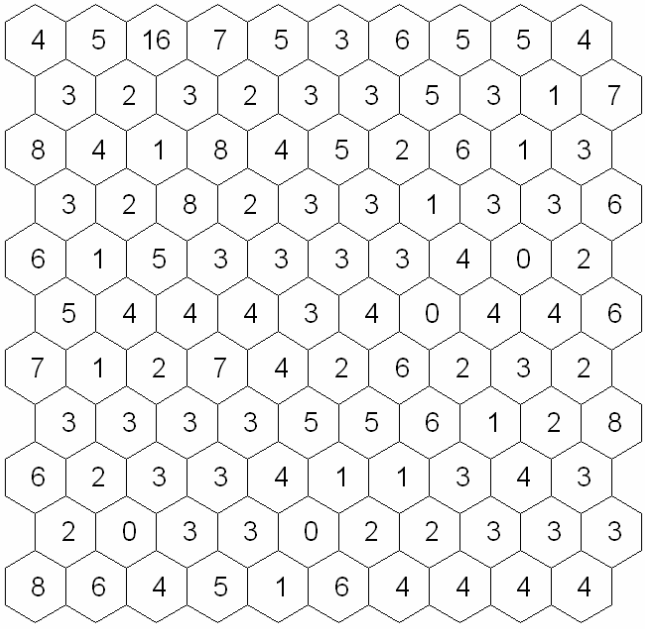


Figure 3.4. Two-dimensional Kohonen map obtained by a network of 11×10 neurons

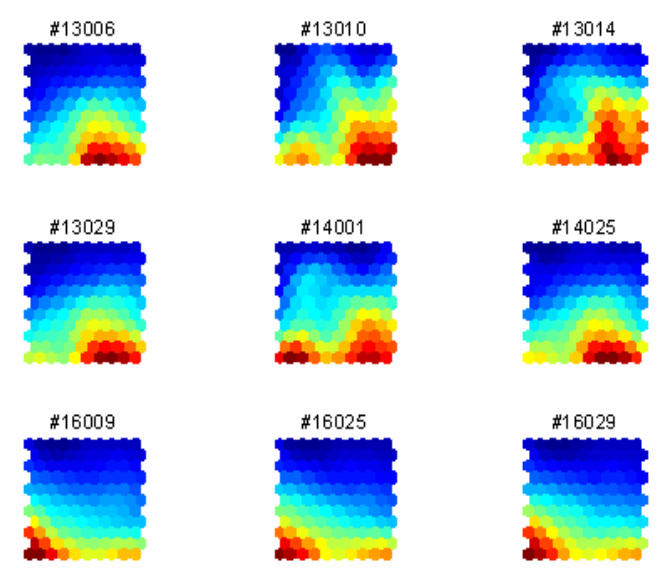


Figure 3.5. Component planes for example precipitation stations

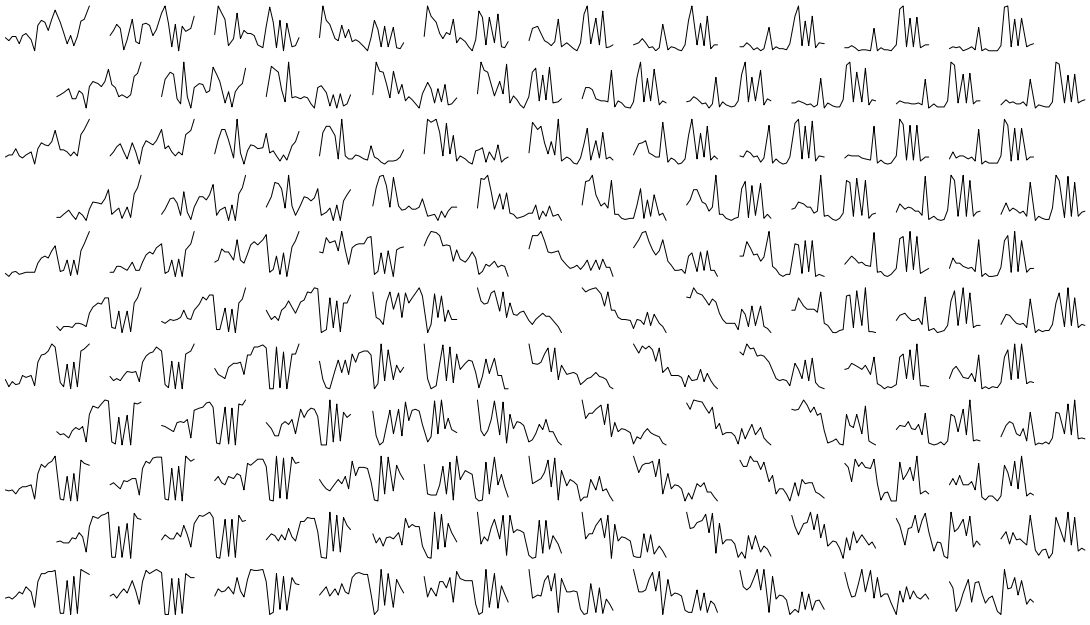


Figure 3.6. Weight vectors of the precipitation fields stored in 11×10 neurons

3.3.2. Feed-forward MLP

The most commonly used ANN is the feed-forward MLP as shown in Figure 3.7. The figure shows a three-layer feed-forward MLP that consists of an input layer, a hidden layer, and an output layer. Each neuron is represented by a circle and each connection weight by a line so that each neuron in a layer is connected to all the neurons of the next layer while the neurons

in one layer are not connected among themselves. Each individual neuron multiplies every input by its connection weight, sums the product, and then passes the sum through a non-linear function called the activation function in order to compute its output. The number of input and output layer neurons depends upon the problem at hand so that the number of neurons in the input layer is equal to the number of input variables (denoted with m) and the number of neurons in the output layer is equal to the number of output variables while the number of neurons in the hidden layer is usually selected via a trial-and-error procedure. Determination of connection weights is called training process. In this thesis, a back-propagation algorithm was used for training the feed-forward MLPs. In a feed-forward MLP, patterns from the inputs presented to the neurons in an input layer are propagated through the network from the input layer to the output layer, i.e. in a forward direction and the outputs from the network are compared with the target values in order to compute the error. Thereafter the calculated error is back-propagated through the network and the connection weights are updated (ASCE Task Committee, 2000a). The training process is repeated until an acceptable convergence is achieved. Subsequently, the network is able to compute outputs given inputs that have not been seen by the network before.

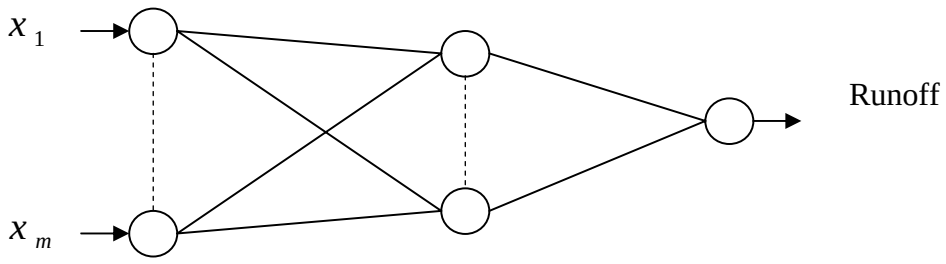


Figure 3.7. Three-layer feed-forward multilayer perceptron (MLP)

3.4. Model performance

In most of studies in this thesis, the correlation coefficient (r), coefficient of determination (R^2), and root mean square error ($RMSE$) were used as performance criteria to evaluate the various ANN models. These criteria were calculated using equations (11), (12), and (13), respectively as follows:

$$r = \frac{\sum_{k=1}^{p}(Q(k)-\overline{Q})(\hat{Q}(k)-\tilde{Q})}{\sqrt{\sum_{k=1}^{p}(Q(k)-\overline{Q})^2 \sum_{k=1}^{p}(\hat{Q}(k)-\tilde{Q})^2}} \qquad (11)$$

$$R^2 = 1 - \frac{\sum_{k=1}^{p} (Q(k) - \hat{Q}(k))^2}{\sum_{k=1}^{p} (Q(k) - \overline{Q})^2} \tag{12}$$

$$RMSE = \left[\frac{\sum_{k=1}^{p} (Q(k) - \hat{Q}(k))^2}{p} \right]^{0.5} \tag{13}$$

where $\hat{Q}(k)$ are the p forecasted (estimated) runoff values, $Q(k)$ are the p observed runoff values, $\overline{Q}$ is the mean of the observed runoff values, and $\tilde{Q}$ is the mean of the forecasted (estimated) runoff values.

