Univ.-Prof. Dr.-Ing. A. Schlenkhoff (Hrsg.)
LuFG Wasserwirtschaft und Wasserbau
FB D – Abteilung Bauingenieurwesen
Bergische Universität Wuppertal

Samer Elabd

Data-Driven Modeling for Water Resources Management (Case study: The Ruhr River Basin)

Bericht Nr. 17, 2010

Vorwort (Hrsg.)

Die Bewirtschaftung von wasserwirtschaftlichen Systemen basiert ganz wesentlich auf einer fundierten Analyse vergangener Zeitreihen und der anschließenden Prognose über die Fortsetzung der Zeitreihe in der Zukunft. Modelle wie von Thomas und Fiering sind in der Hydrologie weit verbreitet und seit langem im Einsatz. Später wurde diese Art von Modellen zunehmend komplexer und allgemein zu sogenannten FARIMA Modellen (Fractional Auto Regressive Integrated Moving Average Modellen) zusammen gefasst. Während bei deterministischen Modellen umfangreiche physikalische Daten über das wasserwirtschaftliche System verfügbar sein oder erhoben werden müssen, werden stochastische Modelle anhand weniger und leicht zu beobachtenden Zeitreihen wie Niederschlag, Temperatur und Abfluss aufgebaut. Dafür müssen die Zeitreihen (Daten) aber eine entsprechend hohe Qualität und ausreichende Länge aufweisen. Die vorliegende Arbeit stellt anhand einer Reihe von Anwendungsbeispielen die Grundlagen und Möglichkeiten verschiedener stochastischer Modelle dar und diskutiert deren Anpassung für den beispielhaften Einsatz für die Talsperren im Einzugsgebiet der Ruhr.

Für die Zeitreihenanalyse oder die Generierung stochastischer Daten werden mittlerweile sehr komplexe mathematische Ansätze verwendet, die weit in die Mathematik und Informationstechnologie hineinreichen, in der Regel über die allgemeinen Kenntnisse der Wasserwirtschaftler hinausgehen und auf Ansätzen wie Künstliche Neuronale Netzwerke, Fuzzy Logic Methoden und Genetische Algorithmen basieren und teilweise auch in Kombination verwendet werden. Zudem sind entsprechend aufwendige Tests und Parameteranpassungen vorzunehmen, um die Modelleigenschaften zu beschreiben und an die Beobachtungen anzupassen. Hier liefert die vorliegende Arbeit einen Beitrag, die Lücke zwischen theoretisch formulierten mathematischen Ansätzen und den pragmatisch ausgerichteten Anwender mit Expertenwissen der wasserwirtschaftlichen Praxis zu schließen.Unter anderem wird die viel diskutierte Persistenz in den Zeitreihen (long memory) auf unterschiedliche Art und Weise dargestellt. Seit Hurst Anfang der 50ziger Jahre diesen Effekt beschrieben hat, wird über die Bestimmung und die Auswirkung für die Wasserwirtschaft diskutiert. Die wichtigste aufgegriffene Frage lautet dabei, wie dieser innere Zusammenhang bei der Generierung von künstlichen Zeitreihen berücksichtigt werden kann. Schließlich werden noch zwei gängige Verfahren zur Untersuchung der Stationarität vorgestellt. Ebenso interessant sind die Untersuchungen zur Zeitreihengenerierung mit Erhaltung der Hauptmomente der Häufigkeitsverteilung und unter Berücksichtigung der zeitlichen Abfolge des Auftretens, also der Korrelation. Es ergeben sich zahlreiche Ansätze, die Möglichkeiten der Zeitreihenanalyse für die Wasserwirtschaft stärker nutzbar zu machen.

Wuppertal, Januar 2011 *Andreas Schlenkhoff*

BERGISCHE
UNIVERSITÄT
WUPPERTAL

Data-Driven Modeling for Water Resources Management (Case study: The Ruhr River Basin)

Vom Fachbereich D (Abteilung Bauingenieurwesen)
der Bergischen Universität Wuppertal

genehmigte

Dissertation

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

von

M.Sc.-Ing. Mohamed Samer Elabd
aus Damiette - Ägypten

Eingereicht am:	26. Oktober 2010
Prüfung am:	22. Dezember 2010
Erster Gutachter:	Univ.-Prof. Dr.-Ing. Andreas Schlenkhoff LuFG Wasserwirtschaft und Wasserbau Bergische Universität Wuppertal
Zweiter Gutachter:	apl. Prof. Dr. rer. nat. habil. Gerd Morgenschweis LuFG Wasserwirtschaft und Wasserbau Bergische Universität Wuppertal
Vorsitzender:	Univ.-Prof. Dr.-agr. Jörg Rinklebe LuFG Boden- und Grundwassermanagement Bergische Universität Wuppertal
Weitere Mitglieder:	Univ.-Prof. Dr.-Ing. Andre Niemann Institut für Wasserbau und Wasserwirtschaft Universität Duisburg-Essen Dr.-Ing. Mario Oertel LuFG Wasserwirtschaft und Wasserbau Bergische Universität Wuppertal

Acknowledgments

I wish to express a deep and sincere gratitude to my advisor, Professor Andreas Schlenkhoff for his guidance, support and valuable advices throughout the study period.

I would like to thank Professor Gerd Morgenschweis for his valuable discussions and help. Furthermore, I want to thank all staff in the Ruhr Association (Ruhrverband) for their cooperation and support during the research period.

I am also grateful to Dr. Mario Oertel, Dr. Daniel Bung (former staff member), Melanie Sichelschmidt and all other staff members at the Institute for Geotechnics, Waste Management & Hydro Sciences, Wuppertal, Germany for their help throughout the study period.

Finally, my special thanks are extended to my parents for their encouragement and support. The most of all, my grateful thanks are due to my wife, Rabab for her help, support and patience, not only during my PhD study, but also throughout our lives.

Wuppertal, September 2010 *Samer Elabd*

Kurzfassung

Die Anwendung von stochastischen Methoden und datenbasierten Modellen zur Analyse und Prognose von Zeitreihen natürlicher Prozesse hat bei der Bewirtschaftung und Verwaltung von Wasserressourcen eine lange Tradition. In der vorliegenden Dissertation wird aufbauend auf den grundlegenden statistischen Methoden eine Reihe von gebräuchlichen stochastischen und datenbasierten Modellen auf die wasserwirtschaftlichen Fragestellungen angewendet. Die Modelle werden gegebenenfalls an die wasserwirtschaftlichen Fragestellungen angepasst oder bezüglich der Periodizität weiterentwickelt. Nach der theoretischen Beschreibung werden die Eignung und die Anwendbarkeit der untersuchten Modelle beispielhaft für die Zuflussdaten einiger Talsperren im Einzugsgebiet der Ruhr dargestellt (Bigge, Henne, Möhne und Sorpe). Hierbei stehen die Methoden und Modelle im Vordergrund der Betrachtung. Eine unmittelbare betriebliche Anwendung ist nicht Gegenstand der Untersuchung.

Diese Arbeit konzentriert sich auf die folgenden Themen:

1. Untersuchung stochastischer Eigenschaften von Zuflüssen
Die stochastischen Eigenschaften der Zuflüsse der Talsperren werden untersucht. Die Zuflüsse werden in Bezug auf die Saisonabhängigkeit, die Trendentwicklung, die Korrelation und die Stationarität betrachtet. Die Ergebnisse der saisonalen Betrachtung der untersuchten Zeitreihen der täglich, 10-täglich und monatlich gemittelten Werte werden dargestellt und zeigen, dass:

- die Saisonalität stark ausgeprägt ist und sich nicht nur auf die Mittelwerte, sondern auf alle statistischen Werte auswirkt,
- der Variationskoeffizient in trockenen Zeiten höhere Werte hat,
- der Schiefekoeffizient ebenfalls höhere Werte in trockenen Zeiten aufweist und umgekehrt,

- die Autokorrelationskoeffizienten bei allen Zuflusszeitreihen in Jahreszeiten mit einem hohen Zufluss niedrig und in Jahreszeiten mit einem geringen Zufluss relativ hoch sind. Dies trifft allerdings nicht für die Zeitreihen der Tagesabflüsse zu, die eine starke Korrelation über die Länge eines Tages aufweisen.

Trenduntersuchungen auf der Basis des saisonabhängigen Mann-Kendall Tests zeigen, dass nur bei der Sorpe ein Abwärtstrend bei allen untersuchten Zuflusszeitreihen mit einem Signifikanzniveau von 5 % angenommen werden kann.

Die erweiterten (Augmented) Dickey-Fuller und Phillips-Perron Unit-Root-Tests werden zur Prüfung der Stationarität, ebenfalls mit einem Signifikanzniveau von 5 %, benutzt. Hierzu werden die logarithmierten und standardisierten Werte der aggregierten (täglichen, 10-täglichen, monatlichen, dreimonatlichen, sechsmonatlichen und jährlichen) Zuflusszeitreihen verwendet. Die Ergebnisse dieser Tests zeigen, dass alle Zuflusszeitreihen stationär zu sein scheinen.

2. Vorhersage täglicher Zuflüsse

Für die Vorhersage der täglichen Zuflüsse zu den Talsperren für den nächsten und übernächsten Tag werden folgende Modellgruppen verwendet: BPNN (Back Propagation Neural Network), ANFIS (Adaptive Neuro-Fuzzy Interference Systems), ARMA (Autoregressive Moving Average) und ARFIMA (Autoregressive Fractionally Integrated Moving Average). Diese Modelle basieren zum Teil auf einer erweiterten Kombination von Autoregression und Moving Average Verfahren sowie auf Methoden von Neuronalen Netzwerken und Ansätzen nach der Methode der Fuzzy Logic. Die Simulationsmodelle werden nach der angenommenen Verfügbarkeit von Eingangsdaten (Zufluss bzw. Zufluss und Niederschlag) in zwei Modellgruppen unterteilt:

Univariate Modelle (Gruppe M1-1 and Gruppe M1-2)

- Verwendete Modellgruppen: BPNN, ANFIS, ARMA und ARFIMA.

Multivariate Modelle (Gruppe 2)

- Verwendete Modellgruppen: BPNN und ANFIS.

Die in dieser Dissertation für das BPNN Modell verwendeten "Trainingsalgorithmen" basieren auf einem Ansatz nach Levenberg-Marquardt. Als Aktivierungsfunktionen werden tan-sigmoid Funktionen für die Neuronen im "Hidden-Layer" und lineare Funktionen für die Output Neuronen verwendet. Der Problematik der parametrischen Überbestimmung wird sowohl bei den BPNN- als auch bei den ANFIS-Modellen durch ein definiertes

Abbruchkriterium (early stopping procedure) begegnet. Die BPNN-Modelle werden im Trainingsstatus mit "nur" einem "hidden layer" ausgestattet. Die Anzahl der Neuronen und der set der „besten" Eingangsvariablen werden mittels einer "trial-and-error" Routine ermittelt. Die beste Anpassung für das ANFIS-Modell wird mit einer weiteren "trial-and-error" Routine ermittelt, wobei die Anfangsparameter aus dem BPNN-Model übernommen werden. Die jeweilige Ordnung für die Ansatzfunktion von "Autoregressiv" und "Moving Average" für die ARMA- und ARFIMA-Modelle wird bis zur fünften Ordnung formuliert. Als Bewertung der Modellgüte wird das AIK-Kriterium (Akaike Information Criterion) verwendet. Die Prognosegüte der ARMA- und ARFIMA-Modelle wird mit dem "Ljung-Box-Test" mit einem Signifikanzniveau von 5 % getestet. Dieser Test zeigt, dass die Nullhypothese (die Annahme eines geeigneten Modellsatzes) nur für die Simulation des täglichen Zuflusses zu der Henne- und Möhnetalsperre bei Verwendung des ARFIMA-Modells verworfen wird. Um die Modellgüten miteinander zu vergleichen, werden folgende Kriterien verwendet: Korrelationskoeffizient, Fehlerquadratmethode, mittlerer relativer Fehler (*AREP*), Index of Agreement, Nash-Sutcliffe-Koeffizient. Der Vergleich zeigt, dass die verwendeten univariaten Modelle mit Ausnahme des *AREP* ähnliche Leistungen aufweisen. BPNN- und ANFIS-Modelle sind dabei für die Prognose der täglichen Zuflüsse geringfügig besser einzuschätzen.

3. Ergänzung und Schließung von Datenlücken in Zuflusszeitreihen
Unterschiedliche, aber gebräuchliche Modellansätze, wie BPNN, ANFIS und GLM (Generalized Linear Model) werden auf ihre Eignung zum Füllen von Datenlücken in den täglichen Zuflusszeitreihen der Talsperren bewertet. Grundlage hierfür ist die hohe Korrelation in den untersuchten Zeitreihen. Der Zufluss zu einer Talsperre wird hierbei anhand der Zuflüsse der drei anderen Talsperren geschätzt, die im Sinne einer Regionalisierung ähnliche Eigenschaften aufweisen. In Bezug auf die vorgeschlagenen Werte weisen BPNN-Modelle kleinere Gesamtabweichungen, gemessen als Fehlerquadrat, zum tatsächlichen Wert auf. Das BPNN-Modell wird daher beispielhaft verwendet, um monatliche Zuflüsse zur Biggetalsperre zu generieren.

4. Generierung monatlicher Zuflüsse
Zum Generieren monatlicher Zuflussdaten werden weiterhin folgende Modelle verwendet: T-F (Thomas-Fiering), Gamma T-F, MC (Monte Carlo) und PHMM (Periodic Hidden Markov Model). Hierbei werden die Zufallszahlen über eine inverse Transformation generiert. Für die Gamma T-F-Modelle muss zudem die Schiefe der Verteilungsfunktion mittels Wilson-Hilferty Transformation abgebildet werden. Die Persistenz der Monatsabflüsse, soweit vorhanden, wird optional über eine Cholesky Decompostion erhalten. Mit diesen

Modellen werden Zeitreihen von 100, 300 und 500 Jahren Länge generiert. Die oben erwähnten statistischen Parameter werden mit denen der beobachteten Zeitreihen verglichen. T-F-, MC- und PHMM-Modelle stellen die statistischen Parameter Mittelwert, Standardabweichung und Schiefekoeffizienten gut bis sehr gut dar und deutlich besser als das Gamma T-F-Modell.

PHMM ist eine in dieser Dissertation neu entwickelte Methode, welche mit Ausnahme der Persistenz alle statistischen Parameter sehr gut abbilden kann. Dies wird besonders gut anhand einer graphischen Auswertung deutlich. Hierfür werden die Methode der Quantil-zu-Quantil-Darstellung und die „Survivor Function Plot"-Methode verwendet. Diese Darstellungen zeigen, dass die MC- und PHMM-Modelle die Verteilungsfunktionen sehr gut wiedergeben können. Dies betrifft auch den Bereich der Extremwerte, wo die PHMM-Methode allerdings besonders überzeugt. Bezüglich der Persistenz bedarf die PHMM-Methode einer Weiterentwicklung.

Aus den so erzeugten Zeitreihen wird beispielhaft ein konsekutiver Satz von fünf Jahren so bestimmt, dass die Zuflusssumme minimal wird und weiteren Untersuchungen zugeführt werden kann.

5. Vorhersage der Fließzeit der Wasserabgaben aus Talsperren

Historische Abflussdaten (15 Minutenzeitreihen) werden zur Abschätzung der Fließzeit der Wasserabgaben aus den Talsperren bis zu einigen Kontrollpegeln flussabwärts benutzt. Die Fließzeit wird zunächst in einem Nicht-Linearen Regressionsmodell (NLR) mit dem Abfluss an dem Kontrollpegel in Relation gestellt. Die geschätzten Fließzeitwerte und die entsprechende Abflussmenge an den Abgabepegeln und an den Kontrollpegeln flussabwärts werden weiterhin in die Modelle ANFIS, BPNN und MLR (Multi Linear Regression) als Datenbasis eingegeben. Wegen der begrenzten Anzahl an auswertbaren Ereignissen werden die Daten sowohl bei der Anpassung als auch bei der Validierung genutzt. Hierfür werden die beiden folgenden Verfahren verwendet: „k-Fold-Cross-Validation (KFCV)" und „Leave-One-Out-Cross-Validation (LOOCV)". Als Ergebnis werden ANFIS-Modelle als eine geeignete Methode für die Fließzeitvorhersage für die Talsperren im Ruhreinzugsgebiet vorgeschlagen.

Der obere und mittlere Flussabschnitt der Ruhr wird zudem mit einem eindimensionalen instationären hydrodynamischen Modell (HEC-RAS) simuliert. Die Ergebnisse werden mit denen des NLR-Modells verglichen und erreichen für mittlere Abflüsse gute Ergebnisse. Bei kleineren Abflüssen versagen beide Modelle bzw. die Standardabweichung erreicht unakzeptable Werte.

Für die Bereitstellung der Datengrundlage, ohne die eine solche Arbeit nicht möglich gewesen wäre, wird dem Ruhrverband besonders gedankt. Für die in dieser Dissertation beschriebenen Modelle wurden Codes und GUIs in MatLab (The MathWorks, Inc.) erstellt.

Summary

Application of stochastic methods and data-driven models to time series analysis and reservoir operation has been a major focus of water resources planning and management. The aim of this study is to investigate the suitability of applying these models in water resources planning and management. Stochastic analysis and data-driven models are applied to the management and operation of reservoirs (case study, the Bigge, Henne, Möhne and Sorpe reservoirs in the Ruhr River basin). The ability of these models to accurately simulate water resources management problems is the first priority of this study. However, an operational application is not the main object.

This thesis is focused on the following topics:

1. The stochastic properties of inflow processes
The stochastic characteristics of the inflow processes of the reservoirs are examined. The inflow processes are investigated for seasonality, trend, long memory and stationarity. The results of the seasonality test of the daily, 10-days and monthly inflow time series show that:

- The inflow time series have a clear seasonality in the mean and standard deviation. Seasons with high mean values have also high standard deviations.
- The coefficient of variation has higher values in the dry periods.
- The higher values of the skewness occur in seasons with low flow and vice versa.
- The autocorrelation coefficients are low for high inflow seasons and high for seasons with low inflow for all inflow time series, except daily inflow time series at lag of one day.

The results of the seasonal Mann-Kendall test at 5 % significance level indicate that a downward trend is only detected for all tested inflow time series of the Sorpe reservoir. The Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) unit root tests are used to test the stationarity of the log-transformed and standardized daily, 10-days, monthly, 3-months, 6-months and annual inflow time series at 5 % significance level. The results show that all inflow time series appear to be stationary by applying log-transformation and standardization to them.

2. Forecasting of daily inflow

The applicability of the backpropagation neural network (BPNN), adaptive neuro-fuzzy inference system (ANFIS), autoregressive moving average (ARMA) and the autoregressive fractional integrated moving average (ARFIMA) models are explored to one-step and two-steps ahead forecasting of the daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs. These models are divided into two groups according to the potential input variables:

Univariate models (group M1-1 and group M1-2)

- The simulation models are the BPNN, ANFIS, ARMA and ARFIMA.
- The potential input variables are the average daily inflow.

Multivariate models (group M2)

- The simulation models are the BPNN and ANFIS.
- The potential input variables are the average daily inflow and the daily rainfall.

The training algorithm that is utilized for all the BPNN models in the dissertation is Levenberg-Marquardt algorithm and the used activation functions are tan-sigmoid and linear functions for the hidden layer neurons and for the output one respectively. The overfitting problem is suppressed in the BPNN and ANFIS models by applying the early stopping procedure. The BPNN models are trained using one hidden layer. The number of neurons in the hidden layer and the optimal input variables in the BPNN models are determined using a trial-and-error procedure. Starting with the input variables of the optimum BPNN models, another trial-and-error procedure is developed to find the ANFIS models which have the best performance. The orders of the autoregressive (AR) and moving average (MA) components in the ARMA and ARFIMA models are determined by trying different values between 0 and 5. The Akaike Information Criterion (AIC) is used to select the best ARMA and ARFIMA models (the models with minimum AIC). The diagnostic of the ARMA and ARFIMA models are tested by applying the Ljung-Box test at 5 % significance level and the results show that the null hypothesis of model adequacy is rejected only for the simulated daily inflow time series of the Henne and Möhne reservoirs using the ARFIMA model.

Different efficiency criteria (correlation coefficient, root mean square error, average relative error percentage, index of agreement and Nash-Sutcliffe coefficient) are used to compare the performance of the models. The comparison shows that the performances of the models, group M1-1 and the models, group M1-2, don't have significantly different performance except for the average relative error percentage (*AREP*). The BPNN and ANFIS models have the minimum values of the *AREP* for all daily inflow time series. The models, group M2, are found to outperform the models group M1-2, in respect of all used efficiency criteria.

3. Filling missing values in daily inflow time series

The efficiency of the BPNN and ANFIS models and the generalized linear model (GLM) for filling the missing values in the daily inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs are explored. High correlation values between the daily inflow time series are detected. Therefore, the inflow of each reservoir is estimated using the inflow data of the other reservoirs as input variables. The BPNN models are trained using one hidden layer with three neurons. ANFIS models with three membership functions (Gaussian type) associated with each input variable are used. The link function and distribution of the response for GLM are selected using a trial-and-error procedure. In respect of the estimated values of the root mean square error (*rmse*), the BPNN models have better performances for filling the missing data. The BPNN model is employed to extend the time series of the monthly inflow into the Bigge reservoir in the period from 11/1960 to 10/1965.

4. Generation of the monthly inflow data

The Thomas-Fiering (T-F), Gamma Thomas-Fiering (Gamma T-F), Monte Carlo (MC) and periodic hidden Markov (PHMM) models are applied to generate monthly inflow data into the Bigge, Henne, Möhne and Sorpe reservoirs. The inverse transform method is used to generate random numbers in the T-F and MC models. However, Wilson-Hilferty transformation is proposed to reproduce skewed noises in the Gamma T-F model. The Cholesky decomposition method is used to preserve month-to-month correlation in the generated monthly inflow data by the MC model. Three monthly inflow time series with lengths 100, 300 and 500 years are generated using T-F, Gamma T-F, MC and PHMM. The statistical parameters (mean, standard deviation, month-to-month correlation and skewness) of the generated monthly inflow are compared with those of the observed one. The results of the comparison show that the T-F, MC and PHMM models reproduce most of the statistical parameters very well. PHMM is a methodology newly developed within this thesis to generate monthly inflow. PHMM has the ability to reproduce all statistical

parameters (except month-to-month correlation) very well. Finally, using the quantile-quantile (Q-Q) and the survivor function plots, the observed and the simulated monthly distributions are graphically compared. These plots indicate the ability of the MC and PHMM models to reproduce the statistical distribution of the observations, in particular the extreme values with superiority of the PHMM. More research is needed to improve the performance of PHMM in preserving month-to-month correlation. A procedure is developed to detect the expected consecutive 5 years that have minimum total inflow using the MC model. The generated monthly inflow time series during the 5 years can be adopted as an inflow scenario for optimization of the reservoir operation.

5. Prediction of the travel time of reservoirs' releases along the Ruhr
Historical flow data (15 minute time series) are used to estimate the travel time of the released flow from the Bigge, Sorpe, Möhne and Henne reservoirs to some downstream gauges. The estimated travel time values are used to build the nonlinear regression (NLR) models to detect the relation between travel time and the flow at each downstream gauge. These NLR models can be easily used to predict the travel time when knowing the flow at the downstream gauge. The estimated travel time values along the reach from gauge Ahausen to gauge Hagen-Hohenlimburg are simulated using the ANFIS, BPNN and multiple linear regression (MLR) models. Due to the limited amounts of the travel time data, k-fold cross validation (KFCV) and leave-one-out cross validation (LOOCV) are used to estimate the generalization errors in the ANFIS, BPNN and MLR models. The values of the generalization error show that the ANFIS model A5 outperforms the other models. The ANFIS model A5 has the following input variables:

1. The increase in the reservoir release (ΔQ_R) with two membership functions (Gaussian type).

2. The discharge at the downstream gauge (Q_D) with three membership functions (Gaussian type).

The upper and middle reaches of the Ruhr River are simulated using the Hydrologic Engineering Center River Analysis System (HEC-RAS) and the results are compared with those of the NLR model. The comparison shows a moderate agreement between the results of the two models.

A graphical user interface (Fliesszeit GUI) is developed using Matlab (The MathWorks, Inc). For any 15 minutes historical flow data this GUI can be used to:

1. Determine the jump points in the flow at the release gauge and the corresponding downstream gauges.

2. Plot the hydrographs at each jump point. These hydrographs can be used to estimate the travel time.

3. The estimated travel time value can be manually entered to update the travel time values that have been estimated previously.

4. The simulation models can be trained using the updated travel time values.

Especially thanks to the Ruhr authority for providing us with data basis without which this work would not have been possible.

Abbreviations

ACF	autocorrelation function
ACVF	autocovariance function
ADF	augmented Dickey-Fuller unit root test
AIC	Akaike information criterion
ANFIS	adaptive neuro-fuzzy inference system
ANN	artificial neural network
AR	autoregressive model
AREP	relative average error percentage
ARFIMA	autoregressive fractional integrated moving average model
ARMA	autoregressive moving average model
BLPs	best linear predictors
BPNN	backpropagation neural network model
cdf	cumulative distribution function
CI	confidence interval
DDM	data-driven modeling
DF	Dickey-Fuller distribution
DP	dynamic programming
EA	evolutionary algorithms
ENSO	El Niño Southern Oscillation
fBm	fractional Brownian motion
fGn	fractional Gaussian noise
FI	fractionally integrated part in the ARFIMA model
FIS	fuzzy inference system
FL	Fuzzy logic
Gamma T-F	Gamma Thomas-Fiering model
GIS	geographic information system
GLM	generalized linear model

GPE	generalized final prediction error
GPH	Geweke and Porter-Hudak estimator of the long memory parameter
GUI	graphical user interface
HEC-RAS	the Hydrologic Engineering Center River Analysis System
HMM	hidden Markov model
KFCV	k-fold cross validation method
LLF	conditional log-likelihood function
LOOCV	leave-one-out cross validation method
LP	linear programming
LRD	long-range dependent
MA	moving average model
MC	Monte Carlo model
M-F	method of fragments
MLE	exact maximum likelihood estimate
MLP	multilayer perceptron model
MLR	multi-linear regression
mse	mean square error
NIC	Network information criterion
NLP	nonlinear programming
NLR	nonlinear regression
NN	neural network
NSC	Nash-Sutcliffe model efficiency coefficient
OLS	ordinary least square analysis
$PACF$	partial autocorrelation function
PAR	periodic autoregressive model
PARMA	periodic autoregressive moving average model
pdf	probability density function
PHMM	periodic hidden Markov model
PP	Phillips-Perron unit root test
Q-Q plot	quantile-quantile plot
R/S	rescaled adjusted range method
RBF-NN	artificial neural networks with radial base function model
REP	relative error percentage
$rmse$	root mean square error
S-C	Spolia-Chan model
SMA	simple moving average

SRD	short-range dependent
T-F	Thomas-Fiering model
TFPW	trend-free pre-whitening
T-T	Two-Tier model
UNET	unsteady NETwork model
VC	Vapnik-Chervonenkis dimension

Contents

Acknowledgments I

Kurzfassung III

Summary IX

Abbreviations XV

List of Figures XXV

List of Tables XXXI

1 Introduction 1

1.1 Introduction . . . 1

1.1.1 The Ruhr River basin . . . 1

1.1.2 Runoff and water demand . . . 1

1.2 Hydrology versus water resources management . . . 3

1.3 Data driven modeling . . . 4

1.3.1 Introduction . . . 4

1.3.2 Machine learning . . . 4

1.3.3 Data sets . . . 5

1.3.4 Popular data-driven methods and typical application . . . 5

1.4 Objectives of the research . . . 7

1.5 Outline of the thesis . . . 7

1.6 References . . . 8

2 Stochastic properties of reservoirs inflow processes **11**

2.1 Introduction . . . 11

2.2 Data used . . . 11

2.3 Seasonality analysis . . . 12

2.3.1 Results of the seasonality analysis 16

2.4 Normalization and standardization 17

2.5 Testing for trend 19

2.5.1 Linear regression method 24

2.5.2 Mann-Kendall test 24

2.5.3 Seasonal Mann-Kendall trend test 26

2.5.4 Results of the linear regression and Mann-Kendall tests 28

2.5.5 Results of the seasonal Mann-Kendall test 30

2.6 Testing for long memory 36

2.6.1 Introduction 36

2.6.2 Presence of long memory 37

2.6.3 Methods for estimation of H and d 38

2.7 Testing for stationarity 45

2.7.1 Introduction 45

2.7.2 Stationarity test methods 45

2.7.3 Augmented Dickey-Fuller unit root test (ADF) 45

2.7.4 Phillips-Perron unit root test (PP) 48

2.7.5 Results of the ADF and PP tests 48

2.8 Conclusions 52

2.9 References 53

3 Daily inflow forecasting **57**

3.1 Introduction 57

3.2 Simulation models 58

3.3 Backpropagation neural network (BPNN) 59

3.3.1 Data preprocessing 60

3.3.2 Training of BPNN models 61

3.3.3 Optimum size of neural networks 62

3.3.4 Improving generalization of the neural networks 63

3.3.5 Activation Functions 65

3.4 Adaptive neuro-fuzzy inference system (ANFIS) 66

3.4.1 ANFIS structure 67

3.4.2 Learning algorithm for ANFIS 68

3.5 Autoregressive moving average ARMA(p,q) processes 68

3.5.1 Introduction to ARMA models 68

3.5.2 ARMA model identification 70

3.5.3 Estimation of the parameters of ARMA model 71

3.5.4 Diagnostic checking 72
3.6 Fractionally integrated ARMA processes (ARFIMA) 73
3.6.1 Introduction to ARFIMA models 73
3.6.2 Estimation of an ARFIMA(p,d,q) model 74
3.7 Forecasting using ARMA models 75
3.8 Models efficiency criteria 77
3.8.1 Average relative error percentage (AREP) 77
3.8.2 Cross correlation coefficient (R) 77
3.8.3 Nash–Sutcliffe model efficiency coefficient (NSC) 78
3.8.4 Index of agreement (g) 78
3.9 Building of the models 79
3.9.1 BPNN models 79
3.9.2 ANFIS models 79
3.9.3 ARMA models 80
3.9.4 ARFIMA models 81
3.9.5 Results of diagnostic checks for ARMA and ARFIMA models 82
3.10 Forecasting performance assessment 83
3.10.1 Univariate models (group M1-1 and group M1-2) 83
3.10.2 Multivariate models (group M2) 84
3.10.3 Univariate vs. multivariate models 84
3.11 Conclusions 93
3.12 References 94

4 Filling in missing data 99
4.1 Introduction 99
4.2 Data used 100
4.3 Selection of the inputs 100
4.4 Missing data estimation models 101
4.4.1 BPNN and ANFIS models 101
4.4.2 Generalized linear model (GLM) 101
4.5 Evaluation of the models 104
4.6 Extension of the monthly inflow data 104
4.7 Conclusions 107
4.8 References 108

5 Monthly inflow generation 111
5.1 Introduction 111

5.2 Data used 113
5.3 Thomas-Fiering model 113
5.4 Gamma Thomas-Fiering model (Gamma T-F) 114
5.5 Monte Carlo model 115
5.6 Random values generation 115
5.7 Lag one month-to-month correlation 117
5.8 Hidden Markov models (HMM) 119
5.8.1 Introduction to HMM 119
5.8.2 HMM for monthly inflow generation 121
5.9 Validation of the models 122
5.9.1 Statistical parameters of the monthly inflow time series 122
5.9.2 Q-Q and survivor function plots for models validation 123
5.10 Consecutive 5 years with minimum inflow 125
5.11 Conclusions 130
5.12 References 136

6 Prediction of the travel time of the reservoirs' releases along the Ruhr and the Lenne 139
6.1 Introduction 139
6.2 Travel time estimation 140
6.3 Nonlinear regression analysis 142
6.4 Travel time estimation using multivariate models 146
6.4.1 Model generalization 146
6.4.2 Adaptive neuro-fuzzy inference system (ANFIS) 147
6.4.3 Backpropagation neural network (BPNN) 155
6.4.4 Multiple linear regression (MLR) 159
6.5 HEC-RAS for travel time prediction 160
6.5.1 Introduction to HEC-RAS 160
6.5.2 Description of the HEC-RAS geometry model 161
6.5.3 Flow scenarios 162
6.5.4 Comparison between HEC-RAS and NLR for travel time prediction . 162
6.6 Conclusions 165
6.7 References 165

7 Conclusions and future work 167
7.1 Conclusions 167
7.2 Future work 168

A Graphical user interfaces 169

A.1 Vorhersage GUI 169

A.1.1 Statistical analysis 170

A.1.2 Monthly inflow generation 171

A.1.3 Critical moving sum 172

A.2 Fehlende Daten GUI 173

A.3 Zuflussprognose GUI 175

A.4 Fliesszeit GUI 176

A.4.1 Prediction of the travel time 176

A.4.2 Updating of training models 178

A.4.3 Estimation of travel time 179

B Graphical presentation of the long memory detection methods 183

B.1 Aggregated variance method 184

B.2 Residuals of regression method 185

B.3 The rescaled adjusted range (R/S) method 186

Resume – Curriculum Vitae 187

List of Figures

1.1 Layout of the Ruhr River 2
1.2 Annual abstracted and exported water in the Ruhr catchment area between 1900 and 2005 (Ruhrverband, 2005) 2
1.3 Learning in data driven modeling (Solomatine, 2002) 5

2.1 Average daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs . . 13
2.2 Seasonal variation in the mean ($\bar{x}$), standard deviation (SD) and coefficient of variation (CV) of the daily inflow time series 18
2.3 Seasonal variation in the skewness (g) of the daily inflow time series 19
2.4 Seasonal variation in the season-to-season autocorrelation coefficient of the daily inflow time series at different lags 20
2.5 Seasonal variation in the $\bar{x}$, standard deviation (SD) and coefficient of variation (CV) of the 10-days inflow time series 21
2.6 Seasonal variation in the skewness (g) of the 10-days inflow time series . . . 21
2.7 Seasonal variation in the season-to-season autocorrelation coefficient of the 10-days inflow time series at different lags 22
2.8 Seasonal variation in the $\bar{x}$, standard deviation (SD) and coefficient of variation (CV) of the monthly inflow time 22
2.9 Seasonal variation in the skewness (g) of the monthly inflow time series . . . 23
2.10 Seasonal variation in the season-to-season autocorrelation coefficient of the monthly inflow time series at different lags 23
2.11 Monthly inflow into the Bigge reservoir (Nov. 1966 - Oct. 2006) 31
2.12 Monthly inflow into the Henne reservoir (Nov. 1960 - Oct. 2006) 32
2.13 Monthly inflow into the Möhne reservoir (Nov. 1960 - Oct. 2008) 33
2.14 Monthly inflow into the Sorpe reservoir (Nov. 1960 - Oct. 2006) 34
2.15 Annual inflow into the Bigge reservoir (1966 - 2006) 34
2.16 Annual inflow into the Henne reservoir (Nov. 1960 - Oct. 2006) 35
2.17 Annual inflow into the Möhne reservoir (Nov. 1960 - Oct 35

2.18 Annual inflow into the Sorpe reservoir (Nov. 1960 - Oct. 2006) 35
2.19 Autocorrelation functions of the inflow time series of the Bigge reservoir . . 44
2.20 Autocorrelation functions of the inflow time series of the Henne reservoir . . 44
2.21 Autocorrelation functions of the inflow time series of the Möhne reservoir . 44
2.22 Autocorrelation functions of the inflow time series of the Sorpe reservoir . . 44

3.1 A typical three-layer feed forward ANN 60
3.2 Early stopping for improving generalization 64
3.3 The unit of an artificial neural network 65
3.4 a) Hard-limit b) Linear c) Log-sigmoid d) Tan-sigmoid The hard-limit, linear, log-sigmoid and tan-sigmoid transfer functions 66
3.5 a) An illustration of the reasoning mechanism for a Sugeno-type model, b) ANFIS architecture 69
3.6 The ACF of the residuals from the ARMA models (models, group M1-1) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs . 85
3.7 The ACF of the residuals from the ARFIMA models (models, group M1-1) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs 85
3.8 The ACF of the residuals from the ARMA models (models, group M1-2) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs . 86
3.9 The ACF of the residuals from the ARFIMA models (models, group M1-2) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs 86
3.10 Comparative one-day-ahead forecasting of the daily inflow into Bigge reservoir (models, group M1-1) 89
3.11 Comparative two-days-ahead forecasting of the daily inflow into the Bigge reservoir (models, group M1-1) 90
3.12 Values of the efficiency criteria vs. the predicted daily inflow (Bigge reservoir) 91
3.13 Values of the efficiency criteria vs. the predicted daily inflow (Henne reservoir) 91
3.14 Values of the efficiency criteria vs. the predicted daily inflow (Möhne reservoir) 92
3.15 Values of the efficiency criteria vs. the predicted daily inflow (Sorpe reservoir) 92

4.1 Cross correlation between the inflow data of each two stations 101
4.2 Comparison between the observed and estimated (using the BPNN models) daily inflow time series 105
4.3 Statistical parameters of the observed and estimated monthly inflow of the Bigge reservoir 106
4.4 The observed and extended monthly inflow of the Bigge reservoir 107

5.1 Plots of the monthly inflow 116

5.2 Cumulative distribution function of a variable X (Gentle, 2003) 118

5.3 The lag one month-to-month correlation of the observed and generated (using the MC model) monthly inflow 120

5.4 a) The transition diagram and b) The transition matrix, for monthly inflow generation HMM 121

5.5 Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Bigge reservoir 126

5.6 Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Henne reservoir 127

5.7 Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Möhne reservoir 128

5.8 Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Sorpe reservoir 129

5.9 The Q-Q plots for the observed and 300 years generated monthly inflow of the Bigge reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 130

5.10 The Q-Q plots for the observed and 300 years generated monthly inflow of the Henne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 131

5.11 The Q-Q plots for the observed and 300 years generated monthly inflow of the Möhne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 131

5.12 The Q-Q plots for the observed and 300 years generated monthly inflow of the Sorpe reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 132

5.13 The survivor function plots for the observed and 300 years generated monthly inflow of the Bigge reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 133

5.14 The survivor function plots for the observed and 300 years generated monthly inflow of the Henne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 133

5.15 The survivor function plots for the observed and 300 years generated monthly inflow of the Möhne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 134
5.16 The survivor function plots for the observed and 300 years generated monthly inflow of the Sorpe reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models 134
5.17 The observed monthly inflow and the generated during the critical consecutive 5 years 135

6.1 List of the release and downstream gauges 141
6.2 Determination of the jump points TRi at the release gauge and TRi at the downstream gauge 142
6.3 Hydrographs at Ahausen, Altena and Hagen-Hohenlimburg (4/12/1996) . . 143
6.4 Hydrographs at Langscheid and Bachum (13/9/1999) 143
6.5 Hydrographs at Günne and Bachum (14/12/1998) 144
6.6 Hydrographs at Meschede-Henne, Meschede-Ruhr and Oeventrop (14/2/1998)144
6.7 Cross correlation between travel time t and ΔQ_R, Q_D and ΔQ_D (case study 2) 144
6.8 The relation between the travel time (t) and downstream flow (Q_D) for all case studies 149
6.9 The ANFIS architecture of the input-output system (model A5) 150
6.10 The values of the AREP for the ANFIS models (data1 - KFCV) 150
6.11 The values of the AREP for the ANFIS models (data2 - KFCV) 151
6.12 The values of the AREP for the ANFIS models (data1 & data2 - LOOCV) . 151
6.13 The values of the AREP for the ANFIS models (data3 - KFCV) 153
6.14 The values of the AREP for the ANFIS models (data4 - KFCV) 154
6.15 TThe values of the AREP for the ANFIS models (data3 - LOOCV) 154
6.16 Architecture of the Backpropagation Neural Network (BPNN) for travel time prediction (model, D5) 156
6.17 The values of the AREP for the BPNN models (data1 - KFCV) 156
6.18 The values of the AREP for the BPNN models (data2 - KFCV) 157
6.19 The values of the AREP for the BPNN models (data3 - KFCV) 157
6.20 The values of the AREP for the BPNN models (data4 - KFCV) 158
6.21 The values of AREP for BPNN models (data1 - LOOCV) 158
6.22 The values of AREP for BPNN models (data3 - LOOCV) 158
6.23 The Geometric data of the HEC-RAS model 161

6.24 The flow at the release gauges and the predicted flow (from the HEC-RAS simulation) at the downstream gauges (scenario Meschede-Henne_3) 163
6.25 Travel time from Henne reservoir to the gauge Oeventrop versus Q_D 164
6.26 Travel time from Sorpe reservoir to the gauge Bachum versus Q_D 164
6.27 Travel time from Möhne reservoir to the gauge Bachum versus Q_D 164

A.1 General overview of the Vorhersage GUI 169
A.2 Vorhersage GUI for seasonality analysis 170
A.3 Vorhersage GUI for monthly inflow generation 172
A.4 Vorhersage GUI for determination of the critical moving sum 173
A.5 General overview of the Fehelend Daten GUI 174
A.6 General overview of the Zuflussprognose GUI 175
A.7 General overview of the Fliesszeit GUI 177
A.8 Prediction of the travel time 177
A.9 Updating of the training models 180
A.10 Performance of the models 180
A.11 Estimation of the travel time 181

B.1 Logarithmic plot of the variance versus the aggregate level (the aggregate variance method) 184
B.2 Logarithmic plot of the residual variance versus the aggregate level (the residuals of regression method) 185
B.3 Logarithmic plot of the R/S variance versus the size of block m (the R/S method) 186

List of Tables

1.1 Data of the main reservoirs in the Ruhr River basin 3

2.1 Statistical characteristics of the raw inflow time series for different timescales 14

2.2 Values of the statistics T_c and Z_{MK} for the 10-days mean inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs 29

2.3 Values of the statistics T_c and Z_{MK} for the monthly mean inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs 30

2.4 Values of the statistics T_c and Z_{MK} for the 3-months, 6-months and annual inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs 36

2.5 Values of the seasonal Mann-Kendall test statistic, Z' for the 10-days, monthly, 3-months and 6-months inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs 36

2.6 Relationships between parameters that capturing long-range dependence. . . 38

2.7 Results of the long-memory detection methods of the original inflow time series 43

2.8 Results of the stationarity tests of the log-transformed inflow time series . . 50

2.9 Results of the stationarity tests of the log-transformed and standardized inflow time series 51

3.1 List of the used data for daily inflow forecasting 59

3.2 Parameters of the BPNN models 80

3.3 Parameters of the ANFIS models 81

3.4 Parameters of the best autoregressivge moving average ARMA(p,q) models 81

3.5 Parameters of the best autoregresssivge fractional integrated moving average ARFIMA(p,d,q) models 82

3.6 Results of the Ljung-Box Q-test 87

3.7 Values of the efficiency criterion parameters for the models, group M1-1 . . 87

3.8 Values of the efficiency criterion parameters for the models, group M1-2 . . 88

3.9 Values of the efficiency criterion parameters for the models, group M2 . . . 88

4.1 Commonly used link functions (De Jong and Heller, 2008). 102

4.2 Types of the distribution of the responses and the corresponding link functions103

4.3 The performances of the missing data estimation models 104

5.1 List of the AREP in the tested statistical parameters 125

6.1 Data of the main reservoirs in the Ruhr River basin 140

6.2 Expected travel time values for different values of the downstream discharges (Q_D) 145

6.3 Cross correlation between the travel time (t) and Q_R, ΔQ_R, Q_D and ΔQ_D . 146

6.4 The number of membership functions corresponds to Q_R, ΔQ_R and Q_D in the ANFIS models 148

6.5 The number of neurons in the hidden layer for the BPNN models 155

6.6 The average values of the AREP for the MLR models 159

6.7 The average values of the AREP for the models A5 and C5 160

6.8 The main flow stations along the upper and middle reaches of the Ruhr River162

6.9 List of the flow at the release gauges and the flow from tributaries and the predicted travel time values for each scenario 163

Chapter 1

Introduction

1.1 Introduction

1.1.1 The Ruhr River basin

Introduction

The Ruhr is a right tributary of the Rhine in North Rhine-Westphalia, with a catchment area of 4485 km^2 and length of 219.50 km. An important economic significance of the Ruhr is now in the drinking water supply of the Ruhr district, as well as in power generation. The Ruhr is also a lifeline to industry and municipalities in North Rhine-Westphalia. The Ruhr valley is a major recreation area for the metropolitan area Rhein-Ruhr. The long-term average discharge of the Ruhr is 79 m^3/s at Mülheim near its mouth. The layout of the Ruhr River basin is displayed in figure 1.1. The figure shows also the main gauges and reservoirs in the basin.

1.1.2 Runoff and water demand

As mentioned before, the average flow rate of the Ruhr at the point of inflow into the Rhine is approximately 80 m^3/s. However, in dry periods, the Ruhr's flow rate can sink to 3.5 m^3/s. By contrast, the Ruhr can increase its flow to a maximum of about 2000 m^3/s during a flood. In contrast to some other districts in Germany, the public water supply in the Ruhr district is not taken directly from the reservoirs by pipelines, but indirectly transported by the Ruhr River to water works (Renz, 1983). The water discharged from reservoirs is used for a variety of purposes. The water demand can be covered from the natural runoff except in the dry periods in which the minimum runoff is guaranteed by water discharged from the reservoirs. The Ruhr River basin contains 15 reservoirs. Table 1.1 lists the data of the main reservoirs (Bigge, Henne, Möhne, Sorpe, Ennepe and Verse). The Ruhr River supplies not only the Ruhr River basin but also the adjacent basins (i.e. the drainage basins of the Emscher, Lippe and Wupper Rivers). About 50 % of the abstracted water is exported to neighboring catchments. Figure 1.2 gives the values of the exported and the total abstracted water from the Ruhr River basin in the time from 1900 to 2005.

