An Exploration of Neural Networks for Real-time Flood Forecasting

Giulia Napolitano

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Geography

September, 2011

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapters 1, 3, 4 and 6 of the thesis contain information based on the following jointly-authored publications, while chapter 7 (conclusions) reflect findings in these papers:

Napolitano, G., See, L.M., Calvo, B., Savi, F. and Heppenstall, A.J. (2009). A conceptual and neural network model for real-time flood forecasting of the Tiber River in Rome. *Physics and Chemistry of the Earth*, 35(3-5), 187-194.

Napolitano, G., Serinaldi, F. and See, L. (2011). Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: an empirical examination. *Journal of Hydrology*, 406(3-4), 199-214.

For the publication by Napolitano et al. (2009), the paper describes a comparison of a neural network rainfall-runoff model with a conceptual hydrological model for the city of Rome. The neural network model was developed by the candidate. The analysis comparing the neural network model with the conceptual model was undertaken by the candidate. This is reported in Chapter 4 with some background information from the paper provided in Chapter 1. Calvo and Savi ran the conceptual model. This is clearly described in the paper and in Chapter 4 but the work is clearly attributed to Calvo and Savi (which they subsequently published in 2009). See assisted in editing the paper. Heppenstall contributed nothing to this paper and never saw the paper before it was published. She obtained a grant to work with La Sapienza and then went on maternity leave. However, her name reflects her funding contributions to See and Napolitano to visit La Sapienza on a number of occasions to complete this work.

For the publication by Napolitano et al. (2011), three main issues were tackled: application of Empirical Mode Decomposition; investigation of performance measures for evaluating models; and an investigation into the effect of the random initialisation of weights of a neural network. The candidate carried out the neural network analysis including application of the decomposition method, analysis of the random initialization of the weights and examination of the performance measures, which form the basis of Chapters 3 and 6. Serinaldi provided valuable advice on the decomposition method and uncertainty; and how to apply to these to the problem. He also provided guidance on which performance measures to select. Serinaldi and See assisted in editing the paper.

Acknowledgements

It has only been with the help of many people that I have been able to complete my PhD research. To all these people, I am deeply grateful.

First of all, I would like to thank Fabrizio Savi for all he has done for me, for teaching me that in life we must not ever give up. Also, for allowing me to meet my supervisor, Linda See, an exceptional person, both humanely and professionally. Thanks Linda, I would not have reached this target without you.

Many thanks to Francesco Serinaldi for having spent time helping and advising me. You are a true friend!

When I arrived in England, I made some really nice friends who helped me with the English language and to pass the IELTS test, fundamental for getting the PhD. So many thanks go to Rahul, Kate, Olga, Emma, Jacqui, Alison, Dianna, Kirk, Adam, John, Dan and Pia. Then there are all my Italian friends who supported me day after day and encouraged me to move forward: Paolo, Patty, Maura, Francesca, Mariella and many more...thank you all!

Finally, my biggest thanks goes to Aldo, who for 4 years has accepted a life divided between Italy and England, and who waited patiently for me when I had to work during the weekends. I will be eternally grateful for this.

"...good!"

Abstract

This thesis examines Artificial Neural Networks (ANNs) for rainfall-runoff modelling. A simple ANN was first developed to predict floods in the city of Rome, located in the Tiber River basin. A rigorous comparison of the ensemble ANN and the conceptual TEVERE model were undertaken for two recent flood events in 2005 and 2008. Both models performed well but the conceptual model was better at overall hydrograph prediction while the ANN performed better for the initial part of the event at longer lead times.

Further experimentation with the ANN model was then undertaken to try to improve the model performance. Additional upstream stations and rainfall inputs were added including hourly totals, effective rainfall and cumulative rainfall. Different methods of normalisation and different ANN training algorithms were also implemented along with four alternative methods for combining the ensemble ANN predictions. The results showed that the ANN was able to extrapolate to the 2008 event.

Finally, Empirical Mode Decomposition was applied to the ANN to examine whether this method has value for ANN rainfall-runoff modelling. At the same time the impact of the random initialisation of the weights of the ANN was investigated for the Potomac River and Clark Fork River catchments in the USA. The EMD was shown to be a valuable tool in detecting signal properties but application to ANN rainfall-runoff modelling was dependent on the nature of the dataset. Overall uncertainty from the random initialisation of weights varied by catchment where uncertainties were shown to be very large at high stream flows.

Finally, a suite of redundant and non-redundant model performance measures were applied consistently to all models. The value of applying a range of redundant and non-redundant measures, as well as benchmark-based methods was demonstrated.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANFIS	Adaptive Neural Fuzzy Inference System
ANN	Artificial Neural Network
ARMA	Autoregressive Moving Average
ASCE	American Society of Civil Engineers
BP	Backpropagation
BR	Bayesian Regularisation
CANN	Cluster-based Artificial Neural Network
CE	Coefficient of Efficiency
EMD	Empirical Mode Decomposition
GA	Genetic Algorithm
GRI	Geometric Reliability Index
LM	Levenberg-Marquardt
IPCC	International Panel on Climate Change
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MdAPE	Median Absolute Percentage Error
MLP	Multi-layer Perceptron
MNN	Modular Neural Network
PANN	Periodic Neural Network
PDIFF	Peak Difference
PI	Persistence Index
PI.MAE	Persistence Index based on MAE
PI.MdAE	Persistence Index based on MdAE
QR	Quantile Regression
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
RMSE	Root Mean Squared Error
RMSPE	Root Mean Squared Percentage Error
RNN	Recurrent Neural Network
SOLO	Self-Organising Linear Output
SOM	Self Organising Map
TANN	Threshold-based Neural Network
TBP-NN	Time Backpropagation Neural Network
TFF	Tevere Flood Forecasting Model

Chapter 1

Introduction

1.1 Background and Context to the Research

There have been many large flood events that have affected a number of different areas around the world, e.g. Central America in 1998; China in 2002 and 2004; Central Europe in 2006; East Africa in 2006; eastern Australia in 2008; Pakistan 2010 (and again very recently this year); Brazil, China, Philippines, Nigeria and north eastern USA in 2011 (Scaruffi, 2011; Wikipedia, 2011). Figure 1.1 provides a compilation of statistics illustrating the magnitude of human and economic losses from events that have occurred between 1980 and 2008. The top 10 flooding disasters in terms of human lives affected and the overall costs are also reported, where it is evident that China has been heavily impacted. In the developed world, Italy, Germany and USA also appear in the top ten list for economic damages incurred.

The effect of climate change in relation to flooding is still unknown but it is thought that the frequency and intensity of extreme events will most likely increase, where extreme precipitation events have already been observed (IPCC, 2007). An increased flood risk in some catchments is also thought to be a likely scenario due to these more intense precipitation events that are predicted to occur. Flood risk is comprised of the flood hazard, the exposure to the hazard and the vulnerability of those exposed (Plate, 2002; Kron, 2003). It is possible to develop strategies that mitigate the risk using both structural and/or non-structural flood protection measures, which are intended to reduce flood frequencies. Structural flood protection measures (e.g. dams, drainage channels, etc.) can be very effective but are not always possible in areas of dense population, e.g. a city centre. For this reason, it is crucial to have timely and accurate flood warnings in order for operational measures to be put into action and to minimise risk to human life and infrastructure damage. Civil protection agencies usually employ a flood warning system that is based on expert knowledge and a physical, conceptual or empirical hydrological model (or a combination of these approaches). Over the last two decades, empirical approaches of a more data-driven or machine learning nature have been reported in the rainfall-runoff literature. These data-driven methods are often comprised of one or more tools from the field of Artificial or Computational Intelligence. Examples include Artificial Neural Networks (ANNs), fuzzy logic, support vector machines and M5 model trees as well hybrids or combinations of these approaches.

Flood - Data and statistics

Flood disasters from 1980 - 2008

Overview

No of events:	2,887
No of people killed:	195,843
Average people killed per year:	6,753
No of people affected:	2,809,481,489
Average people affected per year:	96,878,672
Ecomomic Damage (US$ X 1,000):	397,333,885
Ecomomic Damage per year (US$ X 1,000):	13,701,168

Top ten disasters reported

Affected people

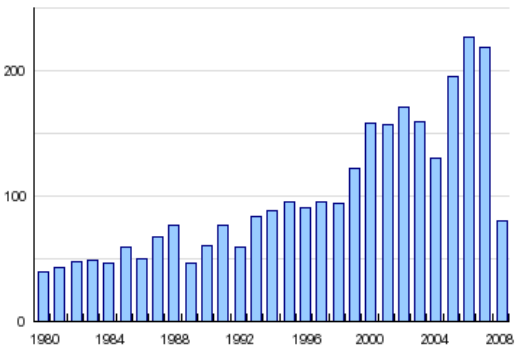
Disaster	Date	Affected (no. of people)
China P Rep	1998	238,973,000
China P Rep	1991	210,232,227
China P Rep	1996	154,634,000
China P Rep	2003	150,146,000
India	1993	128,000,000
China P Rep	1995	114,470,249
China P Rep	2007	105,004,000
China P Rep	1999	101,024,000
China P Rep	1989	100,010,000
China P Rep	2002	80,035,257

Economic damages

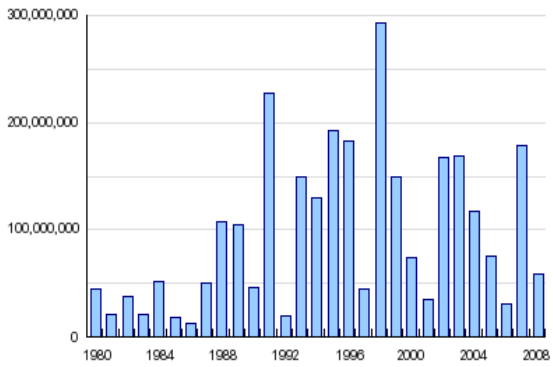
Disaster	Date	Cost (US$ X 1,000)
China P Rep	1998	30,000,000
Korea Dem P Rep	1995	15,000,000
China P Rep	1996	12,600,000
United States	1993	12,000,000
Germany	2002	11,600,000
Italy	1994	9,300,000
China P Rep	1999	8,100,000
Italy	2000	8,000,000
China P Rep	2003	7,890,000
China P Rep	1991	7,500,000

Statistics for flooding and people affected

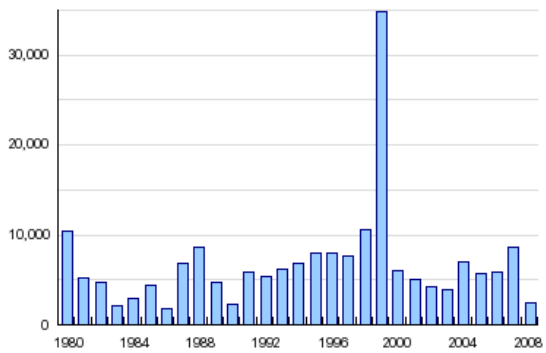
Number of events reported



Number of people affected



Number of people killed



Reported economic damages (US$ in billion)

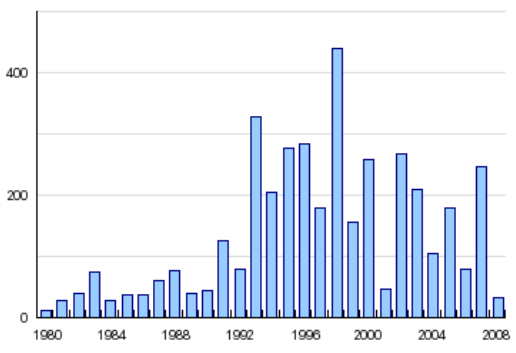


Figure 1.1: Statistics on flooding. Taken from Prevention Web (2011).

Data-driven methods offer an alternative approach to more traditional conceptual and physically-based hydrological models as they are not built using knowledge of the underlying physical processes. Instead these techniques use the data to induce relationships. ANNs, for example, use an input-output training set to learn the relationships from the data. Once the relationship is learned, the model is deterministic and can be used to make forecasts from input data. There are many other advantages of ANNs, which include fast development times, rapid computation times and the ability to generalise to datasets not seen before (Abrahart et al., 2008). Within the area of rainfall-runoff modelling and hydrology more generally, there are hundreds of papers in the academic literature on the application of ANNs (Maier et al., 2010; Abrahart et al., 2010). However, ANNs also have disadvantages where the major one is the black box nature of this method. Although this is true of most empirical models and therefore not entirely a disadvantage of ANNs alone (ASCE, 2000a), it has still meant that hydrologists have been reluctant to use ANNs operationally. In fact there are very few examples of the operational use of ANNs in hydrology as reviewed recently by Macdonald and See (2010). Kneale et al. (1999) developed a user-friendly ANN flood forecasting model, which was integrated with telemetered flow and rainfall data for the North East office of the UK Environment Agency but the system was never used operationally. The chief hydrologist preferred the conceptual River Flow Forecasting System and therefore never properly tested the ANN system (See, 2007, personal communication).

The starting point for this research began through a collaboration between the University of Leeds (Prof Mike Kirkby and Dr Linda See) and La Sapienza University in Rome, Italy (Prof Fabrizio Savi). Prof Savi was interested in comparing his conceptual model with an ANN model. A rigorous comparison would also allow for an assessment of the operational ability of an ANN, particularly at lead times that are meaningful for civil protection agencies. Prof Savi had developed a conceptual model for forecasting floods in the lower Tiber River Basin (subsequently published in Calvo and Salvi, 2009) as this area has been subject to large and frequent flooding. In particular, large areas of these floodplains were inundated in 1937 (when some districts of Rome were also flooded), 1965, 1976, 1992 and more recently in 2005 and 2008 (Natale and Savi, 2004, 2007; Frosini, 1977, Calvo et al., 2007). Natale and Savi (2007) have shown that floods with a return period of just under 200 years will result in overtopping of the river banks and flooding of the northern outskirts of Rome. Risk mitigation strategies can be developed using both structural and non-structural flood protection measures to reduce the possible flood damage. Although structural flood protection measures are usually very effective, they are not always applicable, especially in historical urban areas such

as Rome or in the lower Tiber valley which is densely settled. For this reason, the performance and quality of existing flood warning services must be significantly improved through the development of a real-time operational flood forecasting system. Although such a system does exist for the city of Rome (Todini, 1999), Prof Savi had reservations about the ability and the actual use of the model (Savi, 2007, personal communication).

The Tiber River has been the subject of studies in the past, but these have mainly been concerned with the upper part of the basin and not the lower part that covers the city of Rome. Moramarco et al. (2006) applied an extended Muskingum forecasting model, including lateral inflow contributions, to forecast stage at two gauged river reaches with a lead time of 3.8 hours. The authors concluded that their model can forecast the stage hydrograph with good accuracy when the rating curve at the upstream end is known and lateral inflow occurs in situations where the ratio between the drainage area and the upstream contributing area is small. In another study, Bonafé et al. (1994) applied an ANN to predict daily discharge using precipitation from 26 rain gauges, mean temperature from 13 stations, and mean discharge upstream of the Corbara reservoir. This model was compared to results obtained from the application of the ARX (autoregressive with exogenous input) rainfall-runoff model and a model of persistence. The authors found that the ANN outperformed the other two models, with a Root Mean Squared Error (RMSE) at least 10% smaller than the other two. Tayfur and Moramarco (2007) developed an ANN model for the upper Tiber River basin to predict hourly discharge. They trained a feedforward network with backpropagation on 6 events recorded at three cross sections of the river to predict flow at 4, 8 and 12 hours ahead. They found that the model performed well, especially at a lead time of 8 hours. Calenda et al. (2000) analysed the effect of the precipitation forecast on the real time forecasting of hourly discharges at Ponte Nuovo, by coupling two forecast precipitation models with a lumped rainfall-runoff model. Lead times of 24 hours were forecast but the peak discharge was underestimated by the model. Another study was undertaken by Corradini et al. (2004), who developed a real-time hourly forecasting system at Santa Lucia using a conceptual, semi-distributed rainfall-runoff model. The model simulates the transformation of the effective rainfall into direct runoff by means of the classical Clark translation-routing procedure, while infiltration is schematised by means of two different conceptual models. The values of the parameters of the infiltration models are estimated online via an adaptive calibration, whereas the values of the parameters of the Clark model were estimated offline and did not change during the forecasting. The model does not include forecasts of rainfall, i.e. they assume zero rainfall in the time interval between the forecast start and the forecast lead time. All of

these cited examples show that some relevant work has already been undertaken in this area, but the majority applies to the Upper Tiber basin.

The thesis essentially has three main components in addition to the two literature reviews that provide the necessary context and justification for the research. The first component involves the development of an ANN model for the Tiber River basin. The ANN was then compared to the TEVERE conceptual model of Calvo and Savi (2009) as originally conceived by Prof Savi in 2007. This rigorous comparison represents a real test for ANNs, which is in contrast to what often appears in the academic literature, e.g. comparison with a linear regression model or comparison with other data-driven models such as ARMA models (e.g. Wu and Chau, 2010). If ANNs are ever to be used operationally, their strengths and weaknesses must be clearly understood. Much of the literature is filled with 'hype' and overselling as highlighted by Abrahart et al. (2010) in their review of ANNs in hydrology.

The second component of the thesis follows from the first, i.e. the ANN performed reasonably well in comparison to the conceptual model. However, with a lack of rigorous scientific guidelines on ANN model development, which has been highlighted in the literature on a number of occasions (ASCE, 2000a, b; Maier and Dandy, 2000; Dawson and Wilby, 2001; Maier et al., 2010), the ANN model developed as part of the conceptual model comparison exercise was simple, based on trial and error and the use of a sparse historical data record. The second component is an attempt to improve the model in a number of different ways. The first set of experiments involves adding more data in the form of information from additional upstream stations and rainfall data to see whether this improved the model. With more data, however, the training of the NN was severely impeded. Therefore, attempts were made to build a more parsimonious model using correlation as a way of choosing the model inputs. Experiments were then undertaken to see whether a difference term might improve the model as some limited success with differencing was found by Abrahart and See (2000). Different types of rainfall inputs were used including total rainfall, effective rainfall and cumulative rainfall. Moreover, different methods of normalisation of the inputs and different training algorithms were also used in a range of experiments. The idea was to determine if an empirical pattern would emerge that might provide useful advice for future development of ANN models. Finally, some experimentation was undertaken with ensemble modelling. It was suggested by Anctil (2007) that an ANN should be trained multiple times, e.g. 50, and then an average taken of the outputs in an ensemble approach in order to minimise the effect of the random initialisation of the weights of the ANN. Experimentation was undertaken with a persistence-based

performance measure and the Akaike Information Criterion (AIC) to combine the ensemble.

Another major issue addressed in this thesis is pre-processing. There has been a recent trend to decompose ANNs using wavelet analysis (e.g. Rao and Krishna, 2009; Kisi, 2008; 2009; 2010; Adamowski, 2010). The effect of this decomposition is to extract key components of the time series, which are clearer signals that can be used to better predict the stream flow. In this research, another type of pre-processing method is used, called Empirical Mode Decomposition (EMD), which has not been applied to ANN rainfall-runoff modelling before. The effect of this pre-processing method is critically evaluated in conjunction with another issue, i.e. the random initialisation of the ANN weights. Although 50 runs were averaged to compensate for this effect, plotting the individual ANN ensemble members reveals just how much spread there is in model prediction, particularly at the peaks of flood events. This variation is not reported in the literature. The analysis undertaken in this third component of the thesis therefore highlights an important issue about ANN modelling that needs further attention.

Finally, an issue that runs through all three components is the use of performance measures for ANN model evaluation. There is a reasonable literature on performance measures in hydrology (e.g. Green and Stephenson, 1986; Legates and McCabe, 1999; Dawson et al., 2007) yet there is also little guidance on which measures to choose. In the ANN rainfall-runoff modelling literature, there is little attention paid to the choice of measures. Although Dawson et al. (2007) advocate the use of many measures together, they provide little further guidance to help the ANN modeller. Thus, a review was undertaken of the measures commonly used as well as other measures that have potential for ANN rainfall-runoff modelling. A subset of these measures was then chosen for application and critical evaluation throughout the thesis.

1.2 Aims and Objectives

The overall aim of this research is to examine a number of issues related to the use of ANNs for rainfall-runoff modelling and flood forecasting, in particular the need to rigorously compare ANNs with conceptual/physical models; the use of different performance measures for model evaluation; the problems associated with training ANNs using different starting initialisations; and the use of ensemble methods, all of which have been identified as ongoing issues from a review of the literature. The overall aim of this research will be achieved through the following objectives:

1. To review and critically evaluate the academic literature on ANN rainfall-runoff modelling, which is provided in Chapter 2.
2. To review and evaluate the performance measures which are used to evaluate model performance, choosing a subset for use in this research. This is the subject of Chapter 3.
3. To develop an ANN rainfall-runoff model of the Tiber River basin and compare this with the conceptual TEVERE model. The experiments and the results are described in Chapter 4.
4. To undertake a series of different experiments to improve the basic ANN rainfall-runoff model developed in Chapter 4 including a brief look at ensemble methods. This is the subject of Chapter 5.
5. To apply a pre-processing method called Empirical Mode Decomposition (EMD) to ANN rainfall-runoff modelling and examine the impact of the random weight initialisation of ANNs on the model outcomes. The methods and experiments are applied to two rivers in the USA, which is presented in Chapter 6.
6. To highlight the limitations of the study and to make recommendations for further research, which follow the conclusions presented in Chapter 7.

The value of this thesis derives from a) a rigorous comparison of a conceptual model with an ANN to evaluate the usefulness of ANNs for operational flood forecasting, something which should be reported more often in the literature; b) a critical examination of performance measures, which are not reported with any consistency in the ANN rainfall-runoff modelling literature; c) the application of a novel pre-processing technique called Empirical Mode Decomposition to ANN rainfall-runoff modelling, which has not been tried before; and d) an examination of the impact of the random initialisation of the weights of an ANN, another area where little research has been undertaken to date.

1.3 Thesis Structure

Chapter 2 provides a literature review on NNs and hydrological modelling in order to provide the scientific context for this research. The chapter begins with an overview of hydrological modelling more generally and then focuses on ANNs as an empirical approach. The main issues with ANN model development are highlighted and the main advantages and disadvantages of ANNs for hydrological modelling are then discussed. A review of the literature is then presented as a historical summary, a set of main themes, which have appeared over the last decade with examples of applications, and areas where further research has been recommended. The research undertaken in this thesis is then placed within this broader literature, justifying the research questions that

have been tackled. One of the areas for investigation is in model performance measures. Chapter 3 therefore provides an overview of the main model evaluation measures that have been used in hydrological modelling and in ANN rainfall-runoff modelling. Other measures of potential relevance that have not been used in ANN rainfall-runoff modelling are also discussed. Finally, a suite of measures is chosen for application throughout the rest of the thesis.

Three substantive modelling chapters then follow. Chapters 4 and 5 are concerned with ANN modelling of the Tiber River basin while Chapter 6 considers the Potomac River and the Clark Fork River basin in the USA. Chapter 4 provides an overview of the Tiber River Basin and focuses on flooding in the city of Rome. The conceptual TEVERE model is then presented, which has been used to predict the very large historical floods that occurred in 2005 and 2008. The development of an ANN model to predict both of these flood events is also described. The motivations behind the chapter are to see whether a simple ANN model can be developed that has performance similar or better than the conceptual model, as comparisons with conceptual models are not as common as they should be in the ANN rainfall-runoff literature. This comparison provides a true test of the skill of an ANN for operational flood forecasting. Chapter 5 outlines more than 60 experiments where attempts are made to improve the simple ANN developed in Chapter 4 through the addition of different inputs (i.e. upstream stations and rainfall), different normalisation techniques, different training algorithms and different methods to combine the ensemble of ANN models that is produced when developing an ANN. During the course of this research, it was observed that there is a large variation in model predictions from the ANN ensemble members, particularly when predicting the flood events and especially the peaks. This is function of the random initialisation of the weights, which is barely considered in ANN rainfall-runoff modelling. This motivated further investigation of this topic, which is addressed in Chapter 6. In addition, a pre-processing method referred to as Empirical Mode Decomposition (EMD) is used. Pre-processing methods, in particular, wavelets, are starting to be used much more frequently in ANN rainfall-runoff and other ANN hydrological modelling studies (e.g. Rao and Krishna, 2009; Kisi (2008; 2009; 2010); Adamowski (2010). EMD has not been used before in ANN rainfall-runoff modelling so this research provided an ideal opportunity to apply and critically examine this technique for stream flow forecasting. To use EMD and to also determine whether the random initialisation of weights is also an issue when modelling other catchments, two rivers in the USA with very long time series were chosen, i.e. the Potomac River and the Clark Fork River. These data were freely available for downloading from the US Geological Survey website. Although the error measures chosen in Chapter 3 are

applied consistently throughout the thesis, a very critical examination of them is provided in Chapter 6.

The thesis concludes in Chapter 7, where a summary of the main results is provided, which are related directly back to the aims and objectives that appear in section 1.2 of this introductory chapter. Chapter 7 also highlights the problems that were encountered and the limitations of the research, and then makes suggestions for areas of further study in the future.

1.4 Summary

The context for this research has been presented in this chapter along with the overall aims and objectives. In the next chapter, a literature review of ANNs is provided which covers the basic theory behind this technique, as well as the advantages and disadvantages for hydrological modelling. The considerable body of literature is then reported under a series of main themes that have emerged over the last two decades.

Chapter 2

Artificial Neural Networks for Hydrological Modelling

2.1 Introduction

There has been a great deal of interest in the use of artificial neural networks (ANNs) in hydrology over the last two decades. The purpose of this chapter is to review the application of ANNs within hydrological modelling, in particular with respect to river flow forecasting (or rainfall-runoff modelling). This chapter begins with a very brief overview of hydrological modelling in general and places ANNs within the typology of modelling approaches used. This is followed by an overview of ANNs in terms of definitions, origin, structure, model development, and advantages and disadvantages. This chapter concludes with a review of the major themes that have emerged from the ANN river flow forecasting literature and how the research in this thesis fits within this context.

2.2 Approaches to Hydrological Modelling

Hydrological modelling attempts to represent processes within the hydrological cycle in a simplified manner. These models are used to improve understanding of the processes which underlie the system as well as to make forecasts of the future, e.g. flood events, occurrence of rainfall, river levels, snow melt, evaporation and sediment concentration or volume. The reader is referred to Anderson and Burt (1985) for a range of hydrological modelling application areas.

All models are simplifications of reality and there are many different ways to represent it. Therefore, one can find different approaches to modelling within hydrology. A number of authors (Wilby, 1997; Anderson and Burt, 1985; ASCE, 2000a) have characterised hydrological modelling approaches into three main types: process-based; conceptual; and empirical or data-driven (which includes statistical). Wheater et al. (1993) add a fourth type that is a hybrid of the conceptual and statistical types. Each of the three main types is now briefly reviewed.

2.2.1 Physically-based or Deterministic Models

Physically-based models (Wilby, 1997), also referred to as deterministic models (Anderson and Burt, 1985), represent the physical characteristics of the catchment. The SHE (European Hydrological System) model is the most famous physically-based model. It has been applied to a range of areas including flood forecasting, examining the effects of land use change and ground water modelling (Abbott et al., 1986a, b). Other examples of physically-based models include the Institute of Hydrology

Distributed Model (IHDM) (Beven et al., 1987) and the WATFLOOD model (Kouwen, 1988).

Like global circulation models of the climate, physically-based models represent a catchment as a three dimensional grid. They use the fundamental laws of the conservation of energy and mass to model water movements on the surface and through the unsaturated and saturated zones to the river. A flood hydrograph is then dynamically built from the model runoff (Wood and Connell, 1985). These models incorporate as full an understanding of the catchment processes as possible. Thus, when conditions change, the models can be used to evaluate the impact on runoff or other catchment properties (Anderson and Burt, 1985). Although these models are the most complex and accurate of the three model types, they require a large amount of data and processing time, which is not always available for all catchments. These models therefore have more value for planning than use in real-time forecasting. For these reasons, other more practical approaches to hydrological modelling have been developed.

2.2.2 Conceptual or Lumped Models

The second type of approach is referred to as a conceptual, lumped conceptual or geomorphology-based model (Wilby, 1997; Wood and Connell, 1985; ASCE, 2000a); these are viewed as the most successful model types for rainfall-runoff simulation and flood modelling. These models still have a physical basis but they are structured in such a way as to represent a stream network and the surrounding catchment. These models attempt to represent the main dynamics in the catchment but are characterised by parsimony and computational efficiency (Kavetski et al., 2006), requiring calibration of between 8 to 20 parameters (Blackie and Eeles, 1985). Conceptual models are therefore less demanding compared to physically-based models but require more information than empirical data-driven models. TOPMODEL (Beven and Kirkby, 1997) is a classic example of a simple yet powerful conceptual model. TOPMODEL has been used in numerous applications covering a range of catchment sizes and geographical areas, and continues to be used in research to the present day (e.g. Gallart et al. 2008; Peng et al., 2009; Vincendon et al., 2010; Buytaert and Beven, 2011). Like physically-based models, conceptual models can also be used to examine changes to the catchment, e.g. from land use (Beskow et al., 2011). Other examples of successful conceptual models include the Tank model (Tingsanchali and Gautam, 2000), the United States National Weather Service River Forecasting Model (Wood and Connell, 1985) and the SWAT (Soil and Water Assessment tool) model (Arnold et al., 1998; Arnold and Fohrer, 2005).

2.2.3 Empirical, Data-driven or Black Box Models

This approach, as the name suggests, tries to find an empirical relationship between a set of inputs (e.g. historical data such as rainfall and temperature) and a set of outputs (e.g. runoff). Statistical approaches are also included in this category. These models do not use physical equations, catchment characteristics or other physical parameters (Wilby, 1997; ASCE, 2000a; Anderson and Burt, 1985). One major advantage of this approach is that these models are generally very fast to run and much faster to develop than physically-based or conceptual models. This makes them particularly useful for real-time forecasting. A disadvantage is that they are static, i.e. they cannot take change into account, e.g. changes in land use. However, there are approaches to update the models when new data become available (Wood and Connell, 1985). Unit hydrographs are a classic empirical model (Dooge, 1959), which capture the relationship between rainfall and catchment response. Another example is time series models such as ARMA (Auto-Regressive Moving Average) models (Box and Jenkens, 1970), which are used frequently as a type of empirical model to capture the rainfall-runoff relationship (see e.g. Salas and Obeysekera, 1982; Lin and Lee, 1994).

ANNs are classified as an empirical, data-driven or black box model. Data are fed into the model, a relationship is learned between the inputs and the outputs, and the model is then used to produce a forecast. ANNs do not require any understanding of the physical processes underlying the system. However, some hydrological knowledge is a prerequisite as it guides which kinds of inputs to choose (i.e. rainfall, previous flows, water levels, etc.) and which outputs to forecast (i.e. runoff, stream flows, hydraulic conductivities, etc.). This is a modelling type that does not always appeal to hydrologists who prefer the core of the model to be a dynamic, physically-based, representation of the processes involved. However, some research reported has used sensitivity analysis to determine the most significant inputs to the network (Abrahart et al., 2001; Sudheer, 2005). This allows one to gain a better understanding of what is happening in the model. Other research has looked at opening up the black box to see whether structures within the ANN have hydrological meaning (Wilby et al., 2003; Jain et al., 2004a). Thus, research is underway to interpret ANNs in a physical way. The increasing volume of research on ANNs in hydrology over the last decade alone indicates that ANNs are gaining more and more interest as an empirical hydrological model (see section 2.5). In the next section, an overview of the main concepts behind ANNs is provided.

2.3 Overview of Artificial Neural Networks (ANNs)

There is no universal definition of an ANN in the literature. Zurada (1992) expresses an ANN in a very abstract way as a physical cell-based system that can collect and use

knowledge. Nigrin (1993) refers to an ANN as a circuit of simple neurons that function similar to neurons in the brain. Haykin (1994) emphasises the parallel and distributed nature of an ANN, which emulates the brain through a learning process, with information stored in the synaptic weights that connect the neurons. Finally, the ASCE (2000a) adds the idea of adaptation and generalisation to the definition of an ANN since a trained ANN should be able to make predictions using data it has not seen before. This ability to adapt or generalise is one of the most important features of an ANN.

The original concept of an ANN was developed in 1943 by Warren McCulloch and Walter Pitts, who proposed the conceptualisation of human brain function based on a network of interconnected cells (McCulloch and Pitts, 1943). In 1951, Minsky and Edmonds built the first neural network machine, which was used to follow the progress of a rat through a maze where the neural machine played the role of the rat. The experiments showed that the rat was able to start thinking and that even when one of the physical neurons failed, the system still worked (Simpson, 1990). In 1962, Rosenblatt developed the perceptron, which was a simple arrangement of interconnected artificial neurons in a single layer, along with a learning algorithm (Russell and Norvig, 1995). ANN research then entered a dark phase that lasted almost 20 years. By the late sixties, it became clear that the perceptrons of Rosenblatt could not represent very complex functions, which was demonstrated in a key publication by Minksy and Papert (1969). It was not until the mid-eighties that the main obstacle to ANNs was overcome, i.e. an extra layer was added to the perceptron (henceforth called a multi-layer perceptron - MLP) and an efficient algorithm for learning was developed called backpropagation (BP) (Rumelhart et al., 1986). It has now been proven mathematically that an MLP can approximate any function from a one finite dimensional space to another up to any desired degree of accuracy. This is referred to as universal approximation (Hornik *et al*., 1989). This means that ANNs should theoretically be applicable to any hydrological modelling problem. Before considering how ANNs are developed and applied, the next section will first outline the structure or architecture of ANNs.

2.3.1 Structure of ANNs

ANN structures are described in a number of classic textbooks and papers, e.g. Bishop (1995), Schalkoff (1997) and ASCE (2000a). An ANN is comprised of a series of information processing elements referred to as nodes or neurons. Information is passed between nodes through connections. Weights are then associated with each connection, which represents the magnitude or strength of that connection. Within the

node is a nonlinear transformation function, called an activation function, which is applied to the input signals coming into the node to produce an output signal. The simplest function usually applied is the sigmoid function (see e.g. ASCE, 2000a; Minns and Hall, 1996; Raman and Sunilkumar, 1995; Dawson and Wilby, 2001). The nodes or neurons are then arranged into a series of layers: an input layer; one or more hidden layers; and one output layer (Figure 2.1). A weight matrix *W*, a bias vector *b*, and an activation or transfer function *f* is associated with each hidden layer (Schalkoff, 1997).

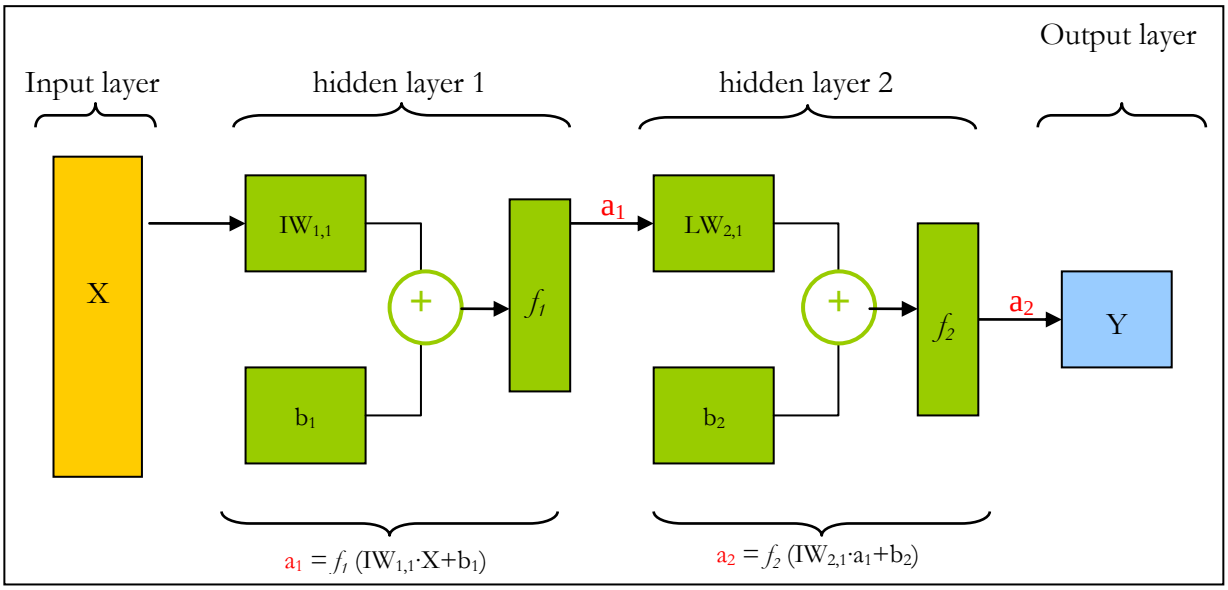


Figure 2.1: Two layer network in abbreviated notation.

The input layer is where external information is received and provided to the network (e.g. antecedent rainfall or runoff) while the output layer produces the forecast (e.g. the river level in 12 hours time). Each node is connected with all other nodes of the previous and the next layer. The representation of nodes in each layer and the interconnections are more clearly shown in Figure 2.2. This is an example of a feedforward network, i.e. the information flows in a forward direction through the network and there are no feedback effects.

The output of each node is obtained by computing the value of the activation function with respect to the product of the input vector and the weight vector, minus the value of the bias associated with that node. It is possible to express the forward processing through the network as a single equation. A network with one hidden layer and K outputs would have the following functional form:

$$f_k(x,y) = w_{k0} + \sum_{j=1}^{q} w_{kj}\, g(w_{j0} + \sum_{i=1}^{p} w_{ji}\, x_i) \qquad (2.1)$$

where p is the number of inputs, q is the number of nodes in the hidden layer, g is the activation function of the hidden layer nodes, and x and w are the weights. The indices i and j correspond to the output node and hidden layer nodes, respectively. However, ANNs are rarely expressed in this manner as the equation is not interpretable.

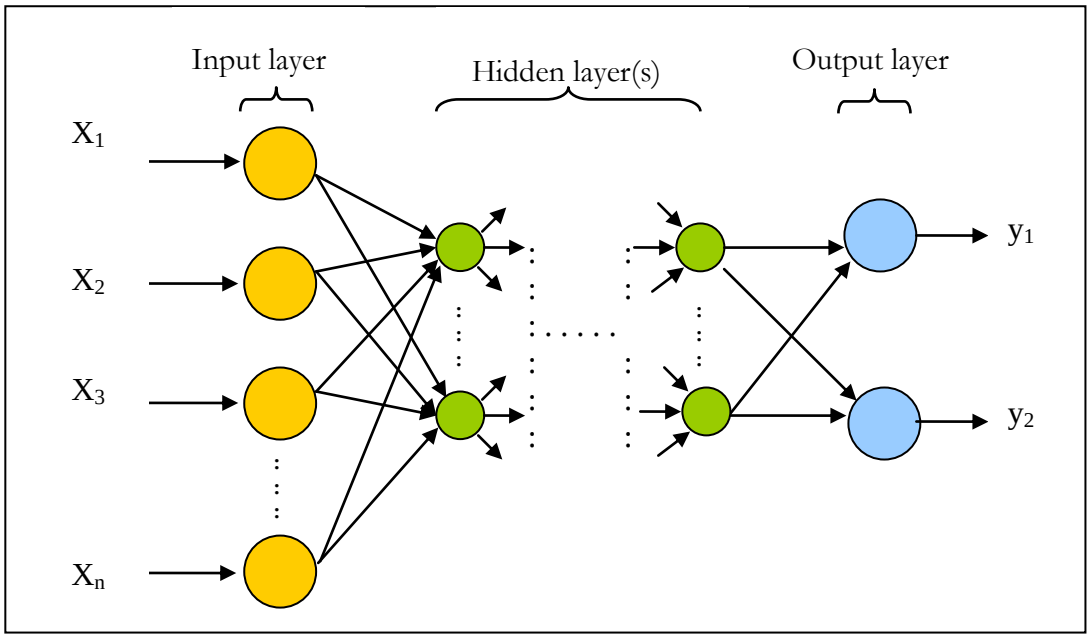


Figure 2.2: Schematic of a multi-layer ANN

2.3.2 Training an ANN

Once the network structure is set, the ANN is then trained. The process of training or learning is used to find the values of the weights W that minimise the error between the inputs and the outputs in the training data set:

$$E_D = \sum_{i=1}^{n}(t_i - a_i)^2 \tag{2.2}$$

where E_D is the sum of the errors squared between the targets, t, and the ANN response, a, for i observations in the input-output dataset.

The training procedure iteratively adjusts the weights of each node until a stopping condition is reached. The initial weights are first randomly selected. During a single training run, the algorithm may fall into a local minimum on the error surface (see Figure 2.3) so it is advisable to train the network several times. Minimising the sum of the errors squared is the most commonly used objective function. This assumes that model errors are normally distributed with mean zero and unknown variance (Velásquez et al., 2006).

To achieve an acceptable level of generalisation by the ANN, the data set is usually divided into three subsets:

- **A training data set:** this is the data set used to train the ANN or which allows the ANN to learn the relationships in the data.
- **A cross training or validation data set:** this is the portion of the overall data set that is reserved to help stop the training process. Otherwise the ANN may overfit the data and lose the ability to generalise to an unseen data set.
- **A testing data set:** this portion of the data is used to test the network on an unseen or independent data set not used during the training process. It is on this data set that the performance of the network is measured.

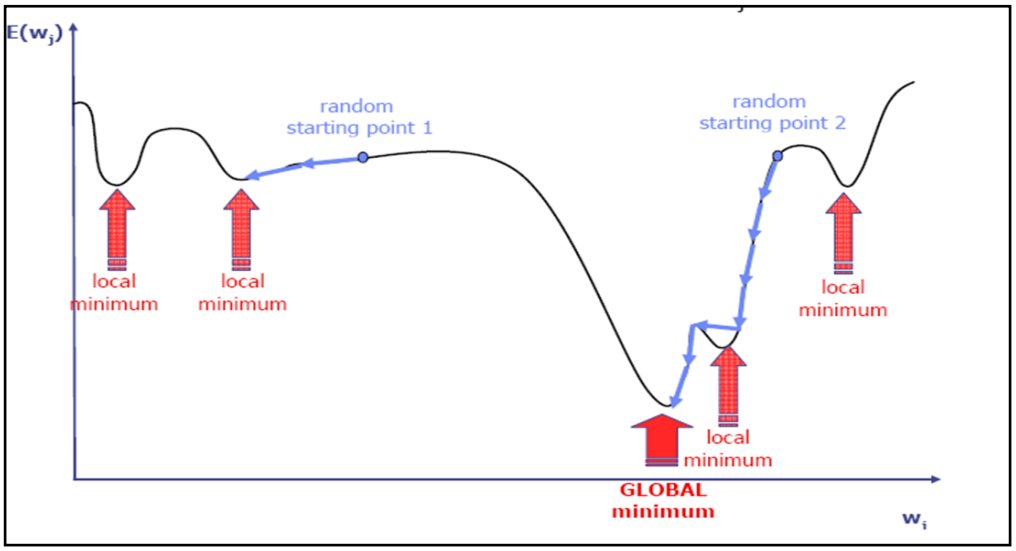


Figure 2.3: Local and global minima of errors

The goal of ANN training is to produce a network with small errors in the training dataset, but which will also performs well on the testing dataset (Foresee and Hagan, 1997). There are two general kinds of training algorithm:

- **Supervised:** the target or known output is available and the network finds the best set of weights by minimising the error between the output and the target.
- **Unsupervised:** only the input data set is given to the network, which then tries to find clusters of similar inputs without any previous knowledge.

Some of the most commonly used supervised training algorithms are:

- **Back-propagation:** Developed originally by Rumelhart et al. (1986), this algorithm updates the ANN weights and biases based on the negative of the gradient. At each iteration, the following equation is applied:

$$x_{k+1} = x_k - \alpha_k g_k \tag{2.3}$$

where x_k is a vector of weights and biases at iteration k, g_k is the gradient, and α_k is the learning rate. This equation is applied through successive iterations in which the error function is reduced until a stopping condition has been reached. The momentum factor ($\propto$) and the learning rate (ε) determine how much each connection weight is adjusted. The first parameter is used to speed up the training of the network, and it prevents possible oscillations in the weights. The second one helps the network to avoid becoming trapped in a local minimum instead of finding a global minimum (ASCE, 2000a) (Figure 2.3). If the learning rate is large, the steps taken are bigger and the training proceeds faster. However, too large a learning rate may result in instability. In addition, the network performance is influenced by the number of training samples used in each step before the weights are updated (Maier and Dandy, 1998b). The weight update equation for the connection weight between nodes *i* and *j* is given by the following equation:

$$\Delta w_{ij}(n) = -\varepsilon \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(n-1) \tag{2.4}$$

where $\Delta w_{ij}(n)$ and $\Delta w_{ij}(n-1)$ are the weight increments between node i and j during the n^{th} and (n -1)$^{\text{th}}$ epoch.

- **Conjugate gradient descent:** Unlike backpropagation, this algorithm propagates the error in a direction orthogonal to the previous step instead of proceeding along the direction of the error gradient. Thus, the equation for updating the weight vector, $W(n+1)$, can be written as a function of the learning rate, ε, the weight vector at the previous time step, $W(n)$, and the gradient vector, $g(n)$ (Fletcher and Reeves, 1964). If $P(n)$ is used to identify the direction vector at the n^{th} iteration of backpropagation, then equation (2.3) could be rewritten as (Haykin, 1994):

$$W(n+1) = W(n) + \varepsilon P(n) \tag{2.4}$$

The initial point is equal to zero, and thus the direction vector is equal to the negative gradient vector $g(n)$. Every subsequent direction vector is calculated from

the current gradient and the previous direction vectors as:

$$P(n+1) = -g(n+1) + \beta(n)P(n) \tag{2.5}$$

where $\beta(n)$ represents a time-dependent parameter, defined by Fletcher and Reeves (1964) as:

$$\beta(n) = \frac{g^T(n+1)g(n+1)}{g^T(n)g(n)} \tag{2.6}$$

- **Radial Basis Function:** Radial basis functions, R_i, have the general form:

$$R_i = \left(-\sum_{i=1}^{n} \frac{\|x_i - c_i\|^2}{2\sigma_{ij}^2} \right) \tag{2.8}$$

where $c^T = [c_{i1},\ c_{i2}, \ldots., c_{in}]$ is the center, and σ_{ij} is the width of the Gaussian function. The centre is chosen from the training set, or using a technique of clustering. The input training dataset is divided into groups where the mean of each group becomes the centre (ASCE, 2000a). The main difference between the RBF network and backpropagation lies in the nonlinearities related to the hidden nodes. Once the basis functions (R_i) in the hidden layer have been found, the network only needs to learn the weights associated with the output layer. The output y of an RBF network is computed as follows:

$$y = f(u) = \sum_{i=1}^{n} w_i R_i(x) + w_0 \tag{2.7}$$

where w_i *is* the connection weight between the hidden neuron and the output neuron, w_0 is the bias, and x *is the* input vector.

- **Cascade correlation algorithm:** developed by Fahlman and Lebiere (1990), it adds hidden nodes to the network one at a time and trains only the output weights. This algorithm adjusts the weights to maximise the overall correlation between the hidden node values and the residual error:

$$S = \sum_{O} \left| \sum_{P} (V_P - \bar{V}) \left(E_{P,O} - \overline{E_O} \right) \right| \tag{2.9}$$

where V_P is the output of the new hidden node for observation P, $\bar{V}$ is the average output over all observations, $E_{P,O}$ is the error for output node O on observation P, and E_O is the average error of the ANN over the training dataset.

- **Levenberg-Marquardt (LM):** The Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) is used to improve the training speed while avoiding the Hessian matrix. If the objective function of the ANN is the sum of squares, then the Hessian matrix can be written as

$$H = J^T J \tag{2.10}$$

and the gradient can be determined as

$$g = J^T e \tag{2.11}$$

where J is the Jacobian matrix that contains the first derivatives of the errors of the ANN with respect to the weights and biases of the network while *e* is a vector containing the ANN errors. Adopting the Jacobian matrix is much easier than using the Hessian matrix because it can be calculated through standard backpropagation (Hagan and Menhaj, 1994). The Levenberg-Marquardt algorithm approximates the Hessian matrix as follows:

$$x_{k+1} = x_k - [J^T J + \mu I]^{-1} J^T e \tag{2.12}$$

When μ is greater than zero, equation 2.12 becomes gradient descent with a decreased step size. Thus, μ is decreased after each iteration and results in a reduction in the objective function.

- **Bayesian Regularization (BR):** BR is an ANN algorithm that updates the weights and bias values using the LM algorithm as a basis. It minimises the squared errors and ANN weights, and then determines a suitable combination of the two in order to produce a network with good capability of generalisation. To do this, BR adds an additional term, i.e. the sum of squares of the network weights, E_W, to the error function, F (Foresee and Hagan, 1997). This approach improves the generalisation capability of the network (Doan and Liong, 2004) and it does not require a validation data set (Doan and Liong, 2004; Hirschen and Schafer, 2006). Using the Bayesian framework developed by MacKay (1992a, b), the optimal weight decay

coefficients can be determined which prevent overfitting of the data in ANN training. To do this, the BR adds an additional term to equation (2.3) to yield:

$$F = \beta E_D + \alpha E_W \tag{2.13}$$

where E_W is the sum of squares of the network weights, and α and β are parameters which are optimised using the Bayesian framework of MacKay (1992a, b). This approach improves the generalisation capability of the network. The weights and biases of the ANN are assumed to be random following a Gaussian distribution of unknown variance (Doan and Liong, 2004).

2.3.3 Types of ANN

The feedforward ANN or MLP is the most commonly used type of ANN (Gallant, 1993). This also holds true in hydrology. Maier et al. (2010) reviewed 210 papers published between 1999 and 2007 that used ANNs in hydrological modelling. They found that 178 out of 210 papers (or roughly 85%) used an MLP. Other types of ANNs include Radial Basis Function Networks (RBFNs), Recurrent Neural Networks (RNNs) and Self-Organizing Maps (SOMs) (Bishop, 1995). RBFNs have architectures that are similar to MLPs. The RBFN differs in the form of the activation function in the hidden layer nodes, which are of a radial basis or Gaussian form. The parameters of the radial basis functions are usually determined first followed by the weights during the training process. When large training datasets are available, RBFNs can be particularly good models to use (Achela et al., 2009; Shamseldin, et al., 2007). Based on the review by Maier et al. (2010), RBFNs have not been used that frequently for hydrological modelling, i.e. less than 20 out of 210 papers (or less than 10%) over the period 1999 to 2007.

Another type of network involves modification of feedforward networks to allow for feedbacks between layers or RNNs. The main success in applying this network has been in handwriting recognition applications (Graves et al., 2009). Different learning algorithms have had to be developed for this type of network (Schmidhuber, 1989; Williams and Zipser, 1994). However, the uptake of RNNs was similar to RBFNs for hydrological modelling, i.e. less than 20 times over the period between 1999 and 2007 (Maier et al., 2010).

A Self Organizing Map (SOM) is a slightly different type of ANN, which was first introduced by Kohonen (1984). SOMs use an unsupervised classification method used to cluster data into similar types. Therefore, only an input data vector is required. These

ANNs are also used for data compression and for visualisation of relationships in the data (Sharma et al., 2007; Vesanto, 1999). However, it is also possible to use SOMs in predictive mode, e.g. Corne et al. (1999) and the SOLO (Self-Organising Linear Output) map developed by Hsu et al. (2002), which simultaneously classifies the input data and makes a prediction. In their simplest form, SOMs are composed of a single two dimensional layer of input neurons of vector *w*. The training algorithm randomly selects an input vector *x* and finds the best matching or winning neuron based on the Euclidean distance between the two vectors *x* and *w*. Then all of the nodes near the winning node (or in a specified neighbourhood) have their weight vectors updated. The next set of input data are then presented to the SOM until a stopping condition is satisfied, e.g. a minimum error is reached (Kohonen, 1984). A disadvantage of the SOM is that there is no clear methodology for the selection of the number of nodes (or clusters) or the values of the learning parameters, e.g. the size of the neighbourhood. Based on the review by Maier et al. (2010), SOMs were used only 10 times during the period 1999 to 2007, making them even less popular than the RBFNs and RNNs for hydrological modelling.