Chapter 4. STUDY AREA AND DATA

4.1. Caspian Sea watershed

In this work a part of the Iranian Caspian Sea watershed bordering Astara in the western Gilan province, to Neka, in the eastern Mazandaran province extending from the Alborz mountains northward to the sea was considered as study area, as shown in Figure 4.1. The Caspian Sea watershed is situated between 35°45' and 38°25´N latitude, and 48°35´ and 54°00´E longitude. This large watershed is divided in terms of topography into two parts: a coastal plain and a mountainous part. Minimum and maximum elevation is -28 m for the Caspian Sea and 5671 m for the Damavand summit (Arabkhedri, 1989; Kalteh, 2002). The Alborz mountains protect the southern parts of Iran from the influence of the Caspian Sea and create special climatic conditions to the study region, as compared to the remaining part of the country. The precipitation along the coastal plain is abundant and approximately homogeneous, with some decrease of annual totals in the plains from west to east. Although most researchers may picture Iran as a desert with mounds of sand, arid climate and no water, northern Iran near the Caspian Sea, has thick forests, green valleys, and mountains such as Alborz that extend east to west in northern Iran.

4.1.1. Relationship between precipitation and altitude in northern Iran

The stations used for this study are shown in Figure 4.1. Monthly total precipitation data were collected from the 24 rain gauge stations with 34-year records. The code of the rain gauge stations, distance from the sea (m), altitude (m) and mean annual precipitation (mm) of the rain gauge stations used for this study are shown in Table 4.1.

Table 4.1. Code of stations, Distance from the sea, Altitude and Mean annual rainfall for 24 rain gauge stations in northern Iran

Code of the stations	Distance from the sea (m)	Altitude (m)	Mean annual precipitation (mm)
13-006	3288	-10	690
13-029	11533	0	689
13-014	86318	1740	858
13-010	87690	2014	562
13-025	50694	200	764
14-001	52055	270	988
15-009	54388	2500	409
13-019	63465	400	592
15-021	3036	0	1009
14-025	5583	0	751
16-009	222	-10	1375
16-029	391	0	1498
16-025	861	0	1167
16-035	1532	0	1525
16-049	8371	110	1115
17-057	17671	-10	1192
17-082	23600	-3	1299
17-060	72222	1750	479
17-029	94332	1000	318
17-526	133768	1800	375
17-007	252143	1650	268
18-031	12958	100	1064
18-017	18787	20	974
18-007	23353	1	1096

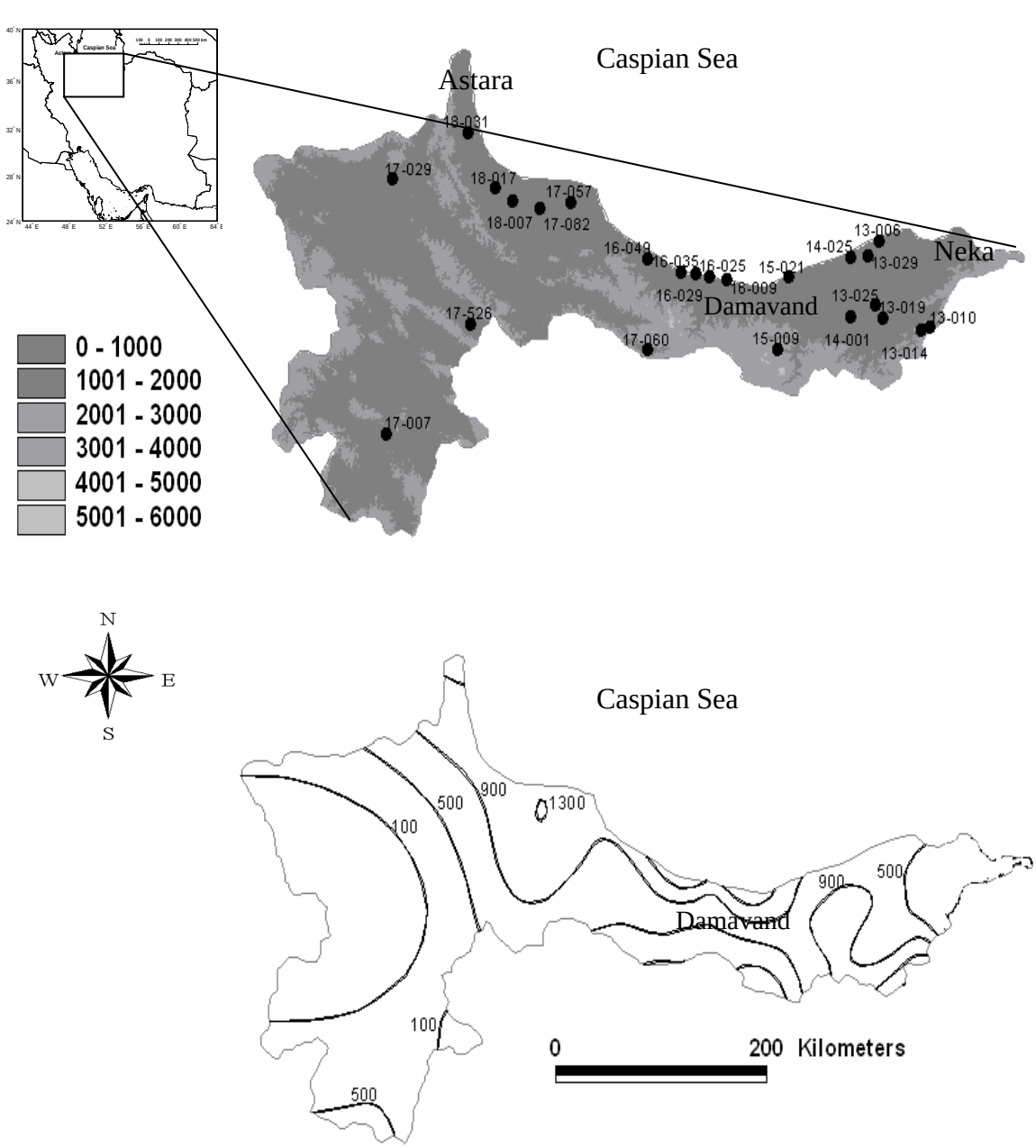


Figure 4.1. Map of Iran, location of precipitation stations in the study area along with topography (m), and mean annual precipitation isolines (mm)

The relationship between precipitation and altitude in northern Iran using the above data is shown in Figure 4.2. The equation in Figure 4.2 indicates that for each 100 meter increase of the altitude, the precipitation decreases 31 mm. This is the inverse of the law of precipitation increasing with altitude.

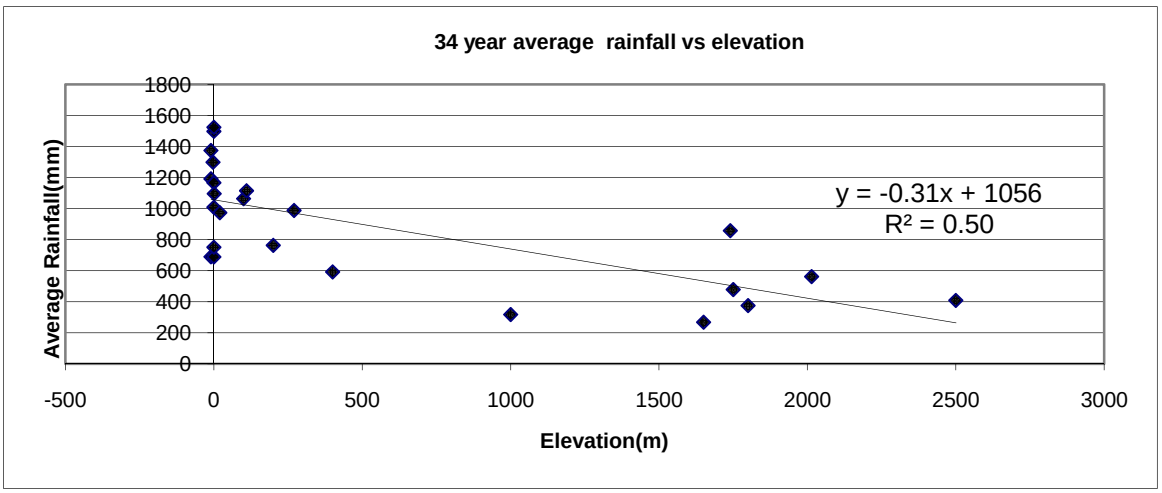


Figure 4.2. Rainfall versus altitude in northern Iran

4.1.2. The data used from Caspian Sea watershed

In ***P3*** monthly precipitation time series, follow hydrological years (October to September) and span the period from 1966-67 to 1999-2000 (34 years), were used to study the precipitation stations regionalization as well as precipitation estimation. The data were published by the Ministry of Energy of Iran. In total, 24 well distributed (not so good in southern and southwestern region) precipitation stations were used as summarized in Table 1 in ***P3*** and shown in Figure 4.1. All precipitation time series were standardized (in order to ensure that they receive equal importance) prior of use by subtracting the mean and dividing by the standard deviation. To summarize, the stations used in ***P3*** are the same as those used in studying the relationship between precipitation and altitude.