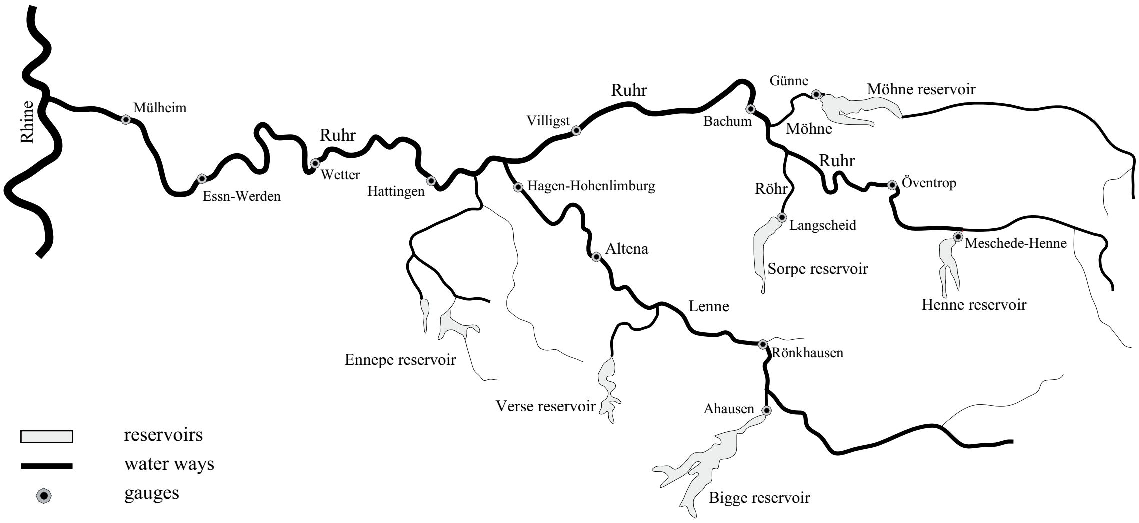


Figure 1.1: *Layout of the Ruhr River*

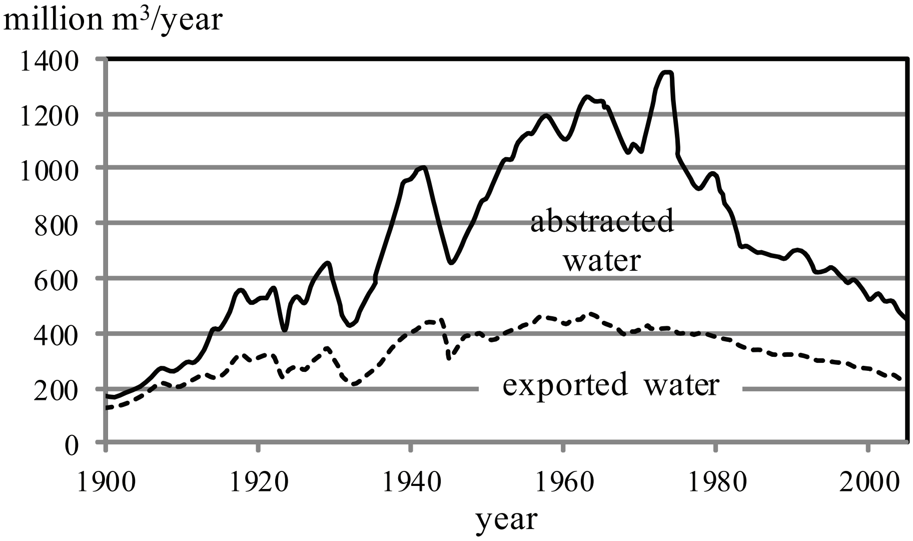


Figure 1.2: *Annual abstracted and exported water in the Ruhr catchment area between 1900 and 2005 (Ruhrverband, 2005)*

Table 1.1: *Data of the main reservoirs in the Ruhr River basin*

	Möhne	Henne	Sorpe	Bigge*	Ennepe	Verse	Total
Catchment area (km^2)	436.37	98.5	100.3	287.43	48.2	24.1	994.9
Gross storage capacity (million m^3)	134.5	38.4	70.37	171.7	12.6	32.8	460.37
Dead storage capacity (million m^3)	6.7	2	3.5	7.5	0	2	21.7
Net storage capacity (million m^3)	127.8	36.4	66.87	164.2	12.6	30.8	438.67
Mean annual inflow (million m^3) **	193	58.11	42.28	238.97	40.02	21.79	594.17

* Lister and Bigge reservoir system.

** see table 2.1 for more details.

1.2 Hydrology versus water resources management

Hydrology and water resources management have a strong impact on the other. Hydrology, economic and political are considered as inputs to water resources management. Comparing with the other factors, the relative importance of hydrology seems to be decreasing. In the face of this, hydrological information plays the most important role in water resources system design and management. Water resources projects need hydrological data for use in their planning, design, construction and optimization. Due to the growing demand for water, it is required to apply an integrated approach to water resources management, incorporating surface and underground water, including return flows and taking into consideration all potential uses: industrial, river navigation, irrigation, municipal and environmental. For this reason, to be able to manage water resources we must know in what quantity, quality and variation they are likely to be in the foreseeable future. The optimum results in the planning and management of water resources systems can be best achieved by an integrated cooperation between those involved in hydrology, water management and water use with those versed in economics, ecology and the social sciences.

1.3 Data driven modeling

1.3.1 Introduction

Different types of models are used in hydrology such as physical models, mathematical models, empirical models, etc. The area of empirical modeling received an important boost due to the availability of data and the development in the area of machine learning. Such models can be called data-driven models. Data-driven modeling (DDM) is based on the analysis of all the data characterizing the system under study (Solomatine and Ostfeld, 2008). A data-driven model of a system can be defined as the model which connects the system state variables (input, internal and output variables) with only a limited knowledge of the details about the "physical" behavior of the system. Hybrid models combine both of data-driven models and physical models. Artificial intelligence, data mining, machine learning, etc. have contributed to develop DDM. Statistical methods, artificial neural networks and fuzzy rule-based systems are the most popular methods used in data-driven modeling of hydrological systems.

1.3.2 Machine learning

Machine learning is the basis of data-driven modeling. It can be defined as the algorithm in which an unknown dependency between a system's inputs and its outputs is determined from the available data (Mitchell, 1998), see figure 1.3. The discovered dependency can be used to predict the system's outputs from the known input values. The learning tasks in data-driven modeling can be divided into the following four types (Solomatine, 2002):

Classification: to find a way of classifying unseen examples.

Association: to identify the association between variables characterizing the system.

Clustering: this process is used to classify objects into relatively larger and meaningful categories.

Regression: where the task constitutes of predicting a real value associated with an input data point.

The task of learning can be classified into two categories: supervised learning and unsupervised learning. Supervised learning requires a set of input-output data values. In contrast to supervised learning, in unsupervised learning there are no target outputs available.

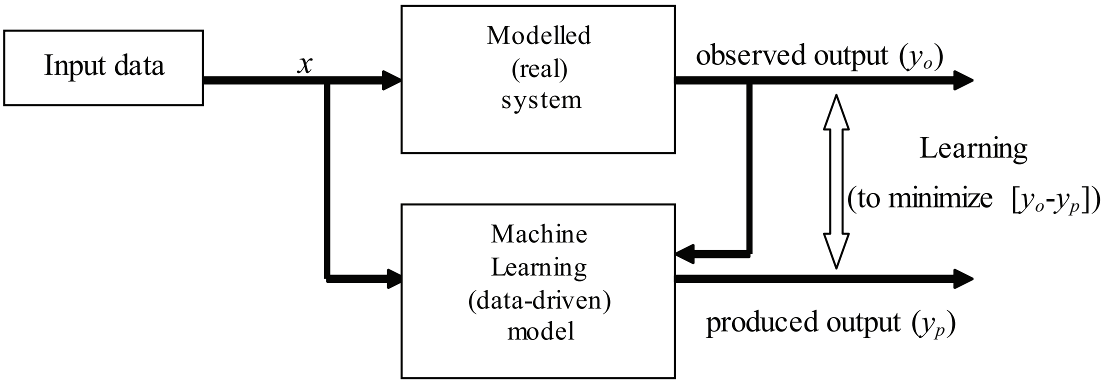


Figure 1.3: *Learning in data driven modeling (Solomatine, 2002)*

1.3.3 Data sets

The available data is usually split into three data sets (training, testing and validation data sets). These data sets should have identical statistical distributions to ensure that these three data sets come from the same population. The training data set is used to train the model (to determine the optimal parameters). The testing procedure is a procedure during which the predictive capability of the model is tested with the testing data set. The validation data set is used to validate the generalization ability of the model and to avoid the overfitting phenomenon. The generalization ability of the models and how the validation data set can be used to avoid the overfitting phenomenon are discussed in detail in chapter 3, section 3.3.4.

1.3.4 Popular data-driven methods and typical application

Data-driven models have proven their applicability to various problems related to river basin management: modeling short-term forecasting and classification of hydrology-related data (Solomatine and Ostfeld, 2008). They gained more popularity in the last decade due to the following advantages (Wang, 2006):

- They can represent arbitrarily complex processes based on mathematical criteria.
- They are easy to apply for different conditions because the modeling and forecasting procedure is usually analogous.
- The analysis of the structure and parameters of data-driven models can sometimes provide useful information on the dynamics of phenomenon of interest.

In the following, some of the popular data-driven approaches are presented:

Regression model

Regression analysis was developed to detect the presence of a mathematical relation between two or more variables subject to random variation. Regression analysis is widely used for prediction and forecasting.

Time series model

Time series analysis became a major tool in different applications in hydrology and water management fields. Time series models can be divided into two sets according to the number of time series in the model as follows:

i) Univariate time series models.

ii) Multivariate time series models.

They are used for building mathematical models to generate synthetic hydrologic records, to forecast hydrologic events, to detect trends and other changes in hydrologic records and to fill in missing data and extend records.

Artificial neural network model

Artificial Neural Network (ANN) is an information processing system that is inspired by the way the brain processes information. Mathematically, ANN can be defined as a complex nonlinear function with many parameters that are adjusted (trained) in such a way that the ANN output becomes similar to the measured output on a known data set.

Fuzzy logic model

Fuzzy logic model (FL) was originally identified by Zadeh (1965). Fuzzy logic is applied in system control and analysis design, because it shortens the time for engineering development. In the case of highly complex systems, fuzzy logic is sometimes the only way to solve the problem.

1.4 Objectives of the research

The objectives of this research are:

1. Investigate the inflow processes of the reservoirs Bigge, Henne, Möhne and Sorpe for trend, seasonality, stationarity and long memory for different timescales (e.g., daily time series; 10-days time series; ...). This is very important to understand the inflow processes, also it is an important task in hydrological modelling.

2. Forecast one-day and two-days ahead daily inflow for short-term operation of the reservoirs.

3. Fill the missing data in the daily inflow time series and extend the daily inflow time series of the Bigge reservoir in the period from 1/11/1960 to 31/10/1965.

4. Generate the monthly inflow process which plays an important role in the optimal operation of the reservoirs. As an application, the generated consecutive 5 years with minimum mean inflow are detected and can be used as an inflow scenario for optimal operation of the reservoirs during the dry periods.

5. Predict travel time to add the users in approximating the time that release from reservoirs (water) may become available to them.

1.5 Outline of the thesis

This thesis consists of seven chapters including the introduction as chapter 1 (the present chapter). The other six chapters can be summarized as follows:

- Chapter 2 discusses the stochastic properties of the inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs.

- In chapter 3, BPNN, ANFIS, ARMA and ARFIMA are used to forecast the daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs.

- In chapter 4, BPNN, ANFIS and GLM are applied for filling in the missing values in the inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs. The best model is used to extend the monthly inflow time series of the Bigge reservoir.

- In chapter 5, we use T-F, Gamma T-F, MC and PHMM models to generate the monthly inflow data into the Bigge, Henne, Möhne and Sorpe reservoirs.

- Chapter 6 discusses the applicability of using ANFIS, BPNN and MLR models for predicting the travel time of reservoirs releases along the Ruhr and Lenne Rivers.
- Chapter 7 summarizes the most relevant conclusions of this dissertation. Future work is also identified.

Each chapter is self-sustained and includes its own references and conclusions.

Four Matlab/GUI (graphical user interface) based simulation tools were developed to implement stochastic analysis and simulate the different models as follows (see appendix A, figures A.1 through A.11):

Vorhersage GUI
To implement the stochastic analysis of the time series and to simulate T-F, Gamma T-F, MC and PHMM models for monthly inflow generation.

Fehlende Daten GUI
To simulate BPNN, ANFIS and GLM models for filling in missing values.

Zuflussprognose GUI
To simulate BPNN, ANFIS, ARMA and ARFIMA models for daily inflow forecasting.

Fliesszeit GUI
To detect the changes in reservoir releases and to simulate ANFIS, BPNN and MLR models for travel time prediction.

1.6 References

Mitchell, T.M., 1998. Machine Learning. McGraw-Hill, New York.

Renz, F.W., 1983. Goals and management of the Ruhr reservoir system since the beginning of our century. IAHS Publication, 147, 637-647.

Ruhrverband, 2005. Jahresbericht Ruhrwassermenge. Ruhrverband, Essen, Deutschland.

Solomatine, D.P., 2002. Data-driven modeling: paradigm; methods, experiences. Proceedings of the 5th International Conference on Hydroinformatics, Carrdiff, UK, 1-5.

Solomatine, D.P., Ostfeld, A., 2008. Data-driven modeling: some past experiences and new approaches. Journal of Hydroinformatics, 10(1), 3-22.

Wang, W., 2006. Stochasticity, nonlinearity and forecasting of streamflow processes. IOS Press, Amsterdam.

Zadeh, L.A., 1965. Fuzzy Sets. Information and Control, 8(3), 338-353.

Chapter 2

Stochastic properties of reservoirs inflow processes

2.1 Introduction

The objective of the stochastic analysis of streamflow processes is to identify and analyze the different components of a given time series. To do this, it is necessary to conduct the following analyses:

- Trend analysis.
- Stationarity analysis.
- Seasonality/Periodicity analysis.
- Long memory analysis.

In this chapter, we investigated the stochastic characteristics of the inflow processes of the Bigge, Henne, Möhne and Sorpe reservoirs for different timescales (e.g., daily time series; 10-days time series; ...)

2.2 Data used

The historical records of the daily inflow of the Bigge (1/11/1966 - 31/10/2006), Henne (1/11/1960 - 31/10/2006), Möhne (1/11/1960 - 31/10/2008) and Sorpe (1/11/1960 - 31/10/2006) reservoirs are used in the present study. The daily inflow time series are compiled in the water year system, which covers the period from 1 November to 31 October. Figures 2.1.a, b, c and d show the plots of the daily inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs respectively.

2.3 Seasonality analysis

Many time series in ecology, hydrology, economic, etc., display seasonality (periodic fluctuation). Hydrological time series often exhibit annual variation. The season may denote a day, a 10-days, a month, etc. In the present study, the available daily inflow data are aggregated to 10-days (36 seasons), monthly (12 seasons), 3-months (4 seasons), 6-months (2 seasons) and annual series by taking the average of the inflow during each timescale. For 10-days inflow time series, all months are assumed to have a length of 30 days (the 31th day is neglected if it exists) and the assumed inflow during the days 29 and 30 February are assumed to be equal to the inflow during the previous day (28 February). Table 2.1 gives the statistical characteristics (mean, $\bar{x}$, standard deviation, SD, skewness, g and lag one autocorrelation function, ACF) for the different timescales for each reservoir.

The following approach is used for the analysis of seasonality in the daily inflow time series, Z, of length, N (e.g., Mitosek, 2000; Martins et al., 2008):

1. Delete the day 29 Feb. of the leap year.
2. Fill in the missing values in the daily inflow time series (see chapter 4).
3. Rewrite the daily inflow time series, Z*, in a matrix form as given by matrix* X *in equation (2.1). This matrix contains 365 columns, each column presents one day and* n *rows (years), each row represents a year (according to matrix* X*,* $n = N/365$*).*

$$X = \begin{vmatrix} x_{1,1} & x_{1,2} & x_{1,i} & x_{1,365} \\ x_{2,1} & x_{2,2} & x_{2,i} & x_{2,365} \\ x_{j,1} & x_{j,2} & x_{j,i} & x_{j,365} \\ x_{n,1} & x_{n,2} & x_{n,i} & x_{n,365} \end{vmatrix} \tag{2.1}$$

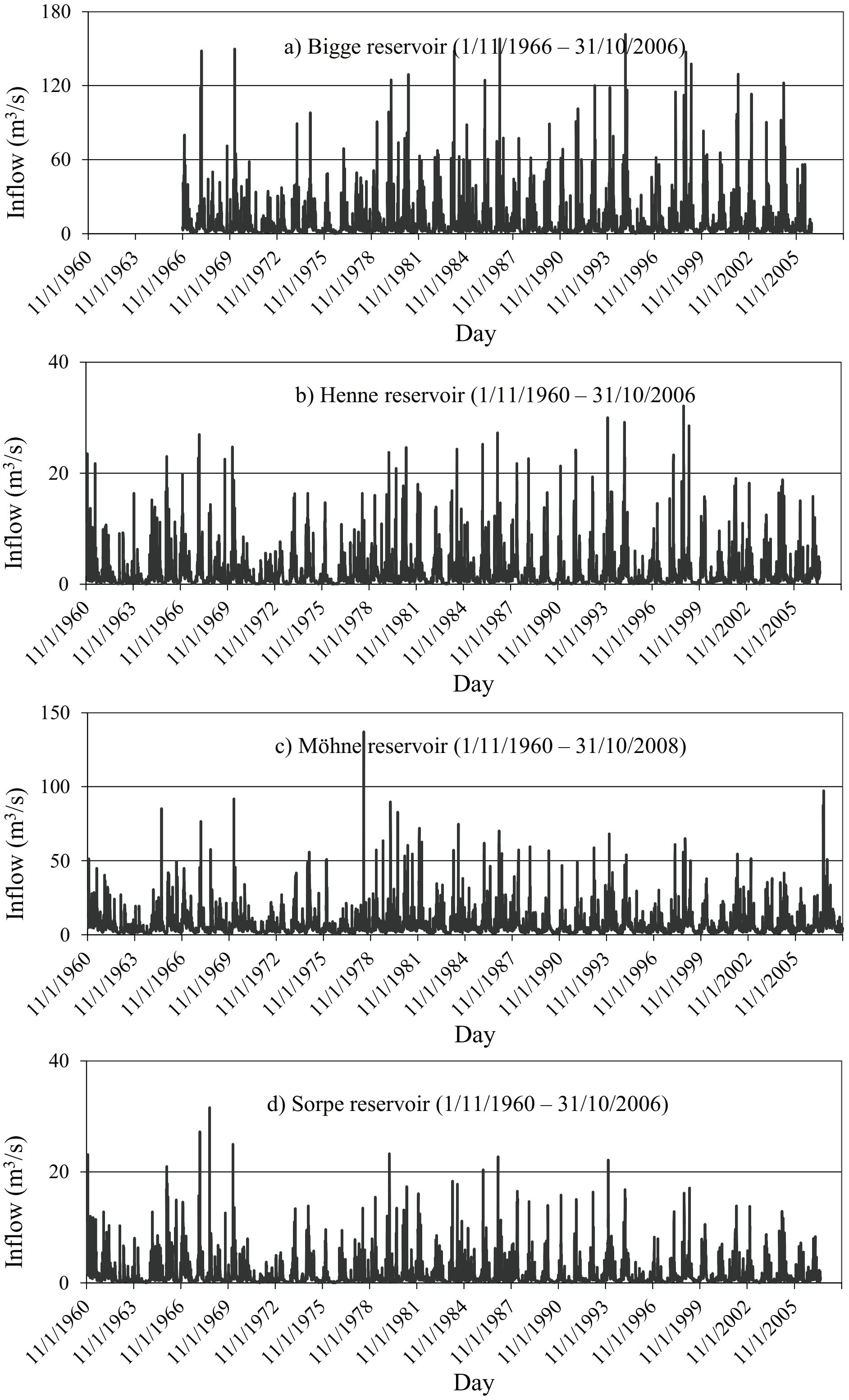


Figure 2.1: *Average daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs*

Table 2.1: *Statistical characteristics of the raw inflow time series for different timescales*

Reservoir	Period	Timescale	$\bar{x}$	SD	g	ACF
			m^3/s	m^3/s		lag 1
Bigge	1/11/1966 - 31/10/2006	daily	7.571	11.465	4.421	0.839
		10-days	7.569	8.706	2.218	0.445
		monthly	7.587	6.649	1.301	0.424
		3-months	7.592	5.156	0.685	0.057
		6-months	7.602	4.538	0.274	-0.671
		annual	7.571	1.766	-0.284	0.09
Henne	1/11/1960 - 31/10/2006	daily	1.833	2.629	3.589	0.906
		10-days	1.833	2.158	2.209	0.52
		monthly	1.837	1.696	1.363	0.392
		3-months	1.84	1.295	0.595	0.052
		6-months	1.841	1.136	0.333	-0.609
		annual	1.833	0.477	-0.25	0.133
Möhne	1/11/1960 - 31/10/2008	daily	6.149	6.995	3.975	0.87
		10-days	6.153	5.654	2.114	0.563
		monthly	6.161	4.609	1.313	0.433
		3-months	6.17	3.587	0.588	0.131
		6-months	6.167	3.091	0.278	-0.417
		annual	6.149	1.552	-0.197	0.182
Sorpe	1/11/1960 - 31/10/2006	daily	1.344	1.968	3.801	0.894
		10-days	1.345	1.623	2.267	0.546
		monthly	1.348	1.296	1.587	0.437
		3-months	1.349	1.016	0.873	0.121
		6-months	1.35	0.906	0.606	-0.463
		annual	1.344	0.462	0.445	0.331

4. *The mean ($\bar{x}_i$), standard deviation (SD_i), the coefficient of variation (CV_i) and the skewness (g_i) for each i^{th} column (day), can be calculated as follows:*

   Mean:

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{j,i} \tag{2.2}$$

   Standard deviation:

$$SD_i = \left(\frac{1}{n}\sum_{j=1}^{n} x_{j,i}^2 - \bar{x}_i^2\right)^{1/2} \tag{2.3}$$

   Coefficient of variation:

$$CV_i = \frac{SD_i}{\bar{x}_i} \tag{2.4}$$

   Skewness:

$$g_i = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_{j,i} - \bar{x}_i)^3}{\left(\frac{1}{n}\sum_{j=1}^{n}(x_{j,i} - \bar{x}_i)^2\right)^{3/2}} \tag{2.5}$$

Given a time series $x_t = x_1, x_2, ..., x_N$, Box and Jenkins (1976) gave a formula to estimate the autocorrelation function (ACF) as follows:

$$\hat{\rho}(k) = c_k/c_o \tag{2.6}$$

where $k = 0, 1, 2, \ldots$ and $c_k = \frac{1}{N-k}\sum_{s}^{N-k}(x_s - \bar{x})(x_{s+k} - \bar{x})$ in which $\bar{x} = \frac{1}{N}\sum_{s}^{N} x_s$.

Taking into consideration the annual cyclicity, Mitosek (2000) used the following formula to estimate the autocorrelation function of the daily inflow (matrix X) between column (season) x_i and column (season) x_{i+k}, where $i = 1, 2, \ldots, 365$ and $k = 0, 1, 2, \ldots, k_{\max} \leq 365$):

$$\hat{\rho}_{i(k)} = \begin{cases} \frac{\frac{1}{n}\sum_{j=1}^{n}(x_{j,i}-\bar{x}_i)(x_{j,i+k}-\bar{x}_{i+k})}{SD_i SD_{i+k}}, & \text{for } i+k \leq 365 \\ \\ \frac{\frac{1}{n-1}\sum_{j=1}^{n-1}(x_{j,i}-\bar{x}_i)(x_{j+1,i+k-365}-\bar{x}_{i+k-365})}{SD_i SD_{i+k-365}}, & \text{for } i+k{>}365 \end{cases} \tag{2.7}$$

where

$$\begin{aligned} \bar{x}_i &= \frac{1}{n}\sum_{j=1}^{n} x_{j,i}, \\ SD_i &= \left(\frac{1}{n}\sum_{j=1}^{n} x_{j,i}^2 - \bar{x}_i^2\right)^{1/2}, \\ \bar{x}_{i+k-365} &= \frac{1}{n-1}\sum_{j=1}^{n} x_{j+1,i+k-365}, \end{aligned}$$

and

$$SD_{i+k-365} = \left(\frac{1}{n-1}\sum_{j=1}^{n-1} x_{j+1,i+k-365}^2 - \bar{x}_{i+k-365}^2\right)^{1/2}$$

The previous procedure is applied also to estimate the seasonal statistical parameters of the 10-days, monthly, 3-months and 6-months inflow time series. Seasonal autocorrelation coefficients are estimated after log-transforming and deseasonalizing the raw time series (see section, 2.4).

2.3.1 Results of the seasonality analysis

Figures 2.2, 2.5 and 2.8 (a, b, c and d) show the variation in the mean, standard deviation and coefficient of variation of the daily, 10-days and monthly inflow time series respectively. From these figures, it is obvious that all of these time series (daily, 10-days and monthly) have a clear seasonality in the mean and standard deviation. Higher values of the mean and standard deviation occur at the same time. For all time series, the coefficient of variation has higher values in the dry periods. The seasonally variations of the skewness coefficient are shown in figures 2.3, 2.6 and 2.9 (a, b, c and d). It is clear that for all time series the higher values of the skewness occur in the seasons with low flows and vice versa. The figures also show that all time series have positive skewness, which indicates a tendency for low flow years to outnumber high flow years. The autocorrelation coefficients of the daily, 10-days and monthly inflow time series are estimated at different lags and the results are plotted in figures 2.4, 2.7 and 2.10 respectively. These results indicate that the autocorrelation coefficients are low for high inflow seasons and high for seasons with low inflow for all inflow time series except for daily inflow time series at lag of one day. For daily inflow time series the lag one day-to-day autocorrelation coefficients are high for all days (seasons).

2.4 Normalization and standardization

The results given in table 2.1, show that all seasonal inflow time series are generally positively skewed (g $>$ 0). We used log-transformation to normalize the inflow time series by taking its logarithm. The log-transformed inflow time series are approximately normal. The seasonality in the mean ($\bar{x}_i$) and standard deviation (SD_i) which is also known as standardization are removed as follows:

$$x_{j,i} = \frac{x_{j,i} - \bar{x}_i}{SD_i} \tag{2.8}$$

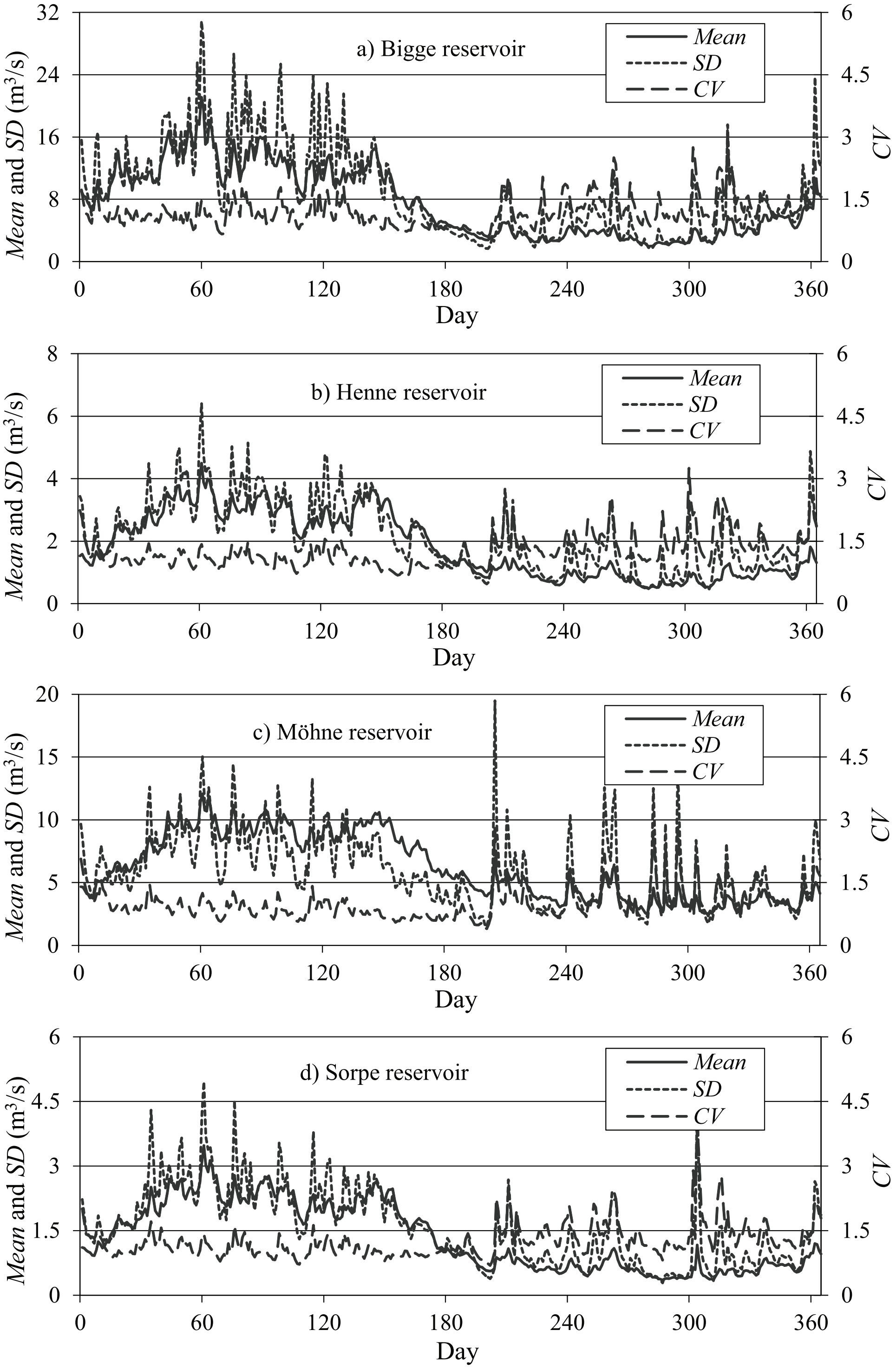


Figure 2.2: *Seasonal variation in the mean ($\bar{x}$), standard deviation (SD) and coefficient of variation (CV) of the daily inflow time series*

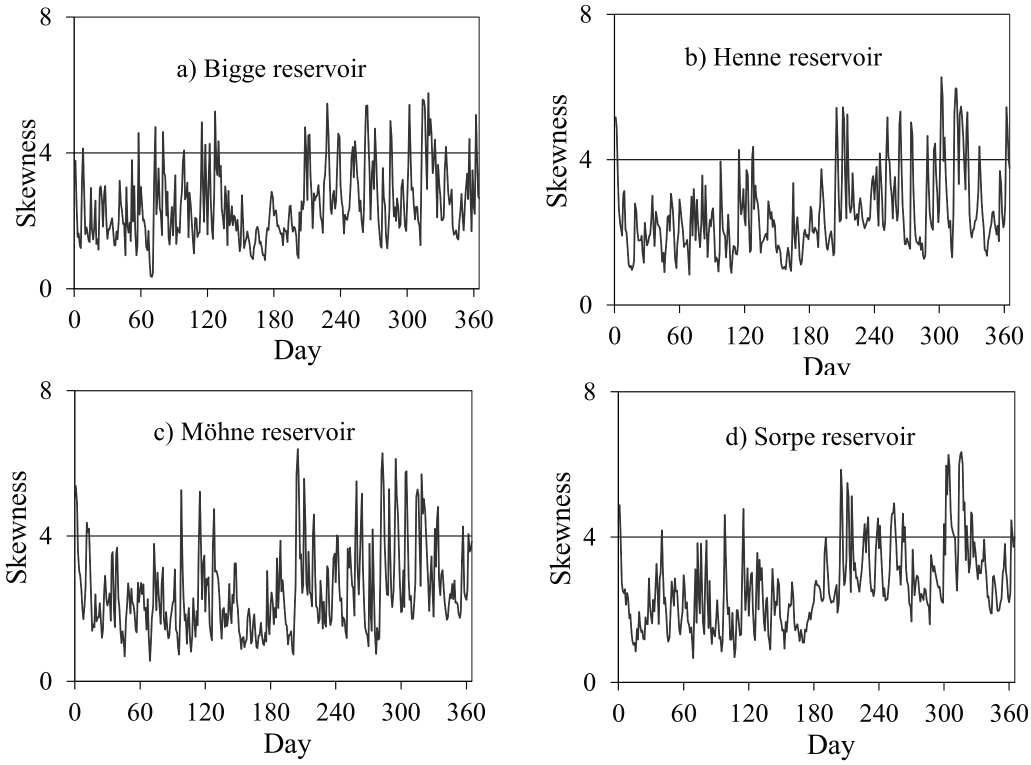


Figure 2.3: *Seasonal variation in the skewness (g) of the daily inflow time series*

2.5 Testing for trend

For a series of observations of a random variable a trend analysis is necessary to determine if their values generally increase or decrease with time. In statistical terms this is a determination of whether the probability distribution from which they arise has changed over time (Helsel and Hirsch, 1992). We used three different test methods to determine the presence of trends in the 10-days, monthly, 3-months, 6-months and annual inflow time series to detect the effect of the climate change and human activities. The linear regression and Mann-Kendall methods are used to detect the presence of trend in each season in the 10-days, monthly, 3-months and 6-months inflow time series also to detect the presence of trend in annual time series. Moreover, 10-days, monthly, 3-months and 6-months inflow time series are tested against the existence of trend using the seasonal Mann-Kendall test.

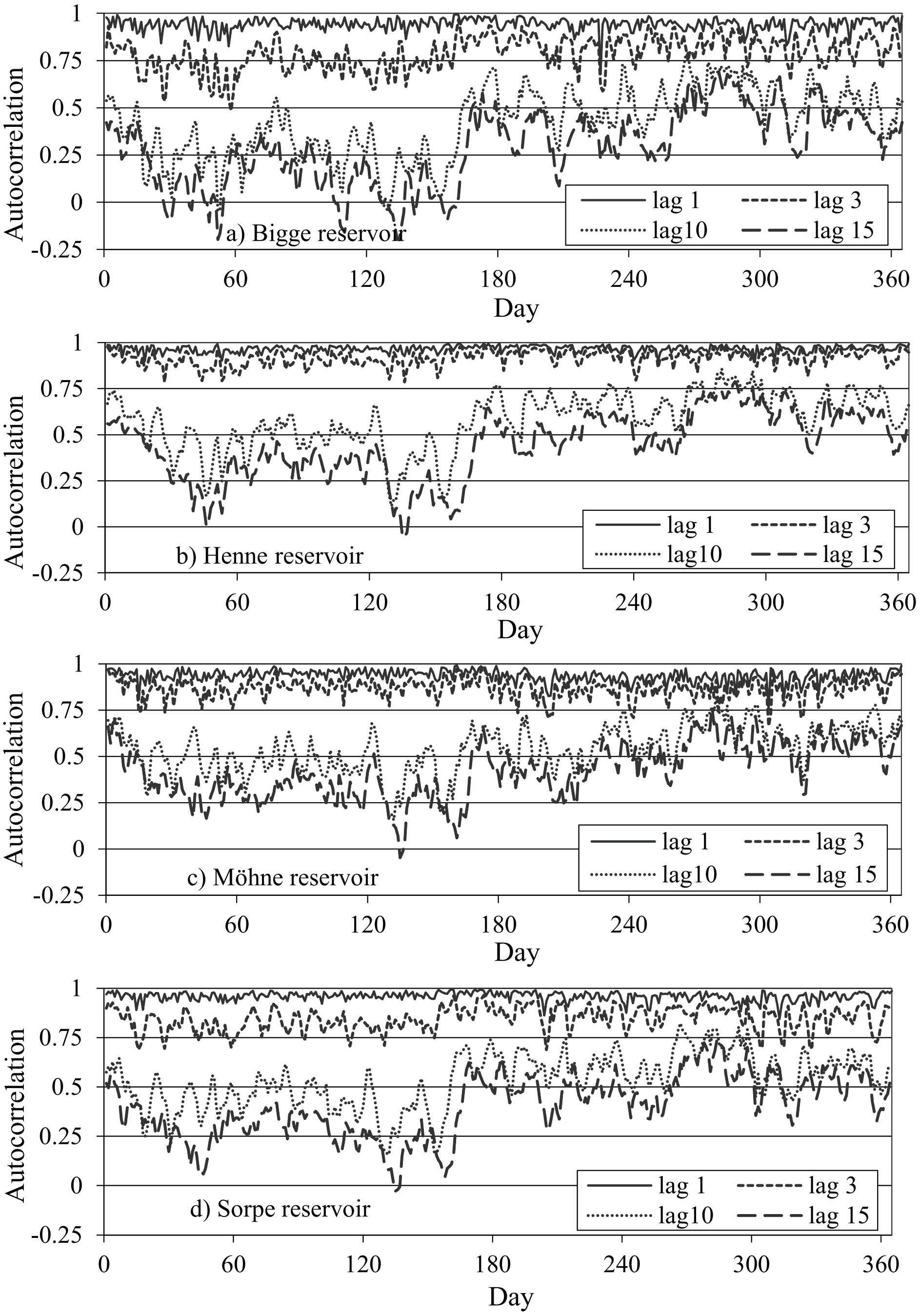


Figure 2.4: *Seasonal variation in the season-to-season autocorrelation coefficient of the daily inflow time series at different lags*

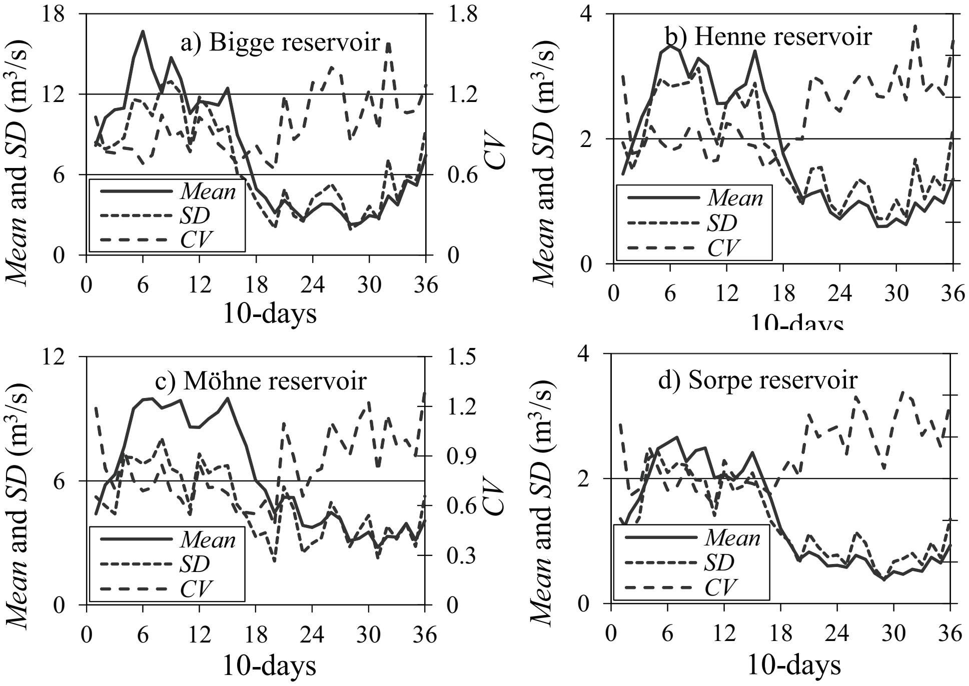


Figure 2.5: *Seasonal variation in the $\bar{x}$, standard deviation (SD) and coefficient of variation (CV) of the 10-days inflow time series*

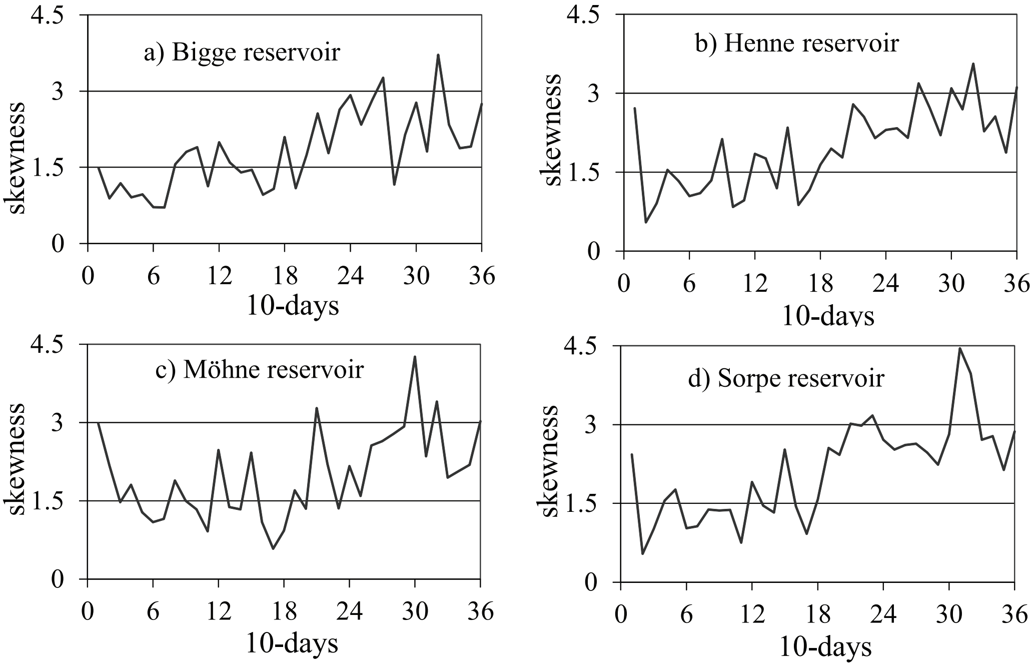


Figure 2.6: *Seasonal variation in the skewness (g) of the 10-days inflow time series*

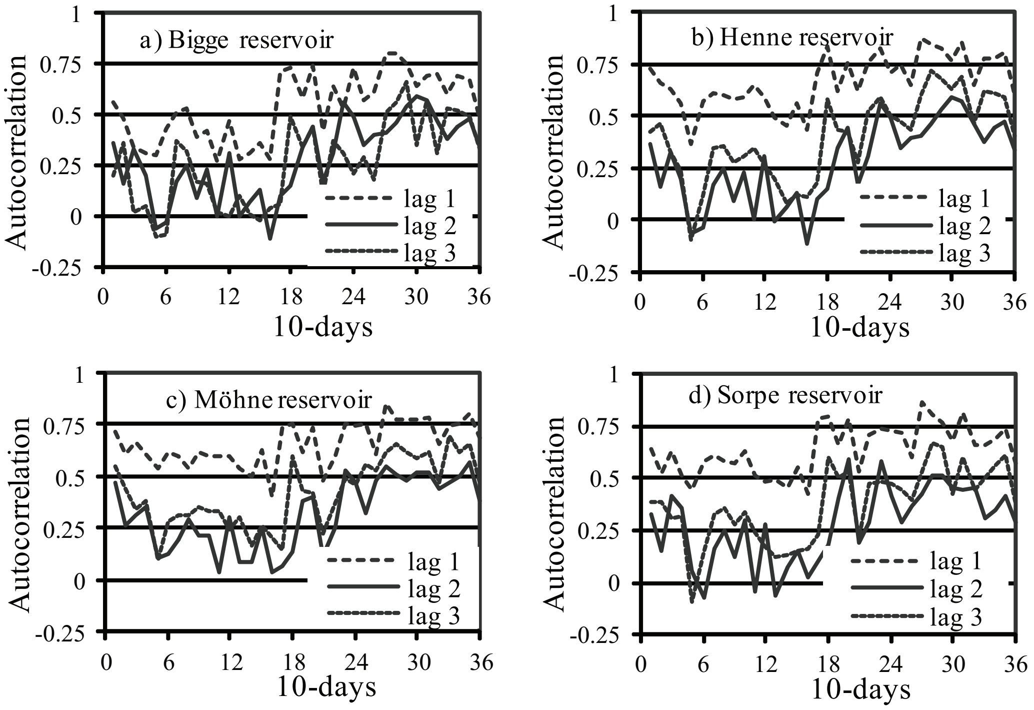


Figure 2.7: *Seasonal variation in the season-to-season autocorrelation coefficient of the 10-days inflow time series at different lags*

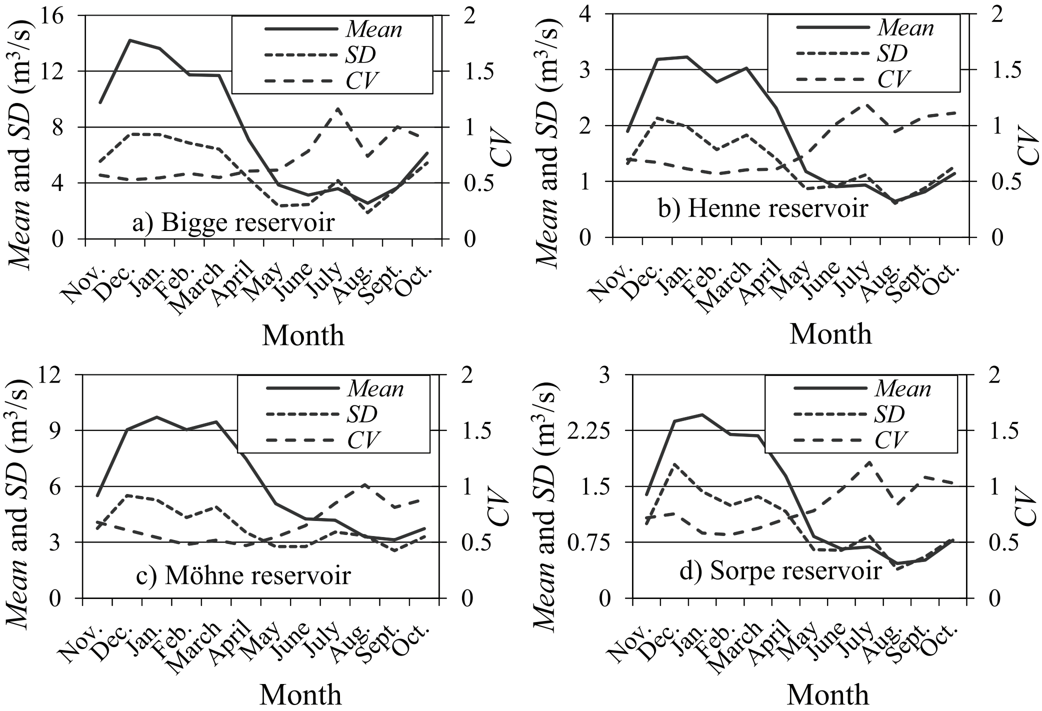


Figure 2.8: *Seasonal variation in the $\bar{x}$, standard deviation (SD) and coefficient of variation (CV) of the monthly inflow time*

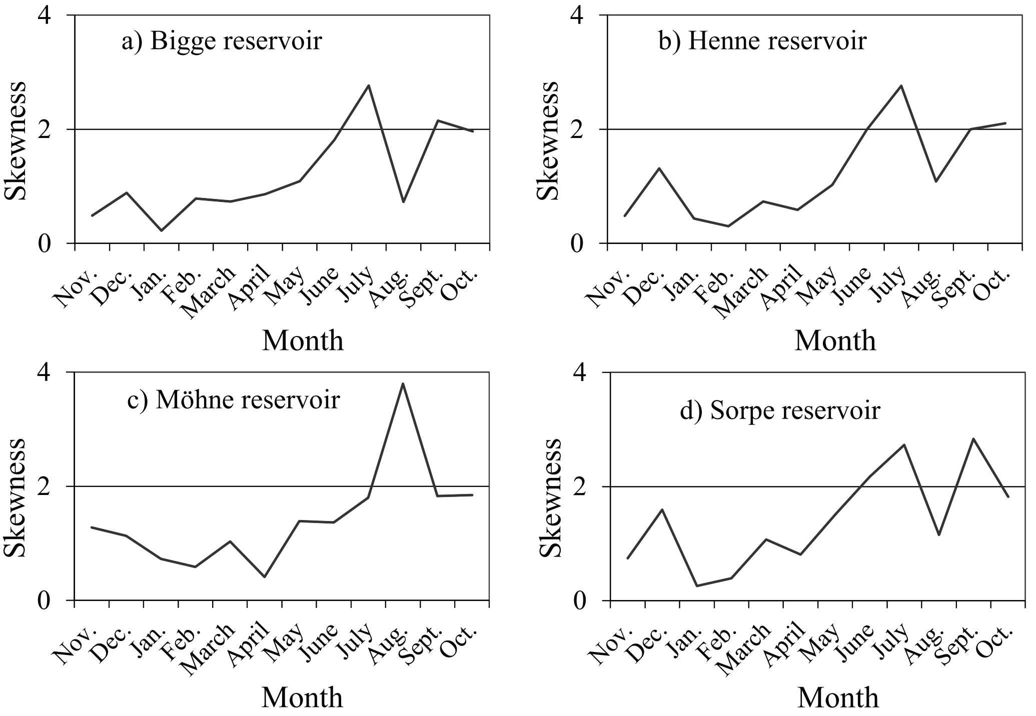


Figure 2.9: Seasonal variation in the skewness (g) of the monthly inflow time series

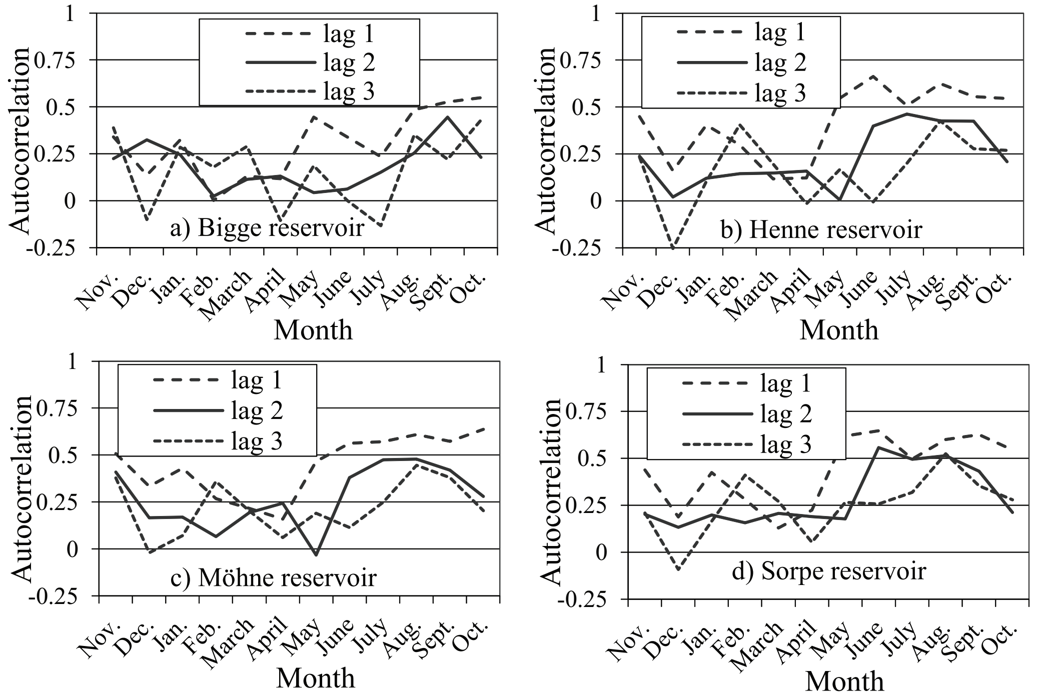


Figure 2.10: Seasonal variation in the season-to-season autocorrelation coefficient of the monthly inflow time series at different lags

2.5.1 Linear regression method

Maidment (1993) introduced a procedure to detect the linear trend in time series. Given a time series x_t, $t = 1, 2, \ldots, N$ and N is the sample size, he presents the following equation to introduce a simple linear trend in x_t,

$$x_t = a + bt \tag{2.9}$$

where a and b are the parameters of the regression model. The hypothesis H_0 (there is no trend) is rejected if

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right| > T_{1-\alpha/2,\upsilon} \tag{2.10}$$

in which r is the cross-correlation coefficient between the sequences $x_1, x_2, \ldots, x_N$ and $1, 2, \ldots, N$ and $T_{1-\alpha/2,\nu}$ is the $1-\alpha/2$ quantile of the student distribution with $\nu = N-2$ degrees of freedom.