2.3.4 Development of an ANN

There is currently very little guidance in the literature on how to develop an optimal ANN for a given application. A number of papers have appeared that attempt to provide some guidance or results from empirical experiments that help to make decisions regarding model development (Dawson and Willby, 1998; Maier and Dandy, 1998a, b; Maier and Dandy, 2000; ASCE, 2000a, b; Dawson and Wilby, 2001). These decisions include things like the choice of model architecture, the number of hidden layers, the number of hidden nodes, the choice of input variables, the choice of activation functions, the choice of training algorithm, and which performance measures to use, the latter of which is dealt with in Chapter 3.

The architecture of an ANN is often selected via trial and error as there is no currently established methodology (ASCE, 2000a; Dawson and Wilby, 2001). Once the type of network is chosen, the number of hidden layers and the number of nodes in each hidden layer must be determined. However, there are no fixed rules so experimentation is therefore often by trial and error. There are, however, some heuristics in the literature. For example, Hecht-Nielsen (1987) suggested the following upper limit for the number of hidden layers to ensure that the network is able to approximate any continuous function:

$$N^H \leq 2N^I + 1 \tag{2.4}$$

where N^H is number of hidden layers and N^I is the number of inputs. Rogers and Dowla (1994) suggested a second relationship to avoid overfitting as follows:

$$N^H \leq \frac{N^{TR}}{N^I + 1} \tag{2.5}$$

where N^{TR} is the number of observations in the training dataset.

Depending upon which training algorithm is chosen, a number of parameters need to be specified, e.g. the learning rate and momentum. The literature reveals that authors have chosen these parameters using trial and error. For example, 0.1 was used as the learning rate by Smith and Eli (1995), Maier and Dandy (1999) and Thirumalaiah and Deo (1998a, b) while values of 0.01 to 0.0005 were employed by Tayfur and Moramarco (2007). High values of momentum (i.e. 0.6 or >1) were chosen by Maier and Dandy (1998; 2000) while lower values of 0.2 were chosen by Thirumalaiah and Deo (1998a, b). Many papers do not even report these parameters as they are deemed to be specific to an application.

Another important decision to make is in the choice of the input variables. The majority of ANN studies have used previous values of river levels or flows and rainfall (e.g. Zealand et al., 1999; See and Openshaw, 1999; Doan and Liong, 2004; Aqil et al., 2007). Total rainfall is normally used although some papers have used effective rainfall in place of total rainfall (e.g. Sajikumar and Thandaveswara, 1999; Jain and Srinivasulu, 2006). Effective rainfall is total precipitation minus losses, which is what actually produces the runoff. However, the problem with effective rainfall is that it is difficult to estimate because it depends on the antecedent moisture conditions of the basin, which changes over time. This would explain why it has not been used very often in ANN rainall-runoff modelling.

Other variables that have been used include temperature (Nayebi et al., 2006); evapotranspiration (Anctil and Rat, 2005); moving average antecedent precipitation (Abrahart, 2001); and soil moisture (Karunanithi et al., 1994). The proper choice of input variables is very important for the efficiency of the ANN (Maier et al. 2010). If too many inputs are included and they are not independent, this increases the size of the network so training will take longer. There is also a greater likelihood of overfitting the training data because the ratio of connection weights to training data increases. Methods for choosing the input variables include trial and error, correlation analysis of the variables (Kumar and Minocha, 2001, Olsson et al., 2004; Sudheer and Jain,

2004), sensitivity analysis (Liong et al., 2000), the time of concentration to determine the amount of antecedent precipitation (Jain, 2005), the F-statistical method (Furundzic, 1998), pruning algorithms (Furundzic, 1998; Maier and Dandy, 2000), Average Mutual Information (Abebe and Price, 2004), genetic algorithms (Bowden et al., 2005a, b; Anctil et al., 2006), Partial Mutual Information (Bowden et al., 2005a, b) and SOMs (Bowden et al., 2005a, b; Toth, 2009). Although many different methods exist, trial and error and correlation analysis are used most frequently (Chaipimonplin, 2010).

Despite the fact that guidance in ANN model development is clearly lacking, ANNs have a series of characteristics that make them very useful for hydrological modelling. The next section summarises the advantages and disadvantages associated with ANNs.

2.4 Advantages and Disadvantages of ANNs for Hydrological Modelling

ANNs owe their information processing capability to their distributed and parallel nature as well as their ability to learn from the data and generalise to situations not seen previously. It is the collective power of the network that allows for the implementation of a surprising number of complex tasks with great efficiency (Reilly and Cooper, 1990). Below is a list of the characteristics or properties of ANNs with reference to hydrological modelling. Some of these characteristics are advantages while others are limitations.

1. **Non-linearity:** Most hydrological problems are non-linear. The interconnection between the neurons in an ANN generates non-linear data processing structures that are distributed across the network. This feature allows intrinsically non-linear processes to be modelled, such as the transformation of rainfall into runoff. For example, Abrahart and See (2007a) showed that ANNs can be used to emulate the outputs of the Xinanjiang Rainfall-Runoff Model, which is a simple non-linear model.

2. **Ability to model input/output relationships:** ANNs do not need an explicit mathematical equation to specify the relationship between the inputs and the outputs. Thus in situations where the processes are not fully understood, ANNs can be used to develop simple models of hydrological value.

3. **Adaptability:** One criticism that is often levelled at ANNs is their static nature. However, ANNs are technically adaptable to change, i.e. capable of adjusting their

weights, as new data become available. This feature makes them particularly useful in the treatment of non-stationary processes, where learning strategies can be designed in real-time so that the model learns continuously. There is some limited research ongoing into dynamic ANNs (e.g. Coulibaly and Baldwin, 2008) but it remains an important area of investigation. However, most studies do not explicitly consider non-stationarity or adaptability.

4. **Rapid construction:** The process of constructing an ANN is very fast relative to conceptual and physically-based models. However, it does require making choices such as which input variables to choose, the architecture of the network, the training algorithm, how much to lag the data to account for travel times, etc. Moreover, the absence of clear guidance on ANN construction and the lack of concrete procedures for determining an optimum network render it a somewhat subjective process at times. Trial and error is often the most commonly used method, which can never be totally exhaustive.

5. **Computationally efficient:** This property is related to the previous one, i.e. once the ANN is trained, it is very fast to run and computationally very efficient (ASCE, 2000a). This also refers to the fact that many of the training algorithms are computationally efficient or improvements have been made through research, e.g. the development of second order methods, which are improvements over first order methods such as backpropagation in terms of training time (Ampazis and Perantonis, 2002).

6. **Less sensitive to noise in the data:** Since ANNs are models of a distributed nature, they technically have a greater ability to handle noise in the input data (Karunanithi et al., 1994; Thirumalaiah and Deo 1998a, b; Zealand et al., 1999, ASCE, 2000a). However, to have an effective ANN that handles the problem of noise, a large amount of data is required to train the network. Thus success relies on the quality and quantity of the data set.

7. **Modularity:** ANNs can be integrated easily into modular architectures to very efficiently solve specific subtasks of the overall problem to which they are best suited in a plug and play type of approach. These sub-tasks may include procedures for pattern recognition, functional approximation, etc. An example is the embedding of an ANN within an expert or decision support system (Bhattacharya et al., 2003).

8. **Black box nature:** The main disadvantage of ANNs is in their black box nature, which makes them less preferable to physically-based and conceptual models, especially for real-time applications. All the major reviews (ASCE 2000b; Maier and Dandy, 2000; Dawson and Wilby, 2001; Maier et al., 2010; Abrahart et al., 2010) have acknowledged that trying to open up the black box and find physical meaning in ANNs is an area of research that needs more attention. Some preliminary studies have begun to address this issue (e.g. Wilby et al., 2003; Sudheer and Jain, 2004), but it continues to be an area where further research is needed.

Many of these advantages have been exploited through the application of ANNs in hydrology, and rainfall-runoff modelling in particular, while the disadvantages remain areas for further research. In the final section, a review of the main themes and applications that have emerged during the past two decades is provided.

2.5 Use of ANNs in Hydrology

This section is organised into three main parts. The first examines the early years when ANN papers first started to appear on rainfall-runoff modelling and river flow forecasting. The second section is thematic and discusses the main areas where research is currently focussed while the third highlights areas where further research is needed. This latter section is based heavily on a series of recent review papers that already make a clear and convincing research agenda for this field (Maier et al., 2010; Abrahart et al. 2010).

2.5.1 Early research into ANN rainfall-runoff modelling

ANN research papers began to appear in the scientific literature around the middle of the nineties. One of the first papers to emerge was by Smith and Eli (1995), who used a simple three layer ANN to predict runoff based on simulated rainfall patterns in a synthetic catchment. Although they described their results as “outstanding”, there were actually some problems with peak and time to peak predictions. However, they were encouraged by the results and concluded their paper with a research question regarding whether the configuration of the network and the network weights could have some relation to the physics of rainfall-runoff. Thus, the disadvantage of an ANN as a black box was recognised from the outset of research in this area. At the same time another paper appeared by Hsu et al. (1995) who developed an ANN for the Leaf River Basin in Mississippi, USA. The ANN was compared to an ARMAX model (ARMA model with eXogenous inputs) and the conceptual SAC-SMA (SACramento Soil Moisture Accounting) model. Although the ANN outperformed the other two models, the authors

were quick to point out that ANNs should not be considered as replacements to conceptual models as they are not physically-based. In Italy, Lorrai and Sechi (1995) used an MLP to model rainfall-runoff using rainfall data and temperature for 30 years in the Araxisi catchment in Sardinia. Raman and Sunikumar (1995) applied an MLP and an ARMA model to predict monthly inflows to two study reservoirs. Another pivotal paper by Minns and Halls (1996) involved the development of an ANN to predict synthetic rainfall-runoff data. They showed that the use of two hidden layers in an ANN only produces slightly better results than a one hidden layer ANN, and therefore drew attention to some of the decision making involved in ANN model development.

The latter half of the nineties saw a number of other studies appear in the literature, which could be classified as case studies or as proof-of-concept demonstrators. They generally involved the application of an ANN to a specific catchment, and the ANNs either performed as well as or outperformed other empirical or conceptual models. For example, Dawson and Wilby (1998) used an ANN to forecast flows of the Amber and Mole Rivers in the UK for a lead time of 6 hours, which was found to be comparable in performance to an existing flood forecasting system in operation. This period also saw different networks being used, e.g. a RBFN by Mason et al. (1996), a temporal backpropagation network (TBP-NN) by Sajikumar and Thandaveswara (1999), and a SOM by See and Openshaw (1999). Moreover, the issue of lead times was considered, e.g. Campolo et al. (1999) examined the effect of ANN predictions as the lead time increased from 1 hour ahead to 5 hours, noting the decrease in performance with increasing forecasting horizon. Golob et al. (1998) used an MLP to predict natural water inflow 2, 4 and 6 hours ahead for the Soca River basin. Papers also started to appear that highlighted the lack of guidance available in developing an ANN model, i.e. the guidelines by Dawson and Wilby (1998) specifically for rainfall-runoff modelling, and the empirical modelling by Maier and Dandy (1998a, b) to try and establish some patterns for model development, albeit in the context of water quality. These studies all highlighted the potential opportunities for this ‘new’ technology in a very positive way. This is very much in line with the enthusiasm that accompanies ‘Innovators’ in the Revised Technological Adoption Life Cycle (Moore, 1991), a framework used by Abrahart et al. (2010) to analyse the current position of ANN technology in hydrological modelling.

In the year 2000, two key reviews appeared by the American Society of Civil Engineers Task Force (ASCE 2000a, b) and Maier and Dandy (2000), which targeted broader areas of hydrology. This was followed shortly afterwards with an additional paper by Dawson and Wilby (2001), who reviewed the state of the art in ANN rainfall-runoff

modelling. All three reviews (ASCE, 2000a; Maier and Dandy, 2000; Dawson and Wilby, 2001) provide an introduction to ANNs and then highlight the main issues surrounding model development such as selection of input variables, optimal division of the input data, data pre-processing methods, etc. The reviews by Maier and Dandy (2000) and Dawson and Wilby (2001) also attempted to provide some guidance on model development in the form of steps that should be followed. Maier and Dandy (2000) reviewed 43 papers while Dawson and Wilby (2001) reviewed 50 papers. Both commented on the lack of rigour in many studies, i.e. many of the decisions were made in an *ad hoc* manner, e.g. choice of appropriate model inputs, and the description of the decisions made was either lacking or poorly described. Dawson and Wilby (2001) highlighted the need for more objective model development methods. The second paper (ASCE, 2000b) reviewed applications of ANNs in hydrology, including a section on rainfall-runoff modelling. The review concluded that ANNs can perform as well as existing hydrological models but it also emphasised the lack of an established methodology for model design and implementation. Also, they drew attention to the fact that ANNs are very data intensive.

All three reviews then made recommendations about where future research efforts should be directed. The ASCE (2000b) review ended with five research questions including whether ANNs can be related to physical processes; whether an optimal training dataset can be identified; whether the training process can be made more adaptive; whether ANNs can improve on time series analysis, which also relates to whether the weights have physical meaning; and whether NNs are good extrapolators, i.e. how well can they perform in situations such as an extreme flood event? The review by Maier and Dandy (2000) and Dawson and Wilby (2001) both mentioned the need to extract knowledge from the connection weights, highlighting the importance of opening up the black box. The review by Maier and Dandy (2000) also argued for the need to develop guidelines that assist in the development of ANN models and when ANNs should be used over alternative approaches. In addition they mentioned incorporation of uncertainty into ANN models, while Dawson and Wilby (2001) suggested that rigorous inter-comparison studies are needed, and that error measures should be developed that penalise overly complex models.

2.5.2 Major Themes in ANN Rainfall-runoff Modelling

This section reviews the research that has taken place over the last decade in the area of ANN rainfall-runoff modelling. During this period, a series of key research themes emerged, as described in the sub-sections that follow.

Theme 1: Continuation of Demonstration and Proof of Concept Studies

Although the reviews by ASCE (2000a, b), Maier and Dandy (2000) and Dawson and Wilby (2001) provided a good set of research questions to address, many papers continued to appear which could be classed as demonstration or proof of concept studies (e.g. Rajurkar et al., 2002; 2004; Birikundavyi et al., 2002; Phien and Kha, 2003; Riad et al., 2004; Pan and Wang, 2005; Khan and Coulibaly, 2006; Sahoo and Ray, 2006; Kisi and Cigizoglu, 2007; Kisi, 2008a; Yazdani et al., 2009; Wu and Chau, 2010; Araujo et al., 2011). Some of this research has involved trying out different types of ANNs (e.g. Bayesian ANNs (Khan and Coulibaly, 2006), RBFs (Sahoo and Ray, 2006) and RNNs (Pan and Wang, 2005)) but these papers mainly serve to provide the same general conclusions as early ANN papers, i.e. that ANNs perform similarly or better than the models against which they were compared. However, there was a definite decrease in the number of these types of papers by 2011, which reflects the fact that simple ANN applications are no longer sufficiently advanced to warrant publication in a peer reviewed journal.

Theme 2: Application of Soft Computing Approaches

Soft computing refers to the integration of different artificial intelligence approaches, i.e. ANNs, fuzzy logic, genetic algorithms (GAs), etc., which work together in a synergistic fashion to produce a better result than individual techniques on their own (Zadeh, 1994; See and Openshaw, 1999). There has been a growing trend in applying neuro-fuzzy and neuro-genetic approaches to rainfall-runoff modelling over the last decade.

The review by Maier and Dandy (2000) suggested that neuro-fuzzy solutions are one potential area for further research, and a reasonable amount of work has been reported in the literature. Neuro-fuzzy solutions use the learning capability of ANNs to generate the rules of a fuzzy model and optimise the parameters (Jang et al., 1997). The solutions are also theoretically interpretable. The main type of neuro-fuzzy model used is the Adaptive Neuro-Fuzzy Inference System (ANFIS) (Jang, 1993). One of the reasons it has been chosen is most likely due to its availability in the fuzzy logic toolbox of the Matlab software (Mathworks, 1994-2011). Papers that have applied ANFIS over the last decade for a range of catchments include: Gautam and Holz (2001); Nayak et al. (2004, 2005); Chau et al. (2005); Chen et al. (2006); Keskin et al. (2006); Aqil et al. (2007); El-Shafie et al. (2007); Firat (2008), Firat and Güngör (2007, 2008); Zounemat-Kermani and Teshnehlab (2008); Dastorani et al. (2009); Keskin and Taylan (2009); Mukerji et al. (2009); Pramanik and Panda (2009). As with the proof of concept applications, the results reported in these studies are positive. However, there are also examples of where ANFIS was not the best performing model when compared with

other ANNs such as a RBFN (e.g. Singh and Deo, 2007). Other types of fuzzy applications in which fuzzy logic was directly integrated into an ANN include the work by Nayak et al. (2007), a Counter Propagation Neural-Fuzzy Network (Nie and Linkens, 1994; Chang and Chen, 2001; Chang et al., 2001, 2008) and the dynamic neuro-fuzzy modelling system of Hong and White (2009), which learns online as the forecasting task takes place. The dynamic system outperformed a regular ANN and an ANFIS fuzzy model in forecasting flows at Waikoropupu Springs in New Zealand. Unfortunately all the papers have one thing in common: they do not attempt to interpret the rules or fuzzy sets in a physical way, despite the fact that Maier and Dandy (2000) originally envisaged the use of neuro-fuzzy models in this way. In this sense neuro-fuzzy models are as black box as regular ANNs.

Neuro-genetic approaches combine the optimisation ability of genetic and evolutionary algorithms with ANNs (Kitano, 1992). Genetic approaches are already commonly used in conceptual model calibration (Nicklow et al., 2010). Normally an ANN model structure is chosen and an iterative learning algorithm is used to adjust the connection weights. Neuro-genetic approaches can be used to find the weights of an ANN or they could be used to evolve the whole structure, e.g. determine the number of hidden nodes, the number of inputs, etc. A number of studies have used a GA to determine the starting weights and then trained the ANN further with more conventional ANN training algorithms such as backpropagation or conjugate gradient descent (Whitley et al., 1990; Shamseldin and O'Connor, 2001; Jain and Srinivasulu, 2004a; Parasuraman and Elshorbagy, 2007; Chen and Chang, 2009; Mukerji et al., 2009; Sedki et al. 2009). As with ANFIS, these studies reported improved performance with the ANNs optimised by a GA compared to ANNs with random initialisation.

Another approach that has been used is the application of evolutionary techniques to optimise the entire network. Examples of ANN rainfall-runoff models bred for the River Ouse in England were undertaken by Dawson et al. (2006b), Abrahart et al. (2007b, c) and Heppenstall et al. (2008). All of these studies used the Symbiotic Adaptive NeuroEvolution (SANE) algorithm (Potter, 1997; Moriarty and Miikkulainen, 1998), in which partial ANN solutions are evolved that cooperate together to breed the best overall ANN. Some of the advantages of using this approach were flexibility in changing the objective function and good performance when compared to ANNs trained in the conventional way.

Other examples of using a GA have been to determine the optimal training set (Bowden et al., 2002; Kamp & Savenije, 2006) and to find a set of optimal inputs

(Bowden et al., 2005a), both of which were raised as areas that need further research (ASCE 2000b; Maier and Dandy, 2000).

Finally Chidthong et al. (2009) built a neuro-fuzzy-genetic system in which the ANN was used to find the parameters of a fuzzy model, where the fuzzy rules were then further optimised by a GA. They tested their model to predict floods in Thailand and Japan. The neuro-fuzzy-genetic system outperformed the neuro-genetic model and provided the best peak prediction while the ANFIS model also performed well.

Theme 3: Modularisation and Ensemble Modelling

In both modularisation and ensemble modelling in the context of ANNs, multiple networks are developed and then combined. The difference between these two approaches revolves around the presence of redundancy (Sharkey, 1999). In ensemble modelling, redundant networks (or those which do the exact same modelling task) are developed while in modularisation, NNs are developed on different components of a system, which are then combined or fused together.

Modularisation has been used in a number of ANN rainfall-runoff modelling studies (e.g. Zhang and Govindaraju, 2000a, b). Minns and Hall (1996) observed that ANNs cannot predict both low and high flow events satisfactorily because different parts of the hydrograph are dominated by different processes. To resolve this problem modularisation was used, i.e. the input-output dataset was split into groups and then each group or sub-set of the data was trained using a separate ANN (Solomatine et al., 2008). For example, See and Openshaw (1999) used a SOM to first classify river level data into events, e.g. low river levels, the rising limb of the hydrograph, the falling limb, etc. and they then trained individual ANNs to predict these events separately, combining them at the very end. A similar exercise was undertaken by Abrahart and See (2000) and Jain and Srinivasulu (2006) in which the hydrograph was decomposed into parts and modelled separately. These approaches have also been referred to as Modular Neural Networks (MNN) in the literature (Wang et al., 2006). Wang et al. (2006) compared three MNNs: a Threshold-based ANN (TANN), a Cluster-based ANN (CANN) and a Periodic ANN (PANN) to predict flows of the Yellow River, China, against a classical ANN. For the TANN, the data were divided by thresholds and a separate ANN was built on the subsets while the data were divided into groups for the CANN using fuzzy c-means. Finally the data were divided by seasons for training the PANN. The results showed that the PANN performed better than the other modular and non-modular approaches.

Another example of modularisation is the use of ANNs to predict conceptual errors in a physically-based model (Toth and Brath, 2002; Abebe and Price, 2004). The ANN acts as one module of a larger system. Toth and Brath (2002) developed an ANN to update the predictions of a conceptual model and have also used ANNs to predict the precipitation that was then fed into a conceptual model. Another example is the combination of statistical models with ANNs, e.g. using an autoregressive model to predict flow and an ANN to predict the AR model errors and vice versa (e.g. Xiong and O'Connor; 2002; Anctil et al., 2003).

It is also possible to take an ensemble modelling approach and train many instances of an ANN on the input-output dataset and then combine these through data fusion (e.g. See and Abrahart, 2001; Abrahart and See, 2002). Alternatively, different model types have also been combined. See and Openshaw (2001) integrated an ANN, a fuzzy logic model, an ARMA model and persistence to create a better overall forecast, while Coulibaly et al. (2005) combined a conceptual model, an ANN and a nearest neighbour model, which were first developed to solve the same problem. The models were then weighted and combined to obtain a better result than using any of the individual models. Ensemble models and modularisation both represent interesting areas for further research in ANN rainfall-runoff modelling However, they are much more intensive in terms of development.

Theme 4: Pre-processing

It is recommended that the input and output variables are standardised before the network is trained. This is often a linear standardisation between the ranges 0.1 to 0.9 or 0.2 to 0.8 (Maier and Dandy, 2000). Unlike statistical models, the data provided to the network do not need to be normally distributed. There are some early examples of where the data have been pre-processed in a statistical manner, e.g. differencing of the data (See and Openshaw, 2000; Abrahart and See, 2002) or use of moving averages (Shamseldin, 1997; Wu et al., 2009, Chaipimonplin et al., 2010). However, the most recent trend in data pre-processing has been in the application of wavelet analysis. Wavelet analysis decomposes a time series into scale independent sub-components or wavelets (Nason and Von Sachs, 1999). The decomposed time series are then provided as inputs to the ANN and recombined to produce a final forecast. The trend in this theme is somewhat similar that of theme 1, i.e. the pre-processing technique was applied and the main conclusions showed that the use of wavelets outperformed the use of ANNs without this pre-processing operation. For example, Rao and Krishna (2009) also found that wavelet analysis with ANNs was better than pure ANN models in simulating daily streamflow and monthly groundwater levels. The same findings are

echoed in Anctil and Tape (2004), Kisi (2008; 2009; 2010), Adamowski (2010), Zhou et al. (2008), Nourani et al. (2009), Rahanam and Noury (2009), Partal (2009) and Wang et al. (2009) for simulating stream flows. Another technique that has appeared even more recently in ANN rainfall-runoff modelling is singular spectrum analysis (SSA), which decomposes the input time series based on a number of components (e.g. trends, periodicities and noise). Wu et al. (2009) found that the SSA enhanced the ANN performance better than wavelets when forecasting the daily discharge of two tributaries of the Yangtze River. SSA in combination with an ANN also performed better when compared to a modular ANN (Wu and Chau, 2011).

Theme 5: Opening up the Black Box

All three review papers (ASCE, 2000b; Maier and Dandy, 2000; Dawson and Wilby, 2001) argued that further research is needed in opening up the black box of the ANN, i.e. to find physical interpretations or meaning in the hidden nodes or weights of the ANN. Rule extraction was the basis for early work in this area (including that outside of hydrology) to try to understand ANN behaviour (Andrews et al. 1995; Benitez et al. 1997; Kingston et al., 2006). Saliency analysis (Abrahart et al., 1999) and sensitivity analysis (Sudheer, 2005) also provided insights into what network inputs were important and how they affected the behaviour of the network. However, the real first attempt at directly examining the behaviour of the hidden nodes was undertaken by Wilby et al. (2003), who developed ANNs for the Test River Basin in the UK. The ANNs were trained to learn the outputs from a calibrated conceptual model. The outputs from the hidden nodes were then plotted against the components of the conceptual model. The results showed that two of the hidden nodes appeared to be capturing baseflow and quickflow when antecedent precipitation and evaporation were included as inputs, while a third hidden node appeared to have some relation to the soil moisture deficit. Similar work followed (Jain et al., 2004b; Sudheer and Jain, 2004; See et al., 2008). The study by Jain et al. (2004b) for the Kentucky River and See et al. (2008) for the Ouse River, UK, both found differentiation of hidden nodes by process, i.e. baseflow, quickflow and infiltration, implying a consistent pattern across catchments. Sudheer and Jain (2004) plotted raw hidden unit outputs against discharge and found that the hidden nodes appeared to correspond to the generation of low, medium and high flows, respectively, for the Narmada River in India. These reported studies show that some progress has been made during the last decade but there is quite clearly a great deal of work still to be done, especially if ANN methods are to become more accepted by hydrologists in the future, particularly for operational forecasting.

Theme 6: Input Variable Selection

Input variable section (as discussed in section 2.3) is one of the decisions that must be

made as part of the development of an ANN. Not all papers report this information systematically. However, trial and error or an *ad hoc* approach is one of the mostly commonly used ways to determine the inputs. This method involves providing the NN with different sets of input variables (e.g. previous flows, rainfall) that are perceived to have a relationship with the outputs (e.g. runoff). Maier et al. (2010) determined that 37 out of 210 papers published between 1999 and 2007 used an *ad hoc* approach. Evidence of the continuing use of this approach can be seen in discharge forecasting on the Huaihe River in China (Li et al., 2009) and by Partal (2009) in the development of different river flow forecasting models. The main problem with this method is that many different combinations must be tried to ensure that an acceptable set of inputs is found.

Although not used often, sensitivity analysis is another method that allows input variables that have little or no effect on the outputs to be removed. An example is the work by Sudheer (2005), who used this approach to show the effect of different inputs on the shape of the hydrograph when modeling runoff on the River Narmada in India. A related approach is referred to as saliency analysis, where a network is first trained and then one input variable is removed at a time, with an analysis of the output after each removal to determine which inputs had little effect on the overall results. Abrahart et al. (2001) applied this approach in developing an ANN rainfall-runoff model and found that it was possible to gain useful knowledge about a number of relevant inputs including previous flow values, rainfall, seasonality, etc.

One of the more frequently used methods is correlation analysis, which was used in 60 out of 210 papers reviewed by Maier et al. (2010). The method provides a way of drastically reducing the number of inputs by eliminating those variables with less than a certain correlation value with the output. Examples of work that have used this approach include Dawson et al. (2006a), Kim et al. (2009) and Jia et al. (2009). The problems with this approach are that: (i) it assumes a linear relationship between the variables, where most hydrological problems are non-linear; (ii) it is unclear which threshold to use to determine whether variables should be included or not; and (iii) it does not take variable independence into account. One method of dealing with the non-linearity and independence of input variables is through the use of Partial Mutual Information (PMI) (Sharma, 2000; Sharma et al., 2000). Although in the context of water quality, Bowden et al. (2005a, b) used the PMI in ANN modelling of the River Murray in South Australia. There is evidence that the PMI is starting to be used, e.g. Hejazi and Cai (2009) and Corzo et al. (2009). Hejazi and Cai (2009) noted improved performance when using the PMI over other input determination methods while

modeling reservoir release in California, while Corzo et al. (2009) used the PMI to select the input variables for an ANN that was used to replace both process-based models and the routing component for the River Meuse catchment. However, the PMI was the only method used so no comparisons with other methods were provided.

Pruning algorithms have also been employed as a method for determining the inputs. Pruning removes unimportant or weak connections between nodes as well as the nodes themselves (Bishop, 1995). The concept behind this algorithm is to start with a fully connected network and to then remove the least significant connections between the inputs and outputs. Abrahart et al. (1999) found that pruning algorithms reduced the total number of connections between 10%-43% whilst retaining good model performance in the development of ANN rainfall-runoff models. Corani and Guariso (2005a) used pruning algorithms to initially reduce the number of inputs by 30 to 40% before training again with an ANN for two catchments in Italy. The authors found that the ANNs trained on the smaller number of inputs generalised better than ANNs developed using all the initial inputs. It is surprising that so few examples exist given the potential of this technique. This may be due to the fact that pruning software is not as readily available as ANN software.

Another method that has potential is a GA, but once again, there are very few examples, e.g. Anctil et al. (2006), who used a GA to determine which rain gauges to include in an ANN rainfall-runoff model. Input determination is clearly an area where methods of optimization will prove valuable.

Theme 7: Other Research

There have also been a number of isolated studies covering a range of topics. Two particular issues of interest that fall outside of the other themes are discussed below.

The first issue is to do with the extrapolation issue. The ASCE (2000b) review asked whether ANNs are good extrapolators. The question is relevant to ANNs as a rainfall-runoff model may be developed over a range of flows and then an extreme event may come along that is larger than any event seen before. Some research has been undertaken to look at this issue. Imrie et al. (2000) modified the Cascade Correlation training algorithm to use a cubic polynomial function in the output layer, so that the ANN could extrapolate outside the range of the training data. Cigizoglu (2003) used an autoregressive model to generate a synthetic time series and used these as inputs to the ANN, thereby improving its ability to reproduce extreme flows more effectively than

training with the observed data. Problems with extrapolation are, however, a function of all empirical models (ASCE, 2000b).

Another area of research has been to incorporate physical parameters into the ANN or add hydrological knowledge explicitly into the development of the model, which is somewhat different to the research undertaken in theme 5. For example, Zhang and Govindaraju (2003) developed an ANN that takes the basin geomorphology into account. The weights between the input nodes and the hidden nodes were equal to the coefficients used in a unit hydrograph, which was an attempt to relate these connections to the number of rivers that connect to the main channel. Pan and Wang (2005) similarly associated the weights of the ANN with Markov parameters of the unit hydrograph. Jain and Indurthy (2003) considered the time of concentration of the basin to determine the number of variables of antecedent rainfall that would go into the model. Jain and Srinivasulu (2004a) developed an ANN that could incorporate conceptual elements by modelling the process of infiltration through the Green-Ampt equation, and by modelling the soil moisture content, evaporation and flow using conceptual techniques. They also estimated the effective rainfall at each time interval to use as a model input to the ANN. Effective rather than total rainfall was also incorporated in previous studies by Sajikumar and Thandaveswara (1999).

The final area to be discussed in this section is the research that has been undertaken on uncertainty. This area is very important yet there is surprisingly little work, which may simply reflect the complexity of this subject area. A recent example is the work by Srivastav et al. (2007), who proposed an approach to determine uncertainty in ANN hydrological models. They used a bootstrapping procedure (resampling with replacement) in which 300 networks were trained to calculate uncertainty bands. They concluded that using performance measures alone does not provide model confidence. Han et al. (2007) proposed two approaches for measuring uncertainty in ANN rainall-runoff modelling. The first method considers the distance between the training data and the predictions. The second method involved looking at responses to see whether the ANN performed in a hydrologically sound way. More recently, Alvisi and Franchini (2011) proposed an ANN rainfall-runoff model with fuzzy weights and biases. The forecast is not deterministic but provides a range or interval of prediction at each time step.

Uncertainty is a subject that has been dealt with more extensively in more tradiational hydrological modelling. In particular Montanari (2011) has provided a comprehensive overview of the uncertainty, recognising two main types, i.e. global, which refers to the

overall uncertainty between the observations and the model output, and individual, which refers to the individual sources of uncertainty. Examples include model structural uncertainty, input uncertainty, etc. A very useful table of the most commonly used uncertainty methods and the type of uncertainty estimates is provided in Montanari (2011) and is reproduced in Figure 2.4.

Table 1 Uncertainty assessment methods in hydrology, along with their classification (see Section 2.17.5) and purpose (see Sections 45.6–45.10)

Assessment method	*Classification*	*Type of uncertainty estimated*
AMALGAM	Nonprobabilistic, parameter estimation	Parameter
BATEA	Probabilistic, parameter estimation, uncertainty assessment, sensitivity analysis	Precipitation induced
BFS	Probabilistic, Bayesian	Global
BMA	Probabilistic, multimodel	Global
DYNIA	Nonprobabilistic, identifiability analysis	Parameter
GLUE	Nonprobabilistic (when an informal likelihood is used), parameter estimation, uncertainty assessment, sensitivity analysis	Global, parameter, data, structural
IBUNE	Probabilistic, parameter estimation, uncertainty assessment, sensitivity analysis	Global, precipitation induced, model structure induced
Machine learning	Nonprobabilistic	Usually global, in principle all
Meta-Gaussian	Probabilistic, data analysis	Global
MOSCEM-UA	Nonprobabilistic, parameter estimation, sensitivity analysis	Parameter
SCE-UA	Probabilistic, parameter estimation	Parameter

Classification is ambiguous in some cases; it distinguishes between probabilistic and nonprobabilistic methods, as well as among the seven categories introduced by Matott *et al.* (2009) (see Section 2.17.5.2).
AMALGAM, a multialgorithm genetically adaptive method for multiobjective optimization; BATEA, Bayesian total error analysis; BFS, Bayesian forecasting system; BMA, Bayesian multimodel analysis; DYNIA, dynamic identifiability analysis; GLUE, generalized likelihood uncertainty estimation; IBUNE, integrated Baysian uncertainty estimator; MOSCEM-UA, multiobjective shuffled complex evolution University of Arizona; SCE-UA, shuffled complex evolution university of Arizona. References for the methods are in the text.

Figure 2.4: Methods and tyes of uncertainty. Taken from: Montanari (2011, p.464).

ANNs fall under ‘Machine learning’ where the type of uncertainty estimated is usually global. However, other types of uncertainty estimates are possible. A salient point was also raised by Montanari (2011). He stressed the need to explain the methodology very clearly for whatever approach is used to evaluate uncertainty so that the results are understandable to the scientific community.

2.5.3 Future Research Areas in ANN Rainfall-Runoff Modelling

In the year 2010, two more reviews appeared. Maier et al. (2010) examined 210 papers published between 1999 and 2007 in the area of water resources. The ANN model development process was then divided into components, e.g. input variable selection methods, data division methods, type of ANN, type of training algorithm, etc. and the papers were systematically analysed to determine which types of methods were used in model development. This allowed the reviewers to look at the most commonly used methods and where improvements could be made. Abrahart et al. (2010) took a different approach. They analysed the number of papers that were published each year in a number of hydrological fields, which clearly showed an increasing number over the last two decades. Roughly 400 papers were published on rainfall-runoff modelling linked to some aspect of ANNs. The authors considered the reasons why ANNs have

not been accepted widely in the hydrological modelling community, using the Revised Technological Adoption Life Cycle (RTALC) (Moore, 1991) as a framework to show how technologies evolve. They argue that the next major challenge is to move ANNs from Innovators to Early Adopters, and after reviewing the main themes in the literature that have emerged over the last two decades, provided a series of recommendations for how to move to the Early Adopter stage. The main areas of relevance to rainfall-runoff modelling that were highlighted by both reviews include:

- **Input variable selection:** The main methods used to date have generally been trial and error and linear correlation. More research is needed into non-linear approaches that take variable independence into account like that of Sharma et al. (2000), Sharma et al. (2000), May et al. (2008) and Fernando et al. (2009).

- **Rigorous inter-comparison studies:** There are currently too many studies that report the results in isolation, e.g. a network of a particular type functioned well on a specific catchment with a particular set of network parameters and a specific dataset. Instead, rigorous inter-comparison studies like the Distributed Model Intercomparison Project (DIMP) (Smith et al., 2004) or the LUCHEM (Land Use Change on Hydrology by Ensemble Modeling) study (Breuer et al., 2009) need to be established. Then it will be possible to develop the type of "collective intelligence" on ANN modelling referred to by Abrahart et al. (2010, p.327).

- **Finding physical meaning in ANNs:** The black box nature of ANNs is a problematic issue when it comes to acceptance of the methodology within the hydrological community. More research needs to be undertaken in this area.

The above listed areas have not been dealt with specifically in this thesis. However, there were two other suggested areas for further research, which are relevant to the thesis:

- **Incorporation of uncertainty:** As mentioned in theme #7, there has been little reporting of uncertainty in ANN rainfall-runoff modelling. Parameter uncertainty is explicitly addressed in the thesis through the calculation of confidence intervals around model predictions using the QR-based additive error model, which is undertaken in Chapter 5. The uncertainty surrounding ANN weight initialisation is also examined in detail in Chapter 6.

- **Further research into hybrid and ensemble methods:** More research is needed in developing hybrid and ensemble methods in hydrological modelling as they hold a great deal of promise in the future. Ensemble methods are dealt with in this thesis in Chapters 5 and 6.

Two other areas were examined in this thesis. The first is the need to compare ANNs with more physically-based models, which is an issue that was raised in the review by Dawson and Wilby (2001). Although there are a number of examples in the literature (as mentioned in relation to the different themes), the majority of papers do not compare ANNs with physically-based or conceptual models. Many studies have used other types of ANNs for comparison or other empirical models. If ANNs are to become more accepted in the traditional hydrological community, a more rigorous comparison with conceptual and physically-based models must be undertaken.

Finally, an area that has been discussed in the broader hydrological literature is the evaluation of hydrological measures, i.e. which performance measures to use in order to evaluate a model (e.g. ASCE, 1993; Legates and McCabe, 1999; Dawson et al., 2007). However, it is an area that has not been extensively dealt with in the ANN rainfall-runoff modelling literature. This thesis will review the performance measures available (Chapter 3) and apply a comprehensive set to the models developed (Chapters 4 to 6).

2.6 Summary

This chapter has provided an overview of the literature on ANNs with particular emphasis on rainfall-runoff modelling. ANNs were first placed within the typology of hydrological modelling methods followed by an introduction to the basics of this technology. This included advantages of ANNs for hydrological modelling as well as the disadvantages. A review of the ANN rainfall-runoff literature was then presented in the form of core themes followed by recommendations for further research, most of which were derived from recent review papers on ANNs in hydrological modelling (Maier et al., 2010; Abrahart et al., 2010). Uncertainty and ensemble modelling were two areas suggested for further research and are addressed in this research. Two additional areas dealt with in the thesis are: a) a greater emphasis on comparison with more physically-based or conceptual models; and b) a critical look at the performance measures used to evaluate ANN models, which are reviewed in the next chapter.

Chapter 3

Performance Measures for Model Evaluation

3.1 Introduction

The performance of hydrological models is commonly assessed by computing a number of measures of performance or goodness-of-fit statistics. This also holds true for ANN rainfall-runoff models. Several performance criteria are described in the literature along with their merits and shortcomings (Legates and McCabe, 1999; Jachner et al., 2007; Dawson et al., 2007; Reusser et al., 2009). Many authors argue that it is important to apply a number of indices, which should be chosen according to the particular needs of each individual application (Dawson et al., 2007). Moreover, it is suggested that the adopted criteria should not be redundant (Gupta et al., 1998; Reusser et al., 2009) and should be sensitive to different types of errors (e.g. errors in peak prediction, errors in timing of the hydrograph prediction, etc.). However, often only redundant indices, which are linked to each other such as the coefficient of determination, the sum of errors squared and the normalised root mean squared error, are applied. Moreover, the 'best' model is usually selected by comparing the values of the performance criteria and sorting the models according to their score without a comparison with a benchmark model and a formal assessment of the significance of the differences, even though these good practices are recognised in the forecasting and hydrological literature (e.g. Makridakis et al., 1998; Siebert, 2001; Brath et al., 2002; Hyndman and Koehler, 2006; Schaefli and Gupta, 2007; Moussa, 2010).

This chapter provides an overview of a range of measures available and how they are computed. These measures can be characterised by those which are absolute (i.e. expressed in the units of the output variable, e.g. metres) or relative (i.e. dimensionless or expressed as a percentage). The relative measures can be further broken down into those which use a reference or benchmark model for comparison (Hyndman and Koelher, 2006), e.g. comparison to the mean. In addition, two measures from economics will be reviewed that may have potential merits for assessing model performance in hydrology. From this review, a set of measures has been chosen that will be used in subsequent chapters of the thesis. In the equations used throughout this chapter, x is the observed series, $\hat{x}$ is the model forecast series, $\bar{x}$ is the mean of the observed series, $\tilde{x}$ is the mean of the forecast series and $\ddot{x}$ is the naïve forecast.

3.2 Absolute Performance Measures

Absolute metrics provide an idea of the absolute differences between observed and

modeled values in the original units of measurement. One of the most commonly calculated measures is the Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum|x - \hat{x}| \tag{3.1}$$

The MAE has no upper limit where 0 indicates a perfect fit between the observed and predicted values. Examples of papers that have used the MAE to assess ANN rainfall-runoff model performance include those by Cannas et al. (2006) and Dawson et al. (2006b).

Since the absolute value of the deviations is used in the MAE, a related measure that takes the sign of the deviations into account is the Mean Error (ME), calculated as:

$$ME = \frac{1}{n}\sum(x - \hat{x}) \tag{3.2}$$

Similar to the MAE, the ME has no upper limit and 0 indicates a perfect fit. Since the ME assigns equal weights to small and high values, this measure can be used to determine possible biases since it is a signed metric. However, a low value of ME may also indicate a situation where the over and under predictions have effectively cancelled each other out and therefore this measure should be used in conjunction with others. Many papers in the past have used a variation of this performance measure, i.e. the square of the deviations (e.g. Karunanithi et al., 1994; Raman and Sunilkumar, 1995; Cigizoglu, 2005; Sahoo and Ray, 2006; Sahoo et al., 2006; Leahy et al., 2008; Partal, 2009).

The MdAE (Median of the Absolute Errors) is simply the median of the absolute deviations between the observed and predicted values and provides information on the distribution of the deviations:

$$MdAE = median\,(|x - \hat{x}|) \tag{3.3}$$

This measure does not appear to have been used in previous research involving ANNs and rainfall-runoff modeling with the exception of Napolitano et al. (2011).

Another commonly reported absolute error measure is the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum(x - \hat{x})^2}{n}} \tag{3.4}$$

As with the MAE and the ME, the RMSE has no upper limit and 0 indicates a perfect fit. As the differences between the observed and predicted values are squared, this measure penalises prediction errors in high flow events compared to low flows, as high flows are generally where the greatest error in model prediction occurs. The RMSE tends to be slightly larger than the MAE where the magnitude of this difference can be used to indicate the extent of outliers in the data (Legates and McCabe, 1999). Many studies can be found that have utilised this measure in assessing ANN rainfall-runoff models (e.g. Smith and Eli, 1995; Dawson and Wilby, 1998; Campolo et al., 1999; Dawson et al., 2000; Corani and Guariso, 2005b; Kumar et al., 2005; Cannas et al., 2006; Dawson et al., 2006b; Sahoo and Ray, 2006; Sahoo et al., 2006; Chidthong et al., 2009; Hejazi and Cai, 2009; Remesan et al., 2009). To allow for comparison of RMSE across different variables or between different catchments, it is possible to normalise this measure by dividing it by the mean of the observed values over the modelling time period as implemented, e.g. by Jain and Srinivasulu (2004a, b).

A variation of the RMSE is the Fourth Root Mean Quadrupled Error (R4MS4E):

$$RMSE = \sqrt[4]{\frac{\sum(x - \hat{x})^4}{n}} \tag{3.5}$$

Similar to the RMSE, it places even higher weight on the largest deviations and therefore penalises models even more than the RMSE for errors in high flow events. This measure was used by Cannas et al. (2006) as one of several measures for the evaluation of different data-driven forecasting models of monthly flows in Sardinia. A similar measure was used by Abrahart and See (2000) in comparing ANN rainfall-runoff models for the Ouse and Wye Rivers in the UK.

A less often reported metric is the Peak Difference (PDIFF):

$$PDIFF = \max(x) - \max(\hat{x}) \tag{3.6}$$

which calculates the highest value predicted in the model and subtracts that from the highest value recorded in the observed dataset. Unlike the other measures it does not attempt to represent the overall level of agreement between the observed and predicted data nor does it consider the temporal relationships between the highest

values. For example, if the dataset is continuous, then the maximum values might be calculated from different flood events. If a single event is being considered, then this measure has potentially more value. However, this measure does indicate whether the model is producing values similar to what is seen in the observed data set. A number of NN rainfall-runoff models have been evaluated using this measure (e.g. Chang and Hwang, 1999; Kerh and Lee, 2006; Chaipimonplin et al., 2010).

3.3 Relative Performance Measures

Relative errors (also known as percentage metrics) introduce a scale check, accounting for the fact that a difference of two, for example, has a much larger impact if the observed value is two rather than 100 (Dawson et al., 2007; Villarini et al., 2008). Four relative measures that correspond directly to the first four absolute measures are the Mean Absolute Percentage Error (MAPE), the Mean Percentage Error (MPE), the MdAPE (the Median Absolute Percentage Error), and the Root Mean Squared Percentage Error (RMSPE), which are calculated as follows:

$$MPE = \frac{1}{n}\sum \frac{(x - \hat{x})}{x} \times 100 \tag{3.7}$$

$$MAPE = \frac{1}{n}\sum \left|\frac{x - \hat{x}}{x}\right| \times 100 \tag{3.8}$$

$$MdAPE = median\left(\left|\frac{x - \hat{x}}{x}\right|\right) \times 100 \tag{3.9}$$

$$RMSPE = \sqrt{\frac{1}{n}\sum \left(\frac{x - \hat{x}}{x}\right)^2} \times 100 \tag{3.10}$$

In addition to these relative measures, another commonly reported measure is the 'Coefficient of Determination' or R-squared (Pearson, 1896):

$$R^2 = \left[\frac{\sum(x - \bar{x})(\hat{x} - \tilde{x})}{\sqrt{\sum(x - \bar{x})^2 \sum(\hat{x} - \tilde{x})^2}}\right]^2 \tag{3.11}$$

The R^2 indicates how much variance is explained by the model. A value of 0 indicates no explanation while 1 is a perfect fit. A number of problems have been identified with this measure as outlined by Legates and McCabe (1999), e.g. high values of goodness of fit can result even when the model is flawed, it is insensitive to the means and variances in the data, and it is very sensitive to outliers. Despite these issues, the R^2 has been used by a number of authors in evaluating their ANN rainfall-runoff models (e.g. Lorrai and Sechi, 1995; Campolo et al., 1999; Dawson and Wilby, 1999; Corani

and Guariso, 2005b; Kumar et al., 2005; Cannas et al., 2006; Dawson et al., 2006b; Kerh and Lee, 2006; Hung et al., 2009; Mukerji et al., 2009).

An improvement over the R^2 statistic and one that has been used very commonly in hydrology (Legates and McCabe, 1999) is the Coefficient of Efficiency (CE) developed by Nash and Sutcliffe (1970):

$$CE = 1 - \frac{\sum(x - \hat{x})^2}{\sum(x - \bar{x})^2} \tag{3.12}$$

The CE generally ranges from 0 to 1 (perfect fit) although negative values are possible. A value of 0 indicates that the model is no better than simply forecasting the mean value. The CE has been commonly reported as a measure used to evaluate ANN rainfall-runoff models (e.g. Minns and Hall, 1996; Chang and Hwang, 1999; Dawson et al., 2006a; Kerh and Lee, 2006; Leahy et al., 2008; Yang and Chen, 2009).

Despite the popularity of CE, Beran (1999) argues that there are better baselines against which model performance should be compared such as persistence or seasonal averages. An example is the Coefficient of Persistence or the Persistence Index (PI) as outlined in the paper by Kitanidis and Bras (1980):

$$PI = 1 - \frac{\sum(x - \hat{x})^2}{\sum(x - \ddot{x})^2} \tag{3.13}$$

The PI has strong similarities to the CE but instead of the observed mean, the last observed record (or the naïve forecast) is used for the purposes of model comparison. If the PI is 0, then the forecasting method performs as well as the naïve forecast. If the PI is greater than 0 up to a value of 1 (perfect fit), then the forecasting method performs better than the naïve model. The higher the value of the PI, the better the model is compared to the naïve forecast. However, if the PI is less than 0, then the forecasting method performs worse than the naïve forecast. There are not many examples of the use of PI although the Hydrotest system of Dawson et al. (2007) automatically calculates this measure. Anctil et al. (2006) used this index to evaluate ANN rainfall-runoff models with differing numbers of rain gauge inputs.

It is also possible to add persistence to the MAE and MdAE measures resulting in:

$$PI.MAE = 1 - \frac{\sum|x - \hat{x}|}{\sum|x - \ddot{x}|} \tag{3.14}$$

$$PI.MdAE = 1 - \frac{median\sum|x - \hat{x}|}{median\sum|x - \ddot{x}|} \tag{3.15}$$

The final measure reviewed in this section is the geometric reliability index (GRI), which measures the accuracy of the simulation within a multiplicative factor (Leggett and Williams, 1981):

$$GRI = \frac{1+Q}{1-Q} \quad where \quad Q = \sqrt{\frac{1}{n}\sum\left(\frac{\hat{x}-x}{\hat{x}+x}\right)^2} \tag{3.16}$$

If the GRI is the value assumed by the index, the observed values fall between 1/GRI and GRI times the corresponding predicted values (Jachner et al., 2007). The GRI measures how wide a cone would be in order to contain the data (Figure 3.1), where small errors would be expected (in absolute terms) for small values, and bigger errors when the absolute value increases.

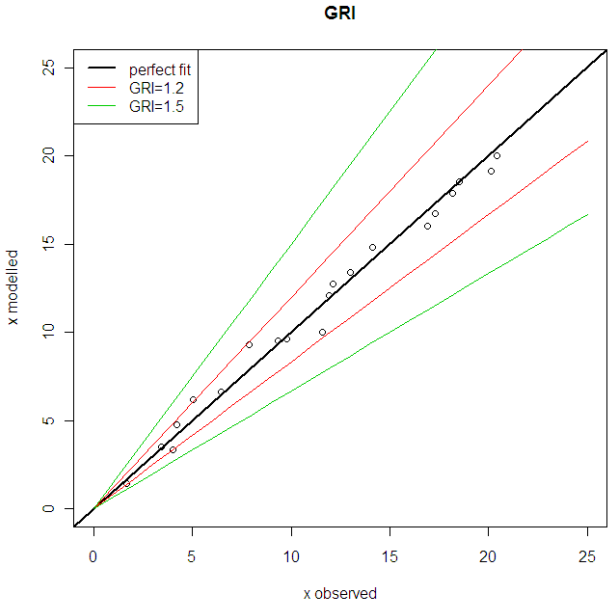


Figure 3.1: Visualisation of the GRI cones around the best fit line

There is no evidence that this measure has been used in the evaluation of ANN rainfall-runoff models with the exception of Napolitano et al. (2011).

3.4 Two Error Measures from Economics

It is also possible to apply formal tests to assess whether two models have equal accuracy, e.g. an ANN compared to a conceptual model or another type of data-driven model. There are two statistical tools which allow for this type of testing that are well-known in econometrics, but have not yet been utilised in hydrology (Laio and Tamea, 2007). These are the sign test (Lehmann, 1975) and the Diebold-Mariano test (Diebold and Mariano, 1995). The details of these tests are provided below.