The study area in ***P1***, ***P4*** and ***P5*** is situated in the eastern part of the watershed (Neka) shown in Figure 4.1. The data used as input and the output variables have been illustrated in each paper. Some description of the rain gauge stations situated in this watershed are shown in Table 4.2. As both precipitation and runoff variables are recorded in some stations, "precipitation" and "runoff" were added within parenthesis in order to clarify.

*Table 4.2. Description of the stations situated in the study area of **P1**, **P4** and **P5***

Stations	Latitude(m)	Longitude(m)	Elevation(m)	Min	Max
13001(precipitation)	1311576.1	4092853	1392	0	81
13004(precipitation)	1275860.5	4087657.2	1038	0	96
13005(precipitation)	1297143	4085876.5	1092	0	110
13005(runoff)	1297143	4085876.5	1092	0	44.5
13007(precipitation)	1371789	4095050.7	3376	0	79
13013(precipitation)	1244301.2	4086715.7	134	0	92
13013(runoff)	1244301.2	4086715.7	134	0	1523.80

Moreover, in order to provide seasonality information to the ANN models, two time series represented by a sine and a cosine curve were introduced as auxiliary input variable. Figures 4.3 and 4.4 represent the annual cycle for daily and monthly data, respectively.

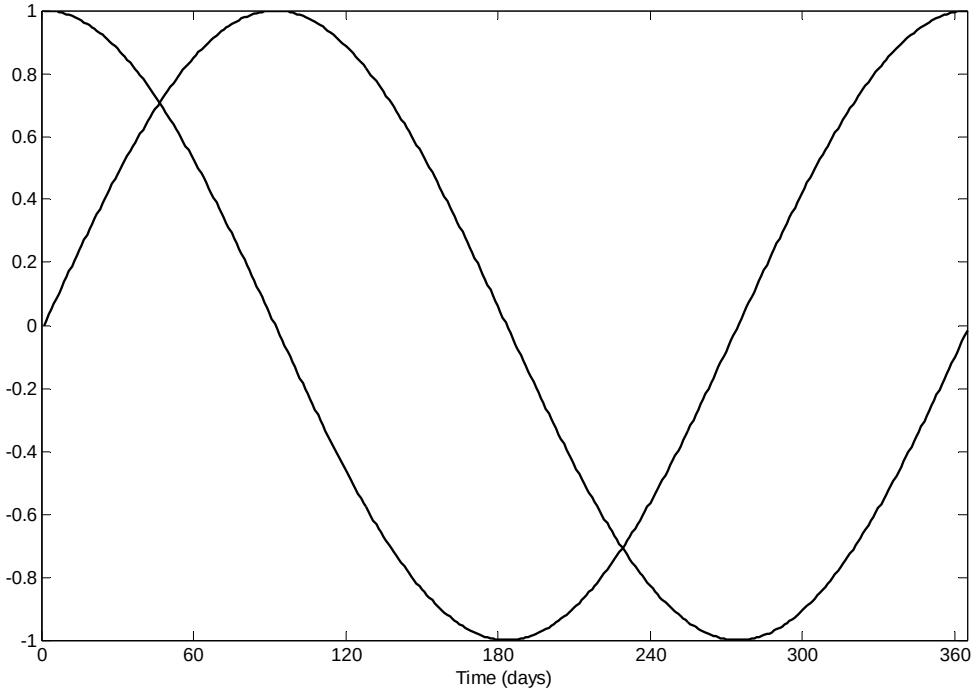


Figure 4.3. Two time series representing the annual cycle for daily data

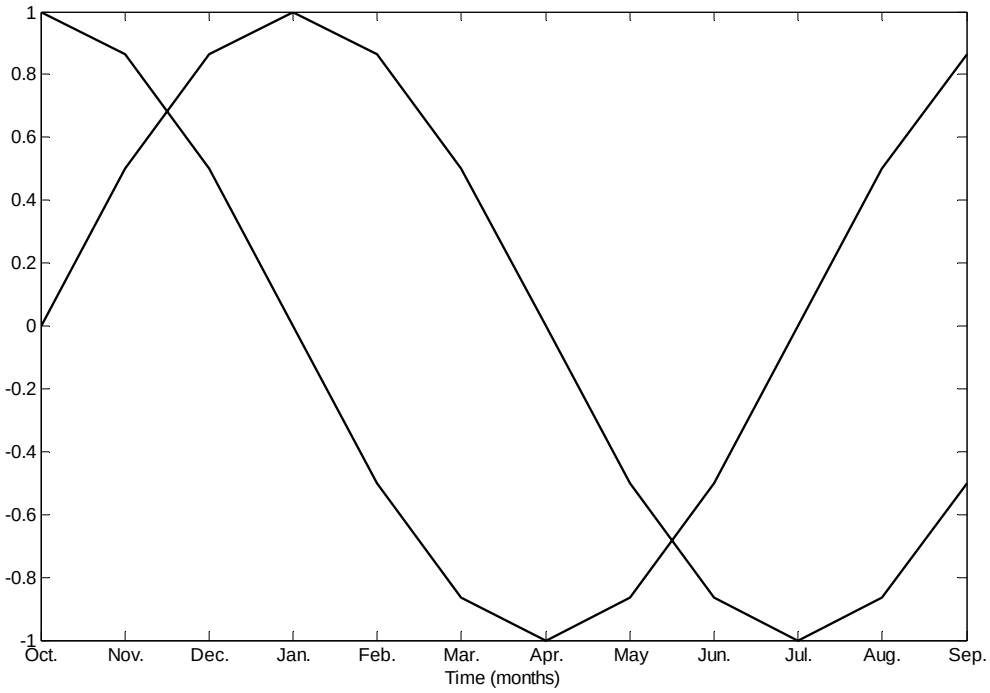


Figure 4.4. Two time series representing the annual cycle for monthly data

4.2. Canadian watershed

In ***P5***, monthly runoff values derived from a watershed situated in Canada were used in order to develop ANN models. Parasuraman *et al.* (2006) used this data in their study for runoff estimation and found good estimation results. These data are available at: http://www.wsc.ec.gc.ca/.

Chapter 5. RESULTS

5.1. Understanding of ANNs

5.1.1. Monthly runoff estimation

In ***P1*** a feed-forward MLP model was developed in order to estimate monthly runoff in a watershed in northern Iran. The scatter diagrams of observed versus estimated monthly runoff for the training and validation periods along with their r and $RMSE$ are shown in Figures 3 and 4 in ***P1***, respectively. It can be noticed that monthly runoff is reasonably well simulated by the feed-forward MLP model.

5.1.2. Understanding of the mechanisms modelled by the feed-forward MLP

As mentioned above, the model developed performs satisfactorily for estimation of monthly runoff. However, there is no idea about how the model came up with this output. In order to get some explanations concerning how the model works, three approaches i.e. the NID, Garson's algorithm and randomization approach were investigated and the results obtained are explained as follows.

Figure 5.1 represents the NID for the feed-forward MLP ANN rainfall-runoff model, in order to depict the influence of each input variable on the output. As can be noticed from the figure, this approach depicts qualitatively the contribution of each input variable via hidden neurons on the output variable. For instance, the temperature variable has positive effects on all neurons in the hidden layer except the second neuron. The outgoing signals from the fifth and sixth hidden neurons have the most positive and negative effects on the output variable, respectively. Moreover, it is possible to depict the interactions among the input variables (described before). For instance, the effect of temperature and runoff variables on the sixth neuron in the hidden layer is opposite. As mentioned earlier, this approach is qualitatively depicts the connection weights of the feed-forward MLP ANN rainfall-runoff model.