2.5.2 Mann-Kendall test

The Mann-Kendall test was originally devised by Mann (1945) as a non-parametric test for trend. Kendall (1975) derived the exact distribution of the test statistic. This test can be used for any form of the distribution function of the data. To compute the Mann-Kendall test statistic S for a time series $x_t, t = 1, 2, \ldots, N$, where N is the sample size, each value $x_i, i = 1, ..., j-1$ is compared with all the subsequent values $x_j, j = 2, ..., N$ as follows:

$$S = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} sgn(x_j - x_i) \tag{2.11}$$

where N is the length of the data set and

$$sgn(x_j - x_i) = \begin{cases} 1 & \text{if} \quad x_j > x_i \\ 0 & \text{if} \quad x_j = x_i \\ -1 & \text{if} \quad x_j < x_i \end{cases} \tag{2.12}$$

Kendall (1975) documented that the statistic (S) is approximately normally distributed for $N \geq 8$. He gave the following formulas to estimate the mean $E(S)$ and the variance $Var(S)$:

$$E(S) = 0 \tag{2.13}$$

$$Var(S) = \frac{N(N-1)(2N+5) - \sum_{p=1}^{n_t} t_p(t_p-1)(2t_p+5)}{18} \tag{2.14}$$

where

t_p = the number of data points in the p^{th} group.

n_t = the number of tied groups (a tied group is a set of sample data having the same value).

The standardized Mann-Kendall test statistic $Z_{MK}$is computed by

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sqrt{Var(S)}} & S < 0 \end{cases} \tag{2.15}$$

The statistic Z_{MK} follows the standard normal distribution with mean of zero and variance of one. The hypothesis H_0 that there is no trend is rejected if

$$|\mathrm{Z}_{MK}| > \mathrm{Z}_{1-\alpha/2} \tag{2.16}$$

where $\mathrm{Z}_{1-\alpha/2}$ is the value read from a standard normal distribution and α is the significance level of the test.

Mann-Kendall test for autocorrelated time series

The presence of positive serial correlation (autocorrelation) inflates the variance $Var(S)$ of the Mann-Kendall statistic (e.g., Hamed and Rao, 1998; Wang, 2006; Martins et al., 2008). Increasing $Var(S)$ will increase the possibility of rejecting the null hypothesis which indicates significant trend (actually there is no trend). On the other hand, the presence of the negative autocorrelation decreases $Var(S)$ and hence decreasing the possibility of rejecting the null hypothesis.

Pre-whitening is used by many researchers to transform the autocorrelated time series into an uncorrelated one (e.g., Fleming and Clarke, 2002; Yue et al., 2002). Donald et al. (2004) gave a modified version of the trend-free pre-whitening (TFPW) which was originally developed by Yue et al. (2002). The modified TFPW is given in the following steps for applying the Mann-Kendall test:

1. Estimate the Mann-Kendall statistic (S) and the monotonic trend (B)

$$B = \text{Median}\left(\frac{x_i - x_j}{i - j}\right) \quad \text{for all } j{<}i, \text{in which } 1 < j < i < N \tag{2.17}$$

2. Remove the monotonic trend (B), using $y_t = x_t - B{\times}t,\ t = 1, 2, \ldots, N$.

3. Estimate the lag-one autocorrelation (r_1) of the de-trended series y_t.

4. If r_1 is not statistically significant at level 5% (the Ljung-Box Q-statistic lack-of-fit hypothesis test is used in the present study), complete the analysis using the results from step 1.

5. If r_1 is statistically significant at level 5%, the de-trended series y_t is pre-whitened through:

$$y_t' = y_t - r_1 y_{t-1} \tag{2.18}$$

6. Add B to the residual series through:

$$y_t'' = y_t' + Bt \tag{2.19}$$

7. Calculate the Mann-Kendall statistic for the series y_t''.

2.5.3 Seasonal Mann-Kendall trend test

Presence of seasonal cycles in time series required special tests for trend. One of these tests is the seasonal Mann-Kendall test which was developed by Hirsch et al. (1982). This test may be used when missing data or tied are present in the time series. The validity of the test does not depend on the data being normally distributed. To use the seasonal Mann-Kendall test, compute the Mann-Kendall test statistic S (equation, 2.11) for each season separately. The estimated values of the Mann-Kendall test statistic are then used to compute the overall statistic (S') as flows:

$$S' = \sum_{j=1}^{p_s} S_j \tag{2.20}$$

where S_j is the Mann-Kendall statistic of the season j ($j = 1, 2, 3, \ldots.., p_s$) and p_s is the number of seasons. The variance of S' can be estimated as:

$$Var(S') \;=\; \sum_{j=1}^{p_s} Var(S_j) \tag{2.21}$$

Hirsch and Slack (1984) used the seasonal Mann-Kendall test with autocorrelated time series and suggested using the following formula to estimate the variance of S':

$$Var(S') \;=\; \sum_{j=1}^{p_s} Var(S_j) + \sum_{g=1}^{p_s-1} \sum_{h=g+1}^{p_s} \sigma_{gh} \tag{2.22}$$

where σ_{gh} is the covariance between the Mann-Kendall test statistic for season g and that for season h. The covariance σ_{gh} can be estimated as in the following procedure:

Assume two matrices X and R which give the sequences of observations over p_s seasons for n years and ranks corresponding to the observations in X respectively.

$$X = \begin{vmatrix} x_{11} & x_{12} & x_{1j} & x_{1p_s} \\ x_{21} & x_{22} & x_{2j} & x_{2p_s} \\ x_{i1} & x_{i2} & x_{ij} & x_{ip_s} \\ x_{n1} & x_{n2} & x_{nj} & x_{np_s} \end{vmatrix} \qquad R = \begin{vmatrix} R_{11} & R_{12} & R_{1j} & R_{1p_s} \\ R_{21} & R_{22} & R_{2j} & R_{2p_s} \\ R_{i1} & R_{i2} & R_{ij} & R_{ip_s} \\ R_{n1} & R_{n2} & R_{nj} & R_{np_s} \end{vmatrix}$$

The ranking is performed among n observations within each season separately. For season j and year i the rank R_{ij} can be expressed as:

$$R_{ij} = \frac{n + 1 + \sum_{k=1}^{n} sgn(x_{ij} - x_{kj})}{2} \tag{2.23}$$

In the case of no missing values, σ_{gh} can be given as follows:

$$\sigma_{gh} = \frac{k_{gh} + 4\sum_{i=1}^{n} R_{ig} R_{ih} - n(n+1)^2}{3} \tag{2.24}$$

where

$$k_{gh} = \sum_{i=1}^{n-1} \sum_{o=i+1}^{n} sgn(x_{og} - x_{ig})(x_{oh} - x_{ih}) \tag{2.25}$$

Using equations (2.20) and (2.22) the standardized seasonal statistic $Z^{'}$ which follows the standard normal distribution with mean of zero and variance of one can be defined as follows:

$$Z' = \begin{cases} \frac{S'-1}{\sqrt{Var(S')}} & S' > 0 \\[2em] 0 & S' = 0 \\[2em] \frac{S'+1}{\sqrt{Var(S')}} & S' < 0 \end{cases} \tag{2.26}$$

2.5.4 Results of the linear regression and Mann-Kendall tests

The results of both of the linear regression and Mann-Kendall tests at 5 % significance level for the 10-days, monthly, 3-months, 6-months and annual inflow time series are given in tables 2.2, 2.3 and 2.4. The computed values of the linear regression test statistic (T_c) and the Mann-Kendall test statistic (Z_{MK}) which are within the range ±1.96 mean that the hypothesis H_0 cannot be rejected (there is no trend). The other values of T_c and Z_c which are out of the range mean that the hypothesis H_0 is rejected (presence of a trend). Table 2.3 shows that using the linear regression test the hypothesis H_0 is rejected for the months April and May for the Henne reservoir, June and July for the Möhne reservoir and April, May, June, July and August for the Sorpe reservoir. The results of the Mann-Kendall test show that the hypothesis H_0 is rejected for the months July for the Henne and Möhne reservoirs and April, June, July and August for the Sorpe reservoir as given in table 2.3. Using both of the linear regression and Mann-Kendall tests, a trend is detected in the 3-months inflow time series in the third season (months May, June and July) for the Henne, Möhne and Sorpe reservoirs (table 2.4). For the 6-months inflow time series (using both of the linear regression and Mann-Kendall tests) and the mean annual inflow time series (using the linear regression test) a downward trend is exhibited only for the Sorpe reservoir as given in table 2.4.

Table 2.2: *Values of the statistics T_c and Z_{MK} for the 10-days mean inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs*

Season	Bigge reservoir				Henne reservoir				Möhne reservoir				Sorpe reservoir			
	T_c	H_0	Z_{MK}	H_0	T_c	H_0	Z_{MK}	H_0	T_c	H0	Z_{MK}	H_0	T_c	H_0	Z_{MK}	H_0
1	1.382	0	1.55	0	1.202	0	1.382	0	0.904	0	0.791	0	0.316	0	0.852	0
2	0.277	0	0.291	0	0.361	0	0.492	0	1.823	0	0.809	0	-0.38	0	-0.36	0
3	-0.873	0	-0.804	0	-0.013	0	0.038	0	0.473	0	0.116	0	-0.922	0	-0.72	0
4	-0.977	0	-1.34	0	-1.919	0	-1.78	0	-1.669	0	-1.751	0	-2.555	1	-1.95	0
5	-0.095	0	0.618	0	-0.482	0	0.019	0	-1.002	0	-0.293	0	-1.49	0	-0.625	0
6	0.711	0	1.2	0	0.848	0	1.875	0	0.177	0	0.738	0	0.294	0	1.231	0
7	1.022	0	0.944	0	0.701	0	0.852	0	0.053	0	0.062	0	0	0	-0.038	0
8	-0.196	0	0.221	0	-0.173	0	0.436	0	-0.271	0	0.044	0	-0.416	0	-0.227	0
9	1.113	0	0.664	0	1.502	0	1.155	0	1.704	0	1.52	0	0.97	0	0.985	0
10	-0.704	0	-0.711	0	-0.079	0	0.17	0	0.034	0	0	0	-0.526	0	-0.095	0
11	2.739	1	1.666	0	0.723	0	0.53	0	0.948	0	1.04	0	0.244	0	0.492	0
12	0.087	0	-0.291	0	1.279	0	0.568	0	1.066	0	1.075	0	0.551	0	0.379	0
13	1.327	0	0.92	0	1.951	0	0.928	0	2.218	1	1.893	0	1.428	0	0.625	0
14	-0.001	0	0.384	0	0.486	0	0.739	0	0.813	0	0.898	0	-0.249	0	0.189	0
15	-0.074	0	0.198	0	0.395	0	0.985	0	1.127	0	1.929	0	-0.17	0	0	0
16	-1.809	0	-1.573	0	-2.197	1	-2.102	1	-1.275	0	-0.72	0	-3.067	1	-2.386	1
17	-1.022	0	-0.384	0	-1.379	0	-1.363	0	-0.489	0	-0.489	0	-2.041	1	-1.629	0
18	-0.486	0	0.641	0	-1.601	0	-0.871	0	-1.519	0	-1.022	0	-2.344	1	-1.136	0
19	0.724	0	1.247	0	-1.584	0	-1.004	0	-0.861	0	-0.649	0	-1.994	1	-0.814	0
20	-0.147	0	0	0	-2.817	1	-2.007	1	-2.306	1	-1.662	0	-2.656	1	-1.553	0
21	-0.074	0	-0.757	0	-0.659	0	-1.666	0	-1.075	0	-1.769	0	-0.623	0	-1.496	0
22	-0.525	0	-0.198	0	-1.159	0	-1.004	0	-0.985	0	-0.489	0	-1.169	0	-0.663	0
23	-1.599	0	-1.27	0	-1.504	0	-1.089	0	-1.825	0	-1.449	0	-2.396	1	-1.894	0
24	-1.47	0	-2.016	1	-2.135	1	-2.537	1	-2.391	1	-3.031	1	-1.962	1	-2.67	1
25	-2.42	1	-1.619	0	-1.974	1	-2.102	1	-2.504	1	-2.231	1	-2.099	1	-2.31	1
26	-0.574	0	-0.501	0	-1.503	0	-1.307	0	-2.449	1	-1.715	0	-1.759	0	-1.326	0
27	-1.069	0	-0.175	0	-1.794	0	-1.647	0	-2.461	1	-2.071	1	-2.451	1	-2.291	1
28	-0.468	0	-0.291	0	-0.261	0	-1.307	0	-0.071	0	-1.911	0	-1.271	0	-1.932	0
29	-1.507	0	-0.501	0	-0.251	0	-0.966	0	0.267	0	-1.182	0	-1.788	0	-1.704	0
30	-0.565	0	0.431	0	-1.68	0	-1.06	0	0.469	0	-0.773	0	-2.598	1	-1.553	0
31	-0.379	0	0.757	0	-1.307	0	-0.568	0	-0.661	0	-0.507	0	-1.966	1	-0.881	0
32	1.031	0	0.711	0	1.65	0	1.193	0	1.242	0	0.827	0	0.535	0	0.227	0
33	-0.131	0	0.524	0	2.009	1	1.78	0	1.36	0	0.595	0	0.57	0	0.644	0
34	-0.537	0	1.037	0	0.09	0	1.856	0	0.825	0	1.946	0	-0.607	0	1.231	0
35	-0.488	0	-0.012	0	0.496	0	1.269	0	0.451	0	0.702	0	-0.247	0	0.227	0
36	0.576	0	1.037	0	0.518	0	0.322	0	0.005	0	0.364	0	0.077	0	0.227	0

The results of the linear regression test and Mann-Kendall test for monthly and annual inflow time series for all reservoirs are visually examined using figures 2.11 through 2.14 for monthly inflow time series and figures 2.15 through 2.18 for annual inflow time series. Visual inspections of the monthly inflow time series show that the months April and May for the Henne reservoir, July for the Möhne reservoir and April, May and August for the Sorpe reservoir exhibit downward trend. Also, the annual inflow time series are visually inspected and no trend was detected except in the mean inflow time series of the Sorpe reservoir (figure 2.18.b) which exhibits a downward trend or shift (shift occurred in 1970). Mann-whitney test is used for testing shift and the null hypothesis of the test is rejected which means presence of shift. The shift is removed using the procedure introduced in Maidment (1993) and the mean inflow time series of the Sorpe reservoir is tested against trend and no trend is detected after removing the shift.

2.5.5 Results of the seasonal Mann-Kendall test

The computed values of the seasonal Mann-Kendall test statistic (Z') are given in table 2.5. The results show that the hypothesis H_0 is rejected only for all tested inflow time series of the Sorpe reservoir. All seasons in the 10-days, monthly, 3-months and 6-months inflow times of Sorpe reservoir are tested against shift and the detected shifts are removed. These seasonal time series are tested against trend after removing the detected shifts and the results are given also in table 2.5. It is clear that after removing the detected shifts no trend is detected except in the 10-days inflow time series.

Table 2.3: *Values of the statistics T_c and Z_{MK} for the monthly mean inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs*

Season	Bigge reservoir				Henne reservoir				Möhne reservoir				Sorpe reservoir			
	T_c	H_0	Z_{MK}	H_0	T_c	H0	Z_{MK}	H0	T_c	H0	Z_{MK}	H_0	T_c	H0	Z_{MK}	H_0
Nov.	0.385	0	0.524	0	0.713	0	0.625	0	1.381	0	0.827	0	-0.419	0	-0.341	0
Dec.	-0.029	0	0.618	0	-0.552	0	0.189	0	-1.026	0	-0.062	0	-1.548	0	-0.492	0
Jan.	1.034	0	1.317	0	1.08	0	1.079	0	0.625	0	0.773	0	0.265	0	0.246	0
Feb.	0.668	0	0.384	0	0.871	0	0.568	0	0.895	0	0.862	0	0.071	0	0.114	0
March	0.674	0	0.781	0	1.368	0	1.363	0	1.848	0	1.929	0	0.397	0	0.454	0
April	-1.445	0	-0.92	0	-2.155	1	-1.932	0	-1.258	0	-1.022	0	-3.097	1	-2.367	1
May	0.198	0	0.408	0	-2.119	1	-1.534	0	-1.715	0	-1.449	0	-2.23	1	-1.458	0
June	-1.614	0	-1.363	0	-1.796	0	-1.647	0	-1.972	1	-1.484	0	-2.318	1	-2.159	1
July	-1.479	0	-0.897	0	-1.941	0	-2.178	1	-2.93	1	-2.231	1	-2.312	1	-2.575	1
Aug.	-1.166	0	-0.548	0	-1.188	0	-1.231	0	0.288	0	-1.431	0	-2.845	1	-2.178	1
Sept.	0.521	0	0.711	0	1.436	0	1.231	0	0.982	0	0.649	0	-0.342	0	0.151	0
Oct.	0.003	0	0.408	0	0.497	0	0.776	0	0.443	0	0.755	0	0.497	0	0.17	0

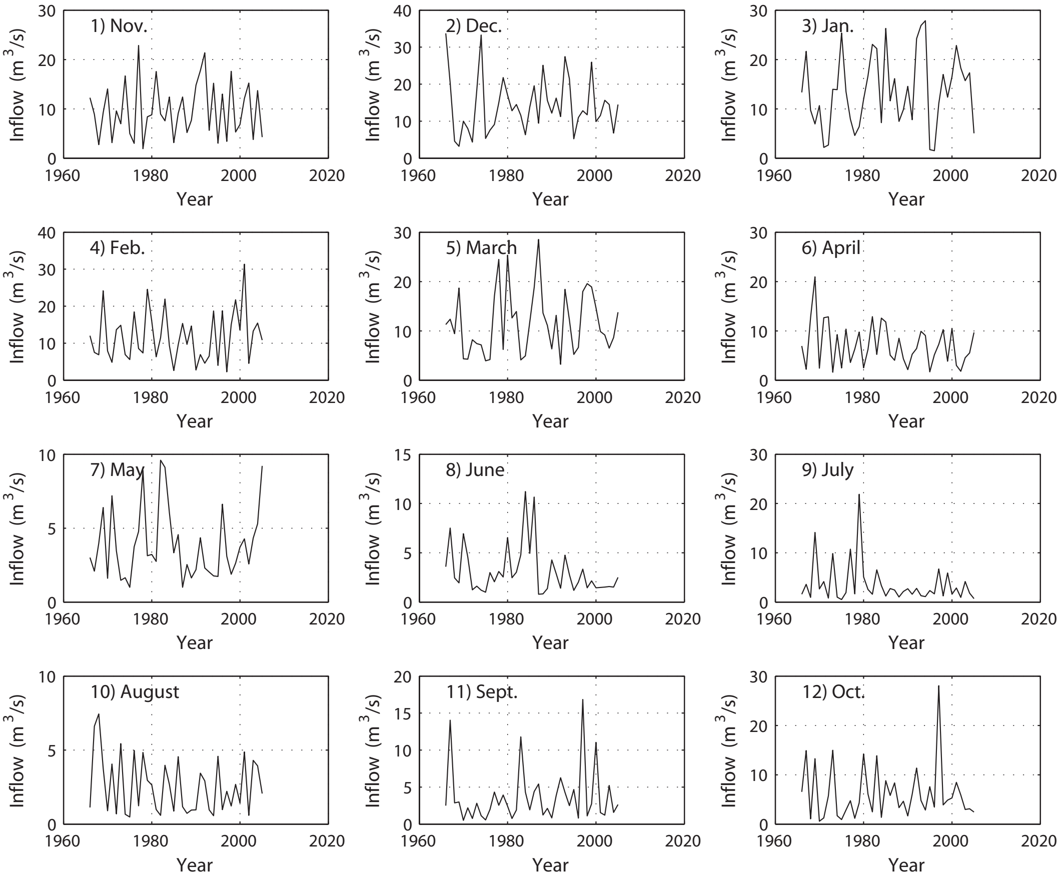


Figure 2.11: *Monthly inflow into the Bigge reservoir (Nov. 1966 - Oct. 2006)*

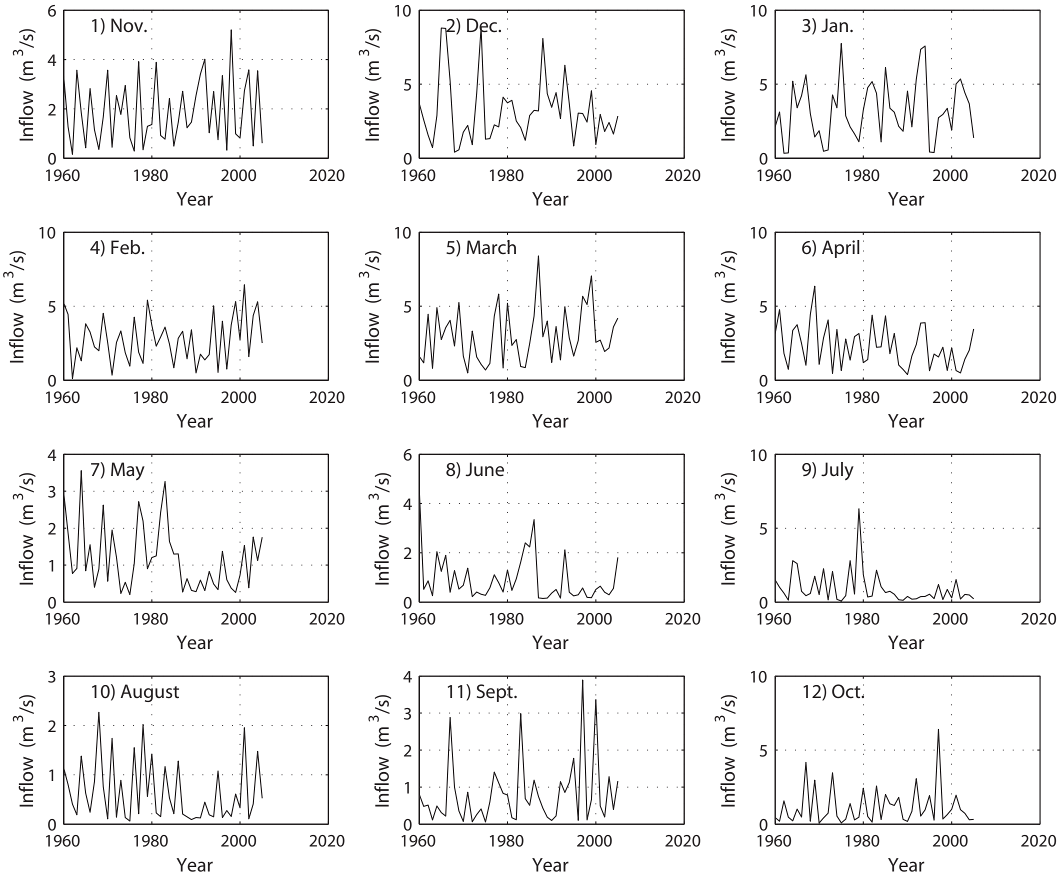


__Figure 2.12:__ Monthly inflow into the Henne reservoir (Nov. 1960 - Oct. 2006)

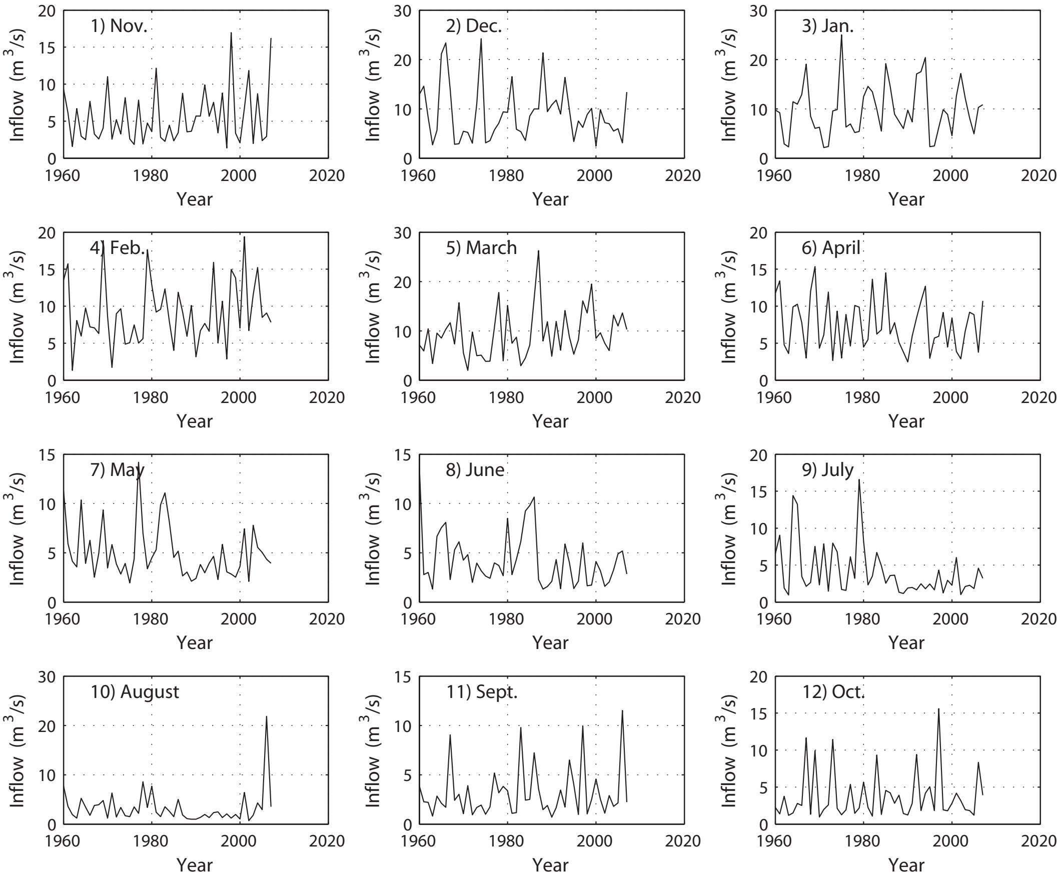


Figure 2.13: *Monthly inflow into the Möhne reservoir (Nov. 1960 - Oct. 2008)*

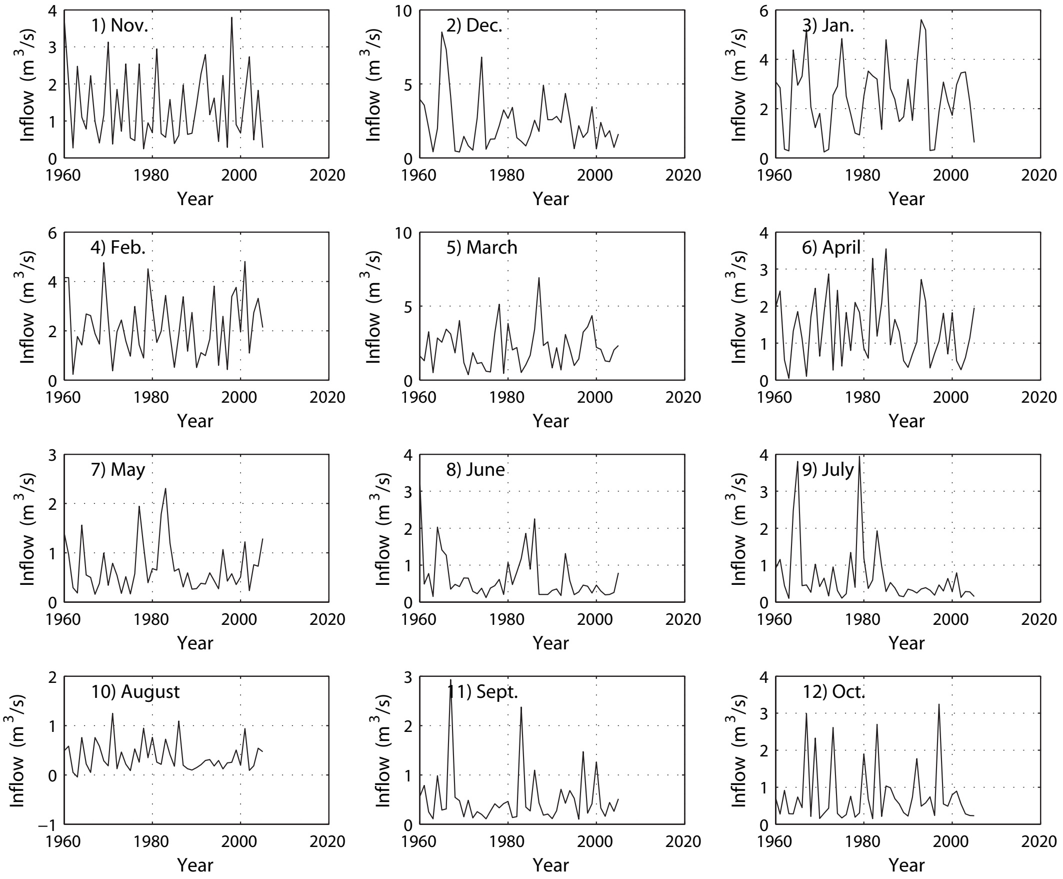


Figure 2.14: *Monthly inflow into the Sorpe reservoir (Nov. 1960 - Oct. 2006)*

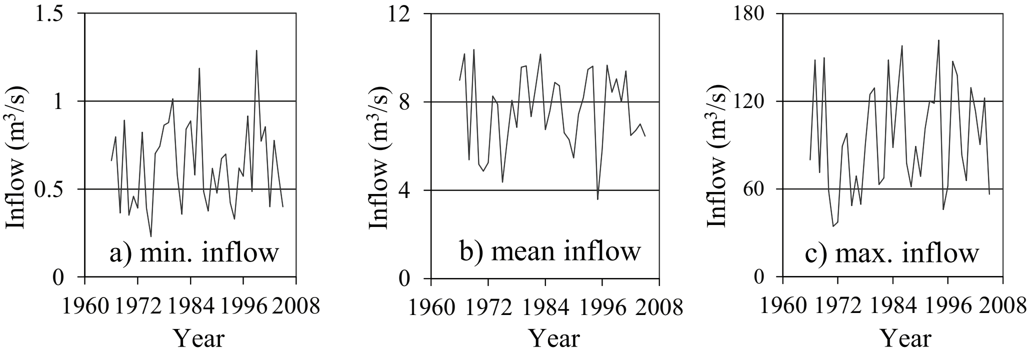


Figure 2.15: *Annual inflow into the Bigge reservoir (1966 - 2006)*

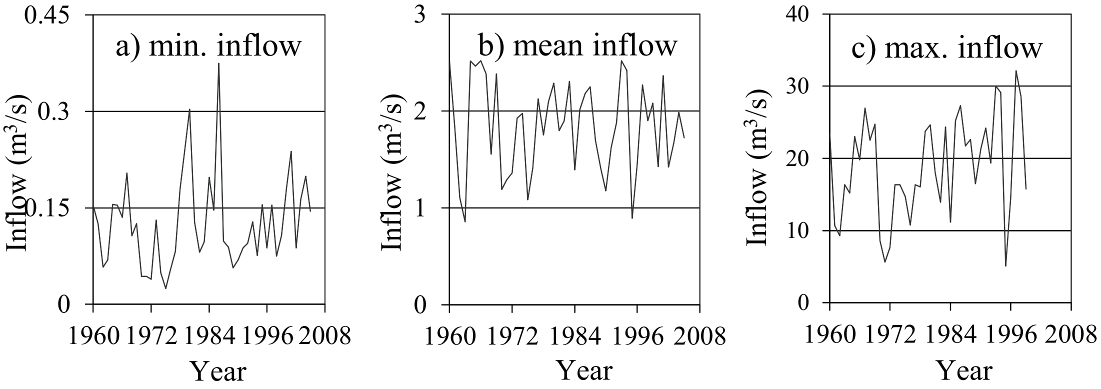


Figure 2.16: *Annual inflow into the Henne reservoir (Nov. 1960 - Oct. 2006)*

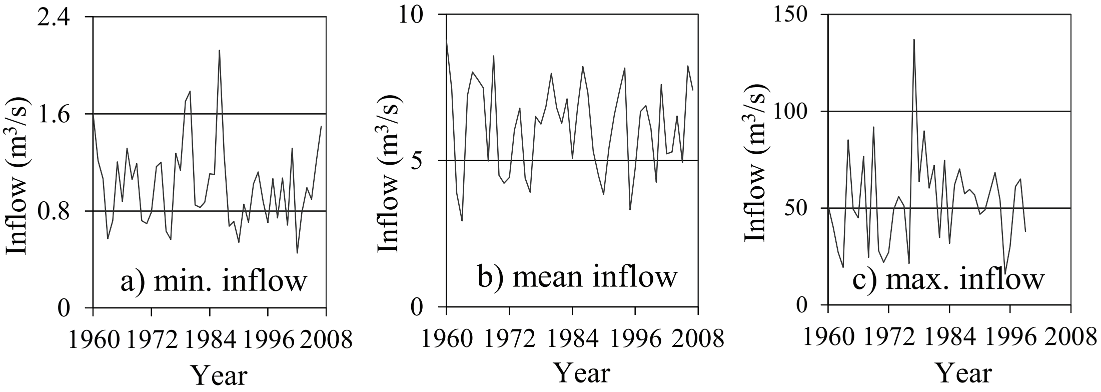


Figure 2.17: *Annual inflow into the Möhne reservoir (Nov. 1960 - Oct*

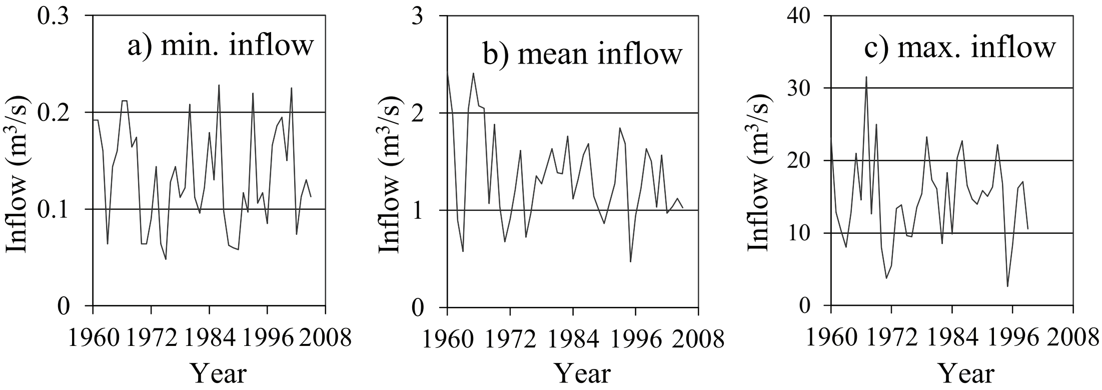


Figure 2.18: *Annual inflow into the Sorpe reservoir (Nov. 1960 - Oct. 2006)*

Table 2.4: *Values of the statistics T_c and Z_{MK} for the 3-months, 6-months and annual inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs*

Season	Bigge reservoir				Henne reservoir				Möhne reservoir				Sorpe reservoir			
	T_c	H_0	Z_{MK}	H_0	T_c	H_0	Z_{MK}	H_0	T_c	H_0	Z_{MK}	H_0	T_c	H_0	Z_{MK}	H_0
3-monthsă																
1	0.708	0	1.014	0	0.526	0	0.814	0	0.256	0	0.4	0	-0.941	0	-0.322	0
2	0.244	0	0.827	0	0.302	0	0.53	0	0.98	0	1.36	0	-1.061	0	-0.985	0
3	-1.695	0	-1.13	0	-2.66	1	-2.045	1	-3.056	1	-2.32	1	-3.027	1	-2.216	1
4	-0.034	0	0.175	0	0.543	0	0.379	0	0.666	0	-0.062	0	-1.013	0	-0.985	0
6-monthsă																
1	0.719	0	0.711	0	0.583	0	0.473	0	0.77	0	0.684	0	-1.287	0	-1.023	0
2	-0.803	0	-0.711	0	-1.354	0	-1.344	0	-1.296	0	-1.395	0	-2.633	1	-1.988	1
annuală																
Min.	0.446	0	0.044	0	1.032	0	1.231	0	-0.673	0	-0.684	0	-0.528	0	-0.616	0
Mean	0.12	0	0.058	0	-0.273	0	-0.379	0	-0.211	0	-0.169	0	-2.172	1	-1.61	0
Max.	1.095	0	1.014	0	1.041	0	1.004	0	0.065	0	0.08	0	-1.099	0	-0.464	0

Table 2.5: *Values of the seasonal Mann-Kendall test statistic, Z' for the 10-days, monthly, 3-months and 6-months inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs*

Time-scale	Bigge		Henne		Möhne		Sorpe			
	Z'	H_0	Z'	H_0	Z'	H_0	original data		after removing shifts	
							Z'	H0	Z'	H_0
10-days	0.43	0	1.3	0	-1.39	0	-3.82	1	-2.16	1
monthly	0.41	0	-0.78	0	-0.54	0	-3.01	1	-1.55	0
3-months	0.44	0	-0.16	0	-0.31	0	-2.25	1	-0.69	0
6-months	0	0	-0.62	0	-0.5	0	-2.14	1	-0.92	0

2.6 Testing for long memory

2.6.1 Introduction

A time series x_t exhibits long memory if the absolute values of autocorrelation are not summable. For a process x_t with autocorrelation function (*ACF*) denoted as ρ_k at lag k, McLeod and Hipel (1978) showed that the processes possess a long memory if

$$\sum_{p=-\infty}^{\infty} |\rho_k| = \infty \tag{2.27}$$

and the processes are called long-range dependent (LRD), long-range correlated or long-memory processes. For some stochastic processes, such as autoregressive and moving average type models the *ACF* decays exponentially and

$$\sum_{p=-\infty}^{\infty} |\rho_k| = \text{constant} < \infty \tag{2.28}$$

In other words the *ACF*, $\rho(k)$ decays to zero exponentially fast as $k \to \infty$. These processes are called short-range dependent (SRD) or short-range correlated.

The first work in the subject of long memory was originally documented by Hurst (1951). Hurst (1951, 1956) analyzed a large number of geophysical time series such as stream-flow, precipitation, temperature and tree-ring series and concluded that the values of the Hurst parameter (H) obtained for the different time series gave a mean of about 0.73.

Since Hurst's work the subject of long-memory time series has subsequently received extensive attention in many diverse fields of application, such as financial time series (e.g., Barkoulas et al., 1999; Teyssière and Abry, 2006), in internet traffic (e.g., Leland et al., 1994; Karagiannis et al., 2004), hydrology and etc.

Eltahir (1996) showed that the change of convergence of atmospheric moisture in the Ethiopian source region of the Nile River due to the El Niño Southern Oscillation (ENSO) is causing a nonstationarity of the mean annual flow of the Nile River. Lohre et al. (2003) discussed long-memory for the flow of Rhine River. The presence of the long-memory in the streamflow time series for the Yellow, Danube, Rhine, Ocmulgee, Umpqua Rivers was detected by Van Gelder et al. (2007).

2.6.2 Presence of long memory

Three measures are commonly used to estimate the strength of long-memory (the presence of long memory):

1. The parameter H, which is also known as the Hurst or self-similarity parameter.
2. The fractional differencing parameter (d), of the autoregressive fractionally integrated moving average, ARFIMA(p,d,q).
3. The power exponent (β) of the power spectrum function $1/f^{\ \beta}$ in the spectral analysis.

Table 2.6 summarizes the relationships between H, d and β (Stroe-Kunold et al., 2009). There are many available methods to test the existence of long-memory and also estimate H, d and β. We used three heuristic methods (the aggregated variance, residuals of regression and rescaled adjusted range R/S methods) to estimate Hurst parameter (H) for the daily, 10-days and monthly inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs. Furthermore, two semiparametric methods are used to test the existence of long-memory for the same data (the Lo's modified R/S statistic and the *GPH* test).

Table 2.6: *Relationships between parameters that capturing long-range dependence.*

	β	H	d
Random walk (ordinary Brownian motion)	2.0	0.5	1.0
White noise (ordinary Gaussian noise)	0	0.5	0
Fractional Gaussian noise (fGn)	[-1.0,1.0]	[0,1.0]	[-0.5,0.5]
	[-1.0,0] antipersistent	[0,0.5] antipersistent	[-0.5,0] antipersistent
	[0,1.0] persistent $H = \frac{\beta+1}{2}$	[0.5,1.0] persistent $H = \frac{2d+1}{2}$	[0,0.5] persistent $\beta = 2d$
Fractional Brownian motion (fBm)	[1,3.0]	[0,1.0]	[0.5,1.5]
	[1.0,2.0] antipersistent	[0,0.5] antipersistent	[0.5,1.5] antipersistent
	[2.0,3.0] persistent $H = \frac{\beta-1}{2}$	[0.5,1.0] persistent $H = \frac{2d-1}{2}$	[1.0,1.5] persistent $\beta = 2d$

2.6.3 Methods for estimation of H and d

Aggregated Variance Method

For a time series $x_t = x_1, x_2, ..., x_N$, Beran (1994) proved that in the presence of long-memory the variance of the sample mean could be expressed as:

$$Var(\bar{x}) \approx cN^{2H-2} \tag{2.29}$$

where $c > 0$ and H is the Hurst parameter. He gave the following procedure to estimate H:

1. Divide the original time series $x_t$into m subseries each of length k $(2 \leq k \leq N/2)$.

2. Calculate the mean of each subseries $x_1(k)$, $x_2(k)$, $\ldots$, $x_m(k)$ and the overall mean

$$\bar{x}(k) = m^{-1} \sum_{j=1}^{m} \bar{x}_j(k) \tag{2.30}$$

where $\bar{x}_j(k)$ is the mean of the subseries j, $j = 1, 2, \ldots, m$.