Let M1 and M2 denote two models to be compared. Two tests are considered to assess if M1 outperforms M2 or vice versa. Both tests are based on the concept of loss-differential:

$$d(t) = g[x(t) - \hat{x}_{M1}(t)] - g[x(t) - \hat{x}_{M2}(t)] \tag{3.17}$$

where g is a function, which is commonly assumed to be $g(\cdot) = |\cdot|$ or $g(\cdot) = (\cdot)^2$. The "equal accuracy" null hypothesis is equivalent to the null hypothesis that the population mean of the loss-differential series is 0. The first test is a classical finite sample sign test (Lehmann, 1975). Assuming that the loss-differential series is independent and identically distributed (iid), the number of positive loss-differential observations in a sample of size n has the binomial distribution with parameters n and 0.5 under the null hypothesis (Diebold and Mariano, 1995). The sign test statistic is therefore:

$$S_2 = \sum_{i=1}^{n} I_+ (d(t)) \tag{3.18}$$

where $I_+\big(d(t)\big) = 1$ if $d(t) > 0$, and otherwise is equal to 0. Note that S2 is insensitive to the choice of g. In large samples, the Studentized version of the sign-test statistic is asymptotically standard normal:

$$S_{2a} = \frac{S_2 - 0.5n}{\sqrt{0.25n}} \sim N(0,1) \tag{3.19}$$

A negative value for the S_{2a} statistic smaller than the standard normal threshold $z_{P=0.025} = -1.96$ indicates that the forecast generated by the M1 model is closer to the observed value than the forecast generated by the M2 model more often than expected by random chance, and it can be concluded that the M1 model generates a more accurate forecast than the M2 model, with a 95% confidence level in a two-sided test. Since loss-differential is commonly serially correlated, the sign test could give biased

results. To avoid this problem, the sign test is performed on subsamples selected by taking loss-differential values separated by a given number of time steps k: $\{d(1), d(1+k), d(1+2k), \ldots\}, \ldots, \{d(k), d(2k), d(3k), \ldots\}$. As the resulting samples are serially independent, a test with size $\propto$ can be obtained by performing k tests, each of size $\propto/k$, on each of the k loss differential sequences and rejecting the null hypothesis if the null is rejected for any of the k samples, exploiting the Bonferroni inequality (Diebold and Mariano, 1995).

The second test, named the Diebold-Mariano test (Diebold and Mariano, 1995), is based on the asymptotic distribution of the sample mean loss-differential, $\bar{d}$. It explicitly accounts for possible autocorrelation of the loss-differential series. The test statistic is:

$$S_1 = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{n}}} \tag{3.20}$$

where $\hat{f}_d(0)$ is an estimate of the spectral density of the loss-differential $f_d(0)$ at zero frequency. As $f_d(0)$ can be deduced from the Fourier transform of the autocorrelation function, it provides a correction for possible serial correlation of loss-differential. Further details on the estimation of $\hat{f}_d(0)$ can be found in Diebold and Mariano (1995). The S_1 test statistic is asymptotically standard Gaussian. A non-zero value of S_1 indicates that the accuracy of the two simulations can be distinguished statistically. Namely, a negative value for S_1 that exceeds the critical threshold would indicate that the first model (M1) generates a more accurate forecast than the second model (M2)

3.5 Choosing a Set of Measures for Assessing Model Performance

Table 3.1 summarises the performance measures chosen in evaluating the models in this research. There is redundancy between some of the measures, which is intentional. The use of redundant indices will provide a mutual validation of the results as well as stressing the importance of using non-redundant indices. Absolute and relative metrics are complemented with the values corresponding to a naïve forecast to highlight the coherence with the results, which implies a direct comparison with benchmark models.

Table 3.1: Summary of performance measures selected for use in the research

Measure	Acronym	Lower Limit	Upper Limit	No error
Mean error	ME	$-\infty$	∞	0
Mean absolute error	MAE	0	∞	0
Median absolute error	MdAE	0	∞	0
Root mean squared error	RMSE	0	∞	0
Peak difference	PDIFF	$-\infty$	∞	0
Mean percentage error	MPE	$-\infty$	∞	0
Mean absolute percentage error	MAPE	0	∞	0
Median absolute percentage error	MdAPE	0	∞	0
Root mean square percentage error	RMSPE	0	∞	0
Coefficient of efficiency	CE	$-\infty$	1	1
Persistence index	PI	$-\infty$	1	1
PI based on MAE	PI.MAE	$-\infty$	1	1
PI based on MdAE	PI.MdAE	$-\infty$	1	1
Geometric reliability index	GRI	1	∞	1
Sign test	S2	-	-	-
Diebold-Mariano Test	S1	-	-	-

Four absolute metrics have been chosen: ME, MAE, MdAE and RMSE, which will provide an idea of the absolute differences between observed and modeled values in their original unit measures. In particular, since the ME is a signed metric, it can be used to determine possible biases. MAE, MdAE and RMSE, on the other hand, are non-negative metrics. Unlike RMSE, the MAE and MdAE are not weighted towards higher or lower magnitude events. The MdAE was chosen as it is less affected by skewed error distributions than the MAE. The R4MS4E is similar enough in nature to the RMSE so will not be used further in the thesis. PDIFF has flaws for continuous modelling and will therefore only be applied to single event modelling.

The relative error measures chosen include: MPE, MAPE, MdAPE and RMSPE, which provide a direct correspondence to the first four absolute metrics (Equations 3.1 to 3.4), and the same comments apply to these measures as above. Indices of relative differences, which compare the errors from the selected model with respect to those from a benchmark or reference model, have also been chosen. The first two error measures are CE and PI. Both use different benchmark models for comparison, i.e. the mean and persistence. Since both CE and PI are based on squared errors, two further measures are added: PI.MAE and PI.MdAE to account for relative errors that equally weight large and small observations. The values of these similarity measures are upper bounded to one and allow for an easy comparison between formal and naïve approaches.

Finally, in the situation where more than one model type is being compared, and the data are of sufficient length, then the sign test and the Diebold-Mariano test will also be applied.

In addition, the GRI (Leggett and Williams, 1981) is chosen because it has not been used to evaluate ANN rainfall-runoff models before.

3.6 Summary

This chapter provided a review of the performance measures commonly used to evaluate ANNs and other types of hydrological model. From these measures, fourteen were chosen to evaluate single models and two additional measures originating from economics, were chosen when comparing the difference between results from two competing models. Some of the measures are redundant but these have been chosen to see whether they provide a consistent message when applied to these models. In Chapters 4 to 6, these measures are systematically applied and compared for fitness of purpose. Visualisation (or graphical comparison) is also used as an additional aid in model evaluation as recommended by Green and Stephenson (1986).

Chapter 4

Comparison of a Conceptual and ANN Rainfall-Runoff Model of the Tiber River

4.1 Introduction

The starting point for this research is an exploration into the capability of ANNs for rainfall-runoff modelling in a large catchment. To provide a real test of the skill of the ANN, the results from this model are compared to the results from a conceptual model. As highlighted at the end of Chapter 2, this is not an exercise that is carried out very frequently, with most researchers choosing linear regression, time series models or other data-driven models for comparison. This chapter will begin with a description of the characteristics of the River Tiber catchment, the data available for modelling and the main flood events that have occurred recently in Rome. The chapter then presents the conceptual TEVERE model followed by the development of an ANN model at the same site. Both models were then run to predict the 2005 and 2008 flood events in Rome at the Ripetta gauging station. The results of these initial experiments are provided, along with a comparison of the two models.

4.2 The Tiber River Basin

The Tiber River is the third longest river in Italy. It has a catchment area of approximately 17,000 km^2 covering around 5% of the country as shown in Figure 4.1. The Tiber has a length of around 400 km and rises in Emilia Romagna on Mount Fumaiolo at 1268 m above sea level, with a discharge of 10 m^3/s. The average discharge of the river is approximately 300 m^3/s. The Tiber flows from the Apennines through Rome and then into the Tyrrhenian Sea. The average basin elevation is 524 m. The main tributaries on the western side are: the Cerfone, Nestore, Paglia and Treia Rivers. On the eastern side, are the Rivers Chiascio, Nera, Farfa and Aniene.

The entire catchment covers 5 regions of Italy as shown in Table 4.1, with negligible areas in two other regions. The catchment can be subdivided into 3 ungauged areas (marked as 5, 7 and 8 in Figure 4.1) and 6 gauged areas (marked 1-4, 6 and 9 in Figure 4.1). The total ungauged area is approximately 2,750 km^2; Kottegoda et al. (2004) found that this area can be considered as a homogeneous rainfall region. In the basin there are 334 municipalities and a total population of 4.5 million inhabitants, with 80% living in the province of Rome.

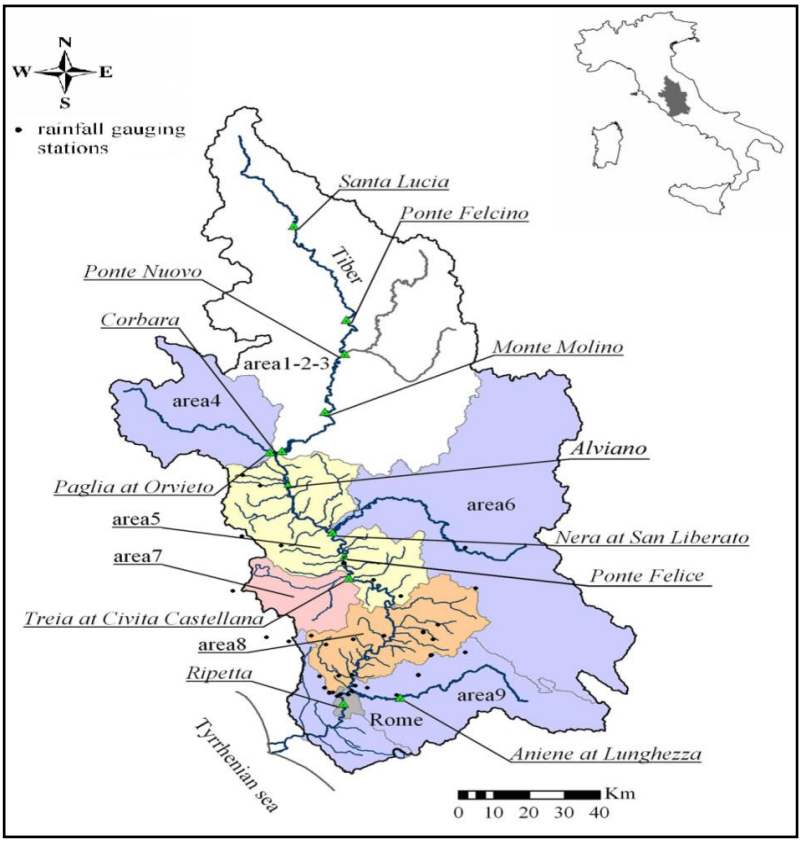


Figure 4.1: The Tiber River Basin

Table 4.1: The percentage area of the Tiber River Basin in each region in Italy

Regions in the Tiber Basin	**Percentage Area**
Emilia Romagna	0.16 %
Toscana	7 %
Umbria	47%
Marche	1.2 %
Abruzzo	3.6 %
Lazio	41%
Vatican City	0.005 %

During the mid-twentieth century, the Corbara dam (Calenda et al., 2009) was built a few kilometres upstream of the Paglia River inlet, which is located approximately 150 km north of Rome as shown in Figure 4.1. Along the Tiber there are many dams with reservoirs (such as the Montedoglio, Chiascio or Alviano reservoirs), but Corbara is considered to be the most important because of its size and capacity. The dam can be used to reduce the peak discharge that occurs downstream in Rome by approximately 300 m^3/s as the dam has an active storage capacity of $135x10^6$ m^3 (Calenda et al., 2009). According to Natale and Savi (2007), this translates to just less than 10% of the peak discharge with a 200 year return period. The travel time of the flood wave from the Corbara dam to the centre of Rome (at the Ripetta gauging station) is between 24

and 30 hours.

Figure 4.2: A photo of Corbara dam taken in 2005. Source: G. Napolitano

4.2.1 Catchment Geology

Four main geomorphological areas can be identified in the catchment: (a) the carbonate Apennine ridge in the eastern and southern area; (b) the Graben of the Tiber, with its deposits of marine and continental sediments in the middle part of the basin; (c) the volcanic mountains of Vulsini, Cimini, Sabatini and Albani, which occupy the south western part of the area; and (d) Terrigenous deposits in flysch facies in the upper part of the catchment (Autorità di bacino del fiume Tevere, 2006). A more detailed presentation of the geology of the basin is provided in Figure 4.3.

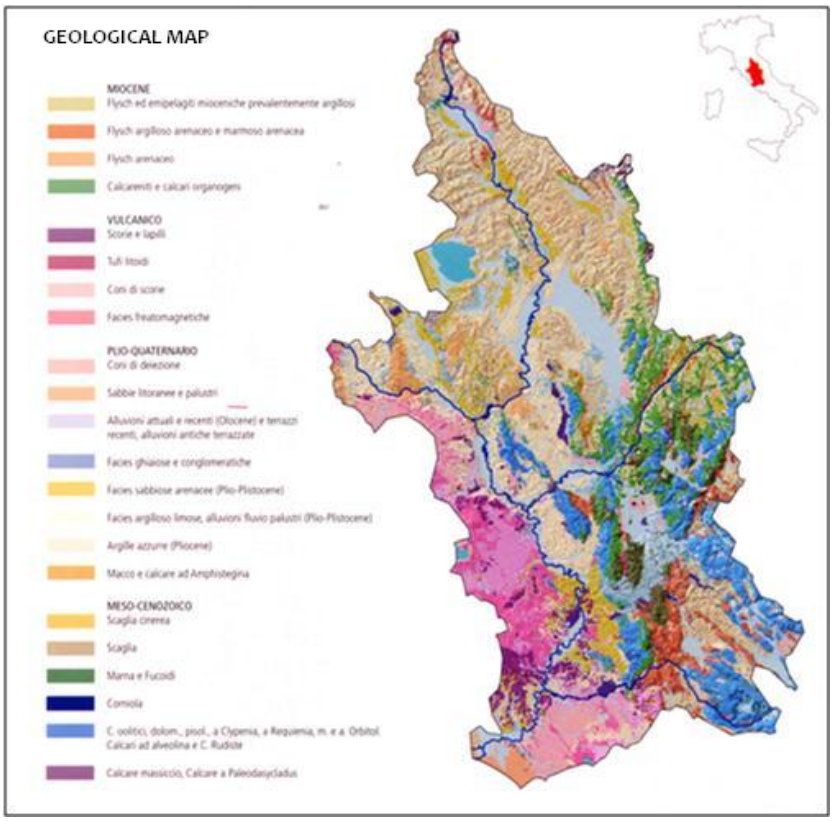


Figure 4.3: Geomorphological map of the Tiber River Basin (Source: Autorità di bacino del Fiume Tevere, 2006)

4.2.2 Land Use

A map showing the land use of the basin is provided in Figure 4.4. Approximately 50% of the area of the basin is used for agriculture, 35% is covered by forests, and only 8% are meadows and pasturelands. Vegetation plays an important role in the defence against erosion of the soil surface. This involves a reduction in sediment transport in the streams, decreasing the runoff, and increasing the concentration time of each sub-basin.

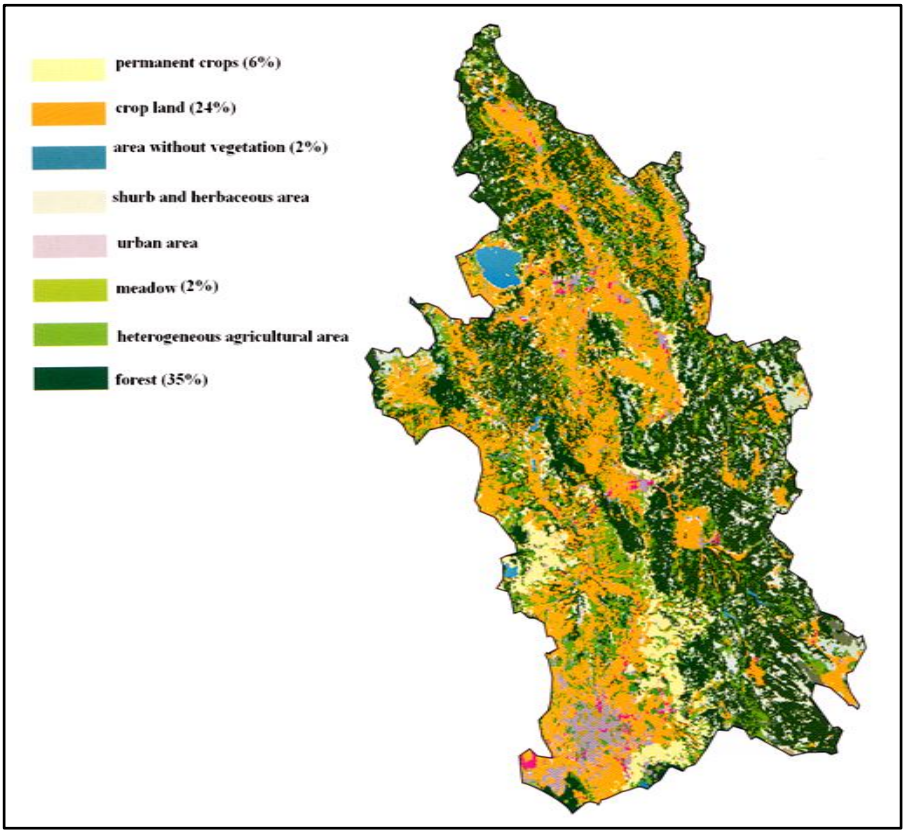


Figure 4.4: Land use map of the Tiber River basin (Source: Autorità di bacino del Fiume Tevere, 2006).

4.2.3 Climate

The rainfall, which is almost uniform throughout the basin, is characterised by two maxima, the first in November and a secondary one in February. The minimum occurs in the summer, usually in July. The distribution of the average annual rainfall is strongly influenced by the orography (as shown Figure 4.5). On the basis of continuous measurements over a period of 50 years, the average annual rainfall has been calculated as 1,050 mm (Autorità di Bacino del Fiume Tevere, 2006). In the case of the Tiber catchment, the rainfall from 1 to 4 days before the flood peak (Bersani and Bencivenga, 2001) or until 6 days before the peak (Palmieri et al., 2001; Remedia et al., 1998) are considered as determinants of the saturation of the land and the state of the river.

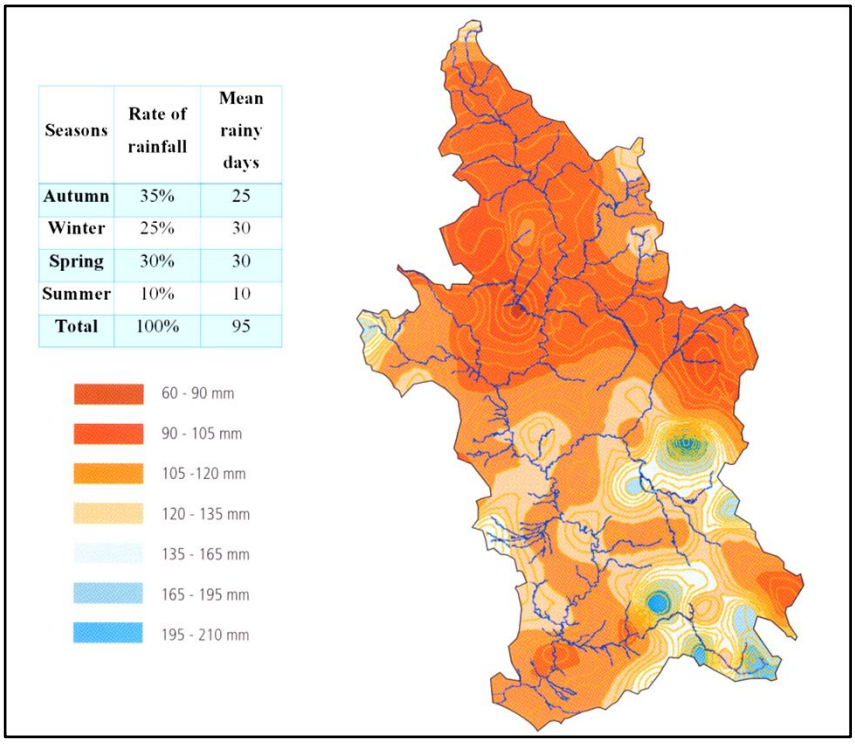


Figure 4.5: Average rainfall map over 50 years of available records (Source: Autorità di Bacino del Fiume Tevere, 2006).

4.2.4 Catchment Hydrology

The Tiber catchment is characterised by the presence of several seasonal streams. This element, plus erosion and solid transport, can cause noticeable variations in the discharge of the River Tiber. In order to control for this effect, several such structures have been built over the past few decades including nine dams, four weirs and twelve bottom sills and drop structures (Autorità di bacino del fiume Tevere, 2006).

The gauging station for further consideration is located at Ripetta in Rome in order to examine how the construction of the Corbara dam, which began operation in 1965 (Natale and Savi, 2007), may have influenced the annual maximum flow at this station. Figure 4.6 shows the annual maximum discharge at Ripetta from 1921 to 2008. The red line highlights the year when the Corbara dam was first in operation.

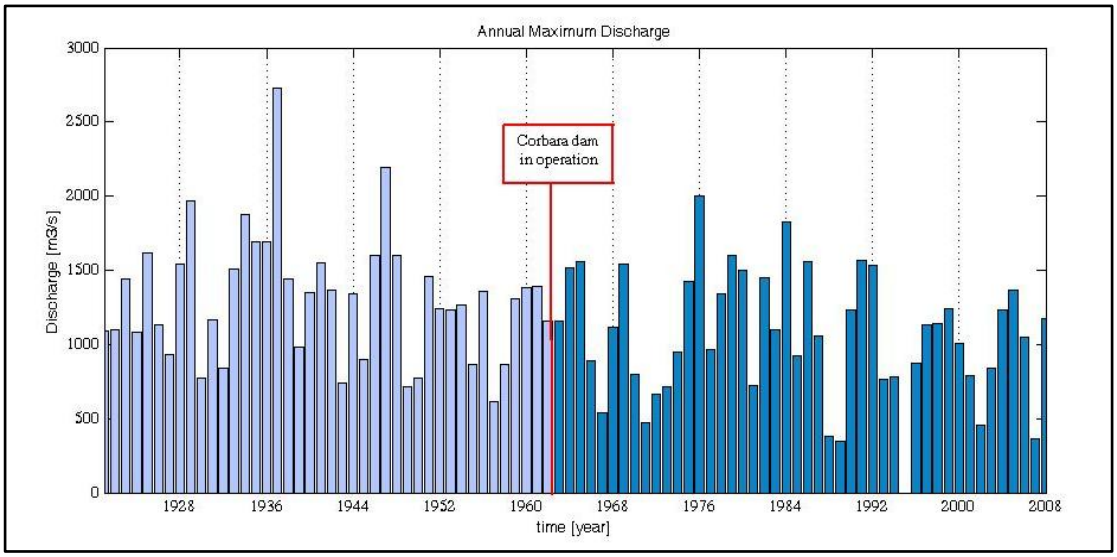


Figure 4.6: Annual maximum discharge at Ripetta gauging station in Rome.

In order to see if the dam has changed the behaviour of the Tiber River, a series of different statistical inhomogeneity tests have been applied including the Run Test (Wald and Wolfowitz, 1940), the Kendall test (Kendall, 1975), the Pearson test (Plackett, 1983) and the Cox Stuart test (Cox and Stuart, 1955). The main purpose of these tests is to evaluate if there has been some change in the trend of the series. Table 4.2 shows the results of these tests.

Table 4.2: Results of Moment Tests (sample size 87) at a significance level of 5%

Moment Tests	**Null hypothesis**	**Reference interval/value**	**z**	**Is the null hypothesis true?**
Run test	The elements of the sequence are mutually independent	[-1.96 ÷ 1.96]	-0.11	YES
Kendall	The elements of the sequence are mutually independent	[-1.96 ÷ 1.96]	-0.10	YES
Pearson	The samples are an independent series of records	1.99	2.84	NO
Cox-Stuart	The series does not have a trend	[13÷25]	11	NO

These tests provide an answer regarding the homogeneity or inhomogeneity of the time series. However, if the sample is not homogeneous, then the test cannot explain the reason for that inhomogeneity, which can only be the result of subjective considerations. The rejection of the null hypothesis of the Pearson test seems to highlight a slight downward trend, noted by Kendall's test and the Cox-Stuart test at 5% significance. In fact, it is well known that the second half of the twentieth century was a period of relative calm for the Tiber River (Calenda et al., 2009).

To test whether the dam has effectively changed the behaviour of the river, the series

of records was divided into two sub-samples: the first with discharge data until 1964 and the second with the data between 1964 until 2008. To detect differences between the mean of the two samples, the T test (parametric) and the Mann-Whitney test (non-parametric) have been applied. In addition, to evaluate the difference in the variance, two parametric tests, i.e. the χ^2 test and the F test, were used. The results are shown in Table 4.3 and indicate that there are significant differences between the series before and after the year 1964.

Table 4.3: Moment tests between the series until 1963 (sample size 43) and from 1964 to 2008 (sample size 44)

MOMENT TESTS	Significance	Null hypothesis	Reference interval/value	Z	Is the null hyp true?
T- test	5%	The means are equals	[-1.66 ÷2.55]	2.55	NO
Mann-Whitney	5%	The samples are from the same population	[-1.96 ÷ 1.96]	-2.61	NO
χ^2 test	5%	The variances are equals	[26.79 ÷ 62.99]	41.39	YES
F- test	10%	The variances are equals	1.664	1.039	YES

The skewness of each sample was also evaluated and is listed in Table 4.4, which shows a marked difference in the asymmetry before and after 1964. The sample with measurements before 1964 is highly asymmetrical, which could be a result of the big floods that occurred in 1937 and 1947, while the asymmetry in the sample of data after 1964 drops almost to zero.

Table 4.4: Skewness value for the series until 1963 (sample size 43) and from 1964 to 2008 (sample size 44)

Statistic parameter	Samples 1921-1962 Size=43	Samples 1963 -2008 Size=44
Skewness	0.978	0.06

4.2.5 Flooding in the City of Rome

The Tiber River flows for over 60 km through the urban area of Rome, passing through 10 town districts out of 19. A population of 1.4 million live in those areas (Autorità di bacino del fiume Tevere, 2006). In particular, Rome and the surrounding areas have been affected by flooding over many centuries. Even though the city centre is quite safe from flooding due to the concrete walls (Muraglioni) that were built along the river in the 20th century (Natale and Savi, 2007), there is still a high possibility that water could overflow the banks around the Milvio bridge (Natale and Savi, 2006) (Figure 4.7). Two more recent events (in 2005 and 2008) are described in more detail in the sections that follow.

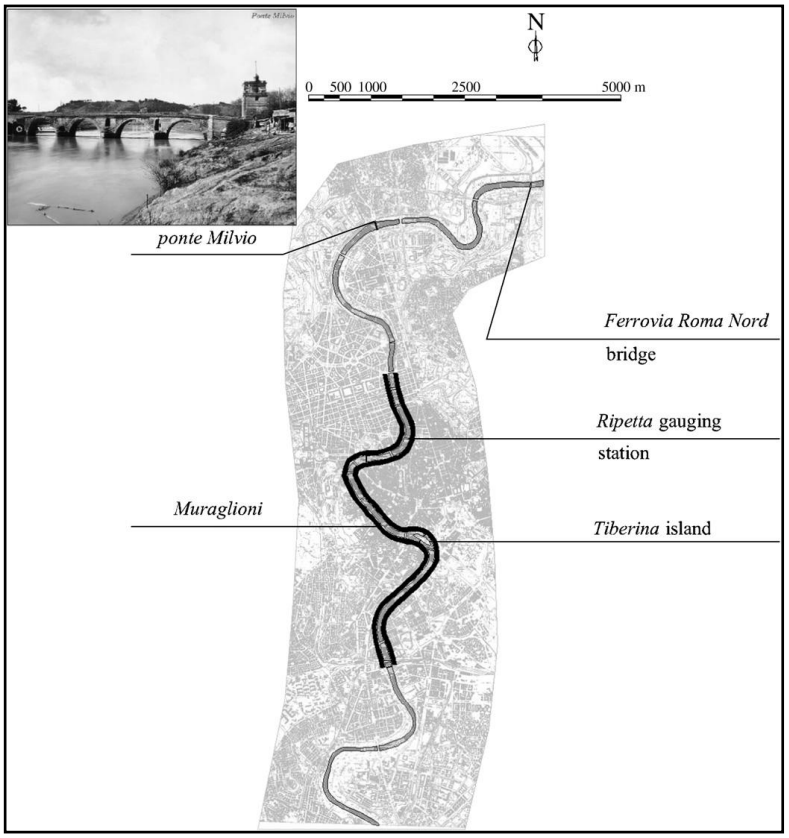


Figure 4.7: Map of the city centre of Rome. Source: Natale and Savi (2007).

4.2.5.1 The 2005 Event

Between 25 November and 9 December 2005, there was a period of persistent rainfall in the Tiber basin, in particular on 25-26 November, 29-30 November, 2-3 December, 5-6 December and 9 December. The cumulative precipitation and the hourly intensity of the rainfall was not particularly high so the main cause of flooding was from: (a) diffuse precipitation covering large spatial and temporal scales within the catchment; and (b) a high degree of soil saturation. As a result, the Corbara reservoir was used to store the flow coming from the upper part of the catchment in order to reduce the effects downstream of the dam. Figure 4.8 shows the observed discharge hydrographs: (a) flowing into the reservoir; (b) released from the Corbara dam; and (c) at the Ripetta gauging station between 26 and 30 November 2005. The main effect of this flooding event was the inundation of the river at Orte, Ponte Felice and at the inlet of the Tiber in the Tyrrhenian Sea (Figure 4.1). In addition, two ships sank along the Tiber in Rome, the power supply in Idroscalo di Ostia was interrupted, and an area in Fiumicino municipality (Passo della Sentinella) was evacuated. The observed peak discharge in Rome was about 1400 m^3/s.

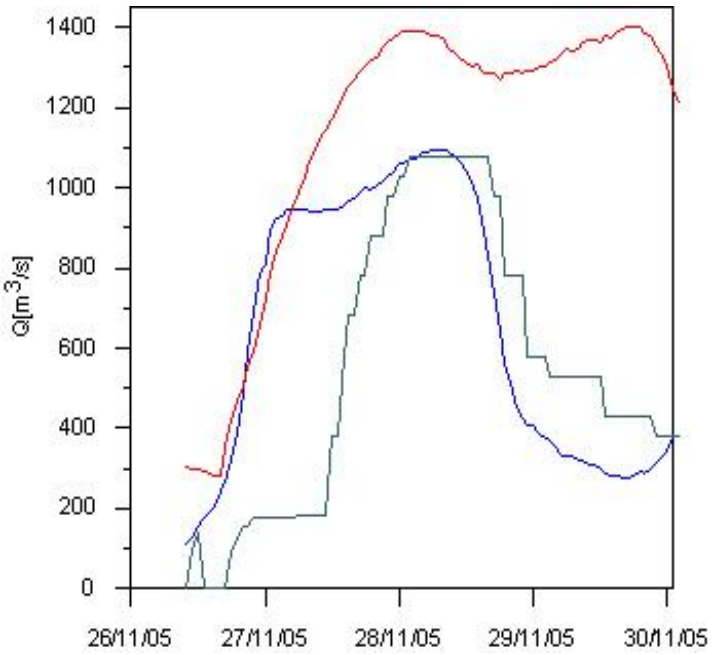


Figure 4.8: Observed discharge hydrographs flowing into the reservoir (in red), released from Corbara dam (blue), and at Ripetta gauging station (green) for the flood event that occurred in November 2005.

4.2.5.2 The 2008 Flood Event

Figure 4.9 shows the 2008 event in terms of precipitation recorded at a number of gauging stations. This event was characterised by 3 distinct periods of persistent rainfall (5-7 December, 10-13 December and 15-16 December 2008). The cumulative precipitation over 6 consecutive days was between 150 to 190 mm over most parts of the basin. The water level at Ripetta rose from 7 m to 11.30 m between 10 and 11 December. At the same time the Corbara reservoir stored most of the rainfall from the central part of the basin. Through the use of floodplains, the rise in the water level at Ripetta was slowed down. The peak (of 12.55 m) was recorded at Ripetta on 13 December 2008. This flood event was about the same magnitude as that which occurred in February 1986 (12.40 m at Ripetta), February 1976 (12.72 m at Ripetta), September 1965 (12.65 m at Ripetta) and December 1964 (12.46 m at Ripetta). The major effects of this flooding event were the inundation of the urban areas of the Aniene River. In addition, two boats were cast off their moorings, obstructing Sant'Angelo bridge in the centre of Rome. This called for the urgent intervention of the Civil Protection Authority and firemen. Both boats were destroyed by controlled explosions, thus enabling the smooth flow of water down the river.

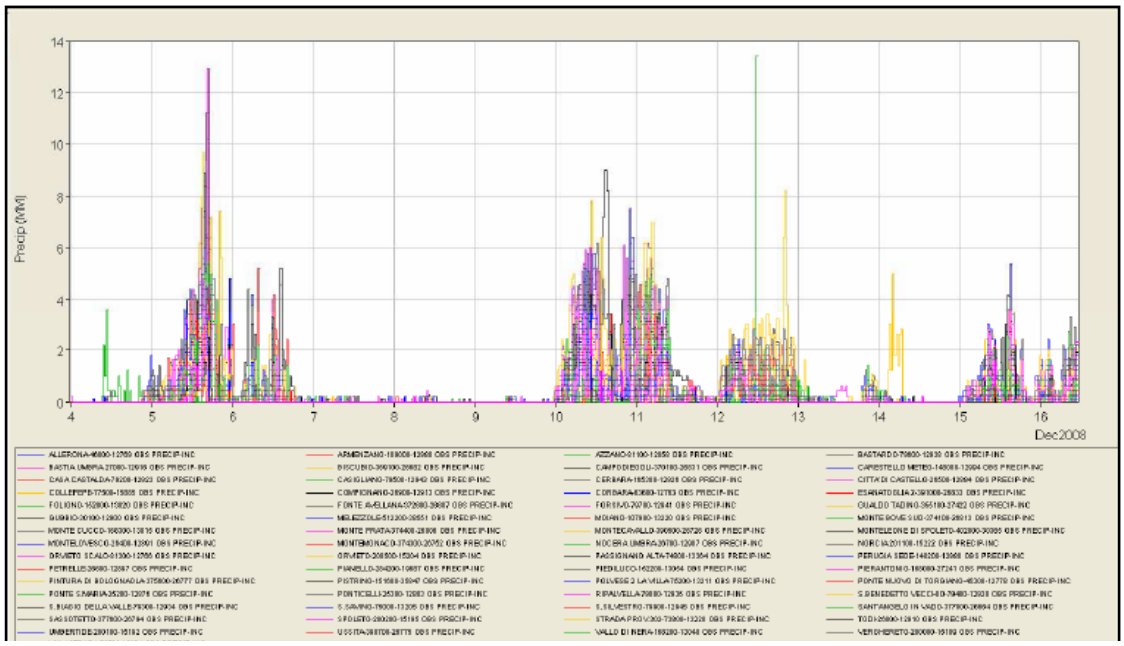


Figure 4.9: Rainfall during the 2008 flooding event for all stations in the middle part of the Tiber basin at a recording interval of 30 minutes. Source: Centro Funzionale Regione Umbria (2009).

4.3 The Conceptual Tevere Flood Forecasting (TFF) Model

The TFF (Tevere Flood Forecasting) model was developed by Calvo and Savi (2009) for real time forecasting by modifying the TEVERE model originally developed by Natale and Savi (2007). The TFF is composed of three sub-models: (i) the TFF BASIN model, which is a conceptual, semi-distributed hydrologic model that transforms rainfall into hourly runoff from the sub-basins of the Tiber River; (ii) the TFF RIVER model, which simulates the propagation of the flood wave from Corbara dam to the Tyrrhenian sea; and (iii) a module for the online calibration and adaptation of the values of the model parameters.

A schematic of the TFF BASIN model is shown in Figure 4.10. Rainfall intensity i is comprised of surface runoff q and the infiltration rate f (Figure 4.10a). The infiltration rate is calculated through application of the mass balance equation to a certain known volume of soil of unit area, considered as a conceptual reservoir (Figure 4.10b): S is the water level in the reservoir and p is the subsurface flow. The fraction αp of the subsurface runoff contributes to the surface runoff, while the remaining fraction $(1-\alpha)p$ is considered lost in deep percolation; the variables f and p depend on S according to the relationships shown in Figures 4.10c and d respectively. The surface runoff is computed as: $q = \max(0, i-f)$. Separate convolution of the two runoff components q and αp, via surface and subsurface Gamma Instantaneous Unit Hydrographs (IUH), where k_{sur}, n_{sur} and k_{sub}, n_{sub} are the parameters of the two IUHs representing surface and subsurface flow, gives the channelised flow Q. The model therefore includes the 13 parameters as shown in Figure 4.10: $k_{sur}, k_{sub}, n_{sur},$

n_{sub}, α, f_i, f_f, S_{if}, S_{ff}, p_i, p_f, S_{ip}, S_{fp}.

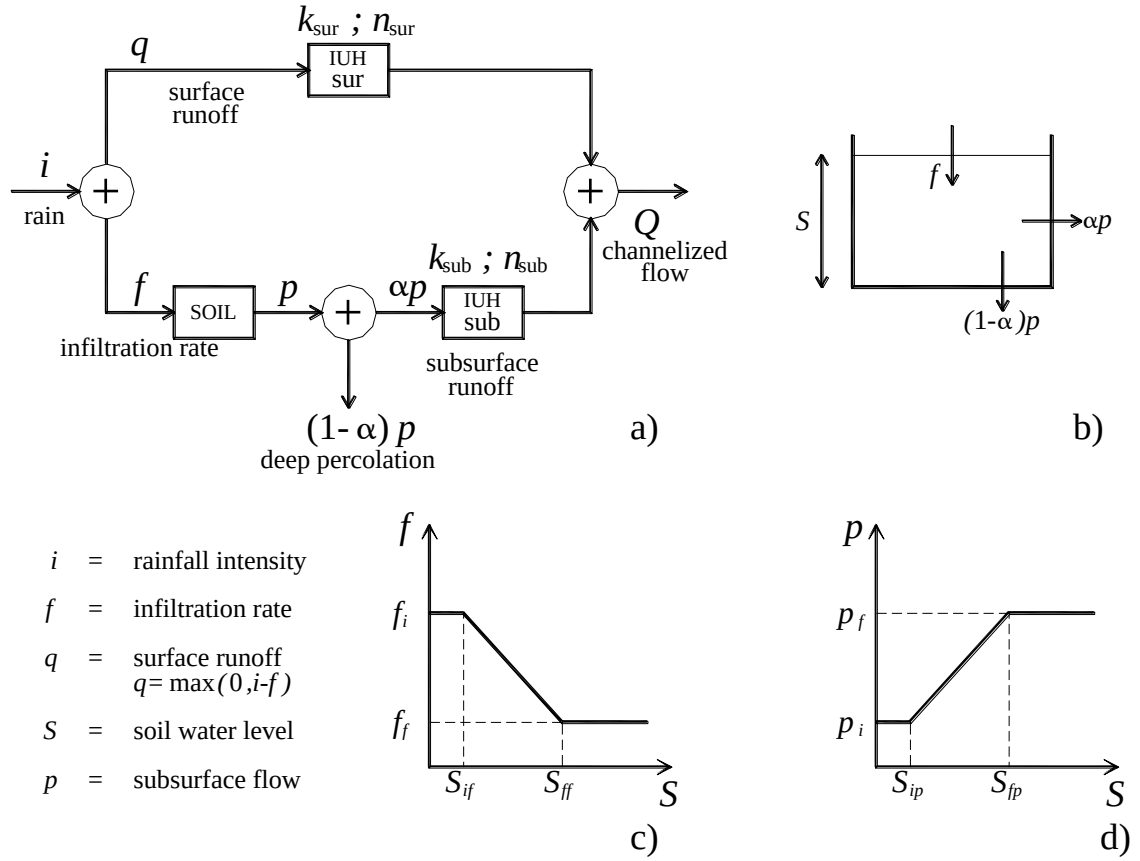


Figure 4.10: An outline of the TFF BASIN model (Source: Napolitano et al., 2009).

The TFF RIVER model simulates the propagation of the flood wave along 8 reaches over which the hydrographic network is divided, including the two short channels around Isola Tibertina, an island in the centre of Rome, and those of Fiumara Grande and Canale Fiumicino (Figure 4.1). For each reach, the 1D, free surface, gradually varied flow equations were integrated (Cunge *et al.*, 1980). The model simulates flow through hydraulic singularities (i.e. bridges, dams, sills and drop structures) along the river by imposing internal boundary conditions, i.e. rating curves, and takes into account the inflows from tributaries. The model parameters (Manning roughness coefficients, discharge coefficients at bridges and dams, minor loss coefficients at nodes) were calibrated offline by simulating historical flood events.

During real time forecasting, the values of the TFF RIVER model parameters are held constant, whereas the values of the TFF BASIN model parameters are calibrated online at each time step. The values of the model parameters for the 41 ungauged sub-basins were estimated through a regional analysis on the basis of hydrological similarity by performing a preliminary offline calibration. Recordings of hourly rainfall and discharge were collected for 70 floods on 12 gauged sub-basins located upstream

and downstream of Corbara dam. These events were simulated and used to calibrate the TFF BASIN model. The values of the model parameters were calibrated by means of a genetic algorithm as proposed by Wang (1991; 1997).

In order to estimate the parameters of the rainfall-runoff model for the 41 ungauged sub-basins, the resulting values of k_{sur} that correlated well with the time of concentration t_c, were computed according to the Giandotti (1934) formula:

$$t_c[hours] = \frac{4\sqrt{\omega} + 1.5L}{0.8\sqrt{\Delta Z}} \tag{4.1}$$

where ω [km^2] is the surface of the sub-basin, L [km] is the length of the main water course, and ΔZ [m] is the difference between the mean sub-basin elevation and the outlet elevation. The values of the other three parameters (n_{sur}, k_{sub}, n_{sub}) of the Gamma Instantaneous Unit Hydrograph (IUH) did not correlate to any morphological parameter and were characterised by their mean values and standard deviation.

The remaining nine parameters (α, f_i, f_f, S_{if}, S_{ff}, p_i, p_f, S_{ip}, S_{fp}), all of which refer to the subsurface flow, were linearly correlated to the value of the SCS Curve Number, CN (SCS, 1985). For each sub-basin, the values of CN varied for different storms due to the variations of the antecedent soil water content. To take into account these variations for each of the nine parameters X_j, two linear correlations were performed. The first provides an estimation of the mean value of the parameter and can be expressed as:

$$X_{mj} = \sigma_j CN_{II} + y_j \qquad j = 1, \ldots\ldots, 9 \tag{4.2}$$

where X_{mj} is the mean value of the generic parameter of the subsurface flow model shown in Figure 4.10 (α, f_i, f_f, S_{if}, S_{ff}, p_i, p_f, S_{ip}, S_{fp}), and CN_{II} is the CN value in normal soil moisture conditions before the storm, which is obtained from maps of soil type and soil use (SCS, 1985). The ratio between the value of the parameter for the storm X_S and the average value $r_{Xj} = X_{sj}/X_{mj}$ was then correlated with the ratio $r_{CN} = CN_S/CN_{II}$ as follows:

$$r_{Xj} = \sigma\gamma_j r_{CN} + \mu_j \qquad j = 1, \ldots\ldots, 9 \tag{4.3}$$

where CN_S is the value of CN for a given storm.

In the real-time flood forecasting procedure, to reduce the number of parameters, the values of the ratio $r_{CN} = CN_S/CN_m$ were kept constant over the three zones 5, 7 and 8 shown in Figure 4.1. The total number of parameters that require online calibration was therefore reduced to 3. During a flood event, these three parameters are calibrated online by minimising the objective function, which is the sum of the squares of the differences between observed and computed water surface elevations in a moving window of 18 hours. For a more detailed explanation of the model, see the paper by Calvo and Savi (2009).

Observed hourly discharge hydrographs were available at four stream gauging stations: Corbara dam, Paglia River at Orvieto, Nera River at San Liberato dam and the Aniene River at Lunghezza (Figure 4.1). The TFF BASIN model was used to calculate the contribution of the tributaries of the sub-basins found in zones 5, 7 and 8 (shown in Figure 4.1).

The full TFF model was used to predict the 2005 and 2008 floods. Confidence intervals of the forecasted values were estimated by means of a Monte Carlo procedure as outlined in Calvo and Savi (2009).

4.4 An ANN Model of the Tiber River

Some ANN modelling of the Tiber River has already been reported in the literature but previous work has been confined to modelling of the upper Tiber and not the lower part of the river as in this research study. For example, Bonafé et al. (1994) applied a feedforward ANN consisting of three layers with three hidden nodes determined by trial and error. The data available covered a period of 4 years from 1/08/1988 to 31/12/1992 including daily precipitation from 26 rain gauge stations, daily mean temperature from 13 thermographic stations, and daily mean discharge at the section of Monte Molino (just upstream of the Corbara reservoir). For training the network, data from 4/08/1988 to 31/12/1991 were employed, and 1992 was then used to validate the performance of the network. The results obtained with the ANN were compared with those obtained from applying the ARX rainfall-runoff model (an autoregressive with exogenous variables model) and the persistence hypothesis to the same catchment and the same data set. The overall conclusion of the authors, based on the efficiency of each model, was that the ANN provided highly accurate runoff reconstructions; the performance was much better compared with the other applied models. The ANN RMSE value was at least 10% smaller than the ARX and the persistence models. In 2007, Tayfur and

Moramarco applied a feedforward ANN trained with backpropagation. The model was trained using 6 events recorded at three cross sections. The model was applied to a 4-h, 8-h and 12-h lead time. The ANN provided a good performance, especially for an 8-h lead time. Tayfur and Moramarco (2007) found that the ANN was able to predict storm events in situations where the travel time of the flood wave is less than the lead time of the prediction. However, neither of these studies compared the ANNs to a conceptual or physically-based model.

In this research, an ANN of the Tiber River was developed to predict the water levels at Ripetta using recordings of hourly water stage from Ripetta and Orte stream gauging stations between 1993 and 2007. The network consisted of 12 lagged inputs from each station, one hidden layer with 10 nodes and 1 output, i.e. the levels at Ripetta station with a lead time of 12 or 18 hours as shown in Figure 4.11.

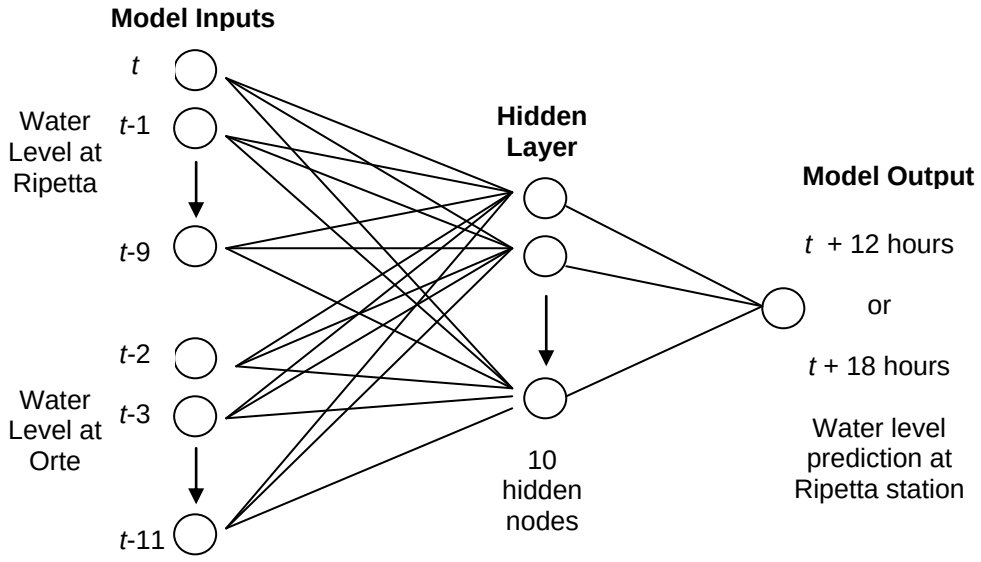


Figure 4.11: A schematic of the ANN for the River Tiber

The configuration of the network, in particular the number of hidden nodes, was determined through trial and error. This consisted of trying networks with a range of hidden nodes and then choosing the best performing one. The network was trained with backpropagation and Bayesian Regularisation algorithms. The training was undertaken 50 times and the results were averaged in order to minimise any variability that could be a function of the random initialisation of the weights (Anctil, 2007). This procedure also means that a validation data set is not required for stopping, i.e. termination of the learning process to avoid overfitting of the data. Therefore, more data could be utilised in the training process. Flood events were extracted from the

historical records while data from 2005 and 2008 were used for independent testing of the network.

4.5 Results

Both the TFF and ANN models were compared in terms of their ability to predict the flood events that occurred in 2005 and 2008. In the forecasting procedure, both models use observed discharge, without additional rainfall forecasting; the TFF model assumes zero rainfall during the lead time (Corradini et al., 2004). Table 4.5 summarises the statistical characteristics of the observed and computed time series from each model. Since the forecasted time series of the TFF is smaller than the ANN, two different statistics are reported in Table 4.5: the first one is computed for the duration of the TFF forecasting, the second one (in parenthesis) is for the duration of the ANN forecasting. The overall statistics are similar for both models at a forecasting horizon of 12 hours. However, for a lead time of 18 hours (i.e. L_T=18 h), the ANN statistics agree with the observed ones better than the TFF model.

Table 4.5: Descriptive statistics for the observed and computed water levels for the TFF and ANN models

Stats	**Event**	**Observed**	**Computed**	**TFF**	**Computed**	**ANN**
		2005	**L_T=12 h**	**L_T=18 h**	**L_T=12 h**	**L_T=18 h**
Mean	2005	11.48 m (10.90 m)	11.48	11.41 m	11.56 m (10.92 m)	11.23 m (10.54 m)
Min		10.59 m (7.17 m)	10.48	9.96 m	10.86 m (7.81 m)	10.14 m (6.76 m)
Max		11.80 m	11.88 m	12.02 m	11.88 m	11.61 m
Std dev		0.27 m (1.04 m)	0. 32 m	0.62 m	0.31 m (1.06 m)	0.37 m (1. 10 m)
Skewness		-1.36 (-1.98)	-1.63	-1.60	-1.19 -1.45	-1.75 (-1.47)
Mean	2008	11.24 m (10.63 m)	10.96 m	10.55 m	10.93 m (10.41 m)	10.70 m (10.23 m)
Min		8.50 m (6.82 m)	8.45 m	7.00 m	8.31 m (6.69 m)	8.22 m (6.87 m)
Max		13.16 m	13.16 m	13.08 m	13.15 m	13.22 m
Std dev		1.49 m (1.96 m)	1.57 m	1.80 m	1.55 m (1.91 m)	1.50 m (1.81 m)
Skewness		-0.35 (-0.46)	-0.26	-0.18	-0.36 (-0.39)	-0.11 (-0.23)

Tables 4.6 and 4.7 provide the computed performance measures outlined in Chapter 3 for the two models and two lead times. Unfortunately the times series was too short and not independent so the sign test and Diebold-Mariano test to compare the two models could not be applied. The PDIFF was employed as these are single flood events and not a continuous record.

Starting with the flood event in 2005 (Table 4.6) and a lead time of 12 hours, the

absolute error measures indicate good performance for both the TFF and ANN models with an average error of less than 30 cm. However, the ANN model shows worse performance than the conceptual model with the exception of the ME. As this is a signed statistic, the over- and under-predictions may simply be cancelling each other out as mentioned in Chapter 3. The difference in peak prediction is similar for both models. In terms of the relative measures, a consistent message can be found, i.e. the ANN outperforms the conceptual model in terms of the MPE but is worse in the MAPE, MdAPE and RMSPE. The measures relating to benchmark models show very high values of CE, which denotes a very good performance according to Shamseldin (1997) and Dawson et al. (2007). The PI also indicates that the model is considerably better than a naïve forecast. However, the other PI similarity measures (PI.MAE and PI.MdAE) are lower than 1 for both models but higher for the conceptual TFF model. The PI.MdAE for the ANN indicates that it is only slightly better than an MdAE naïve forecast. The GRI indicates a very small cone around the line of best fit and provides little to help differentiate between the performance of the two models. This may be due to the small number of data points in the forecasts. Overall, both models perform well for a lead time of 12 hours but the conceptual model is better based on an examination of these measures.