However, in order to obtain a quantitative estimate of the contribution of each input variable on the output, the Garson's algorithm was used. Figure 6 in ***P1*** shows the results obtained from Garson's algorithm. As can be noticed, the input variable contributions ranged from 4.12% to 19.73 %. The highest contribution belongs to the temperature variable measured at station 13005 and the lowest contribution belongs to the rainfall variable measured at station 13009. It does not seem to be appropriate to compare the NID and Garson's algorithm, because the former represents visually the connection weights flowing from input variables via hidden neurons to the output neuron while the latter provides the relative importance of each input variable on the output without consideration of the direction of the connection weights.

It is also interesting to get information about significant input variable (s) on the output. To do so, a randomization approach (by following the procedures mentioned before) was used and it was found that the runoff input variable was the only significant input variable on the output. Considering that the NID involves the visualization of all connection weights flowing from the input to the output layer, it seems that this visualization will be problematic when the number of input variables is many. Therefore, the randomization approach would be able to assist the modeller by reducing the number of input variables and keeping only those variables that are significant on the output variable.

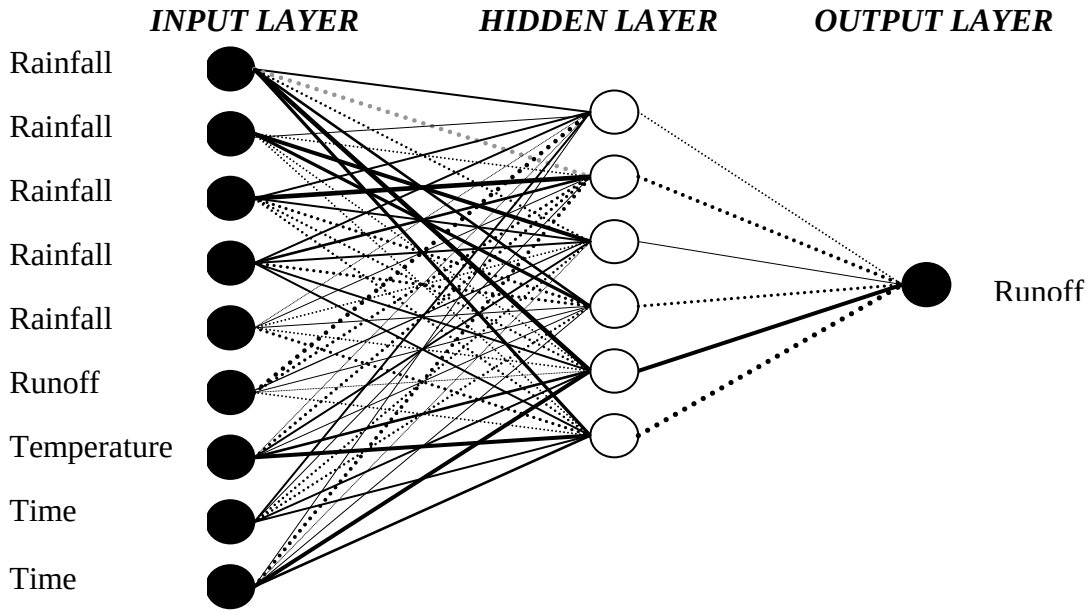


Figure 5.1. Neural Interpretation Diagram (NID) for artificial neural network (ANN) rainfall-runoff modelling. The line thickness represents the magnitude of the connection weights and the format of the lines as dotted (negative effect) and solid (positive effect) represents the direction of the connection weights.

5.2. Missing values

5.2.1. Regionalization

The application of the SOM algorithm for regionalization purpose ("input determination") was introduced in ***P3***, in which the component planes visualization technique was used in order to depict visually the correlation among precipitation stations (Figure 4 in ***P3***). Each component plane represents the values of one variable in each map neuron. Consequently, the similarity of variables may be revealed by similar patterns in identical locations on the component planes. Although the human eye has a fascinating capability to identify similarity among various objects, the paper argues that the re-organization of the component planes facilitates the interpretation. For this purpose, a SOM was used in which the location of each projected component plane corresponds to its best matching neuron [Figure 5(a) in ***P3***], so that highly similar precipitation stations are mapped near each other. The map is divided into three homogenous sub-areas, so that stations in the upper right corner of the map are located in the southern and southwestern parts of the study area, stations to the left on the map are mainly located in the eastern part of the study area and stations located in the lower right corner of the map are located along the coastal line of the study area (Figure 1 in ***P3*** or Figure 4.1). The grouping results obtained were also validated using hierarchical cluster analysis with the Ward algorithm and a Euclidean distance measure in order to cluster variables (precipitation stations).

5.2.2. Estimation of monthly precipitation

After finding homogenous groups of stations using the SOM, one representative station from each homogeneous group was selected for presentation of the estimation results based on the SOM and MLP models. To obtain objective results, the SOM and MLP networks were trained by repeatedly eliminating one-year, two-year, five-year, and finally, ten-year precipitation data from each of the stations used for estimation. Consequently, all data were used for training and all data were used for validating in a so-called jack-knifing procedure. To clarify the jack-knifing procedure; e.g. in the case of 1-year, the whole time series was divided into consecutive 1-year periods so that the 1-year periods were consecutively excluded from the training and used for validation, hence all data not removed was used for the training and the excluded part was used for validation. The same way was also conducted for 2-, 5-, and 10-year precipitation periods.

Figure 5.2 presents an example of observed and estimated monthly precipitation based on a SOM model for a representative station (13006). As can be noticed from Figure 5.2, the estimation performs well for small and intermediate monthly precipitation while for extreme values, precipitation is often underestimated. However, the paper takes into account the regionalization results by training the SOM using only homogeneous data which indeed improves the estimation results as shown in Figure 5.3. As can be noticed, the estimation obtained from the SOM model after regionalization, shows a closer tracking of the precipitation data on all portions.

A summary of the results for all excluded periods is shown in Table 2 in ***P3***. It indicates that the use of regionalized data generally improves estimation results. A notable improvement of the estimation results can be seen for station 17526. This station is part of the sparse network for the large south and southwestern regions. Here, the regionalization improves the results greatly. The SOM appears to be able to efficiently capture the overall pattern particular for this region (using regionalized data as input) and also to interpolate missing data with quality.

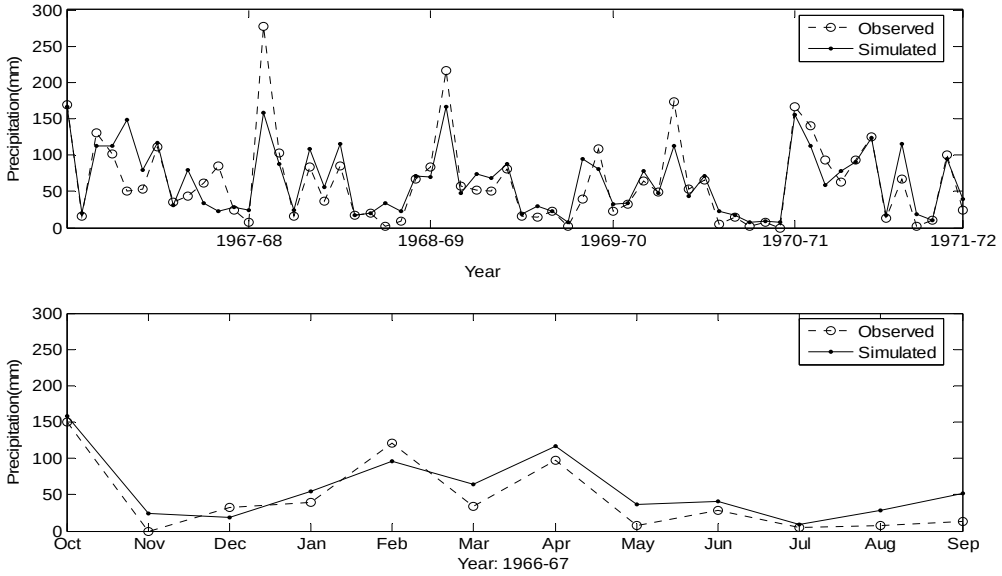


Figure 5.2. Observed and estimated validated 1-year precipitation for station 13006 without regionalization