3. Calculate the sample variance, $Var(k)$of the m samples

$$Var(k) = (m-1)^{-1}\sum_{j=1}^{m}(\bar{x}_j(k) - \bar{x}(k))^2 \tag{2.31}$$

4. Repeat steps from 1 to 3 using successive values of m.

5. For each value of k, plot $Var(k)$ against log k.

6. The plotted points from step 5 are expected to be scattered around a straight line with negative slope $2H - 2$.

Residuals of Regression method

One of the more recent methods for estimating the strength of long-memory is the residuals of regression method due to Peng et al. (1994). For time series $x_t = \{x_i,\, i = 1, 2, \ldots, N\}$, the residuals of regression method comprises the following steps (Taqqu et al., 1995):

1. Divide the time series x_t into blocks of size n.

2. Compute the partial sums within each of the blocks. Call the partial sums within a block, Y_t, $t = 1, 2, \ldots, n$.

3. Fit a least square line $a + bt$ to these partial sums.

4. Compute the sample variance of residuals$Var(re)$

$$Var(re) = \frac{1}{n}\sum_{t=1}^{n}(Y_t - a - bt)^2 \tag{2.32}$$

5. Repeat steps 2, 3 and 4 for all blocks.

6. Compute the average variance $\overline{Var(re)}$.

7. Repeat all previous steps for different values of n.

8. Plot $\overline{Var(re)}$ versus n on log-log plot.

9. The result from step 8 is a straight line with slope $2H$.

The rescaled adjusted range (R/S) method

The R/S statistic was developed by Hurst (1951) and discussed in details in Mandelbrot and Wallis (1969), Mandelbrot (1975) and Mandelbrot and Taqqu (1979). For a time series $x_t = \{x_i, \mathrm{i} = 1, 2, \ldots, N\}$, the following steps can be used to estimate H using R/S statistic (Taqqu et al., 1995; Alptekin, 2006):

1. Divide the time series $x_t = \{x_i,\ i = 1,\ 2,\ \ldots,\ N\}$ into m subperiods, each of size $n = N/m$. Then any element in each subperiod can be defined as $x_{i,j}$, where $i = 1, 2, \ldots, n$ which denotes the number of the elements in each subperiod and $j = 1, 2, \ldots, m$ denotes the index of the subperiod.

2. The mean of each subperiod is estimated as $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$. Estimation of the deviation from the mean, $Y_{i,j} = x_{i,j} - \bar{x}_j$ with standard deviation

$$S_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_{i,j}^2} \tag{2.33}$$

3. For each subperiod j, the R/S statistic is estimated as follows:

$$\left(\frac{R}{S}\right)_j = \frac{1}{S_j}\left[\max_{1\leq k\leq n}\sum_{i=1}^{k} Y_{i,j} - \min_{1\leq k\leq n}\sum_{i=1}^{k} Y_{i,j}\right] \tag{2.34}$$

4. For fractional Gaussian noise or fractional ARIMA,

$$E\left[\frac{R}{S}(n)\right] \sim C_H n^H \tag{2.35}$$

   as $n \to \infty$, where $(R/S)_n$ is the average of the estimated $(R/S)_j$ and C_H is another positive, finite constant dependent on n.

5. Plot $\log[(R/S)_n]$ versus $\log n$.

6. Repeat steps from 1 to 6 by increasing n to the next integer value until $n = N/2$.

7. The Hurst parameter (H) is the slope of the fitted line of the plotted points from step 3.

To check the effectiveness of the R/S test, Wang (2006) generated ten simulations of an AR(1) model, ten simulations of an ARFIMA(0,d,0) and ten simulations of an ARFIMA(1,d,0). The results indicated that the R/S analysis is not a very reliable method in the presence of short-range dependence.

Lo's modified R/S statistic

Lo (1991) discussed the robustness of the R/S statistic and concluded that the R/S statistic is not robust to short memory dependence. For a time series, x_t with length N, Lo (1991) estimated the statistic $V_N(q)$ which is a modification to the R/S statistic as follows:

$$V_N(q) = N^{-1/2} Q_N(q) \tag{2.36}$$

where q is the truncation lag and

$$Q_N(q) = \frac{1}{S_q}\left[\max_{1\leq k\leq N}\sum_{i=1}^{k}(x_i-\bar{x}) - \min_{1\leq k\leq N}\sum_{i=1}^{k}(x_i-\bar{x})\right] \tag{2.37}$$

in which

$$S_q = \left(\frac{1}{N}\sum_{j=1}^{N}(x_j-\bar{x})^2 + \frac{2}{N}\sum_{j=1}^{q}w_j(q)\left[\sum_{i=j+1}^{N}(x_i-\bar{x})(x_{i-j}-\bar{x})\right]\right)^{1/2}, \quad q < N \tag{2.38}$$

where $\bar{x}$ is the sample mean of the time series and the weights $w_j(q)$ can be defined as $w_j(q) = 1 - \frac{j}{(q+1)}$ $q< N$.

In the case of no long-memory, Lo (1991) showed that given the right value of q, the distribution of $V_N(q)$is asymptotic to that of

$$W_1 = \max_{0\leq t\leq 1} W_0(t) - \min_{0\leq t\leq 1} W_0(t) \tag{2.39}$$

where W_0 is the standard Brownian bridge ($W_0(t) = B(t) - tB(1)$, in which B denotes the standard Brownian motion). The distribution of the random variable W_1 is given by (Kennedy, 1976) as:

$$P(W_1 \leq a) = 1 + 2\sum_{i=1}^{\infty}\left(1 - 4a^2i^2\right)e^{-2a^2i^2}, \qquad a \geq 0 \tag{2.40}$$

which follows that $P\{W_1 \in [0.809, 1.862]\} = 0.95$.

Lo used the interval [0.809, 1.862] as the 95% (asymptotic) acceptance region for testing the null hypothesis, $\mathrm{H}_0 = \{$no long-memory, i.e., $H = 0.5\}$ against the long memory alternative $\mathrm{H}_1 = \{$there is long-memory, i.e., $0.5 < H < 1.0\}$.

In the present work, the truncation lag (q) was estimated using the formula given by Lo (1991) as follows:

$$q = \left[\left(\frac{3N}{2}\right)^{1/3}\left(\frac{2\,\frown{\rho}}{1-\frown{\rho}^2}\right)^{2/3}\right] \tag{2.41}$$

where $\frown{\rho}$ is the lag one autocorrelation function.

GPH test

The *GPH* estimator is proposed by Geweke and Porter-Hudak (1983) and it is probably the most used estimator of the long memory parameter. *GPH* estimator is based on the autoregressive fractionally integrated moving average (ARFIMA) model. The ARFIMA(p,d,q) model is discussed in details in chapter 3, section 3.6. Given a fractionally integrated process $\{x_t\}$ its spectral density is given by:

$$f(\omega, d) = [2\sin(\omega/2)]^{-2d} f_u(\omega) \tag{2.42}$$

where ω is the Fourier frequency, $f_u(\omega)$is the spectral density corresponding to a stationary short memory disturbance (u_t) with zero mean and $\omega_j = 2\pi j/N$, j = 1, 2, ..., $N/2$, where N is the sample size. Taking the logarithm of the spectral density (equation, 2.42), the following equation can be obtained:

$$\ln f(\omega_j, d) = \ln f_u(0) - d \ln\left[4\sin^2(\omega_j/2)\right] + \ln\left[f_u(\omega_j)/f_u(0)\right] \tag{2.43}$$

The fractional difference parameter (d) can be estimated by the regression equations constructed from equation (2.43). Using a periodogram estimate of $f(\omega_j, d)$, Geweke Porter-Hudak (1983) showed that if the number of frequencies used in the regression equation (2.43) is a function g(N) (a positive integer) of the sample size N where g(N) = N^α with $0 < \alpha < 1$, the least squares estimate of $\hat{d}$ using the above regression is asymptotically normally distributed in large samples. Under the null hypothesis of no long-memory (d = 0), the t-statistic

$$t_{d=0} = \hat{d} \times \left(\frac{\pi^2}{6 \times \sum_{j=1}^{g(N)} (U_j - \bar{U})^2}\right)^{-1/2} \tag{2.44}$$

has limiting standard normal distribution where $U_j = \ln\left[4\sin^2(\omega_j/2)\right]$ and $\bar{U}$ is the sample mean of U_j, j = 1, 2, ..., $g(N)$.

Results of long memory estimation methods

Before conducting the long memory analysis, we seek to remove the trend (if exists) out of the original inflow time series. We removed the existing trend in the time series by fitting a trend line to the data and subtract the calculated trend component at each time out of the original time series. The results of the aggregate variance, residuals of regression and R/S methods (heuristic methods) are plotted in figures B.1, B.2 and B.3 (appendix B) respectively. The results of the heuristic methods and semiparametric tests are summarized in table 2.7 and can be concluded in the following points:

- The daily inflow time series of all reservoirs exhibit long-memory ($H > 0.5$) using the aggregate variance, residuals of regression and R/S methods.

- The 10-days inflow time series of the Möhne and Sorpe reservoirs are found to be long-memory process using the aggregate variance, residuals of regression and R/S methods.

- Long-memory is detected in the monthly inflow time series of all reservoirs using the R/S method.

- According to the results of the Lo's modified R/S test, the null hypothesis of no long-memory is rejected only for daily and 10-days inflow time series of the Bigge and Henne reservoirs, also for the monthly inflow time series of Henne reservoir.

- The results of the GPH test show that the null hypothesis of no long-memory is rejected only for daily inflow time series of the Bigge, Möhne and Sorpe reservoirs.

Long memory detection using autocorrelation function

If the time series exhibits long-memory, the autocorrelation function decreases to be zero at long lags (Gil-Alana, 2006), see equation (2.27). Figures 2.19, 2.20, 2.21 and 2.22 show the autocorrelation function of the inflow time series (daily, 10-days and monthly) of the Bigge, Henne, Möhne and Sorpe reservoirs respectively. It is obvious that for all reservoirs the autocorrelation function of the daily inflow time series decay more slowly than the 10-days and monthly inflow time series. As mentioned before slowly decaying of the autocorrelation function is an indication of the presence of long-memory which consistents with the results of the GPH test except those for the daily inflow time series of the Henne reservoir.

Table 2.7: *Results of the long-memory detection methods of the original inflow time series*

Reservoir	Time series	Hurst parameter (H)			Lo′s test		GPH test	
		Aggregate variance	Residuals of regression	R/S	lag	V	$t_{d=0}$	d
Bigge	daily	0.743	0.87	0.672	88	0.793	2.525*	0.159
	10-days	0.417	0.677	0.492	13	0.739	0.272	0.034
	monthly	0.337	0.34	0.522	9	0.985	-0.145	-0.025
Henne	daily	0.761	0.871	0.68	135	0.679	0.858	0.052
	10-days	0.415	0.642	0.543	17	0.69	-0.173	-0.02
	monthly	0.337	0.418	0.603	8	0.805	-0.501	-0.083
Möhne	daily	0.771	0.885	0.728	110	0.822	2.155*	0.128
	10-days	0.546	0.704	0.609	19	0.809	0.188	0.022
	monthly	0.492	0.555	0.704	9	0.878	-0.081	-0.013
Sorpe**	daily	0.775	0.883	0.71	125	1.008	2.507	0.151
	10-days	0.501	0.658	0.567	18	0.893	0.267	0.031
	monthly	0.462	0.487	0.64	9	0.943	-0.323	-0.053

* t-statistic in the GPH test is not contained in the interval [-1,960, 1.960] which means rejection of the null hypothesis of no long memory.

** the detected shifts are removed from the 10-days and monthly inflow time series of the Sorpe reservoir before applying this test.

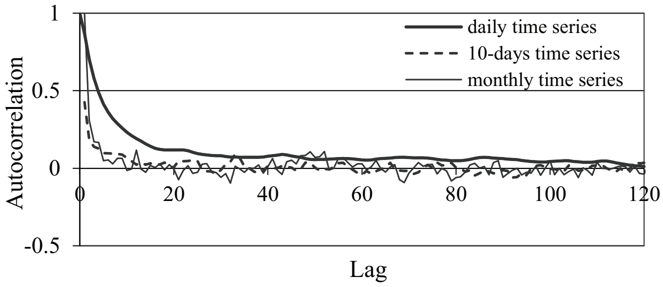


Figure 2.19: *Autocorrelation functions of the inflow time series of the Bigge reservoir*

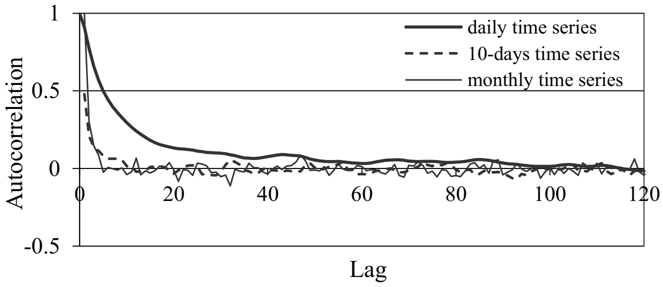


Figure 2.20: *Autocorrelation functions of the inflow time series of the Henne reservoir*

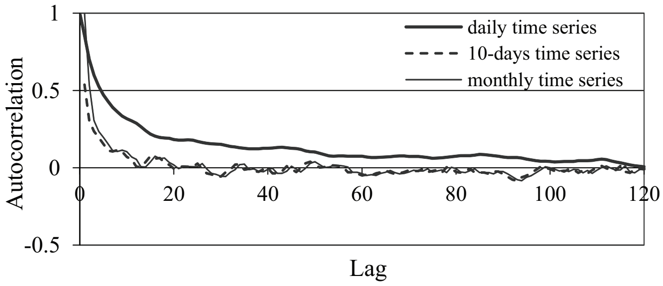


Figure 2.21: *Autocorrelation functions of the inflow time series of the Möhne reservoir*

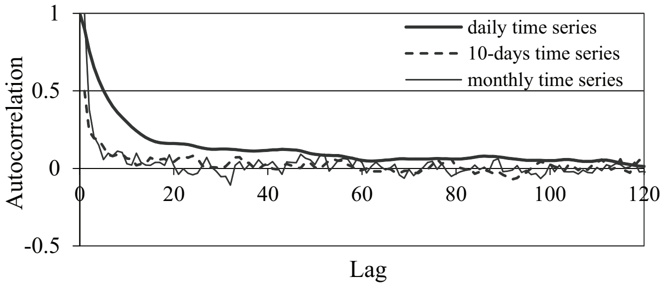


Figure 2.22: *Autocorrelation functions of the inflow time series of the Sorpe reservoir*

2.7 Testing for stationarity

2.7.1 Introduction

According to Shumway and Stoffer (2006) a strictly stationary time series, $\{x_t\}$, $t = 1, 2, \ldots, N$ is defined as the time series for which the probabilistic behavior of every set of values $\{x_{t1}, x_{t2}, \ldots, x_{tk}\}$ is identical to that of the time shifted set $\{x_{t1+h}, x_{t2+h}, \ldots, x_{tk+h}\}$. That is,

$$P(x_{t1} = c_1, ..., x_{tk} = c_k) = P(x_{t1+h} = c_1, ..., x_{tk+h} = c_k) \tag{2.45}$$

for all $k = 1, 2, ...$, all time points $t_1, t_2, \ldots, t_k$, all numbers $c_1, c_2, \ldots, c_k$ and all time shifts $h = 0, \pm1, \pm2, ...$.

Time series process $\{x_t\}$ is assumed to be stationary if one or more of its statistical properties such as mean, variance, autocorrelation, etc., do not depend on time.

2.7.2 Stationarity test methods

Two groups of methods can be used for testing time series for stationarity. The first group is based on the statistics of the full sequence (Otache et al., 2008; Wang, 2006). The second group is based on the statistics of different segments of the time series (Chen and Rao, 2002). In the present study, we applied two unit root tests (of the first group) for testing the inflow time series for stationarity. These tests are the Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) unit root tests.

2.7.3 Augmented Dickey-Fuller unit root test (ADF)

The Augmented Dickey-Fuller (ADF) unit root test is first proposed by Dickey and Fuller (1979). It is conducted by estimating the ordinary least square analysis (OLS). Consider a simple autoregressive model of order one AR(1) for a given time series $x_t = x_1, x_2, ..., x_N$

$$x_t = \phi x_{t-1} + \varepsilon_t \,\text{where}\, \varepsilon_t \sim iid(0, \sigma^2) \tag{2.46}$$

The hypotheses of the test are

$$H_0 : \phi = 1 \quad \Rightarrow \quad x_t \sim I(1)$$
$$H_1 : |\phi| < 1 \quad \Rightarrow \quad x_t \sim I(0)$$

If $\phi = 1$ the time series x_t is nonstationary and known as random walk process. Acceptance of the alternative hypothesis ($|\phi\ | < 1$) implies that the time series x_t is stationary. The test statistic t_ϕ for testing the null hypothesis that $\phi = 1$ is calculated as:

$$t_\phi = \frac{\hat{\phi} - 1}{\hat{\sigma}_{\hat{\phi}}} \tag{2.47}$$

where $\hat{\phi}$ is the least squares estimate and is estimated as:

$$\hat{\phi} = \left(\sum_{t=2}^{N} x_{t-1}^2 \right)^{-1} \sum_{t=2}^{N} x_t x_{t-1} \tag{2.48}$$

and $\hat{\sigma}_{\hat{\phi}}$ is the usual OLS for the estimated coefficient, $\hat{\sigma}_{\hat{\phi}} = S_e \left(\sum_{t=2}^{N} x_{t-1}^2 \right)^{1/2}$ in which S_e is the standard deviation of the OLS estimate of the residual in the retrogression model of equation (2.46) and estimated as:

$$S_e = \left(\frac{1}{N-2} \sum_{t=2}^{N} \left(x_t^2 - \hat{\phi} x_{t-1} \right)^2 \right)^{1/2} \tag{2.49}$$

The limiting distribution of the statistic t_ϕ under $\phi = 1$ is called the Dickey-Fuller (DF) distribution. Fuller (1976) gave a set of tables of the percentiles of the limiting distribution of t_ϕ under $\phi = 1$. The unit root test which is described above is valid only if the observed time series x_t is well characterized by an AR(1) with white noise errors. Most of time series have a complicated dynamic structure that cannot be captured by a simple AR(1) model. Said and Dickey (1984) augmented the basic autoregressive unit root test to accommodate general ARMA(p,q) models with unknown orders and their test is referred to as the augmented Dickey-Fuller (ADF) test. The ADF test is based on estimating OLS regression model which includes a time trend

$$x_t = C + \phi x_{t-1} + \delta t + \sum_{j=1}^{p} \zeta_j \Delta x_{t-j} + \varepsilon_t \tag{2.50}$$

for some constant C, AR(1) coefficient $\phi < 1$, time trend stationary coefficient δ and Δ is the lag operator. Lag p indicates the number of lagged changes or first differences in x_t that are included in the OLS regression model.

Optimal lag length for the ADF test

In equation (2.50) the lagged differences (Δx_{t-j})are included to ensure that the residuals (ε_t) are well behaved. The specification of the lag length (p) is an important practical issue for the implementation of the ADF test. If p is too small, the size of the test changes in an unknown manner due to the remaining serial correlation and if too many lags are included, the power of the test will suffer. Two procedures are used in the present study to determine the optimum lag p as follows:

1. Set an upper bound p_{max} for p. Schwert (1989) recommends a maximum lag as

   $$p_{\max} = 12(N/100)^{1/4} \tag{2.51}$$

   where N is the number of observations.

   Select $p \leq p_{max}$ that produces the minimum Akaike information criterion, AIC (Akaike, 1974). The AIC for the estimated model is defined by the following equation:

   $$\text{AIC} = \frac{2m}{N} + \log\left(\hat{\sigma}^2\right) \tag{2.52}$$

   in which N is the number of observations, $\hat{\sigma}^2$ is the estimated noise variance and m is the number of parameters.

2. Using the procedure which is suggested by Gallet (2003). First, set an upper bound p_{max} for p (e.g., that was suggested by Schwert, 1989). Next, estimate the ADF test regression with $p = p_{max}$. If the last lag term is significant (i.e., the absolute value of the t-statistic exceeds 1.645), then p remains at p_{max} and perform the unit root test. Otherwise, reduce the lag length by one. Stop the process when the coefficient of the last lag term is significant or p is set equal to zero.

2.7.4 Phillips-Perron unit root test (PP)

An approach to detect the presence of unit roots in the data was suggested by Phillips (1987) and Phillips and Perron (1988). They modified the Dickey–Fuller test by including additional lagged variables as regressors in the model on which the test is based. In this approach, the effect of autocorrelation present was captured by making a non-parametric correction to t-test statistic. Considering the time series model which was introduced in equation (2.46), the PP test-statistic under the null-hypothesis of which the Dickey-Fuller test is a special case can be given as follows:

$$t_{PP} = (s_\varepsilon / s_{NK})\, t_\phi - \left(\frac{1}{2}\right) (s_{NK}^2 - s_\varepsilon^2) \left\{ s_{NK} \left[N^{-2} \sum_{t=1}^{N} x_{t-1}^2 \right]^{1/2} \right\}^{-1} \tag{2.53}$$

where $s_\varepsilon^2 = N^{-1} \sum_{t=1}^{N} \varepsilon_t^2$, and $s_{NK}^2 = N^{-1} \sum_{t=1}^{N} \varepsilon_t^2 + 2N^{-1} \sum_{t=1}^{K} \sum_{t=j+1}^{N} \varepsilon_t \varepsilon_{t-j}$, are the consistent estimators of $\sigma_\varepsilon^2 = \lim_{N\to\infty} E\left[N^{-1}\left(\sum_{t=1}^{N} \varepsilon_t^2\right)\right]$ and $\sigma^2 = \lim_{N\to\infty} N^{-1} \sum_{t=1}^{N} E(\varepsilon_t^2)$, respectively in which K is the truncation lag.

The importance of truncation lag is to ensure that the autocorrelation is fully captured. The lag truncation (K) is taken to be the integer value of $4(N/100)^{2/9}$ (Newey and West, 1987).

2.7.5 Results of the ADF and PP tests

The results of the ADF and PP tests at 5 % significance level for the log-transformed inflow time series at different timescales (daily, 10-days, monthly, 3-months, 6-months and annual) are given in table 2.8 and can be summarized as follows:

daily and 10-days inflow time series:

1. In favor of the stationarity alternative the null hypothesis is rejected for daily and 10-days inflow time series of all reservoirs.

monthly, 3-months and 6-months inflow time series:

2. For all inflow time series of the Bigge reservoir the null hypothesis of the ADF test cannot be rejected (lags are estimated using procedure 1, section 2.7.3).

3. The null hypothesis of the ADF test is rejected for all inflow time series of the Henne reservoir except for the 6-months one (lags are estimated using procedure 2 section 2.7.3)
4. The null hypothesis of the ADF test cannot be rejected for all inflow time series of the Möhne reservoir.
5. For all inflow time series of the Sorpe reservoir, the null hypothesis of the ADF and PP tests are rejected.

annual time series:

6. The null hypothesis of the ADF and PP tests cannot be rejected for all annual inflow time series of all reservoirs except that of the Sorpe reservoir.
7. The null hypothesis of the PP test is rejected for the annual inflow time series of the Sorpe reservoir.

Table 2.9 gives the results of the ADF and PP tests for the log-transformed and standardized inflow time series at different timescales. The results show that the null hypothesis is rejected for all inflow time series of all reservoirs which means that all inflow time series appear to be stationary by applying log-transformation and standardization to them.

Table 2.8: *Results of the stationarity tests of the log-transformed inflow time series*

Time	Test	Bigge			Henne			Möhne			Sorpe		
series		lag	pvalue	TestStat	lag	pvalue	TestStat	lag	pvalue	TestStat	lag	pvalue	TestStat
daily	Dickey-Fuller	41*	< 0.01	-5.59	39	< 0.01	-11.33	34	< 0.01	-5.32	36	< 0.01	-10.45
		42**		-5.58	44		-14.41	44		-4.823	44		-9.45
	Phillips-Perron	13		-10.86	13		-16.15	43		-9.101	13		-16.23
10-days	Dickey-Fuller	24	0.041	-2.026	11		-12.46	25	0.044	-2.001	24		-6.421
		24		-2.026	11		-12.46	25	0.044	-2.001	25		-6.266
	Phillips-Perron	8	< 0.01	-7.978	8		-14.75	8	< 0.01	-5.986	8		15.15
monthly	Dickey-Fuller	12	0.318	-0.911	11		-5.157	14	0.271	-1.095	11		-4.073
		7	0.037	-2.072	14		-5.174	8	0.071	-1.78	19		-4.733
	Phillips-Perron	6	< 0.01	-4.925	6		-11.48	6	< 0.01	-4.202	6		-11.15
3months	Dickey-Fuller	14	0.429	-0.604	11	0.011	-2.539	15	0.71	0.161	11		-3.503
		1	< 0.01	-3.351	14	0.048	-1.958	2	0.082	-1.712	13		-3.568
	Phillips-Perron	5	< 0.01	-3.279	5	< 0.01	-11.17	5	< 0.01	-2.936	5		-11.4
6months	Dickey-Fuller	12	0.58	-0.187	5	0.029	-2.173	10	0.609	-0.108	11	0.01	-2.608
		0	< 0.01	-3.291	6	0.059	-1.871	1	0.29	-0.982	11	0.01	-2.608
	Phillips-Perron	4	< 0.01	-2.992	4	< 0.01	-12.33	4	0.017	-2.396	4	< 0.01	-14.45
annual	Dickey-Fuller	8	0.728	0.234	7	0.442	-0.557	7	0.603	-0.117	7	0.05	-1.945
		0	0.398	-0.678	1	0.099	-1.615	0	0.372	-0.751	6	0.098	-1.62
	Phillips-Perron	4	0.436	-0.573	4	0.05	-1.945	4	0.448	-0.544	4	< 0.01	-4.283

* lag is estimated using procedure 1 which is given in section 2.7.3

** lag is estimated using procedure 2 which is given in section 2.7.3

Table 2.9: *Results of the stationarity tests of the log-transformed and standardized inflow time series*

Time	Test	Bigge			Henne			Möhne			Sorpe		
series		lag	pvalue	TestStat	lag	pvalue	TestStat	lag	pvalue	TestStat	lag	pvalue	TestStat
daily	Dickey-Fuller	4	< 0.01	-20.52	7	< 0.01	-37.83	34	< 0.01	-12.8	38	< 0.01	-19.77
		42		-11.99	44		-17.21	44		-11.92	44		-18.36
	Phillips-Perron	13		-21.41	13		-112.5	13		-23.44	13		-156.87
10-days	Dickey-Fuller	4		-11.8	19		-9.28	7		-9.797	2		-24.05
		24		-6.476	25		-8.05	25		-7.545	25		-8.036
	Phillips-Perron	8		-18.1	8		-40.16	8		-17.65	8		-41.58
monthly	Dickey-Fuller	4		-8.43	7		-8.08	12		-6.402	11		7.065
		18		-4.768	19		-3.92	19		-5.316	10		-7.049
	Phillips-Perron	6		-14.59	6		-21.43	6		-13.56	6		-24.37
3months	Dickey-Fuller	1		-7.69	4		-5.212	1		-7.972	4		-5.693
		0		-9.523	14		-2.95	0		-10.5	14		-3.08
	Phillips-Perron	5		-9.492	5		-12.09	5		-10.47	5		-13.63
6months	Dickey-Fuller	1		-5.84	2		-5.165	2	< 0.01	-5.462	2		-5.032
		0		-8.163	8		-2.95	11	0.012	-2.524	8		-2.86
	Phillips-Perron	4		-8.17	4		-9.76	4	< 0.01	-8.366	4		-12.45
annual	Dickey-Fuller	6	0.042	-2.026	4		-3.346	6	0.013	-2.524	6	0.017	-2.4079
		6	0.042	-2.026	2		-5.128	6		-2.524	4	0.014	-2.498
	Phillips-Perron	4	< 0.01	-5.519	4		-5.924	4	< 0.01	-5.69	4	< 0.01	-5.321

2.8 Conclusions

As a result of the stochastic analysis of the inflow processes of the Bigge, Henne, Möhne and Sorpe reservoirs, it can be concluded that:

1. All daily, 10-days and monthly inflow time series have a clear seasonality in the mean and standard deviation. Seasons with high mean values have also high standard deviations. For all time series, the coefficients of variation and skewness have higher values in dry periods. The autocorrelation coefficients are low for high inflow seasons and high for seasons with low inflow for all inflow time series except for daily inflow time series at lag of one day.

2. All inflow time series are generally positively skewed. Log-transformation is applied to normalize the inflow time series.

3. We applied the linear regression, Mann-Kendall and seasonal Mann-Kendall tests to determine the presence of trend in the inflow time series at 5 significance level. The hypothesis H_0 of the linear regression test is rejected (presence of trend) for the inflow time series of the months:

   - April and May for the Henne reservoir.
   - June and July for the Möhne reservoir.
   - April, May, June, July and August for the Sorpe reservoir.

   According to the results of the Mann-Kendall test the hypothesis H_0 is rejected for the inflow time series of the month July for the Henne and Möhne reservoirs and the months April, June, July and August for the Sorpe reservoir. The results of the linear regression and Mann-Kendall tests show that a downward trend is detected in the 3-months inflow time series in the third season (months May, June and July) for the Henne, Möhne and Sorpe reservoirs. A downward trend is exhibited for the 6-months and annual inflow time series only for the Sorpe reservoir. Using the seasonal Mann-Kendall test, a downward trend is detected in all tested inflow time series of the Sorpe reservoir only. All seasons in the 10-days, monthly, 3-months and 6-months inflow times of Sorpe reservoir are test against shift (occurred at 1970) and the detected shifts are removed. These seasonal time series are tested against trend after removing the detected shifts and a trend is detected only in the 10-days inflow time series.

4. The results of the GPH test show that the null hypothesis of no long-memory is rejected for the daily inflow time series of the Bigge, Möhne and Sorpe reservoirs.

5. The ADF and PP unit root tests are applied for the log-transformed and for the log-transformed and standardized inflow time series at different timescales. The results of ADF and PP tests show that the null hypothesis is rejected for all inflow time series of all reservoirs after applying log-transformation and standardization. This means that all inflow time series appear to be stationary by applying log-transformation and standardization to them.

2.9 References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

Alptekin, N., 2006. Long memory analysis of USD/TRL exchange rate. International Journal of Social Sciences, 1(2), 111–116.

Barkoulas, J.T., Baum, C.B., Caglayan, M., 1999. Long memory or structural breaks: can either explain nonstationary real exchange rates under the current float?. Journal of International Financial Markets, Institutions and Money, 9, 359–376.

Beran, J., 1994. Statistics for long–memory processes. Chapman & Hall, London.

Box, G.E.P., Jenkins, G.M., 1976. Time series analysis: forecasting and control. San Francisco: Holden–Day.

Chen, H–L., Rao, A.R., 2002. Testing hydrologic time series for stationarity. Journal of Hydrologic Engineering, 7(2), 129–136.

Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74, 427–431.

Donald, H.B., Juraj, M.C., 2004. Hydrological trends and variability in the Liard River basin. Hydrological Sciences Journal, 49(1), 53–67.

Eltahir, E. A. B., 1996. El Niòo and the natural variability in the flow of the Nile River. Water Resources Research, 32(1), 131–137.

Fleming, S.W., Clarke, G.K.C., 2002. Autoregressive noise, deserialization and trend detection and quantification in annual river discharge time series. Canadian Water Resources Journal, 27, 335–354.

Fuller, W., 1976. Introduction to statistical time series. Wiley & Sons, New York.

Gallet, C.A., 2003. Convergence of market shares in the U.S. cigarette industry. The Journal of Applied Business Research, 19(4), 33–37.

Geweke, J., Porter–Hudak, S., 1983. The estimation and application of long memory time series models. Journal of Time Series Analysis, 4, 221–238.

Gil–Alana, L.A., 2006. Fractional integration in daily stock market indexes. Review of Financial Economics, 15, 28–48.

Hamed, K. H., Rao, A. R., 1998. A modified Mann–Kendall trend test for autocorrelated data. Journal of Hydrology, 204(1), 182–196.

Helsel, D. R., Hirsch, R. M., 1992. Statistical methods in water resources. Elsevier Publishers, New York.

Hirsch, R.M., Slack, J.R., 1984. A nonparametric trend test for seasonal data with serial dependence. Water Resources Research, 20, 727–732.

Hirsch, R.M., Slack, J.R., Smith, R.A., 1982. Techniques of trend analysis for monthly water quality data. Water Resources Research, 18, 107–121.

Hurst, H.E., 1951. Long–term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers, 116, 770–799.

Hurst, H.E., 1956. Methods of using long term storage in reservoirs. Proceedings of the Institute of Civil Engineers I, 519-543.

Karagiannis, T., Molle, M., Faloutsos, M., 2004. Long–range dependence ten years of internet traffic modeling. IEEE Internet Computing, 8(5), 57–64.

Kendall, M. G., 1975. Rank correlation methods. Charles Griffin, London.

Kennedy, D., 1976. The distribution of the maximum Brownian excursion. Journal of Applied Probability, 13, 371–376.

Leland, W.E., Taqqu, M.S, Willinger, W., Daniel V. Wilson, D.V., 1994. On the self–similar nature of Ethernet traffic (Extended Version). IEE/ACM Transactions on Networking, 2(1), 1–15.

Lo, A.W., 1991. Long–term memory in stock market prices. Econometrica, 59(5), 1279–1313.

Lohre, M., Sibbertsen, P., Könning, T., 2003. Modeling water flow of the Rhine River using seasonal long memory. Water Resources Research, 39(5), 1132–1138.

Maidment, D.R., 1993. Handbook of hydrology. McGraw–Hill, New York.

Mandelbrot, B.B., 1975. Limit theorems of the self–normalized range for weakly and strongly dependant processes. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 31(1), 271–285.

Mandelbrot, B.B., Taqqu, M.S., 1979. Robust R/S analysis of long run serial correlation. In Proceedings of the 42nd Session of the International Statistical Institute, volume 48 of Bulletin of the International Statistical Institute, 69–104.

Mandelbrot, B.B., Wallis, J.R., 1969. Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. Water Resources Research, 5, 967–987.

Mann, H.B., 1945. Nonparametric tests against trend. Econometrica, 13, 245–259.

Martins, Y.O., et al., 2008. Analysis of stochastic characteristics of the Benue River flow process. Chinese Journal of Oceanology and Limnology, 26(2), 142–151.

McLeod, A.I., Hipel, K.W., 1978. Preservation of the resealed adjusted range: 1. A reassessment of the Hurst phenomenon. Water Resources Research, 14, 491–508.

Mitosek, H.T., 2000. On stochastic properties of daily river flow processes. Journal of Hydrology, 228, 188–205.

Newey, W.K., West, K.D., 1987. A Simple, positive semi–definite, Heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55, 703–708.

Otache, M.Y., Bakir, M., Zhijia, L, 2008. Analysis of stochastic characteristics of the Benue River flow process. Chinese Journal of Oceanology and Limnology, 26(2), 142–151.

Peng, C.K., et al., 1994. Mosaic organization of DNA nucleotides. Physical Review E, 49, 1685–1689.

Phillips, P.C.B., 1987. Time Series Regression with a unit root. Econometrica, Econometric Society, 55(2), 277–301

Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika, 75(2), 335–346.

Said, S.E., Dickey, D. A., 1984. Testing for unit roots in autoregressive–moving average models of unknown order. Biometrika, 71, 599–607.

Schwert, W., 1989. Tests for unit roots: A Monte Carlo investigation. Journal of Business and Economic Statistics, 7, 147–159.

Shumway, R.H., Stoffer, D.S., 2006. Time series analysis and its applications: With R examples. 2nd edition, Springer, New York.

Stroe–Kunold, E., Stadnytska, T., Werner, J., Braun, S., 2009. Estimating long–range dependence in time series: An evaluation of estimators implemented in R. Behavior Research Methods, 41, 909–923.

Taqqu, M.S., Teverovsky, V., Willinger, W., 1995. Estimators for long range dependence: an empirical study. Fractals, 3, 785–798.

Teyssière, G., Abry, P., 2006. Wavelet analysis of nonlinear long–range dependent processes. Applications to financial time series. In: Teyssière G, Kirman A (eds) Long memory in economics. Springer, New York, 173–238.

Van Gelder, P.H.A.J.M., Wang, W., Vrijling, J. K., 2007. Statistical estimation methods for extreme hydrological events. Vasiliev, O.F., et al., (eds.), Extreme hydrological events: New concepts for security, 78, 199–252.

Wang, W., 2006. Stochasticity, nonlinearity and forecasting of streamflow processes. IOS Press, Amsterdam.

Yue, S., Pilon, P.J., Phinney, B., Cavadias, G., 2002. The influence of autocorrelation on the ability to detect trend in hydrological series. Hydrological Processes, 16(9), 1807–1829.

Chapter 3

Daily inflow forecasting

3.1 Introduction

The purpose of this chapter is to apply the backpropagation neural network (BPNN) and adaptive neuro-fuzzy inference system (ANFIS) models to forecast daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs. The autoregressive moving average (ARMA) and autoregressive fractional integrated moving average (ARFIMA) models are also applied to forecast the daily inflow into the same reservoirs and the results of the four models are compared.

The complex nature of the hydrological systems and the ability of the artificial neural networks (ANN) and the fuzzy logic (FL) to model nonlinear processes lead to use ANNs and FL in many branches of hydrology.
Coulibaly et al. (2001) compared the performance of the dynamic networks on reservoir inflow prediction and showed that all these models demonstrate significant improvement in prediction accuracy over the traditional multilayer perceptron (MLP) model. Ahmed and Sarma (2007) developed three synthetic streamflow generation models, namely, Thomas-Fiering, ARMA(2,0) and ANN-based models. They used the three models to generate 100 years of synthetic monthly flow and reported that the ANN-based model outperforms the other two models. Xu and Li (2002) formulated an ANN model to forecast 1- to 7-hours ahead inflow into a hydropower reservoir and concluded that this model forecasted the small and medium inflow values satisfactorily. Dawson and Wilby (2001) used MLP models to predict 6-hours river flow from precipitation data and the results showed that the MLP forecasts accurately the medium values of flow but overestimates the large ones.

Jain et al. (1999) used ANN models for reservoir inflow prediction. They compared the results of the ANN models with those of the classical time series models and found that the ANN models had better results. Smith and Eli (1995) applied a backpropagation ANN model to predict discharge and time to peak over a hypothetical watershed. El-Shafie et al. (2007) presented an adaptive neuro-fuzzy inference system (ANFIS) model for the inflow forecasting of the Nile River at Aswan high dam. Khadangi et al. (2009) used ANFIS, multi-linear regression (MLR) and artificial neural networks with radial base function (RBF-NN) models to simulate daily river flow time series of Mahabad River in northwest of Iran. The results of the ANFIS models are compared with those of the MLR and RBF-NN models and the results of ANFIS were the more accurate.

Moatmari et al. (1999) introduced a fractionally differenced autoregressive moving average (ARFIMA) model with periodical parameters, for modeling data affected by long memory and seasonal non-stationarity. Wang et al. (2008) applied ARMA and ARFIMA to daily average discharge series of medium-sized watersheds. They showed that both the ARMA and the ARFIMA models work well in forecasting short-term daily average discharges and the performance of the ARFIMA model was generally slightly better than that of the ARMA model.

3.2 Simulation models

The simulation models are divided into two groups according to the potential input variables as follows:

Univariate models (group M1-1 and group M1-2):

- The simulation models are BPNN, ANFIS, ARMA and ARFIMA.
- The potential input variables are the average daily inflow at days $t-1, t-2, t-3, t-4$ and $t-5$ ($x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}$ and x_{t-5} respectively).
- The output variable is the average daily inflow at day $t(x_t)$.

Multivariate models (group M2):

- The simulation models are BPNN and ANFIS.
- The potential input variables are the average daily inflow at days $t-1, t-2, t-3, t-4$ and $t-5$ ($x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}$ and x_{t-5} respectively) and the daily rainfall at days $t-1$ and $t-2$ (D_{t-1} and D_{t-2} respectively).

- The output variable is the average daily inflow at day $t(x_t)$.

Table 3.1 gives more details about the simulation models. The performance of the models, group M1-2 and the models, group M2, will compared in terms of the selected performance criteria.

Table 3.1: *List of the used data for daily inflow forecasting*

Reservoir	Models group	No. of potential input variables (m)	Used data
Bigge	M1-1	5	1 Nov. 1966 - 31 Oct. 2006
	M1-2		1 Nov. 1990 - 31 Oct. 2006
	M2	7	
Henne	M1-1	5	1 Nov. 1960 - 31 Oct. 2006
	M1-2		1 Nov. 1990 - 31 Oct. 2000
	M2	7	
Möhne	M1-1	5	1 Nov. 1960 - 31 Oct. 2008
	M1-2		1 Nov. 1990 - 31 Oct. 2006
	M2	7	
Sorpe	M1-1	5	1 Nov. 1960 - 31 Oct. 2006
	M1-2		1 Nov. 1990 - 31 Oct. 2000
	M2	7	

3.3 Backpropagation neural network (BPNN)

Multilayer perceptron (MLP) backpropagation neural network (BPNN) is applied in the present work. Figure 3.1 shows a typical three-layer feed forward ANN in which i, j and k denote nodes input layer, hidden layer and output layer, respectively and w is the weight of the nodes. Subscripts of w specify the connections between the nodes. For example, w_{ij} is the weight between nodes i and j.

In this network, the input data are fed to input nodes and then they will pass to the hidden nodes after multiplying by a weight. A hidden layer adds up the weighted input received from the input nodes, associates it with the bias and then passes the result on through a nonlinear transfer function (see section 3.3.5). The output node does the same operation as that of a hidden layer.

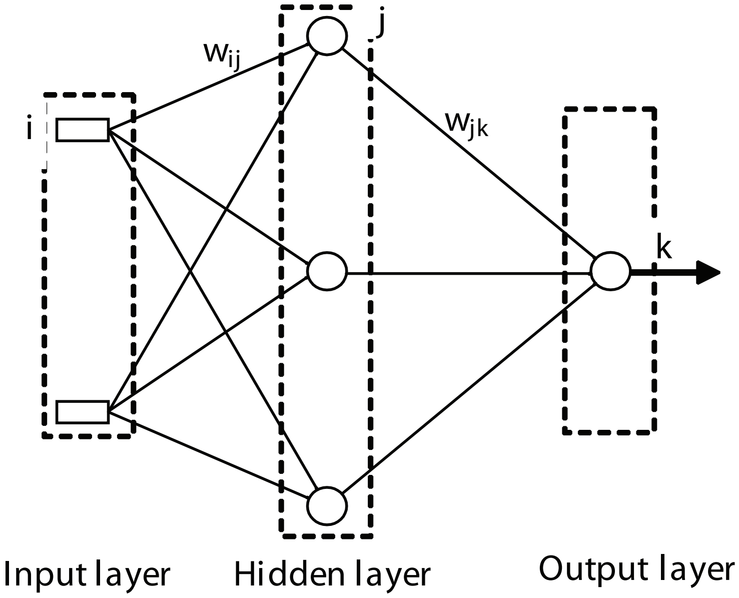


Figure 3.1: *A typical three-layer feed forward ANN*

3.3.1 Data preprocessing

In data preprocessing, variables are usually scaled so that important variables with small magnitudes are not overshadowed by those with larger magnitudes, in other words to give the same importance for all variables. It is also advisable to remove the detected trends in the data to improve the accuracy of the model (Zahng and Qi, 2005). Any one of the following methods can be used to scale the inputs and targets (Matlab.a, 2008):

Min-max method

The Min-max method is an approach to scale the inputs and the targets to be in the range [-1, 1]. For a given data set $\{x_i\}, i = 1, 2, \ldots, N$ the min-max method can be applied to calculate the rescaled data set $\{y_i, i = 1, 2, \ldots, N$ as follows:

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

in which $x_{\min}$ and $x_{\max}$ are the minimum and the maximum of $\{x_i\}$ respectively.

Mean-standard deviation method

The second approach for scaling network inputs and targets is to normalize the mean and standard deviation of the training set. The following equation illustrates the use of this approach to scale a data set $\{x_i\}, i = 1, 2, \ldots, N$

$$y_i = \frac{x_i - x_{mean}}{SD_x} \tag{3.2}$$

where x_{mean} and SD_x are the mean and standard deviation of $\{x_i\}$ respectively.

3.3.2 Training of BPNN models

The aim of the training process is to relatively adjust the weights and biases of the network to minimize the network performance function. There are several different training algorithms for feed forward networks which use the gradient of the performance function to determine how to adjust the weights to minimize performance. Backpropagation procedure is used to determine the gradient, which involves performing computations backward through the network. Levenberg–Marquardt backpropagation training algorithm is used in the present study.

Levenberg–Marquardt backpropagation training algorithm
Assume a network with performance function $\eta(w)$as the sum of the nonlinear least squares between the observed and the predicted outputs, defined by (Coulibaly et al., 2000):

$$\eta(w) = \frac{1}{2}\sum_{i=1}^{n_p}\sum_{k=1}^{F}(x_{ik} - \hat{x}_{ik})^2 \tag{3.3}$$

where w represents the weights and biases of the network, n_p is the total number of training patterns, F is the total number of neurons in the output layer, x_{ik} represents the observed output (target output) for the input i at the output neuron k, $\hat{x}_{ik}$ represents the actual output for the input i at the output neuron k.

The original Levenberg-Marquardt algorithm is suggested by Levenberg (1944) and later modified by Marquardt (1963) to solve the nonlinear least squares problems. The following steps are involved in training a neural network in batch mode using the Levenberg-Marquardt algorithm (Coulibaly et al., 2000; Matlab.a, 2008; Ranga Suri and Nagabhushan, 2002):

1. Present all inputs to the network and compute the corresponding network outputs, errors and the mean square error over all inputs, $\eta(w)$ as in equation (3.3).