Table 4.6: Statistical error measures comparing observed and computed water levels for the event in 2005. The ME, MAE, MdAE, RMSE and PDIFF are in metres. The remaining relative measures are dimensionless.

Parameters	**TFF**		**ANN**	
	L_T=12 h	**L_T=18 h**	**L_T=12 h**	**L_T=18 h**
ME	0.09	0.18	-0.01	0.32
MAE	0.12	0.26	0.30	0.36
MdAE	0.13	0.26	0.28	0.26
RMSE	0.14	0.37	0.35	0.45
PDIFF	-0.08	-0.22	-0.08	0.19
MPE	0.39	1.71	-0.11	3.0
MAPE	2.20	3.55	3.33	3.13
MdAPE	1.12	2.20	2.70	2.47
RMSPE	3.58	5.30	4.39	4.98
CE	0.96	0.94	0.93	0.93
PI	0.96	0.87	0.88	0.80
PI.MAE	0.68	0.63	0.46	0.50
PI.MdAE	0.63	0.39	0.03	0.13
GRI	1.00	1.00	1.00	1.00

At a lead time of 18 hours, the difference between the conceptual model and the ANN is less pronounced. In terms of the absolute measures, the conceptual model shows better performance but similar values for the MdAE. This time the peak difference is also underestimated in the conceptual model but overestimated by the ANN, which is a preferred result from a forecasting point of view. The relative measures show a mixed result, i.e. the conceptual model performs better in terms of the MPE and MdAPE but

not in terms of the MAPE. The CE is very similar at this lead time, which continues to highlight the fact that the mean is not the best benchmark model to use. The PI is still high in both models indicating better forecasting ability than a naïve model. The other PI-based similarity measures also show the same degradation as at a lead time of 12 hours. The same comments apply to the GRI.

Table 4.7 contains the same set of statistics for both models and lead times but for the 2008 flood event this time. It should be noted that the 2008 flood event is an event higher than that which appeared in the training dataset. It is therefore expected that the ANN might perform much worse than the conceptual model. Interestingly, this is not the case. Considering a lead time of 12 hours, the absolute measures indicate pretty similar performances for both models with a better performance in the RMSE for the ANN, i.e. the measure which is biased towards higher flow events. The PDIFF is also very good for both models with a slightly better performance by the ANN. The PDIFF does not, however, indicate anything about time to peak, which is discussed in the next section on visual inspection of the hydrographs.

Table 4.7: Statistical error measures comparing observed and computed water levels for the event in 2008. The ME, MAE, MdAE, RMSE and PDIFF are in metres. The remaining relative measures are dimensionless.

Parameters	**TFF**		**ANN**	
	L_T=12 h	**L_T=18 h**	**L_T=12 h**	**L_T=18 h**
ME	0.24	0.11	0.27	0.26
MAE	0.42	0.74	0.39	0.58
MdAE	0.30	1.03	0.43	0.44
RMSE	0.62	1.01	0.48	0.76
PDIFF	0.04	0.08	0.01	-0.06
MPE	2.06	0.37	2.05	1.89
MAPE	4.11	5.18	5.32	5.83
MdAPE	3.25	5.89	4.56	4.55
RMSPE	6.69	12.71	6.43	7.50
CE	0.74	0.17	0.90	0.85
PI	0.81	0.61	0.87	0.82
PI.MAE	0.60	0.34	0.48	0.58
PI.MdAE	0.52	0.05	0.32	0.51
GRI	1.00	1.00	1.00	1.00

The relative measures indicate similar performance in terms of the MPE but a superior performance by the conceptual model in the remaining relative measures. In terms of benchmark measures, the CE is much higher for the ANN and the PI is higher. However, when the PI.MAE and PI.MdAE are considered, the ANN performs worse. The values are higher, however, for this event compared to the 2005 event. Finally the same comments are relevant to the GRI as outlined for the 2005 event.

For a lead time of 18 hours, the ANN shows superior performance in all of the absolute

measures except for the ME. The PDIFF is slightly underestimated by the ANN but both models perform well according to this measure. The relative measures again indicate a mixed performance with the conceptual model doing better on some measures and vice versa. The CE for the conceptual model is very poor while it is good for the ANN. If this was the only measure used to evaluate the models, then the picture of model performance would not be correct. This highlights the danger of using only one measure. The ANN performs better in terms of all the PI-based measurements and the PI.MdAE indicates the conceptual model is similar to the use of the naïve MdAE. The GRI remains consistently at 1.0 and therefore provides no differentiation between models.

In general, the evaluation of the model performance based on these error measures is positive. It indicates that these models are able to forecast the level at lead times of 12 and 18 hours with reasonable accuracy, which is a useful length of time for implementing civil protection measures. However, for the longer lead time, the ANN model seems to outperform the conceptual model although not in all evaluation measures. This analysis highlights the danger of using only one measure to evaluate model performance such as the CE. It also shows that models do not perform consistently well across the measures, and appropriate measures should therefore be chosen to highlight the desired skill of a model.

As mentioned at the end of Chapter 3, visual inspection of the hydrographs is also undertaken to complement the quantitative analysis via performance measures. Figures 4.12 and 4.13 show the envelope of the forecasted water levels computed by the TFF model for a lead time of 12 and 18 hours, respectively, at Ripetta gauging station in Rome for the 2005 event. The 90% confidence intervals of the forecasted water levels are also provided in Figure 4.12. Moreover, in both of these figures, the average rainfall hyetograph over the medium-lower Tiber basin (zones 5, 7 and 8 in Figure 4.1) is shown.

The conceptual model accurately forecasts the observed hydrograph once the event has progressed to 18 hours beyond the start of the event, where the first 18 hours are required for calibration. Confidence intervals of the forecast levels are about 30 cm wide (Figure 4.12). If a greater lead time is considered, i.e. 18 hours (Figure 4.13), the model performances decay essentially in the forecasting of the rising limb of the hydrograph. This is mainly due to the assumption of zero rainfall in the forecasting period, which means that the contributions of the rainfall that occurred at the end of 26 November (Figure 4.13) in the forecasting of the water levels observed at the beginning

of the flood (27 November) are ignored.

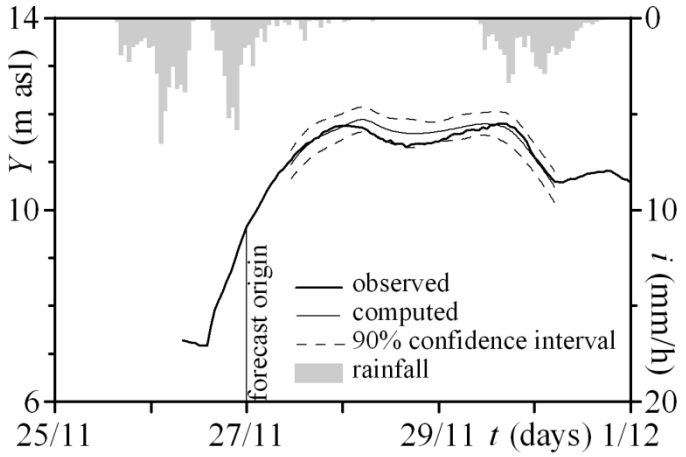


Figure 4.12: Comparison between the observed and 12h forecasted water levels from the TFF model for the November 2005 flood. Source: Napolitano et al. (2009).

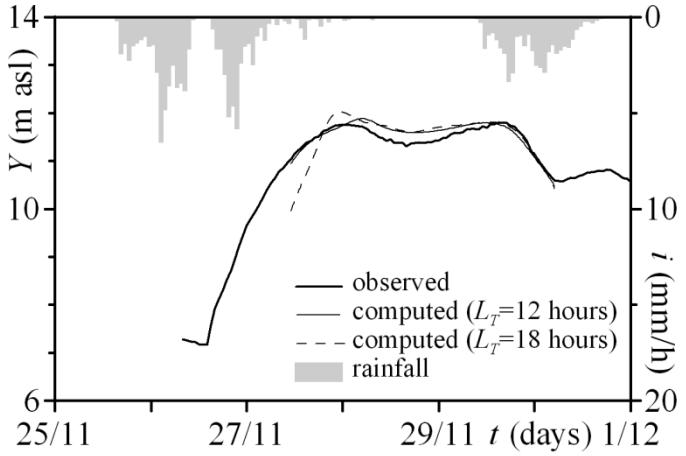


Figure 4.13: Comparison between the observed and forecasted water levels with two different lead times from the TFF model for the November 2005 flood. Source: Napolitano et al. (2009).

The results of the ANN model are shown in Figure 4.14 for lead times of 12 and 18 hours. The ANN model is able to forecast the rising limb of the hydrograph accurately but it forecasts both the observed peaks slightly early. For a lead time of 18 hours, the results degrade slightly. The rising limb shows a delay in prediction and the peak is slightly underpredicted. However, the general shape of the hydrograph is reproduced satisfactorily for real world applications.

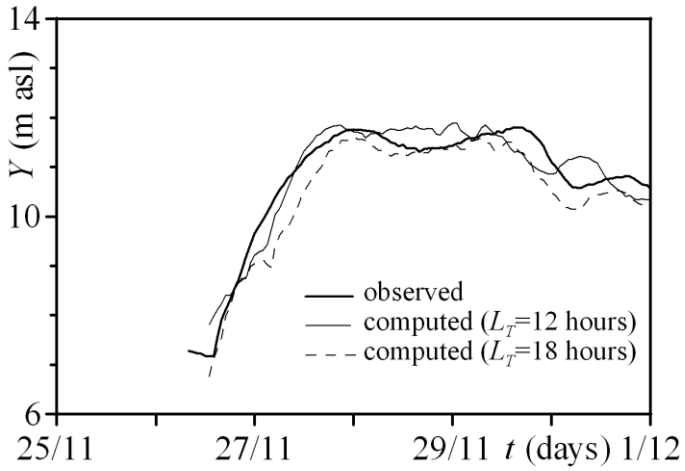


Figure 4.14: Comparison between observed and forecasted water levels with different lead times from the ANN model for the November 2005 flood. Source: Napolitano et al. (2009).

Figures 4.15, 4.16 and 4.17 refer to the 2008 flood. In Figures 4.15 and 4.16 the forecasted water level hydrographs computed by means of the TFF model are compared with the observed one. In Figure 4.15 (L_T=12 hours), the 90% confidence intervals of the forecasted water levels (about 0.5 m wide) are also shown. The TFF model fails to forecast the rising limb of the hydrograph but correctly forecasts the peak. The same occurs with an increase in the lead time (Figure 4.16). In this case the effects of the assumption of zero rainfall during the lead time clearly emerge: at the beginning of the forecasting, i.e. the heavy rainfalls which occurred in the first 12 hours of 11 December, are not considered by the model and so the rising limb of the hydrograph at Ripetta is significantly underestimated. The same occurs for the second peak on 16 December.

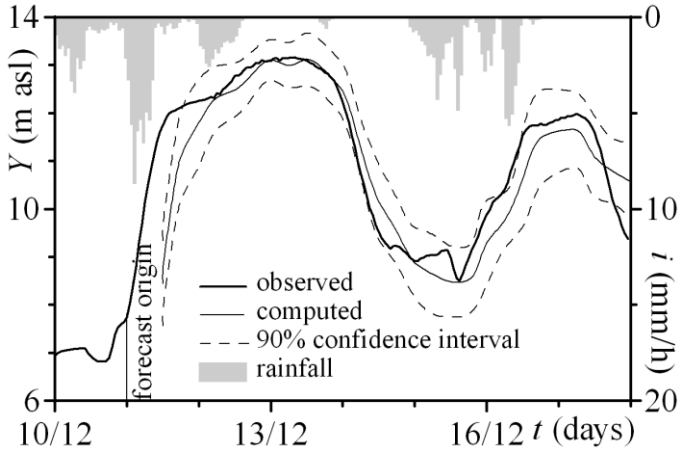


Figure 4.15: Comparison between the observed and 12h forecasted water levels for the TFF model for the December 2008 flood. Source: Napolitano et al. (2009).

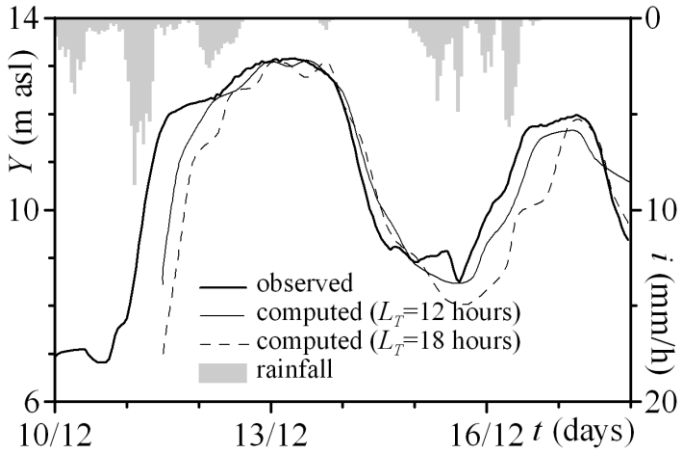


Figure 4.16: Comparison between the observed and forecasted water levels with different lead times for the TFF model for the December 2008 flood. Source: Napolitano et al. (2009).

The forecasted water levels computed for different lead times by means of the ANN model are compared with the observed one in Figure 4.17. In this figure, two results computed with L_T=12 hours are reported. The first one (which is plotted as a grey line labelled as 'computed (L_T=12 hours) wt') excludes the 2008 flood from the training period, which is the most relevant flood in the period where continuous recordings are available (1993-2008). In this case the ANN model significantly underestimates the maximum water levels by more than 0.60 m. This is unsurprising as the ANN model has not seen such a large event before. This also highlights the problem with the use of the PDIFF measure which indicates a better performance than that shown by the graphs.

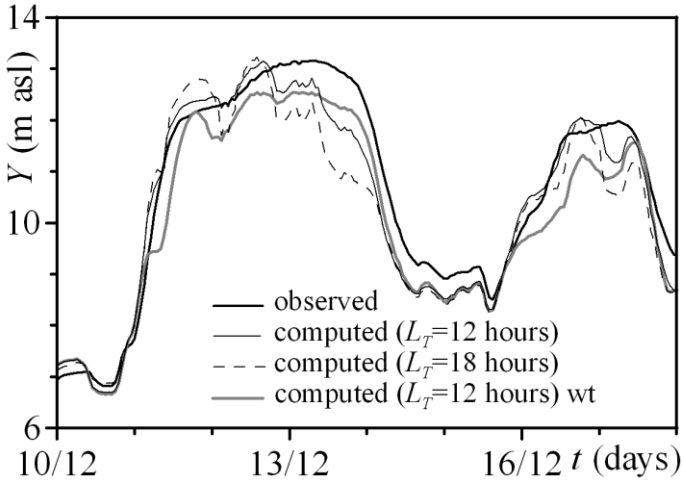


Figure 4.17: Comparison between observed and forecasted water levels with different lead times for the ANN model for the December 2008 flood. Source: Napolitano et al. (2009).

The second line (labelled as 'computed (L_T=12 hours)') includes the 2008 event. This is not a true test of the ANN since the 2008 event is no longer in the independent dataset. However, it was merely to see the effect on the result. In this case, the model performance increases significantly, although there are still problems in predicting the

later part of the first event. Model performances degrade slightly if the longer lead time is considered, but the ANN is generally able to produce acceptable results for real world applications.

4.6 Summary

This chapter has compared a simple ANN model of the Tiber River basin at Ripetta in the city of Rome with a conceptual model developed for the same area. Both the conceptual TFF and ANN models were found to provide good model performance for a lead time of 12 hours. Although the TFF model demonstrates more accurate forecasting of the peak, it requires a calibration window of 18 hours so cannot predict the rising limb of the hydrograph. Moreover, the assumption of zero rainfall in the lead time may cause an underestimation in the rising limb of the hydrograph and a rapid decay of the model performance as the lead time increases.

The ANN model is simpler to construct and train and allows reliable forecasting for a longer lead time of 18 hours, albeit with a slight degradation in performance. The main drawback of the ANN model results in the limited duration of the training period (continuous recordings of hourly water levels are available in the last 16 years only) so that only a few relevant floods are included. As a consequence the model may fail in forecasting extreme floods. However, both models proved to be reliable for real world applications. Given the cheap nature of computing power, it would be possible to run the ANN and TFF model together. Running in parallel, the models provide a complete hydrograph as well as giving forecasters additional confidence from two forecasts of the peak. This is an operational issue, and as mentioned in Chapter 2, there is currently a reluctance to use black box ANN models for real-time flood prediction. This may, however, change in the future.

It is clear that performance measures should only be used with visual inspection of model performance in predicting the hydrographs. Moreover, a range of performance measures should be used to gain a better idea of model performance. PDIFF is not necessarily a good measure as it does not indicate time to peak so will not be used further in this research. The use of CE either produced very optimistic results regarding model performance or very poor ones so should definitely be used in combination with other measures. The GRI was not informative so will also not be used further in modelling of the Tiber River. The models did not perform consistently across all measures but it was possible to gain on overall picture of model performance by using a suite of measures.

The ANN model developed in this chapter was simple. It included only two stations as inputs and no rainfall. In the next chapter, a series of experiments are devised to add more complexity to the ANN model. Ensemble modelling is also considered as multiple ANNs are trained when using the Bayesian Regularisation algorithm.

Chapter 5

Further ANN Model Experimentation

5.1 Introduction

An initial ANN model was presented in Chapter 4 and compared to the conceptual TEVERE model in predicting two recent flood events 12 and 18 hours ahead, which affected the city of Rome in 2005 and 2008. The ANN model performed reasonably well in comparison to the conceptual model but the conceptual model was better able to predict the overall shape of the hydrograph for both events.

This chapter considers ways in which the ANN model can be improved through a number of different experiments such as the addition of more inputs, lengthening the time series, pre-processing of the data, and methods for handling the ensemble, which arises through the need to train the ANN several times to account for the random initialisation of the weights. The first eight experiments are presented for the 2008 flood event since this is the event that requires the ANN to extrapolate outside the range of data on which it was trained. This also provides a true test of whether the ANN is useful for real-time forecasting in this catchment. The latter experiments are presented for both 2005 and 2008 since the idea is to determine the effect of model decisions on more than one event and therefore whether some generalisation of the results can be made. A lead time of 12 hours is used consistently in these experiments. The calculation of confidence limits around the model predictions concludes this chapter.

5.2 Adding Additional Upstream Stations and Rainfall to the ANN Model

The initial ANN developed in Chapter 4 used a limited number of upstream stations and no rainfall data. Thus the ANN acted more as a routing model than a true rainfall-runoff model. These first set of three experiments attempts to improve the ANN through the addition of more upstream stations and rainfall data. In the first experiment ANNs are trained using hourly water levels from gauging stations at Ripetta, Lunghezza, Alviano, Ponte Felice and Orte using data between 2005 and 2008, excluding the 2008 event. The network consisted of 12 lagged inputs from each of these stations, one hidden layer with 10 nodes and 1 output, i.e. the levels at Ripetta station with a lead time of 12 hours. The lag times for the inputs were determined by the average travel times between the upstream stations and Ripetta, which are listed in Table 5.1. As with the initial ANN developed in Chapter 4, the configuration of the network, in particular the number of hidden nodes, was determined through trial and error. Similarly, the network was trained with a Bayesian Regularization algorithm 50 times, and the results were

averaged to minimise any variability that could be a function of the random initialisation of the weights (Anctil, 2007). No validation dataset was therefore required, which is beneficial in this modelling exercise as the amount of historical data available at all of the stations was limited and more data could then be utilised for training.

Table 5.1: Travel times between Ripetta and upstream stations

Station	Travel time
Lunghezza	6 hours
Alviano	20 hours
Ponte Felice	12 hours
Orte	12 hours
Area 5	18 hours

Further experiments were then undertaken through the addition of rainfall inputs. In the second experiment, hourly rainfall was added from area 5 (Figure 4.1) as a driver to the ANN model. Four stations were averaged at Castel Cellesi, Poggio Mirteto, Ponte Felice and Monte Fiascone. The travel time from area 5 to Ripetta is 18 hours (Table 5.1).

In the third experiment, effective rainfall was added, i.e. rainfall that actually reaches the river. Few other studies have tried this with the exception of Sajikumar and Thandaveswara (1999) and Jain and Srinivasulu (2004a). Sajikumar and Thandaveswara (1999) used only effective rainfall as in input since the paper was concerned with a comparison of different kinds of ANN and the effect of the length of the training dataset. Thus, no conclusions could be drawn on the use of effective rainfall vs total rainfall. Jain and Srinivasulu (2004a), on the other hand, compared an ANN developed using total rainfall with what they referred to as a gray box ANN that used effective rainfall and other conceptual model inputs. The gray box model outperformed the ANN using total rainfall. Therefore, some success with this type of rainfall input has been reported in the literature.

The effective rainfall in this study was calculated using the TFF BASIN model (see section 4.3). As with the modelling undertaken in Chapter 4, the same error measures were calculated for each experiment except for PDIFF and GRI as explained in section 4.6. A summary of experiments 1 to 3 is provided in Table 5.2.

Table 5.2: An outline of three experiments to test the effect of additional inputs to the ANN

Expt	Inputs	Output
1	Ripetta (t), Lunghezza (t to t-5), Alviano (t to t-19), Ponte Felice (t to t-11) and Orte (t to t-11)	Water level at Ripetta 12 hours ahead for the 2008 flood event
2	Above + Hourly rainfall average from area 5 (t to t-17)	
3	Above + Effective rainfall from area 5 (t to t-17)	

The results of the three experiments in terms of the performance measures are provided in Table 5.3. The results from Chapter 4 for the ANN are provided as a reference. The first thing to note is the grey shading in the table, which indicates the experiment with the best result according to individual performance measures. It is clear that adding rainfall improves the model, which is logical since rainfall is a driver of runoff. The model with total rainfall outperformed on all performance measures with the exception of the MPE. Experiment #3, which involved adding effective rainfall, did not produce a better result than using total rainfall unlike that found by Jain and Srinivasulu (2004a). Experiment #1, which involved adding more upstream stations, also showed improvements over the simple model developed in Chapter 4. This is also unsurprising as additional information was provided to the network to aid in modelling the flood wave at Ripetta.

Table 5.3: Performance measures for Expts #1 to 3. Grey shading denotes the best performing model.

Type of Measure	Performance Measure	Expt #1 No rainfall	Expt #2 Total rainfall	Expt #3 Effective rainfall	Results from Chapter 4
Absolute	ME	0.074	0.012	-0.294	0.27
	MAE	0.361	0.286	0.587	0.39
	MdAE	0.251	0.172	0.440	0.43
	RMSE	0.495	0.422	0.766	0.48
Relative	MPE	0.131	-0.352	-3.660	2.05
	MAPE	4.439	3.230	6.323	5.32
	MdAPE	3.420	2.034	4.253	4.56
	RMSPE	5.860	5.530	8.892	6.43
	CE	0.941	0.954	0.822	0.90
	PI	0.721	0.881	0.828	0.87
	PI.MAE	0.452	0.656	0.583	0.48
	PI.MdAE	0.449	0.626	0.582	0.32

Visual inspection of the hydrographs was then undertaken, which are provided in Figures 5.1 and 5.2. Figure 5.1 shows the observed and predicted values for the 2008 flood event for Experiment #1 while Experiments #2 and #3 are shown in Figures 5.2 a and b respectively. The model prediction is shown as an average of the 50 ANN model runs (as a blue line) but the individual model runs are also plotted as thin black lines. This reveals some interesting behaviour. Firstly it is surprising how wide the model predictions are at the peaks, implying that weight initialisation is having a much more profound effect than originally anticipated. Moreover, when hourly rainfall is used as an input to the model (Figure 5.2a), the 50 runs show an even wider spread than that shown in Figure 5.1, where no rainfall was used. The results continue to be even more pronounced when effective rainfall was used in place of hourly rainfall (Figure 5.2b). This means that the ANN is very sensitive to the rainfall pattern. In fact, both the hourly and effective rainfall are characterised by an impulsive pattern (plotted at the top of

Figures 5.2 a and b), which appears to be having a noticeable effect on the different ANNs across the 50 runs. However, once the 50 runs are averaged, the ANN still outperforms models without rainfall. Looking at predictions, however, reveals that the addition of effective is resulting in a good prediction, although somewhat oscillatory. It predicts the hydrograph slightly early and then overpredicts the peak. The shape, in general, is reasonably predicted. Visual inspection, therefore, confirms the findings of Jain and Srinivasulu (2004a).

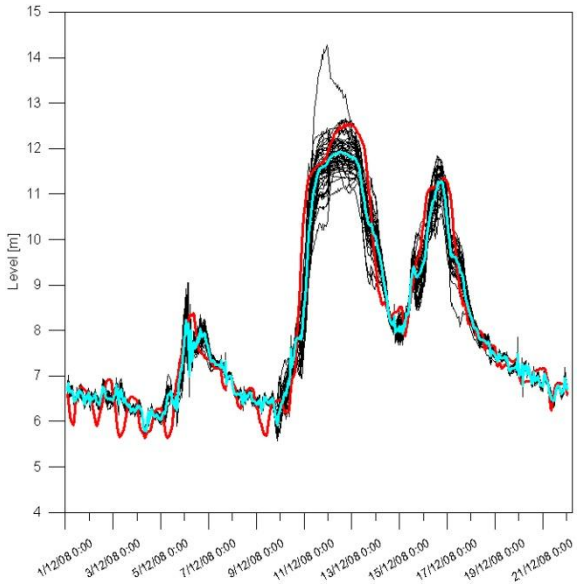


Figure 5.1: 2008 flood event computed without rainfall (Expt #1) where the red line is the observed and the blue is the mean of the 50 simulations, shown in black.

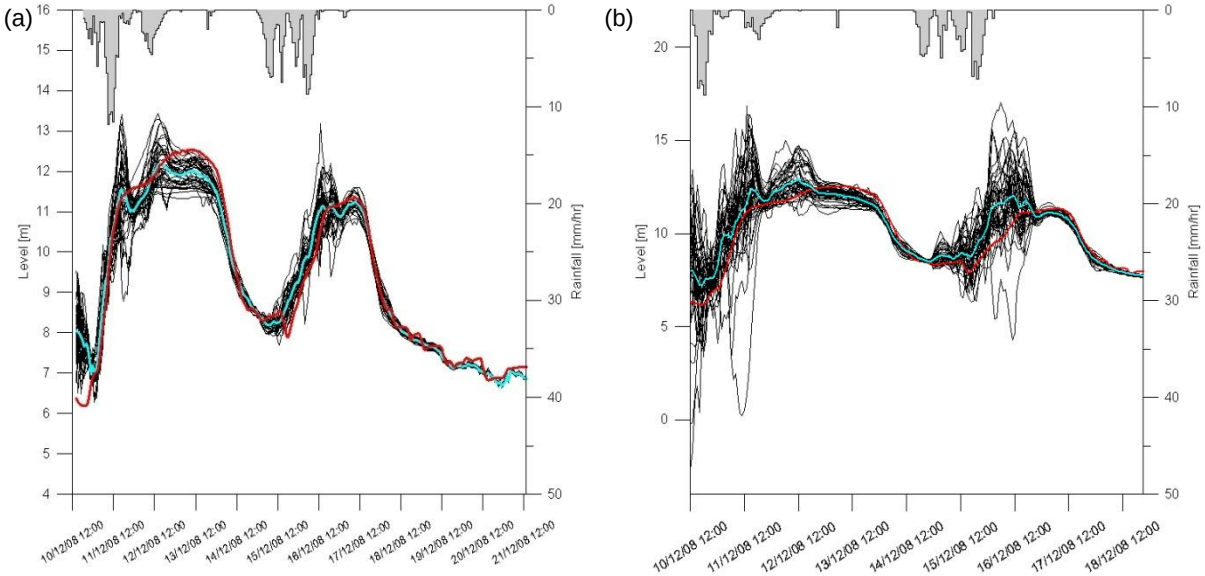


Figure 5.2: The 2008 flood event computed with (a) hourly rainfall (Expt #2) and (b) effective rainfall (Expt #3), where the red line is the observed and the blue is the mean of the 50 individual simulations, which are shown in black. Observed rainfall is shown in grey at the top of (a) and effective rainfall is shown at the top of (b).

In the next set of experiments, correlation coefficients are used to create a more parsimonious model.

5.3 Creating a More Parsimonious Model

Correlation coefficients were calculated between Ripetta at t+12 hours and the upstream stations. The highest correlations were those with the gauging station at Orte. The correlations between Ripetta and other stations varied from about 0.4 to 0.7. In addition, average rainfall from areas 5 to 7 was used. Therefore two experiments were tried as set out in Table 5.4, one with many upstream stations and one with only Orte. Trying to find a more parsimonious model has the advantage of reducing the training time significantly. The data set for training consisted of 4 years of data (2004-2007) while the output was once again the water level at Ripetta gauging station 12 hours ahead for the 2008 flood.

Table 5.4: An outline of two experiments to test parsimony

Expt	Inputs	Output
4	Alviano (t to t-19), Ponte Felice (t to t-11), Lunghezza (t to t-5), Average rainfall from areas 5 to 7 (t to t-17), Ripetta (t)	Water level at Ripetta 12 hours ahead for the 2008 flood event
5	Orte (t to t-11), Average rainfall from areas 5 to 7 (t to t-17), Ripetta (t)	

Table 5.5 contains the performance measures for Experiments #4 and 5 with the grey shading indicating the best performing model based on different performance measures. The absolute and relative measures consistently indicate that Experiment #4 is better with the exception of RMSE and its counterpart RMSPE while the benchmark measures, however, indicate superior performance by the model in Experiment #5. Thus based on the quantitative assessment, it is difficult to say whether the parsimonious model is better than the non-parsimonious one because of the contradiction between the non-redundant measures. A visual inspection is therefore a necessity.

Table 5.5: Performance measures for Expts #4 and 5. Grey shading denotes the best performing model.

Type of Measure	Performance Measure	Expt #4	Expt #5
Absolute	ME	0.193	0.405
	MAE	0.538	0.584
	MdAE	0.371	0.447
	RMSE	0.743	0.717
Relative	MPE	1.740	3.585
	MAPE	5.745	5.958
	MdAPE	4.094	5.439
	RMSPE	7.781	7.029
	CE	0.892	0.893
	PI	0.661	0.720
	PI.MAE	0.402	0.408
	PI.MdAE	0.314	0.296

Figure 5.3 shows that Experiment #4 with more inputs has a broader range of predictions across the 50 runs than the more parsimonious model. This may simply be because the ANN has much more information to handle, which is propagated through the random initialisation of the ANN weights. The average predictions are generally similar although the model with the full inputs is slightly better on the rising limb of the event.

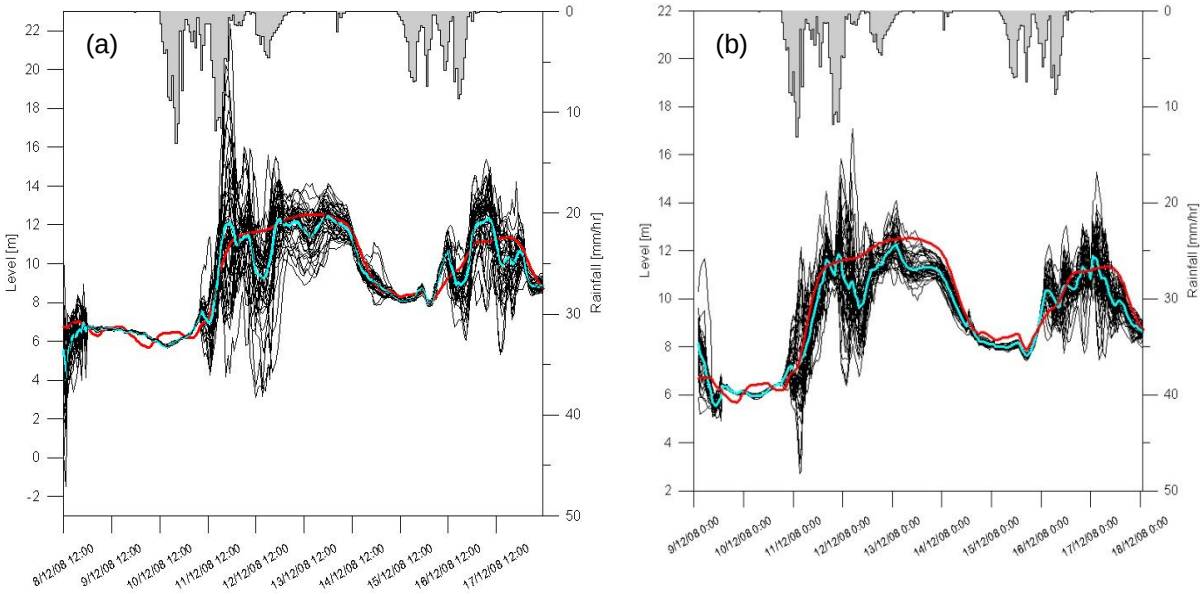


Figure 5.3: The results for (a) Expt#4 and (b) Expt #5 for a lead time of 12 hours for 2008 flood event. The red line is the observed, the blue is the average of the 50 simulations, which are shown in black, and observed rainfall is shown at the top of each figure in grey.

The results from this analysis are somewhat inconclusive. For this reason, the more parsimonious model (with shorter training times) is used in later experiments.

5.4 Adding a Difference Term

Another method was tried, which deals specifically with stationarity. Each station was lagged by its travel time (t_t) and then an additional input was used, i.e. the 12 hour difference (or DELTA) between the level at time t and t-12 for each station. Two experiments were run to test out this configuration and are listed in Table 5.6.

Table 5.6: An outline of two experiments in which DELTA inputs were used

Expt	Inputs	Output
6	Alviano (delta 12), Ponte Felice (delta 12), Lunghezza (delta 12), Average rainfall, Ripetta (delta 12)	Water level at Ripetta 12 hours ahead for the 2008 flood event
7	Orte (delta 12), Average rainfall, Ripetta (delta 12)	

The results of the two experiments in terms of the performance measures are provided in Table 5.7. The grey shading clearly shows that the model in Experiment #6 outperforms the one in Experiment #7, i.e. the less parsimonious model on the majority of performance measures with the exception of the RMSPE and the PI. However, if

these experiments are compared to Experiments #4 and #5, the DELTA terms have not improved the model performance. Instead these ANN models show worsening performance. Figure 5.4 confirms this result. The average of the 50 runs for both models is now underpredicting the observed peak.

Table 5.7: Performance measures for Expts #6 and #7. Grey shading denotes the best performing model.

Type of Measure	Performance Measure	Expt #6	Expt #7
Absolute	ME	0.451	0.638
	MAE	0.601	0.711
	MdAE	0.399	0.532
	RMSE	0.857	0.890
Relative	MPE	4.163	5.909
	MAPE	6.243	6.943
	MdAPE	4.609	6.417
	RMSPE	8.317	8.160
	CE	0.856	0.836
	PI	0.550	0.570
	PI.MAE	0.332	0.280
	PI.MdAE	0.260	0.163

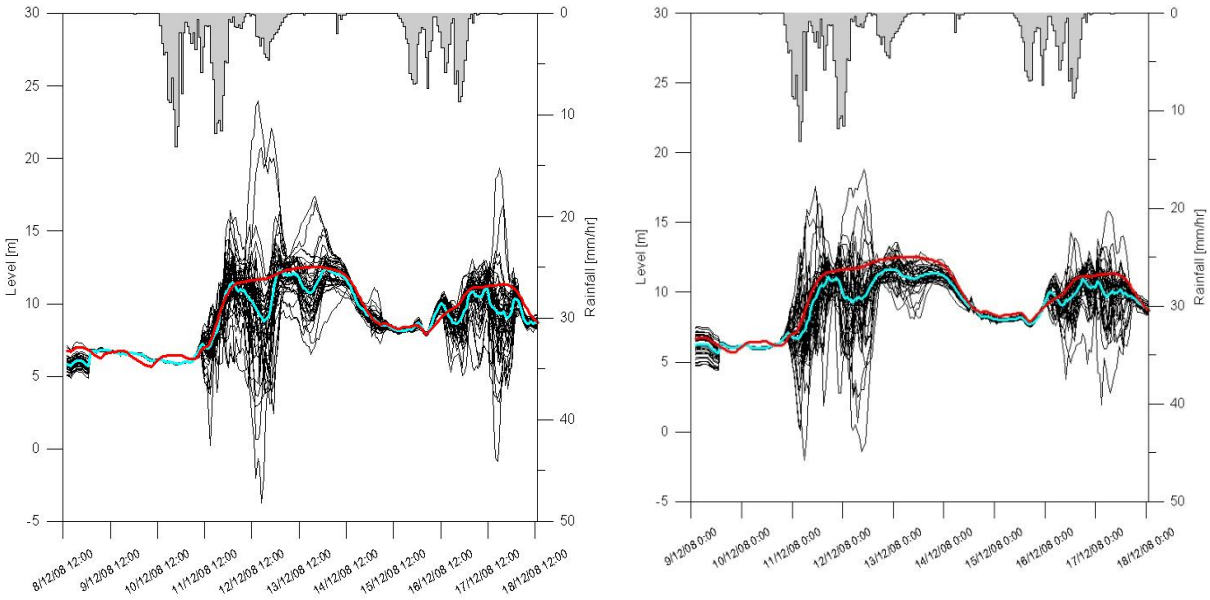


Figure 5.4: Results for a) Expt #6 and b) Expt #7 for the 2008 flood event. The red line is the observed and the blue is the mean of the 50 simulations, which are shown individually in black. Observed hourly rainfall is shown in grey at the top of both figures.

5.5 Lengthening the Times Series

Additional data were obtained for the period between 1993 and 2008. As the dataset was too long for the ANN to handle, flood events were chosen from the series using the criteria that events must exceed 7m. Two days of data before the peak and three days after were also extracted. In addition, cumulative rainfall was used instead of hourly or effective rainfall. It was thought that the smoothing effect of a cumulative rainfall series

might reduce the range of predictions across the 50 runs at the peaks. Only one experiment is reported (Table 5.8) using the upstream station with the highest correlation with Ripetta.

Table 5.8: An outline of one experiment in which cumulative rainfall was used

Expt	Inputs	Output
8	Orte (t to t-11), Cumulative average rainfall, Ripetta (t)	Water level at Ripetta 12 hours ahead for the 2008 flood event

The performance measures for this experiment are listed in Table 5.9 and the results for the 2008 flood event are shown in Figure 5.5. The results are much improved compared to Expts #4 to 7. The ANN forecasts are good on the rising limb of the hydrograph although the peaks are still underpredicted. Thus, effective rainfall in combination with all the upstream station appears to produces the best model overall.

Table 5.9: Performance measures for Expt #8

Type of Measure	**Performance Measure**	**Expt #8**
Absolute	ME	0.162
	MAE	0.296
	MdAE	0.275
	RMSE	0.366
Relative	MPE	1.632
	MAPE	3.611
	MdAPE	3.401
	RMSPE	4.367
	CE	0.968
	PI	0.885
	PI.MAE	0.601
	PI.MdAE	0.426

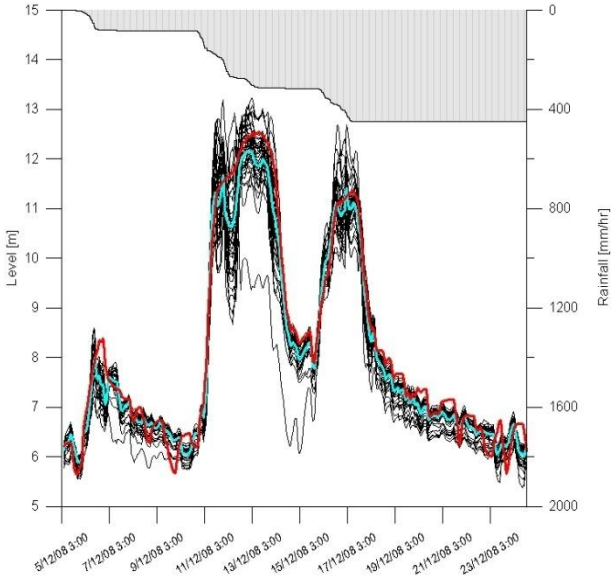


Figure 5.5: Results for Expt #8 for the 2008 flood event. The red line is the observed and the blue is the mean of the 50 simulations, which are shown individually in black. Cumulative rainfall is shown in grey at the top of the figure.

In the next section, the method of normalisation of the input data is the subject of further experimentation.

5.6 Changing the Method of Normalisation

In this set of experiments the focus is to use a continuous dataset rather than individual events. Therefore, a shorter time period of 2004 to 2008 is selected to examine the effect of changes to the method of normalisation. Moreover, the ANN is trained 100 times instead of 50. Normalisation, as explained in Chapter 2, is a simple method of pre-processing. Two of the most commonly used methods are the mean and standard deviation (hereafter referred to as MapStd), and the minimum and maximum (or MapMinMax). For the pre-processing method that uses the mean and standard deviation, the inputs are pre-processed so that the input signals have a mean of 0 and a standard deviation of 1 as follows:

$$\bar{x}_i = \frac{x_i - \rho(x_i)}{\sigma(x_i)} \tag{5.1}$$

for i =1,.., n , where x_i are the inputs, $\bar{x}$ are the normalized network input signals, $\rho(x_i)$ and $\sigma(x_i)$ are, respectively, an estimation of the mean and the standard deviation of the input variables. The output signals are then unnormalised as follows:

$$y_i = \rho\,(y_i) + \sigma(y_i)\overline{y_i} \tag{5.2}$$

for i =1,..,m, where y_i are the unnormalised outputs, $\overline{y_i}$ are the normalised network outputs, and $\rho(y_i)$ and $\sigma(y_i)$ are estimates of the mean and standard deviation of the outputs variables, respectively.

For the MapMinMax method, the inputs are scaled to the range [-1; 1] as follows:

$$\bar{x}_i = 2\frac{x_i - min(x_i)}{max(x_i) - min(x_i)} - 1 \tag{5.3}$$

where $min(x_i)$ and $max(x_i)$ are the minimum and the maximum values of the input vectors, respectively. The output is then post-processed as:

$$y_i = 0.5(\overline{y_i} + 1)\big(max(y_i) - min(y_i)\big) + min(y_i) \tag{5.4}$$

where $min(y_i)$ and $max(y_i)$ are the minimum and the maximum values of the unnormalised output, respectively. However, this was modified in this particular set of

experiments to scale the inputs to the range [-0.95; 0.05]. The reason for this was to determine whether forcing the ANN to train within a narrower range might lead to better extrapolation to extreme events like that seen in 2008. Table 5.10 lists the eight experiments that are used to examine the effects of the normalisation method. This time the impact on the 2005 event is also included in the experiments.

Table 5.10: An outline of eight experiments in which two methods of normalisation were examined

Expt	Inputs	Output	Method of Normalisation
9	Orte (t to t-11), Ripetta(t), No rainfall	Ripetta at t+12, 2005 event	MapMinMax
10	Orte (t to t-11), Ripetta(t), No rainfall		MapStd
11	Orte (t to t-11), Ripetta(t), Cumulative average rainfall		MapMinMax
12	Orte (t to t-11), Ripetta(t), Cumulative average rainfall		MapStd
13	Orte (t to t-11), Ripetta(t), No rainfall	Ripetta at t+12, 2008 event	MapMinMax
14	Orte (t to t-11), Ripetta(t), No rainfall		MapStd
15	Orte (t to t-11), Ripetta(t), Cumulative average rainfall		MapMinMax
16	Orte (t to t-11), Ripetta(t), Cumulative average rainfall		MapStd

Considering first the 2005 event, the flood hydrographs are provided in Figure 5.6 and the performance measures are provided in Table 5.11.

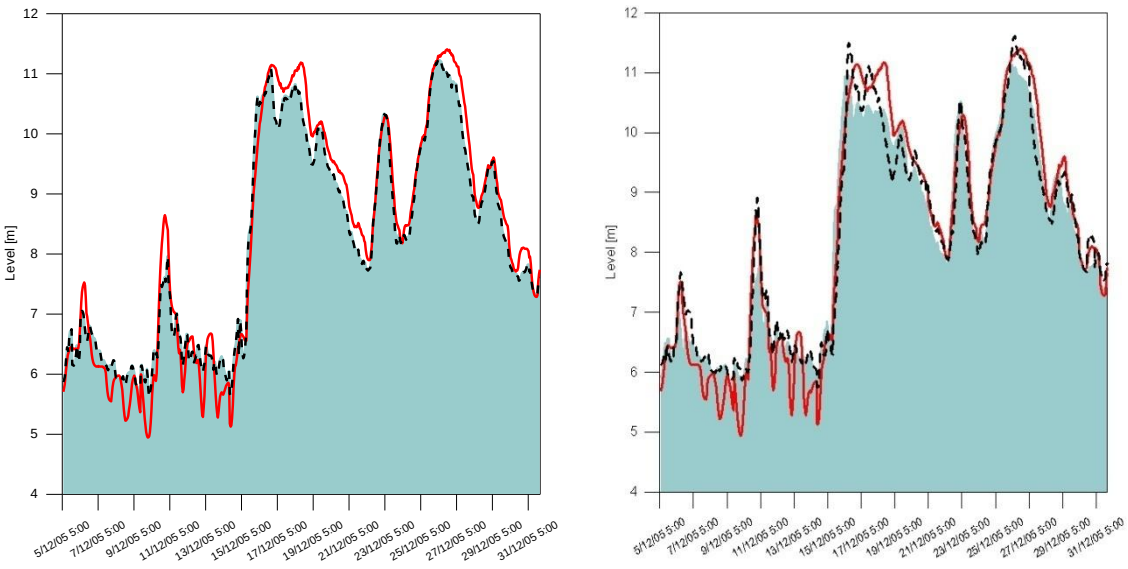


Figure 5.6: Results for a) Expts #9 and 10 and b) Expts #11 and 12 for the 2005 flood event. The red line is the observed, the dotted black line is the MapStd method and the light blue filled in area is the MapMinMax method.

Table 5.11: Performance measures for Expts #9 to 12 for the 2005 flood event. Grey shading denotes the best performing model overall while bold denotes the best performance between pairs of experiments, i.e. between Expts # 9 and #10, and between Expts # 11 and #12.

		Without rainfall		With rainfall	
Type of Measure	**Performance Measure**	**Expt #9 MapMinMax**	**Expt #10 MapStd**	**Expt #11 MapMinMax**	**Expt #12 MapStd**
Absolute	ME	**0.089**	**0.089**	0.087	**0.014**
	MAE	0.312	**0.311**	0.365	**0.333**
	MdAE	**0.288**	**0.288**	0.308	**0.250**
	RMSE	0.383	**0.381**	0.440	**0.431**
Relative	MPE	0.363	**0.354**	**0.219**	-0.728
	MAPE	4.069	**4.059**	4.637	**4.375**
	MdAPE	**3.478**	3.562	3.895	**3.042**
	RMSPE	5.216	**5.201**	**5.824**	5.976
	CE	0.960	**0.961**	0.947	**0.950**
	PI	0.796	**0.799**	0.730	**0.741**
	PI.MAE	0.529	**0.531**	0.449	**0.498**
	PI.MdAE	**0.456**	**0.456**	0.419	**0.528**

Looking at Figure 5.6a, which considers the two normalisation methods when no rainfall is used as an input to the ANN, there is almost no difference in the predictions. The rising limbs of the hydrographs are predicted well but the peaks are slightly underpredicted. Figure 5.6b compares the same two normalisation methods but with rainfall added to the ANN. This time the MapStd method outperforms the MapMinMax method and slightly overpredicts the peaks while the MapMinMax methods leads to an underprediction. This is confirmed when examining the performance measures in Table 5.11. When no rainfall is used (Experiments #9 and #10), the boldface shows that the MapStd method (Experiment #10) performs better. However, the numbers are virtually identical between experiments so there is very little difference between the two methods. When considering the addition of rainfall (Experiments #11 and #12), the MapStd method outperforms the MapMinMax method across the majority of performance measures with the exception of MPE and RMSPE. Moreover, if the grey shading is examined, which indicates the model that performs best across all four experiments, the MapStd method with no rainfall (Experiment #10) would be chosen. Yet the hydrographs would suggest that Experiment #12 is the best as the peaks are slightly overestimated, which is better from an operational flood forecasting perspective. Thus both quantitative measures and visual inspection of the hydrograph are once again necessary.

Turning to the 2008 flood event, the hydrographs are provided in Figure 5.7 and the performance measures are listed in Table 5.12.

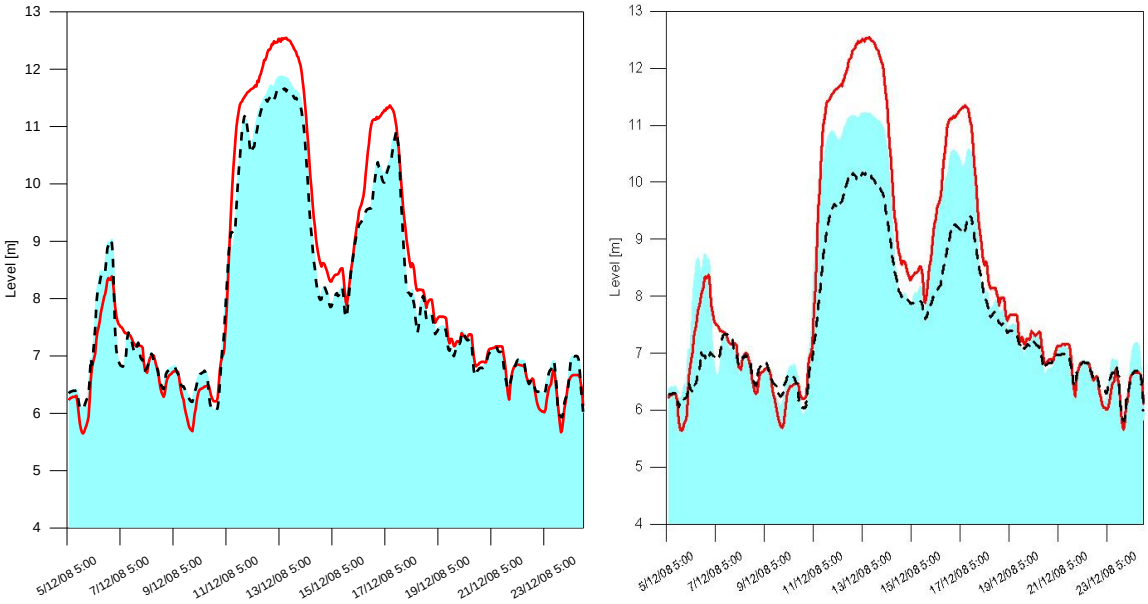


Figure 5.7: Results for a) Expts #13 and 14 and b) Expts #15 and 16 for the 2008 flood event. The red line is the observed, the dotted blue line is the MapStd method and the light blue filled in area is the MapMinMax method.

Once again, the MapStd method and MapMinMax methods are very similar when no rainfall is added to the ANN. A clear underprediction of the peak can also be seen. When adding rainfall, the MapMinMax method now outperforms the MapStd method although both still considerably underpredict the peak. The performance measures in Table 5.12 confirm these results. Without rainfall, the MapMinMax method appears to outperform the MapStd method but the numbers are very similar across both experiments. With rainfall, the MapMinMax method clealy outperforms the MapStd method across all performance measures.

Table 5.12: Performance measures for Expts #13 to 16 for the 2008 flood event. Grey shading denotes the best performing model overall while bold denotes best performance between pairs of experiments, i.e. between Expts #13 and #14, and between Expts #15 and #16.