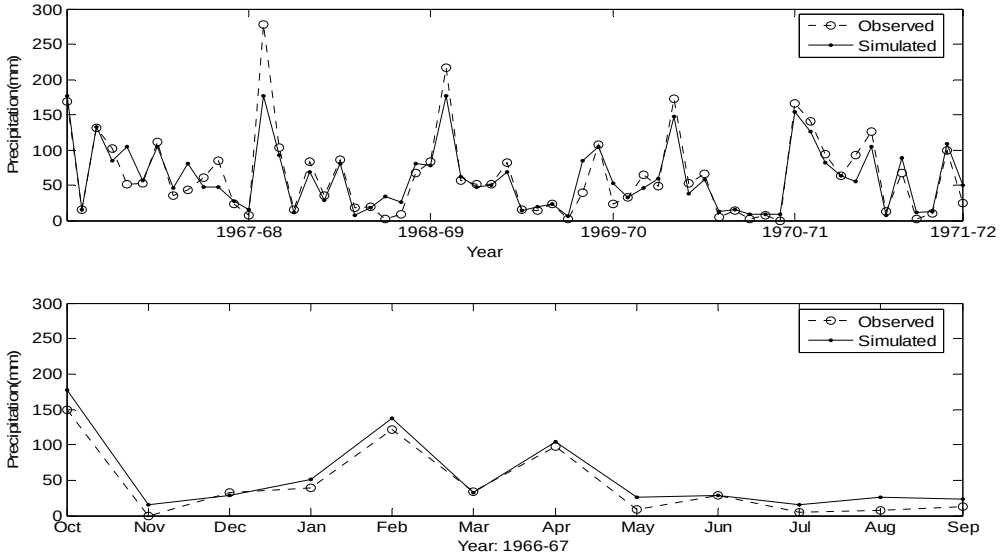


Figure 5.3. Observed and estimated validated 1-year precipitation for station 13006 with regionalization

The paper also compares the SOM estimation results with those from a well-known MLP model. A summary of results for all excluded periods for the MLP is shown in Table 3 in ***P3***. It indicates that for non-regionalized data input, the performance of the MLP is somewhat better than that achieved by the SOM. With increasing length of excluded data, however, the SOM results approach those of the MLP. For regionalized data, however, the SOM is superior for all stations and most of the excluded time periods. The MLP can not improve

estimation results much for station 17526 using regionalized data. However, as mentioned above, for this station, the SOM results were greatly improved for regionalized data input.

5.2.3. Imputation of missing values in a daily precipitation-runoff database

In ***P4***, the SOM, MLP, MNN, REGEM and the MI methods were used in order to fill in (impute) the gaps in a daily precipitation-runoff database. The standard deviation and mean of the variables under investigation before and after filling-in the gaps along with the change (%) of these parameters after filling-in the gaps were used in order to evaluate the models (see Table 3 in ***P4***).

It can be noticed from the table that all statistics were severely violated on the runoff variable of station 13013. This indicates the importance of the quality of data and the amounts of missing values in gap filling-in methods. The results obtained in ***P4*** show that multivariate methods are also dependent on the variable under investigation, since Junninen *et al.* (2004) concluded that univariate methods are dependent on the gap length as well as variable under study. Moreover, the results indicate that there is a substantial difference between two ANN methods used in ***P4***, the SOM and the MLP, in terms of handling missing value problem particularly this difference is significant with regard to the data quality of the variables under investigation. In addition, a single SOM model is able to handle all missing data patterns together hence, there is no need to develop separate models for each missing data pattern. Moreover, the SOM model provides more information about the variables under investigation via e.g. component plane visualization technique. This technique was used for regionalization ("input determination") in ***P3***, as explained before in this chapter.

One can also notice from the table that the MNN performs almost as well as other used methods which may indicate that such a simple method is very important, especially it has some advantages such as being computationally inexpensive as well as using existing values in the database to impute missing values hence it does not generate new values.

As stated previously, the most important aspect of MI is that with the statistical analyses of imputed databases and thereafter combining them one can obtain statistically valid inference about the database with missing values, hence point prediction of the missing values is not the main goal. Rather, it is to prepare the complete database for the needs of the ultimate user's in order to obtain statistically valid inference of particular importance. In ***P4***, Table 4 shows the univariate statistical inferences of 5 imputed databases. By using MI technique, we have created 5 versions from the original incomplete precipitation-runoff process database; thereafter the hydrologist can conduct the preferred analyses of each database separately and then combine the results by averaging over the 5 analyses in order to obtain a single inference from the original incomplete precipitation-runoff database which reflects the uncertainty attached to the missing data. These final single inferences, obtained by averaging of the results are shown in Table 4 in ***P4***, and MI results are shown in Table 3 in ***P4***.

The bivariate correlation coefficients between the runoff variable at station 13013 and other variables are shown in Table 5 in ***P4***. Since the station 13013 is located at the outlet of the watershed, the correlation coefficients of this variable with others were calculated. As can be noticed from the table, the correlation coefficients have decreased for all the methods, except at 13005(R) for REGEM and MI, compared to the complete-case (CC) analysis. Increasing correlation coefficients at 13005(R) by REGEM and MI can be explained by the fact that

these methods carry the assumption of multivariate normality. With reference to Table 3 in ***P4***, it can be noticed that changes of the standard deviation after filling-in the gaps for this variable with these methods are very high compared to other methods. Unsurprisingly, the MLP method provides the poorest performance compared to others which is consistent with the results obtained from the univariate statistics from the former analysis. Even though the correlation coefficients are not very high, the point that different methods, except MLP, provide approximately similar results may give confidence about the applicability of the methods in the context of precipitation-runoff process missing data filling-in. As indicated previously as well as in ***P4***, the missing data problem in our database is very complicated such that the fractions of missing values in most of the variables are great hence it would certainly affect the performance of the methods.

To summarize, it was not the main objective of the study in ***P4*** to make a detailed comparison of the methods. Rather the study aimed to illustrate the applicability of various gap filling-in methods on the real missing values in precipitation-runoff process database which is required to hydrologists for further studies such as runoff prediction.

All the methods used have advantages and disadvantages hence the objective was not to introduce a single method to fill in missing values in hydrological applications. However, in ***P4***, the SOM and MNN methods seem to be better compared to the others. Another important objective in ***P4*** was that, according to the authors' knowledge, it was the first time that the MI method for imputing missing values in hydrological application was practised.

5.3. Rainfall-runoff modelling

In ***P5***, the completed database obtained by means of SOM (in ***P4***) was used to model rainfall-runoff relationship in northern Iran. Moreover, a Canadian watershed database was also used to model runoff (see Chapter 4).

5.3.1. Decomposition of input patterns

According to the literature review in ***P5***, it seems that clustering or grouping or segmentation or decomposition of the modelling domain leads to improved modelling performance. Consequently, SOM models were used to group of input patterns prior to the development of input-output mappings. In ***P5***, three different one-dimensional SOM models were developed for both an Iranian and a Canadian watershed. They were used to explore the possibility of decomposition of the input patterns into different number of clusters as follows: the first model, with two neurons in the Kohonen layer or SOM(2), was developed in order to study the possibility of decomposition of the input patterns into two clusters, the second model, with three neurons in the Kohonen layer or SOM(3), was developed in order to study the possibility of decomposition of the input patterns into three clusters, and finally the third model, with four neurons in the Kohonen layer or SOM(4), was developed in order to study the possibility of decomposition of the input patterns into four clusters. The plots of each SOM model for both the Iranian and the Canadian watershed are shown in Figures 3(a) and (b) in ***P5***, respectively. The results obtained by grouping (decomposition) of input patterns using SOM for both watersheds along with their interpretation [Figures 3(a) and (b) in ***P5***] are summarized in the following paragraphs.