2. Compute the Jacobian matrix, $J(w)$ where $J^T(w)J(w)$ referred to as the Hessian matrix.

3. Estimate Δw using the Levenberg-Marquardt algorithm which uses an approximation to the Hessian matrix in the following Newton-like update as:

$$\Delta w = \left[J^T(w)J(w) + \mu I\right]^{-1} J^T(w)e(w) \tag{3.4}$$

where μ is training parameter, I is the identity matrix, and *e(w)* is a residual error vector of size $F \times n_p$ and is calculated as follows:

$$e(w) = \begin{pmatrix} d_{11} - a_{11} \\ d_{12} - a_{12} \\ .... \\ d_{21} - a_{21} \\ d_{22} - a_{22} \\ .... \\ .... \\ d_{Fn_p} - a_{Fn_p} \end{pmatrix}$$

4. Compute the error using $w + \Delta w$. If this new error is smaller than that computed in step 1, then reduce the training parameter μ by μ^{-1}. Let $w = w + \Delta w$ and go back to step 1. If the error is not reduced, then increase μ by $\mu^{+1}$and resume the training from step 3. The parameters μ^{-1} and $\mu^{+1}$are predefined by the user and typically set to 0.1 and 10 respectively.

The algorithm is assumed to have converged when the norm of the gradient $g = J^T e(w)$ is less than some predetermined value, or when the error is reduced to some error goal.

3.3.3 Optimum size of neural networks

Choosing the optimum network size is one of the most important challenges that face the neural network designers. There is no way to determine the best number of hidden units without training several networks and estimating the generalization error of each network. Few hidden units, lead to high training error and high generalization error due to underfitting and high statistical bias. There are many methods for estimating generalization error. However too many hidden units produce low training error but still have high generalization error due to overfitting and high variance. In the following are some of these methods:

- *Split-sample* is the most commonly used method for estimating generalization error. We divide the available data into three sets (training x_{train}, validation $x_{validate}$ and test x_{test} sets). The training set is then used to train the models and the test set to test the generalization ability of the model.

- *Cross-validation* is an improvement on split-sample validation that allows us to use all of the data for training.

- *Bootstrapping* is an improvement on cross-validation that often provides better estimates of generalization error.

Larsen (1994) suggested optimizing the architecture by selecting the model with minimal estimated averaged generalization error. He concluded that the network architecture with minimal estimated average generalization error is selected as being optimal. Several methods are available for determining the optimal network size, e.g., the network information criterion, NIC (Murata, et al., 1994), the generalized final prediction error, GPE (Moody, 1992) and the Vapnik-Chervonenkis, VC dimension (Bartlett and Maass, 2003). Lawrence et al. (1996) preformed a large scale numerical study on the optimal size and they showed that a solution near the optimal solution is often not obtained. A neural network can have any number of hidden layers, but in general, one hidden layer is sufficient (Berry and Linoff, 2004). Wang et al. (2005) conducted systematic numerical experiments to study the effect of the network size on the performance of the neural network and they found that there is no relationship between the network size and its performance.

3.3.4 Improving generalization of the neural networks

Samarasinghe (2006) discussed the following approaches to avoid underfitting and overfitting (improve generalization of Neural Networks) for a given fixed amount of training data:

Regularization

In regularization, to keep the weights away from getting large, a regularized performance index (W) is used instead of the mean square error (mse) by adding a sum of square weights $\{{w_j}^2\}$, $j = 1, 2, \ldots, m$ as given in following equation:

$$W \; = \gamma \times mse + (1-\gamma)\sum_{j=1}^{m} w_j^2 \tag{3.5}$$

Where γ is the performance ratio and w_j is a weight in the total set of m weights in the network.

Early stopping

A model overfits if it has too much flexibility that is expressed by the number of free parameters (i.e., weights). At this point the network seems to get better and better, i.e., the error on the training set (x_{train}) decreases but actually it begins to get worse, i.e., the error on the validation set ($x_{validate}$) increases as shown in figure 3.2.

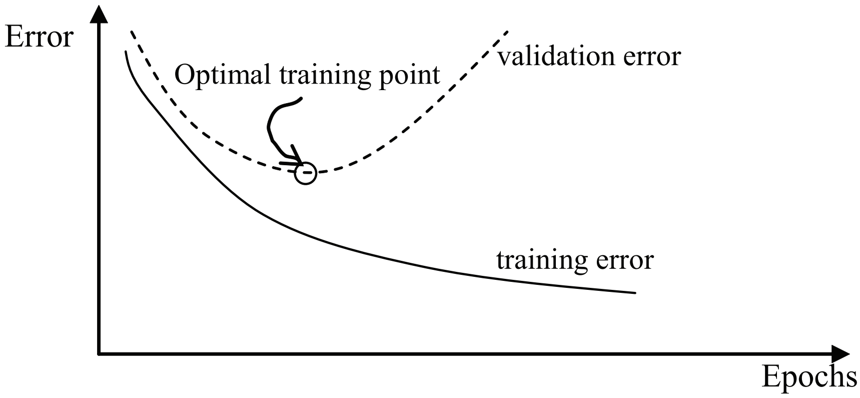


Figure 3.2: *Early stopping for improving generalization*

Early stopping proceeds as follows (Matlab.a, 2008):

- Divide the available data x into three sets, the training set x_{train}, the validation set $x_{validate}$ and the test set x_{test}.
- Use x_{train} for computing the gradient and updating the network weights and biases.
- Start the training process and monitor the error of $x_{validate}$. During the initial phase of training, the error of $x_{validate}$ decreases as does the error of x_{train}.
- Stop the training processes when the validation error increases (beginning of overfitting).
- Return the weights and biases at the minimum of the validation error.
- Test the generalization of the network using the test data set x_{test}.

We can get a good approximation of the estimated error, if x_{train}, $x_{validate}$ and x_{test} fully reflect the probability distribution of the observed data (Rojas, 1996).

Exhaustive search

This is the simplest way but the most time consuming. It uses a trial-and-error method to search the optimum number of hidden layers and number of neurons in each hidden layer. Estimate the error on the validation set ($x_{validate}$) for each trial. The optimum model is the network that gives the minimum error in the validation set.

3.3.5 Activation Functions

The behavior of an artificial neural network (ANN) depends on both the weights and the input-output function (transfer function) that is specified for the units. This process is sometimes called as activation function which has two parts as shown in figure 3.3 (Berry and Linoff, 2004). The first part is the combination function that merges all the inputs into a single value. Weighted sum is the most common combination function, where each input is multiplied by its weight and the products are added together to produce w_p. A bias is then added to w_p to produce wt_j' which is the output of the combination function. The bias is much like a weight except that it has a constant input of 1. The bias is also called the threshold term and is defined as the input to a neuron in the absence of any other inputs (Detienne et al., 2003). This sum wt_j' is passed to the transfer function, f (the second part of the activation function) to get the neuron's output (α). Transfer functions are needed to introduce nonlinearity into the network to make multilayer networks so powerful. The most commonly used transfer functions are hard-limit, linear, log-sigmoid and tan-sigmoid (figure 3.4).

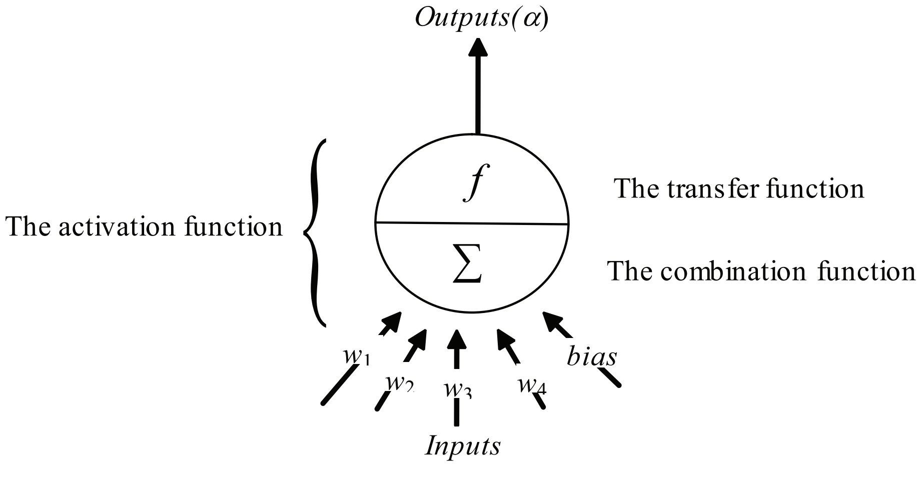


Figure 3.3: *The unit of an artificial neural network*

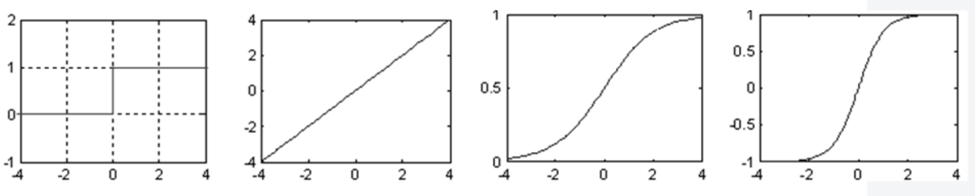


Figure 3.4: *a) Hard-limit b) Linear c) Log-sigmoid d) Tan-sigmoid The hard-limit, linear, log-sigmoid and tan-sigmoid transfer functions*

3.4 Adaptive neuro-fuzzy inference system (ANFIS)

Adaptive neuro-fuzzy inference system (ANFIS) is employed to simulate the daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs. The concept of fuzzy logic was emerged in the development of the theory of fuzzy sets by Zadeh (1965). Fuzzy logic allows intermediate values to be defined between conventional evaluations like true/false, yes/no (Hellmann, 2001). It is a form of multi-valued logic that is derived from fuzzy set theory to deal with reasoning that is approximate rather than precise. A fuzzy set is any set that allows its members to have different grades of membership in the interval [0,1]. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. Mamdani (Mamdani and Assilian, 1975) and Sugeno (Takagi and Sugeno, 1985) are two types of fuzzy inference systems. The main difference between Mamdani and Sugeno is that the Sugeno output membership functions are either linear or constant. Jang (1993) described the architecture of adaptive-network-based fuzzy inference system (ANFIS) based on the Sugeno inference system type. ANFIS is a hybrid intelligent system which has the ability of fuzzy logic (FL) to reason with neural network (NN) to learn. The goal of ANFIS is to find a model which will simulate correctly the inputs with the outputs. Fuzzy inference system (FIS) is a knowledge representation where each fuzzy rule describes a local behavior of the system. ANFIS is the network structure that implements FIS and employs hybrid-learning rules to train (Loukas, 2001). One of the best features of ANFIS is that it pre-processes all the data into several membership functions before mapping the data into an adaptive neuro structure. This pre-processing feature allows ANFIS to converge faster and better.

3.4.1 ANFIS structure

For simplicity, a fuzzy inference system has two inputs x and y and one output is assumed. For a first-order Sugeno fuzzy model, a common rule set with two fuzzy if–then rules is defined as:

1. If x is A_1 and y is B_1, then $f_1 = p_1 x + q_1\ y + r_1$

2. If x is A_2 and y is B_2, then $f_2 = p_2 x + q_2\ y + r_2$

in which p_1, p_2, q_1, q_2, r_1 and r_2 are linear parameters (consequent parameters) and A_1, A_2, B_1 and B_2 are nonlinear parameters (premise parameters).

Figure 3.5.a illustrates the reasoning mechanism for this Sugeno model, the corresponding equivalent ANFIS architecture is shown in figure 3.5.b, where the nodes of the same layer have similar functions, as described next (Jang, 1993; Li and et al., 2007):

Layer 1. Is the fuzzy layer, in which x and y are the input of nodes A_1, A_2, B_1 and B_2. The membership relationship between the output and input functions of this layer can expressed as:

$$\begin{aligned} O_{1,i} &= \mu A_i(x) \qquad & i = 1,2 \\ O_{1,j} &= \mu B_j(y) \qquad & j = 1,2 \end{aligned} \tag{3.6}$$

where $O_{1,i}$ and $O_{1,j}$ denote the output functions and μ_{Ai} and μ_{Bj} denote the membership functions. The generalized symmetric Gaussian function is used in the present study as membership function and it is given as follows

$$\mu(z) = e^{\frac{-(z-c)^2}{2b^2}} \tag{3.7}$$

where b and c are the parameters of the function which determine the shape of the function. Every node in this layer is an adaptive node. Parameters in this layer are called parameters or nonlinear parameters.

Layer 2. Every node in this layer is a fixed node labeled Π, whose output is the product of all the incoming signals. The output w_1 and w_2 are the weight functions of the next layer.

Layer 3. Every node in this layer is a fixed node labeled Ω. The i^{th} node calculates the ratio of the i^{th} rule's firing strength. Thus the outputs of this layer are called normalized firing strengths.

Layer 4. Every node i in this layer is an adaptive node. Parameters in this layer are referred to as consequent parameters (linear parameters).

Layer 5. The single node in this layer is a fixed node labeled (Σ) which computes the overall output as the summation of all incoming signals.

3.4.2 Learning algorithm for ANFIS

The learning algorithm which is used in the present study for ANFIS is a hybrid algorithm, which is a combination of gradient descent and the least-squares method. More specifically, in the forward pass of the hybrid learning algorithm, node outputs go forward until layer 4 and the consequent parameters are identified by the least-squares method (Jang, 1993). In the backward pass, the error signals propagate backwards and the premise parameters (nonlinear parameters) are updated by gradient descent. The consequent parameters are optimized under the condition that the premise parameters are fixed. The proposed hybrid approach converges much faster since it reduces the search space dimensions of the original pure backpropagation method used in neural networks. The overall output can be expressed as a linear combination of the consequent parameters (see figure 3.5.a). The error, η, which is used to train the above-mentioned ANFIS is defined as:

$$\eta = \sum_{k=1}^{n_p} (x_k - \hat{x}_k)^2 \tag{3.8}$$

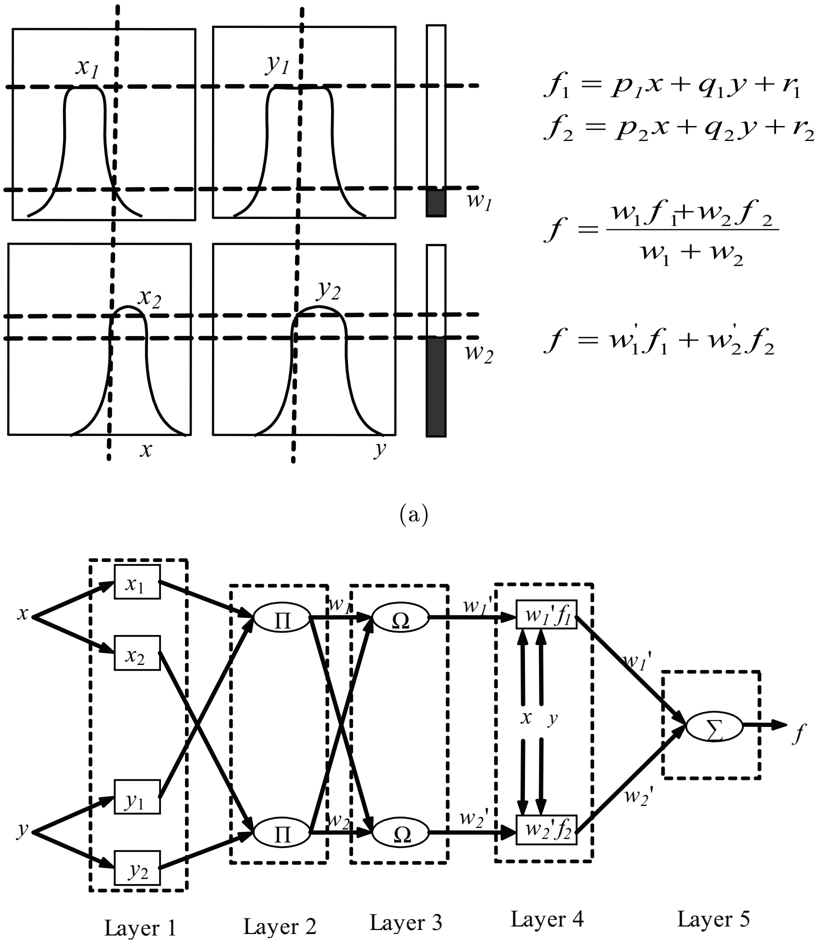
where x_k and $\hat{x}_k$ are the k^{th} desired and estimated output, respectively and n_p is the total number of pairs (inputs–outputs) of data in the training set.

3.5 Autoregressive moving average ARMA(p,q) processes

3.5.1 Introduction to ARMA models

The autoregressive-moving-average (ARMA) models are mathematical models of the persistence, or autocorrelation, in a time series. They are widely used in hydrology, economics and many other fields.

The ARMA model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually then referred to as the ARMA(p,q) model where p is the order of the autoregressive part and q is the order of the moving average part. Brockwell



(a)

(b)

Figure 3.5: *a) An illustration of the reasoning mechanism for a Sugeno-type model, b) ANFIS architecture*

and Davis (2002) mentioned that $\{x_t\}$, $t = 1, 2, \ldots, n$ is an ARMA(p,q) process if $\{x_t\}$ is stationary and if for every t,

$$x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{3.9}$$

where $\{\varepsilon_t\}$ is the error term with zero mean and variance σ^2. The error term $\{\varepsilon_t\}$ is assumed to be identically and independently distributed *i.i.d.*

The process $\{x_t\}$ is said an ARMA(p,q) process with mean ţ if $(x_t - \mu)$ is an ARMA (p,q) process. It is convenient to use the more concise form of equation (3.9)

$$\phi(B)x_t = \theta(B)\varepsilon_t \tag{3.10}$$

where $\phi(\cdot)$ and $\theta(\cdot)$ are the p^{th} and q^{th}-degree polynomials

$$\phi(B) \; = (1 - \phi_1 B - \phi_2 B^2 \; \cdots - \; \phi_p B^p) \tag{3.11}$$

and

$$\theta(B) \; = (1 + \theta_1 B + \theta_2 B^2 \; \cdots + \; \theta_q B^q) \tag{3.12}$$

and B is the backward shift operator ($B^j x_t = x_{t-j}$, $B^j \varepsilon_t = \varepsilon_{t-j}$, $j = 0, \pm 1, \cdots$). The time series $\{x_t\}$ is said to be an autoregressive process of order p or AR(p) if $\theta(B) \equiv 1$, and moving–average process of order q or MA(q) if $\phi(B) \equiv 1$. When neither p nor q is zero, an ARMA(p,q) model is sometimes referred to as a "mixed model". The main stages in setting up an ARMA forecasting model are as follows (Box and Jenkins, 1976):

1. ARMA model identification
2. Model parameters estimation
3. Diagnostic checking

3.5.2 ARMA model identification

There are various methods and criteria for selecting the orders p and q of an ARMA(p,q). Autocorrelation function (ACF) and partial autocorrelation function ($PACF$) can provide powerful tools to determine the order of a pure autoregressive (AR) or moving average (MA) process. Another method to select the orders p and q is using the so-called information criteria. Two information criteria for statistical model identification are proposed by Akaike (1974) and (1979). These two information criteria are known as Akaike information criteria, AIC (Akaike, 1974) and Bayseian information criteria, BIC (Akaike, 1979) and are defined as follows:

$$AIC = \frac{2m}{n} + \log\left(\hat{\sigma}^2\right) \tag{3.13}$$

in which n is the number of observations, $\hat{\sigma}^2$ is the estimated noise variance which is usually obtained from maximum likelihood and m is the number of parameters.

$$BIC = \frac{m \log n}{n} + \log\left(\hat{\sigma}^2\right) \tag{3.14}$$

The aim of using *AIC* or *BIC* is to balance the risks of underfitting (selecting orders smaller than the true orders) and overfitting (selecting orders larger than the true orders) by minimizing the estimated value of *AIC* or *BIC* (Mathimatica, 2007). It is difficult to compare the relative performance of the two criteria (Wang and Langari, 1995), however Wang and Libert (1994) suggested that the *BIC* is superior to the *AIC* in identifying the ARMA models in some aspects.

3.5.3 Estimation of the parameters of ARMA model

We used the maximum likelihood estimation method to estimate the parameters of the ARMA models. Maximum likelihood estimation aims to find the most likely values of distribution parameters for a set of data by maximizing the value of what is called the likelihood function. This likelihood function is largely based on the probability density function (*pdf*) for a given distribution. An ARMA(p,q) model simulates an observed data $\{x_t\}$ t = 1, 2, ..., n as follows:

$$x_t = \mu + \phi_1 x_{t-1} + \ \cdots + \phi_p x_{t-p} - \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ \cdots + \ \theta_q \varepsilon_{t-q} \tag{3.15}$$

where $\varepsilon_t \sim i.i.d.\ N\left(0, \sigma^2\right)$.

The goal is to estimate population parameters Φ where

$$\Phi = (\mu, \phi_1, \phi_2, \cdots, \phi_p, \theta_1, \cdots \theta_2,\ \theta_q, \sigma^2) \tag{3.16}$$

The maximum likelihood estimate of Φ is the value that maximize the probability density

$$f_{x_n, x_{n-1}, \ldots, x_2, x_1}(x_n, x_{n-1}, ..., x_2, x_1; \Phi) \tag{3.17}$$

The approximation to the likelihood functions for an autoregressive and moving average processes conditions on the initial values of the x's and the ϵ's respectively (Hamilton, 1994). Take the initial values for $x_o \equiv (x_o,\ x_{-1},\ \ldots,\ x_{-p+1})$ to be equal to their actual values and estimate the initial values $\epsilon_o \equiv (\epsilon_o,\ \epsilon_{-1},\ \ldots,\ \epsilon_{-q+1})$, the sequence $\{\epsilon_1,\ \epsilon_2,\ \ldots,$ $\epsilon_n\}$ can be calculated from $\{x_1,\ x_2,\ \ldots,\ x_n\}$ using the following formula:

$$\varepsilon_t = x_t - \phi_1 x_{t-1} - \ \cdots - \phi_p x_{t-p} - \theta_1 \varepsilon_{t-1} - \ \cdots - \ \theta_q \varepsilon_{t-q} \tag{3.18}$$

for $t = 1, 2, 3, \ldots, n$. If ϵ_t is Gaussian, the conditional log-likelihood function (LLF) is then

$$\begin{aligned} LLF &= \log f_{x_n, x_{n-1}, \ldots, x_1 | x_0, \varepsilon_0}(x_n, x_{n-1}, \ldots, x_1 | \, x_0, \varepsilon_0; \Phi) \\ &= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^2) - \textstyle\sum_{t=1}^{n} \frac{\varepsilon_t^2}{2\sigma^2} \end{aligned} \tag{3.19}$$

Evaluation of the log-likelihood function is conditioned, or based on a set of pre-sample observations. For this reason, the log-likelihood objective functions shown here are referred to as conditional log-likelihood functions. The iterative numerical optimization method can be used to estimate the optimum Φ.

3.5.4 Diagnostic checking

The model selection criterion (AIC) is used to choose the "best" ARMA models. The AIC penalizes the models for their complexity; therefore, diagnostic checking is needed to ensure that the residuals of the models are random. Based on the expected properties of the residuals, $\varepsilon_t(t = 1, 2, \ldots, n)$ we used the following diagnostic checks (Brockwell and Davis, 2002):

The sample autocorrelation function (ACF) of the residuals

If the ARMA model effectively describes the persistence of the data, then the model residuals should be random or uncorrelated in time. The ACF of the residuals can be examined by scanning it to see if any individual coefficients fall outside some specified confidence interval (CI) around zero. Assume r_k is the autocorrelation coefficient of the residuals at lag k. The appropriate CI for r_k can be found by referring to a normal distribution *cdf*, then the 95% confidence interval for r_k is $\pm 1.96/n^{0.5}$. The noise hypothesis that observed residuals are consistent with *i.i.d.* noise can be rejected if more than two or three out of 40 fall outside the bounds $\pm 1.96/n^{0.5}$ ($n_{out}/m = 2/40$ or $3/40$) where n_{out} is the number of the estimated ACF which fall out the bounds and m is the number of coefficients to test autocorrelation. We can also reject the noise hypothesis if one of the estimated ACF of the residuals falls far outside the bounds.

Test for Randomness of the Residuals

Instead of looking at the residual autocorrelations one at a time, it is possible to carry out what is called a portmanteau lack-of-fit test (Chatfield, 2004). The classical portmanteau test statistic is the one proposed by Box and Pierce (1970)

$$Q_{BP} = n \sum_{k=1}^{m} r_k^2 \tag{3.20}$$

where n is the number of observations, m is the number of coefficients to test autocorrelation (25 is a reasonable number for many time series, Li, 2004) and r_k is the sample autocorrelation of order k of residual. Under the null hypothesis that the ARMA model is adequate, $Q_{\mathrm{BP}}$is distributed as a $\chi^2$with $(m-p-q)$degrees of freedom. We reject the null hypothesis at level α if $Q_{\mathrm{BP}} > \chi^2_{1-\alpha}(m-p-q)$,where $\chi^2_{1-\alpha}(m-p-q)$is the $1-\alpha$ quantile of the chi-squared distribution with $(m-p-q)$degrees of freedom. Ljung and Box (1978) suggested an alternative formula in which $Q_{\mathrm{BP}}$is replaced by

$$Q_{\mathrm{LB}} = n(n+2)\sum_{k=1}^{m} r_k^2/(n-m) \tag{3.21}$$

Choice of m equal to 20 is somewhat arbitrary, 25 is a reasonable number for many series.

3.6 Fractionally integrated ARMA processes (ARFIMA)

3.6.1 Introduction to ARFIMA models

The autocorrelation function $\rho(\cdot)$ of an ARMA process at lag h converges rapidly to zero as $h \to \infty$ in the sense that there exists C > 1 such that (Brockwell and Davis, 2002):

$$c^h\rho(h) \to 0 \quad \text{as} \quad h \to \infty \tag{3.22}$$

Fractionally integrated ARMA processes (ARFIMA) are stationary processes with much more slowly decreasing autocorrelation function. Brockwell and Davis (2002) defined ARFIMA processes as autoregressive integrated moving-average ARIMA(p,d,q) processes with $-0.5 \le d \le 0.5$, satisfy difference equations of the form

$$\phi(B)(1-B)^d X_t = \theta(B)\varepsilon_t \tag{3.23}$$

where $\phi(\cdot)$,$\theta(\cdot)$ and B are predefined in section 3.5.1 and $\{\varepsilon_t\}$ is an *i.i.d.* process with mean 0 and variance σ^2. The operator $(1-B)^d$is defined by binomial expansion

$$(1-B)^d = \sum_{j=0}^{\infty} \pi_j B^j \tag{3.24}$$

where $\pi_j = \prod\limits_{0<k\le j} \frac{k-1-d}{k}, \qquad j = 1,\, 2,\, \cdots.$

If the parameter d is in the range [-0.5, 0], the process is said to be antipersistent; if $d = 0$ the process is either a short memory process or a white noise; finally for d in the range [0,0.5] the process exhibits long memory (Coli et al., 2005).

3.6.2 Estimation of an ARFIMA(p,d,q) model

Maximum likelihood estimation may be used to estimate the parameters ϕ, θ, and $d$of ARFIMA (p,d,q) process. Haslett and Raftery (1989) proposed a fast and accurate method to calculate the exact maximum likelihood estimate (MLE). Sowell (1992) described how to compute the exact maximum likelihood estimate (MLE) for a stationary ARFIMA model with $-0.5 \leq d \leq 0.5$. Brockwell and Davis (2002) suggested that it is much simpler to estimate the parameters of the of ARFIMA (p,d,q) process by maximizing the Whittle approximation instead of maximizing the exact Gaussian likelihood.

ARFIMA(p,d,q) process can be regarded as an ARMA(p,q) process driven by fractional integrated noise (Brockwell and Davis, 2002). The process defined in equation (3.23) can be replaced by the following two equations

$$\phi(B)X_t = \theta(B)W_t \tag{3.25}$$

and

$$(1-B)^d W_t = \varepsilon_t \tag{3.26}$$

Reisen et al. (2001) gave the following procedure to build an ARFIMA(p,d,q) for the $\{X_t\}$ process defined in equation (3.23):

1. Estimation of d in ARFIMA(p,d,q) using the GPH method given in chapter 2, section 2.6.3 or any other method and denote the estimate by $\hat{d}$.

2. Calculation of $\hat{U}_t = (1-B)^{\hat{d}} X_t$.

3. Identification and estimation of ϕ and θ in the ARMA(p,q) process $\phi(B)\hat{U}_t = \theta(B)\varepsilon_t$, using Box-Jenkins modeling (Box et al., 1976) or the maximum likelihood estimation method (section, 3.5.3).

4. Calculation of $\hat{W}_t = \frac{\hat{\phi}(B)}{\hat{\theta}(B)} X_t$, in which $\hat{\phi}(B)$ and $\hat{\theta}(B)$ are the estimated values of $\phi(B)$ and $\theta(B)$ from the previous step.

5. Re-estimation of d using in the ARFIMA(p,d,q) model $(1-B)^{\hat{d}}\hat{W}_t = \varepsilon_t$.

6. Using the new estimate of d, repeat steps 2 to 5 until the estimates of parameters d, ϕ and θ converge.

We wrote a Matlab code to estimate the parameters of ARFIMA(p,d,q) using the previous procedure.

3.7 Forecasting using ARMA models

The goal of the forecasting process is to predict future values of a time series, x_{n+m}, $m = 1$, 2, ..., based on the data collected to the present, $x = \{x_n, x_{n-1}, \dots, x_1\}$. Let $\{x_t\}$ be a stationary process with expectation, $E[\{x_t\}]$=0 and autocovariance function (γ). An observation x_{n+1} beyond the end of a series can be estimated as linear combination of x_1, $x_2, \dots, x_n$, i.e.,

$$\hat{x}_{n+1} = \sum_{j=1}^{n} \Phi_{n_j} x_j \tag{3.27}$$

such that the expectation of the mean squared error is given as:

$$E\left|x_{n+1} - \hat{x}_{n+1}\right|^2 \tag{3.28}$$

Best linear predictors (BLPs) of equation (3.27) are the predictors that minimize the mean square prediction error (equation, 3.28). Using the projection theorem, we can rewrite equation (3.28) as (Shumway and Stoffer, 2006)

$$E\left[\left(x_{n+1} - \sum_{j=1}^{n} \Phi_{n_j} x_{n+1-j}\right) x_{n+1-k}\right] = 0, \qquad k = 1, 2, ..., n \tag{3.29}$$

which can be written as

$$\sum_{j=1}^{n} \Phi_{n_j} \gamma\left(k - j\right) = \gamma\left(k\right), \qquad k = 1, 2, ..., n \tag{3.30}$$

Combining all the equations of the autocovariances in a matrix form, we have

$$\gamma_n = M_n \Phi_n \tag{3.31}$$

Due to the projection theorem, if M_n is nonsingular, the elements of $\Phi_n$are unique and are given by

$$\Phi_n = M_n^{-1} \gamma_n \tag{3.32}$$

The mean square one-step-ahead prediction error is

$$P_{n+1}^{n} = E(x_{n+1} - x_{n+1}^{n})^2 = \gamma(0) - \gamma{\prime}_n M_n^{-1} \gamma_n \tag{3.33}$$

Shumway and Stoffer (2006) used the Durbin-Levinson algorithm to solve equations (3.32) and (3.33) iteratively as follows:

$$\Phi_{00} = 0, \qquad p_1^0 = \gamma(0) \tag{3.34}$$

For $n \geq 1$,

$$\Phi_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1} \Phi_{n-1,k}\rho(n-k)}{1 - \sum_{k=1}^{n-1} \Phi_{n-1,k}\rho(k)}, \qquad p_{n+1}^{n} = p_n^{n-1}(1 - \Phi_{nn}^2), \tag{3.35}$$

where for $n \geq 2$, $\Phi_{nk} = \Phi_{n-1,k} - \Phi_{nn}\Phi_{n-1,n-k}, \qquad k = 1, 2, \cdots, n-1$ and $\rho(n) = \frac{\gamma(n)}{\gamma(0)}$, is the autocorrelation.

Estimation of autocovariances $(ACVF)$

The calculation of the $ACVF$ of an ARIMA and ARFIMA processes is a crucial aspect in the implementation of the Durbin-Levinson algorithms (Palma, 2007). The method used to estimate autocovariances $\gamma(.)$ of ARMA(p, q) follows Brockwell and Davis (1991). By multiplying both sides of equation (3.9) by x_{t-k} and taking the expectations we obtained

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = \sigma^2 \textstyle\sum_{k \leq i \leq q} \theta_i \psi_{i-k}, \qquad 0 \leq k < \max(p, q+1) \tag{3.36}$$

and

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0 \qquad k \geq \max(p, q+1) \tag{3.37}$$

where

$$\psi_j - \sum_{0<s\leq j} \phi_s \psi_{j-s} = \theta_j, \qquad 0 \leq j < \max(p, q+1) \tag{3.38}$$

and

$$\psi_j - \sum_{0<s\leq p} \phi_s \psi_{j-s} = 0, \qquad j \geq \max(p, q+1) \tag{3.39}$$

Sowell (1992) derived an expression for the computation of $ACVF$ of an ARFIMA process which involves hypergeometric functions. Palma (2007) used the so-called splitting method to estimate $ACVF$ of an ARFIMA process. He decomposed the ARFIMA model into its ARMA and its fractionally integrated (FI) parts. Then, the $ACVF$ of the corresponding ARFIMA process is given by

$$\gamma(k) \approx \sum_{h=-m}^{m} \gamma_0(h)\gamma_{\text{ARMA}}(k-h) \tag{3.40}$$

where $\gamma_0(.)$ is the $ACVF$ of the fractional noise ARFIMA(0,d,0) and $\gamma_{ARMA}(.)$ is the $ACVF$ of the ARMA component. $ACVF$ of ARFIMA(0,d,0) process is given by

$$\gamma_0(h) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} \frac{\Gamma(h+d)}{\Gamma(1+h-d)} \tag{3.41}$$

where $\Gamma(.)$ is the gamma function.

3.8 Models efficiency criteria

We used the following efficiency criteria as a mathematical estimate of the error between the predicted and observed daily inflow data to evaluate the performances of the models:

3.8.1 Average relative error percentage (AREP)

$$AREP \quad = \quad 100 \times \frac{1}{n} \sum_{j=1}^{n} |(x_j - \hat{x}_j)/x_j| \tag{3.42}$$

in which n is the total number of the observed data and $x_j$and $\hat{x}_j$ are the observed and predicted daily inflow respectively.

3.8.2 Cross correlation coefficient ($\boldsymbol{R}$)

$$R \quad = \quad \frac{\sum_{j=1}^{n}(x_j - \bar{x})(\hat{x}_j - \bar{\hat{x}})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(\hat{x}_j - \bar{\hat{x}})^2}} \tag{3.43}$$

where $\bar{x}$ and $\bar{\hat{x}}$ are the means of the observed and predicted daily inflow data respectively.

3.8.3 Nash–Sutcliffe model efficiency coefficient (NSC)

Nash and Sutcliffe (1970) proposed a coefficient, NSC to estimate the efficiency of the fit which takes values from 1 (best fit) to -∞. Values of NSC lower than zero indicate that the mean of the observed data are as accurate as the model predictions. The disadvantage of using the efficiency coefficient (NSC) is that the larger values in a time series are strongly overestimated whereas lower values are neglected (Krause et al., 2005). To reduce the problem of the squared differences, Krause et al. (2005) suggested estimating the efficiency coefficient (NSC) using the log-transformed values of the observed x_j and the predicted $\hat{x}_j$ inflow data as follows

$$NSC \; = 1 - \frac{\sum_{j=1}^{n}(x_j - \hat{x}_j)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{3.44}$$

They have also estimated another form of the efficiency coefficient (NSC) using the differences between the observed and predicted values as relative deviations to reduce the influence of high and low values as follows:

$$NSC_{rel} \; = 1 - \frac{\sum_{j=1}^{n}\left(\frac{x_j - \hat{x}}{x_j}\right)^2}{\sum_{j=1}^{n}\left(\frac{x_j - \bar{x}}{\bar{x}}\right)^2} \tag{3.45}$$

3.8.4 Index of agreement (g)

The index of agreement (g) was proposed by Willmot (1981) to overcome the insensitivity of NSC to differences in the observed and predicted means and variances. The range of g lies between 0 (no correlation) and 1 (perfect fit) and is defined as:

$$g \; = 1 - \frac{\sum_{j=1}^{n}(x_j - \hat{x}_j)^2}{\sum_{j=1}^{n}(|x_j - \bar{x}| + |\hat{x}_j - \bar{x}|)^2} \tag{3.46}$$

3.9 Building of the models

3.9.1 BPNN models

The Levenberg-Marquardt is used as training algorithm and tan-sigmoid and linear functions as transfer functions for the hidden layer neurons and for the output one respectively. Early stopping procedure is employed to prevent overfitting of the BPNNs in the present work by dividing the available data x_t into three sets x_{train}, $x_{vaidate}$ and x_{test} with 60 %, 20 % and 20 % percents from available data respectively. One hidden layer is assumed to be sufficient to simulate the training data (Berry and Linoff, 2004). We used trial-and-error procedure to determine the optimum number of neurons in the hidden layer and to select the input variables that give the best performance. Selection of input variables for the BPNN model has been done after making extensive trials with different combinations of input variables as in the following procedure:

1. Train the BPNN model using each input variables combination with number of neurons (F), $m_s \leq F \leq 1.5 \times m_s$ where m_s is the number of the input variables in each combination. The number of combinations (N_s) is equal to $m!\times$2-1 where m is the number of the potential input variables (see table 3.1).

2. Use the previous step to determine the optimum F corresponds to the each input variables combination.

3. Now, choose the BPNN model with the best performance as the optimal one.

Table 3.2 gives a list of the input-output variables and the number of neurons in the hidden layer for the optimum models for each reservoir.

3.9.2 ANFIS models

Selection of the ANFIS model with best performance is the aim of this section. Starting with the input variables of the optimum BPNN models, the following procedure is assumed to find out the optimum ANFIS models:

1. Use the input variables of the best BPNN models (table 3.2) as potential input variables of the ANFIS models. Take the number of membership functions for each input variable to be 2 except for input variables at day t-1 (if exist) for which three membership functions are used.

2. Train the model using each input variables combination and estimate its performance η (equation, 3.8).

3. Compare between the performances of the different trained models and choose the model with the best performance to be the optimum one.

Early stopping criteria provided by the validation data sets are used to prevent overfitting of the ANFIS models. A list of the input-output variables and the number of membership functions for each input variable for the best ANFIS models is given in table 3.3. Both the BPNN and ANFIS models are trained and simulated using Matlab 7.5 developed by the Math Works Inc, Natick, Massachusetts.

Table 3.2: *Parameters of the BPNN models*

Reservoir	Models group	Optimum inputs	Optimum no. of neurons	Output
Bigge	M1-1	x_{t-1}, x_{t-2}, x_{t-3}	4	x_t
	M1-2	x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}	6	x_t
	M2	x_{t-1}, x_{t-2}, D_{t-1}, D_{t-2}	5	x_t
Henne	M1-1	x_{t-1}, x_{t-2}	7	x_t
	M1-2	x_{t-1}, x_{t-2}, x_{t-3}	5	x_t
	M2	x_{t-1}, x_{t-2}, x_{t-3}, D_{t-1}, D_{t-2}	9	x_t
Möhne	M1-1	x_{t-1}, x_{t-2}, x_{t-3}	6	x_t
	M1-2	x_{t-1}, x_{t-2}, x_{t-3}	5	x_t
	M2	x_{t-1}, x_{t-2}, D_{t-1}	4	x_t
Sorpe	M1-1	x_{t-1}, x_{t-2}, $x_{t-3}3$, x_{t-4}	3	x_t
	M1-2	x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}	5	x_t
	M2	x_{t-1}, x_{t-2}, D_{t-1}	6	x_t

3.9.3 ARMA models

For mixed models it is often difficult to determine the exact orders of the ARMA. In the present work, we assumed that neither the AR nor MA component has a higher order than 5 ($0 \leq p \leq 5$ and $0 \leq q \leq 5$).

We tried different ARMA models with different orders of AR and MA components for each data series and compared the *AIC* values of the models. The ARMA models with the minimum *AIC* are assumed to be the best models. Table 3.4 gives the parameters of the best ARMA models (models, group M1-1 and models, group M1-2) for each reservoir.

***Table 3.3:** Parameters of the ANFIS models*

Reservoir	Models group	Inputs	No. of membership functions	Output
Bigge	M1-1	x_{t-1}, x_{t-2}, x_{t-3}	3 - 2 - 2*	x_t
	M1-2	xt-1, x_{t-2}, x_{t-3}, x_{t-4}	3 - 2 - 2 - 2	x_t
	M2	x_{t-1}, x_{t-2}, D_{t-1}, D_{t-2}	3 - 2 - 3 - 2	x_t
Henne	M1-1	x_{t-1}, x_{t-2}	3 - 2	x_t
	M1-2	x_{t-1}, x_{t-2}, x_{t-3}	3 - 2 - 2	x_t
	M2	x_{t-1}, x_{t-2}, x_{t-3}, D_{t-1}, D_{t-2}	3 - 2 - 2 - 3 - 2	x_t
Möhne	M1-1	x_{t-1}, x_{t-2}, x_{t-3}	3 - 2 - 2	x_t
	M1-2	x_{t-1}, x_{t-2}, x_{t-3}	3 - 2 - 2	x_t
	M2	x_{t-1}, x_{t-2}, D_{t-1}	3 - 2 - 3	x_t
Sorpe	M1-1	x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}	3 - 2 - 2 - 2	x_t
	M1-2	x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}	3 - 2 - 2 - 2	x_t
	M2	x_{t-1}, x_{t-2}, D_{t-1}	3 - 2 - 3	x_t

* the number of membership functions for input variables x_{t-1}, x_{t-2}, x_{t-3} are 3, 2 and 2 respectively.

***Table 3.4:** Parameters of the best autoregressivge moving average ARMA(p,q) models*

Reservoir	Models group	Parameters of the model					
Bigge	M1-1	ϕ	-0.859	-	-	-	-
		θ	0.258	-0.084	-0.047	-0.051	-0.036
	M1-2	ϕ	-1.092	0.334	-0.142	0.044	-
		θ	-0.003	-	-	-	-
Henne	M1-1	ϕ	-1.626	0.861	-0.221	0.062	-0.029
		θ	-0.339	-0.145	-	-	-
	M1-2	ϕ	0.52	-0.949	-0.465	0.162	-
		θ	1.714	0.763	-0.01	-	-
Möhne	M1-1	ϕ	-2.058	1.716	-1.34	0.929	-0.244
		θ	-1.097	0.445	-0.594	0.188	0.085
	M1-2	ϕ	-1.363	0.041	0.327	-0.03	0.03
		θ	-0.477	-0.465	-	-	-
Sorpe	M1-1	ϕ	-0.884	-	-	-	-
		θ	0.314	-0.124	-0.073	-0.024	-
	M1-2	ϕ	0.073	-0.881	0.032	-	-
		θ	1.237	0.266	0.071	0.048	-

3.9.4 ARFIMA models

The procedure described in section 3.6.2 is used to estimate the parameters of the ARFIMA model. The procedure is repeated for different orders of AR and MA components ($0 \leq p \leq 5$) and ($0 \leq q \leq 5$).

The model with the minimum AIC is assumed as the best ARFIMA model. Table 3.5 gives the estimated parameters of the best ARFIMA models.

Table 3.5: *Parameters of the best autoregresssivge fractional integrated moving average ARFIMA(p,d,q) models*

Reservoir	Models group	Parameters of the model					
Bigge	M1-1	ϕ	-0.663	-	-	-	-
		θ	0.286	-	-	-	-
		d	0.173				
	M1-2	ϕ	-0.777	0.018	-	-	-
		θ	0.184	-0.066	-	-	-
		d	0.126				
Henne	M1-1	ϕ	-0.6479	-	-	-	-
		θ	0.3842	-	-	-	-
		d	0.2576				
	M1-2	ϕ	-0.648	-	-	-	-
		θ	0.3842	-	-	-	-
		d	0.1896				
Möhne	M1-1	ϕ	-0.685	0.129	-0.072	-	-
		θ	-	-	-	-	-
		d	0.278				
	M1-2	ϕ	-0.5	-	-	-	-
		θ	-	-	-	-	-
		d	0.3948				
Sorpe	M1-1	ϕ	-0.7584	-	-	-	-
		θ	0.2228	-0.1535	-0.0471	-	-
		d	0.2141				
	M1-2	ϕ	-0.774	-	-	-	-
		θ	0.166	-	-	-	-
		d	0.2238				

3.9.5 Results of diagnostic checks for ARMA and ARFIMA models

We tested the diagnostic of the ARMA(p,q) and ARFIMA(p, d,q) models using the the Ljung-Box test and the ACF of the residuals. The results of the Ljung-Box test at 5% significance level are given in table 3.6. The results show that except for the ARFIMA models for the daily inflow into the Henne and Möhne reservoirs, all p-values are greater than 0.05, then the null hypothesis of models adequacy cannot be rejected at this significance level.

Figures 3.6 and 3.7 plot the ACF of the residuals obtained from the ARMA and ARFIMA models (models, group M1-1) respectively for daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs. The plots of the ACF for the ARMA and ARFIMA models (models group M1-2), are shown in figure 3.8 and 3.9 respectively. The estimated n_{out}/m values are displayed on the plots of ACF of the residuals. The graphs and the estimated n_{out}/m values show that on the basis of the autocorrelation function there is no cause to reject the fitted models in the following cases:

1. The ARMA(p,q) and ARFIMA(p,d,q) models for the daily inflow into the Bigge and Sorpe reservoirs (models, group M1-1 and models, group M1-2).