		Without rainfall		With rainfall	
Type of Measure	**Performance Measure**	**Expt #13 MapMinMax**	**Expt #14 MapStd**	**Expt #15 MapMinMax**	**Expt #16 MapStd**
Absolute	ME	**0.208**	0.219	**0.371**	0.644
	MAE	**0.367**	**0.376**	**0.476**	0.732
	MdAE	**0.269**	0.270	**0.187**	0.368
	RMSE	**0.488**	0.504	**0.709**	1.087
Relative	MPE	**1.712**	1.790	**3.647**	5.980
	MAPE	**4.103**	4.172	**5.246**	7.429
	MdAPE	3.787	**3.751**	**2.824**	4.997
	RMSPE	**5.106**	5.178	**7.420**	10.075
	CE	**0.942**	0.939	**0.818**	0.715
	PI	**0.808**	0.795	**0.404**	0.047
	PI.MAE	**0.543**	0.531	**0.276**	0.088
	PI.MdAE	**0.485**	0.482	**0.584**	0.295

Therefore, the results show that MapStd is better for the 2005 event (although only marginally so) and that the MapMinMax method is better for the 2008 event (although peak predictions were very poor). This is most likely a result of the [-0.95; 0.05] range chosen over which to normalise. Instead of helping the ANN extrapolate, the narrow range had an adverse impact on model performance, in particular in peak prediction. Moreover, the two different normalisation methods do not perform consistently across events so it is not so straightforward to make a recommendation about which normalisation method to use. However, as the MapMinMax method performed better for the 2008 event, this method will continue to be used throughout the rest of the thesis. The 2008 event is outside of the range of the training data. Perhaps in that situation the MapMinMax works better as a normalisation method.

5.7 Experimentation with Different Training Algorithms

In this set of experiments, different training algorithms were used. Up to now, all experiments used Bayesian Regularisation. In addition, the Levenberg-Marquardt backpropagation and the BFGS quasi-Newton backpropagation algorithms were also employed. A set of twelve different experiments is outlined in Table 5.13. Experiments continue to use the more parsimonious form of the model, i.e. inclusion of only the upstream gauging station Orte and rainfall. Moreover, Experiments #20 to #22 and Experiments #26 to #28 include the 2005 and 2008 flood events in the training dataset. This is not a true test of the ANN as the independent data must not be in the training data. However, it is merely to see the effect on the results, i.e. does seeing these events (and their peaks) make a significant difference to the results?

Table 5.13: An outline of twelve experiments in which different training algorithms were used

Expt	Inputs	Output	Training Method
17	Orte (t to t-11), Ripetta(t), Cumulative average rainfall	Ripetta at t+12, 2005 event	BR
18			BFGS
19			LM
20	Orte (t to t-11), Ripetta(t), Cumulative average rainfall, 2005 and 2008 in the training data set		BR
21			BFGS
22			LM
23	Orte (t to t-11), Ripetta(t), Cumulative average rainfall	Ripetta at t+12 2008 event	BR
24			BFGS
25			LM
26	Orte (t to t-11), Ripetta(t), Cumulative average rainfall, 2005 and 2008 in the training data set		BR
27			BFGS
28			LM

Note: Expt#17 is the same as Expt #11; Expt #23 is the same as Expt#23

Table 5.14 contains the performance measures for the six experiments pertaining to the 2005 flood event and Figure 5.8 contains the flood hydrographs. Using the quantitative performance measures from Table 5.14, Experiment #17, which uses the BR algorithm, performs the worst overall with some measures favouring Experiment #18 (BFGS algorithm) and others better in Experiment #19 (LM algorithm). When adding the 2005 and 2008 flood events in the training dataset, the LM performs best overall. However, looking at the absolute differences between the measures reveals very little difference between them.

Table 5.14: Performance measures for Expts #17 to 22 for the 2005 flood event. Grey shading denotes the best performing model.

Performance Measure	**Orte, Ripetta and Rainfall**			**Orte, Ripetta, Rainfall and 2005/2008 events in the calibration**		
	Expt #17 BR	**Expt #18 BFGS**	**Expt #19 LM**	**Expt #20 BR**	**Expt #21 BFGS**	**Expt #22 LM**
ME	0.087	**0.078**	0.086	0.051	**0.044**	0.052
MAE	0.365	**0.356**	**0.356**	0.346	0.346	**0.339**
MdAE	0.308	**0.290**	0.307	0.307	**0.300**	0.304
RMSE	0.440	0.429	**0.424**	0.416	0.417	**0.403**
MPE	0.219	**0.116**	0.197	**-0.105**	-0.225	-0.109
MAPE	4.637	4.545	**4.534**	4.470	4.481	**4.387**
MdAPE	3.895	**3.653**	3.778	3.738	3.705	**3.697**
RMSPE	5.824	5.765	**5.683**	5.687	5.719	**5.541**
CE	0.947	0.950	**0.951**	0.953	0.953	**0.956**
PI	0.730	0.744	**0.750**	0.759	0.758	**0.775**
PI.MAE	0.449	**0.463**	**0.463**	0.478	0.478	**0.488**
PI.MdAE	0.419	**0.452**	0.420	0.421	**0.434**	0.426

Figure 5.8 shows a reasonably similar performance between the different algorithms and in the experiments where the 2005 and 2008 flood events have been added. The timing of the rising limb is very good but there are still issues with underprediction of the peak in the event. Looking at the individual members of the ensemble, the variation in prediction does increase at the peak across the ensemble but there is a generally a small band of variation around the rest of the hydrograph.

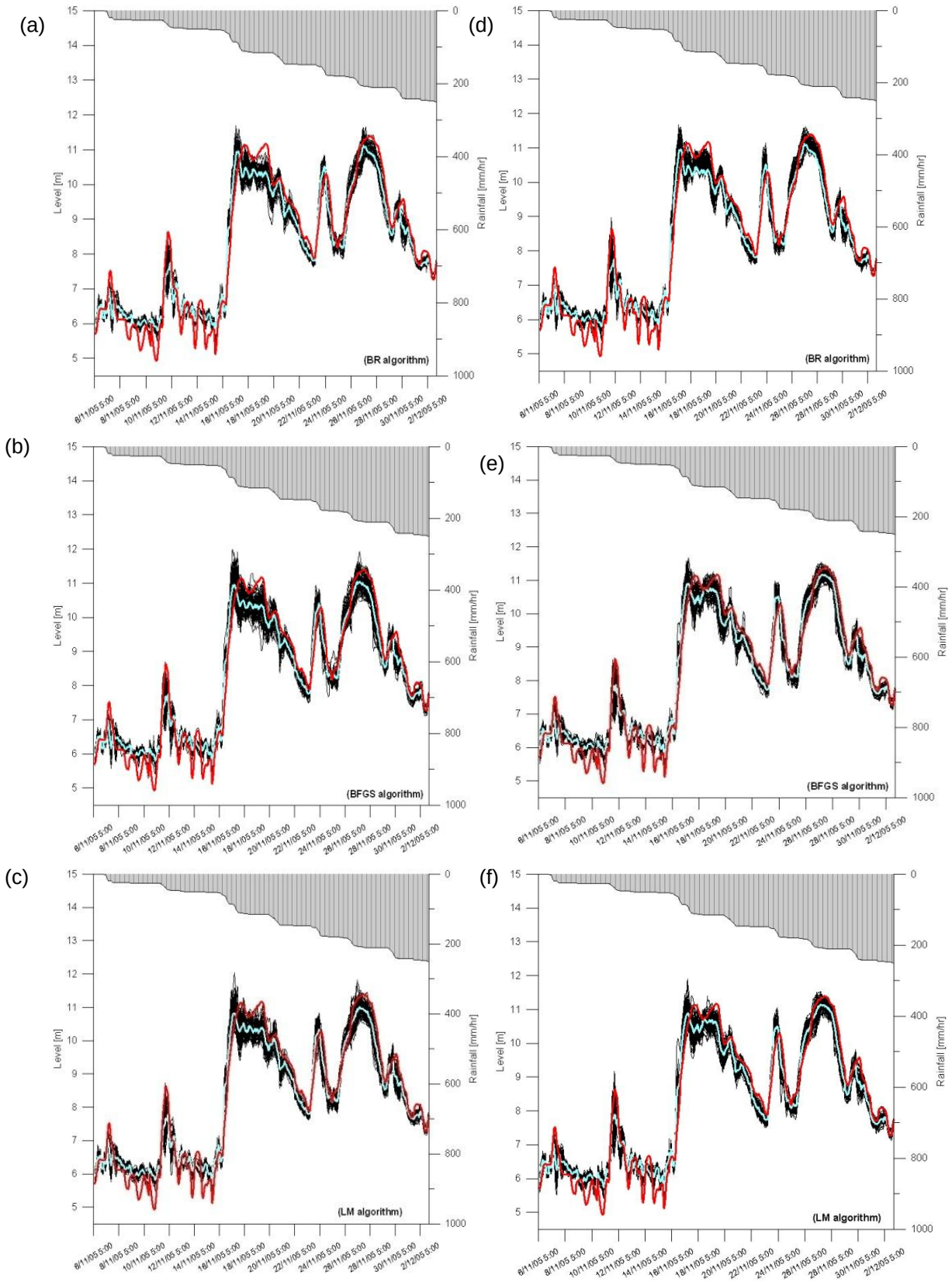


Figure 5.8: Results for a) Expts #17; b) Expt #18; c) Expt #19; d) Expt #20; e) Expt #21; and f) Expt #22 for the 2005 flood event. The red line is the observed and the blue line is the average of the predictions, shown individually in black. Cumulative rainfall is plotted on the top of each graph.

The same is repeated in Table 5.15 and Figure 5.9 for the 2008 flood event. This time the BFGS algorithm performs better compared to the BR and LM in most performance measures for the case where the 2005 and 2008 flood events have not been included in the training data set. When they have been included, then the BR algorithm generally outperforms the others.

Table 5.15: Performance measures for Expts #23 to 28 for the 2008 flood event. Grey shading denotes the best performing model.

Performance Measure	**Orte and Rainfall**			Orte+rainfall 2008+2005 in calibration		
	Expt #23 BR	**Expt #24 BFGS**	**Expt #25 LM**	**Expt #26 BR**	**Expt #27 BFGS**	**Expt #28 LM**
ME	0.371	**0.309**	0.347	**0.389**	0.426	0.417
MAE	0.476	**0.430**	0.459	**0.519**	0.527	0.521
MdAE	**0.187**	0.246	0.228	0.250	0.249	**0.241**
RMSE	0.709	**0.601**	0.693	**0.765**	0.789	0.783
MPE	3.647	**2.764**	3.096	**3.413**	3.809	3.712
MAPE	5.246	**4.633**	4.808	**5.402**	5.409	5.349
MdAPE	**2.824**	3.449	3.192	3.726	**3.335**	3.474
RMSPE	7.420	**5.864**	6.451	**7.152**	7.254	7.182
CE	0.818	**0.913**	0.884	**0.859**	0.850	0.852
PI	0.404	**0.708**	0.613	**0.528**	0.498	0.505
PI.MAE	0.276	**0.464**	0.428	**0.353**	0.343	0.351
PI.MdAE	**0.584**	0.529	0.564	0.522	0.523	**0.538**

Examining the hydrographs in Figure 5.9 reveals a much wider amount of variation amongst the ensemble members for the 2008 flood event. This is even more evident when the 2005 and 2008 flood events were included for the BFGS algorithm but less so for the BR and LM algorithms. Looking purely at the hydrographs, the best performing experiment is #25 if the basis of judgement is the rising limb and the peak prediction before rapidly decreasing and missing the rest of the event. It the basis of judgement is the volume of the hydrograph, then BR in Experiment #23 would be the best one. The quantitative measures and the visual inspection are therefore quite different in this example. Moreover, there is currently no generalisable pattern regarding the best algorithm to use. Thus, a greater impact on the results would most likely come from better inputs to the ANN than changing the algorithm per se.

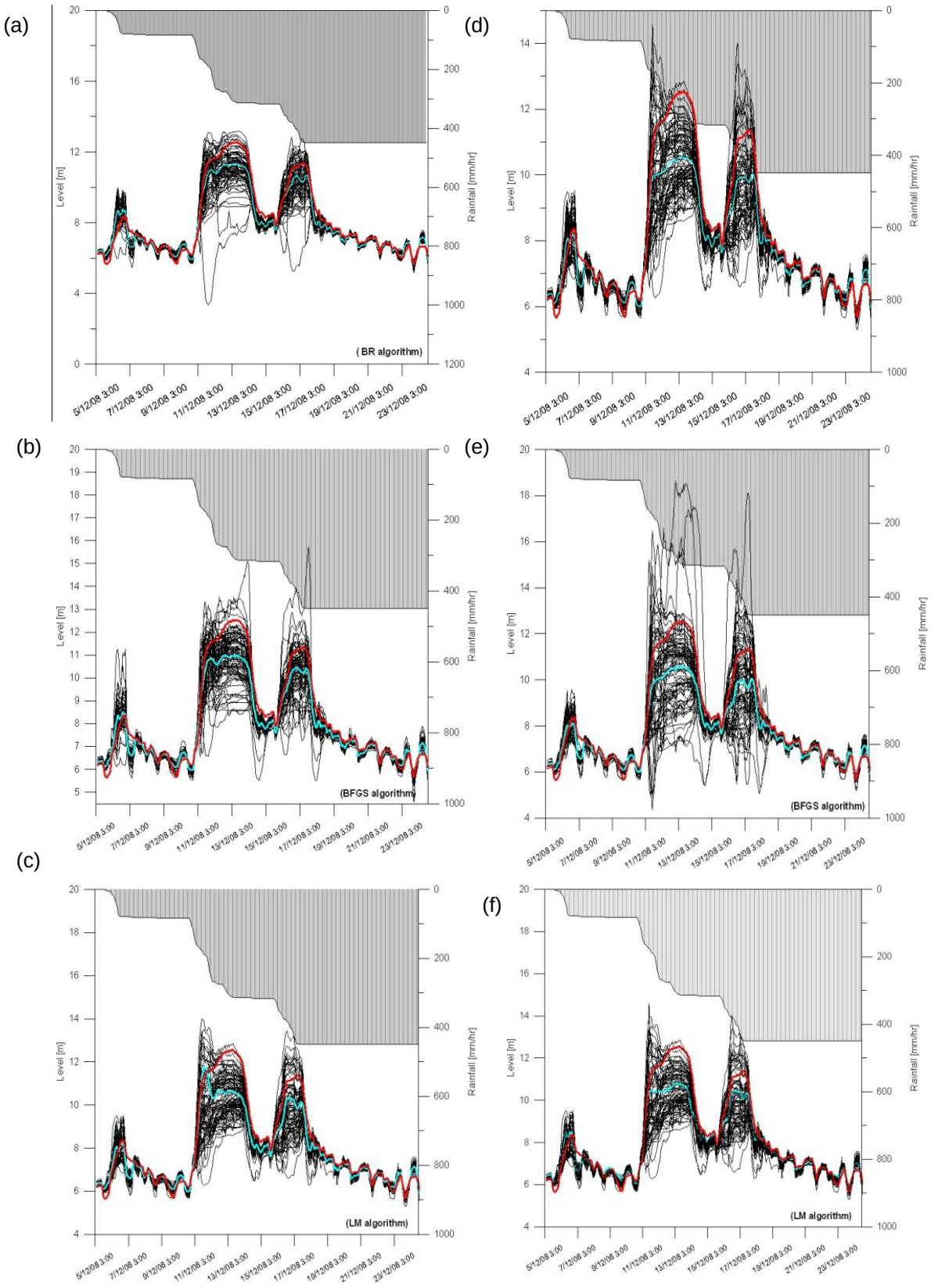


Figure 5.9: Results for a) Expts #23; b) Expt #24; c) Expt #25; d) Expt #26; e) Expt #27; f) Expt #28 for the 2008 flood event. The red line is the observed and the blue line is the average of the predictions. Cumulative rainfall is plotted on the top of each graph.

5.8 Exploring PI as an Alternative Method to Combine the Ensemble

Up to this point, the ANN ensemble model runs (of either 50 or 100) have simply been averaged in each experiment as this is one valid way to combine the ensemble members. However, there are many other methods available to combine the ensemble including the use of ANNs (e.g. See and Abrahart, 2001). In this set of experiments the PI performance measure was used with a threshold of 0.65 to choose the best ANN models from the ensemble. These chosen ANN runs were then further averaged to create a single ANN prediction per experiment. This approach has been taken in a series of experiments as listed in Table 5.16.

Table 5.16: An outline of twelve experiments in which the best simulations were chosen using a PI threshold

Expt	Inputs	Output	Training Method
29	Orte (t to t-11), Ripetta(t), Cumulative average rainfall	Ripetta at t+12, 2005 event	BR
30			BFGS
31			LM
32	Orte (t to t-11), Ripetta(t), Cumulative average rainfall, 2005 and 2008 in the training data set		BR
33			BFGS
34			LM
35	Orte (t to t-11), Ripetta(t), Cumulative average rainfall	Ripetta at t+12 2008 event	BR
36			BFGS
37			LM
38	Orte (t to t-11), Ripetta(t), Cumulative average rainfall, 2005 and 2008 in the training data set		BR
39			BFGS
40			LM

Figure 5.10 contains the hydrographs for the 2005 flood event while Table 5.17 contains the corresponding performance measures. Examination of Figure 5.10 shows that there is very little difference between the experiments except that when the 2005 and 2008 flood events are included in the training, the three training algorithms produced almost identical results.

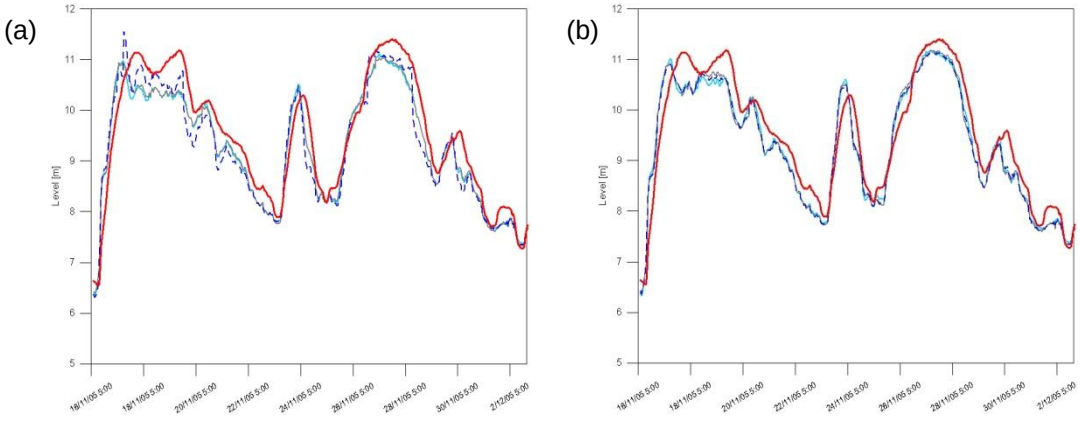


Figure 5.10: (a) Expts # 29 to 31 and b) Expts #32 to 34 for the 2005 flood event where the red line is the observed, the dotted blue line is the LM algorithm, the light blue line is the BR algorithm and the gray line is the BFGS algorithm.

The performance measures in Table 5.17 confirm this result. For Experiments #29 to #31, the BFGS algorithm provides the best overall result while for Experiments #32 to #34, the LM algorithm performed best overall. However, the performance measures are quite similar across the different algorithms for both sets of experiments with and without the inclusion of 2005 and 2008 in the training dataset. This is perhaps not surprising because the final extra row of Table 5.17 shows the number of runs with PI greater than 0.65. This shows that most of the runs were kept using this criterion for selection.

Table 5.17: Performance measures for Expts #29 to 34 for the 2005 flood event. Grey shading denotes the best performing model while bold denotes best performance between triplets of experiments, i.e. between Expts #29 to #31, and between Expts #32 and #34.

Performance Measure	**Orte, Ripetta and Rainfall**			**Orte, Ripetta, Rainfall and 2005/2008 events in the calibration**		
	Expt #29 BR	**Expt #30 BFGS**	**Expt #31 LM**	**Expt #32 BR**	**Expt #33 BFGS**	**Expt #34 LM**
ME	0.088	**0.078**	0.104	0.051	**0.044**	0.052
MAE	0.363	**0.355**	0.371	0.346	0.346	**0.339**
MdAE	0.310	**0.289**	0.314	0.307	0.299	**0.304**
RMSE	0.436	**0.427**	0.465	0.416	0.417	**0.402**
MPE	0.246	**0.123**	0.463	-0.105	-0.224	**-0.103**
MAPE	4.606	**4.532**	4.698	4.470	4.475	**4.383**
MdAPE	3.889	3.673	**3.665**	3.738	3.700	**3.688**
RMSPE	5.769	**5.737**	6.014	5.687	5.711	**5.533**
CE	0.948	**0.951**	0.941	0.953	0.953	**0.956**
PI	0.736	**0.747**	0.699	0.759	0.759	**0.775**
PI.MAE	0.452	**0.464**	0.440	0.478	0.478	**0.489**
PI.MdAE	0.415	**0.455**	0.408	0.421	**0.436**	0.426
Number of runs with PI > 0.65	91/100	94/100	97/100	100/100	97/100	99/100

The results for the experiments concerned with modelling the 2008 event appear in Figure 5.11 and Table 5.18.

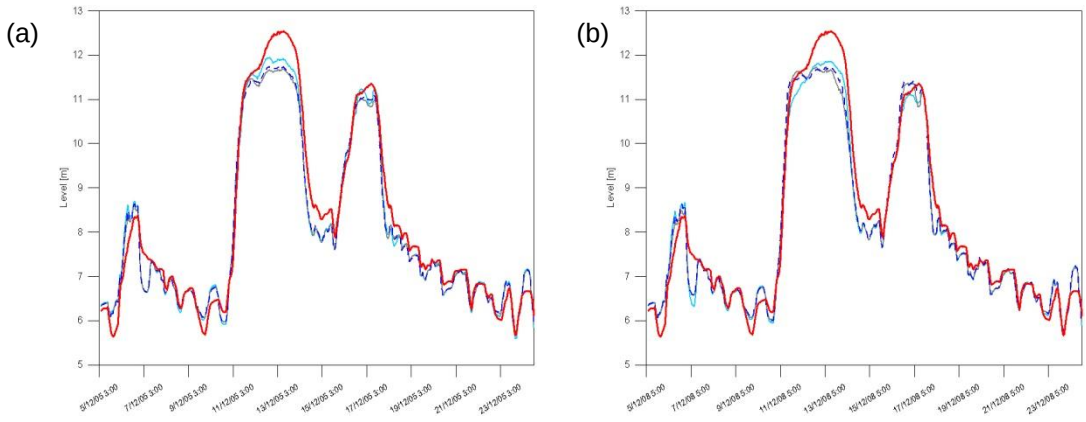


Figure 5.11 a) Expts # 35 to 37 and b) Expts #38 to 40 for the 2008 flood event where the red line is the observed, the dotted blue line is the LM algorithm, the light blue line is the BR algorithm and the gray line is the BFGS algorithm.

Regardless of whether the 2005 and 2008 events are used in the training dataset, the results for the peak event in 2008 are the same, i.e. an underprediction. In fact there is little to differentiate Figure 5.11a from Figure 5.11b based on an examination of the hydrographs alone.

The quantitative measures in Table 5.18 indicate that for Experiments #35 to #37, the BR outperforms the other two training algorithms although the values are quite similar across experiments. For Experiments #38 to #40 where the 2005 and 2008 flood events were included in the training data, the result is much more mixed even across redundant measures. Moreover, looking at the number of runs with a PI greater than 0.65, the number of runs included is generally less than 50% for Experiments #35 to #37 and less than 25% for Experiments #38 to #40. This indicates that the forecasting problem is obviously much harder for the 2008 flood event and the impact of the random initialisation of the weights is larger. Comparing the results to Table 5.15 where all the runs were included in the average, the results are actually much improved. This implies that in more difficult non-linear problems, the effect of weight initialisation will be larger and a simple averaging across the whole ensemble is not recommended.

Table 5.18: Performance measures for Expts #35 to 40 for the 2008 flood event. Grey shading denotes the best performing model while bold denotes best performance between triplets of experiments, i.e. between Expts #35 to #37, and between Expts #38 and #40.

Performance Measure	**Orte, Ripetta and Rainfall**			**Orte, Ripetta, Rainfall and 2005/2008 events in the calibration**		
	Expt #35 BR	**Expt #36 BFGS**	**Expt #37 LM**	**Expt #38 BR**	**Expt #39 BFGS**	**Expt #40 LM**
ME	**0.118**	0.167	0.147	0.137	0.133	**0.101**
MAE	**0.275**	0.295	0.291	**0.275**	0.282	0.282
MdAE	**0.206**	0.219	0.223	0.210	**0.181**	0.187
RMSE	**0.356**	0.387	0.378	**0.369**	0.399	0.384
MPE	**1.066**	1.515	1.308	1.224	1.150	**0.806**
MAPE	**3.366**	3.453	3.436	3.299	**3.276**	3.311
MdAPE	**2.656**	2.750	2.737	2.593	**2.353**	2.396
RMSPE	**4.323**	4.360	4.328	4.374	4.357	**4.317**
CE	**0.969**	0.964	0.965	**0.967**	0.962	0.964
PI	**0.898**	0.879	0.885	**0.890**	0.871	0.881
PI.MAE	**0.658**	0.633	0.638	**0.657**	0.649	0.649
PI.MdAE	**0.605**	0.581	0.573	0.597	**0.653**	0.642
Number of runs with PI > 0.65	45/100	43/100	61/100	25/100	26/100	21/100

5.9 Ensemble Modelling using the Akaike Information Criterion

In this final section, two different methods were employed to combine the ensemble ANN models based on the Akaike Information Criterion (AIC), which was developed by Hirotsugu Akaike (Akaike, 1973). This criterion is a tool for statistical model selection (Panchal et al., 2010). The AIC is calculated as follows:

$$AIC = n \times ln\left(\frac{RSS}{n}\right) + 2 \times k \tag{5.5}$$

where the RSS is the residual sum of squares, n is the number of samples in the data and k is the number of free parameters for each model considered. For a small sample size ($n/k < 40$), Sagiura (1978) proposed the following expression for the evaluation of the AIC:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \tag{5.6}$$

The AIC has two important characteristics. The first is that the AIC will equate to a maximum likelihood solution if the number of parameters in the model is fixed; the second is that if the model is performed with several numbers of parameters, the AIC can be used to select the model that fits the best but with the smallest number of parameters (Zhao et al., 2008). In order to compare models, the Delta AIC (Δ_i) can be calculated as follows:

$$\Delta_i = AIC_i - \min AIC \tag{5.7}$$

where the AIC_i is the AIC value of the i^{th} model and $\min AIC$ is the minimum value of all AIC values. The smallest value of the AIC and the Delta AIC correspond to the best model while the largest values represent the worst model. The Delta AIC is then used to calculate the weights (w_i) of each model as follows:

$$w_i = \frac{exp^{\left(-\frac{\Delta_i}{2}\right)}}{\sum_{j=1}^{n} exp^{(-\Delta_j/2)}} \tag{5.8}$$

where $\sum w_i = 1$ and w_i represents the probability that the model performs better compared with all other ensemble members (Burnham and Anderson, 2002). In ensemble modelling, the model predictions can be weighted by w_i and linearly summed to create a single model prediction. Zhao et al. (2008) proposed a modified Delta AIC, which is expressed as:

$$(\Delta_m)_i = 1 + \frac{\Delta_i - \Delta_{min}}{\Delta_{max} - \Delta_{min}} \times \beta \tag{5.9}$$

where β is a constant that measures the diversity of the model output in the training dataset:

$$\beta = \frac{Div_{max}}{Div_{min}} \tag{5.10}$$

and Div_i is defined as:

$$Div_i = \sqrt{\frac{(O_i - \bar{O})^2}{N_{tr}}} \tag{5.11}$$

where N_{tr} is the number of samples in the training dataset. From this, the modified Akaike weights (ω_i) can be calculated as follows:

$$\omega_i = \frac{1/(\Delta_m)_i}{\sum_{j=1}^{m} 1/(\Delta_m)_j} \tag{5.12}$$

Table 5.19 outlines 24 experiments that are undertaken to compare the effect of using the AIC and modified AIC to combine the ANN ensembles relative to the average. The experiments consider two variations of input variables, three training methods and two flood events, each of which use the AIC and modified AIC to combine the ANN ensemble.

Table 5.19: Twenty-four experiments with the AIC and the modified AIC for ensemble combination

Expt	Inputs	Training Method	Output	Ensemble Combination
41	Orte (t to t-11), Ripetta(t), Cumulative average rainfall	BR	Ripetta at t+12, 2005 event	AIC
42				Modified AIC
43		BFGS		AIC
44				Modified AIC
45		LM		AIC
46				Modified AIC
47	Orte (t to t-11), Ripetta(t), Cumulative average rainfall, 2005 and 2008 in the training data set	BR		AIC
48				Modified AIC
49		BFGS		AIC
50				Modified AIC
51		LM		AIC
52				Modified AIC
53	Orte (t to t-11), Ripetta(t), Cumulative average rainfall	BR	Ripetta at t+12 2008 event	AIC
54				Modified AIC
55		BFGS		AIC
56				Modified AIC
57		LM		AIC
58				Modified AIC
59	Orte (t to t-11), Ripetta(t), Cumulative average rainfall, 2005 and 2008 in the training data set	BR		AIC
60				Modified AIC
61		BFGS		AIC
62				Modified AIC
63		LM		AIC
64				Modified AIC

Table 5.20 contains the performance measures for the first 6 experiments to predict the 2005 flood event using Orte, Ripetta and rainfall only as inputs while Table 5.21 contains the second 6 experiments in which the 2005 and 2008 flood events were included in the training data set. The corresponding hydrographs for all of these experiments are provided in Figure 5.12. Overall, the modified AIC generally outperforms the pure AIC. Secondly the LM algorithm performs better overall although this is truer for the set of experiments that include the 2005 and 2008 flood events in the training data set. However, if these experiments are compared to the ensemble combination via averaging and the PI threshold, the experiments using AIC and modified AIC are generally worse for this event.

Table 5.20: Performance measures for Expts #41 to 46 for the 2005 flood event for inputs of Orte and rainfall only. Grey shading denotes the best performing model overall while bold denotes the best performance between pairs of experiments, i.e. between Expts #41 and #42, Expts #43 and #44 and between Expts #45 and #46.

Type of Measure	Performance Measure	BR		BFGS		LM	
		Expt #41 AIC	Expt #42 Modified AIC	Expt #43 AIC	Expt #44 Modified AIC	Expt #45 AIC	Expt 46 Modified AIC
Absolute	ME	**0.060**	0.091	**-0.023**	0.080	**0.056**	0.088
	MAE	**0.364**	0.365	0.380	**0.357**	0.403	**0.355**
	MdAE	0.317	**0.313**	0.331	**0.294**	0.322	**0.310**
	RMSE	0.444	**0.440**	0.487	**0.429**	0.520	**0.423**
Relative	MPE	**-0.066**	0.266	-0.939	**0.141**	**-0.137**	0.217
	MAPE	4.673	**4.632**	4.748	**4.553**	5.054	**4.524**
	MdAPE	**3.881**	**3.881**	3.750	**3.704**	4.012	**3.810**
	RMSPE	5.962	**5.813**	6.174	**5.766**	6.581	**5.669**
	CE	**0.947**	**0.947**	0.936	**0.950**	0.927	**0.952**
	PI	0.726	**0.731**	0.671	**0.744**	0.625	**0.752**
	PI.MAE	**0.450**	0.449	0.426	**0.461**	0.392	**0.464**
	PI.MdAE	0.401	**0.409**	0.375	**0.446**	0.393	**0.416**

Table 5.21: Performance measures for Expts #47 to 52 for the 2005 flood event for for Orte+rainfall 2008+2005 in calibration. Grey shading denotes the best performing model overall while bold denotes the best performance between pairs of experiments, i.e. between Expts #47 and #48, Expts #49 and #50 and between Expts #51 and #52.

Type of Measure	Performance Measure	BR		BFGS		LM	
		Expt #47 AIC	Expt #48 Modified AIC	Expt #49 AIC	Expt #50 Modified AIC	Expt #51 AIC	Expt #52 Modified AIC
Absolute	ME	**0.049**	0.053	**0.035**	0.045	**0.031**	0.053
	MAE	0.362	**0.345**	0.365	**0.344**	0.373	**0.337**
	MdAE	0.305	**0.304**	0.300	**0.297**	**0.294**	0.303
	RMSE	0.452	**0.414**	0.465	**0.413**	0.472	**0.400**
Relative	MPE	-0.108	**-0.082**	-0.332	**-0.215**	**-0.383**	-0.090
	MAPE	4.625	**4.457**	4.719	**4.454**	4.828	**4.367**
	MdAPE	3.743	**3.661**	3.776	**3.638**	**3.630**	3.710
	RMSPE	5.883	**5.662**	6.301	**5.676**	6.384	**5.506**
	CE	0.945	**0.954**	**0.941**	0.954	0.940	**0.957**
	PI	0.717	**0.762**	**0.699**	0.763	0.691	**0.778**
	PI.MAE	0.454	**0.480**	**0.449**	0.481	0.437	**0.491**
	PI.MdAE	0.424	**0.427**	**0.433**	0.439	**0.445**	0.429

Figure 5.12 shows relatively similar predictions across all the different experiments. Thus there is little to differentiate the hydrograph predictions by either training algorithm or input data for the 2005 flood event.

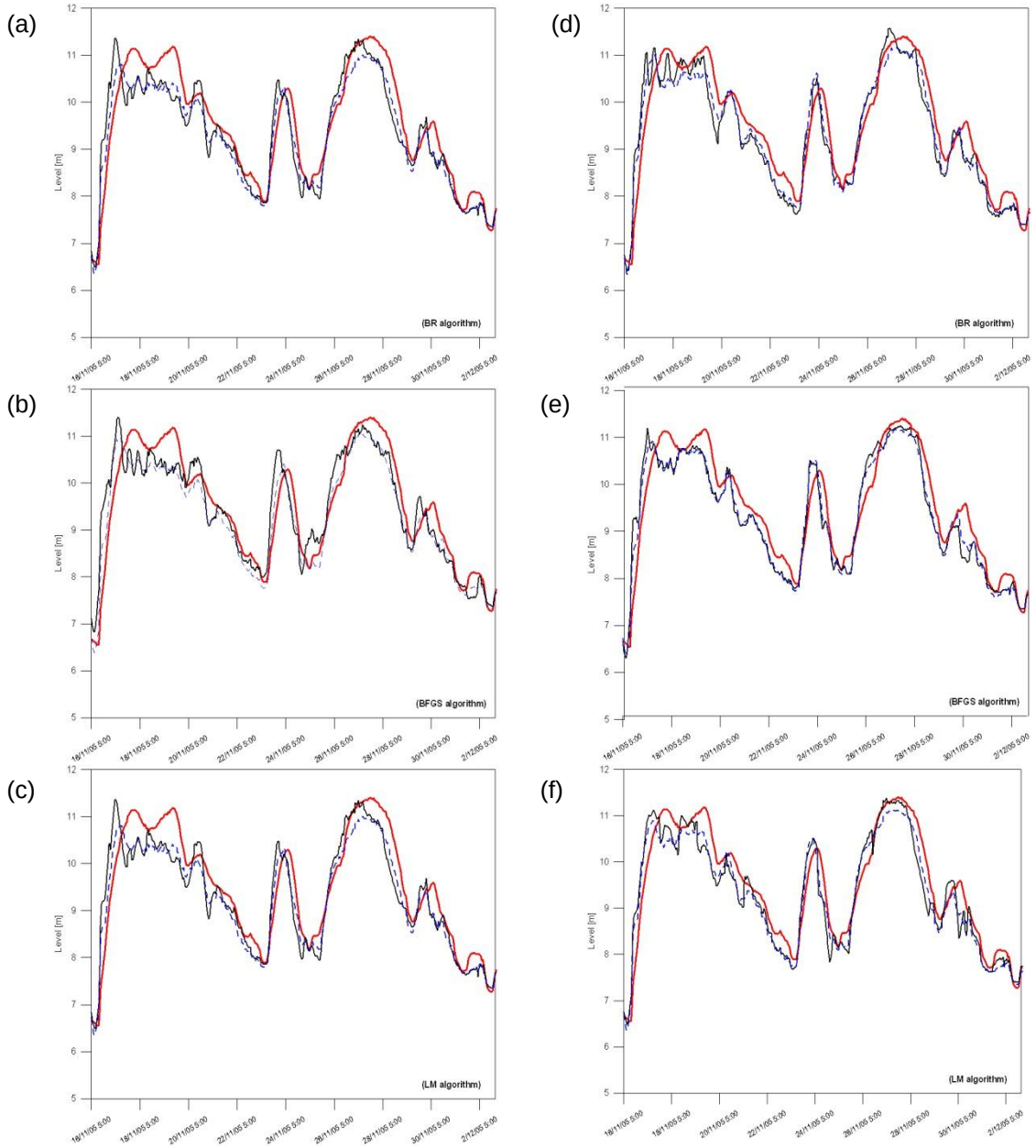


Figure 5.12: Results for a) Expts #41 and #42; b) Expts #43 and #44; c) Expts #45 and 46; d) Expts #47 and #48; e) Expts #49 and #50; and f) Expts #51 and #52 for the 2008 flood event where the red line is the observed, the black solid line is the weighted average AIC and the dotted blue line is the weighted average modified AIC.

The same performance measures and flood hydrographs are available in Table 5.22, Table 5.23 and Figure 5.13 for prediction of the 2008 flood event. This time the performance measures indicate the opposite, i.e. the original AIC generally outperforms the modified AIC across all the experiments. This is also very apparent from Figure 5.13 where the peak prediction is much worse for the modified AIC.

Table 5.22: Performance measures for Expts #53 to 58 for the 2008 flood event for inputs of Orte and rainfall only. Grey shading denotes the best performing model overall while bold denotes the best performance between pairs of experiments, i.e. between Expts #53 and #54, Expts #55 and #56 and between Expts #57 and #58.

Type of Measure	Performance Measure	BR		BFGS		LM	
		Expt #53 AIC	Expt #54 Modified AIC	Expt #55 AIC	Expt #56 Modified AIC	Expt #57 AIC	Exp t#58 Modified AIC
Absolute	ME	**0.046**	0.346	**0.139**	0.367	**-0.044**	0.330
	MAE	**0.305**	0.482	**0.315**	0.484	**0.375**	0.451
	MdAE	**0.182**	0.284	**0.181**	0.261	**0.213**	0.264
	RMSE	**0.421**	0.666	**0.476**	0.692	**0.519**	0.628
Relative	MPE	**0.157**	3.090	**1.253**	3.300	**-0.711**	2.958
	MAPE	**3.567**	5.153	**3.509**	5.106	**4.330**	4.811
	MdAPE	**2.402**	4.084	**2.304**	3.649	**2.625**	3.690
	RMSPE	**4.774**	6.456	**4.842**	6.562	**5.892**	6.058
	CE	**0.957**	0.893	**0.945**	0.884	**0.935**	0.905
	PI	**0.857**	0.642	**0.817**	0.614	**0.783**	0.682
	PI.MAE	**0.619**	0.400	**0.608**	0.397	**0.533**	0.438
	PI.MdAE	**0.651**	0.455	**0.653**	0.501	**0.591**	0.494

Table 5.23: Performance measures for Expts #59 to 64 for the 2008 flood event for inputs Orte+rainfall 2008+2005 in the calibration. Grey shading denotes the best performing model overall while bold denotes the best performance between pairs of experiments, i.e. between Expts #59 and #60, Expts #61 and #62 and between Expts #63 and #64.

Type of Measure	Performance Measure	BR		BFGS		LM	
		Expt #59 AIC	Expt #60 Modified AIC	Expt #61 AIC	Expt #62 Modified AIC	Expt #63 AIC	Expt #64 Modified AIC
Absolute	ME	**0.130**	0.465	**0.064**	0.521	**0.041**	0.274
	MAE	**0.337**	0.588	**0.341**	0.616	**0.334**	0.515
	MdAE	**0.229**	0.249	**0.185**	0.250	**0.197**	0.239
	RMSE	**0.461**	0.900	**0.493**	0.950	**0.465**	0.815
Relative	MPE	**1.118**	4.090	**0.303**	4.684	**0.171**	2.477
	MAPE	**3.904**	5.989	**3.824**	6.203	**3.797**	5.315
	MdAPE	**2.848**	3.714	**2.509**	3.406	**2.489**	3.416
	RMSPE	**5.042**	8.198	**5.114**	8.586	**4.978**	7.489
	CE	**0.949**	0.805	**0.941**	0.782	**0.948**	0.840
	PI	**0.829**	0.347	**0.804**	0.272	**0.825**	0.464
	PI.MAE	**0.581**	0.268	**0.576**	0.232	**0.584**	0.358
	PI.MdAE	**0.562**	0.523	**0.645**	0.520	**0.623**	0.542

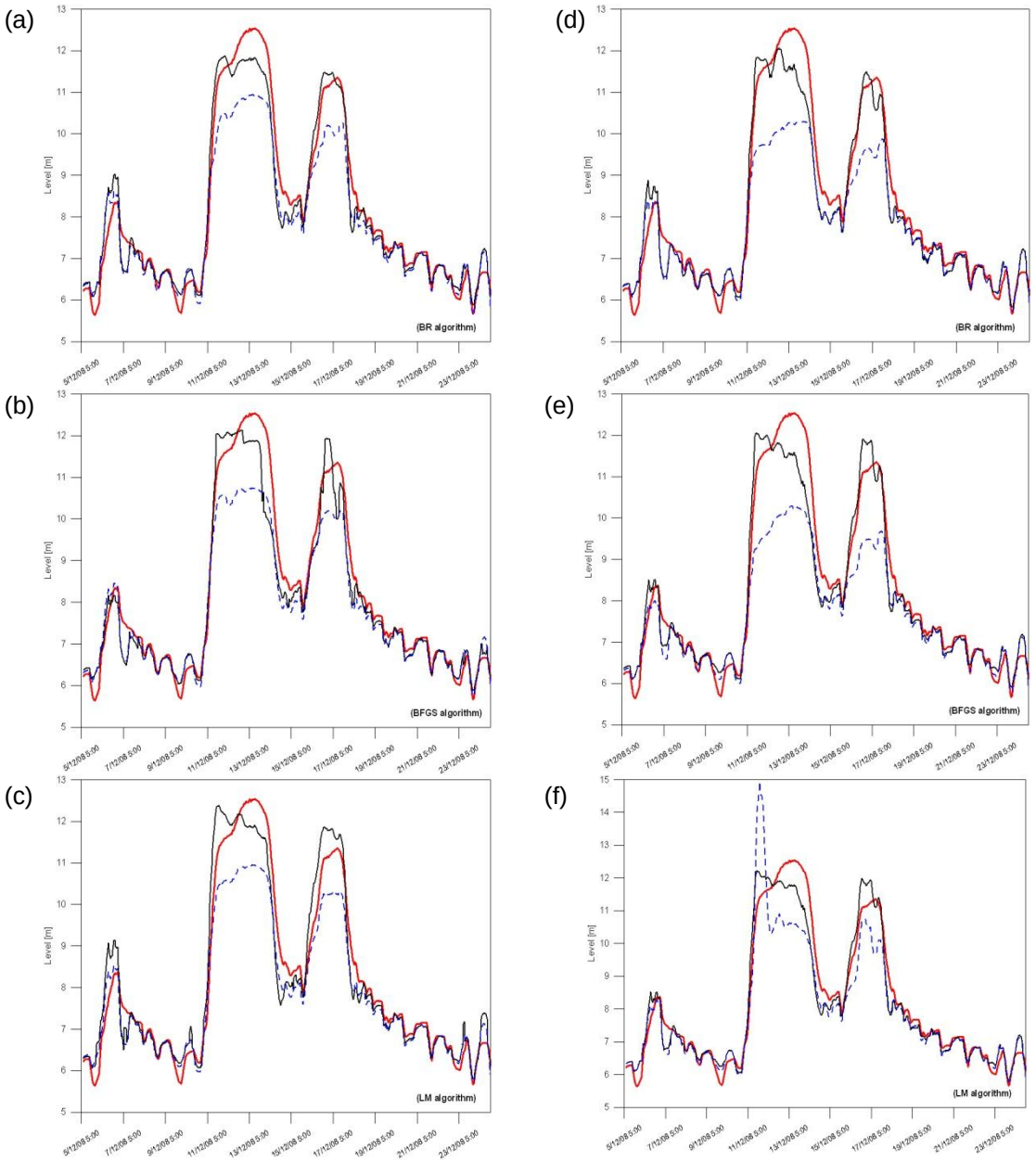


Figure 5.13: Results for a) Expts #53 and #54; b) Expts #55 and #56; c) Expts #57 and 58; d) Expts #59 and #60; e) Expts #61 and #62; and f) Expts #63 and #64 for the 2008 flood event where the red line is the observed, the black solid line is the weighted average AIC and the dotted blue line is the weighted average modified AIC.

Comparing the results to Table 5.15 (i.e. pure averaging of the ensemble), the results for the AIC are better overall. However, comparing the results with the PI threshold approach (Table 5.18), the results are more mixed. In the experiments where the 2005 and 2008 floods were not used in the training data, some of the absolute and relative measures indicate better performance for the PI while others indicate better performance for the AI, indicating some contradictions between these redundant measures. The benchmark based measures, however, indicate better performance when using PI to select the model runs. For the experiments where the 2005 and 2008

flood events were included in the training dataset, the PI approach is generally better than the AIC.

5.10 Calculation of the Confidence Limits of the Predictions

Although a number of different ANN have been evaluated as part of this research, no confidence limits were calculated other than by Savi for the conceptual TEVERE model in Chapter 4. In order to examine the uncertainty of the model predictions, the statistical behaviour of the forecast error of Experment #17 to Experiment #28 has been analysed and modelled. First the Normal Quantile Transformation (NQT) was applied to the model forecasts and errors (i.e. observations minus the forecast) to check the applicability of the NQT-based methodology suggested by Montanari and Brath (2004) and further developed by Montanari and Grossi (2008). To illustrate a valid example for all the distributions analysed, Figure 5.14 shows the NQT of Experiment #18, where it is evident that the NQT effectively transforms the marginal distributions into Gaussian distributions, but the joint distribution is not elliptic. This result is expected as the marginal transformations do not transform the structure of dependence of the data. Since the joint density of the NQT variables is far from the elliptical shape (evident from the contour plot in Figure 5.14), the application of linear regression related to the Gaussian framework is deemed unsuitable for these data, as results returned could be biased.

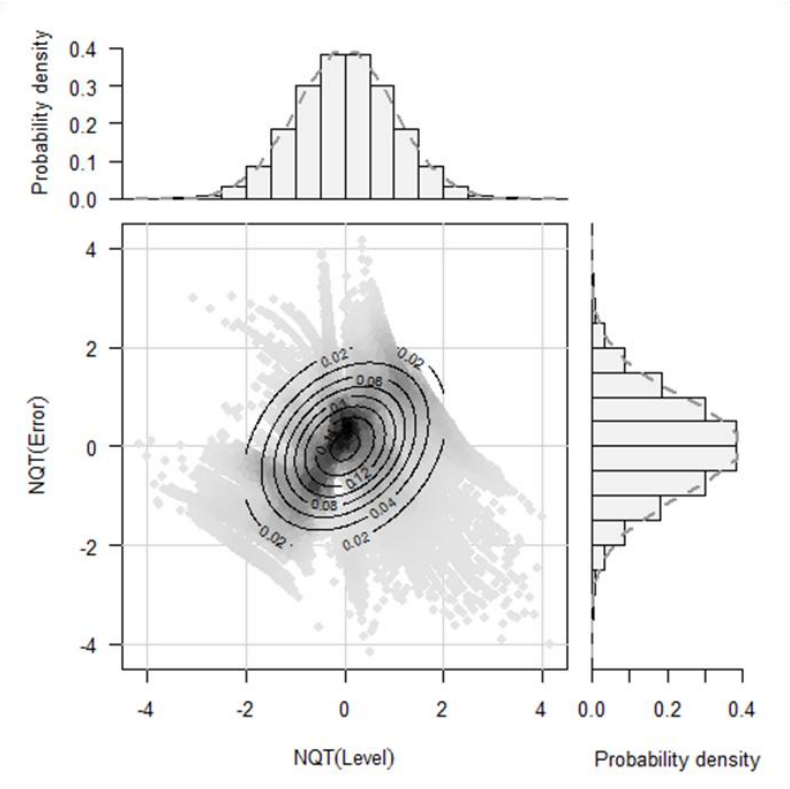


Figure 5.14 NQT applied to Expt #18. The figure shows the empirical joint histogram and the isolines of the joint Gaussian density function with a correlation parameter equal to the empirical correlation computed on the data. At the top and right are the histograms of the NQT variables and the corresponding standard Gaussian density functions (dashed lines).

Alternative methods can be applied instead of linear regression in the NQT transformed space. A simple alternative is Quantile Regression (QR; Koenker and Bassett, 1978), which is a flexible technique to define the conditional quantiles without making any assumptions about the marginal and joint distributions of the data. The QR algorithm can identify the band containing a certain percentage of the corresponding error for each value of the water level forecast. The data do not require transformation and the approach is non-parametric. The QR was employed to determine the 90% confidence band around the 'best' forecast, where best is defined through the AIC weighting technique described in section 5.9.

Figure 5.15 contains the scatterplot of the errors versus the model forecast for Experiment #18. The blue lines are the 5% and 95% conditional percentiles for each value of the forecast. It shows that about 90% of the error values fall within this band. The shape is a result of a spline with 25 degrees of freedom. Even though more refined algorithms are available choosing the degrees of freedom, a simple trade-off between parsimony and accuracy was adopted. As QR is a data-driven approach, it is flexible but there is loss in terms of extrapolation ability. Figure 5.15 shows that the upper limit follows the diagonal pattern that characterises the errors for high forecast values. The typical diagonal stripes shown by the data are an artefact resulting from the error definition (observation minus forecast). The spread of values in Figure 5.15 does not describe a genuine joint distribution of two quantities stochastically correlated, but it is influenced by the functional relationship between the forecasts and the errors.

This aforementioned behaviour results in an upper bound to the upper confidence limit of the forecast, which is clearly shown in Figure 5.16, where the observed are plotted against the forecasted levels. From Figure 5.16 it is evident that the forecast is upper bounded by the maximum observed level. The extrapolation of the upper bound requires a parametric model, and could be achieved by replacing the splines with parametric models such as low-order polynomials, which introduce minimal statistical assumptions, or by the NQT approach, introducing the strong assumption of a bivariate Gaussian distribution of the data. In this context, the QR approach based on a B-spline was applied to keep the procedure data-driven and coherent with an ANN approach, which avoids statistical hypotheses), and to account for the fact that the dependence structure of the clouds in Figures 5.15.and 5.16 is influenced by the artefact related to the error definition. The results are summarised in Figures 5.17 and 5.18.