The results of the clustering for the Iranian watershed was that the SOM (2) clustered the training input patterns (240 patterns) into two clusters consisting of 119 and 121 patterns, respectively. The first cluster corresponds to lower mean values compared to the second cluster hence the former will probably be associated with low runoff values and the latter with high runoff values. To examine, the mean and standard deviation of the corresponding output patterns for the first were calculated and found to be 4.41 and 5.76 respectively. The corresponding values for the second cluster were found to be 6.44 and 3.62. It was found that the first and second cluster correspond to low and high runoff values, respectively. The SOM (3) clustered the training input patterns into three clusters consisting of 99, 41, and 100 patterns, respectively. The first cluster corresponds to lower mean values and the third cluster to higher mean values while the second lies in between. To examine, the mean and standard deviation of the corresponding output patterns for the first, second, and third cluster were calculated and 2.90, 4.70; 7.74, 5.55; and 7.00, 3.54; were obtained, respectively. In terms of these statistics the first, second and third cluster correspond to low, high and medium runoff values, respectively. One can notice that these results are physically unusual as the third cluster with higher mean values exhibits medium runoff magnitude while the second cluster which lay between the other two clusters exhibits high runoff magnitude. Finally, the SOM (4) clustered the training input patterns into four clusters consisting of 79, 41, 40, and 80 patterns, respectively. The first cluster corresponds to lower mean values and the fourth cluster corresponds to higher mean values while the second and third clusters lie in between. To examine, the mean and standard deviation of the corresponding output patterns for the first, second, third, and fourth cluster were calculated and 3.05, 5.20; 3.08, 2.33; 5.41, 2.74; and 9.02, 4.23; were obtained, respectively. The first, second, third and fourth cluster correspond to low, medium, medium and high runoff values, respectively which is physically plausible. It may be mentioned that for the above SOM clusters, judgment of the different runoff magnitudes were based on their mean values. In order to make a comparison with the descriptive statistics (the mean and standard deviation) above, it must be mentioned that these statistics for the training output patterns without such a clustering or decomposition are 5.44, and 4.90, respectively.

In ***P5*** the same numbers of SOM models were developed for the Canadian watershed, in order to explore the possibility of decomposition of the input patterns into different number of clusters from two to four segments or clusters and to compare the results obtained with the Iranian watershed. The results of this clustering can be summarized as follows: the SOM (2) clustered the training input patterns (240 patterns) into two clusters consisting of 119 and 121 patterns, respectively. The first cluster corresponds to lower mean values compared to the second cluster hence the former will probably be associated with low runoff values while the latter with high runoff values. To check, the mean and standard deviation of the corresponding output patterns for the first and second cluster were calculated and 93.82, 55.40; and 177.10, 98.56; were obtained, respectively. It was seen that the first and second cluster correspond to low and high runoff values, respectively. The SOM (3) clustered the training input patterns into three clusters consisting of 112, 26, and 102 patterns, respectively. The first cluster corresponds to lower mean values and the third cluster to higher mean values while the second lies in between. To validate, the mean and standard deviation of the corresponding output patterns for the first, second, and third cluster were calculated and 91.11, 55.89; 117.82, 55.85; and 189.47, 99.17; were obtained, respectively. The first, second and third cluster correspond to low, medium and high runoff values, respectively which is physically plausible. Finally, the SOM (4) clustered the training input patterns into four clusters consisting of 80, 39, 40, and 81 patterns, respectively. The first

cluster corresponds to lower mean values and the fourth cluster corresponds to higher mean values while the second and third clusters lie in between. To check, the mean and standard deviation of the corresponding output patterns for the first, second, third, and fourth cluster were calculated and 96.98, 63.59; 87.32, 32.56; 116.17, 61.25; and 207.19, 99.84; were obtained, respectively. The first, second, third and fourth cluster correspond to medium, low, medium and high runoff values, respectively. The results for the third and fourth cluster are physically plausible but the results for cluster one and two are not. However, it must be mentioned that these statistics for the training output patterns without such a clustering or decomposition are 135.81, and 90.19, respectively. It is worthy to mention that although it was possible to characterize the SOM results somehow, it is not always easy to interpret the results due to the fact that ANNs, and among them SOMs, are black-box models.

5.3.2. Monthly runoff forecasting

Once the clustering of the input patterns was achieved using SOM, the next step was mapping of the input patterns in each cluster or group to the corresponding outputs, corresponding to different physics in the watersheds, using feed-forward MLP and SSOM models. In other words, separate feed-forward MLP and SSOM models were developed in order to map input patterns of each cluster or group to the corresponding outputs. The results obtained by doing so, for both the Iranian and the Canadian watershed are summarized in the following paragraphs.

The results of the various MLP and SSOM models, based on the model performance criteria in Chapter 3, in forecasting monthly runoff for the Iranian and Canadian watersheds are presented in Tables 5.1 and 5.2, respectively. It can be noticed from Table 5.1 that in the case of MLP models, the best performance was achieved by the two and four cluster MLP models with the two cluster models doing somewhat better. It is worthy to mention that although the four cluster MLP models performed better during training compared to the two cluster MLP models, the four cluster MLP models were not able to generalize better during validation which may indicate the importance of more principled procedures for determination of the number of neurons for decomposition. However, all of the modular MLP models outperform the single MLP model during validation which indicates that the process under investigation is not homogenous hence the importance of decomposition of the process. In the case of SSOM models for the Iranian watershed, Table 5.1, the best performance was achieved by the four cluster SSOM models. It is worthy to mention that although the single SSOM model performed slightly better than the four cluster SSOM models during training the four cluster SSOM models outperformed it during validation hence again indicating the importance of decomposition of the process. As can be noticed from Table 5.1, the performance of the SSOM models during training was better than that for the MLP models both before decomposition and after decomposition while after decomposition during validation period the MLP models outperform the SSOM models. However, it can also be noticed from Table 5.1 that the performance of the single SSOM model is better than that of the single MLP model for both the training and the validation period. Considering these results, in the selection of best model between the single MLP and single SSOM model the latter was selected while in terms of decomposition results the two cluster MLP models is considered to be the best model for forecasting monthly runoff for the Iranian watershed. The performance of these selected models during validation period is shown in Figure 5.4(a) and (b), respectively.

In the case of the Canadian watershed (Table 5.2) the models for forecasting monthly runoff performed differently compared to the Iranian watershed results. For instance, although the decomposition procedure improved the modelling performance during training for MLP models, it could not improve the generalization ability of the models as the performance of the single MLP model is slightly better than that of the two cluster MLP model and substantially better than those of the three and four cluster MLP models. However, in the case of the SSOM models, the decomposition procedure improved the performance of the models so that there are improvements in forecasting performance by using the two, three, and four cluster SSOM models compared to the single SSOM model. One can notice from Table 5.2 that the performance of the SSOM models is mostly better than those of MLP models after decomposition for both training and validation. However, the performance of the single MLP and single SSOM models is almost similar with a small advantage for the single MLP. By considering these results, the single MLP and four cluster SSOM models are considered to be the best models for forecasting monthly runoff in the Canadian watershed. The performance of these models during validation is shown in Figure 5.5(a) and (b), respectively.

Table 5.1. The performance of models in forecasting monthly runoff for the Iranian watershed

	Training			Validation		
Model	r	R^2	RMSE	r	R^2	RMSE
MLP models						
Single MLP	0.652	0.426	3.706	0.383	0.132	6.469
Two MLPs	0.630	0.395	3.803	0.441	0.193	6.237
Three MLPs	0.642	0.412	3.748	0.401	0.160	6.362
Four MLPs	0.671	0.450	3.627	0.447	0.192	6.241
SSOM models						
Single SSOM	0.733	0.534	3.337	0.403	0.159	6.366
Two SSOMs	0.723	0.516	3.402	0.378	0.138	6.448
Three SSOMs	0.709	0.499	3.460	0.394	0.152	6.395
Four SSOMs	0.722	0.520	3.387	0.406	0.163	6.353

Table 5.2. The performance of models in forecasting monthly runoff for the Canadian watershed

	Training			Validation		
Model	r	R^2	RMSE	r	R^2	RMSE
MLP models						
Single MLP	0.850	0.718	47.773	0.800	0.615	47.658
Two MLPs	0.864	0.743	45.553	0.805	0.612	47.886
Three MLPs	0.867	0.752	44.776	0.797	0.583	49.639
Four MLPs	0.873	0.758	44.208	0.772	0.559	51.033
SSOM models						
Single SSOM	0.909	0.794	37.696	0.824	0.604	48.363
Two SSOMs	0.904	0.815	38.693	0.799	0.605	48.298
Three SSOMs	0.914	0.833	36.678	0.808	0.631	46.654
Four SSOMs	0.919	0.842	35.748	0.815	0.635	46.445