2. The ARMA(p,q) model for the daily inflow into the Henne reservoir (models, group M1-1).

3.10 Forecasting performance assessment

In the present section we investigate the forecasting ability of the BPNN, ANFIS, ARMA and ARFIMA models by assessing their forecasting performances using the models efficiency criteria discussed in section 3.8. The forecasting performances of the different models are compared and the results may be summarized as follows:

3.10.1 Univariate models (group M1-1 and group M1-2)

The efficiency criteria for the models, group M1-1 and group M1-2 (one-day-ahead and two-days-ahead forecasting) are given in tables 3.7 and 3.8 respectively. The results show that in terms of all efficiency criteria except the *AREP*, the BPNN, ANFIS, ARMA and ARFIMA models have similar performances but the first two are slightly better. The BPNN and ANFIS models have the minimum *AREP* values in forecasting one-day-ahead of all daily inflow time series (except for this of the Sorpe reservoir, group M1-2). According to the values of the *AREP*, the ANFIS models outperform BPNN models in forecasting one-day-ahead of all daily inflow time series except:

- The daily inflow time series of the Bigge and Henne reservoirs (the models, group M1-1).

- The daily inflow time series of the Henne reservoir (the models, group M1-2).

Plots of observed daily inflow and one-day-ahead and two-days-ahead forecast hydrographs for period from 1-11-1999 to 31-10-2000 for the models, group M1-1 are shown in figures 3.10 and 3.11 respectively. To save space, only the plots for the Bigge reservoir are displayed here.

3.10.2 Multivariate models (group M2)

Table 3.9 lists the efficiency criteria for the models, group M2. The results show that the BPNN and ANFIS models don't have significant difference in the performances except the *AREP*. The values of the *AREP* show that the ANFIS models outperform the BPNN models in forecasting of:

1. One-day-ahead daily inflow into the Henne and Sorpe reservoirs.
2. Two-days-ahead daily inflow into the Bigge and Sorpe reservoirs.

3.10.3 Univariate vs. multivariate models

The performances of the models, group M1-2 are compared with those of the models, group M2. The results of the comparison show a clear superiority of the models, group M2 for one-day-ahead forecasting. For two-days-ahead forecasting, the performances of the models, group M2 are slightly better. The following procedure is used to compare between the performances of the BPNN model, (the models group M1-2) and the BPNN model (the models, group M2):

1. Sort the observed and the one-day-ahead forecasted daily inflow data in ascending order according to observed one.
2. Divide the data into sets with equal size.
3. Estimate the performance for each daily inflow group.

Figures 3.12, 3.13, 3.14 and 3.15 show the values of the efficiency criteria vs. the forecasted daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs respectively. It is clear that the models, group M2 outperform the models, group M1-2 in forecasting daily inflow with values more than average daily inflow (the dotted vertical line) for all efficiency criteria (especially the *rmse* and *AREP*). These figures can be used to guess the expected efficiency criteria for the forecasted daily inflow.

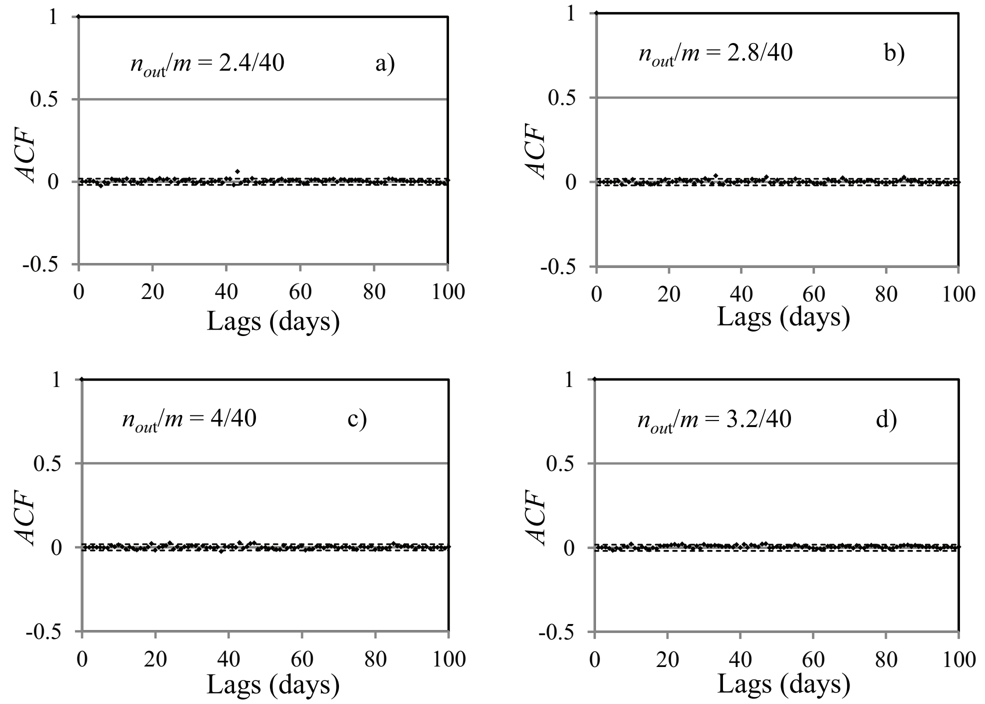


Figure 3.6: *The ACF of the residuals from the ARMA models (models, group M1-1) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs*

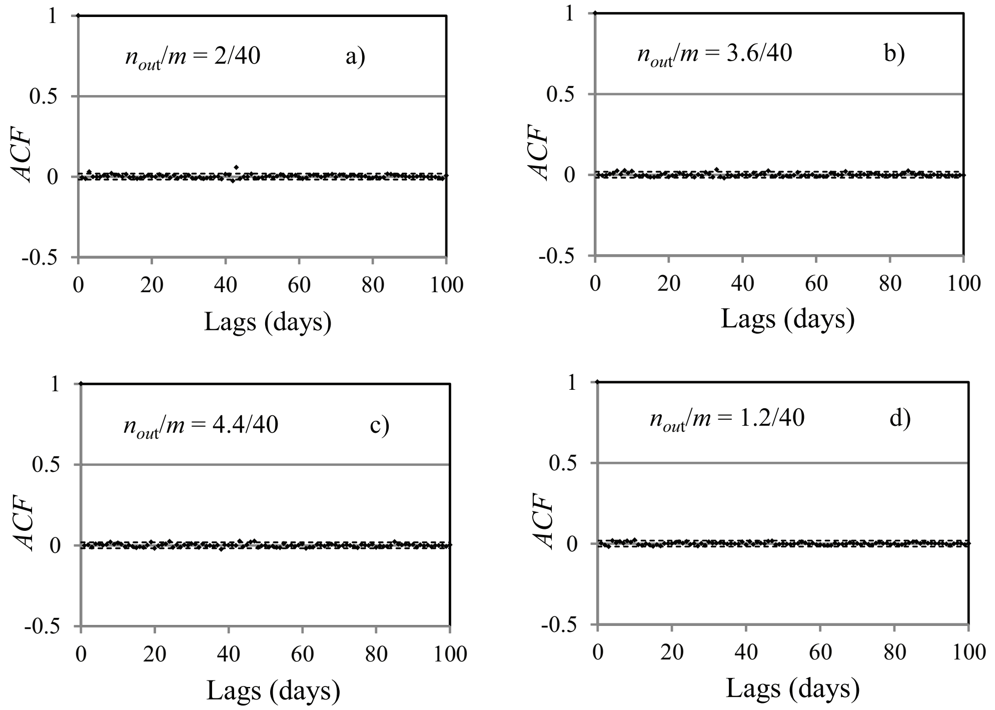


Figure 3.7: *The ACF of the residuals from the ARFIMA models (models, group M1-1) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs*

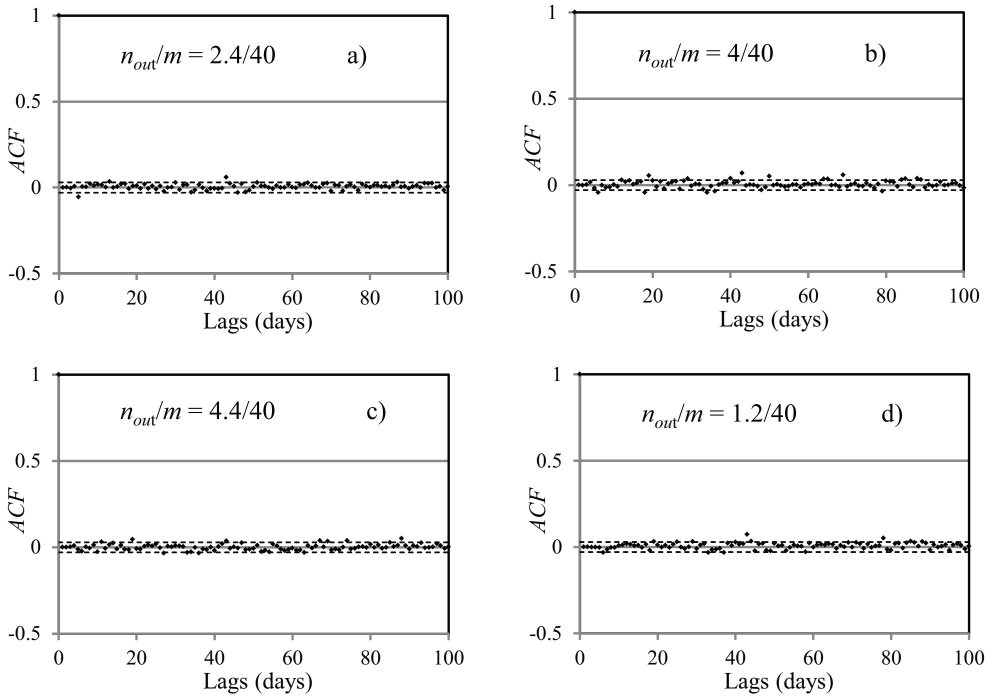


Figure 3.8: *The ACF of the residuals from the ARMA models (models, group M1-2) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs*

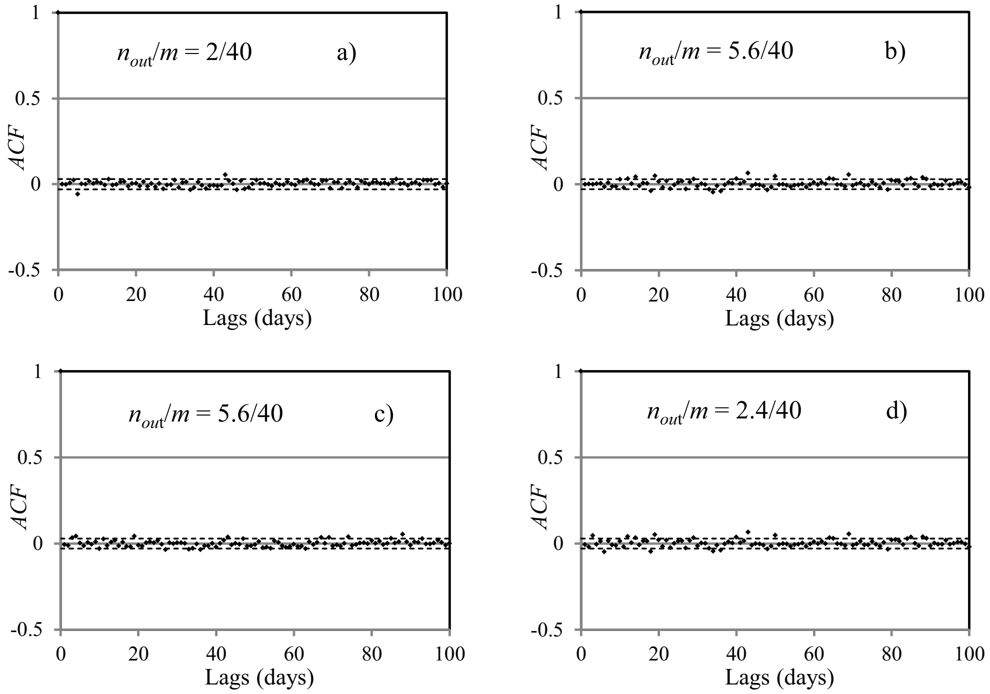


Figure 3.9: *The ACF of the residuals from the ARFIMA models (models, group M1-2) for daily inflow into the a) Bigge, b) Henne, c) Möhne and d) Sorpe reservoirs*

Table 3.6: *Results of the Ljung-Box Q-test*

Reservoirs	Models group M1-1				Models group M1-2			
	ARMA		ARFIMA		ARMA		ARFIMA	
	p-value	Q statistic*	p-value	Q statistic*	p-value	Q statistic*	p-value	Q statistic*
Bigge	0.1665	31.71	0.2374	29.66	0.1219	33.38	0.1739	31.47
Henne	0.5021	24.3	0.0235	40.9	0.0718	35.99	0.004	47.67
Möhne	0.0624	36.64	0.0078	45.28	0.1489	32.32	0.0158	42.5
Sorpe	0.0743	35.83	0.5105	24.154	0.9441	14.88	0.9427	14.94

* critical value of Q statistic is 37.70.

Table 3.7: *Values of the efficiency criterion parameters for the models, group M1-1*

Lead time	Reservoir	Model	R	$rmse$	$AREP$	g	NSC	NSC_{rel}
one-day-ahead	Bigge	BPNN	0.878	5.486	22.873	0.942	0.881	0.932
		ANFIS	0.868	5.672	24.975	0.938	0.883	0.924
		ARMA	0.863	5.832	30.694	0.936	0.887	0.929
		ARFIMA	0.83	6.515	47.028	0.92	0.812	0.836
	Henne	BPNN	0.935	0.984	15.249	0.968	0.935	0.969
		ANFIS	0.928	1.031	17.827	0.965	0.951	0.975
		ARMA	0.929	1.025	25.583	0.966	0.929	0.949
		ARFIMA*	0.916	1.113	30.47	0.96	0.914	0.934
	Möhne	BPNN	0.886	3.071	22.662	0.946	0.883	0.916
		ANFIS	0.887	3.047	19.497	0.947	0.899	0.939
		ARMA	0.897	2.93	22.61	0.951	0.888	0.93
		ARFIMA*	0.896	2.93	22.153	0.951	0.89	0.932
	Sorpe	BPNN	0.936	0.605	22.204	0.969	0.91	0.942
		ANFIS	0.933	0.622	17.765	0.967	0.924	0.956
		ARMA	0.931	0.625	25.784	0.967	0.913	0.941
		ARFIMA	0.906	0.727	25.609	0.955	0.914	0.946
two-days-ahead	Bigge	BPNN	0.703	8.126	46.773	0.874	0.757	0.799
		ANFIS	0.702	8.144	46.809	0.873	0.746	0.799
		ARMA	0.682	8.476	56.149	0.865	0.734	0.765
		ARFIMA	0.669	8.571	69.618	0.862	0.659	0.626
	Henne	BPNN	0.805	1.643	42.068	0.912	0.838	0.86
		ANFIS	0.803	1.65	35.51	0.911	0.863	0.911
		ARMA	0.788	1.704	51.522	0.905	0.804	0.803
		ARFIMA*	0.775	1.748	54.49	0.9	0.79	0.789
	Möhne	BPNN	0.756	4.336	32.821	0.893	0.785	0.847
		ANFIS	0.756	4.336	32.304	0.893	0.791	0.865
		ARMA	0.76	4.3	37.293	0.894	0.768	0.827
		ARFIMA*	0.759	4.305	36.386	0.894	0.773	0.837
	Sorpe	BPNN	0.823	0.977	39.308	0.919	0.808	0.844
		ANFIS	0.821	0.983	35.377	0.918	0.822	0.881
		ARMA	0.806	1.016	49.404	0.912	0.783	0.797
		ARFIMA	0.792	1.056	41.237	0.905	0.818	0.863

* the fitted model is rejected in the diagnostic checking step.

Table 3.8: *Values of the efficiency criterion parameters for the models, group M1-2*

Lead time	Reservoir	Model	R	$rmse$	$AREP$	g	NSC	NSC_{rel}
one-day-ahead	Bigge	BPNN	0.901	4.197	32.434	0.953	0.86	0.895
		ANFIS	0.897	4.257	27.328	0.952	0.877	0.929
		ARMA	0.887	4.447	33.549	0.947	0.855	0.905
		ARFIMA	0.846	5.142	39.581	0.929	0.827	0.874
	Henne	BPNN	0.946	0.966	24.319	0.974	0.949	0.94
		ANFIS	0.939	1.02	27.435	0.971	0.943	0.931
		ARMA	0.946	0.993	24.083	0.974	0.951	0.934
		ARFIMA*	0.931	1.126	29.983	0.967	0.936	0.903
	Möhne	BPNN	0.894	2.532	21.536	0.95	0.887	0.926
		ANFIS	0.892	2.551	19.385	0.949	0.892	0.936
		ARMA	0.896	2.502	22.35	0.951	0.88	0.918
		ARFIMA*	0.897	2.5	22.014	0.951	0.881	0.921
	Sorpe	BPNN	0.947	0.619	18.775	0.975	0.95	0.961
		ANFIS	0.937	0.682	14.904	0.969	0.953	0.969
		ARMA	0.954	0.593	14.78	0.978	0.958	0.97
		ARFIMA	0.944	0.655	17.033	0.973	0.951	0.962
two-days-ahead	Bigge	BPNN	0.664	7.318	50.577	0.857	0.713	0.799
		ANFIS	0.659	7.419	47.814	0.853	0.696	0.793
		ARMA	0.655	7.399	62.822	0.854	0.658	0.659
		ARFIMA	0.634	7.557	63.652	0.847	0.649	0.657
	Henne	BPNN	0.83	1.671	34.759	0.923	0.903	0.889
		ANFIS	0.817	1.73	44.732	0.917	0.884	0.849
		ARMA	0.836	1.695	44.983	0.925	0.878	0.757
		ARFIMA*	0.825	1.743	50.41	0.921	0.863	0.712
	Möhne	BPNN	0.749	3.789	37.785	0.888	0.749	0.813
		ANFIS	0.755	3.722	32.505	0.892	0.771	0.855
		ARMA	0.768	3.628	37.333	0.897	0.753	0.773
		ARFIMA*	0.763	3.653	35.903	0.896	0.758	0.802
	Sorpe	BPNN	0.839	1.055	23.561	0.926	0.9	0.927
		ANFIS	0.83	1.09	29.452	0.921	0.877	0.898
		ARMA	0.867	0.992	24.979	0.938	0.903	0.92
		ARFIMA	0.859	1.019	27.901	0.935	0.892	0.901

* the fitted model is rejected in the diagnostic checking step.

Table 3.9: *Values of the efficiency criterion parameters for the models, group M2*

Lead time	Reservoir	Model	R	$rmse$	$AREP$	g	NSC	NSC_{rel}
one-day-ahead	Bigge	BPNN	0.944	3.253	18.787	0.972	0.907	0.936
		ANFIS	0.928	3.633	22.944	0.965	0.883	0.623
	Henne	BPNN	0.966	0.767	20.466	0.984	0.94	0.936
		ANFIS	0.951	1.009	19.505	0.972	0.962	0.929
	Möhne	BPNN	0.939	1.948	13.602	0.97	0.938	0.95
		ANFIS	0.928	2.112	14.03	0.965	0.933	0.93
	Sorpe	BPNN	0.973	0.439	13.998	0.987	0.965	0.97
		ANFIS	0.968	0.485	11.428	0.984	0.97	0.978
two-days-ahead	Bigge	BPNN	0.707	6.908	49.258	0.872	0.665	0.74
		ANFIS	0.694	7.071	43.775	0.866	0.736	0.709
	Henne	BPNN	0.838	1.631	34.259	0.926	0.881	0.827
		ANFIS	0.834	1.653	38.798	0.924	0.895	0.836
	Möhne	BPNN	0.786	3.499	28.849	0.904	0.79	0.851
		ANFIS	0.78	3.573	29.714	0.9	0.797	0.833
	Sorpe	BPNN	0.851	1.021	32.118	0.931	0.85	0.766
		ANFIS	0.844	1.04	24.475	0.928	0.891	0.91

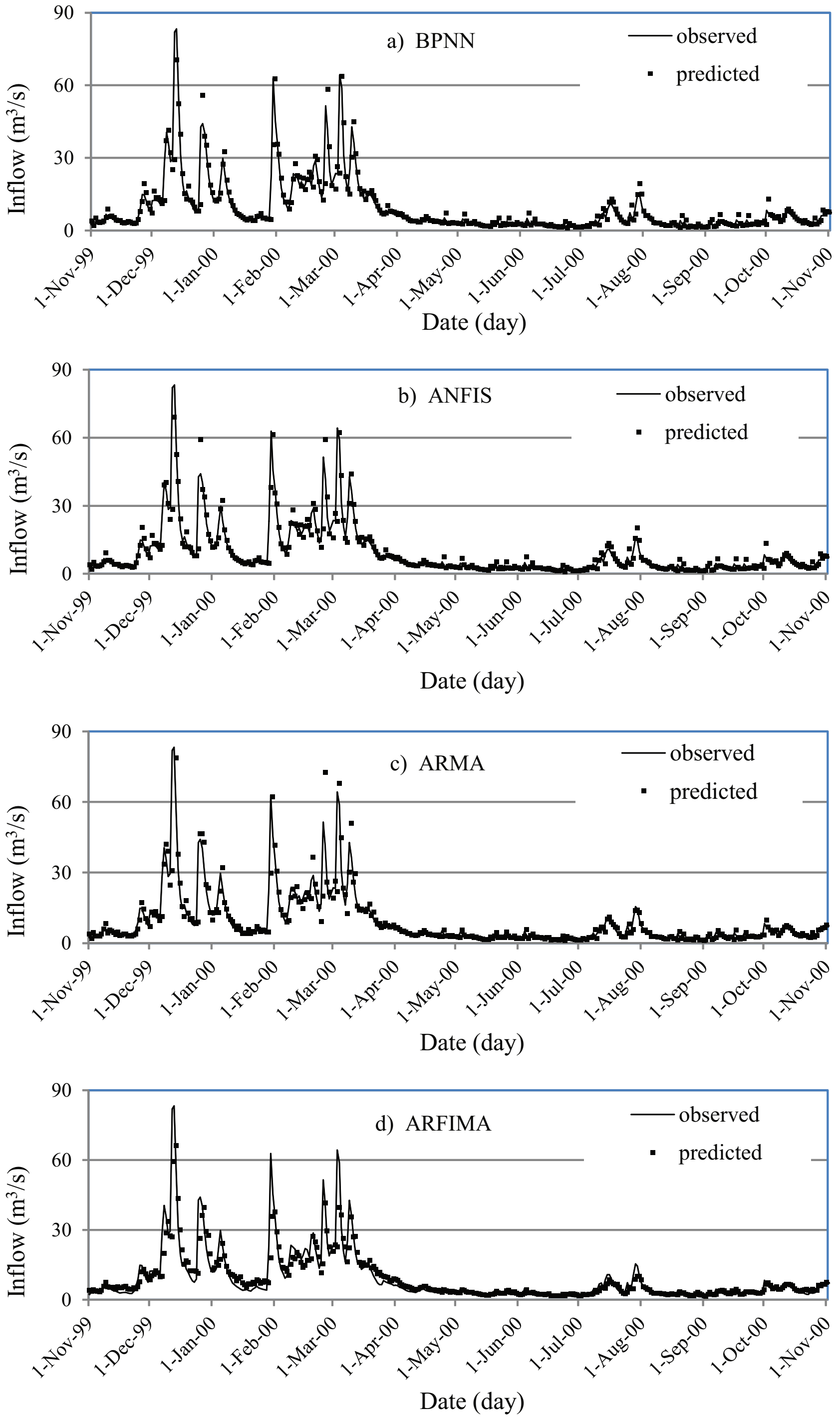


Figure 3.10: *Comparative one-day-ahead forecasting of the daily inflow into Bigge reservoir (models, group M1-1)*

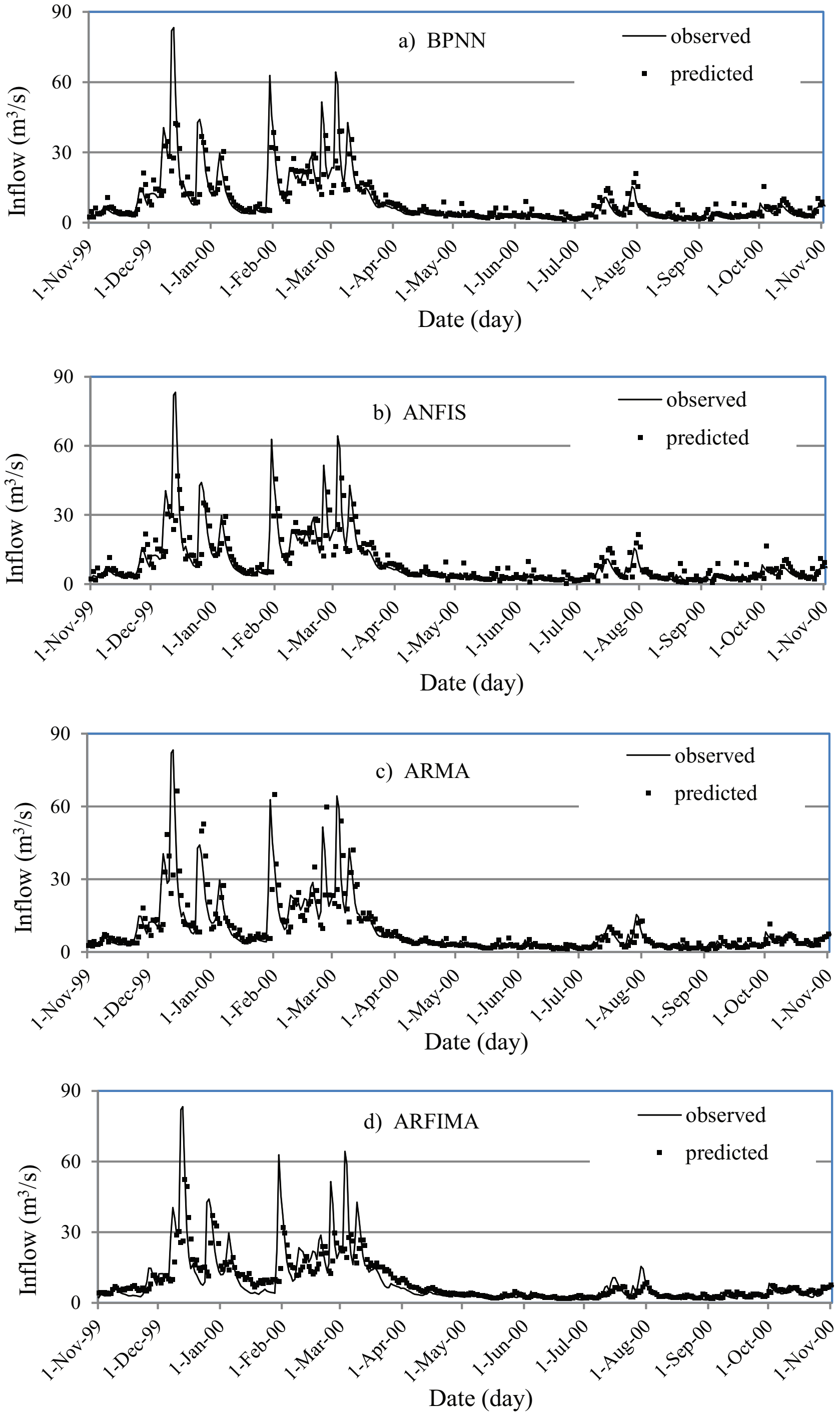


Figure 3.11: *Comparative two-days-ahead forecasting of the daily inflow into the Bigge reservoir (models, group M1-1)*

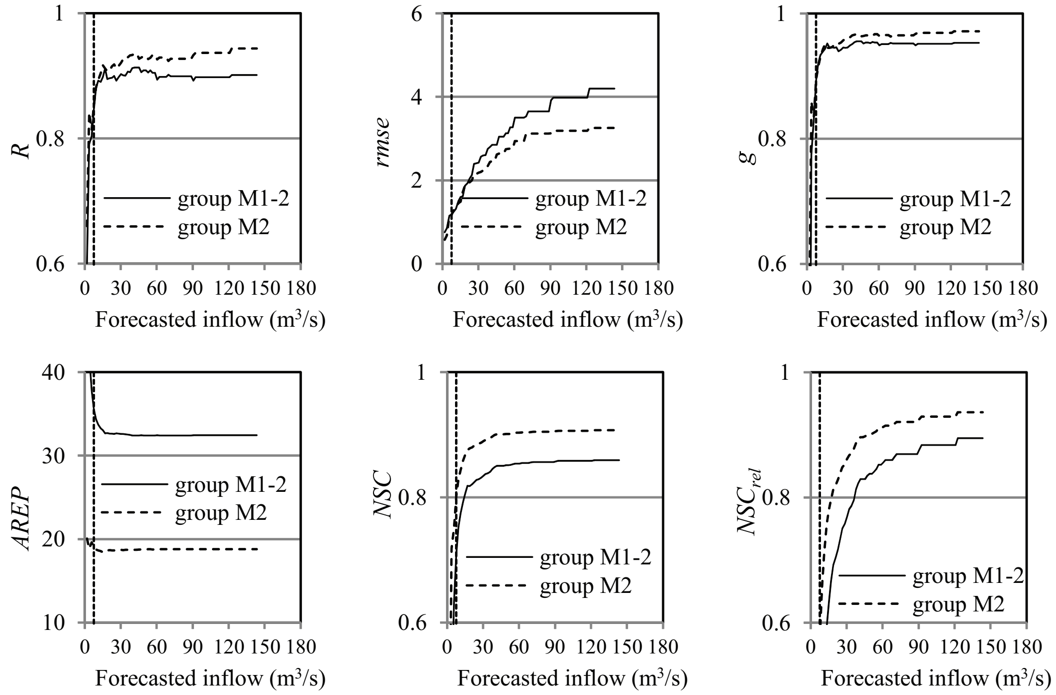


Figure 3.12: Values of the efficiency criteria vs. the predicted daily inflow (Bigge reservoir)

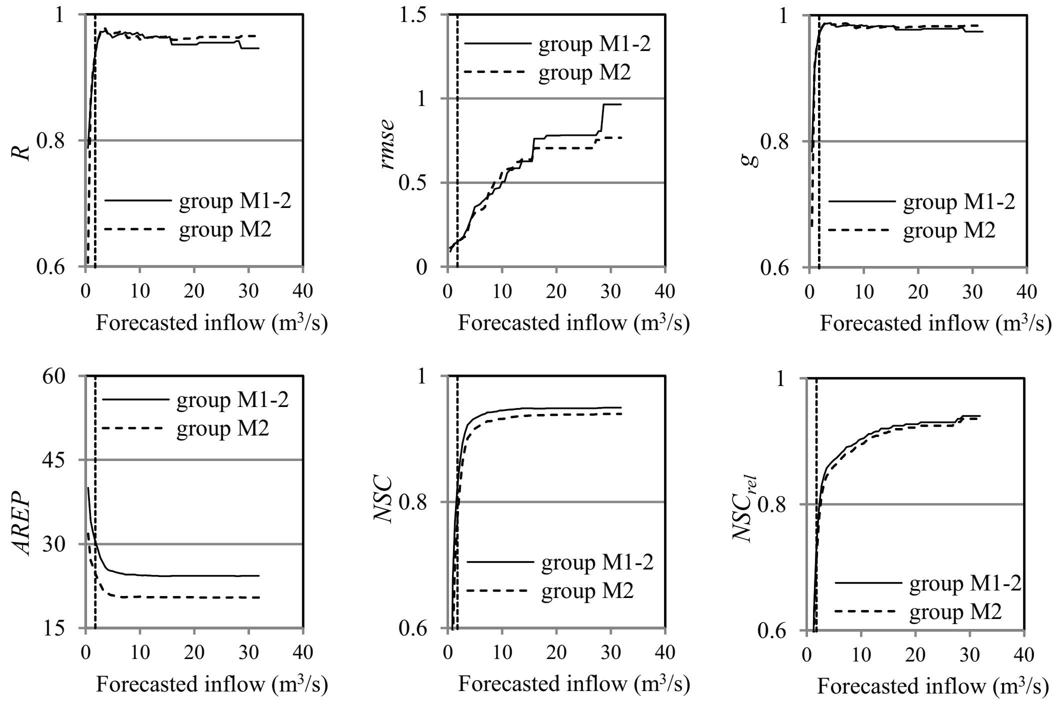


Figure 3.13: Values of the efficiency criteria vs. the predicted daily inflow (Henne reservoir)

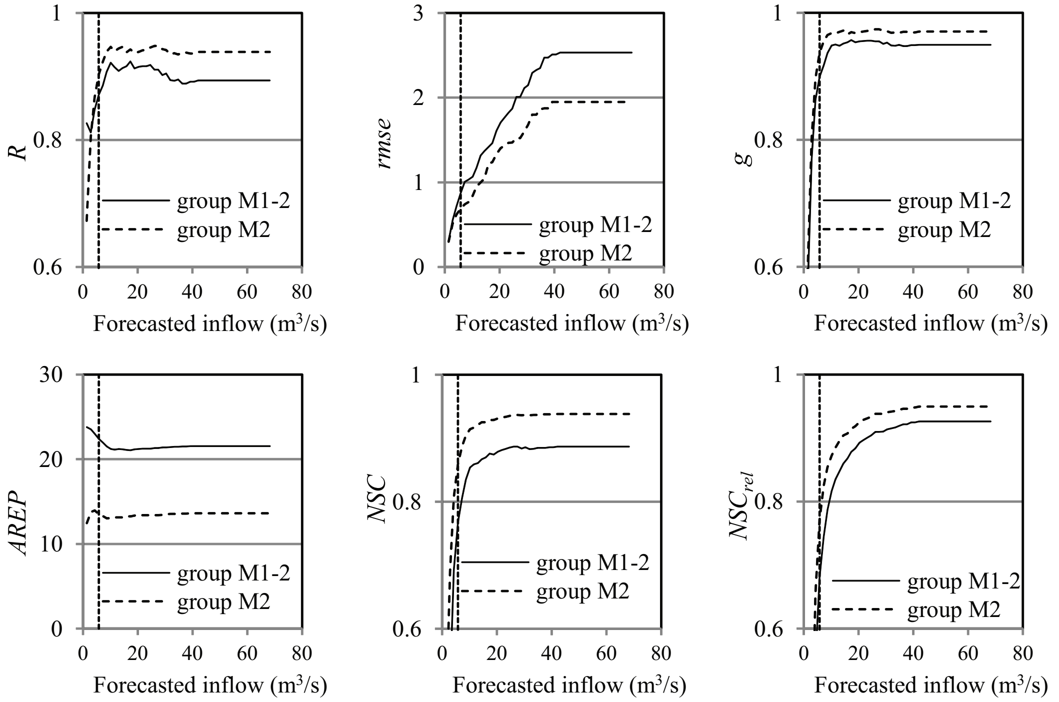


Figure 3.14: *Values of the efficiency criteria vs. the predicted daily inflow (Möhne reservoir)*

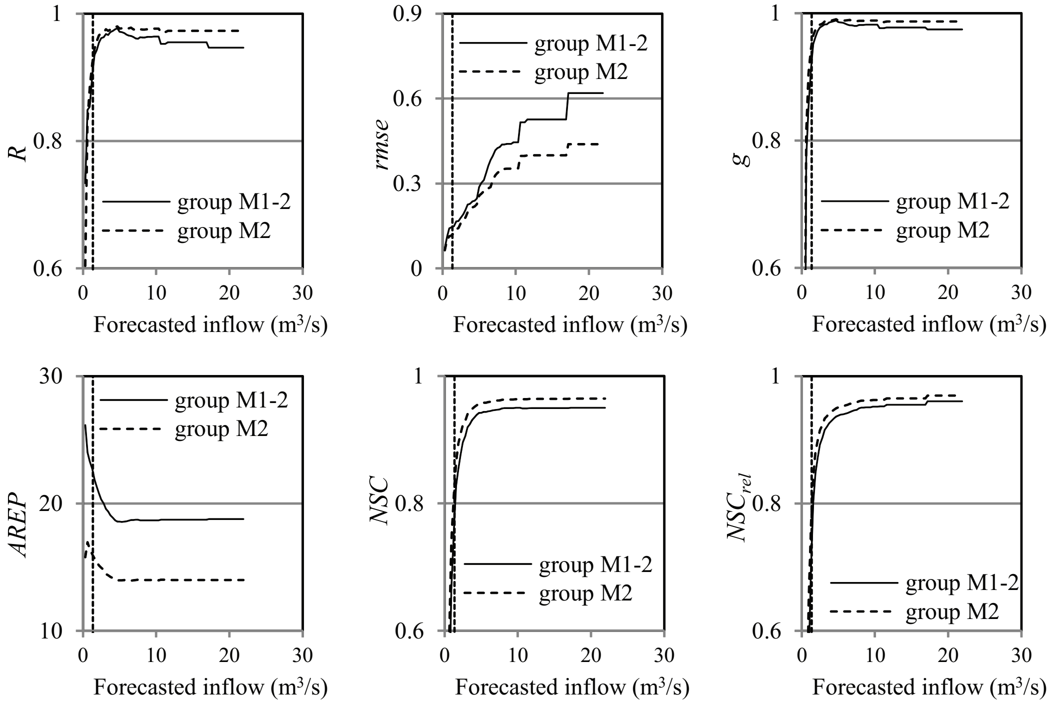


Figure 3.15: *Values of the efficiency criteria vs. the predicted daily inflow (Sorpe reservoir)*

3.11 Conclusions

The BPNN, ANFIS, ARMA and ARFIMA models are used to forecast daily inflow into the Bigge, Henne, Möhne and Sorpe reservoirs. The models are divided into univariate models (the models group M1-1 and group M1-2) and multivariate models (the models group M2) according to the potential input variables. The BPNN, ANFIS, ARMA and ARFIMA are the simulation models in the models, group M1-1 and group M1-2 and the average daily inflow are the potential input variables, however in the models, group M2 the BPNN and ANFIS are the simulation models and the average daily inflow and rainfall are the potential input variables.

Early stopping procedure is applied to prevent overfitting in the BPNN and ANFIS models. One hidden layer is assumed to be sufficient to simulate the training data using the BPNN models. A trial-and-error procedure is employed to determine the number of neurons in the hidden layer and to select the input variables that give the best performance. Starting with the input variables of the optimum BPNN models, different ANFIS models are trained to find out the best one.

Different ARMA and ARFIMA models are tried with different orders ($0 \leq p \leq 5$ and $0 \leq q \leq 5$) of the AR and MA components for each data series and the AIC values of the models are compared to select the best models. The ACF of the residuals and the Ljung-Box test are used to test the diagnostic of the ARMA(p,q) and ARFIMA(p,d,q) models. The results of the Ljung-Box test at 5 % significance level show that the null hypothesis of model adequacy cannot be rejected for all simulated daily inflow time series except that of the Henne and Möhne reservoirs which are simulated using the ARFIMA models.

The performances of the models group M1-1 and group M1-2 show that these models don't have significant difference in the performances except for the average relative error percentage ($AREP$). The BPNN and ANFIS models have the minimum values of the $AREP$ for all daily inflow time series.

The models, group M2, outperform the models, group M1-2 in respect of all used efficiency criteria.

3.12 References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

Ahmed, J.A., Sarma A.K., 2007. Artificial neural network model for synthetic streamflow generation. Water Resources Managent, 21(16), 1015–1029.

Akaike, H., 1979. A Bayesian extension of the minimum *AIC* procedure of autoregressive model fitting. Biometrika, 66(2), 237–242.

Bartlett P. L., Maass, W., 2003. Vapnik–Chervonenkis dimension of neural nets. In the handbook of brain theory and neural networks, M. A. Arbib, editor, pages 1188–1192. MIT Press (Cambridge), 2nd edition.

Berry, M.J., Linoff, G., 2004. Data mining techniques: For marketing, sales and customer support. Wiley & Sons, New York.

Box, G.E.P., Jenkins, G.M., 1976. Time series analysis: forecasting and control. San Francisco: Holden–Day.

Box, G.E.P., Pierce, D.A., 1970. Distribution of residual autocorrelations in autoregressive integrated moving average time series. Journal of American Statistical Association, 65, 1509–1526.

Brockwell, P.J., Davis, R.A., 2002. Introduction to time series and forecasting. 2nd edition, Springer, New York.

Brockwell, P.J., Davis, R.A., 1991. Time series: Theory and methods. 2nd edition, Springer, New York.

Chatfield, C., 2004. The analysis of time series: An introduction. Chapman & Hall/CRC.

Coli, M., Fontanella, L., Granturco, M., 2005. Parametric estimation for ARFIMA models via spectral methods. Statistical Methods & Applications, 14, 11–27.

Coulibaly, P., Anctil, F., Bobeé, B., 2001. Multivariate reservoir inflow forecasting using temporal neural networks. Journal of Hydrologic Engineering, 6(5), 367–376.

Coulibalya, P., Anctilb, F., Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. Journal of Hydrology 230, 244–257.

Dawson, C.W., Wilby, R.L., 2001. Hydrological modeling using artificial neural network. Progress in Physical Geography, l(2), 80–108.

Detienne, K.B., Detienne, D.H., Joshi, S.A., 2003. Neural networks as statistical tools for business researchers. Organizational Research Methods, 6(2), 236–265.

El–Shafie A., Tahal M.R., Noureldin A., 2007. A neuro–fuzzy model for inflow forecasting of the Nile River at Aswan high dam. Water Resources Management, 21(3), 533–556.

Hamilton, J.D., 1994. Time series analysis. Princeton University Press.

Haslett, J., Raftery, A.E., 1989. Space–time modeling with long–memory dependence: Assessing Ireland's wind power resource. Applied Statistics, 38, 1–50.

Hellmann, M., 2001. Fuzzy logic introduction. Epsilon Nought Radar Remote Sensing Tutorials.

Jain, S.K., Das, A., Srivastava, D.K., 1999. Application of ANN for reservoir inflow prediction and operation. Journal of Water Resources Planning and Management, 125(5), 263–271.

Jang, J.–S.R., 1993. ANFIS: Adaptive–Network–based fuzzy inference systems. IEEE Transactions on Systems, Man and Cybernetics, 23(3), 665–685.

Khadangi, E., Madvar, H. R., Ebadzadeh, M.M., 2009. Comparison of ANFIS and RBF models in daily stream flow forecasting. Computer, Control and Communication, IC4 2009, 2nd International Conference, 17(18), 1–6.

Krause, P., Boyle, D., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5, 89–97.

Larsen, J., 1994. Optimizing neural network architectures using generalization error estimators. Radiophysics and Quantum Electronics, 37(9), 729–740.

Lawrence, S., Giles, C., Tsoi., A., 1996. What size neural network gives optimal generalization? convergence properties of backpropagation. University of Maryland Technical Report, UMIACS–TR–96–22.

Levenberg, K., 1944. A method for the solution of certain nonlinear problems in least squares. The Quarterly of Applied Mathematics, 2, 164–168.

Li, W., 2004. Diagnostic checks in time series. Chapman & Hall/CRC.

Li, X.–X., Huang, H., Liu, C.–H. 2007. The Application of an ANFIS and BP neural network method in vehicle shift decision. 12^{th} IFToMM World Congress, Besançon (France).

Ljung, G.M., Box, G.E.P., 1978. On a measure of a lack of fit in time series models. Biometrika, 65, 297–303.

Loukas, Y.L., 2001. Adaptive neuro–fuzzy inference system: an instant and architecture–free predictor for improved QSAR studies. Journal of medicinal chemistry, 44(17), 2772–2783.

Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man–Machine Studies, 7(1), 1–13.

Marquardt, D., 1963. An algorithm for least–squares estimation of nonlinear parameters. SIAM Journal on Applied Mathematics, 11, 431–441.

Mathimatica, 2007. Time series: Part 1. User's guide to time series. Wolfram Research, Inc.

Matlab.a, 2008. Neural network toolbox 6, User's guide. The MathWorks, Inc.

Moatmari, A., Longoni, M., Rosso, R., 1999. A seasonal long–memory stochastic model for the simulation of daily river flows. Physics and Chemistry of the Earth (B), 24(4), 319–324.

Moody, J.E., 1992. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In: John E. MOODY, Steve J. HANSON and Richard P. LIPPMANN, eds. Advances in Neural Information Processing Systems 4. San Mateo, CA: Morgan Kaufmann, 847–854.

Murata, N., Yoshizawa, S., Amari, S., 1994. Network information criterion–determining the number of hidden units for an artificial neural network model. IEEE Transactions on Neural Networks, 5(6), 865–872.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, part I–A discussion of principles. Journal of Hydrology, 10(3), 282–290.

Palma, W., 2007. Long–memory time series: Theory and methods. Wiley–Interscience, New Jersey.

Ranga Suri, N.N.R., P. Nagabhushan, P., 2002. Parallel Levenberg–Marquardt–based neural network training on Linux clusters: A case study. ICVGIP, 3rd Indian Conference on Computer Vision, Graphics and Image Processing.

Reisen, V., Abraham, B., Lopes, S., 2001. Estimation of parameters in ARFIMA processes: A simulation study. Communications in Statistics, Simulation and Computation, 30(4), 787 – 803.

Rojas, R., 1996. Neural networks: A systematic introduction. Springer, New York.

Samarasinghe, S., 2006. Neural networks for applied sciences and engineering. From fundamentals to complex pattern recognition. Auerbach Publications. USA.

Shumway, R.H., Stoffer, D.S., 2006. Time series analysis and its applications: With R examples. 2nd edition, Springer, New York.

Smith, J., Eli, R., N., 1995. Neural–network models of rainfall–runoff process. Journal of Water Resources Planning and Management, 121(6), 499–508.

Sowell, F., 1992. Maximum likelihood estimation of stationary univariate fractionally integrated time series models. Journal of Econometrics, 53, 165–188.

Takagi T, Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man and Cybernetics, 15(1), 116–132.

Wang, L., et al., 2005. Optimal size of a feedforward neural network: How much does it matter?. Autonomic and Autonomous Systems and International Conference on Networking and Services.

Wang, L., Langari, R., 1995. Building Sugeno–type models using fuzzy discretization and orthogonal parameter estimation techniques. IEEE Transactions on Fuzzy Systems, 3(4), 454–458.

Wang, L., Libert, G.A., 1994. Combining pattern recognition techniques with Akaike's information criteria for identifying ARMA models. IEEE Transactions on Signal Processing, 42(6), 1388–1396.

Wang, W., Chen, X., Ma, J., HUO, S., 2008. Comparing univariate ARMA and ARFIMA model for forecasting daily streamflows. Hydrological Research in China: Process Studies, Modeling Approaches and Applications, IAHS Publications, 322, 213–219.

Willmot, C.J., 1981. On the validation of models. Physical Geography, 2, 184–194.

Xu, Z.X., Li, J.Y., 2002. Short–term inflow forecasting using an artificial neural network model. Hydrological Processes, 16(12), 2423–2439.