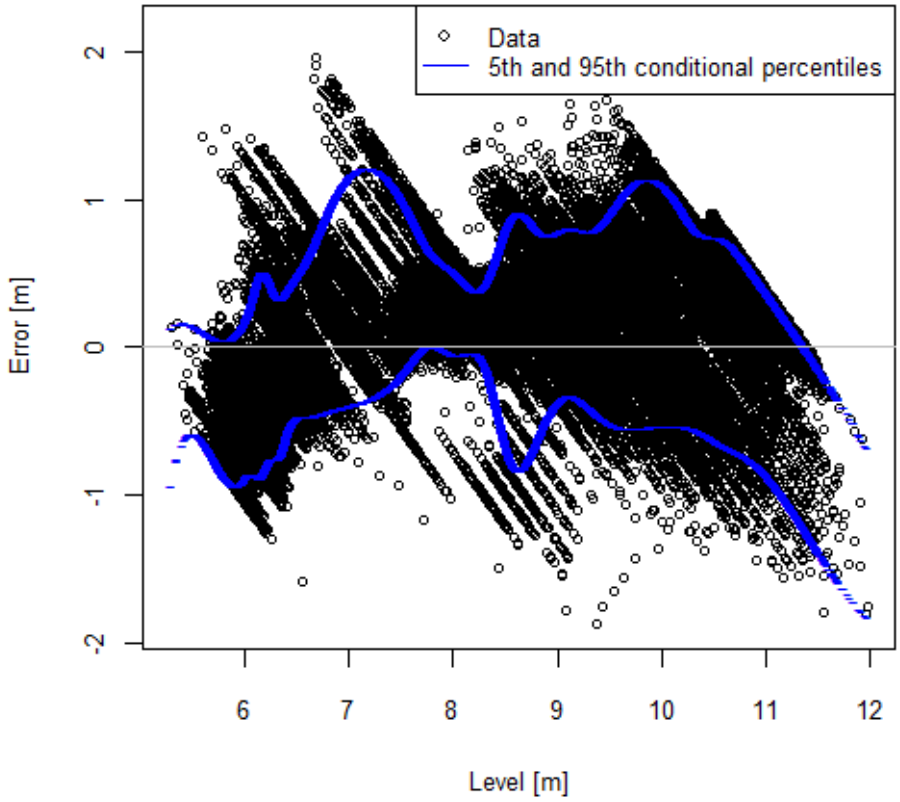


Figure 5.15: Quantile Regression with a spline of 25 degrees of freedom applied to Expt #18 for the error plotted againt the forecasted water levels

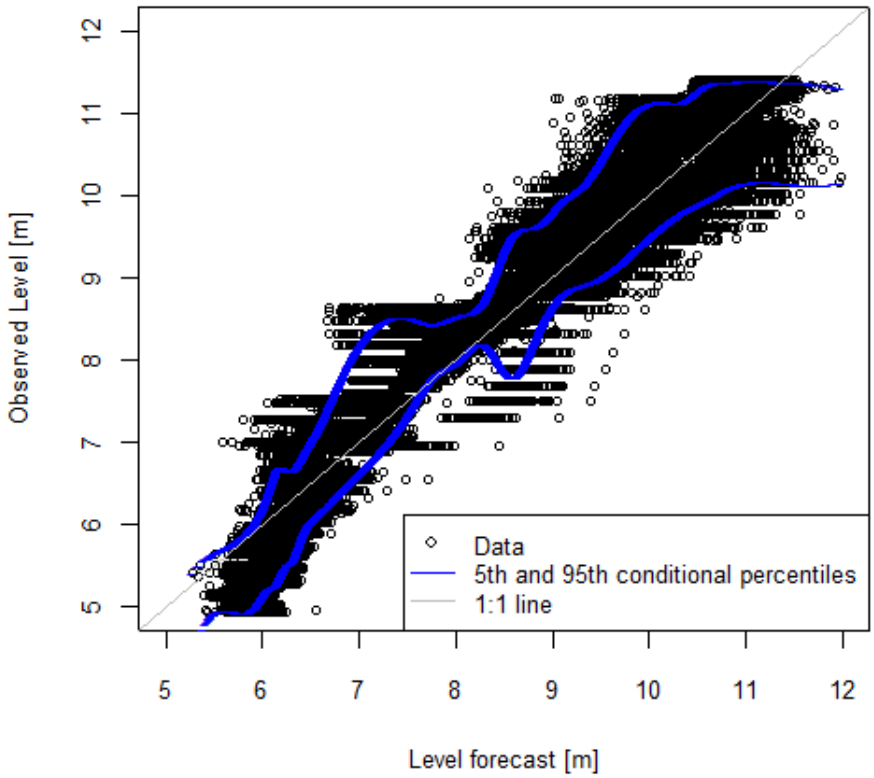


Figure 5.16: Quantile Regression with a spline of 25 degrees of freedom for Expt #18 for the observed versus forecast water levels

The confidence bands are plotted in Figures 5.17 for the 2005 flood event and Figures 5.18 for the 2008 flood event where the individual graphs correspond to Experiments #29 to #40. Figures 5.17 (a-f) show that reasonably similar bands are associated with the different algorithms and the experiments, where the 2005 and 2008 flood events have been added. Looking at Figure 5.18, where the 2008 event is presented, the behaviour of the confidence band is similar to the 2005 event. For middle level values. the bands describe the non-parametric patterns provided by QR. For high level values, especially for the 2008 event, the upper bound is bounded and constant. This is expected in light of the behaviour described in Figures 5.15 and 5.16. For high level forecasts, there is a range of values where the upper limit is constant and equal to the observed maximum and the error varies linearly with the level. For the 2008 event, the upper bound corresponds to a range of forecast values larger than that of 2005 (where the figures are not shown here). Thus the lack of extrapolation that characterises QR is more evident.

Finally, it is worth pointing out that the upper bound of the confidence bands should not interpreted as a lack of uncertainty. On the contrary, they reflect the lack of information (i.e. the complete ignorance and uncertainty) concerning the phenomenon under study when non-parametric techniques are used. While the parametric methods try to advance hypotheses on the behaviour of the process beyond the observed range, the nonparametric results reflect the attitude described by the Greek term Epoché (ἐποχή, epokhē := "suspension"), which describes the theoretical moment where all judgments about the existence of the external world, and consequently all actions in the world, are suspended. Thus, the upper bounds represent the place where the non-parametric methods suspend their explanatory activity and communicate their inability to provide any piece of information about the unknown (unobserved) phenomenology of the studied process.

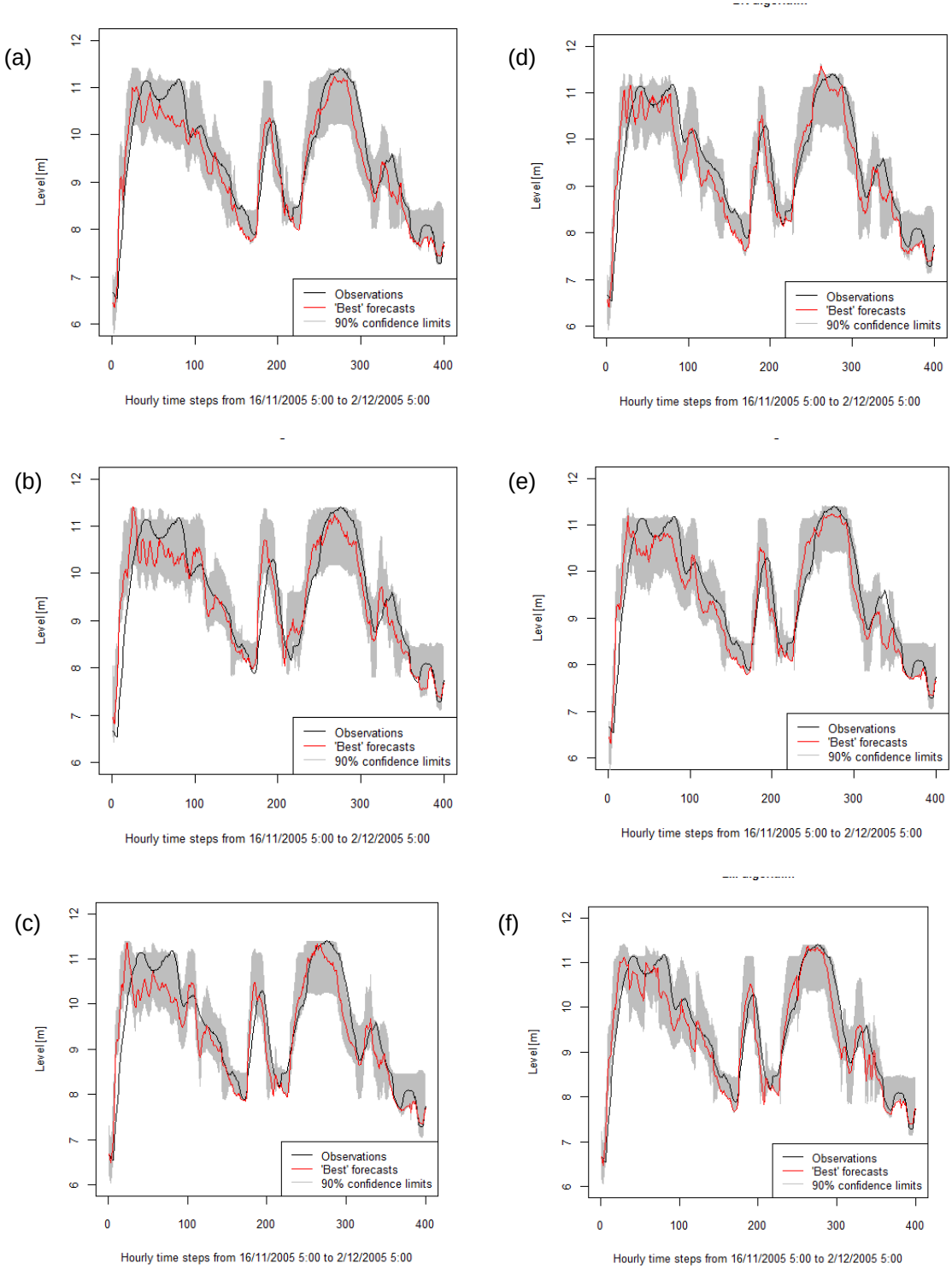


Figure 5.17: Confidence intervals (90%) shown in gray for a) to c) Expts # 29 to #31 and d) to f) Expts #32 to #34 for the 2005 flood event where the red line is the observed and the black line is the average of the simulations determined by the PI threshold method.

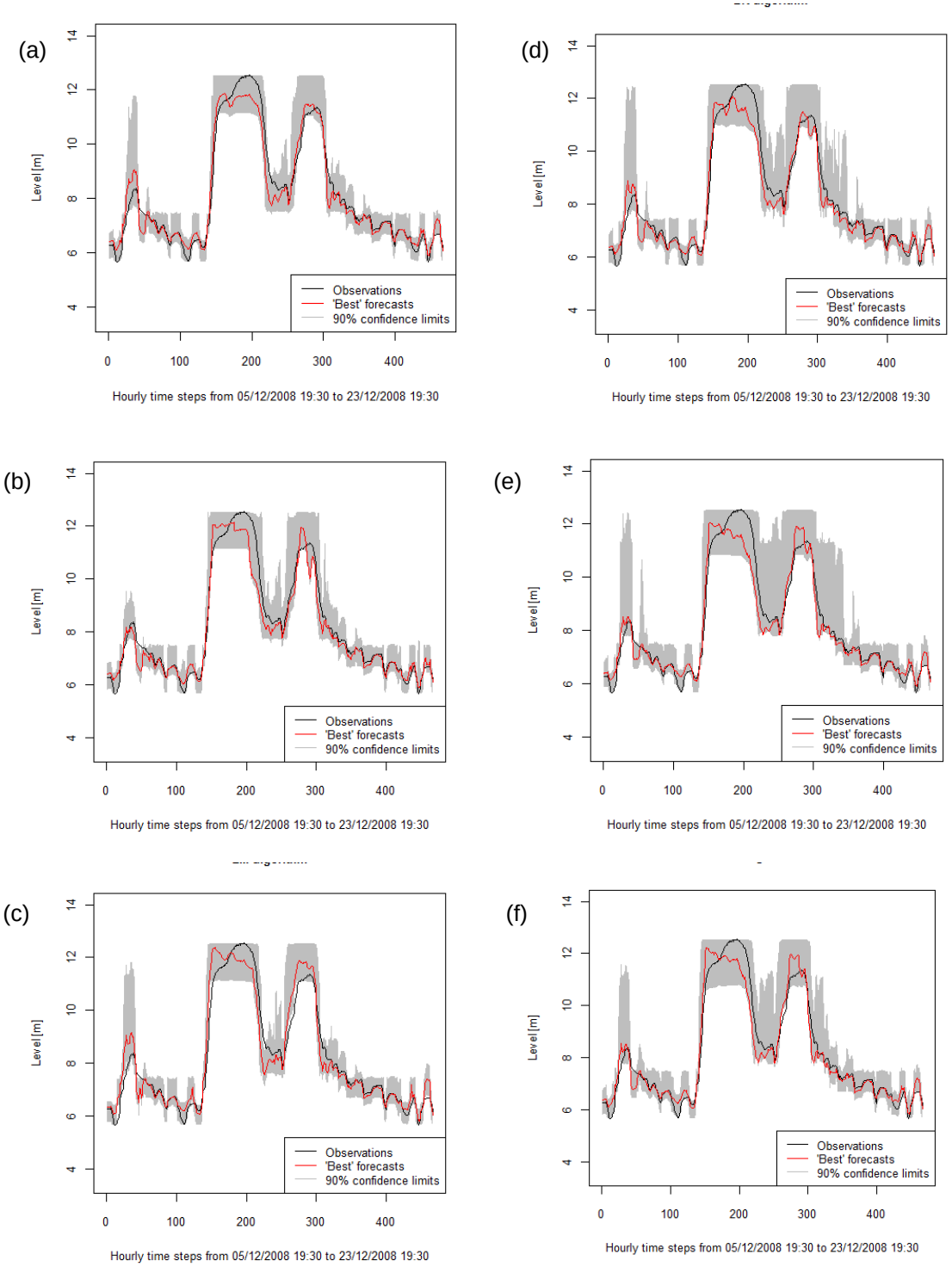


Figure 5.18: Confidence intervals (90%) shown in gray for a) to c) Expts # 35 to #37 and d) to f) Expts #38 to #40 for the 2008 flood event where the red line is the observed and the black line is the simulations determined by the AIC.

5.11 Discussion

A number of different experiments were undertaken with the overall aim of improving the ANN model developed in Chapter 4. The first set of experiments were focussed on model predictions for the 2008 event, as this event was much harder to predict than the 2005 event, and the peak is outside the range of data encountered in the training data set. Therefore, part of the experiments were designed to see whether the ANN could

extrapolate to a more extreme event. The first set of experiments involved adding more upstream stations and rainfall. Both clearly improved the model performance in terms of evaluation measures and a visual inspection of the hydrograph. The addition of rainfall resulted in a prediction that oscillated much more around the hydrograph. Moreover, an examination of the 50 ensemble members showed a greater spread of predictions indicating that rainfall increased the impact of the random initialisation of the weights of the NN. The effective rainfall created even a greater oscillation but looking at the hydrograph, the model overpredicted the peak (which is better from an operational point of view) compared to the hourly rainfall, where evidence of underprediction can be seen. Thus it is clear from these experiments that the NN can extrapolate and predict events outside the range previously seen in training. This relates directly back to the literature review (Chapter 2) in which extrapolation was raised by the ASCE (2000b) review and the subject of a few research studies discussed in the theme 'Other Research'. These sets of experiments also highlighted how much variation there is between ensemble members, especially at the peaks. These observations led to further investigation of the impact of the random initialisation of the weights in Chapter 6.

Other experiments considered the development of a more parsimonious model; accounting for nonstationarity by adding a difference term to the inputs; and adding more data from the time series including cumulative rainfall. These were attempts to simplify the ANN and consider the effects of other types of inputs and more data. It showed that it is possible to create a more parsimonious model but there are tradeoffs in performance. The addition of cumulative rainfall did reduce the oscillations observed with hourly and effective rainfall. However, the best results were still obtained by adding rainfall and more upstream stations to the model.

The next set of experiments worked with the parsimonious model and considered the 2005 and 2008 flood events. The idea was to examine the impact of methods of normalisation and training algorithms on the model results. With such little guidance available on model development from the literature on ANN rainfall-runoff modelling (Maier and Dandy, 2000; Dawson and Wilby, 2001), these experiments were attempts to look for generalisable patterns in normalisation and training algorithms. For the 2005 flood event, normalisation using the mean and standard deviation outperformed normalisation using the minimum and maximum. However, the difference between the two methods in terms of performance measures and the hydrographs was very small. The opposite was found for the 2008 flood event. Thus the latter method appeared to

work better when predicting more extreme events than those contained in the training dataset. For the training algorithms, there was no discernible pattern.

The final sets of experiments considered the method used to combine the ensemble. Averaging was used in all previous experiments. Three other methods were examined: use of a PI threshold to select the best models before averaging; AIC; and a modified AIC after Zhao et al. (2008) whereby the models were weighted and linearly combined. For the 2005 flood event, the average and PI approaches were similar, mostly because the number of ensemble members chosen using this method was the majority of members. The AIC showed an improvement over the average and PI approaches but only on some of the performance measures. For the 2008 event, the PI showed an improvement over the average. This more complex event outside the range of the training data clearly benefited from a more refined method of ensemble combination. The AIC also resulted in improvements compared to a pure average, but the results were more mixed when compared to the PI. Thus, in general, the average works very well. Using the PI or AIC will, however, improve the results. The modified AIC performed better than the AIC for the 2005 flood event but not for the 2008 event. Thus the results do not entirely support the findings of Zhao et al. (2008). Confidence limits were then calculated around the model predictions for Experiments #29 to #40, illustrating some of the problems that surround the calculation of uncertainty for non-parametric methods such as ANNs.

5.12 Summary

This chapter considered ways in which the ANN model developed in Chapter 4 could be improved. The experiments undertaken in the chapter showed that the addition of upstream stations and rainfall did result in an ANN model that could better predict the peak of the 2008 event, indicating an ability to extrapolate. This provides further evidence to existing studies dealing with this issue as reviewed in Chapter 2.

An interesting observation made during these experiments is the large degree to which the ensemble members differ in their predictions, particularly at the peak of flood events. The averaging of the members results in good predictions but the variation between ensemble members are clearly the result of a random initialisation of the weights of the ANN prior to training. Thus in the next chapter, this issue is investigated in more detail.

Chapter 6

Investigation of ANN Initialisation and Ensemble Methods using Empirical Mode Decomposition (EMD)

6.1 Introduction

It was clearly shown in Chapter 5 that 100 random initialisations of an ANN when using Bayesian Regularisation can produce quite different results, particularly in forecasting flood events. This chapter further explores the impact of random weight initialisation on 1-day ahead hindcasts for two rivers in the USA. At the same time, a pre-processing technique called Empirical Mode Decomposition (EMD; Huang et al., 1998) is used with the ensemble ANN approach. EMD decomposes a time series into its intrinsic components, which are generally more regular than the original time series. Each component is then modelled individually and finally combined to reconstruct the series in the original space. Other decomposition methods have been used, e.g. Wang et al. (2006) used classical decomposition and hybrid ANNs to model daily discharge time series while Adamowski and Sun (2010) applied wavelet decomposition as a pre-processing operation and then used ANNs to carry out an ensemble forecast. Wavelet analysis was discussed as part of theme four in the ANN rainfall-runoff modelling literature (section 2.4.5). EMD has not been used before in ANN rainfall runoff-modelling although it has been used to decompose times series of crude oil spot prices (Yu et al., 2008). The authors then modelled the resulting components using ARIMA models and ANNs, which produced a better result than modelling the original time series using either of the two methods alone. The data driven nature of EMD is compatible with the nonparametric nature of ANNs. Hence their combination provides a very flexible modelling strategy.

In this chapter, different catchments were chosen rather than the Tiber River to see whether similar issues arise with random weight initialisation. Moreover, long, time series were available for experimentation in these different catchments, which was required for application of the EMD. The concepts underlying EMD are first introduced in Section 6.2 while section 6.3 describes the catchments and datasets used in this modelling exercise. Section 6.4 outlines the experiments undertaken while the results are presented in sections 6.5 and 6.6 for the Potomac and Clark Fork Rivers, respectively. A discussion of the main findings is provided in section 6.7 and is followed by a summary in section 6.8.

6.2 Empirical Mode Decomposition (EMD)

EMD is an adaptive method for signal analysis introduced by Huang et al. (1998) and is

designed specifically for application to nonlinear and nonstationary data. Combined with Hilbert spectral analyisis (HSA), EMD produces the so-called Hilbert-Huang transform (HHT), which represents a new *a posteriori* and data-driven paradigm for analysing data (Huang and Wu, 2008). HHT acts similarly to wavelets but the basis function is not fixed *a priori*, it accounts for nonlinearity, and its theoretical basis is not mathematical, but empirical.

The EMD algorithm decomposes a time series $x(t)$ into a set of band limited and orthogonal functions $c_i(t), i = 1, \ldots, L$, called intrinsic mode functions (IMFs) and a remaining part $r(t)$ called the 'residue', which represents a monotonic pattern, and can be considered as the overall trend of the original series. The summation of the IMFs and residue returns the original series, with unavoidable but generally negligible numerical errors:

$$x(t) = \sum c_i\, t + r(t) \tag{6.1}$$

An important feature of the IMFs is that their Hilbert transform is consistent with physically meaningful definitions of instantaneous frequency and amplitude, providing a more physically meaningful time-frequency-energy description of a time series (Huang and Wu, 2008). This property (denoted as adaptivity) plays a fundamental role in the analysis of nonlinear and nonstationary data, as only the adaptation to the local variations of the data can fully account for the physics of the underlying processes. As discussed by Huang et al. (1998), the use of a predetermined basis to fit all the phenomena causes a 'harmonic distortion' in the Fourier analysis of nonlinear processes. An easy way to generate the necessary adaptive basis is to derive the basis from the data. Moreover, as a nonlinear system does not admit an explanation by superposition, and any linear expansion for a nonlinear system, such as that shown in Equation 6.1, does not make physical sense, it is worth pointing out that the aim of the HHT is not to provide a physically meaningful linear expansion but individual components in the linear system, which can have physical meaning related to the full nonlinear system. In this sense, HTT is able to capture important features of nonstationary and nonlinear signals such as Lorenz, Rössler and other nonlinear chaotic systems (Huang et al., 1998; Kijewski-Correa and Kareem, 2007; Lee and Ouarda, 2011b).

Each IMF is a time series that satisfies the following two characteristics: (1) the number of extrema (i.e. the number of maxima and minima) and the number of zeros crossing differs at most by one; and (2) at any point, the mean values of the envelopes defined by the smooth curves passing through all the local maxima and all the local minima,

respectively, is zero. The IMFs are defined by an iterative process called ‘sifting’, which is the core of the EMD algorithm. It serves to eliminate riding waves, and to make the wave-profiles more symmetric. The sifting procedure can be summarised as follows (Huang et al., 1998; Flandrin et al., 2004; Yang et al., 2010):

1. Identify all extrema of the signal $x(t)$ (i.e., local minima and maxima);
2. Interpolate between the minima (and the maxima) by means of cubic splines or more refined methods (Pegram et al., 2008), producing a lower and upper envelope, e_{min} and e_{max}, respectively;
3. Compute the average of the envelopes $m = [e_{max} + e_{min}]/2$;
4. Extract the detail $h_1(t) = x(t) - m(t)$; if $m(t)$ is equal to zero or smaller then a fixed threshold, h_1 is retained as the first IMF and labeled as c_1; otherwise, steps 1-3 are repeated treating $h_1(t)$ as data, and so forth until $h_1(t)$ fulfills the properties of an IMF (number of zero crossings, and zero mean);
5. Take the difference $r_1(t) = x(t) - c_1(t)$: the procedure finishes if the number of extrema of $r_1(t)$ is not larger than two; otherwise, treat $r_1(t)$ as the new $x(t)$ and repeat steps 1-4 to define the next IMF.

Since the procedure is data-driven and relatively new, there are some drawbacks to the method. For example, the use of a spline to determine the envelope of extrema is not the only method available (Pegram et al., 2008) and can be affected by serious problems of fitting near the ends of the signal. The stopping criterion in step 4 is also somewhat subjective. However, Huang and Wu (2008) have described these and other issues, which they have tackled and partially solved. Therefore, despite these limitations, EMD and HHT have been applied successfully in several fields (Huang and Shen, 2005; Huang and Attoh-Okine, 2005; Huang and Wu, 2008) as they provide a better representation of the local behaviour of the data compared to the Fourier and wavelet transforms. Moreover, it has been shown that IMFs can have physical meaning (e.g., Coughlin and Tung, 2004; Zhen-Shan and Xian, 2007).

Despite the growing interest in HHT and EMD in economics (Huang et al., 2003; Zhang et al., 2008, 2009) and geophysical and climatic studies (Franzke, 2009; Crockett and Gillmore, 2010; Fauchereau et al., 2008; Jackson and Mound, 2010; Kataoka et al., 2009; Solé et al., 2009), there are only a few hydrological applications to date. Sinclair and Pegram (2005) applied a two-dimensional EMD to separate high and low frequency components of radar measured rainfall fields in order to assess the temporal persistence of the low frequency components. Pegram et al. (2008) suggested using cubic splines instead of rational splines as one way of improving the EMD procedure

described earlier, where the potential benefits were demonstrated on three series of annual rainfall totals from three different locations around the world. Huang et al. (2009) studied daily time series of flow from two French rivers (Seine and Wimereux) and showed that both rivers are likely to be influenced by the same maritime climate regime of Northern France through the analysis of the correlation among the large scale IMF modes. Franceschini and Tsai (2010) applied the HHT method to analyze four time series in the Niagara River: flow, water temperature, and incoming concentrations of two polycyclic aromatic hydrocarbons (fluoranthene and chrysene). Lee and Ouarda (2010) used EMD to extract nonstationary oscillations of scaled precipitation and the North Atlantic Oscillation index. In order to perform long-term forecasts, Lee and Ouarda (2010) modelled the most important components with a nonstationary oscillation resampling technique, the sum of unselected components by k-nearest neighbour resampling or autoregressive models, and the EMD residual by trend fitting techniques. Lee and Ouarda (2011b) applied the same modelling strategy to forecast global surface temperature anomalies, which were also analysed by Lee and Ouarda (2011a) to separate the driving climatic signals from noise.

Such an approach represents a generalisation of the “divide and conquer” approach used by Yu et al. (2008) and Yang et al. (2010) to forecast crude oil price time series and climatic time series, respectively. In particular, Yu et al. (2008) modeled each IMF with ANNs and ARIMA models, and then compared the ensemble and non-ensemble approaches, concluding that the EMD-ANN was more accurate than the other competitors in terms of the root mean squared error (RMSE) of 1-day ahead forecasts. Yang et al. (2010) modeled each component with ANNs and reached similar conclusions for lead times of 1 to 6 steps in the future, using Pearson’s product moment correlation coefficient as a measure of performance. A hybrid EMD-ANN approach was also used by Hui and Xinxia (2010) to model a 40-year monthly runoff sequence from a hydrological station in Handan City (China), but no details were provided of the ANN structure and no statistical tests or indices were used to assess the model performance.

6.3 Catchments and Data Availability

Datasets from two different catchments were used in this study. The first dataset consists of 115 years of mean daily stream flow (in m^3/s) from the Potomac River at Point of Rocks, Maryland, USA (US Geological Survey 303 station ID 01638500), with a drainage area of 24,996 km^2. This dataset is one of the longest US discharge time series available. The Jennings Randolph Dam, completed in 1981, controls less than 2% of the catchment (Villarini et al., 2009) and therefore has a relatively small impact

on the low flow of the Potomac River. Low flows were also affected slightly by the Stony River Reservoir from 1913 to July 1981, the by Savage River Reservoir since December 1950, and extensively at times by hydroelectric plants. Moreover, despite near-total deforestation in the late 19th and early 20th centuries related to agricultural practices (e.g., Bonan, 1999), the Potomac River at Point of Rocks provides one of the most natural annual peak records for a river of its size in the USA (Villarini et al., 2009). The annual peak record for the 20th century was studied by Villarini et al. (2009), who did not find evidence of non-stationarity. However, mean daily stream flow series convey more information and are more complex signals than the series of annual peaks.

The second dataset consists of 80 years of mean daily stream flow (in m^3/s) from the Clark Fork River below Missoula, Montana, USA (US Geological Survey station ID 12353000), with a drainage area of 23,317 km^2. Clark Fork is the largest river by volume in Montana. Based on the USGS quality control protocols, the quality of the records is classified as good (the difference between the data and the actual stream flow is within 5%) except for estimated daily discharge, which is fair (the difference between the data and the actual stream flow is within 8%). Diversions for irrigation of about 951 km^2 occur upstream from the station.

Table 6.1 contains summary statistics for the dataset as a whole (1 October 1895 to 30 September 2009 for the Potomac and 1 October 1929 to 30 September 2009 for Clark Fork), for the training datasets (1 October 1895 to 30 September 1979 for the Potomac and 1 October 1929 to 30 September 1989 for Clark River) and for the testing datasets (1 October 1979 to 30 September 2009 for the Potomac and 1 October 1989 to 30 September 2009 for Clark River) separately.

For the Potomac River, the average discharge and the quantiles with probability $P \leq 0.9$ in the training dataset are slightly shifted compared to those of the testing dataset, probably owing to the effects of the Stony River Dam and Jennings Randolph Dam, operating until and since 1981, respectively. For Clark River, the average discharge and the quantiles with probability $0.1 \leq P \leq 0.9$ in the training dataset are slightly higher compared to those of the testing period.

Table 6.1: Summary statistics (in m3/s) for the Potomac and Clark Fork Rivers. The symbol xP, with p={0.1,0.25,0.5,0.75,0.9}, denotes the quantile with nonexceedance probability P.

Statistic	Potomac River			Clark Fork River		
	Full set	Training	Testing	Full set	Training	Testing
Min	15.3	15.3	20.4	16.4	16.4	19.1
$x_{0.1}$	48.1	46.1	51.2	45.8	46.1	45.6
$x_{0.25}$	73.9	73.9	79.5	58.3	59.4	56.6
$x_{0.50}$	152.8	152.3	161.9	78.1	79.8	73.3
$x_{0.75}$	311.3	302.8	322.6	147.7	148.8	144.6
$x_{0.90}$	585.8	574.5	622.6	373.6	384.9	339.6
Max	12282.2	12282.2	8150.4	1531.0	1471.6	1531.0
Mean	269.2	264.9	281.0	150.2	153.5	140.3
Std Dev	384.8	386.1	381.0	177.2	182.3	160.6

6.4 Experimental Set Up

The Potomac River data cover the period 1 Oct 1895 to 30 Sep 2009. Daily data from 1 Oct 1895 to 30 Sep 1979 (85 years, ≈ 74% of the records) were used for training, while the remaining 30 years were used for testing (≈ 26% of the records). A three-layer feedforward ANN with a logistic activation function was chosen, where the input layer takes the previous 5 days as inputs, the hidden layer contains 5 nodes, and the output layer predicts the flow 1-day ahead. This architecture was chosen by trial and error and is coherent with those selected by e.g. Cannas et al. (2006), Wang et al. (2006) and Adamowski and Sun (2010) for other daily stream flow series. Other more complex structures were tried but no definitive improvements were found. This was further complicated by which performance index was chosen to determine improvements and the effect of the random initialisation of the weights. As the ratio of the number of input variables and the number of network weights is far larger than 50, ANNs should not be prone to either underfitting or overfitting if their training is not stopped appropriately (Amari et al., 1997; Wang et al., 2006). Therefore, training was stopped when the training error reached a sufficiently small value or when changes in the training error remained small. If these conditions were not fulfilled, the training was stopped after 1000 epochs.

Similar remarks hold for the Clark Fork River data, which span from 1 Oct 1929 to 30 Sep 2009. The first 60 years (75% of the records from 1 Oct 1929 to 30 Sep 1989) were used for training, whereas the remaining period (20 years from 1 Oct 1989 to 30 Sep 2009) was used for testing. For these data, the trials made little difference and therefore to be consistent, the same architecture was used for the Potomac River data to examine and compare the impact of the signal structure discussed in the following sections.

The same strategy was applied to model each EMD component for the EMD-ANN and both datasets; however, for the EMD-ANN, at each forecasting step, it was necessary

to update the previous five days of all IMFs moving from the training set to the validation set. Since the decomposition was originally performed on the training set, the IMFs of the validation set must be computed. A possible strategy is to repeat the decomposition of the training set including the new observations that become progressively available. Such an approach is not efficient for two reasons. First, the EMD algorithm is affected by problems of fitting near the ends of the signal; therefore, the IMF values corresponding to the new observations which are progressively added at the end of the signal are not reliable. Second, the number L of the IMFs is related to the sample size n according to the approximate relation $L + 1 \approx \log_2 n$, as EMD acts as a dyadic filter bank (e.g. Flandrin et al., 2004; Wu and Huang, 2004). For instance, the number of components for the Potomac training set is 14 ($\log_2 30$ = 14.9), where the last component is the residual trend. To apply the 14 fitted ANNs to the testing set, 14 new IMF values are needed for each observation progressively added in the testing period. These 14 values can only be obtained by decomposing a series with size n. Hence, to avoid the above-mentioned problems, EMD was applied to the last 85 years of stream flows, and used the last 30 years of each component as the basis to test the EMD-ANN approach. In this manner, the required number of IMFs is obtained (i.e. 14) for each observation of the testing set and avoids the problem whereby the IMF values are affected by end effects of the EMD method (Yang et al., 2010); nevertheless, the resulting IMF values used to forecast the stream flow of a generic day t in the testing period are computed using information from future stream flow values, which would not be available on day t in a forecasting exercise. The testing stage has therefore been configured as a hindcast experiment rather than a real forecast.

It should be noted that, in the EMD-ANN approach, the hindcasts of each component at each time step are compounded to obtain the hindcasts of the original stream flows by a simple summation, as this procedure is consistent with the underlying idea of EMD. Yu et al. (2008) suggest that the components can be combined using an adaptive linear ANN; however, experimentation showed that this approach does not provide significant improvements and introduces further complexity. Finally, to assess the impact of the ANN parameter uncertainty, the ANN and EMD-ANN were trained 100 times with different initial weights randomly generated by a uniform distribution defined on the range [−0.5, 0.5], which is in agreement with the guidelines suggested by Haykin (1999, pp. 182-184). The EMD analysis and ANN modeling were performed in R (R Development Core Team, 2009) using the freely available packages EMD (Kim and Oh, 2008) and nnet (Venables and Ripley, 2002).

6.5 Potomac River Results

6.5.1 Preliminary Analysis

Figure 1 shows the 14 components given by the EMD for the training dataset along with the ANN values resulting from the training procedure (based on the minimisation of the RMSE), with a fixed set of initial weights. On the right are scatterplots of the values of the EMD components versus the corresponding ANN fitted values. Figure 6.1 shows that the ANNs provide an excellent fit to components 3-14, whereas the errors are more evident for components 1-3, which are characterised by higher mean frequencies.

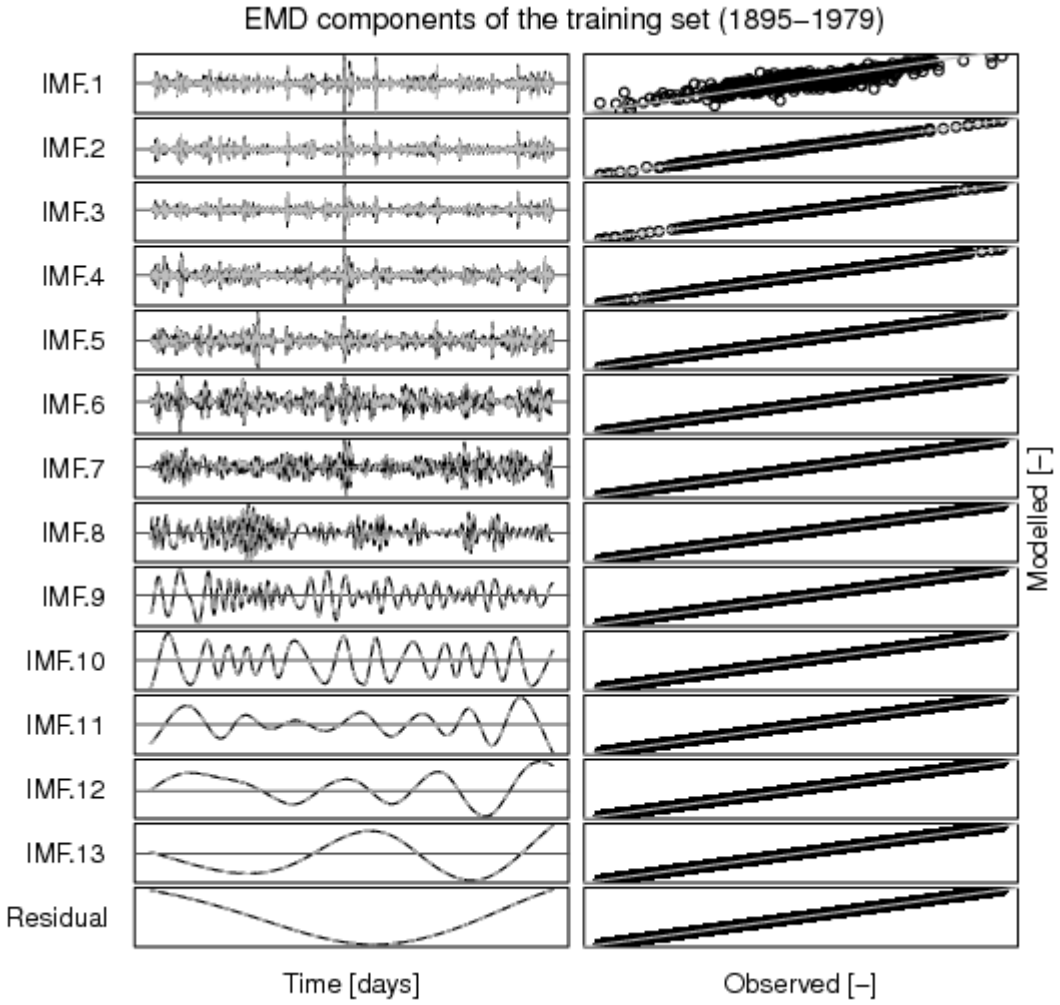


Figure 6.1: On the left are the EMD components extracted from the mean daily discharge time series of the Potomac River spanning from 1895 to 1979 (training set; black lines), along with 1-day ahead hindcasts obtained by the fitted ANNs (dashed gray line). On the right are scatter plots of observations versus ANN modeled hindcasts. The 1:1 gray lines denote a perfect fit. Source: Napolitano et al. (2011).

Figure 6.2 shows the mean period ω of each IMF versus the number of IMFs in a loglinear plane: the points are aligned along a straight line with slope $\gamma \approx 1.98$, fulfilling the relationship $\omega = \gamma^L$. This confirms that the EMD acts as a dyadic filter bank ($\gamma = 2 \approx 1.98$), as for white noise (Wu and Huang, 2004), fractional Gaussian noise (Flandrin et al., 2004) and turbulence time series (Huang et al., 2008). The seventh component appears to represent the annual cycle as the mean period is ≈ 304 days. The plot of the energies and cumulated energies in Figure 6.2 points out that about 80% of the energy (variability) is explained by the first four IMFs, whose mean periods are smaller than 44 days. Moreover, a peak of energy is given by the seventh component. To summarise, the time series is dominated by intra-monthly and annual dynamics, whereas processes at the other scales (between monthly and annual, and interannual) contribute only marginally to the overall energy of the stream flow process.

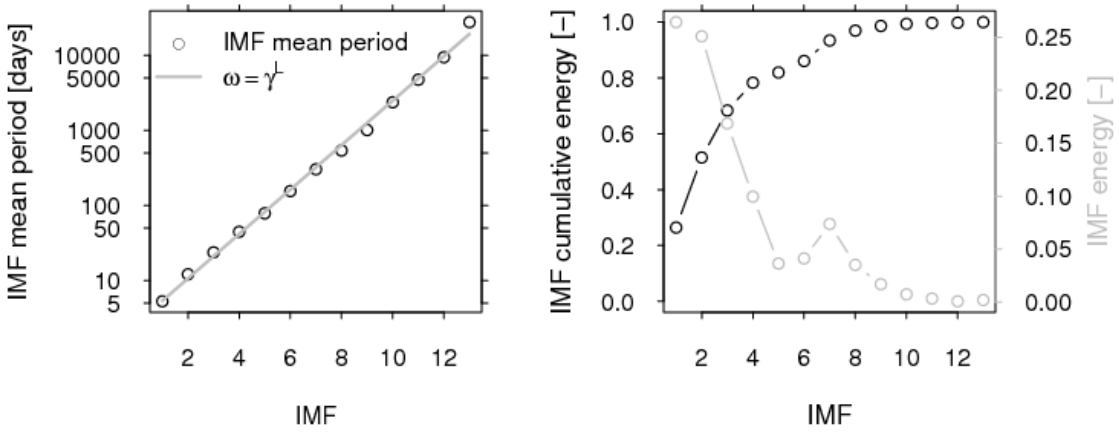


Figure 6.2: On the left is the scaling relationship between the number of IMFs and the corresponding mean periods for the components shown in Figure 6.1. On the right is the non-dimensional energy (gray) and cumulative energy (black) for each IMF shown in Figure 6.1. Source: Napolitano et al. (2011).

Figure 6.3 shows the 1-day ahead hindcasts of each EMD component for the testing set (1980-2009). Similar to the training dataset, the ANNs perform very well for components 3-14, whereas the errors for components 1-3 are comparable to that obtained in the training period.

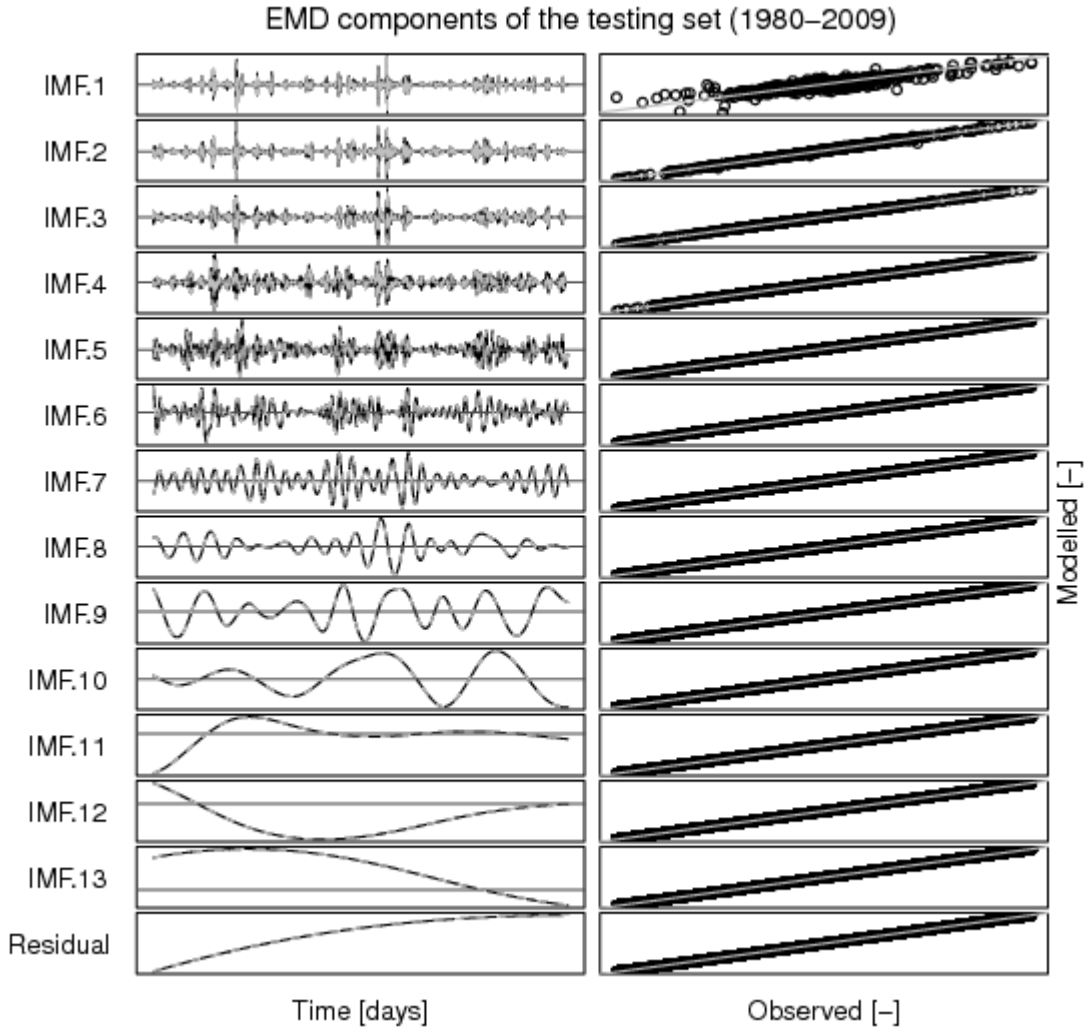


Figure 6.3: Mean daily discharge time series of the Potomac River spanning from 1980 to 2009 for the test data set. Source: Napolitano et al. (2011).

Figures 6.4a-b show the discharge observed during the testing period along with ANN hindcasts and EMD-ANN values obtained by summing up the modeled components shown in Figure 6.3. Figures 6.4c-d show the scatter plots of the final ANN and EMD-ANN hindcasts versus the observed stream flow of the test period for one example of a training run.

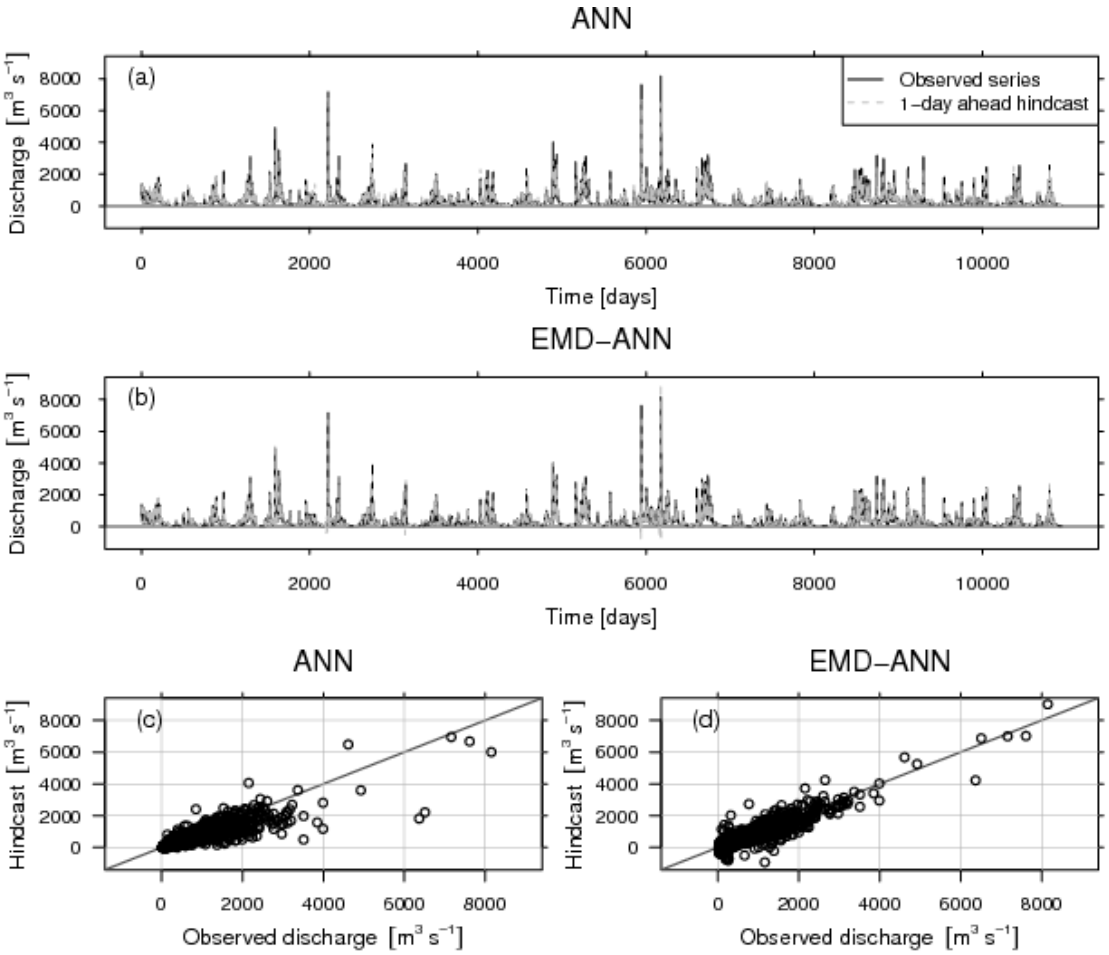


Figure 6.4: The Potomac River discharge time series of the test period along with 1-day ahead hindcasts obtained by (a) a simple ANN and (b) the EMD–ANN model. Scatterplots of the observed discharge vs hindcasts computed by (c) the ANN and (d) the EMD–ANN. Source: Napolitano et al. (2011).

6.5.2 Analysis of the Weight Initialisation Uncertainty

Figures 6.5a-b focus on one example year and show the records and the corresponding 1-day ahead hindcasts given by 100 ANN and EMD-ANN training runs with different randomly generated initial weights. These plots illustrate that the ANN seems to be slightly biased for small stream flows, whereas the EMD-ANN shows a larger uncertainty than the ANN for high values. This behaviour is better highlighted in Figures 6.5c-f, which show the time series of the difference between hindcasts and observations (Figures 6.5c-d), and the series of the differences of the minimum and maximum hindcast at each time step $\Delta(t) = (max_j \hat{x}_j(t) - min_j \hat{x}_j(t)$ for j = 1, ...,100 (Figures e-f).

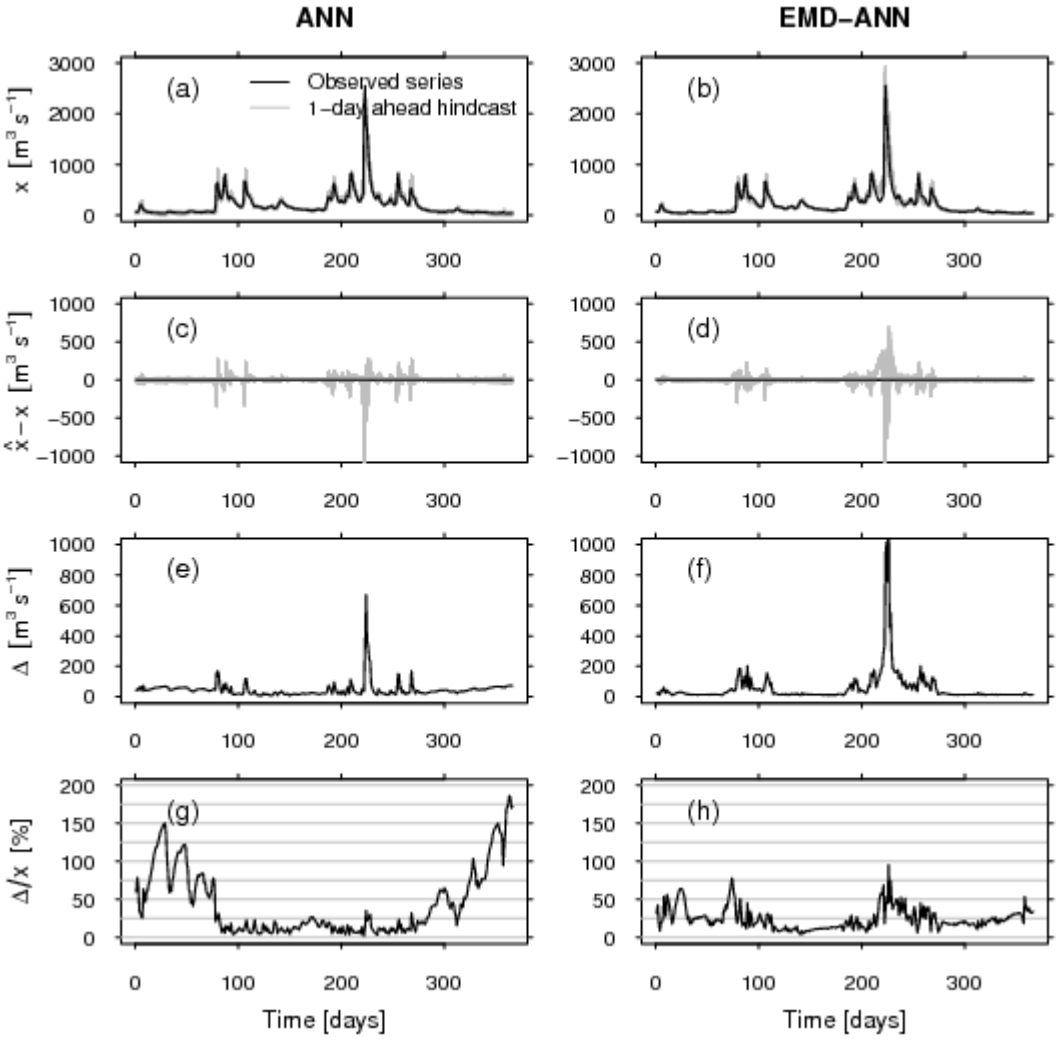


Figure 6.5: Potomac River mean daily discharge from October 2008 to September 2009, and 100 1-day ahead hindcast series from (a) the ANN and (b) the EMD-ANN obtained from 100 sets of initial random weights. Figures (c) and (d) contain the time series of the differences $\hat{x}(t) - x(t)$ corresponding to the time series in (a) and (b). Figures (e) and (f) contain the time series of the differences, $\Delta(t) = \left(\max_j \hat{x}_j(t) - \min_j \hat{x}_j(t)\right)$, $j = 1, \ldots, 100$, which point out the variability of the hindcast at each time step. Figures (g) and (h) are the time series of $\Delta(t)\% = \Delta(t)/x(t) * 100$. Source: Napolitano et al. (2011).