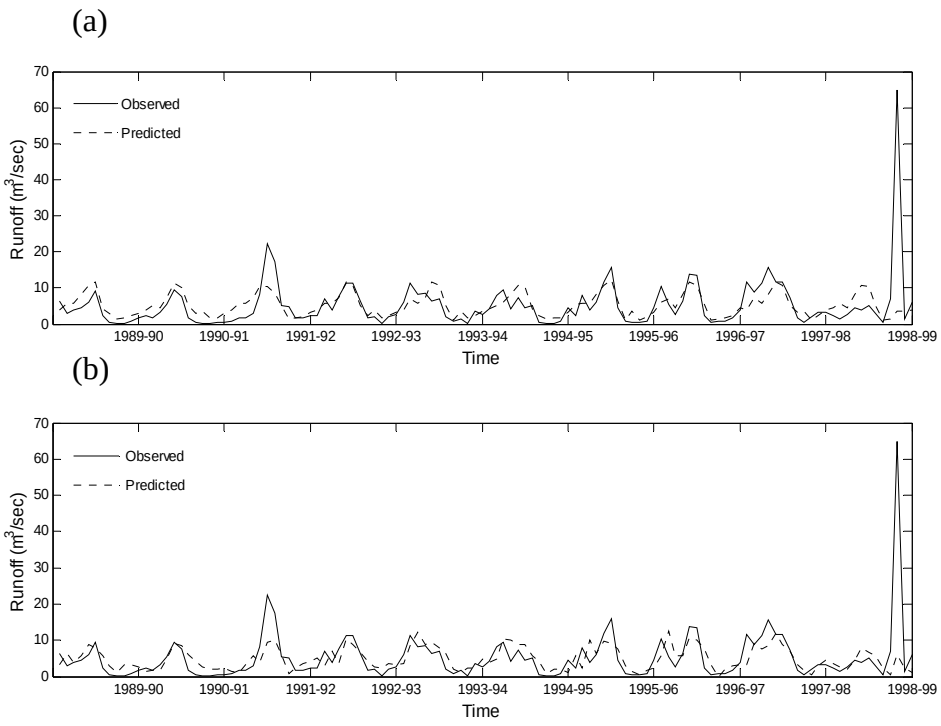


Figure 5.4. Observed and predicted runoff for the Iranian watershed from (a) single cluster SSOM model, and (b) two cluster MLP model

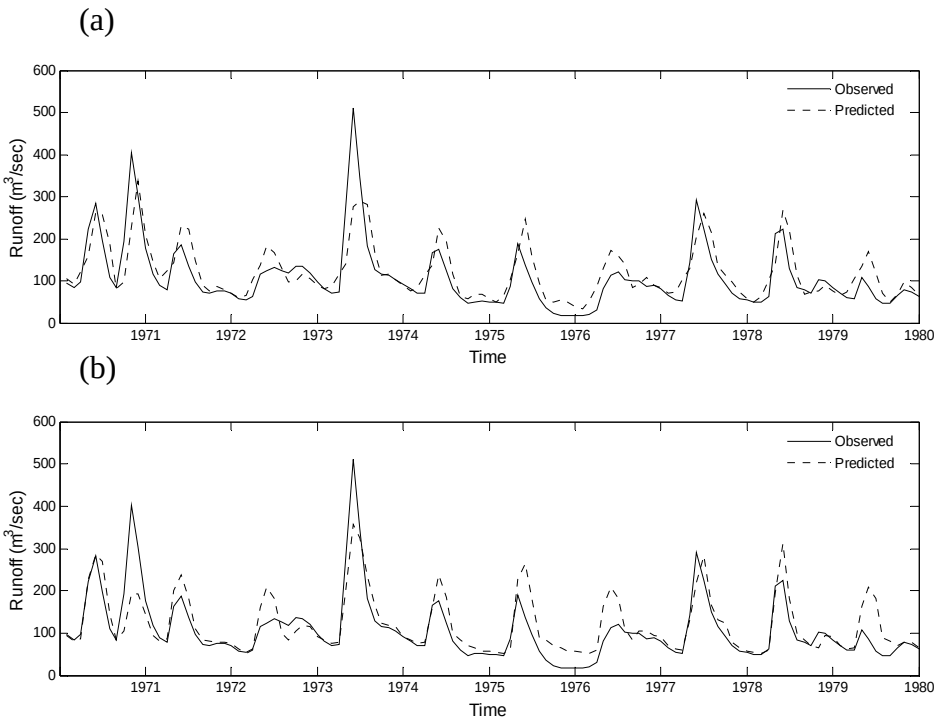


Figure 5.5. Observed and predicted runoff for the Canadian watershed from (a) single cluster MLP model, and (b) four cluster SSOM model

Chapter 6. CONCLUSIONS

Based on ***P1***, it can be concluded that ANN models, which are normally considered as black-box models, are powerful not only in modelling of hydrological processes reasonably well but they are also able to reveal some information concerning how they end up with this output. Therefore, this property would help them in obtaining even wider acceptability among hydrologists who are usually interested in gaining information from the model.

P2 indicated that applications of SOM algorithm in analysis, estimation and prediction of hydrological processes have increased tremendously over the last decade. This is due to the ability of the SOM in clustering large amounts of data as well as to preserve their topological structure. Therefore, the SOM algorithm can be considered as an alternative approach to well-known feed-forward MLP.

The successful application of SOM and feed-forward MLP ANN models in estimation of missing values in the context of monthly precipitation and daily precipitation-runoff databases in ***P3*** and ***P4***, respectively, indicated their suitability in solving such problems. However, the SOM and the feed-forward MLP differ in handling missing values or in an estimation problem as below:

The SOM works best for homogenous data (after regionalization with SOM), thus, it is important to test the data for homogeneity and if necessary perform a regionalization prior to the use of SOM.

For non-regionalized data (before applying SOM for the regionalization) the MLP appears to give slightly better results. However, for homogeneous data, the SOM clearly gives the best results.

Although the MLP performs better for the non-homogenous data, the SOM can be used to homogenize the data prior to application and thus has the overall advantage.

As mentioned in ***P4***, the MLP needs to train separate models for different combinations of missing variables which may lead to incoherence between filled-in values while a single SOM model is able to handle missing values regardless of their location.

Moreover, our study in ***P4*** indicated that there is a substantial difference between the SOM and MLP in filling-in the gaps, particularly for a variable that contains extreme values.

In addition to the above, the SOM has the advantage of providing information via e.g. component plane visualization technique from the trained network parameters. For instance, in ***P3*** the technique was used for regionalization (“input determination”) of the precipitation stations.

However, the comparison of the above ANN models with those of MNN, REGEM and MI methods in estimating missing values in a daily precipitation-runoff database reveal that the results obtained from five methods, i.e. the SOM, MLP, MNN, REGEM and MI, show that most of the methods yield similar results. Interestingly, the MNN which is an easy method for implementation and computationally cheap, performs as well as the other methods used in ***P4*** for filling-in the gaps even for the variable of which contains extreme values. It uses

existing values in the database in order to fill in missing values. Consequently, the MNN does not generate new values.

Although most studies tend to result in the conclusion that there is no method that is superior in all applications, there seems to be an agreement that one should use statistically principled methods. This speaks in favour of REGEM and MI. MI has the great advantage of giving an estimate of the uncertainties involved in the imputation. However this comes at a cost. Both REGEM and MI build on the assumption that the data are multivariate normal. The results indicate that there may significant bias if this assumption is violated. Thus it seems safer to rely on either SOM or MNN.

Although most studies conclude that there is no imputation methodology that can be characterized as superior to others, it can be suggested that for a rainfall runoff database, the SOM and MNN provide the most robust and reliable results.

The results obtained from the investigation of whether the modelling performance concerning forecasting monthly runoff can be improved by means of decomposition of ANN models, where the decomposition of the process under investigation was achieved by means of SOM algorithm, indicated improvements, but they were rather small.

Although the idea that a modular model would be better at responding to the different physical processes at different stages of the hydrograph is highly impressive, the results obtained indicated that there is a need for more principled procedures in order to decompose the models. As the results obtained by using different number of segments varied, it indicates the importance of the number of neurons (segments) which are usually selected in the SOM algorithm via a trial-and-error procedure. It was shown that the division of the process under investigation into further divisions may not produce better performance but the optimum number of divisions is of prime importance in order to model the process successfully. Therefore, finding an appropriate approach to dividing the processes under investigation into an optimum number of segments is a general need in this regard.