Zadeh, L.A., 1965. Fuzzy sets. Information and Control, 8(3), 338–353.

Zhang, G.P., Qi, M., 2005. Neural network forecasting for seasnal and trend time series. European Journal of Operational Research, 160, 501–514.

Chapter 4

Filling in missing data

4.1 Introduction

We proposed three models to fill in the missing data in the daily inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs. These models are the backpropagation neural networks (BPNN), the adaptive neuro-fuzzy inference system (ANFIS) and the generalized linear model (GLM). The performances of the models are compared and the model with the best performance is also applied to extend the daily inflow into the Bigge reservoir in the period from 1/11/1960 to 31/10/1965.

Machine learning techniques such as artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) models have been used to solve different hydrology and water resources problems. Some of these applications were presented in the previous chapter. Only a limited number of reports and researches related to the use of ANN and ANFIS in filling in missing data are available. An ANN- based model for estimating missing values in a multivariate data set was reported by Gupta and Lam (1996). He found that the performances of the ANN-based models are better than those obtained by iterative regression analysis. Khalil et al. (2001) investigated the concepts of ANN and seasonal groups and their characteristics for the estimation of missing data values in monthly streamflow.

Dastorani and Wright (2003) completed a research project on flow estimation for ungauged catchments using neural networks. Dastorani and Wright (2004) employed ANN to optimize the results of a hydrodynamic approach for river flow prediction. Based on correlation factors, Petersen et al. (2008) used rainfall time series and torrent flows to calculate discharges at Mongalla, Sudan. The results provided a near reality flow time series extending the available Mongalla data for the period from 1984 to 1996.

Dastorani et al. (2009) investigated the capabilities of ANN and ANFIS to fill in the gaps of hydrological data series measured in some stations in Iran. The ANFIS model showed superiority in the accuracy of estimation the missing data. The results of the ANN models also showed a good level of accuracy.

No reports or researches related to the use of GLM for filling in missing data are available. GLM are widely used for rainfall simulation (e.g., Chandler and Wheater, 2002; Yang et al., 2005; Little et al., 2009).

4.2 Data used

The average daily inflow data of the Henne, Möhne and Sorpe reservoirs in the period from 1-11-1960 to 1-10-2006 and of the Bigge reservoir in the period from 1-11-1965 to 1-10-2006 are used to train the models. The available inflow data are divided into two sets. The first set is used to train the models and the second to validate them.

4.3 Selection of the inputs

We denoted the average daily inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs as *QB*, *QH, QM* and *QS* respectively. The cross correlation between the inflow time series (*QB*, *QH, QM* and *QS*) are shown in figure 4.1. The figure shows a high correlation between the inflow time series of all reservoirs (ranges from 0.865 to 0.964). High cross correlations between the inflow time series of all reservoirs indicate that they are related to each other. Due to that the missing data in an inflow time series is estimated using the other inflow time series as input variables. For example, to estimate the missing data in *QB*, we used *QH, QM* and *QS* as inputs.

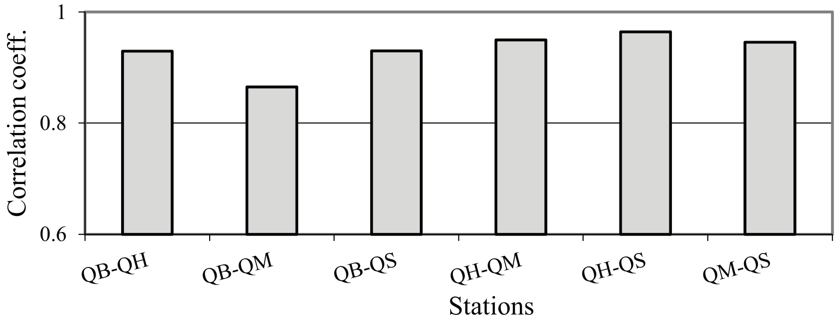


Figure 4.1: *Cross correlation between the inflow data of each two stations*

4.4 Missing data estimation models

4.4.1 BPNN and ANFIS models

Both backpropagation neural networks (BPNN) and adaptive neuro-fuzzy inference system (ANFIS) models are discussed in detail in the previous chapter. BPNN models with one hidden layer of three neurons and ANFIS models with three Gaussian membership functions associated with each input variable are used in the present study.

4.4.2 Generalized linear model (GLM)

The generalized linear models were defined by Nelder and Wedderburn (1972) and Wedderburn (1974) as an extension of the traditional linear regression model to data with non-normal responses. The monograph by McCullagh and Nelder (1989) was the first monograph on this topic.

Assume the observations $y_i, i = 1, 2, \ldots, N,$ the traditional linear regression model is of the form

$$y_i = \beta X_i + \varepsilon_i, \quad i = 1, 2, \cdots, N \tag{4.1}$$

where y_i is the response variable for X_i, X (N by k) is the model matrix,β (k by 1) is a vector of coefficients and the residuals ε (N by 1) are *i.i.d.* N(0,σ^2). The vector of coefficient β is estimated by least squares fit to the data. The response y_i has a normal distribution with mean μ.

In GLM, at each set of values for the predictors, the response y_i has a distribution that may be normal, binomial, Poisson, gamma, or inverse Gaussian, with parameters including a mean μ. There are three components that are common to all GLM (Agresti 2002):

The random component which refers to the probability distribution of the response y_i is assumed to be a member of the exponential family of distributions.

The systematic component is a linear predictor similar to that in the linear models,

$$\eta_i = \beta X_i \tag{4.2}$$

The third component of GLM is a *monotonic link function* $g(\cdot)$ that connects the random and systematic components. Let $\mu_i = E(y_i), \;\; i = 1,\, 2,\, \cdots, N$,the model links $\mu_i$to η_i by $g(\mu_i) = \eta_i$. Thus, the mean can be expressed as the inversely linked linear predictor,

$$\mu_i = g^{-1}(\eta_i) \tag{4.3}$$

The commonly used link functions $g(\mu)$ are given in table 4.1. If $g(\mu) = \theta$ then g is called the canonical link corresponding to $a(\theta)$.

Table 4.1: *Commonly used link functions (De Jong and Heller, 2008).*

Link function	$g(\mu)$	Canonical link for
identity	μ	normal
log	$\ln\, \mu$	Poission
power	μ^p	Gamma (p=-1)
		inverse Gaussian (p=-2)
logit	$\ln \frac{\mu}{1-\mu}$	binomial

The choice of the link function is very important to find an appropriate generalized linear model (GLM).

The optimal distributions of the responses and the corresponding link functions are detected using a trial-and-error procedure. Table 4.2 lists the type of the distribution of the response y_i for each daily inflow time series and the corresponding link function.

Table 4.2: *Types of the distribution of the responses and the corresponding link functions*

Reservoir	Distribution function	Link function
Bigge	gamma	identity
Henne	gamma	identity
Möhne	normal	identity
Sorpe	gamma	identity

Maximum likelihood estimation of GLM

Nelder and Wedderburn (1972) proposed an iteratively reweighted least squares method for maximum likelihood estimation of the GLM parameters. Dobson (2002) obtained maximum likelihood estimators of the parameters of generalized linear models (GLM) by an iterative weighted least squares procedure (see also, Charnes et al., 1976). He suggested solving the following equation iteratively to estimate the parameters of the GLMs

$$X^T W X b^{(m)} = X^T W z \tag{4.4}$$

where W is the $N \times N$ diagonal matrix with elements

$$w_{ii} = \frac{1}{var(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \tag{4.5}$$

and $b^{(m)}$ is the vector of estimates of the parameters β_1, ..., β_k at the m^{th} iteration and z has elements

$$z_i = \sum_{j=1}^{k} x_{ij} b_j^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 \tag{4.6}$$

with $\mu_i$and $\frac{\partial \eta_i}{\partial \mu_i}$ evaluated at $b^{(m-1)}$.

The following procedure can be used to solve equation, 4.4 (Dobson, 2002):

1. Use some initial approximation $b^{(0)}$ to evaluate z and W.

2. Solve equation (4.4) to give $b^{(1)}$.

3. Use $b^{(1)}$ to obtain better approximations for z and W.

4. Repeat step 3 until adequate convergence is achieved. In other words when the difference between successive approximations b$^{(m-1)}$ and b$^{(m)}$ is sufficiently small, b$^{(m)}$ is taken as the maximum likelihood estimate.

Table 4.3: *The performances of the missing data estimation models*

	Bigge		Möhne		Henne		Sorpe	
	R	$rmse$	R	$rmse$	R	$rmse$	R	$rmse$
BPNN	0.889	5.327	0.953	0.789	0.932	2.28	0.974	0.402
ANFIS	0.88	5.531	0.941	0.868	0.933	2.283	0.971	0.417
GLM	0.881	5.398	0.953	0.8	0.928	2.346	0.968	0.432

4.5 Evaluation of the models

The performances of the BPNN and ANFIS and GLM are evaluated by estimating the correlation coefficient (R) and the root mean square error ($rmse$). The root mean square error is calculated as follows:

$$rmse = \sqrt{\frac{1}{m}\sum_{i=1}^{i=m}(x_i - \hat{x}_i)^2} \tag{4.7}$$

where m is the number of observations in the validation data set, x_i is the observed daily inflow and $\hat{x}_i$ is the estimated daily inflow.

The estimated values of R and $rmse$ for the validation data sets are given in table 4.3. The results indicate that there is no significant difference in the estimated R and $rmse$ values among all the estimated daily inflow time series using the three models but BPNN models are slightly better. Due to that we assumed the BPNN model is better than the ANFIS and GLM to fill in the missing data. Figure 4.2 shows the observed vs. the estimated daily inflow using the BPNN models.

4.6 Extension of the monthly inflow data

We used the BPNN model to estimate daily inflow into the Bigge reservoir in the period from 1/11/1965 to 31/10/2006. The monthly inflow is computed by estimating the average value of the daily average inflow during each month. The statistical parameters (mean, standard deviation, skeweness and lag one month-to-month correlation) of the estimated monthly inflow time series are compared to those of the observed one and the results are shown in figure 4.3. By inspecting the results of figure 4.3, it can be seen that the BPNN model was found to produce satisfactory results. It preserves well the monthly mean, standard deviation and skewness coefficient of the observed monthly inflow time series as well as the lag one month-to-month correlation.

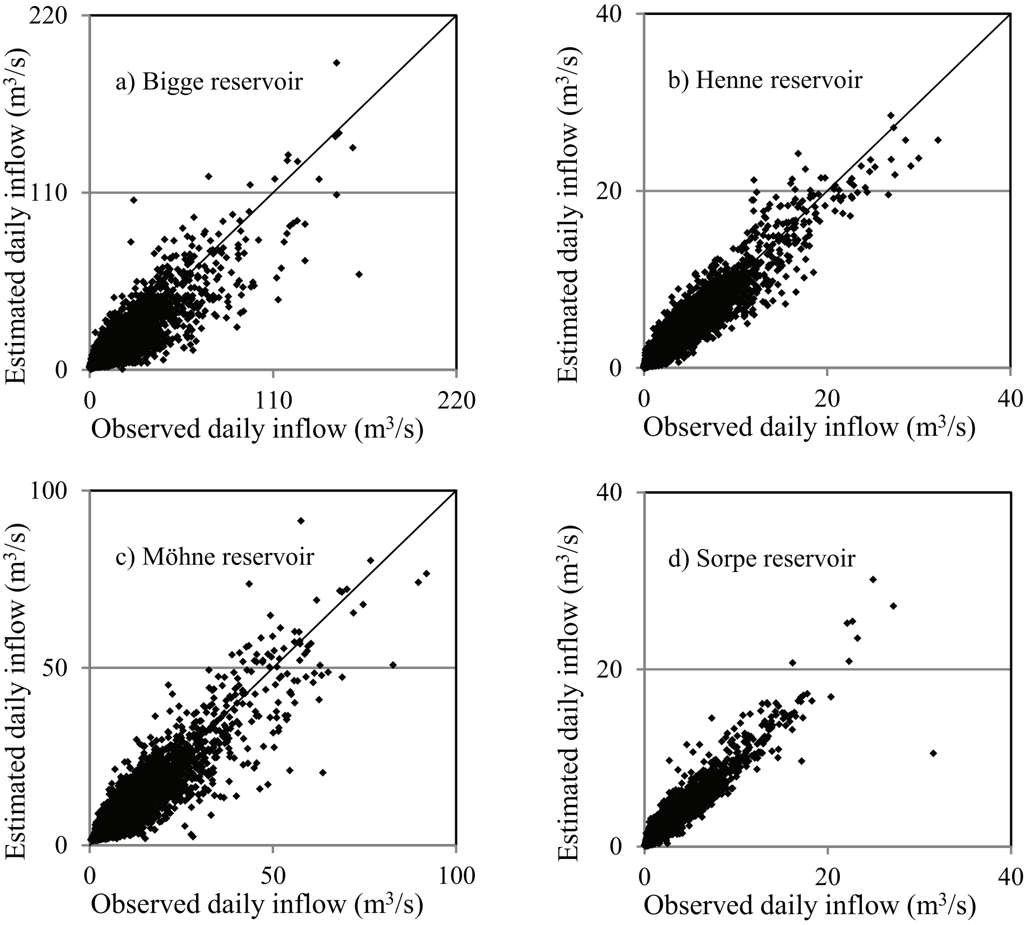


Figure 4.2: *Comparison between the observed and estimated (using the BPNN models) daily inflow time series*

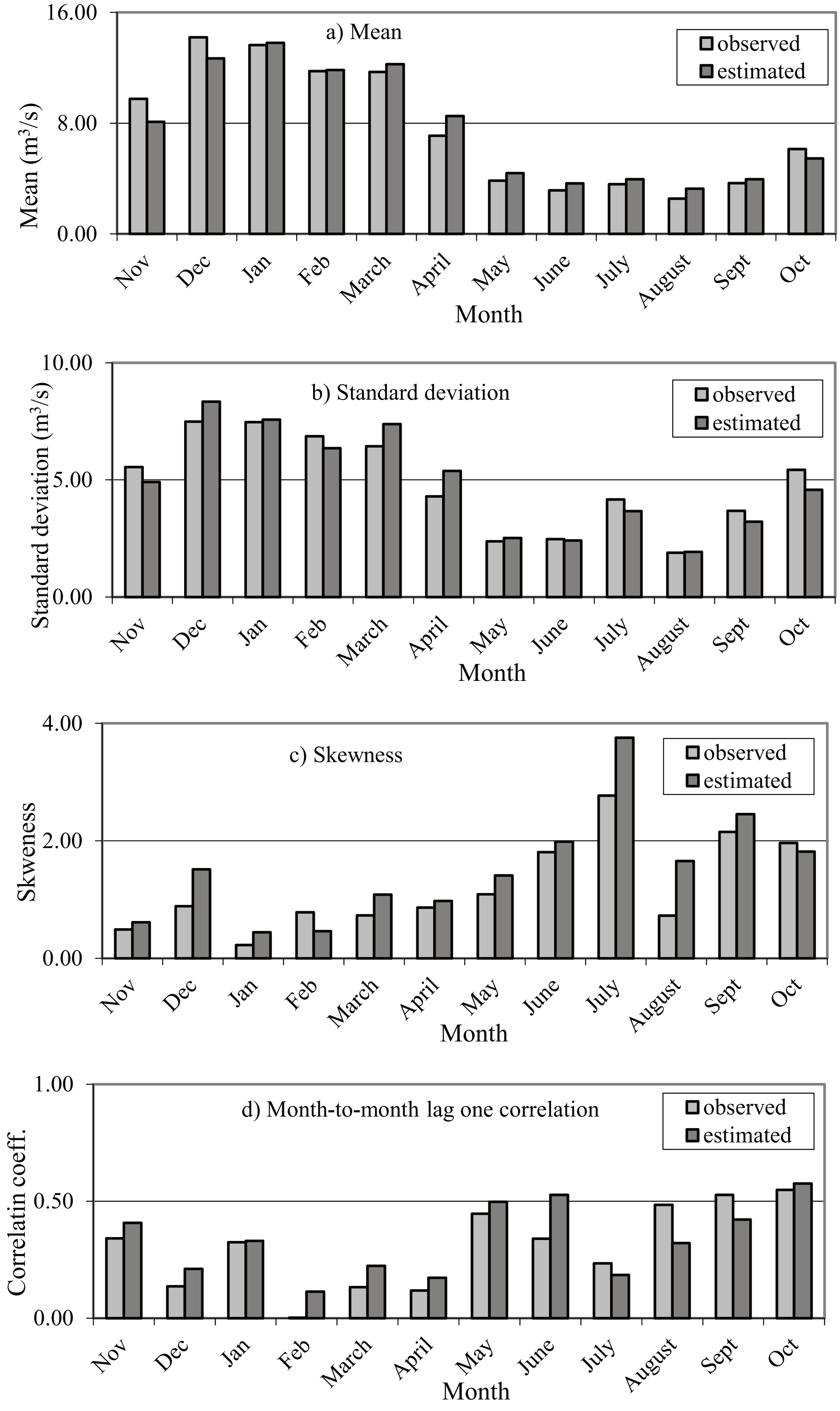


Figure 4.3: *Statistical parameters of the observed and estimated monthly inflow of the Bigge reservoir*

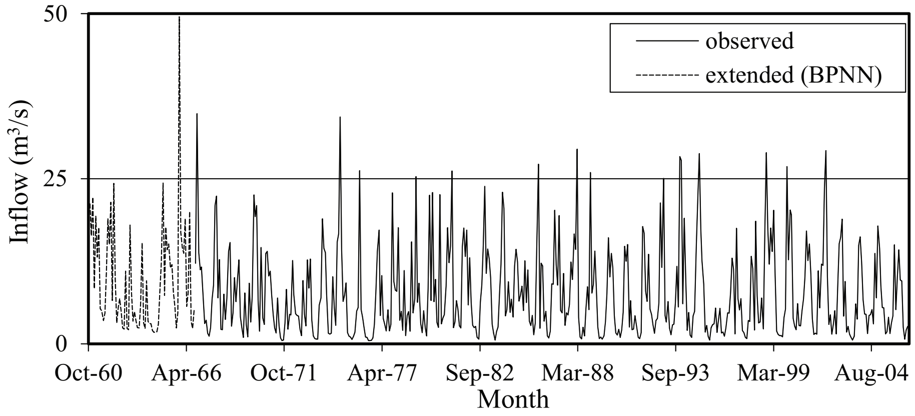


Figure 4.4: *The observed and extended monthly inflow of the Bigge reservoir*

From the above mentioned results, it is clear that the BPNN model can be used to extend the daily inflow data of the Bigge reservoir in the period from 1/11/1960 to 31/10/1965. The extended monthly inflow in the period from 11/1960 to 10/1965 and the observed one are plotted in figure 4.4.

4.7 Conclusions

We explored the efficiency of the BPNN and ANFIS models and the GLM for filling in the missing values in the daily inflow time series. High correlations between the inflow time series (QA, QH, QM and QS) are detected. The inflow of each reservoir is estimated using the inflow of the other reservoirs as input variables.

The BPNN models are trained using one hidden layer with three neurons. The ANFIS models with three membership functions (with Gaussian type) associated with each input are used. Trial-and-error procedure is applied to select the link function and the distribution of the response for the GLM.

Two efficiency criteria (R and $rmse$) are used to compare the performances of the models. The results show that there is no significant difference in the performances of the models however the BPNN models have slightly better performances in filling in missing data.

To ensure the ability of the BPNN model to extend the monthly inflow data of the Bigge reservoir in the period from 11/1960 to 10/1965, the BPNN model is used to estimate the daily inflow into the Bigge reservoir in the period from 1/11/1965 to 31/10/2006. The mean, standard deviation, skeweness and lag one month-to-month correlation of the estimated monthly inflow time series are compared to those of the observed one and the results show that the BPNN model produce satisfactory results.

4.8 References

Agresti, A., 2002. Categorical data analysis. 2nd edition, Wiley–Interscience, New Jersey.

Chandler, R.E., Wheater, H.S. 2002. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. Water Resources Research, 38(10), 1192–1203.

Charnes, A., Frome, E.L., Yu, P.L., 1976. The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. Journal of the American Statistical Association, 71(353), 169–171.

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

Dastorani, M.T., Wright, N.G., 2004. A hydrodynamic/neural network approach for enhanced river flow prediction. International Journal of Civil Engineering, 2(3), 141–148.

Dastorani, M.T., Wright, N.G., 2003. Flow estimation for ungauged catchments using a neural network method. In Proceedings of the 6th international river engineering conference. Ahwaz, Iran.

De Jong, P., Heller, G.Z., 2008. Generalized linear models for insurance data. Cambridge University Press.

Dobson, A.J. 2002. An introduction to generalized linear models. 2nd edition, Chapman & Hall, London.

Gupta, A., Lam., M., 1996. Estimating missing values using neural networks. Journal of the Operation Research Society, 47, 229–238.

Khalil, M., Panua, U.S., Lennox, W.C., 2001. Groups and neural networks based streamflow data infilling procedures. Journal of Hydrology, 241, 153–176.

Little, M.A., McSharry, P.E., Taylor, J.W., 2009. Generalized linear models for site–specific density forecasting of UK daily rainfall. Monthly Weather Review, 137(3), 1029–1045.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models. 2nd edition, Chapman & Hall, London.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A, 135(3), 370–384.

Petersen, G., Bast, H., Fohrer, N., 2008. Estimation of ungauged Bahr el Jebel flows based on upstream water levels and large scale spatial rainfall data. Advances in Geosciences, 18, 9–13.

Wedderburn, R.W.M., 1974. Quasi–1ike1ihood functions, generalized linear models and the Gauss–Newton method. Biometrika, 61, 439–447.

Yang, C., Chandler, R.E., Isham, V.S., Wheater, H.S., 2005. Spatial–temporal rainfall simulation using generalized linear models. Water Resources Research, 41(11), 1–13.

Chapter 5

Monthly inflow generation

5.1 Introduction

Time series of inflow data are essential for proper planning, design and operation of many water resources systems. However, presently for most of the reservoirs, the measured length of inflow data is limited. Streamflow generation procedures play an important role for obtaining reliable estimate of flow statistics.

In this chapter, we apply four models to generate monthly inflow data into the Bigge, Henne, Möhne and Sorpe reservoirs. These models are the Thomas-Fiering (T-F) model, the Gamma Thomas-Fiering (Gamma T-F) model, the Monte Carlo (MC) model and the hidden Markov model with periodic states (PHMM). The results of the T-F, Gamma T-F, MC and PHMM models are discussed and compared to choose the best model for monthly inflow generation. We developed a procedure to detect the consecutive 5 years that have minimum total inflow. The predicted critical consecutive 5 years monthly inflow time series can be used as an inflow scenario for optimal operation of the reservoirs.

Several stochastic models have been proposed for modeling hydrological time series and generating synthetic streamflows. This synthetic flow should resemble the historical time series. Selection of the appropriate model needs to consider the hydrologic characteristics, data availability and the statistical properties (Kim and et al., 2004). Phien and Khan (1981) used Thomas-Fiering (T-F) and Spolia-Chander (S-C) models for monthly streamflow generation. They compared the results of the two models based on the reproduction of the historical record in terms of several important statistics such as the mean, standard deviation, skewness coefficient, correlation coefficient and the reservoir storage components and they proved that the Thomas-Fiering model is superior to the Spolia-Chander model. Alhassoun et al. (1997) generated the monthly evaporation sequences for ten selected stations in the Saudi-Arabia. They used three autoregressive models for generating: method of fragments (M-F), Thomas-Fiering (T-F) model and Two-Tier (T-T) model. The performances of the models were evaluated by comparing the statistical parameters of the generated sequences with those of the historical data. The T-F model gave the best representation of the mean, standard deviation, skewness and lags one autocorrelation of the monthly evaporation sequences. Raman and Sunilkumar (1995) employed an artificial neural network (ANN) models and autoregressive moving average (ARMA) model for the synthesis of monthly inflow for two reservoirs in the Bharathapuzha basin in south India. They concluded that the results obtained using the neural network model compared well in the mean with those obtained using the autoregressive model. Ochoa-Riveria et al. (2002) presented an ANN based model for multivariate streamflow generation. The model consists of two components, the neural network (NN) deterministic component and a random component which is assumed to be normally distributed. They compared the results of the ANN based models to those of a lag two autoregressive AR(2) and concluded that the ANN represents a promising modeling alternative for simulation purposes. Sarma and Ahmed (2002) and (2004) used an ANN model for synthetic streamflow generation of a Himalayan River called Pagladia. However, the model could not able to generate a good synthetic streamflow series. Kim and et al. (2004) used Monte Carlo (MC), lag-one autoregressive AR(1) and Thomas-Fiering (T-F) models for the annual and monthly streamflow simulations. They repeated the simulation using stochastic models by bootstrap resampling scheme (bootstrapped MC, AR(1) and T-F models) and showed that the bootstrapped stochastic models are much better than the stochastic models for the simulation study. The simulated series by the bootstrapped stochastic models reproduced the skewness coefficient and the probability density function of observed series very well.

Celeste et al. (2004) used T-F model to determine monthly inflow scenarios for the watershed of the reservoir that supplies the city of Matsuyama, Japan. They used the generated scenarios for optimal operation of the reservoir. A periodic autoregressive moving average (PARMA) model was adapted and applied by Mendes et al. (2007) for monthly synthetic streamflow generation. They assumed that the streamflow of a certain month depends explicitly on the streamflow of the prior month, the streamflow of the same month in the previous year and in the preceding year, as well as on a random noise. They concluded that the adopted model reproduces properly annual and periodic statistics values of generated series.

The interest in HMM application has increased in the recent years. They are applied by many researches to stochastic hydrology, particularly in climate variability application. Zucchini and Guttorp (1991) applied a hidden Markov model (HMM) for the analysis of rainfall occurrences at several sites. Charles et al. (1999) and Bellone et al. (2000) extended the work of Zucchini and Guttorp (1991) to rainfall amounts by relating local precipitation to atmospheric circulation. They used atmospheric data to modify the transition probabilities of the Markov process. Betrò et al. (2008) applied a homogeneous HMM to daily rainfall data collected at four pluviometric stations in Central-East Sardinia. They introduced mixtures of Weibull distributions to model positive rainfall amounts.

5.2 Data used

The historical records of the monthly inflow into the Bigge (after extending the data, see chapter 4), Henne, Möhne and Sorpe reservoirs in the period from 1961 to 2006 are used to train the models. The monthly inflow is computed by estimating the average value of the average daily inflow during each month. Figures 5.1.a, b, c and d show the plots of the monthly inflow time series of the Bigge, Henne, Möhne and Sorpe reservoirs respectively.

5.3 Thomas-Fiering model

Autoregressive (AR) models were used by many researchers to reproduce the statistical properties of the hydrologic time series. An autoregressive model of order p model is given as:

$$y_{\mathrm{t}} = \mu + \sum_{j=1}^{p} \phi_j (y_{t-j} - \mu) + \varepsilon_{\mathrm{t}} \tag{5.1}$$

in which ϕ_1, ϕ_2, $\cdots$, $\phi_p$are the parameters of the model, μ is the mean and ϵ_t is an uncorrelated normal variable with mean zero and variance $\sigma^2(\epsilon)$. Uncorrelated means there is no correlation between ϵ_t and y_{t-1},y_{t-2},$\dots$, y_{t--p}(Maidment, 1993). A periodic time series model with a periodic hydrologic process $y_{\nu,\tau}$ inwhich ν is the year and τ is the season, τ = 1, 2, ..., ω and ω is the number of seasons in the year (for example $\omega = 12$ months) can be defined as:

$$y_{\nu,\tau} = \mu_\tau + \sum_{j=1}^{p} \phi_{j,\tau} (y_{\nu,\tau-j} - \mu_{\tau-j}) + \varepsilon_{\nu,\tau} \tag{5.2}$$

and the model is often denoted as PAR(p). The PAR(1) model arises by making $p = 1$ in equation (5.2) as:

$$y_{\nu,\tau} = \mu_\tau + \phi_{1,\tau}(y_{\nu,\tau-1} - \mu_{\tau-1}) + \varepsilon_{\nu,\tau} \tag{5.3}$$

Thomas and Fiering (1962) used this model to simulate monthly streamflow. We used the Thomas-Fiering (T-F) model in the present study to generate monthly inflow time series. The T-F model presents a set of 12 regression equations which can be expressed as follows (Phien and Ruksasilp, 1981):

$$q_{\nu,\tau} = q'_\tau + \frac{r_\tau S_\tau (q_{\nu,\tau-1} - q'_{\tau-1})}{S_{\tau-1}} + Z_{\nu,\tau} S_\tau \sqrt{(1 - r_\tau^2)} \tag{5.4}$$

where

$q_{\nu,\tau-1}$ is the $i - 1^{th}$ value for the $\tau - 1^{th}$ month,

$q_{\nu,\tau}$ is the i^{th} simulated value for the τ^{th} month,

$q\prime_\tau$ and $q\prime_{\tau-1}$ are the mean monthly values during the τ and $\tau - 1^{th}$ months respectively,

r_τ is the cross correlation between the monthly values during the $\tau - 1$ and τ^{th} months respectively,

$S_{\tau-1}$ and S_τ are the standard deviations of monthly values during the $\tau - 1$ and τ^{th} months respectively and

$Z_{\nu,\tau}$ is a random Normal deviate N(0,1).

In the T-F model, the effects of seasonality on the variability of the data are accounted for by considering month-to-month variation in the average value and month-to-month coefficient of correlation.

5.4 Gamma Thomas-Fiering model (Gamma T-F)

Wilson and Hilferty (1931) gave the following transformation to estimate the skewed deviate ($Zr_{\nu,\tau}$), from the normal deviate ($Z_{\nu,\tau}$) to deal with skewed data

$$Zr_{\nu,\tau} = \frac{2}{gt_\tau}\left[1 + \frac{gt_\tau Z_{\nu,\tau}}{6} + \frac{gt_\tau^2}{36}\right] - \frac{2}{gt_\tau} \tag{5.5}$$

where gt_τ = the coefficient of skewness of the random skewed deviate during the τ^{th} month.

Thomas and Burden (1963) derived the following formula to estimate gt_τ

$$gt_\tau = \frac{g_\tau - r_\tau^3 g_{\tau-1}}{(1-r_\tau^2)^{3/2}} \tag{5.6}$$

where g_τ and $g_{\tau-1}$ are the seasonal coefficients of skewness during the τ^{th} and τ-1^{th}months respectively.

By replacing $Z_{\nu,\tau}$in equation (5.4) with $Zr_{\nu,\tau}$,the Gamma T-F model can be presented as follows:

$$q_{\nu,\tau} = q'_\tau + \frac{r_\tau S_\tau (q_{\nu,\tau-1} - q'_{\tau-1})}{S_{\tau-1}} + Zr_{\nu,\tau} S_\tau \sqrt{(1-r_\tau^2)} \tag{5.7}$$

5.5 Monte Carlo model

The Monte Carlo simulation is a method for obtaining the probability distribution of an output given the probability distribution of one ore more inputs (Maidment, 1993). Equation (5.8) is assumed here to introduce the monthly inflow as a function of q'_τ and S_τ

$$q_{\nu,\tau} = q'_\tau + R_{\nu,\tau} S_\tau \tag{5.8}$$

where

$q_{\nu,\tau}$	is the simulated value for the τ^{th} month,
$q\prime_\tau$	is the mean monthly value during the τ^{th} month,
$R_{\nu,\tau}$	is a random value, and
S_τ	is the standard deviation of monthly value during the τ^{th} month.

5.6 Random values generation

As previously presented in chapter 2, the skewness values of the monthly inflow time series into the Bigge, Henne, Möhne and Sorpe reservoirs are positive (see figure 2.9). Positive values of skewness indicate that these time series are not normally distributed. Non-normality of time series means non-normality of the distributions of Z and R (equations, 5.4 and 5.8 respectively). We assumed the following procedure to generate Z and R in the T-F and MC models respectively:

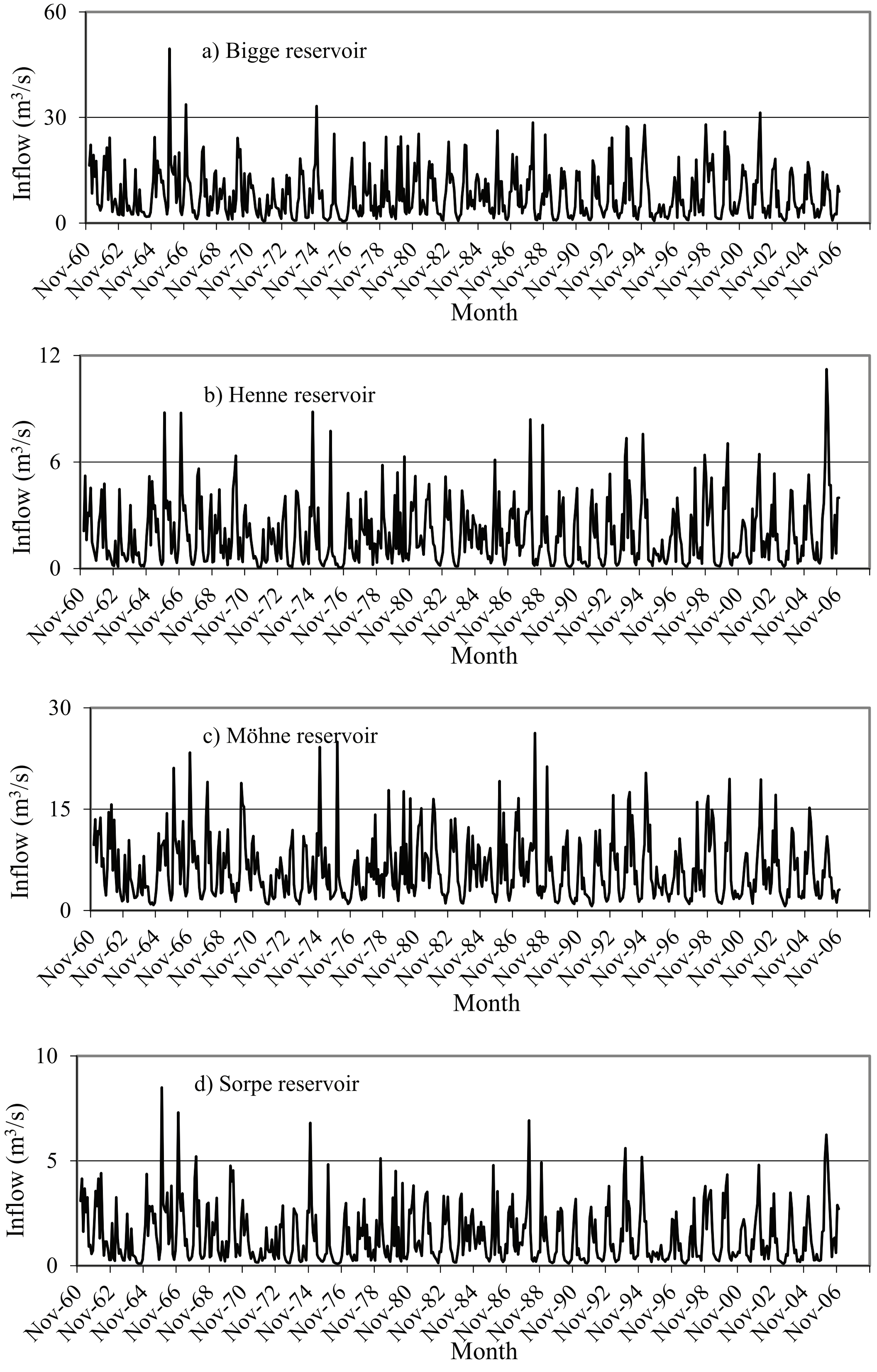


Figure 5.1: *Plots of the monthly inflow*

1. Estimate the random values Z and R from the observed data and denote them by Z_{obs} and R_{obs} respectively.

2. Use the inverse transform method to generate random values from Z_{obs} or R_{obs}.

The previous procedure is employed to generate the random values Z and R in the T-F and MC models respectively. The random value in the Gamma T-F model Z_r is generated from the normal distribution and then the Wilson-Hilferty transformation is applied (equation, 5.5) .

Inverse transform method

The inverse transform method is a method for generating random numbers from any probability distribution given its cumulative distribution function (cdf). As shown in figure 5.2 suppose $X = (x_1, x_2, \ldots, x_n)$ where $x_1 < x_2 < \ldots < x_n$ then the cumulative distribution function (cdf) of X, $F(x)$ can be given as

$$F(x_i) = \begin{cases} 0 & x_i < x_1 \\ \sum_{j=1}^{i} P(X = x_j) & x_i \leq x_{i+1},\ i \leq n-1 \\ 1 & x_i \geq x_n \end{cases} \tag{5.9}$$

The following algorithm is used to generate a discrete random number y from X (Hisashi, 2004):

1. Generate a random number $u\,(0 \leq u \leq 1)$from the uniform distribution.

2. The random draw of X is given by x_i if $F(x_{i-1}) \leq u < F\,(x_i)$ where $F(x_i) = \sum_{j=1}^{i} P(X = x_j)$ and $F(x_o) = 0$ (see figure 5.2) .

3. The generated random number y is equal to x_i.

5.7 Lag one month-to-month correlation

Generation of monthly inflow data using the Monte Carlo model depends on the generated random values $R_{\nu,\tau}$. Month-to-month correlations are not considered in equation, 5.8 (the MC model), which means that the generated random values must be correlated to preserve the correlation in the generated inflow data. The Cholesky decomposition method is used to preserve the month-to-month correlation in the generated inflow data.

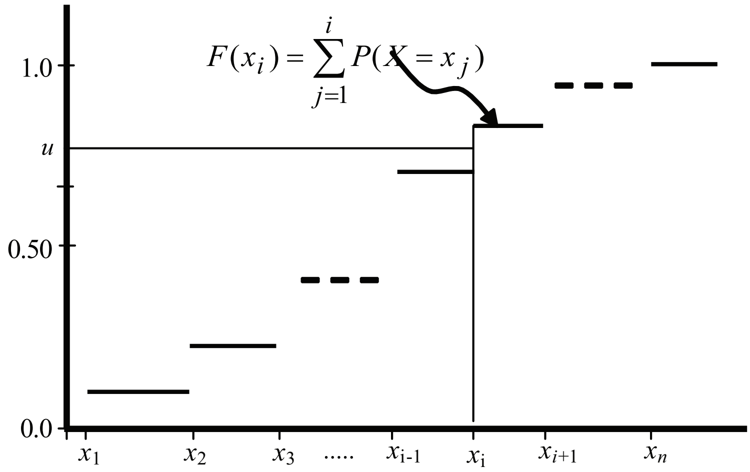


Figure 5.2: *Cumulative distribution function of a variable X (Gentle, 2003)*

Generating multiple sequences of correlated random values

Generation of correlated random numbers with a given correlation matrix, C (month-to-month correlation) is done by finding a matrix F such that:

$$F^T F = C \tag{5.10}$$

The Cholesky decomposition of the correlation matrix is the most common methods to solve equation 5.10. Correlated random numbers, Rc can be generated from uncorrelated numbers R by multiplying R with F

$$Rc = RF \tag{5.11}$$

To preserve the lag one month-to-month correlation, the random value R in the MC model (equation, 5.8) is replaced by the correlated random value Rc as follows:

$$q_{\nu,\tau} = q'_\tau + Rc_{\nu,\tau} S_\tau \tag{5.12}$$

The MC model is applied to generate 300 years monthly inflow data of the Bigge, Henne, Möhne and Sorpe reservoirs. Figures 5.3.a, b, c and d show the lag one month-to-month correlation of the observed and the generated monthly inflow data of the Bigge, Henne, Möhne and Sorpe reservoirs respectively. It is obvious that there is no significance lag one month-to-month correlation (the values are nearly equal to zero) for the generated inflow data by using the MC model with random values R (without using Cholesky decomposition). In contrast, the values of lag one month-to-month correlation of the generated inflow data by using the MC model with random value Rc (after using Cholesky decomposition) are very close to those of the observed data. The Monte Carlo models with random values Rc are used in the present study to preserve the lag one month-to month correlation.

5.8 Hidden Markov models (HMM)

5.8.1 Introduction to HMM

The Hidden Markov model (HMM) is the model with a sequence of observed emissions (E) and unobserved sequences of states (S). Moving from one state to another depends on the matrix of transition probabilities (A). The probability to move from state S_i to S_j is denoted by a_{ij}. The outcome emitted by each state depends on the matrix of emission probabilities (B).

To define an HMM we have to know (Rabiner, 1989):

1. The hidden states are S_1, S_2, $\ldots$, S_N where N is the number of states in the model.
2. The hidden state transition matrix (A) of size $N \times N$. The sum of the entries of each row of A is equal to 1. Each element of A is denoted as:

   $$a_{ij} = P\left[d_{t+1} = S_j | d_t = S_i\right], \qquad 1 \leq i, j \leq N \tag{5.13}$$

   where d_t is the hidden state at time t.
3. The set of distinct observations per state $G = \{g_1,\, g_2,\, \ldots,\, g_M\}$ where M is number of observations.
4. The observation probability distribution in state j, $B = \{b_j(k)\}$,where

   $$b_j(k) = P\left[g_k \,\mathrm{at}\, t \,| d_t = S_j\,\right] \qquad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{5.14}$$

5. The initial state distribution $\pi = \{\pi_i\}$where

   $$\pi_i = P\left[d_1 = S_i\right], \quad 1 \leq i \leq N \tag{5.15}$$

Given appropriate values of N, M, A, B and π, the following procedure is introduced by Rabiner (1989) to use HMM as a generator to give an observation sequence $\{O_t\}$, $t =$ 1, 2, ..., $T$where T is the number of observations in the sequence and $O_t \in \; G$:

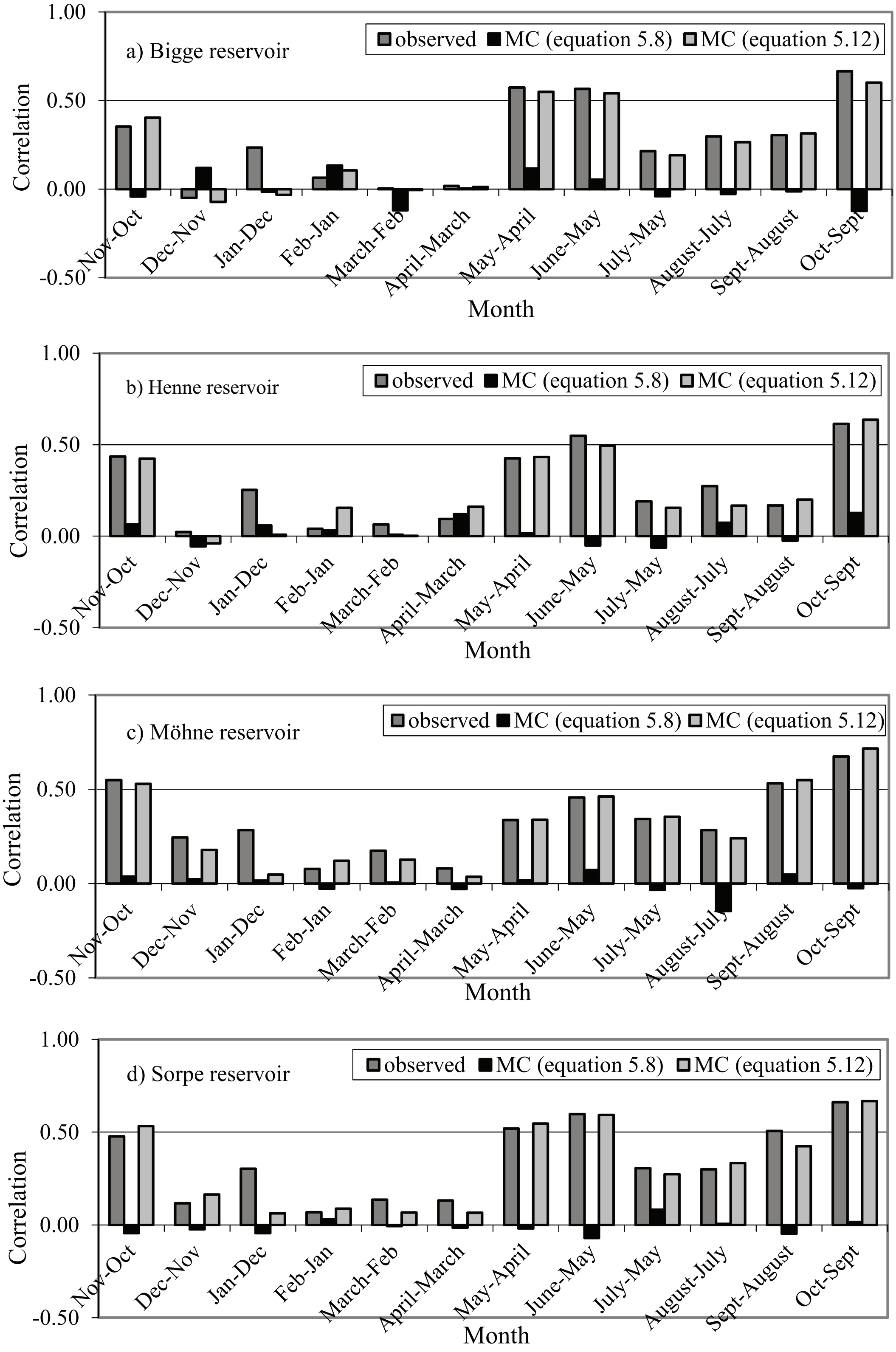


Figure 5.3: *The lag one month-to-month correlation of the observed and generated (using the MC model) monthly inflow*

1. Assume an initial state distribution π and choose the corresponding initial state $d_1 = S_i,\ 1 \leq i \leq N$.

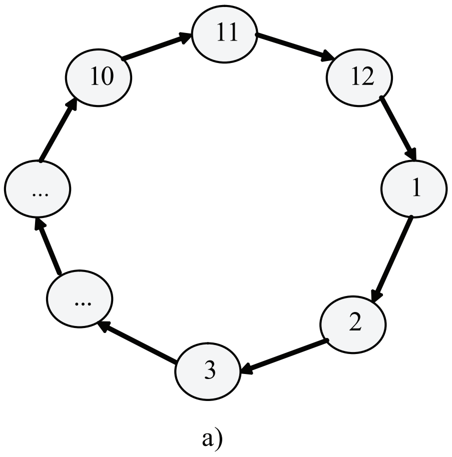
2. Set t = 1 and choose $O_t = g_k$ according to the observation probability distribution in state S_i, i.e., $b_i(k),\ 1 \leq k \leq M$.