As expected, the highest absolute uncertainty corresponds to the discharge around the peaks, where the process dynamics evolve quickly and the models experience difficulty in following the changing patterns. Figures 6.5g-h represent the values of $\Delta(t)\% = \Delta(t)/x(t) * 100$, which illustrates the weight of the parameter uncertainty relative to the magnitude of the observations. As can be expected, for small values, the uncertainty can be very high, as small differences from small observations can result in high percentage differences. However, for discharge around the peaks (the most interesting for flood forecasting), the ANN parameter uncertainty may generate a variability whose width can reach 25-50% of the recorded values. This point was further investigated by computing $\Delta(t)\%$ values corresponding to the observations above three thresholds defined as the 20th, 50th and 80th percentiles of the records. For all thresholds, the

mean value of $\Delta(t)\%$ is ≈ 15% for the ANN, and ≈ 35% for the EMD-ANN, whereas the maximum values may reach ≈ 1500% (ANN, all thresholds), ≈ 4000% (EMD-ANN, 20th percentile), and ≈ 1000% (EMD-ANN, 80th percentile).

6.5.3 Performance Analysis

The performance measures are provided in Figure 6.6. The first row of boxplots refers to the absolute measures (ME, MAE, MdAE, and RMSE). As the ME is a sign measure allowing both negative and positive values, it shows that the ANN produces hindcasts strongly biased compared to the EMD-ANN, which in turn is comparable with the naïve hindcasts. This behaviour can be ascribed to the bias that characterises the small values produced by the ANN (Figure 6.5). Focusing on absolute errors (MAE and MdAE), contrasting results are obtained. As these metrics do not average out positive and negative values, they highlight the overall errors. The MAE indicates that the ANN performs better than the EMD-ANN, and both are better than the naïve hindcasts, whereas the MdAE leads to opposite conclusions. The MdAE values are smaller than the MAE, meaning that absolute errors are highly skewed. Thus, the two boxplots for MdAE and MAE show that the EMD-ANN performs better than the ANN with respect to small errors (which usually refer to small stream flows), whereas the opposite occurs when the focus is on middle sized errors. Recalling that the RMSE emphasises high errors, and these are likely related to stream flow peaks (and their neighbours), the boxplot for RMSE shows that the EMD-ANN is more accurate than the ANN near peak discharge. Note also that the models perform better than naïve hindcasts for mid to high errors (MAE, RMSE), and worse than the naïve hindcasts for small errors (MdAE). As small (high) errors usually refer to small (high) stream flows, it can be deduced that the naïve hindcast is acceptable for calm (base flow) periods, whereas it is outperformed by models during the rising and recession limbs. This result is not surprising; however, it can be highlighted only by analysing the behaviour of these different metrics.

The MPE and MdAPE are coherent with their absolute counterparts (ME, MdAE), whereas the MAPE shows that the ANN and EMD-ANN are outperformed by the naïve hindcast. A similar behaviour is shown by the RMSPE, which, in addition, highlights a switch between the ANN and EMD-ANN. In terms of percentage measures, the models are invariably outperformed by the naïve hindcast. However, this behaviour should not lead to erroneous conclusions. As shown by the absolute metrics, the naïve hindcast performs better than model for small errors, which tend to occur for periods with smaller stream flows. In these periods the small but systematic errors given by the ANN and EMD-ANN result in very high percentage errors (see also Figures 6.5g-h),

which strongly contribute to the final values of the percentage metrics. The widely used CE shows values between 0.8 and 0.9, which denotes a fairly good performance according to Shamseldin (1997) and Dawson et al. (2007). The high values of CE should not lead to overoptimistic conclusions, as the index adopts the overall mean as a reference model, which provides very poor hindcasts. Moreover, as seasonality is a dominant component, models able to reproduce this property tend to exhibit high CE values. More reliable similarity measures seem to be PI, PI.MAE and PI.MdAE, which involve the naïve hindcast as a reference. PI values are almost all smaller than ≈ 0.5, PI.MAE values do not exceed ≈ 0.3, whereas PI.MAE are negative, denoting that the models are systematically worse than the naïve hindcast.

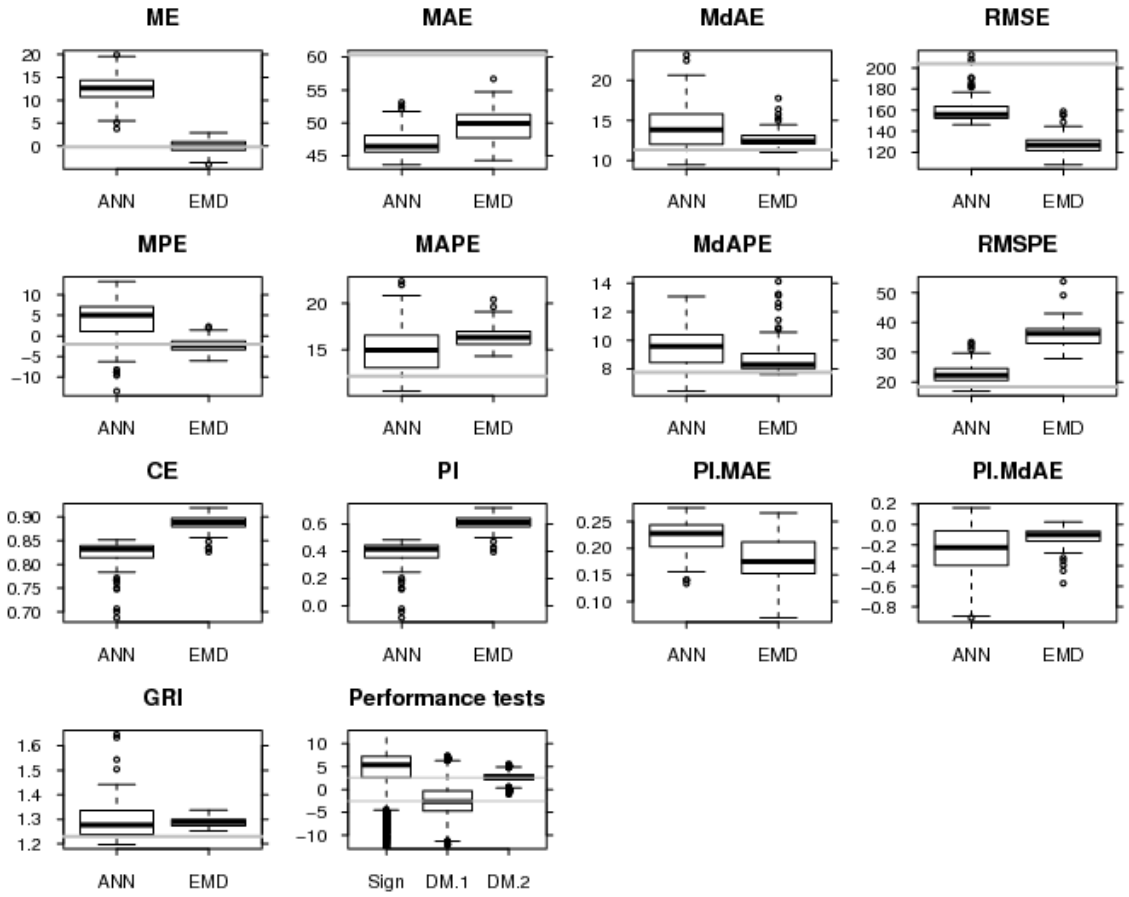


Figure 6.6: Box-plots of the performance measure for the Potomac River. Each box-plot summarises the 100 values of each criterion computed on the ANN and the EMD–ANN series. The gray lines in the boxplots for ME, MAE, MdAE, RMSE, MPE, MAPE, MdAPE, RMSPE, and GRI refer to the reference value corresponding to the naïve hindcast. The gray lines in the boxplot labeled 'Performance tests' refer to the 0.05th and 99.5th percentiles of the standard normal distribution, which define the 99% confidence interval of the test statistics under the null hypothesis for two-sided tests. Source: Napolitano et al. (2011).

It should be noted that these similarity measures provide a piece of information similar to the corresponding absolute measures (RMSE, MAE and MdAE). For example, both the PI.MdAE and MdAE plots show that ANN and EMD-ANN perform worse than the naïve hindcasts. The plot of MdAE reveals information about the absolute difference between the models and the naïve hindcast, whereas the PI.MdAE quantifies the

relative difference (or similarity). The GRI index shows that the naïve hindcast yields errors within [≈ 1/1.25, ≈ 1.25] times the observed values, whereas the errors produced by the ANN and EMD-ANN fall within a wider range: [≈ 1/1.30, ≈ 1.30] times the observed values (on average).

Finally, the sign test and the Diebold-Mariano test were applied to assess the significance of the differences between the ANN and EMD-ANN. The Diebold-Mariano test was applied by using absolute and squared errors (denoted as DB.1 and DB.2). The test statistics were computed for the loss differential sequences resulting from all possible 100 * (100 − 1)/2 combinations of ANN and EMD-ANN series. The Ljung-Box test applied to loss-differential sub-samples confirms that these series are serially uncorrelated, allowing a proper computation of the sign test statistic. The box-plots show that sign test statistic (Sign), DB.1, and DB.2 fall outside the two-sided 99% critical region [−2.58, 2.58] (denoted by grey lines 82%, 56%, and 54% of times, respectively). Hence, the differences between the ANN and EMD-ANN models should be considered significant. However, the negative values of DB.1 denote that the ANN performs significantly better than the EMD-ANN, whereas the DB.2 values suggest opposite conclusions. Since DB.1 relies on absolute errors, while DB.2 on squared errors, this result is coherent with the MAE and RMSE.

6.6 Clark Fork River Results

6.6.1 Preliminary Analysis

For the Clark Fork River training data, the decomposition returns 13 components (12 IMFs plus the residual trend), which is smaller than that expected from the dyadic filter. Figure 6.7 shows that the mean period ω increases with the number of IMFs following approximately the relationship $\omega = \gamma^L$ with $\gamma \approx 2.02$. This confirms that the EMD acts as a dyadic filter bank also for this example. Similar to the Potomac River data, the seventh component appears to represent the annual cycle as the mean period is ≈ 359 days. However, unlike the Potomac River, the plot of the energies and cumulated energies in Figure 6.7 indicates that ≈ 80% of the energy is explained by three IMFs (5, 6 and 7), whose mean periods are ≈ 120, ≈ 211, ≈ 359 days, respectively. Therefore, the Clark Fork River time series is dominated by dynamics acting between seasonal and annual scales.

As the high frequency components are the most difficult to model properly, they represent the main source of error in the modelling stage (as is shown in the previous section for the Potomac River). However, if they contribute a small amount of energy (i.e. a small amplitude) to the overall signal (as in the case of the Clark Fork River),

corresponding modelling errors are expected to be small.

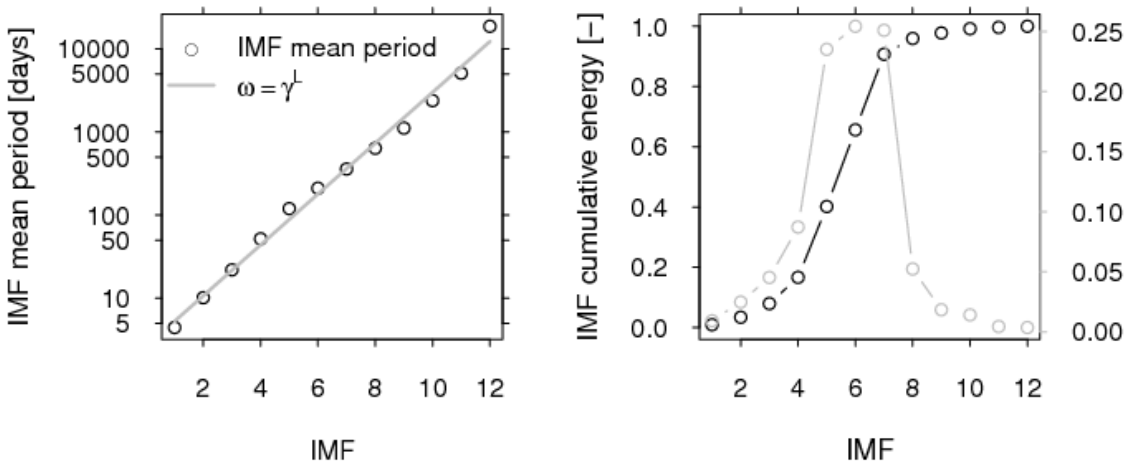


Figure 6.7: On the left is the scaling relationship between the number of IMFs and the corresponding mean periods for the components of the Clark Fork River time series. On the right is the non-dimensional energy (gray) and cumulative energy (black) for each IMF. Source: Napolitano et al. (2011).

Figure 6.8 confirms this statement, showing that the agreement of the ANN and EMD-ANN hindcasts with the discharge observed during the testing period for the Clark Fork River is generally better than the Potomac River results, assuming the same model architecture. Therefore, the structure of the signal plays a key role for the modelling results, and EMD can help to point out in advance possible difficulties which can arise in the modelling stage.

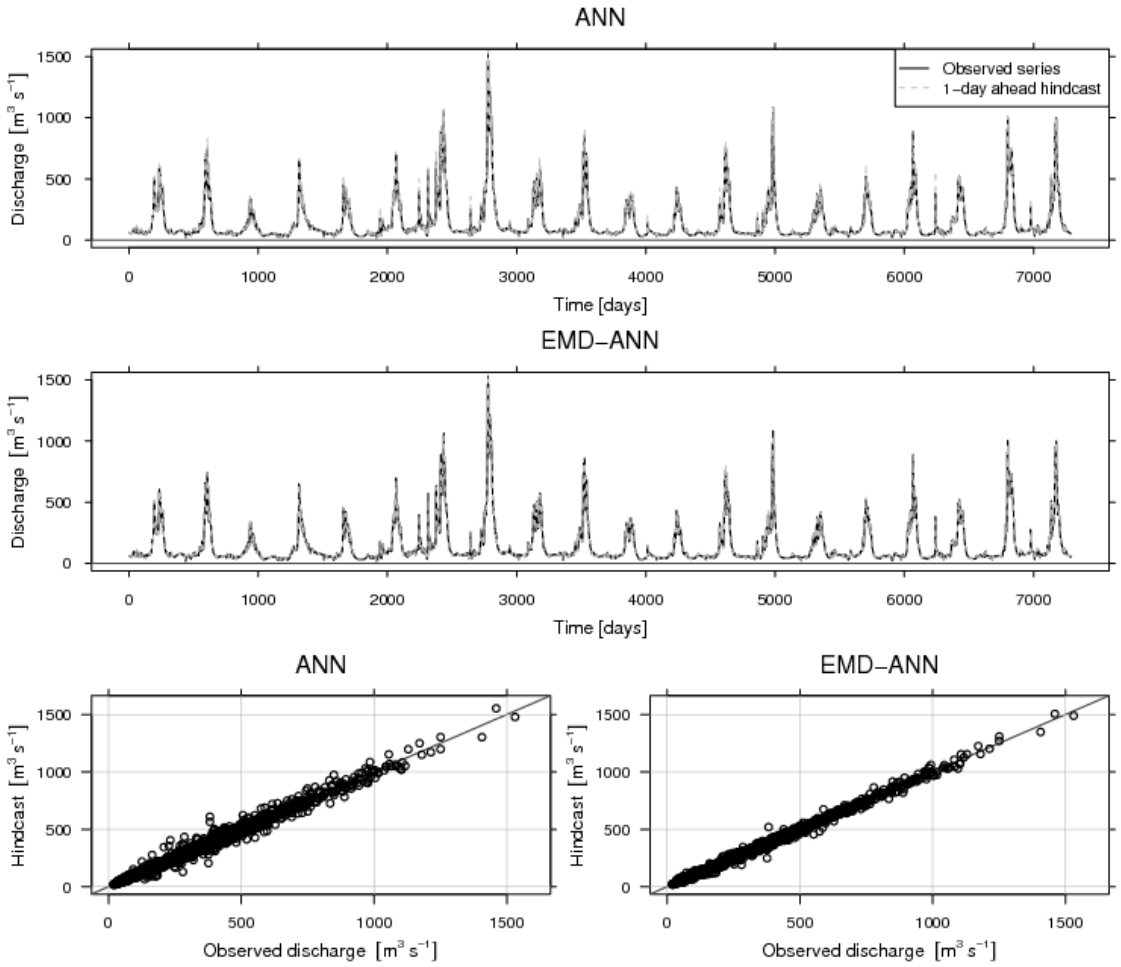


Figure 6.8: The Clark Fork River discharge time series of the test period along with 1-day ahead hindcasts obtained by (a) a simple ANN and (b) the EMD–ANN model. Scatterplots of the observed discharge vs hindcasts computed by (c) the ANN and (d) the EMD–ANN. Source: Napolitano et al. (2011).

6.6.2 Analysis of Weight Initialisation Uncertainty

Figure 6.9 corresponds to Figure 6.5. In particular, Figures 6.9e-h point out that the absolute and percentage width of the error bands related to the random initialisation of the weights, for the ANN is generally smaller than that of the EMD-ANN for both peak and calm periods. As expected, the highest absolute uncertainty corresponds to the discharge around the peaks. However, in terms of percentage, the variability is very small (less than 5%) for the ANN and ≈ 10−25% for the EMD-ANN around the main peak. As for the Potomac River, this point was further investigated by computing $\Delta(t)\%$ values corresponding to the observations above three thresholds defined as the 20th, 50th and 80th percentiles of the records. For all thresholds, the mean value of $\Delta(t)\%$ is ≈ 3% for the ANN, and ≈ 17% for the EMD-ANN, whereas the maximum values may reach ≈ 115% (ANN, all thresholds), ≈ 1000% (EMD-ANN, all thresholds).

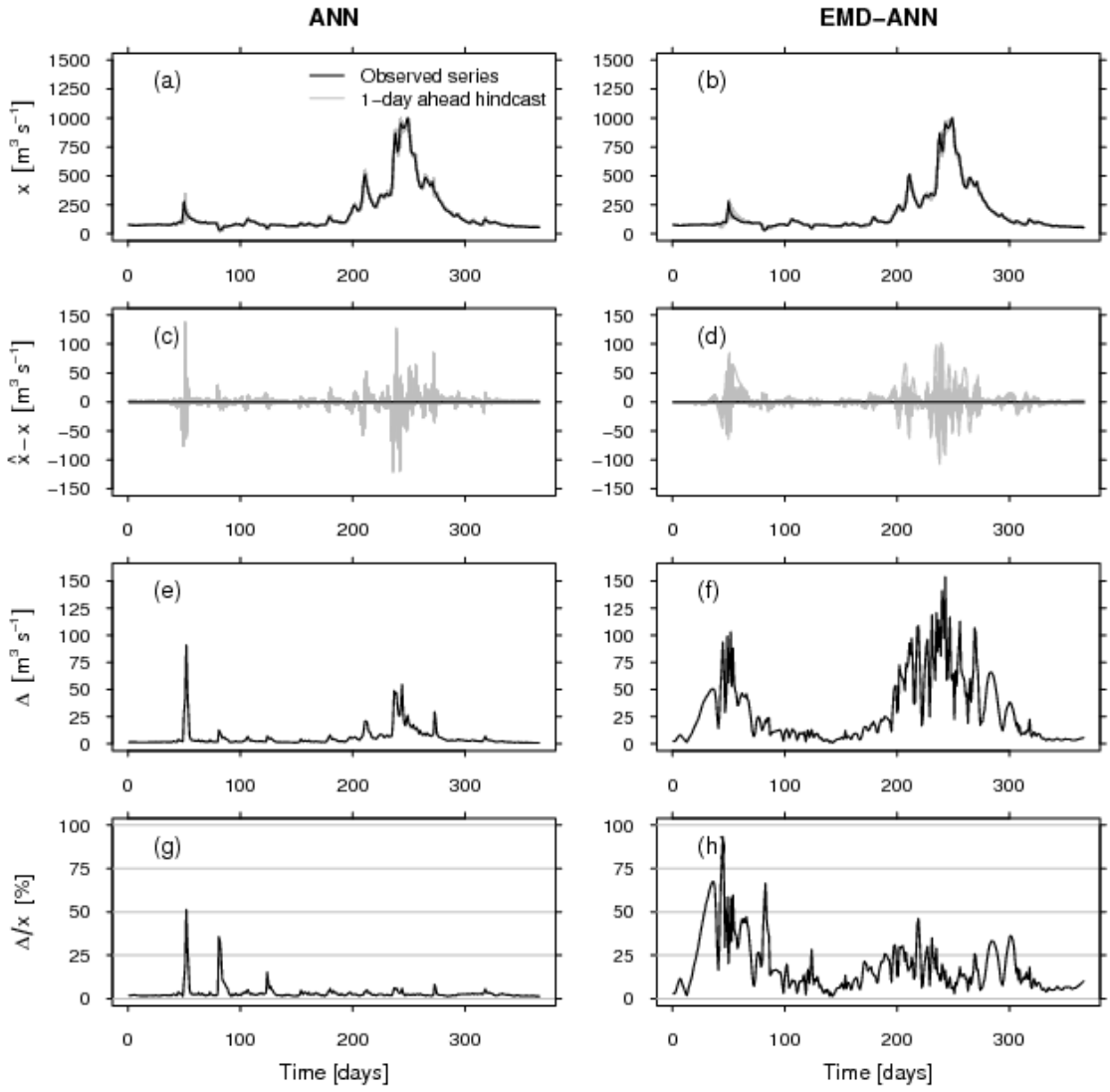


Figure 6.9: The Clark Fork River mean daily discharge from October 2008 to September 2009, and 100 1-day ahead hindcast series from (a) the ANN and (b) the EMD-ANN obtained from 100 sets of initial random weights. Figures (c) and (d) contain the time series of the differences $\hat{x}(t) - x(t)$ corresponding to the time series in (a) and (b). Figures (e) and (f) contain the time series of the differences, $\Delta(t) = \left(\max_j \hat{x}_j(t) - \min_j \hat{x}_j(t)\right)$, $j = 1, \ldots, 100$, which point out the variability of the hindcast at each time step. Figures (g) and (h) are the time series of $\Delta(t)\% = \Delta(t)/x(t) * 100$. Source: Napolitano et al. (2011).

6.6.3 Performance Analysis

The performance measures are displayed in Figure 6.10. The ME shows that the EMD-ANN tends to produce hindcasts more biased than the ANN. Therefore, in this case, the EMD-ANN gives small values that are more biased than the corresponding ANN values. On the contrary, the MAE, MdAE and RMSE indicate that the EMD-ANN outperforms the ANN. As for the Potomac River data, the MdAE values are smaller than the MAE, meaning that the absolute errors are skewed. Thus, the three box-pots for MdAE, MAE and RMSE show that the EMD-ANN performs better than the ANN with respect to small errors (which usually refer to small stream flows), middle sized errors and high errors (which are likely related to stream flow peaks and their neighbours).

Both models outperform the naïve hindcasts in terms of MdAE, MAE and RMSE. All percentage indices (MPE, MAPE, MdAPE and RMSPE) are coherent with their absolute counterparts. For the Clark Fork River data, the bias corresponding to small values is generally small resulting in small percentage errors (see also Figures 6.9g-h). This confirms the good performance of the models across the whole range of discharge values.

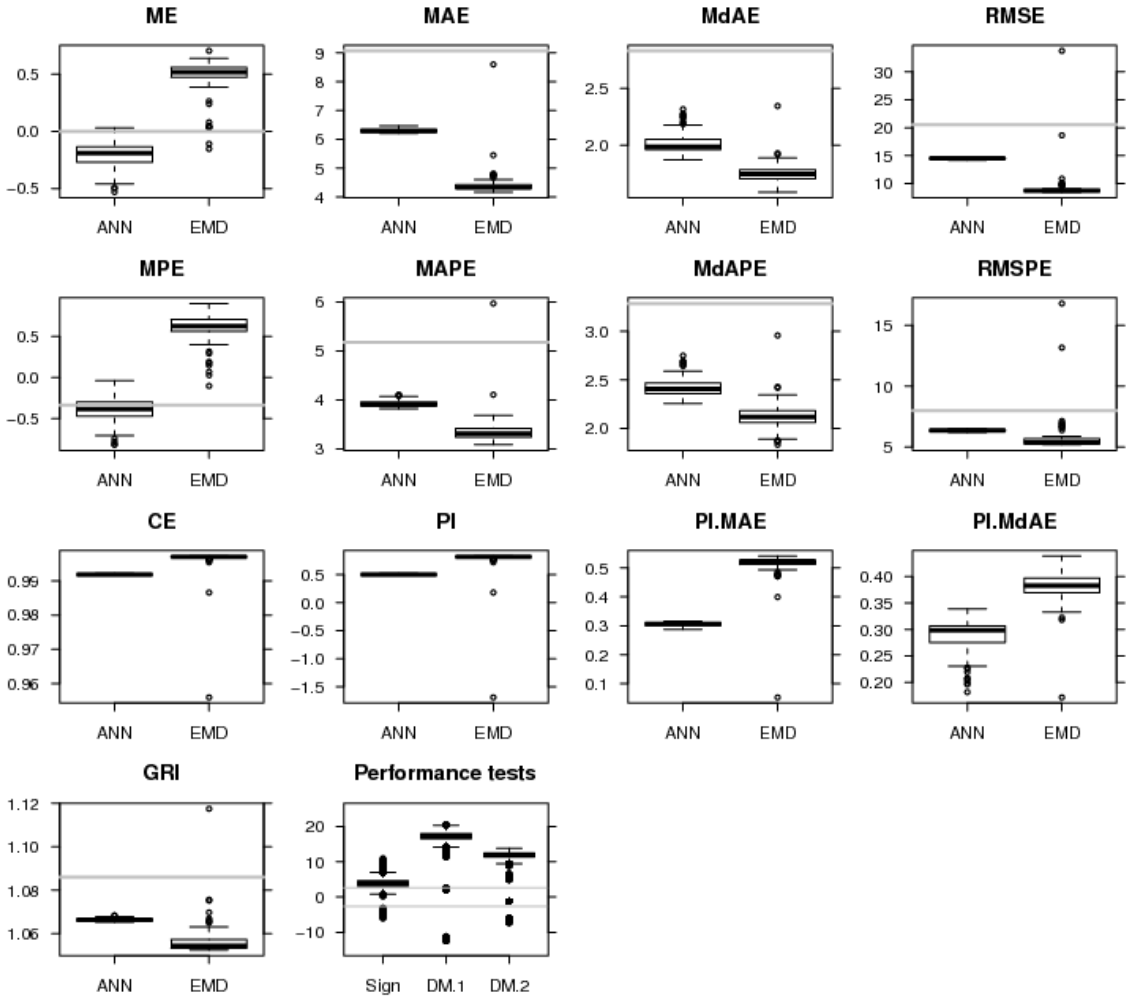


Figure 6.10: Box-plots of the performance measure for the Clark Fork River. Each box-plot summarises the 100 values of each criterion computed on the ANN and the EMD–ANN series. The gray lines in the boxplots for ME, MAE, MdAE, RMSE, MPE, MAPE, MdAPE, RMSPE, and GRI refer to the reference value corresponding to the naïve hindcast. The gray lines in the boxplot labeled 'Performance tests' refer to the 0.05th and 99.5th percentiles of the standard normal distribution, which define the 99% confidence interval of the test statistics under the null hypothesis for two-sided tests. Source: Napolitano et al. (2011).

The CE shows values greater than 0.99, which denotes a good performance according to Shamseldin (1997) and Dawson et al. (2007). As for the Potomac River, the high values of CE only denote that the models strongly outperform the overall mean as a reference model. When a more representative naïve option is assumed (as for PI, PI.MAE and PI.MdAE), the similarity measures do not exceed the value 0.6. As previously mentioned, these similarity measures provide a piece of information similar

to the corresponding absolute measures (RMSE, MAE and MdAE). The GRI index shows that the naïve hindcast yields errors within [≈ 1/1.09, ≈ 1.09] times the observed values, whereas the errors produced by the ANN and EMD-ANN fall within a smaller range: [≈ 1/1.07, ≈ 1.07] and [≈ 1/1.05, ≈ 1.05] times the observed values (on average), respectively. Finally, the sign test and the Diebold-Mariano test were applied following the same approach as for the Potomac River. The box-plots show that all the test statistics fall outside the two-sided 99% critical region [−2.58, 2.58] (denoted by grey lines) 11% (Sign), 1% (DB.1), and 1% (DB.2) of times, respectively. Hence, the differences between the ANN and EMD-ANN models should be considered statistically significant. The positive values of the test statistics indicate that EMD-ANN performs significantly better than the ANN. This result is coherent with the MAE, MdAE and RMSE as well as the corresponding similarity measures.

6.7 Discussion

These experiments have compared the one day-ahead hindcast performance of two modelling strategies: a multilayer perceptron feedforward ANN and an ensemble counterpart deduced by decomposing the signal via the EMD technique. The analysis quantified the output uncertainty corresponding to the random initialisation. A number of redundant and non-redundant measures of performance, and formal tests were applied to assess the statistical significance of the differences between the considered models was used as set out in Chapter 3. The analyses were carried out on two long daily stream flow series with good quality data, and free from evident abrupt changes and trends, to emphasise the differences of the models on real but well-behaved data.

The preliminary analyses of the signal components obtained by EMD show that some intrinsic modes are characterised by physical meaningful mean periods. The energy of each component allows for a better understanding of the contribution of each extracted mode to the overall signal and to help foresee possible difficulties and potential solutions to a certain extent in the modelling stage. In particular, as the high-frequency components are generally difficult to model, if they exhibit high energy, larger output errors can be expected than those related to signals that are dominated by mid-low frequency modes. Therefore, the results further point to EMD as a valuable tool for detecting signal properties and can be useful in improving the modelling.

The analysis of the output uncertainty caused by the random initialisation of the ANN weights shows that this source of variability can exhibit significant width, depending on the overall model performance. When the model closely fits the data, the mean values of the uncertainty width can be negligible (≈ 3% of the observed discharge, on average,

for the ANN and Clark Fork River data) or not (≈ 15 − 35% of the observed discharge, on average, for the ANN and Potomac River data, and EMD-ANN), whereas the maximum values are always very large (from ≈ 100% to ≈ 4000%). In general, attempts to merge EMD and ANN in ensemble models produced a greater output uncertainty in the estimates at the mid to high stream flow values (values above different thresholds), which is expected and can be ascribed to the propagation of the modelling error throughout the IMF components.

A further observation can be made about the weight initialisation. The initialisation of weights is essentially related to the mathematical nature of ANNs. In this context, Giustolisi and Laucelli (2005) and Giustolisi and Simeone (2006) proposed an optimisation approach, which considers the weights as a decision variable and aims at finding the optimal set of parameter values. Focusing on the nature of real-world data Wang et al. (2006) argued that attempts at choosing the best ANN model is not very sound for a number of different reasons, namely: (1) in real-world cases, the testing set is not observed yet and it is not certain whether the best model for the training and/or validation datasets is also the best one for the testing set; (2) even though the training dataset is large, it cannot take possible future changes (such as climatic or anthropogenic) into account; (3) the chosen model is the best one according to some performance metric or criterion, where different criteria can give different model scores. Therefore, as far as point estimates/forecasts are concerned, a possible strategy to overcome the weight uncertainty is the model averaging suggested by Wang et al. (2006), who trained 10 models, took the five best ones, and then took their average as the final output. On the other hand, the uncertainty related to the different final configurations of the weights should be accounted for in a comprehensive assessment of the overall uncertainty. In this case, the errors resulting from fitting several ANNs with different initial weights can be used to build an error model, such as the parametric meta-Gaussian error model suggested by Montanari and Brath (2004), or resorting to techniques that treat the weights and biases of ANNs as fuzzy numbers rather than crisp numbers (Alvisi and Franchini, 2011). The rationale of these approaches is partly related to the final scope, which moves progressively from a point estimation to an interval estimation. These issues are beyond the scope of this research. Instead, this research has shown the possible magnitude of the error resulting from the random initialisation, highlighting its relation to signal properties.

The study of the model performance shows that the larger uncertainty of the EMD-ANN can result in smaller bias and better hindcast performance. For the Potomac River data, the ensemble approach provided less biased results for low flows. However,

focusing on the overall model accuracy, the performance scores provided contrasting results that depend on the chosen performance measure. In general, measures based on signed errors and squared errors tend to favour the EMD-ANN, whereas indices based on absolute errors indicate that the ANN models are better. These contrasting results reflect the intrinsic difficulties in modelling a signal with highly energetic high frequency components. For the Clark Fork River stream flow signal, which is dominated by mid-low frequency modes, the EMD-ANN exhibited a significant improvement with respect to the ANN. As the EMD-ANN is much more complex than the ANN, this result could be expected; however, the Potomac case study shows that a more complex model can fail to improve the results. The EMD analysis helped to recognise that the nature of the data plays a key role in the propagation of error. The EMD-ANN outperforms the ANN when the error of the high-frequency modes is small (Clark Fork River data), whereas it can be outperformed when the error of the high-frequency modes is large. Therefore, the success of an ensemble approach over a non-ensemble depends to some extent on the properties of the high frequency components of the signal.

The results show that non-redundant indices can provide discordant results (as for Potomac River data), whereas the redundant measures can support each other and further illustrate that the use of other non-redundant indices is necessary for a fair model assessment. Therefore, the application of several appropriate measures and the comparison with simple benchmark models should become standard practice for a fair model assessment. In particular, the relationships between absolute metrics, deviance measures and similarity measures must be taken into account to avoid the use of criteria which appear to be different, but provide similar information.

6.8 Summary

This chapter explicitly considered the impact of random weight initialisation on the results of the ANN. This is an important issue as there has been little research done on this topic. Moreover, many papers simply report the results from one optimal ANN chosen through trial and error without considering the effect of weight initialisation. The chapter showed that the impact can be large. The chapter also investigated a pre-processing technique called empirical mode decomposition, which has not been applied to ANN rainfall-runoff modelling before. EMD provides a good method for initial investigation of a time series. However, model performance is a function of the error in the high-frequency modes. When this error was small, the EMD-ANN outperformed the ANN while the opposite was true when the error was large. Thus, this technique must be used with particular knowledge of the properties of the specific time series and

catchment being modelled. Moreover, the EMD-ANN generally resulted in a greater output uncertainty in the medium to high flow predictions. Finally, the suite of performance measures outlined in Chapter 3 was systematically applied to the two catchments. The results showed that the use of redundant indices help to support each other in providing the same message but non-redundant indices can show conflicting results. Therefore, it is imperative to apply several measures to each modelling exercise including comparison with benchmark models in order to fully understand the model performance. The next chapter concludes the thesis and provides recommendations for further research.

Chapter 7

Conclusions and Recommendations for Further Research

7.1 Introduction

This chapter summarises the findings of the research in relation to the aims and objectives as set out in section 1.2 in order to show how they have been achieved. The contributions of the research are also highlighted in this first section. The second section contains a discussion of the limitations of the research and any problems encountered during the study. A set of recommendations in the form of a short research agenda for ANN rainfall-runoff research comprises the final section of this chapter and of the thesis.

7.2 Summary of the Research Findings

The overall aim of this thesis was to examine a number of issues related to ANN rainfall-runoff modelling that have been identified from a review of the literature, in particular the need to rigorously compare ANNs with conceptual/physical models; the use of different performance measures for model evaluation; the problems associated with training ANNs using different random weight initialisations; and the use of ensemble methods, all of which have been identified as ongoing issues from a review of the literature. Below are the objectives of the research as stated in section 1.2. Following each objective is a description of how the objective has been achieved and the significance of the research findings.

Objective 1: To review and critically evaluate the academic literature on ANN rainfall-runoff modelling

A comprehensive review of ANNs was undertaken in Chapter 2. The first part of the chapter placed ANNs in the schema of approaches to hydrological modelling. A brief history and definitions were then provided. This was followed by on overview of ANN structures and model development. This section clearly highlighted the general lack of guidance that exists, where most aspects of model development are undertaken via trial and error or based on heuristics from the literature. The advantages and disadvantages were then provided as a critique of this method, where the disadvantages drive some of the research happening in this area. Finally, the existing body of literature was reviewed starting with the initial phase of activity in ANN rainfall-runoff modelling. Since the appearance of the first paper around 1985, there have been hundreds of papers published on this topic. For this reason, the research review was

divided into the main themes that emerged from the literature review. An indication of the volume of research in this area is also clear from the review papers that have appeared around the year 2000 and more recently in 2010. From these review papers, the areas that were recommended for further research were summarised. From this list, two key areas relevant to this research were discussed, i.e. the need for considerations of uncertainty and further research into ensemble methods. Both of these concepts were then considered within objectives 4 and 5.

Two other important areas were flagged. These were not part of any individual theme but based on observations from the research reviewed. The first was a greater need to compare ANNs with conceptual and/or physically-based models. This is the only way that hydrologists will be convinced that ANNs are a viable technology for the operational environment. This is far from currently established and ANNs may actually not end up being the most appropriate tool. However, research in this direction is still required. Secondly, there was no consistent pattern in the application of performance measures to evaluate ANN rainfall-runoff models. This has been further elaborated in Objectives 2 and 6.

Objective 2: To review and evaluate the measures that are used to evaluate model performance, choosing a subset for use in the research

An overview of the most commonly used performance measures was provided in Chapter 3. These were cateogorised into absolute, relative and those based on benchmark models. From this set of measures, a subset was chosen including five absolute measures (ME, MAE, MdAE, RMSE and PDIFF), four relative measures, which correspond to the first four absolute measures (MPE, MPAE, MPdAE and RMSPE) and five benchmark measures (CE, PI, PI.MAE, PI.MdAE and GRI). Two further measures were introduced for comparing two different models against one another, which are used in economics but have not been applied in hydrology before. These are the sign test and the Diebold-Mariano test. The list chosen was intended to be comprehensive as well as including both redundant and non-redundant measures. The idea was to see whether there is consistency of message between measures, consistency of measures between models, and to determine the general utility of the measures. The results of the application of these different measures highlighted situations in which the performance measures were in complete agreement, situations in which the absolute and relative measures agreed but the benchmark-based measures did not, and situations where no discernible pattern was revealed when comparing sets of experiments. Therefore, the application of several appropriate measures and the comparison with simple benchmark models should become standard

practice for a fair model assessment. Moreover, visual inspection of the hydrograph proved to be valuable in some instances where the situation based on the evaluation measures alone was not incisive.

Objective 3: To develop an ANN rainfall-runoff model of the Tiber River basin and compare this with the conceptual TEVERE model

Prior to development of the ANN model, an overview of the Tiber River basin was provided including geology, land use, climate, etc. The flood events in 2005 and 2008 that affected the city of Rome, and were predicted by both the ANN and conceptual model, were described in detail. An ANN model was then built to predict these two flood events at Ripetta gauging station in Rome using historical water level data at Ripetta and one upstream station at Orte for a lead time of 12 and 18 hours. No rainfall was used in the model. Trial and error was used to find the optimal configuration resulting in 10 hidden nodes. Bayesian Regularisation was used to train the ANN because this algorithm does not require a validation dataset to avoid overtraining. It also has the advantage that more data can be used for both training and testing, which was necessary due to the small amount of data available for development of the ANN. This is primarily because flood events were extracted from the historical record and this greatly reduced the amount of data available. The networks were trained 50 times according to guidance provided by Anctil (2007) to compensate for variations in model predictions caused by the random initialisation of the weights. The conceptual model of Calvo and Savi (2009) was also described, which was then specifically run to predict the 2005 and 2008 events for the same lead times of 12 and 18 hours.

The two models were then compared using the suite of performance measures chosen as part of Objective 2 with the exception of the sign and Diebold-Mariano tests. Unfortunately the length of the model predictions was too short and they were not independent so the tests were inapplicable. Visual inspection of the hydrographs was also undertaken. Overall both models generally performed well. At a lead time of 12 hours, the conceptual model was superior. This is reflected in the suite of performance measures applied. Examination of the hydrographs showed excellent correspondence by the conceptual model. For the ANN, the prediction of the rising limb was late but the rest of the hydrograph was predicted well. At a lead time of 18 hours, the conceptual model was late in predicting the rising limb while the ANN produced a better result on the rising limb (although also slightly late) but then the rest of the hydrograph was not predicted as well as the conceptual model.

Although the conceptual model outperformed the ANN, there are two major issues with

the model: a) the lack of forecasted rainfall resulted in a late prediction for the 2008 event; b) the model requires 18 hours for calibration so the first 18 hours of any flood event will not be predicted. This conceptual model can therefore only be used on catchments with very long lead times. The ANN is able to predict the first 18 hours so one could see an obvious synergy between the use of the two models together. Moreover, the ANN did not use any rainfall in this first main experiment and it had obvious issues with predicting the latter part of the hydrograph. It is therefore clear that improvements are possible but after the research undertaken in this chapter, one could only recommend the conceptual model or a combination of the conceptual and ANN model working together.

Objective 4: To undertake a series of different experiments to improve the basic ANN rainfall-runoff model developed as part of Objective 3 and briefly examine methods of ensemble combination

This objective was addressed in Chapter 5 in which different experiments were undertaken with the overall aim of improving the ANN model developed in Chapter 4. The first set of experiments were designed to see whether the model could extrapolate to the 2008 event, which was harder to predict than the 2005 event because an event of such a magnitude was not present in the training dataset. The experiments involved adding more upstream stations and rainfall. Both clearly improved the model performance in terms of both the quantitative evaluation measures and a visual inspection of the hydrograph. Thus, the objective was achieved and the ANN was able to better predict the 2008 event. During the course of these experiments, the large degree of variation between ensemble members in their predictions, especially at the peaks, was observed. As a result, the impact of the random initialisation of the ANN weights was further investigated in Objective 5.

The second part of the objective concerns methods of ensemble combination. A simple average was employed as originally suggested by Anctil (2007), who also used BR to train ANNs in a hydrological context. This second part of Objective 4 was acheived through the application of three different methods of ensemble combination: use of a PI threshold to select the best models before averaging; the AIC; and a modified AIC after Zhao et al. (2008). In the case of the AIC and modified AIC, the ensemble members were weighted and then linearly combined. The results showed that the average generally worked well. However, for the 2008 flood event, both the PI and AIC-based ensemble combination worked better. Thus, a more refined ensemble combination method is recommended.

Objective 5: To apply a pre-processing method called Empirical Mode Decomposition (EMD) to ANN rainfall-runoff modelling and examine the impact of random weight initialisation on the ANN model outcomes

Empirical Mode Decomposition (EMD) is a method of decomposing a time series (Huang et al., 1998) where the individual signals can then be modelled separately and recombined to create a single prediction. This method has not been applied before to ANN rainfall-runoff modelling so the application undertaken in this thesis and now published in Napolitano et al. (2011) represents a significant scientific contribution to the literature on ANN rainfall-runoff modelling. The experiments undertaken as part of achieving this objective involved the comparison of one day ahead hindcasts of an ANN and an ensemble counterpart produced by decomposing the signal via the EMD technique. The methods were applied to two rivers with long time series in the USA: the Potomac River and the Clark Fork River. The EMD was first used to analyse the time series of each river. The analysis demonstrated that EMD can be used as a valuable tool for detecting signal properties, which can be used to improve the modelling.

The output uncertainty caused by the random initialisation of the ANN weights was shown to be large at the highest values (from ≈ 100% to ≈ 4000%) and then varied depending upon the dataset, e.g. 3% of the observed discharge, on average, for the ANN and Clark Fork River data or (≈ 15% to 35% of the observed discharge, on average, for the ANN and Potomac River data, and EMD-ANN). In general, attempts to merge the EMD and ANN in ensemble models produced a greater output uncertainty in the estimates at the mid to high stream flow values (values above different thresholds), which is expected and can be ascribed to the propagation of the modelling error throughout the IMF components.

Objective 6: To highlight the limitations of the study and to make recommendations for further research

The limitations and recommendations for further research are outlined in the next section, which comprises a short research agenda for the future. This section also concludes the thesis.

Thus, the six objectives set out in section 1.2 were achieved. In general the research demonstrated that ANN models can be built that have good performance from an operational perspective although a complementary approach with the conceptual model would be more beneficial than using one model alone. The research also highlighted the importance of the random initialisation of the weights and the need to

use a suite a performance measures to provide a comprehensive model evaluation. Finally EMD was shown to be a technique that has a great deal of potential for ANN rainfall-runoff modelling.

7.3 Limitations and Problems Encountered During the Research

One major limitation of this study regarded the data available for modelling on the River Tiber. Although there is theoretically a large network of rain gauges and rainfall stations, not all of the data for the stations were available and some stations did not record hourly data. When station data were available, the records were incredibly messy. This required reformatting and cleaning of the data (as there was inconsistency between stations) and filling in missing records when possible. Issues with the data were another reason why experiments with a more parsimonious model were undertaken. It is clear that adding more upstream stations often improved the ANN. However, to have complete records at all stations for the same event was quite rare. Thus, the use of more stations meant less data for training and testing because of missing data. ANNs require a considerable amount of data. Conceptual models have a definite advantage over data-driven models when the available data are sparse. This was another reason for using the long time series for the Potomac River and Clark Fork River to continue experimentation in Chapter 6 on EMD and the random weight initialisation issue. However, it was also good to examine these issues on different catchments to consider how transferable the conclusions are to different areas.

A second problem was lack of access to the conceptual TEVERE model due to the unfortunate death of Prof Savi part way through this research. This meant that further experimentation with the conceptual model beyond what appears in Chapter 4 was not possible. For example, plans to couple an ANN rainfall forecasting model with the TEVERE model had to be abandoned. Other plans to couple elements of the conceptual model like that undertaken by Corzo et al. (2009) or examine the hidden nodes of an ANN using a conceptual model (e.g. Wilby et al., 2003) were simply not possible. Thus, a shift in the research direction was taken partway through the PhD.

7.4 Recommendations for Further Research

A number of areas for further research can be recommended from this research study. These include the following:

1. Research is required that specifically addresses the operational capability of ANNs. The literature is full of examples of how ANNs outperform existing models (see Chapter 2) but the movement from academic study to operation is not going to

happen unless individuals or organisations are brave enough to really test out this technology. The development and maintenance of real-time flood forecasting systems cost an immense amount of money. The development of an ANN flood forecasting system would be a very small investment in comparison.

There are a number of ways that could be suggested for encouraging this technology transfer. The first is that ANNs need to be part of mainstream hydrological modelling education, embedded in undergraduate and postgraduate courses. If there is sufficient understanding of how these tools work and what they can and cannot do, then when students who took these courses go onto work in civil protection agencies, they may be more inclined to try alternative approaches. Secondly, agencies that fund research need to be lobbied in order to add ANN modelling to their research agendas. Finally, research should also consider ANNs as complementary rather than competitive approaches to traditional hydrological modelling as suggested in section 7.2 under the Objective 3 summary. As ANNs have a fast computation time and a low burden of operation, they would happily sit alongside a conceptual model. When forecasters have more than one piece of evidence on which to make operational decisions, it should then be much easier to convince decision makers to act at higher levels. However, this requires ANN rainfall-runoff researchers to be much more proactive in trying to transfer the technology to the operational environment. However, many express little interest in this (See, 2008, personal communication) as evidenced by a recent survey of operational ANN applications in the water resources industry (Macdonald and See, 2010).

2. More research is needed to examine EMD as both a diagnostic tool to better understand a time series (or a particular catchment) and as a modelling tool to be used in combination with ANNs. As this was the first attempt at applying EMD to ANN rainfall-runoff modelling, there are many further developments possible in this field.

3. More research needs to be undertaken in methods of uncertainty for ANN models. Although some methods are available, more explicit guidance must be developed to help ANN rainfall-runoff modellers to calculate and report uncertainty as standard practice.

4. The impact of the random initialisation of weights of the ANN needs further investigation. The majority of studies in the literature do not take this into account, reporting only the best single model produced. As models are often only tested on

a small dataset to evaluate model performance, the single instance of an ANN cannot really be trusted given the range of model predictions that were observed across the ensemble members, especially at the peak of a flood event. Studies that look at the sensitivity of ANN models to multiple independent datasets and compare these to ensemble model performance should be undertaken. This will require catchments with long time series or could be undertaken using synthetically generated datasets with different properties and distributions. This research should consider the effect on the performance measures as well as the resulting uncertainty in the model predictions.

5. A suite of performance measures should be used to assess model performance including absolute, relative and benchmark-based models. The PDIFF was not a very useful error measure so a better measure should be devised, e.g. difference between the model prediction at the highest observed point of the peak or at a level that triggers an operational event. Visual inspection of the hydrograph proved crucial in some experiments and should always be part of the performance evaluation. More operational measures could also be devised that better evaluate the usefulness of ANN methods for an operational environment rather than global measures that only provide some idea of overall goodness-of-fit. Some research that attempts to standardise practice in this area is therefore recommended.

References

Abbott, M. B., J. C. Bathurst, J. A. Cunge, P. E. O. Connell, and J. Rasmussen (1986a). An Introduction to the European Hydrological System-Systeme Hydrologique Europeen, "She", 1: History and Philosophy of a Physically-Based Distributed Modelling System, *Journal of Hydrology*, *87*, 45-59.

Abbott, M. B., J. C. Bathurst, J. A. Cunge, P. E. O. Connell, and J. Rasmussen (1986b). An Introduction to the European Hydrological System-Systeme Hydrologique Europeen, "She", 2: Structure of a Physically-Based Distributed Modelling System, *Journal of Hydrology*, *87*, 61-77.

Abebe, A.J. and Price, P.K. (2004). Information theory and neural networks for managing uncertainty in flood routing. *Journal of Computing in Civil Engineering*, 18(4), 373-380.

Abrahart, R.J., See, L. and P.E. Kneale (1999). Using pruning algorithms and genetic algorithms to optimise networks architectures and forecasting inputs in a neural network rainfall-runoff model. *Journal of Hydroinformatics*, 1, 103-113.

Abrahart R.J. (2001). Single-model-bootstrap applied to neural network rainfall-runoff forecasting. In: *Proceedings of the 6th International Conference on GeoComputation.*

Abrahart, R.J. and See, L.M. (2000). Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrological Processes*, 14(11-12), 2157-2172.

Abrahart R.J. and See, L. (2002). Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. *Hydrology and Earth System Sciences*, 6, 655-670.

Abrahart R.J., See L. and Kneale P.E. (2001). Applying saliency analysis to neural network rainfall-runoff modelling. *Computers and Geosciences*, 27, 921-928.

Abrahart, R.J. and See, L.M. (2007a). Neural network emulation of a rainfall-runoff model. *Hydrology Earth System Sciences*, 11, 1563-1579.

Abrahart, R.J., Heppenstall, A.J. and See, L.M. (2007b). Timing error correction procedures applied to neural network rainfall-runoff modelling. *Hydrological Sciences Journal*, 52(3), 414-431.

Abrahart, R.J., See, L. and Dawson, C.W. (2008). Neural network hydroinformatics: Maintaining scientific rigour. In: Abrahart, R.J., See, L. and Solomatine, D.P. (eds) *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, pp. 33-47. Springer-Verlag, Heidelberg.

Abrahart, R.J., See, L.M. and Heppenstall, A.J. (2007c). Neuroevolution applied to river level forecasting under winter flood and drought conditions. *Journal of Intelligent Systems*, 16(4), 373-386.

Abrahart, R.J., See, L.M., Dawson, C.W., Shamseldin, A.Y. and Wilby, R.L. (2010). Nearly two decades of neural network hydrological modelling. In: Sivakumar, B. and Berndtsson, R. (Eds.) *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*. World Scientific, New Jersey, USA.

Achela, D., Fernando, K. and. Shamseldin, A.Y. (2009). Investigation of internal functioning of the radial-basis-function neural networks for river flow forecasting models. *Journal of Hydrologic Engineering*, 14, 3, 286-292.

Adamowski, J. and Sun, K. (2010). Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. *Journal of Hydrology*, 390(1-2), 85-91.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on information theory, pp. 267-281.

Alvisi, S. and Franchini, M. (2011). Fuzzy neural networks for water level and discharge forecasting with uncertainty. *Environmental Modelling & Software*, 26, 523–537.

Amari, S., Murata, N., Muller, K. R., Finke, M. and Yang, H.H. (1997). Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks*, 8(5), 985-996.

American Society of Civil Engineers (1993). Criteria for evaluation of watershed models. *Journal of Irrigation Drainage Engineering*, 119(3), 429-442.