Concerning the comparison between the performance of feed-forward MLP and SSOM models on forecasting monthly runoff, the results indicated that the SSOMs consistently performed better during training while during validation there were small differences in performance which may indicate the application of the SSOM model as an alternative to the feed-forward MLP model. However, it may be mentioned that the SSOM and generally SOM model produce the output results based on more network parameters than the feed-forward MLP. Both SOM and feed-forward MLP model include trial-and-error procedures. It is worthy to mention that the visualization ability of SOM is a fascinating feature that helps the modeller to further investigate the processes.

REFERENCES

Allison, P.D., 1998. *Multiple imputation for missing data: A cautionary tale.* Retrieved 15 July 2006 (available at: http://www.ssc.upenn.edu/~allison/).

Arabkhedri, M., 1989. *Investigation of maximum floods in northern Alborz watersheds.* MSc Thesis. Tehran University, Tehran, Iran (in Persian).

ASCE Task Committee on application of Artificial Neural Networks in Hydrology. 2000a. **Artificial neural networks in hydrology. I: preliminary concepts.** Journal of Hydrologic Engineering 5(2) 115-123.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology. 2000b. **Artificial neural networks in hydrology. II: hydrologic applications.** Journal of Hydrologic Engineering 5(2) 124-137.

Corne, S., Murray, T., Openshaw, S., See, L., Turton, I., 1999. **Using computational intelligence techniques to model subglacial water systems.** Journal of Geographical Systems 1, 37-60.

Dawson, C.W., Wilby, R.L., 1998. **An artificial neural network approach to rainfall-runoff modeling.** Hydrological Sciences Journal 43(1), 47-65.

Dawson, C.W., Wilby, R.L., 2001. **Hydrological modelling using artificial neural networks.** Progress in Physical Geography 25(1) 80-108.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. **Maximum likelihood from incomplete data via the EM algorithm.** Journal of the Royal Statistical Society 39(1), 1-38.

Dixon, J.K., 1979. **Pattern recognition with partly missing data.** IEEE Transactions on Systems, Man, & Cybernetics 10(SMC-9), 617-621.

Frane, J.W., 1976. **Some simple procedures for handling missing data in multivariate analysis.** Psychometrika 41(3), 409-415.

Garson, G.D., 1991. **Interpreting neural-network connection weights.** Artificial Intelligence Expert 6, 47-51.

Giustolisi, O., Laucelli, D., 2005. **Improving generalization of artificial neural networks in rainfall-runoff modelling.** Hydrological Sciences Journal 50(3), 439-457.

Haykin, S., 1999. *Neural networks: a comprehensive foundation.* Prentice Hall Upper Saddle River, New Jersey, USA.

Hsu, K.L., Gupta, H.V., Gao, X., Sorooshian, S., Imam, B., 2002. **Self-organizing linear output map (SOLO): an artificial neural network suitable for hydrologic modeling and analysis.** Water Resources Research 38(12), 1302, doi: 10.1029/2001WR000795.

Hsu, K.L., Gupta, H.V., Sorooshian, S., 1995. **Artificial neural network modeling of the rainfall-runoff process.** Water Resources Research 31(10), 2517-2530.

Jain, A., Srinivasulu, S., 2006. **Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques.** Journal of Hydrology 317, 291-306.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. **Methods for imputation of missing values in air quality data sets.** Atmospheric Environment 38, 2895-2907.

Kalteh, A.M., 2002. *Investigation of effective factors on sediment yield in Namak Lake and Caspian Sea watersheds.* MSc Thesis. Tehran University, Tehran, Iran (in Persian).

Kalteh, A.M., Berndtsson, R., 2007. **Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP).** Hydrological Sciences Journal 52(2), 305-317.

Kalteh, A.M., Hjorth, P., Berndtsson, R., 2007. **Review of self-organizing map (SOM) in water resources: analysis, modelling, and application.** Environmental Modelling & Software (submitted).

Kohonen, T., 1982a. **Analysis of a simple self-organizing process.** Biological Cybernetics 44, 135-140.

Kohonen, T., 1982b. **Self-organized formation of topologically correct feature maps.** Biological Cybernetics 43, 59-69.

Kohonen, T., 2001. *Self-Organizing Maps.* Springer-Verlag, Berlin.

Little, R.J.A., Rubin, D.B., 2002. *Statistical analysis with missing data.* Wiley, New York.

Lorrai, M., Sechi, G.M., 1995. **Neural nets for modelling rainfall-runoff transformations.** Water Resources Management 9, 299-313.

Maier, H.R., Dandy, G.C., 2000. **Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications.** Environmental Modelling & Software 15, 101-124.

Minns, A.W., Hall, M.J., 1996. **Artificial neural networks as rainfall runoff models.** Hydrological Sciences Journal 41(3), 399-417.

Murao, H., Nishikawa, I., Kitamura, S., Yamada, M., Xie, P., 1993. **A hybrid neural network system for the rainfall estimation using satellite imagery.** Proceedings of International Joint Conference on Neural Networks, IEEE press, 1211-1214.

Olden, J.D., Jackson, D.A., 2002. **Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks.** Ecological Modelling 154, 135-150.

Özesmi, S.L., Özesmi, U., 1999. **An artificial neural network approach to spatial habitat modelling with interspecific interaction.** Ecological Modelling 116, 15-31.

Parasuraman, K., Elshorbagy, A., Carey, S.K., 2006. **Spiking modular neural networks: a neural network modeling approach for hydrological processes.** Water Resources Research 42(5), W05412, doi: 10.1029/2005WR004317.

Patrician, P.A., 2002. **Multiple imputation for missing data.** Research in Nursing & Health 25, 76-84.

Pigott, T.D., 2001. **A review of methods for missing data.** Educational Research & Evaluation 7(4), 353-383.

Rajurkar, M.P., Kothyari, U.C., Chaube, U.C., 2002. **Artificial neural networks for daily rainfall-runoff modelling.** Hydrological Sciences Journal 47(6), 865-877.

Ramirez, M.C.V., Velho, H.F.D.C., Ferreira, N.J., 2005. **Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region.** Journal of Hydrology 301, 146-162.

Rubin, D.B., 1996. **Multiple imputation after 18+ years.** Journal of the American Statistical Association 91, 473-489.

Schafer, J.L., 1999. *NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2.* Software for Windows 95/98/NT, available at: http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer, J.L., Graham, J.W., 2002. **Missing data: our view of the state of the art.** Psychological Methods 7(2), 147-177.

Schneider, T., 2001. **Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values.** Journal of Climate 14, 853-871.

Tokar, A.S., Johnson, P.A., 1999. **Rainfall-runoff modeling using artificial neural networks.** Journal of Hydrologic Engineering 4(3), 232-239.

Tsikriktsis, N., 2005. **A review of techniques for treating missing data in OM survey research.** Journal of Operations Management 24, 53-62.

Vesanto, J., 1999. **SOM-based data visualization methods.** Intelligent Data Analysis 3,111-126.

Vesanto, J., 2002. *Data exploration process based on the self-organizing map.* Ph.D. Thesis, available at: http://lib.tkk.fi/Diss/2002/isbn9512258978/isbn9512258978.pdf.

Wilby, R.L., Abrahart, R.J., Dawson, C.W., 2003. **Detection of conceptual model rainfall-runoff processes inside an artificial neural network.** Hydrological Sciences Journal 48(2), 163-181.

Zoppou, C., Roberts, S., Hegland, M., 2000. **Spatial and temporal rainfall approximation using additive models.** ANZIAM Journal 42(E), C1599-C1611.

Appendix. PAPERS

The following papers are included in the thesis:

I. Kalteh, Aman Mohammad; (2007). **Rainfall-runoff modelling using artificial neural networks (ANNs): modelling and understanding.** (manuscript).

II. Kalteh, Aman Mohammad; Hjorth, Peder; Berndtsson, Ronny; (2007). **Review of self-organizing map (SOM) in water resources: analysis, modelling, and application.** *Environmental Modelling & Software* (submitted)*.

III. Kalteh, Aman Mohammad; Berndtsson, Ronny; (2007). **Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP).** *Hydrological Sciences Journal*, 52(2), 305-317.

IV. Kalteh, Aman Mohammad; Hjorth, Peder; (2007). **Imputation of missing values in a precipitation-runoff process database.** *Nordic Hydrology* (submitted)**.

V. Kalteh, Aman Mohammad; Hjorth, Peder; (2007). **Monthly runoff forecasting by means of artificial neural networks (ANNs).** *Hydrological Sciences Journal* (submitted).