American Society of Civil Engineers (2000a). Artificial neural networks in hydrology. I: Preliminary concepts, *Journal of Hydrologic Engineering*, 5(2), 115-123.

American Society of Civil Engineers (2000b). Artificial neural networks in hydrology. II: Hydrologic applications, *Journal of Hydrologic Engineering*, 5(2), 124-137.

Ampazis, N. and Perantonis, S.J.(2002). Two highly efficient second-order algorithms for training feedforward networks. *IEEE Transactions on Neural Networks*, 13(5), 1064-1074.

Anctil, F. (2007). Tools for the assessment of hydrological ensemble forecasts. In: *Proceedings of the International Workshop on Advances in Hydroinformatics 2007*, P. Coulibaly (Ed.), 4-7 June 2007, Niagara Falls, Canada.

Anctil, F. and Rat, A. (2005). Evaluation of neural network stream flow forecasting on 47 watersheds. *Journal of Hydrologic Engineering*, 10(1), 85-88.

Anctil, F. and Tape, D.G. (2004). An exploration of artificial neural network rainfall-runoff forecasting combined with wavelet decomposition. *Journal of Environmental Engineering and Science*, 3(S1), S121-S128.

Anctil, F., Lauzon, N., Andreassian, V., Oudin, L. and Perrin, C. (2006). Improvement of rainfall-runoff forecasts through mean areal rainfall optimization. *Journal of Hydrology*, 328, 717-725.

Anctil, F., Perrin, C. and Andreassian, V. (2003). ANN output updating of lumped conceptual rainfall/runoff forecasting models. *Journal of the American Water Resources Association*, 39, 1269-1279.

Anderson, M.G., and T.P. Burt (1985). Modelling Strategies, in *Hydrological Forecasting*. Anderson, M.G. and Burt, T.P. (eds.), pp.1-13, John Wiley & Sons Ltd., Suffolk.

Andrews, R., Diederich, J. and Tickle, A.B. (1995). A survey and critique of techniques for extracting rules from trained neural networks. *Knowledge Based Systems*, 8, 373-389.

Aqil, M., Kita, I., Yano, A. and Nishiyama, S. (2007). Neural networks for real time catchment flow modeling and prediction. *Water Resources Management*, 21(10), 1781-1796.

Araujo, P., Astray, G., Ferrerio-Lage, J.A., Mejuto, J. C., Rodriguez-Suarez, J. A. and Soto, B. (2011). Multilayer perceptron neural network for flow prediction. *Journal of Environmental Monitoring*, 13(1), 35-41.

Arnold, J.G. and Fohrer, N. (2005). SWAT2000: current capabilities and research opportunities in applied watershed modeling. *Hydrologic Processes*, 19, 563-572.

Arnold, J.G., Srinivasan, R., Muttiah, R.S. and Williams, J.R. (1998). Large area hydrologic modeling and assessment Part I: model development. *Journal of the American Water Resources Association*, 34(1), 73-89.

Autorità di bacino del Fiume Tevere (2006). *Il fiume Tevere a Roma*-portolano, Edizioni Ambiente.

Benitez, J.M., Castro, J.L., Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, 8(5), 1156-1164.

Beran, M. (1999). Hydrograph prediction – How much skill? *Hydrology and Earth System Sciences*, 3(2), 305-307.

Bersani, P. and Bencivenga, M. (2001). Le piene del tevere aroma dal V secolo a.c. all'anno 2000. Presidenza del Consiglio dei Ministri Dipartimento per i Servizi Technici Nazionali Servizio Idrografico e Mareografico Nazionale.

Beskow, S., Mello, C.R., Norton, L.D. and da Silva, A.M. (2011). Performance of a distributed semi-conceptual hydrological model under tropical watershed conditions. *Catena*, 86(3), 160-171.

Beven, K., Calver, A. and Morris, E.M. (1987). *Institute of Hydrology Distributed Model*. Internal Report. Institute of Hydrology, Wallingford.

Beven, K.J. and Kirby, M.J. (1979). A physically based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24, 43-69.

Bhattacharya, B., Lobbrecht, A.H. and Solomatine, D.P. (2003). Neural networks and reinforcement learning in control of water systems. *ASCE Journal of Water*

Resources Planning and Management, 129(6), 458-465.

Birikundavyi, S., Labib, R., Trung, H.T. and Rousselle, J. (2002). Performance of neural networks in daily streamflow forecasting. *Journal of Hydrological Engineering*, 7(5), 392-298.

Bishop, C.M. (1995). *Neural Networks for Pattern Recogni*tion. Oxford University Press, Oxford UK.

Blackie, J.R. and Eeles, C.W.O. (1985). Lumped catchment models. In *Hydrological Forecasting*, Anderson, M.G. and Burt, T.P. (eds.), pp.311-345, John Wiley & Sons. Ltd., Suffolk.

Bonan, G.B. (1999). Frost followed the plow: Impacts of deforestation on the climate of the United States. *Ecological Applications*, 9(4), 1305-1315.

Bonafé, A., Galeati, G. and Sforna, F. (1994). Neural networks for daily mean flow forecasting. In: Blain, V.W.R. and Katsifarakis, K.L. (eds.) Hydrologic Engineering Woftware, Vol 1. Computational Mechanics Publications, Southampton, UK, pp 131-138.

Bowden, G.J., Maier, H.R. and Dandy, G.C. (2002) Optimal division of data for neural network models in water resources applications. *Water Resources Research*, 38(2), 2-1- 2-11.

Bowden, G.J., Dandy, G.C. and Maier, H.R. (2005a). Input determination for neural network models in water resources applications. Part 1 -- Background and methodology. *Journal of Hydrology*, 301, 75 92.

Bowden, G.J., Maier, H.R. and Dandy, G.C. (2005b). Input determination for neural network models in water resources ppplications. Part 2. Case study: Forecasting salinity in a river. *Journal of Hydrology*, 301, 93-107.

Box, G.E.P. and Jenkins, G.M. (1970). Time Series Analysis for Forecasting and Control. Holden Day, San Francisco, USA.

Brath, A., Montanari, A. and Toth, E. (2002). Neural networks and nonparametric methods for improving realtime flood forecasting through conceptual hydrological models. *Hydrology and Earth System Sciences*, 6 (4), 627-640.

Breuer, L., Huisman, J.A., Willems, P., Bormann, H., Bronstert, A., Croke, B.F.W, Frede, H., Gräff, T., Hubrechts, L., Jakeman, A.J., Kite, G.A., Lanini, J., Leavesley, G., Lettenmaier, D.P., Lindstrom, G., Seibert, J., Sivapalan, M. and Viney, N.R. (2009) Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use. *Advances in Water Resources*, 32, 129-146.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*, Second Ed., Springer, New York.

Buytaert, W. and Beven, K. (2011) Models as multiple working hypotheses:

hydrological simulation of tropical alpine wetlands. *Hydrological Processes*, 25(11), 1784-1799.

Calenda, G., Mancini, C.P. and Volpi, E. (2009). Selection of the probabilistic model of extreme floods: The case of the River Tiber in Rome. *Journal of Hydrology*, 371,1–11.

Calvo, B. and Savi, F. (2009). Real-time flood forecasting of the Tiber river in Rome. *Natural Hazards*, 50(3), 461-477.

Campolo, M., Andreussi, P. and Soldati, A. (1999). River flood forecasting with a neural network model. *Water Resources Research*, *35*, 1191-1197.

Cannas, B., Fanni, A., See, L. and Sias, G. (2006). Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. *Physics and Chemistry of the Earth*, 31, 1164-1171.

Centro Funzionale Regione Umbria (2009). *Evento 4-16 Dicembre 2008 – Rapporto Preliminare.*

Chaipimonplin T., See, L.M. and Kneale, P.E. (2010). Using radar data to extend the lead time of neural network forecasting on the River Ping. *Disaster Advances*, v.3(3), 33-45.

Chaipimonplin, T. (2010). *Improving Neural Network Flood Forecasting Models*. Unpublished PhD thesis. School of Geography, University of Leeds.

Chang, F.J. and Hwang, Y.Y. (1999). A self-organization algorithm for real-time flood forecast. *Hydrological Processes*, 13, 123-138.

Chang, F-J. and Chen, Y-C. (2001). A counterpropagation fuzzy-neural network modeling approach to real time streamflow prediction. *Journal of Hydrology*, 245, 153-164.

Chang, F-J., Chang, K-Y. and Chang, L-C. (2008). Counterpropagation fuzzy-neural network for city flood control system. *Journal of Hydrology* 358, 24-34.

Chang, F-J., Hu, H-F. and Chen, Y-C. (2001). Counterpropagation fuzzy-neural network for streamflow reconstruction. *Hydrological Processes*, 15, 219-232.

Chau, K.W., Wu, C.L. and Li, Y.S. (2005). Comparison of several flood forecasting models in Yangtze River. *Journal of Hydrologic Engineering*, 10(6), 485-491.

Chen, S-H., Lin, Y-H., Chang, L-C. and Chang, F-J. (2006). The strategy of building a flood forecast model by neuro-fuzzy method. *Hydrological Processes*, 20, 1525-1540.

Chen, Y.H. and Chang, F.J. (2009). Evolutionary artificial neural networks for hydrological systems forecasting. *Journal of Hydrology*, 367, 125-137.

Chidthong, Y., Tanaka, H. and Supharatid, S. (2009). Developing a hybrid multi-model for peak flood forecasting. *Hydrological Processes*, 23(12), 1725-1738.

Cigizoglu, H.K. (2003). Incorporation of ARMA models into flow forecasting by artificial neural networks. *Environmetrics*, 14(4), 417-427.

Cigizoglu, H.K. (2005). Generalized regression neural network in monthly flow forecasting. *Civil Engineering and Environmental Systems.* 22, 71-84.

Corani, G. and Guariso, G. (2005a). An application of pruning in the design of neural networks for real time flood forecasting. *Neural Computing and Applications,* 14, 66-77.

Corani, G. and Guariso, G. (2005b). Coupling fuzzy modeling and neural networks for river flood prediction. *IEEE Transactions on Systems, Man and Cybernetics-PartC: Applications and Reviews*, 35(3), 382-390.

Corne, S., Murray, T., Openshaw, S., See, L. and Turton, I. (1999). Using artificial intelligence techniques to model sub-glacial water systems. *Journal of Geographical Systems*, 1, 37-60.

Corradini, C., Morbidelli, R., Saltalippi, C. and Melone, F. (2004). Flood forecasting and infiltration modelling. *Hydrological Sciences Journal*, 49(2), 227- 236.

Corzo, G.A., Solomatine, D.P. Hidayat, de Wit, M.,Werner, M., Uhlenbrook, S. and Price, R.K. (2009). Combining semi-distributed process-based and data-driven models in flow simulation: a case study of the Meuse river basin. *Hydrology and Earth System Sciences*, 13, 1619-1634.

Coughlin, K.T. and Tung, K.K. (2004). 11-Year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Advances in Space Research*, 34, 323-329.

Coulibaly, P. and Baldwin, C.K. (2008). Dynamic neural networks for nonstationary hydrological time series modeling. In: Abrahart, R.J., See, L.M., Solomatine, D.P. (Eds.) *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, pp. 71-85. Springer-Verlag, Heidelberg.

Coulibaly, P., Haché, M., Fortin, V. and Bobée, B. (2005). Improving daily reservoir inflow forecasts with model combination. *Journal of Hydrologic Engineering*, 10(2), 92-99.

Cox, D.R. and Stuart, A. (1955). Some quick tests for trend and dispersion. *Biometrika*, 42, 80-95.

Crockett, R.G.M. and Gillmore, G.K. (2010). Spectral-decomposition techniques for the identification of radon anomalies temporally associated with earthquakes occurring in the UK in 2002 and 2008. *Natural Hazards and Earth System Sciences*, 10, 1079-1084.

Cunge, J., Holly, F.M. and Verwey, A. (1980). *Practical Aspects of Computational River Hydraulics*. Boston MA: Pittman Publishers.

Dastorani, M.T., Moghadamnia, A., Piri J. and Rico-Ramirez, M. (2009). Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment*, 166, 421-434.

Dawson C.W., Abrahart R.J., Shamseldin A.Y. and Wilby R.L., (2006b). Flood Estimation at ungauged sites using artificial neural networks. *Journal of Hydrology*, 319, 391-409.

Dawson, C.W. and Wilby, R.L. (1998). An artificial neural network approach to rainfall runoff modelling. *Hydrological Sciences Journal*, *43*, 47-66.

Dawson, C.W. and Wilby, R.L. (1999). A comparison of artificial neural networks used for river flow forecasting. *Hydrology and Earth System Sciences*, 3, 529-540.

Dawson, C.W., Brown, M.R. and Wilby, R.L. (2000). Inductive learning approaches to rainfall-runoff modelling. *International Journal of Neural Systems,* 10, 43-57.

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

Dawson, C.W., Abrahart, R.J. and See, L. (2007). HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling and Software*, 22, 1034-1052.

Dawson, C.W., Abrahart, R.J., Shamseldin, A.Y. and Wilby, R.L. (2006a). Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology,* 319, 391-409.

Dawson, C.W., See, L., Abrahart, R.J. and Heppenstall, A.J. (2006b). Symbiotic adaptive neuro-evolution applied to rainfall-runoff modelling in northern England. *Neural Networks*, 19(2), 236-247.

Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253-263.

Doan, C.D. and Liong, S.-Y. (2004). Generalization for multilayer neural network bayesian regularization or early stopping. In: Proceedings of Asia Pacific Association of Hydrology and Water Resources 2nd Conference. Singapore.

Dooge, J.C.I. (1959). A general theory of the unit hydrograph. *Journal of Geophysical Research,* 64(2), 241-256.

El-Shafie, A., Taha, M.R. and Noureldin, A. (2007). A neuro-fuzzy model for inflow forecasting of the Nile river at Aswan high dam. *Water Resources Management*, 21, 533-556.

Fahlman, S. E. and C. Lebiere (1990). The cascade-correlation learning architecture. In Touretzky, D.S. (ed.) *Advances in Neural Information Processing Systems 2*. Morgan-Kaufmann, Los Altos CA.

Fauchereau, N., Pegram, G.G.S. and Sinclair, S. (2008). Empirical Mode Decomposition on the sphere: application to the spatial scales of surface

temperature variations. *Hydrology and Earth System Sciences*, 12, 933-941.

Firat, M. (2008). Comparison of Artificial Intelligence Techniques for river flow forecasting. *Hydrology and Earth System Sciences*, 12, 123-139.

Firat, M. and Güngör, M. (2007). River flow estimation using adaptive neuro fuzzy inference system. *Mathematics and Computers in Simulation*, 75, 87-96.

Firat, M. and Güngor, M. (2008). Hydrological time-series modelling using an adaptive neuro-fuzzy inference system. *Hydrological Processes*, 22, 2122-2132.

Flandrin, P., Rilling, G. and Goncalves, P. (2004). Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11 (2), 112-114.

Fletcher, R. and Reeves, C. (1964). Function minimization by conjugate gradients. *Computer Journal*, 7, 81-84.

Foresee, F.D. and Hagan, M.T. (1997). Gauss-Newton approximation to Bayesian regularization. In: Proceedings of the 1997 International Joint Conference on Neural Networks, pp.1930-1935.

Franceschini, S. and Tsai, C.W. (2010). Application of Hilbert-Huang transform method for analyzing toxic concentrations in the Niagara River. *Journal of Hydrologic Engineering*, 15(2), 90-96.

Franzke, C. (2009). Multi-scale analysis of teleconnection indices: climate noise and nonlinear trend analysis. *Nonlinear Processes in Geophysics*, 16, 65-76.

Furundzic, D. (1998). Application example of neural networks for time series analysis: rainfall runoff modelling. *Signal Processing*, 64, 383-396.

Gallant, S. (1993). *Neural Network Learning and Expert Systems*. MIT, Massachusetts.

Gallart, F., Latron, J., Llorens, P. and Beven, K.J. (2008). Upscaling discrete internal observations for obtaining catchment-averaged TOPMODEL parameters in a small Mediterranean mountain basin. *Physics and Chemistry of the Earth*, 33(17-18), 1090-1094.

Garbrecht Jurgen D., 2006. Comparison of three alternative ANN designs for monthly rainfall-runoff simulation. *Journal of Hydrologic Engineering 11(5), 502-505.*

Gautam, D.K. and Holz, K.P. (2001). Rainfall-runoff modelling using adaptive neuro-fuzzy systems. *Journal of Hydroinformatics*, 3, 3-10.

Giandotti, M. (1934). Previsione delle piene e delle magre neu corsi d'acqua, Ministero LL.PP., Servizio Idrografico Italiano, Memorie e studi idrografici, 8(2).

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

Giustolisi, O. and Simeone, V. (2006). Multi-Objective strategy in artificial neural network construction. *Hydrological Sciences Journal*, 51(3), 502-523.

Golob, R., Stokelj, T. and Grgic, D. (1998). Neural-network-based water inflow forecasting. *Control Engineering Practice*, 6, 593-600.

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H. and J. Schmidhuber, J. (2009). A novel connectionist system for improved unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5), 855-868.

Green, I.R.A. and Stephenson, D. (1986). Criteria for comparison of single event models. *Hydrological Sciences Journal*, 31(3), 395-411.

Gupta, H.V., Sorooshian, S. and Yapo, P.O. (1998). Toward improved calibration of hydrological models: multiple and noncommensurable measures of information. *Water Resources Research*, 34(4), 999-1018.

Hagan, M. and Menhaj, M. (1994). Training feedforward networks with the Marquardt Algorithm. *IEEE Transactions on Neural Networks*, 5(6), 989-993.

Han, D., Kwong, T. and Li, S. (2007). Uncertainties in real-time flood forecasting with neural networks, *Hydrological Processes*, 21, 223-228.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. NY, Macmillan.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, 2nd Edition. Prentice Hall.

Hecht-Nielsen, R., (1987). Counterpropagation networks. *Applied Optics*, 26(23), 4979-4984.

Hejazi, M.I. and X. Cai (2009). Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. *Advances in Water Resources*, 32, 582-593.

Heppenstall, A.J., See, L.M., Abrahart, R.J. and Dawson, C.W. (2008). Neural Network Hydrological Modelling: An Evolutionary Approach. In: Abrahart, R.J., See, L. and Abrahart R.J. (eds.) *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, pp.321-332. Springer-Verlag, Heidelberg.

Hirschen, K. and Schafer, M. (2006). Bayesian regularization neural network for optimizing fluid flow processes. *Comput. Methods Appl. Mech. Engrg.,* 195, 481-500.

Hong, Y-S.T and White, P.A. (2009) Hydrological modeling using a dynamic neuro-fuzzy system with on-line and local learning algorithm, *Advances in Water Resources*, 32(1), 110-119.

Hornik, K., Stinchcombe, M. and H. White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366.

Hsu K.L., Gao, X.G. and Sorooshian, S. (1995). Artificial neural network modelling of the rainfall runoff process. *Water Resources Research*, 31(10), 2517-2530.

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

http://www.mathworks.com/products/fuzzylogic/

Hu, T.S., Lam, K.C. and Ng, T.G. (2001). River flow forecasting with a range dependent network. *Hydrological Science Journal*, 45(5), 729-745.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N., Tung, C.C. and Liu, H.H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A*, 454(1971), 903-995.

Huang, N.E. and Attoh-Okine, N.O. (2005). *The Hilbert-Huang Transform in Engineering*. Taylor & Francis.

Huang, N.E. and Shen, S.S. (2005). The Hilbert-Huang Transform and its Applications. World Scientific Publishing, Singapore.

Huang, N.E. and Wu, Z. (2008). A review on Hilbert-Huang transform: Method and its applications to geophysical studies. *Reviews of Geophysics*, 46, RG2006.

Huang, N.E., Wu, M.-L., Qu, W., Long, S.R., Shen, S.S.P. and Zhang, J.E. (2003). Applications of Hilbert-Huang transform to non-stationary financial time series analysis. *Applied Stochastic Models in Business and Industry*, 19, 245-268.

Huang, Y., Schmitt, F.G., Lu, Z. and Liu, Y. (2008). An amplitude-frequency study of turbulent scaling intermittency using Hilbert spectral analysis. *Europhysics Letters*, 84, 40010.

Huang, Y., Schmitt, F.G., Lu, Z. and Liu, Y. (2009). Analysis of daily river flow fluctuations using empirical mode decomposition and arbitrary order Hilbert spectral analysis. *Journal of Hydrology*, 373, 103-111.

Hui, S. and Xinxia, L. (2010). Multi-scale rbf prediction model of runoff based on emd method. In: *2010 Third International Conference on Information and Computing (ICIC)*, Vol. 3. pp. 296-299.

Hung, N.Q., Babel, M.S., Weesakul, S. and Tripathi, N.K. (2009). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. *Hydrology and Earth System Sciences*, 13, 1413-1425.

Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, 679-688.

Imrie, C.E., Durucan, S. and Korre, A. (2000). River flow prediction using artificial neural networks: generalisation beyond the calibration range. *Journal of Hydrology, 233, 138-153.*

Jachner, S., van den Boogaart, K.G. and Petzoldt, T. (2007). Statistical methods for the qualitative assessment of dynamic models with time delay. *Journal of Statistical Software*, 22 (8), 1-30.

Jackson, L.P. and Mound, J.E. (2010). Geomagnetic variation on decadal time scales: What can we learn from Empirical Mode Decomposition? *Geophysical Research Letters*, 37, L14307.

Jain A. (2005). Comment on 'Comparison of static-feedforward and dynamic-feedback neural networks for rainfall-runoff modeling' by Chiang Y.M., Chang L.C. and Chang F.J., 2004. Journal of Hydrology 290, 297–311. *Journal of Hydrology 314, 207–211.*

Jain, A. and Indurthy, S.K.V.P. (2003). Comparative analysis of event based rainfall-runoff modeling techniques - deterministic, statistical, and artificial neural networks. *Journal of Hydrological Engineering*, 8(2), 93-98.

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

Jain, A. and Srinivasulu, S. (2004a). Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms, and artificial neural network techniques. *Water Recourses Research*, 40(4), W04302. doi: 10.1029/2003WR002355.

Jain, A., Sudheer, K.P. and Srinivasulu, S., (2004b). Identification of physical processes inherent in artificial neural network rainfall runoff models. *Hydrological Processes*, 18(3), 571-581.

Jang, J-S.R. (1993). ANFIS: Adaptive-Network-based Fuzzy Inference System. *IEEE Transactions on Systems, Man and Cybernetics*, 23, 665-685.

Jang, J-S.R., Sun, C.-T. and Mizutani, E. (1997). *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. Prentice-Hall, Upper Saddle River, NJ.

Jia, Y., Z. Hongli, N. Cunwen, J. Yunzhong, G. Hong, X. Zhi, Z. Xueli, and Z. Zhixin (2009). A WebGIS-based System for Rainfall-Runoff Prediction and Real-Time Water Resources Assessment for Beijing. *Computers & Geosciences*, 35, 1517-1528.

Kamp, R. and Savenije, H.H.G. (2006). Optimising training data for ANNs with genetic algorithms. *Hydrology and Earth System Sciences*, 10, 603-608.

Karunanithi, N., Grenney, W. J., Member, Asce, Whitley, D. and Bovee, K. (1994). Neural networks for river flow prediction. *Journal of Computing in Civil Engineering*, 8, 201-220.

Karunanithi, N., Grenney, W.J., Whitley, D. and Bovee, K. (1994). Neural networks for river flow prediction. *Journal of Computing in Civil Engineering*, 8(2), 201-220.

Kataoka, R., Miyoshi, Y. and Morioka, A. (2009). Hilbert-Huang Transform of geomagnetic pulsations at auroral expansion onset. *Journal of Geophysical Research*, 114, A09202.

Kavetski, D., Kuczera, G. and Franks, S.W. (2006). Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. *Journal of Hydrology*, *320*, 173-186.

Kendall M.G. (1975). *Rank Correlation Methods*. London UK: Charles Griffin.

Kerh, T. and Lee, C.S. (2006). Neural networks forecasting of flood discharge at an unmeasured station using river upstream information. *Advances in Engineering Software*, 37, 533-543.

Keskin, M.E. and Taylan, D. (2009) Artificial models for interbasin flow prediction in southern Turkey, *Journal of Hydrologic Engineering*, 14(7), 752-758.

Khan, M.S. and Coulibaly, P. (2006). Bayesian neural network for rainfall-runoff modeling. *Water Resources Research*, 42(7), W07409, DOI: 10.1029/2005WR003971.

Kijewski-Correa, T. and Kareem, A. (2007). Using multi-objective genetic algorithm for svm construction. *Journal of Engineering Mechanics*, 133(2), 238-245.

Kim, D. and Oh, H.-S. (2008). EMD: Empirical Mode Decomposition and Hilbert Spectral Analysis. R package version 1.2.0. URL http://dasan.sejong.ac.kr/~dhkim/software emd.html

Kim, S., Kim, J.-H. and Park, K.-B. (2009). Neural networks models for the flood forecasting and disaster prevention systems in the small catchment. *Disaster Advances*, 2, 51-63.

Kingston, G.B., Maier, H.R. and Lambert, M.F. (2006). A probabilistic method to assist knowledge extraction from artificial neural networks used for hydrological prediction. *Mathematical and Computer Modelling*, 44(5-6), 499-512.

Kisi, Ö. (2004). River flow modelling using artificial neural networks. *Journal of Hydrologic Engineering*, 9(1), 60-63.

Kisi, Ö. (2008a). River flow forecasting and estimation using different artificial neural network techniques, *Hydrology Research* 39(1), 27-40.

Kişi, Ö. (2008b). Stream flow forecasting using neuro-wavelet technique. *Hydrological Processes*, 22, 4142-4152.

Kişi, Ö. (2009). Neural networks and wavelet conjunction model for intermittent streamflow forecasting. *ASCE Journal of Hydrological Engineering*, 14, 773-782.

Kişi, Ö. (2010). Wavelet regression model for short-term streamflow forecasting. *Journal of Hydrology*, 389, 344-353.

Kisi, Ö. and Cigizoglu, H.K. (2007) Comparison of different ANN techniques in river flow prediction, *Civil Engineering and Environmental Systems*, 24(3), 211-231.

Kitanidis, P.K. and Bras, R.L. (1980). Real-time forecasting with a conceptual hydrologic model: 2. Application and results. *Water Resources Research*, 16(6), 1034-1044.

Kitano, H. (1992) Neurogenetic learning: an integrated method of designing and training neural networks using genetic algorithms. Center for Machine Translation, Carnegie Mellon University.

Klemes, V. (1973). Watershed as semi infinite storage reservoir. *ASCE Journal Irrigation and Drain. Div. 99 IR4, 477-491.*

Koenker, R. and Bassett Jr., G. (1978). Regression quantiles. *Econometrica*, 46(1), 33-50.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Springer, Berlin.

Kottegoda, N.T., Natale, L., and Raiteri, E. (2004). Some considerations of periodicity and persistence in daily rainfalls. *Journal of Hydrology*, 296, 23-37.

Kouwen, N. (1988). WATFLOOD: A micro-computer based flood forecasting system based on real-time weather radar. *Canadian Water Resources Journal*, 13(1):62-77.

Kumar, A. and Minocha, V.K. (2001). Discussion on “Rainfall runoff modelling using artificial neural networks” by Tokar and Johnson, 1999. *Journal of Hydrologic Engineering 6(2)*, 176-177.

Kumar, A.R.S., Sudheer, K.P., Jain, S.K. and Agarwal, P.K. (2005). Rainfall-runoff modelling using artificial neural networks: comparison of network types. *Hydrological Processes,* 19, 1277-1291.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11, 1267-1277.

Leahy, P., Kiely, G. and Corcoran, G. (2008). Structural optimisation and input selection of an artificial neural network for river level prediction. *Journal of Hydrology*, 355, 192-201.

Lee, T. and Ouarda, T.B.M.J., (2011a). An EMD and PCA hybrid approach for separating noise from signal, and signal in climate change detection. *International Journal of Climatology*, doi:10.1002/joc.2299.

Lee, T. and Ouarda, T.B.M.J., (2011b). Prediction of climate nonstationary oscillation processes with empirical mode decomposition. *Journal of Geophysical Research*, 116, D06107.

Lee, T., Ouarda and T.B.M.J. (2010). Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. *Journal of*

Geophysical Research, 115, D13107.

Legates, D.R. and McCabe, G.J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1), 233-241.

Leggett, R.W. and Williams, L.R. (1981). A reliability index for models. *Ecological Modelling*, 13, 303-312.

Lehmann, E.L. (1975). Nonparametrics, Statistical Methods Based on Ranks. McGraw-Hill, San Francisco.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly Applied Mathematics*, 2, 164-168.

Li, Z., Deng, P. and Dong, J. (2009). Application of artificial neural network in rainfall-runoff model. In *Proceedings of the International Conference on Hydrological Changes and Management from Headwaters to the Ocean*, CRC Press, Kyoto, Japan.

Lin, G.F. and Lee, F.C. (1994). Assessment of aggregated hydrologic time-series modeling. *Journal of Hydrology*, 156(1-4), 447-458.

Liong, S.-Y., Lim, W.-H. and Paudyal, G.N., (2000). River stage forecasting in Bangladesh: neural network approach. *Journal of Computing in Civil Engineering*, 14(1), 1-8.

Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-303.

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

Macdonald, A. and See, L. (2010). *A Review of Artificial Neural Networks*. Environment Agency, UK.

MacKay, D.J.C. (1992). Bayesian interpolation. *Neural Computation*, 4(3), 415- 447.

MacKay, D.J.C. (1992a). A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3), 448-472.

Maier, R.H. and Dandy, G.C. (1998a). The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. *Environmental Modelling & Software*, 13, 193-209.

Maier, R.H. and Dandy, G.C. (1998b). Understanding the behaviour and optimising the performance of back-propagation neural networks: an empirical study. *Environmental Modelling & Software*, 13, 179-191.

Maier H.R. and Dandy G.C. (1999). Empirical comparison of various methods for training feedforward neural networks for salinity forecasting. *Water Resources Research*, 35(8), 2591-2596.

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

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

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). *Forecasting: methods and applications*, 3rd Edition. Wiley.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal Applied Mathematics*, 11, 431-441.

Mason, J.C., Price, R.K., and Tem'me, A. (1996). A neural network model of rainfall-runoff using radial basis functions, *Journal of Hydraulic Research*, 34(4), 537-548.

Mathworks (1994-2011). Fuzzy Logic Toolbox.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity*. Bulletin of Mathematical Biophysics*, 7, 115-133.

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

Minsky, M. and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge.

Montanari, A. (2011). Uncertainty of hydrological predictions. In Wilderer, P.A. (Ed.). *Treatise on Water Science*, Elsevier.

Montanari, A. and Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, 40, W01106.

Montanari, A. and Grossi, G. (2008). Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resources Research*, 44, W00B08 doi:10.1029/ 2008WR006897.

Moore, G.A. (1991). *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers*, HarperCollins Publishers, New York.

Moriarty, D. E., & Miikkulainen, R. (1998). Forming neural networks through efficient and adaptive coevolution. Evolutionary Computation, 5, 373–399.

Moussa, R. (2010). When monstrosity can be beautiful while normality can be ugly: assessing the performance of event-based flood models. *Hydrological Sciences Journal*, 55, 1074-1084.

Mukerji, A., Chatterjee, C. and Raghuwanshi, N.S. (2009). Flood forecasting using ANN, Neuro-Fuzzy and Neuro-GA models. *Journal of Hydrologic Engineering*, 14, 647-652.

Napolitano, G., See, L.M., Calvo, B., Savi, F. and Heppenstall, A.J. (2009). A conceptual and neural network model for real-time flood forecasting of the Tiber

River in Rome. *Physics and Chemistry of the Earth*, 35(3-5), 187-194.

Napolitano, G., Serinaldi, F. and See, L. (2011). Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: an empirical examination. *Journal of Hydrology*, 406(3-4), 199-214.

Nash, J. and Sutcliffe, J. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10, 282-290.

Nason, G.P. and Von Sachs, R. (1999). Wavelets in time series analysis. *Phil Trans Roy Soc*, 357, 2511-2526.

Natale, L. and Savi, F. (2006). Structural measures to protect Rome from floods along Tiber River. In: *Proceedings of the 17th IASTED International Conference on Modelling and Simulation*, 557-560.

Natale, L. and Savi, F. (2007). Monte Carlo analysis of probability of inundation of Rome. *Environmental Modelling and Software*, 22(10), 1409-1416.

Nayak, P.C., Sudheer, K.P. and Jain, S.K. (2007). Rainfall-runoff modeling through hybrid intelligent system. *Water Resources Research*, 43, W07415. doi:10.1029/2006WR004930.

Nayak, P.C., Sudheer, K.P., Rangan, D.M. and Ramasastri, K.S. (2004). A neuro-fuzzy computing technique for modeling hydrological time series. *Journal of Hydrology*, 291, 52-66.

Nayak, P.C., Sudheer, K.P., Rangan, D.M. and Ramasastri, K.S. (2005). Short-term flood forecasting with a neurofuzzy model. *Water Resources Research*, 41, W04004. doi:10.1029/2004WR003562

Nayebi, M., Khalili, D.Amin, S. and Zand-Parsa, S. (2006). Daily stream flow prediction capability of artificial neural networks as influenced by minimum air temperature data. *Biosystems Engineering*, 95(4), 557-567.

Nicklow, J., Reed, P., Savic, D, Dessalegne, T., Harrell, L., Minsker, B. Ostfeld, A., Singh, A. and Zechman, E. (2010). State of the art for Genetic Algorithms and beyond in water resources planning and nanagement. *Journal of Water Resources Planning and Management-ASCE*, 136(4), 412-432.

Nie, J. and Linkens, D.A. (1994). Fast self-learning multivariable fuzzy controllers constructed from a modified CPN network. *International Journal of Control*, 60, 369-393.

Nigrin, A. (1993). *Neural Networks for Pattern Recognition*. Cambridge, MA, The MIT Press.

Nourani, V., Komasi, M. and Mano, A. (2009). A multivariate ANN-wavelet approach for rainfall-runoff modelling. *Water Resources Management*, 23, 2877-2894.

O'Connor K. M., 1976. A discrete linear cascade model for hydrology. *Journal of hydrology 29, 203-242.*

Olsson, J., Uvo, C.B., Jinno, K., Kawamura, A., Nishiyama, K., Koreeda, N., Nakashima, T. and Morita, O. (2004). Neural networks for rainfall forecasting by atmospheric downscaling. *Journal of Hydrologic Engineering,* 9(1), 1-12.

Palmieri, S., Bencivenga, M., Bersani, P., Siani, A.M. and Casale, G.R. (2001). Hydrometeorological aspects of Tiber basin storms. In: *Proceedings of the 3rd Plinius Conference on Mediterranean storms*. Baja Sardinia.

Pan, T.Y. and Wang, R.Y. (2005). Using recurrent neural networks to reconstruct rainfall-runoff processes. *Hydrological Processes*, 19(18), 3603-3619.

Panchal, G., Ganatra, A., Kosta, Y.P. and Panchal, D. (2010). Searching most efficient neural network architecture using Akaike's Information Criterion (AIC). *International Journal of Computer Applications*, 1(5), 41-44.

Parasuraman, K. and Elshorbagy, A. (2007) Cluster-based hydrologic prediction using genetic algorithm-trained neural networks. Journal of Hydrologic Engineering, 12(1), 52-62.

Partal, T. (2009). River flow forecasting using different artificial neural network algorithms and wavelet transform. *Canadian Journal of Civil Engineering*, 36, 26-39.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Phil Trans R Soc Lond Series A,* 187*,* 253-318.

Pegram, G.G.S., Peel, M.C. and McMahon, T.A. (2008). Empirical mode decomposition using rational splines: an application to rainfall time series. *Proceedings of the Royal Society of London A*, 464 (2094), 1483-1501.

Peng, D., Zhijia, L. and Fan, X. (2009). Application of TOPMODEL in Buliu River catchment, Pearl River basin and comparison with Xin'anjiang model. *Hupo Kexue*, 21(3), 441-444.

Phien, H.N. and Kha, N.D.A. (2003). Flood forecasting for the upper reach of the Red River Basin, North Vietnam. *Water SA*, 29(3), 267-272.

Plackett, R.L. (1983). Karl Pearson and the Chi-Squared Test. *International Statistical Review (International Statistical Institute (ISI))*, 51(1), 59–72.

Potter, M.A. (1997). The design and analysis of a computational model of cooperative coevolution. Unpublished PhD thesis: George Mason University.

Pramanik N. and Panda R.K. (2009) Application of neural network and adaptive neuro-fuzzy inference systems for river flow prediction, *Hydrological Sciences Journal*, 54(2), 247–260.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL http://www.R-project.org

Rahnama, M.B. and Noury, M. (2008). Developing of Halil River rainfall-runoff model, using conjunction of wavelet transform and artificial neural networks. *Research Journal of Environmental Sciences*, 2, 385-292.

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

Rajurkar, M.P., Kothyari, U.C. and Chaube, U.C. (2004). Modelling of the daily rainfall-runoff relationship with artificial neural network. *Journal of Hydrology 285, 96–113.*

Raman, H. and Sunilkumar, N. (1995). Multivariate modelling of water resources time series using artificial neural networks. *Hydrological Sciences Journal*, 40, 145-163.

Rao, Y.R.S and Krishna, B. (2009). Modelling hydrological time series data using wavelet neural network analysis. *New Approaches to Hydrological Prediction in Data-sparse Regions*. IAHS Publ. 333, 101-111.

Reilly D.L. and Cooper L.N. (1990). An overview of neural networks: early models to real world systems. Zornetzer, S.F., Davis, J.L., Lau, C. (eds.) *An Introduction to Neural and Electronic Networks.* Academic Press, New York, 227–248.

Remedia, G., Alessandroni, M.G. and Mangianti, F. (1998). Le piene eccezionali del fiume Tevere a Roma Ripetta. *Università degli Studi di L'Aquila, Dip. di Ingegneria delle Strutture, delle Acque e del Terreno (DISAT n. 3).*

Remesan, R., Shamim, M.A., Han, D. and Mathew, J. (2009). Runoff prediction using an intergrated hybrid modelling scheme. *Journal of Hydrology,* 372, 48-60.

Reusser, D. E., Blume, T., Schaefli, B. and Zehe, E. (2009). Analysing the temporal dynamics of model performance for hydrological models. *Hydrology and Earth System Sciences*, 13, 999-1018.

Riad, S., Mania, J., Bouchaou, L. and Najjar, Y. (2004). Rainfall-runoff model using an artificial neural network approach. *Mathematical and Computer Modelling*, 40(7-8), 839-846.

Rogers, L.L. and Dowla, F.U. (1994). Optimal groundwater remediation using artificial neural networks with parallel solute transport. *Water Resources Research,* 30(2), 458-481.

Rumelhart, D.E., Hinton, G.E. and Willliams, R.J. (1986). Learning internal representations by error propagation. In Rumelhart, D.E. and McClelland, J.L. (eds.) *Parallel Distributing Processing: Explorations in the Microstructure of Cognition*, vol.1, Cambridge, MA, MIT Press.

Russell, S. and Norvig, P. (1995). *Artficial Intelligence: A Modern Approach*. Prentice-Hall, Saddle River, NJ.

Sahoo, G.B. and Ray, C. (2006). Flow forecasting for a Hawaiian stream using rating curves and neural networks. *Journal of Hydrology*, 317, 63-80.

Sahoo, G.B., Ray, C. and Carlo, E.H.D. (2006). Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. *Journal of Hydrology*, 327, 525-538.

Sajikumar, N. and Thandaveswara, B.S. (1999). A non-linear rainfall-runoff model using an artificial neural network. *Journal of Hydrology*, 216, 32-55.

Salas, J.D. and Obeysekera, J.T.B. (1982). ARMA model identification of hydrologic time-series. *Water Resources Research*, 18(4), 1011-1021.

Schaefli, B. and Gupta, H.V. (2007). Do Nash values have value? *Hydrological Processes*, 21, 2075-2080.

Schalkoff, R. (1997). *Artificial Neural Networks*, McGraw-Hill.

Schmidhuber, J. (1989). A local learning algorithm for dynamic feedforward and recurrent networks. *Connection Science*, 1(4), 403-412.

Sedki, A., Ouazar, D. and El Mazoudi, E. (2009). Evolving neural network using real coded genetic algorithm for daily rainfall-runoff forecasting. *Expert Systems with Applications*, 36(3), 4523-4527.

See L. and Openshaw S. (1999). Applying soft computing approaches to river level forecasting. *Hydrological Science Journal*, 44(5), 763-778.

See, L. and Abrahart, R.J. (2001). Multi-model data fusion for hydrological forecasting. *Computers & Geosciences*, 27, 987-994.

See, L. and Openshaw, S. (2000) A hybrid multi-model approach to river level forecasting. *Hydrological Sciences Journal*, 45, 523-536.

See, L., Jain, A., Dawson, C.W. and Abrahart, R.J. (2008). Visualisation of Hidden Neuron Behaviour in a Neural Network Rainfall-Runoff Model. In: Abrahart, R.J., See, L. and Abrahart R.J. (eds.) *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, pp.87-99. Springer-Verlag, Heidelberg.

Shamseldin, A.Y. (1997). Application of a neural network technique to rainfall-runoff modelling. *Journal of Hydrology*, 199, 272-294.

Shamseldin, A.Y., and O'Connor, K.M. (2001). *A Non-linear Neural Network Technique for Updating of River Flow Forecasts*. *Hydrology and Earth System Sciences*, 5(4), 577-597.

Shamseldin, A.Y., O'Connor, K.M., and Nasr, A.E. (2007). A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models. *Hydrological Sciences Journal*, 52, 896–916.

Sharkey, A.J.C. (1999). Multi-net systems. In: Sharkey, A.J.C (Ed.) *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, pp. 3-30. Springer-Verlag.

Sharma, A. (2000). Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1-A strategy for system predictor identification. *Journal of Hydrology*, 239, 232-239.

Sharma, A., Luk, K.C., Cordery, I. and Lall, U. (2000). Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2-Predictor identification of quarterly rainfall using ocean-atmosphere information. *Journal of Hydrology*, 239, 240-248.

Sharma, D.K., Gaur, L. and Okunbor, D. (2007). Image compression and feature extraction using Kohonen's self-organizing map neural network. *Journal of Strategic E-Commerce*. 5(1-2). http://news-business.vlex.com/vid/compression-extraction-kohonen-neural-55414661.

Siebert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, 15, 1063-1064.

Sugiura, N. (1978). Further analysis of data by Akaike information criterion and finite corrections. *Communications in Statistics Part A - Theory and Methods*, 7(1), 13-26.

Simpson, P.K. (1990). *Neural Systems: Foundations Paradigms, Applications and Implementations*. Pergamon Press.

Sinclair, S. and Pegram, G.G.S. (2005). Empirical Mode Decomposition in 2-d space and time: a tool for space-time rainfall analysis and nowcasting. *Hydrology and Earth System Sciences*, 9, 127-137.

Singh, P. and Deo, M.C. (2007) Suitability of different neural networks in daily flow forecasting. *Applied Soft Computing*, 7(3), 968-978.

Smith, J. and Eli, R.N. (1995). Neural-network models of rainfall-runoff process. *Journal of Water Resources Planning and Management*, 121, 499-508.

Smith, M.B., Georgakakos, K.P. and Liang, X. (2004). The distributed model intercomparison project (DIMP). *Journal of Hydrology*, 298, 1-3.

Soil Conservation Service (SCS) (1985). *National Engineering Handbook, Section 4 - Hydrology*. Littleton, Colorado: Water Resources Publication.

Solé, J., Turiel, A., Estrada, M., Llebot, C., Blasco, D., Camp, J., Delgado, M., Fernandez-Tejedor, M. and Diogene, J. (2009). Climatic forcing on hydrography of a Mediterranean bay (Alfacs Bay). *Continental Shelf Research*, 29, 1786-1800.

Solomatine, D.P. (2008). Combining machine learning and domain knowledge in modular modeling. In: Abrahart, R.J., See, L. and Abrahart R.J. (eds) *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, Springer-Verlag, Heidelberg.

Srivastav, R.K., Sudheer, K.P. and Chaubey, I. (2007). A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network

hydrologic models. *Water Resources Research*, 43, 1-12.

Sudheer, K.P. (2005). Knowledge extraction from trained neural network river flow models. *Journal of Hydrologic Engineering*, 10(4), 264-269.

Sudheer, K.P. and Jain, A. (2004). Explaining the internal behaviour of artificial neural network river flow models. *Hydrological Process*, 118(4), 833-844.

Tayfur, G. and Moramarco, T. (2007). Forecasting flood hydrographs at Tiber River basin. International Congress on River Basin Management. Antalya (Turkey).

Thirumalaiah, K. and Deo, M. C. (1998a) River stage forecasting using artificial neural network. *Journal of Hydrologic Engineering*, 13, 101-11.

Thirumalaih, K. and Deo, M.C. (1998b). Real time flood forecasting using neural networks. *Computer Aided Civil and Infrastructure Engineering*, 13(2), 101-111.

Tingsanchali, T. and Gautam, M.R. (2000). Application of Tank, Nam, Arma and neural network models to flood forecasting, *Journal of Hydrological Processes*, *14*, 2473-2487.

Tokar S. and Markus, M. (2000). Precipitation-runoff modelling using artificial neural networks and conceptual models. *Journal of Hydrologic Engineering*, 5(2), 156-161.

Toth, E. (2009). Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting. *Hydrolology Earth System Sciences*, 13, 1555–1566.

Toth, E. and Brath, A. (2002). Flood forecasting using artificial neural networks in black-box and conceptual rainfall-runoff modelling. The International Environmental Modelling and Software Society.

Velásquez, D., Dyner, I. and Souza, R. (2006). Tendencias en la predicción y estimación de los intervalos de confianza usando modelos de redes neuronales aplicados a series temporales. *Dyna, Año*, 73(149), 141-147.

Venables, W.N. and Ripley, B.D. (2002). Modern Applied Statistics with S, 4th Edition. Springer, New York, ISBN 0-387-95457-0. URL http://www.stats.ox.ac.uk/pub/MASS4.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2), 111-126.

Villarini, G., Serinaldi, F. and Krajewski, W.F. (2008). Modeling radar-rainfall estimation uncertainties using parametric and non-parametric approaches. *Advances in Water Resources*, 31, 1674-1686.

Villarini, G., Serinaldi, F., Smith, J.A. and Krajewski, W.F. (2009). On the stationarity of annual flood peaks in the continental United States during the 20th century. *Water Resources Research*, 45, W08417.

Vincendon, B., Ducrocq, V., Saulnier, G.-M., (2011). Benefit of coupling the ISBA land

surface model with a TOPMODEL hydrological model version dedicated to Mediterranean flash-floods. *Journal of Hydrology*, 394(1-2), 256-266.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathemathical Statistics*, 11, 147-162.

Wang, Q.J. (1991). The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resources Research*, 27(9), 2467-2471.

Wang, Q.J. (1997). Using genetic algorithms to optimise model parameters, *Environmental Modelling & Software*, 12(1), 27-34.

Wang, W., Jin, J. and Li, Y. (2009) Prediction of inflow at three gorges dam in Yangtze River with wavelet network model. *Water Resources Management*, 23, 2791-2803.

Wang, W., Van Gelder, P.H., Vrijling, J.K. and Ma, J. (2006). Forecasting daily streamflow using hybrid ANN models. *Journal of Hydrology*, 324, 383-399.

Wang, W-C., Chau, K-W., Cheng, C-T and Qiu, L. (2009). A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series, *Journal of Hydrology*, 374(3-4), 294-306.

Wheater, H.S., Jakeman, A.J. and Beven, K.J. (1993). Progress and directions in rainfall-runoff modelling. In *Modelling Change in Environmental Systems*, Jakeman, A.J., Beck, M.B. and McAleer, M.J. (eds.) pp.101-132, Wiley, Chichester.

Whitley, D., Starkweather, T., & Bogart, C. (1990). Genetic algorithms and neural networks: Optimizing connections and connectivity. Parallel Computing, 14, 347-361.

Wilby, R. L. (1997). Contemporary hydrology: towards holistic environmental science. In *Hydrological Modelling in Practice*, Watts, G. (ed.) John Wiley & Sons, Chichester.

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

Williams, R.J. and Zipser, D. (1994). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Chauvin, Y. and Rumelhart, D. (eds.) *Back-propagation: Theory, Architectures and Applications*. Hillsdale, NJ: Erbaum.

Wood, E.F. and Connell, P.E.O. (1985). Real-time forecasting. In *Hydrological Forecasting*, M.G. Anderson and T.P. Burt (eds.), pp. 505-558, John Wiley & Sons, Suffolk.

Wu, C.L. and Chau, K.W. (2010) Data-driven models for monthly streamflow time series prediction. *Engineering Applications of Artificial Intelligence*, 23(8), 1350-1367.

Wu, C.L. and Chau, K.W. (2011). Rainfall-runoff modeling using artificial neural network coupled with singular spectrum analysis. *Journal of Hydrology*, 399(3-4), 394-409.

Wu, C.L., Chau, K.W. and Li, Y.S. (2009). Methods to improve neural network performance in daily flows prediction. *Journal of Hydrology*, 372, 80-93.

Wu, Z.H. and Huang, N.E. (2004). A study of the characteristics of white noise using the empirical mode decomposition method. Proceedings of the Royal Society of London A, 460(2046), 1597-1611.

Xiong, L. and O'Connor, K.M. (2002). Comparison of four updating models for real-time river flow forecasting. *Hydrological Sciences Journal*, 47, 621-639.

Yang, C.-C. and Chen, C.-S. (2009). Application of integrated back-propagation network and self organizing map for flood forecasting. *Hydrological Processes,* 23, 1313-1323.

Yang, P.C., Wang, G.L., Bian, J.C. and Zhou, X.J. (2010). The prediction of non-stationary climate series based on empirical mode decomposition. *Advances in Atmospheric Sciences*, 27(4), 845-854.

Yazdani, M.R., Saghafian, B., Mahdian, M.H. and Soltani, S. (2009). Monthly runoff estimation using artificial neural networks. Journal of Agricultural Science and Technology, 11(3), 355-362.

Yu, L., Wang, S. and Lai, K.K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics*, 30, 2623-2635.

Zadeh, L. (1994). Fuzzy sets. *Information and Control*, 8, 338-353.

Zealand, C.M., Burn, D.H. and Simonovic, S.P. (1999). Short term stream flow forecasting using artificial neural networks. *Journal of Hydrology*, 214, 32-48.

Zhao, Z., Zhang, Y. and Liao, H. (2008). Design of ensemble neural network using the Akaike information criteria. *Engineering Applications of Artificial Intelligence*, 21, 1182-1188.

Zhang, B. and Govindaraju R.S. (2000b). Prediction of watershed runoff using Bayesian concepts and modular neural networks. *Water Resources Research*, 36 (3), 753-762.

Zhang, B. and Govindaraju, R.S. (2000a). Modular neural network for watershed runoff. In Govindaraju, R.S. and Ramanchandra, R.A. (eds.), *Artificial Neural Networks in Hydrology.* Kluwer Academic Publisher, The Netherlands.

Zhang, B. and Govindaraju, R.S. (2003). Geomorphology-Based artificial neural networks (GANNs) for estimation of direct runoff over watersheds. *Journal of Hydrology*, 273, 18-34.

Zhang, X., Lai, K.K. and Wang, S.-Y. (2008). A new approach for crude oil price analysis based on empirical mode decomposition. *Energy Economics*, 30(3), 905-918.

Zhang, X., Yu, L., Wang, S. and Lai, K.K. (2009). Estimating the impact of extreme events on crude oil price: An EMD-based event analysis method. *Energy Economics*, 31, 768-778.

Zhen-Shan, L. and Xian, S. (2007). Multi-scale analysis of global temperature changes and trend of a drop in temperature in the next 20 years. Meteorology and Atmospheric Physics, 95, 115-121.

Zhou, H.C., Peng, Y. and Liang, G-H. (2008). The research of monthly discharge predictor-corrector model based on wavelet decomposition. *Water Resources Management*, 22(1), 217–227.

Zounemat-Kermani, M. and Teshnehlab, M. (2008). Using adaptive neuro-fuzzy inference system for hydrological time series prediction. *Applied Soft Computing*, 8, 928-936.

Zurada, J.M. (1992). Introduction to Artificial Neural Systems. Boston: PWS Publishing Company.