3. Go to new state d_{t+1}= S_j according to the state transition probability for state S_i, i.e., a_{ij}.

4. Set t = t +1; return to step 3 if t < T; otherwise terminate the procedure.

5.8.2 HMM for monthly inflow generation

We classified the observed monthly inflow time series in 12 states (11, 12, 1, 2, ..., 9, 10) each state corresponds to a specific month. The transition diagram and the transition matrix for monthly inflow generation HMM model are shown in figures 5.4.a and 5.4.b respectively. From figure 5.4.a it is clear that each state i has 12 periods which means that the model returns to state i in multiples of 12 time steps. We denoted our model by periodic hidden Markov model (PHMM).



$$A = \begin{array}{c|cccccccccccc} & 11 & 12 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \hline 11 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 10 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

b)

Figure 5.4: *a) The transition diagram and b) The transition matrix, for monthly inflow generation HMM*

HMM codes are available in the Statistical toolbox, Matlab.b (2008). We used a Matlab code hmmdestimate to estimate the transition and emission matrices and a modified version of hmmgenerate (monthly_flow_hm) to generate simulated sequences as follows:

1. The code generates a set of random numbers $\{Rn_i\}$, i = 1, 2,...,T from an uniform distribution.

2. The first generated random number Rn1 is compared to the emission probability matrix of the state 11. The code emits an observation g_k if $b_{11}(k-1) < b_{11}(k) < Rn_1$ and then transits to the next state.

3. Repeat step 2 for all values of $\{Rn_i\}$ to generate T values of monthly inflow.

5.9 Validation of the models

The synthetic streamflow generation models should preserve the statistical parameters of the observed data. Model validation is the most important step in the model building sequence to check ability of the model to reproduce the important statistical parameters of the observed data. Two methods are used in the present study to validate the models as follows:

1. Comparing the statistical parameters (mean, standard deviation, month-to-month correlation and skewness) of the generated monthly inflow with those of the observed one (Alhassoun et al., 1997; Kim and et al., 2004; Ahmad and Sarma 2007).

2. Using visual validation, we fitted the quantile-quantile (Q-Q) and survivor function plots of the observed inflow data against the corresponding generated one.

5.9.1 Statistical parameters of the monthly inflow time series

The performances of the models are evaluated using the following statistical parameters:

q' is the monthly mean inflow,
SD is the standard deviation in the monthly inflow,
r is the lag one month-to-month correlation, and
g is the skewness.

The average relative error percentage ($AREP$) is used to test the ability of the models to reproduce the statistical parameters of the observed data. The $AREP$ is the average value of the estimated relative error percentage (REP) for the 12 months of the year. For example, the REP in the mean q' during the τ^{th} month is estimated as follows:

$$REP_{\tau} = 100 \times \left| \frac{q'_{\tau} - \hat{q}_{\tau}}{q'_{\tau}} \right| \tag{5.16}$$

where q'_{τ} and $\hat{q}_{\tau}$ are the mean monthly value during the τ^{th} month of the observed and generated inflow data respectively. Then the $AREP$ is given by

$$AREP = \frac{1}{12} \sum_{\tau=1}^{12} REP \tag{5.17}$$

We generated three monthly inflow time series with lengths 100, 300 and 500 years using T-F, Gamma T-F, MC and PHMM. The synthetic monthly inflow time series are compared with the observed one using the $AREP$ in the statistical parameters mean, standard deviation, month-to-month correlation and skewness. Figures 5.5, 5.6, 5.7 and 5.8 show the statistical parameters of the observed and the 300 years generated monthly inflow into the Bigge, Henne, Möhne and Sorpe reservoirs respectively. Table 5.1 lists the estimated values of $AREP$ in each statistic for each model. In respect of the estimated $AREP$, we observe the following for most synthetically generated series:

1. In respect of the mean value of the monthly inflow, the series generated by the T-F, MC and PHMM models are found to be quite closer to the observed series.

2. The MC and PHMM models reproduce the monthly standard deviation better than the other models.

3. The T-F and MC models are superior to the other models in terms of the monthly month-to-month correlation.

4. Compared with the other models, the PHMM model preserves the monthly skewness very well.

From the above results, it is clear that most of the statistics of the simulated series by the T-F model, MC model and PHMM reproduce those of observed one fairly well, especially, the MC model. PHMM preserves the mean, standard deviation and skewness very well however it cannot preserve month-to-month correlation.

5.9.2 Q-Q and survivor function plots for models validation

Quantile-quantile (Q-Q) and survivor function plots are used to check the ability of the T-F, Gamma T-F, MC and PHMM models to generate realistic monthly inflow data.

Quantile-Quantile Plot

Quantile-quantile (Q-Q) plot is a visualization technique to determine whether two samples come from the same distribution family. They are scatter plots of quantiles computed from each sample, with a line drawn between the first and third quartiles (the 25 percentile and 75 percentile respectively). If the data falls near the line, it is reasonable to assume that the two samples come from the same distribution.

Survivor function plot

The empirical survivor function is a non-parametric tool for analyzing survival data. The empirical survivor function is an estimate of the probability of survival past time y, which does not depend on distributional assumption.

The most common way to estimate the survivor function is the Kaplan-Meier method. This method was first proposed by Kaplan and Meier (1958). To use this method, the survival times are assumed to be independent. This method can be presented in the following steps (Tableman and Kim, 2004):

1. Order the k observed survival times by increasing magnitude, $0 \leq y_{(1)} \leq y_{(2)} \leq ... \leq y_{(k)}$.
2. Define t_i as the i_{th} unique value in the series $0, y_{(1)}, y_{(2)}, ...y_{(k)}, i = 1, 2, ..., m$.
3. Define n_i, as the number of subjects at risk until just before time t_i, $i = 1, 2, ..., m$.
4. Define l_i as the number of subjects which fail at time t_i, $i = 1, 2, ..., m$.
5. For each $i = 1, 2, ..., m$ compute
$$\widehat{S}_i = \left(\frac{n_i - l_i}{n_i}\right) \widehat{S}_{i-1}\ fori > 1 \tag{5.18}$$
6. The estimate of the survivor function is then
$$\widehat{S}_Y\ (y) = \widehat{S}_i \qquad t_i \leq y \leq t_{i+1} \tag{5.19}$$

Results of the visual validation methods

We used Q-Q and survivor function plots to compare the distribution of the 300 years generated monthly inflow time series with the distribution of the observed one. Figures 5.9, 5.10, 5.11 and 5.12 show the Q-Q plots of the observed versus the generated inflow data for the Bigge, Henne, Möhne and Sorpe reservoirs respectively. It is clear that the Q-Q plots show a clear superiority of the MC and PHMM models for all reservoirs especially in generating high inflow events.

The survivor function plots of the observed versus the generated inflow data for the Bigge, Henne, Möhne and Sorpe reservoirs are shown in figures 5.13, 5.14, 5.15 and 5.16 respectively. It is to be seen that the survivor function plots for the generated inflow by the PMHH model are very close to those of the observed one. Also, the plots show that the series generated by the MC models are found to be better than that generated by the T-F and Gamma T-F models in respect of the survivor function plots.

Table 5.1: *List of the AREP in the tested statistical parameters*

		100 years				300 years				500 years			
		Bigge	Henne	Möhne	Sorpe	Bigge	Henne	Möhne	Sorpe	Bigge	Henne	Möhne	Sorpe
T-F	qt'	2.5	2.93	1.49	1.91	2.41	2.52	1.07	2.41	2.14	2.42	1.09	2.38
	s	11.08	10.97	14.69	16.16	11.51	10.06	13.25	15.58	11.12	10.67	14.47	15.43
	r	14.11	15.52	11.72	18.71	12.12	13.59	13.34	15.74	14.89	16.05	11.68	13.31
	g	22.12	16.26	16.63	22.69	24.06	15.42	14.56	22.71	21.85	15.87	15.78	23.95
Gamma T-F	qt'	14.03	12.78	7.42	16.26	14.4	12.75	8.04	17.14	14.85	13.61	7.94	16.87
	s	7.81	5.18	5.9	8.65	7.43	4.22	4.11	8.57	6.89	4.8	4.02	7.94
	r	30.82	32.43	23.24	28.16	32.82	35.05	23.11	30.29	33.52	30.03	23.03	30
	g	34.76	20.56	14.58	24.97	32.39	23.4	12.41	24.36	35.68	20.24	14.11	27.28
MC	qt'	1.18	2.39	3.07	3.7	1.19	1.76	2.35	3.59	2.18	2.52	1.62	3.04
	s	5.22	3.72	3.75	4.31	3.92	4.33	3.9	3.8	3.06	3.38	3.87	4.22
	r	16.5	16.41	15.9	20.2	16.71	17.76	12.57	15.95	18.36	16.57	12.22	15.7
	g	20.58	15.3	21.17	20.68	16.76	11.66	17.32	16.42	17.2	12.39	14.71	16.32
PHMM	qt'	2	2.07	2.37	3.39	1.32	3.09	0.88	2.82	0.77	3.7	0.97	3.37
	s	2.6	6.72	4.22	4.68	2.19	7.09	1.75	4.05	1.52	6.57	1.43	4.56
	r	100.66	92.3	100.66	97.1	99.72	99.94	98.96	100.02	101.63	101.77	98.83	99.11
	g	12.06	26.21	7.17	9.34	5.63	26.86	3.46	8.1	4.48	28.51	2.55	8.24

5.10 Consecutive 5 years with minimum inflow

Selection of the best monthly inflow generation model is an important requirement for improving the operation of the reservoirs in the Ruhr River basin. As mentioned in the previous section, the Monte Carlo MC is assumed to be better than the other models (the T-F, Gamma T-F and PHMM models) for monthly inflow generation. We used the simple moving average (*SMA*) to determine the critical consecutive 5 years with minimum *SMA* as follows:

1. Generate 1000 years of monthly inflow using the MC model.
2. Estimate the 5 years *SMA* of the generated monthly inflow

$$SMA_5 = \sum_{i-4}^{i} Q_i \qquad i = 5, 6, \cdots, 1000 \tag{5.20}$$

   where Q is the total annual inflow (million m^3).

3. Determine the minimum value of SMA_5 and denote it as $minSMA_5(j)$, j = 1, 2, . . . ,10000.

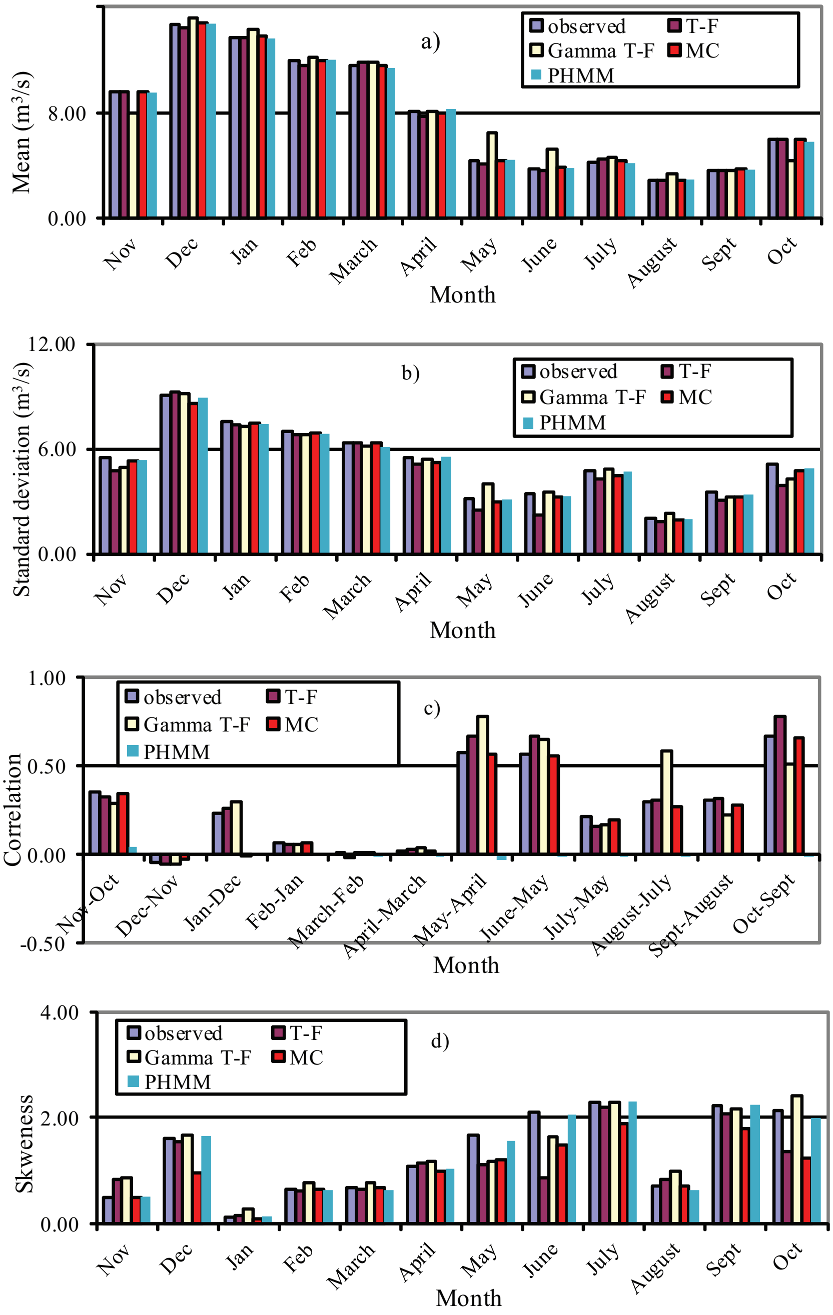


Figure 5.5: *Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Bigge reservoir*

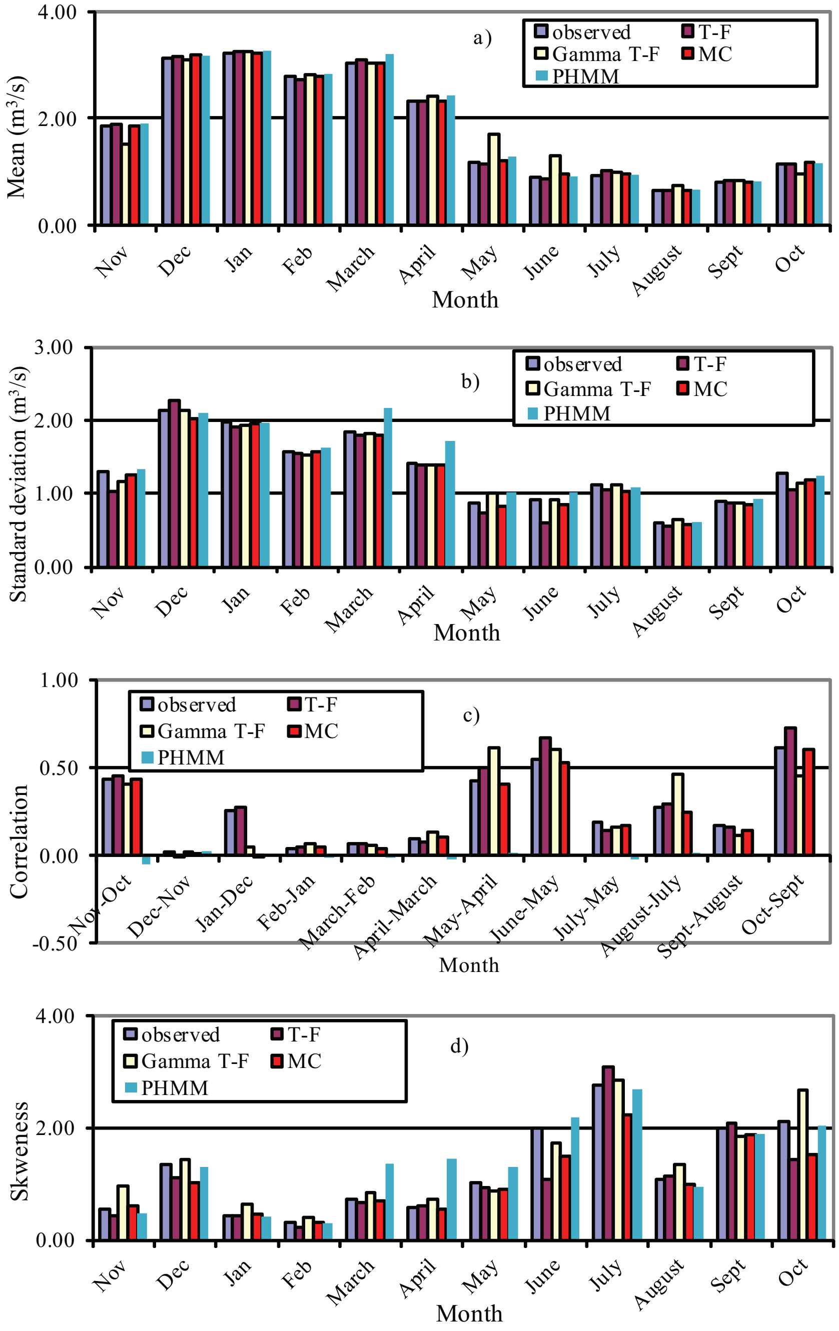


Figure 5.6: *Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Henne reservoir*

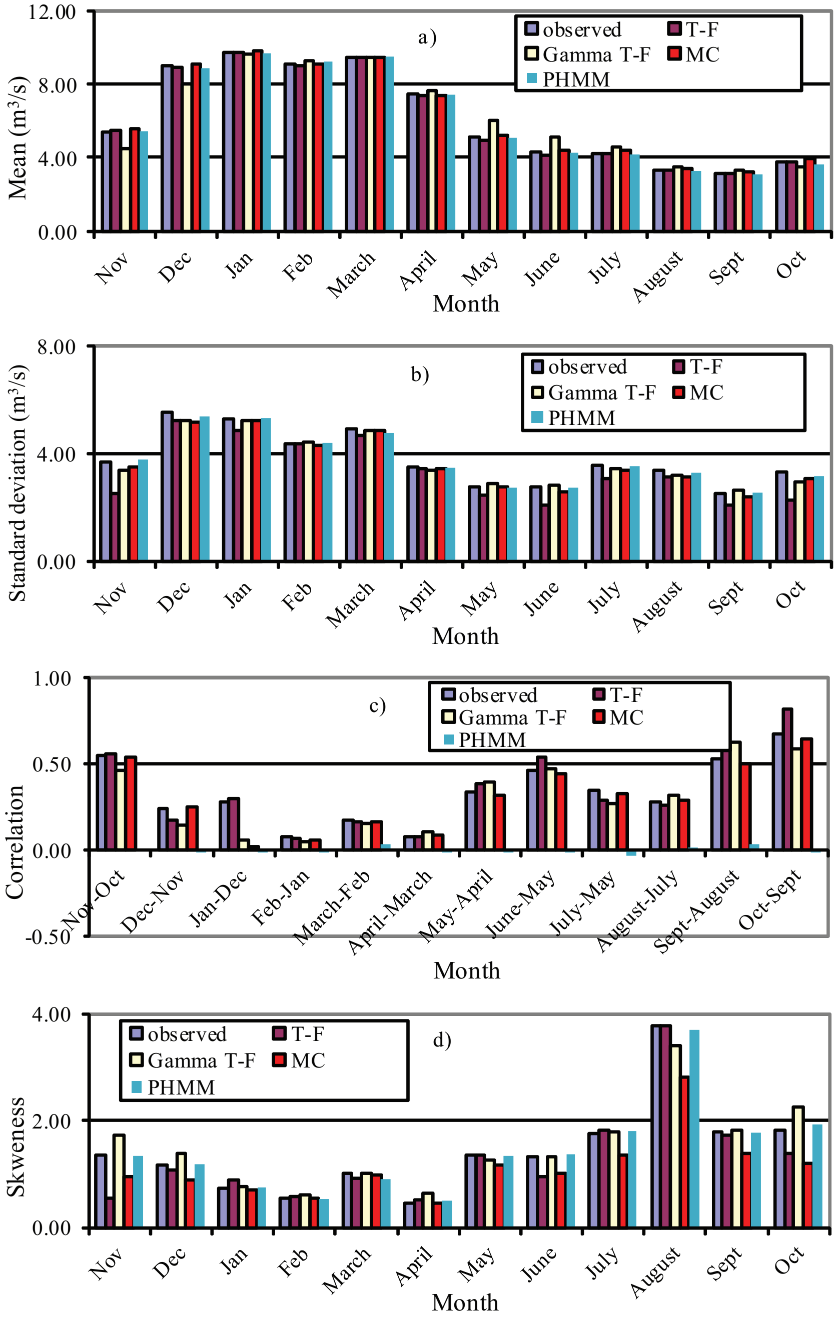


Figure 5.7: *Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Möhne reservoir*

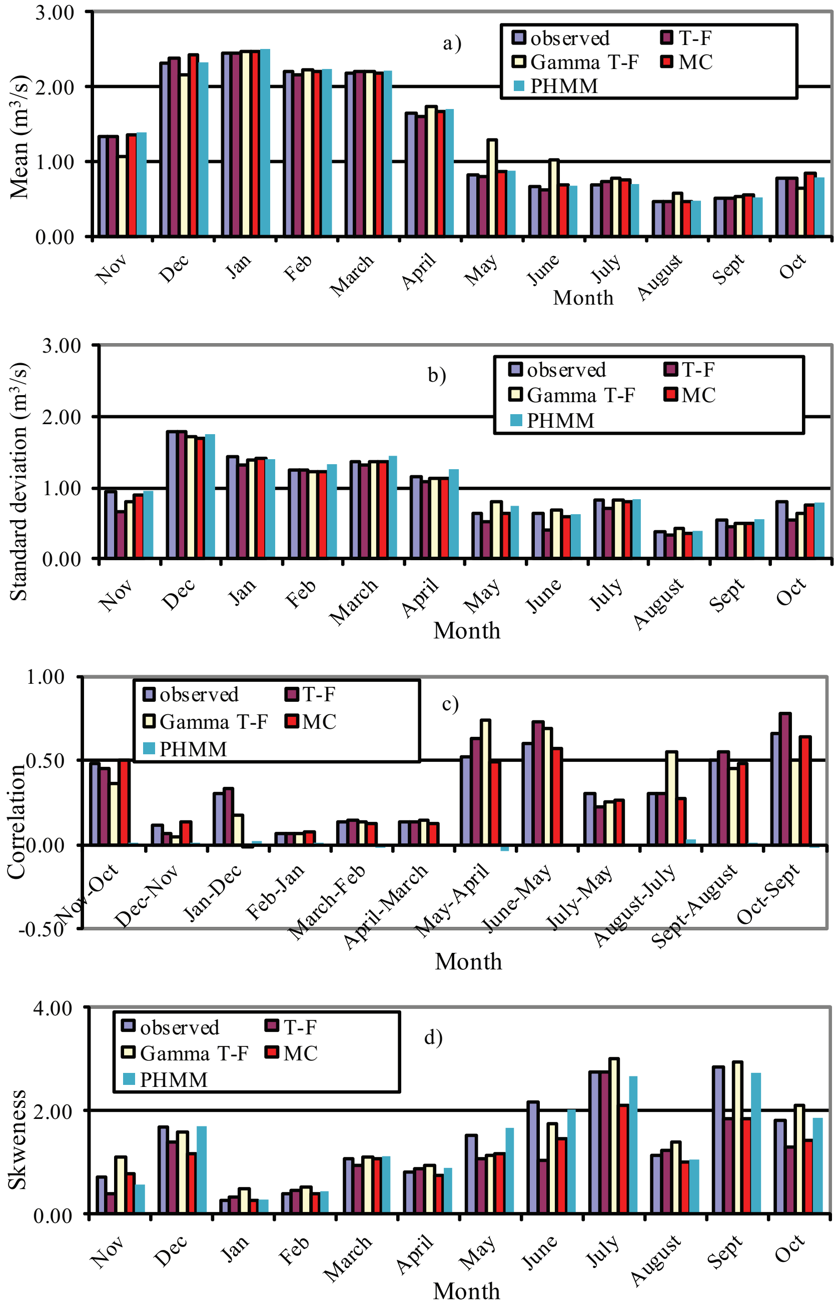


Figure 5.8: *Plots of the a) monthly mean, b) standard deviation, c) lag-1 month-to-month correlation and d) skewness of the observed and 300 years generated monthly inflow of the Sorpe reservoir*

4. Repeat steps from 1 to 3, 10000 times.

The critical consecutive 5 years are the consecutive 5 years with the minimum $\{minSMA_5\}$. The monthly inflow during the critical consecutive 5 years can be used as an inflow scenario for optimal operation of the reservoirs during the dry periods. The observed monthly inflow and the generated during the critical consecutive 5 years (the 5 years with minimum $minSMA_5$) for the Bigge, Henne, Möhne and Sorpe reservoirs are shown in figures 5.17.a, b, c and d respectively.

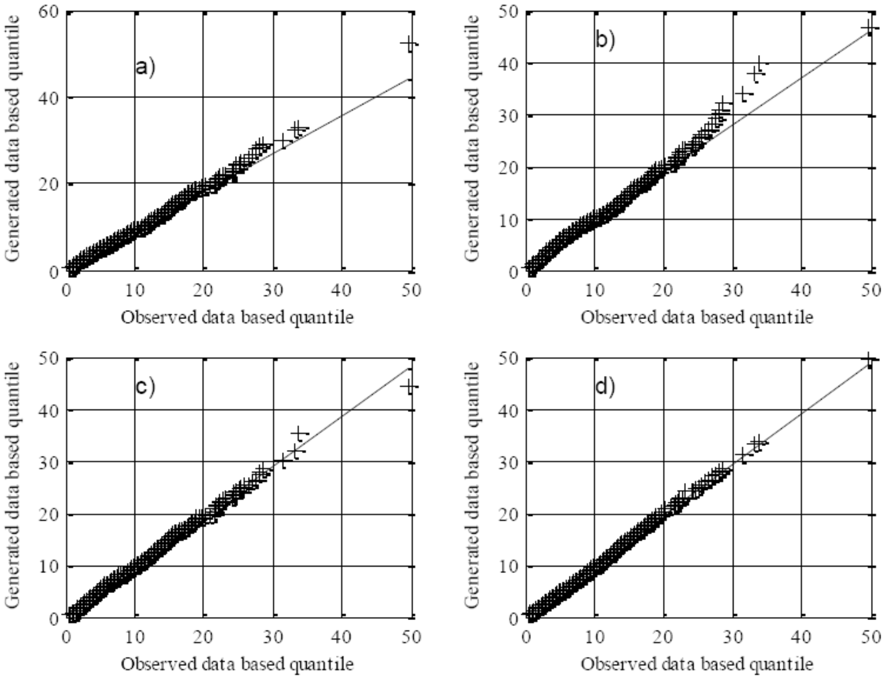


Figure 5.9: *The Q-Q plots for the observed and 300 years generated monthly inflow of the Bigge reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

5.11 Conclusions

T-F, Gamma T-F, MC and PHMM are used to generate 100, 300 and 500 years of monthly inflow into the Bigge, Henne, Möhne and Sorpe reservoirs. The inverse transform method is applied to generate random numbers in the Thomas- Fiering and Monte Carlo models to deal with skewed data however Wilson-Hilferty transformation is proposed to reproduce skewed noises in the Gamma Thomas-Fiering model. The month-to-month correlation in the generated inflow data is preserved in the MC model by applying the Cholesky decomposition method. The statistical parameters (the mean, standard deviation, month-to-month

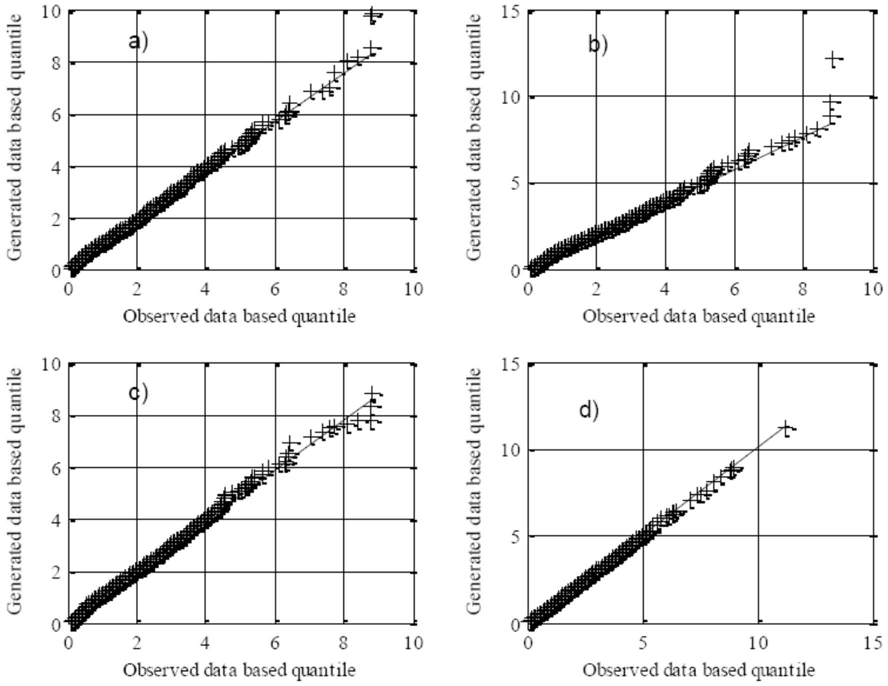


Figure 5.10: *The Q-Q plots for the observed and 300 years generated monthly inflow of the Henne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

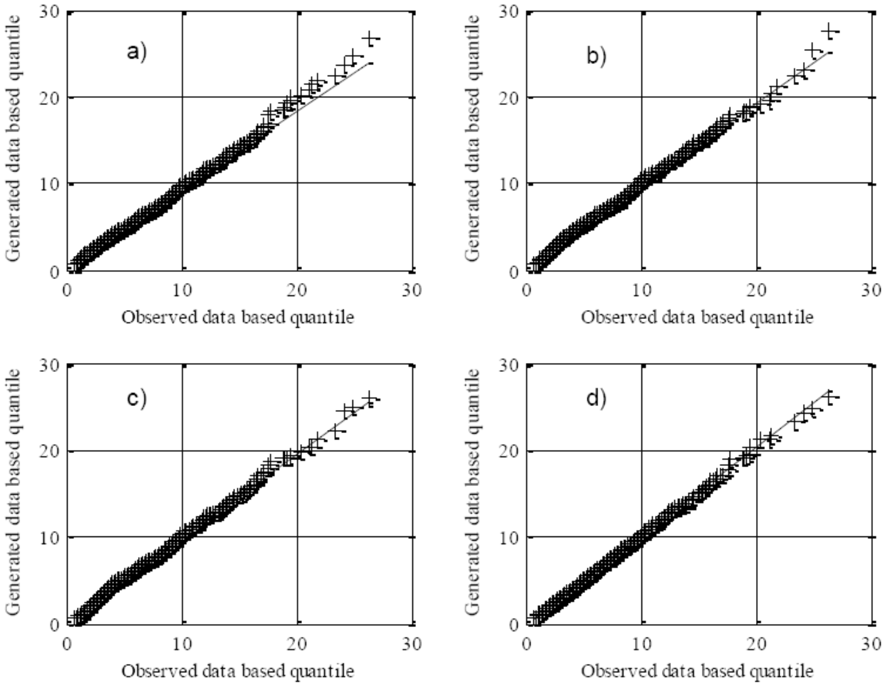


Figure 5.11: *The Q-Q plots for the observed and 300 years generated monthly inflow of the Möhne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

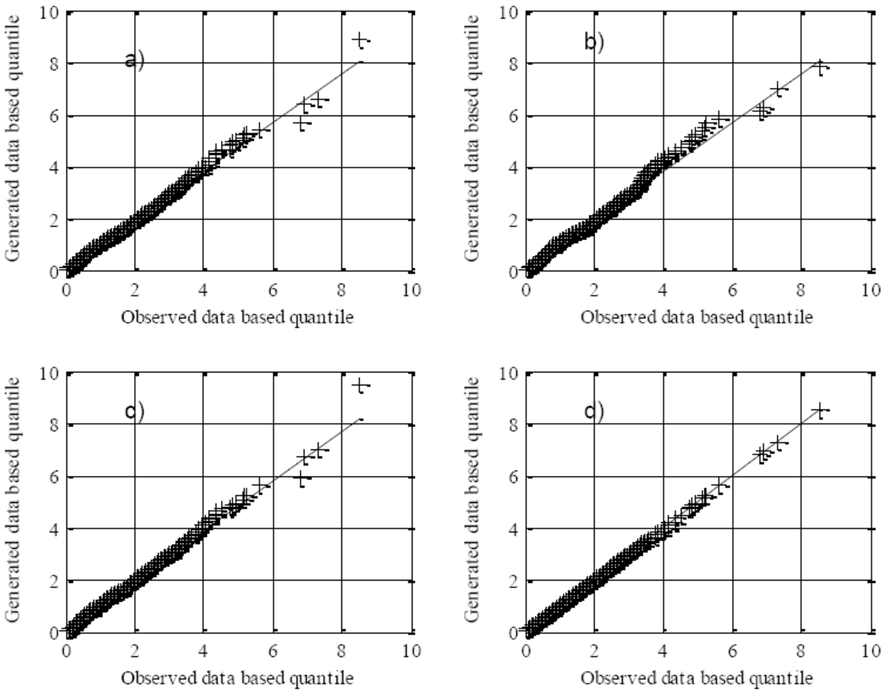


Figure 5.12: *The Q-Q plots for the observed and 300 years generated monthly inflow of the Sorpe reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

correlation and skewness) of the generated monthly inflow time series are compared with those of the observed one and the results of the comparison show that:

1. Using of the inverse transform method to generate random values improves the performance of the T-F model comparing with those of the Gamma T-F.

2. The T-F, MC and PHMM models reproduce the mean of the monthly inflow very well.

3. The MC and PHMM models reproduce the monthly standard deviation better than the other models.

4. The T-F and MC models are superior to the other models in terms of the monthly month-to-month correlation.

5. The PHMM model preserves the monthly skewness very well.

The abilities of the models to reproduce monthly inflow are also evaluated using the Q-Q and survivor function plots. The Q-Q plots show that the MC and PHMM models outperform the other models especially for generating high inflow events. The survivor function plots for the generated inflow by the PMHH are very close to those of the observed

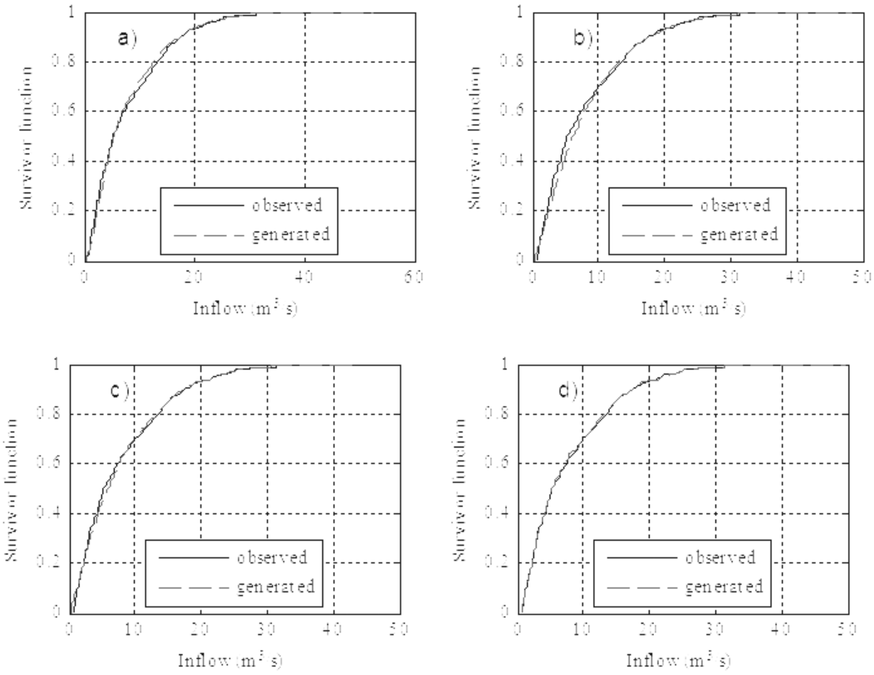


Figure 5.13: *The survivor function plots for the observed and 300 years generated monthly inflow of the Bigge reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

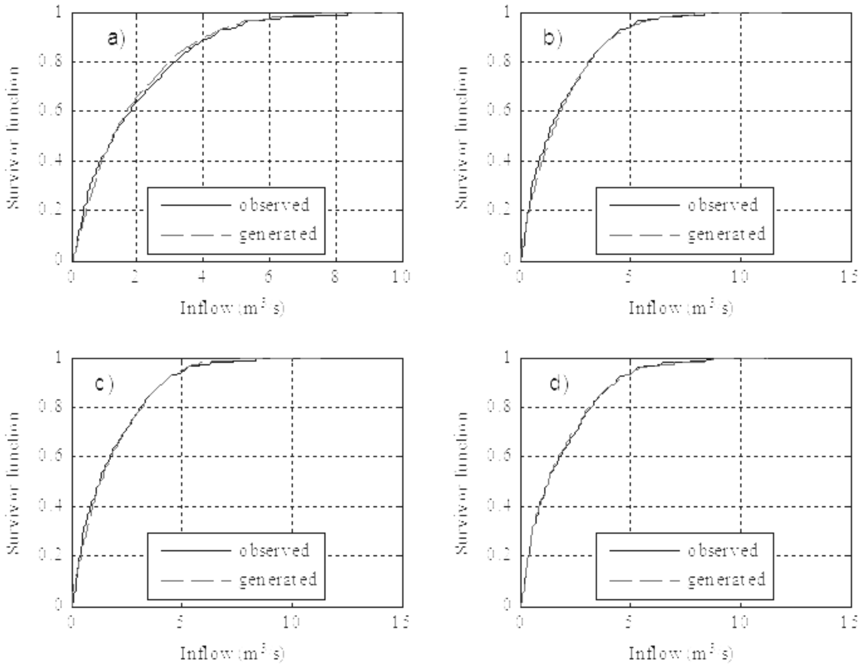


Figure 5.14: *The survivor function plots for the observed and 300 years generated monthly inflow of the Henne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

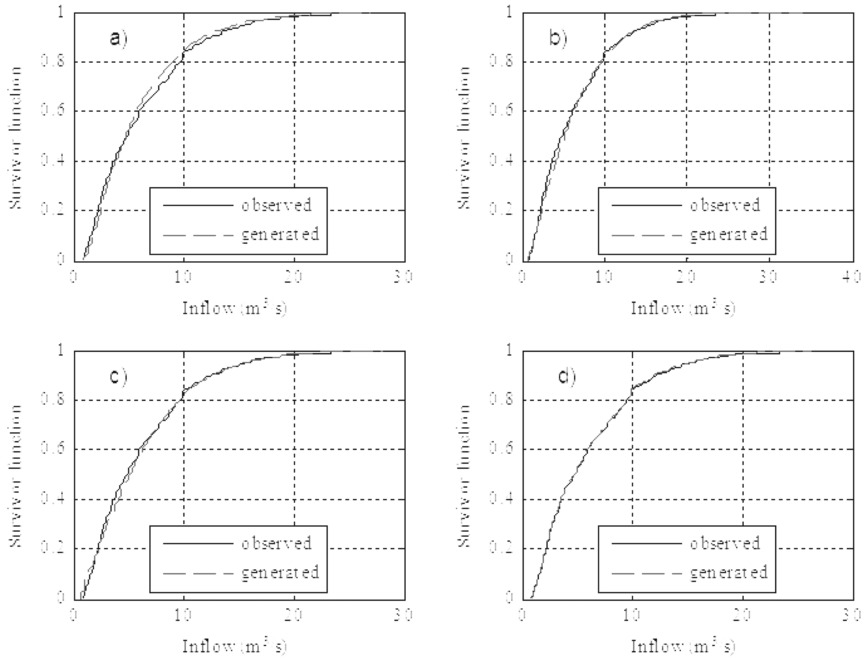


Figure 5.15: *The survivor function plots for the observed and 300 years generated monthly inflow of the Möhne reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

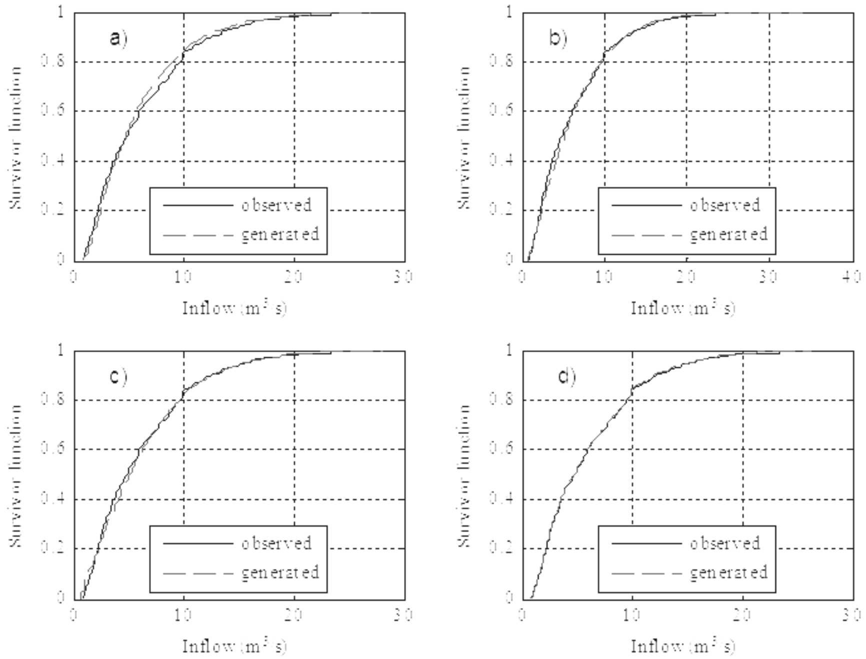


Figure 5.16: *The survivor function plots for the observed and 300 years generated monthly inflow of the Sorpe reservoir using the a) T-F, b) Gamma T-F, c) MC and d) PHMM models*

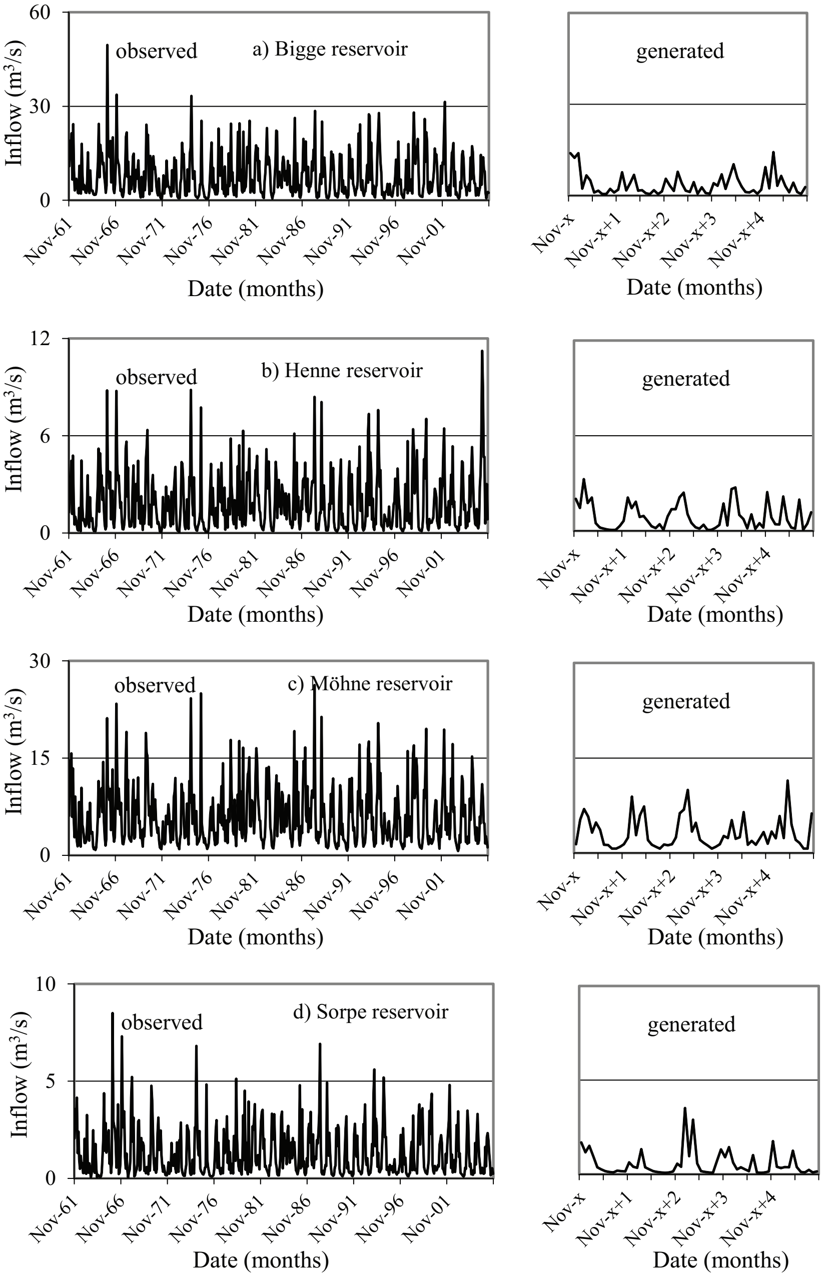


Figure 5.17: *The observed monthly inflow and the generated during the critical consecutive 5 years*

one. We assumed that the MC model is better than the other models (the T-F, Gamma T-F and PHMM) in monthly inflow generation. A procedure is developed to detect the consecutive 5 years that have minimum total inflow. The generated monthly inflow during the critical consecutive 5 years can be used as an inflow scenario for optimal operation of the reservoirs during the dry periods.

5.12 References

Alhassoun, S., Sendil, U., Al-Othman, A.A., Negm, A.M., 1997. Stochastic generation on annual and monthly evaporation in Saudi Arabia. Candian Water Resources Journal, 22(2), 141-154.

Bellone, E., Hughes, J.P., Guttorp, P.A., 2000. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Climate Research, 15, 1-12.

Betrò, B., Bodini, A., Cossu, Q. A., 2008. Using a hidden Markov model to analyse extreme rainfall events in Central-East Sardinia. Environmetrics, 19(7), 702-713.

Celeste, A.B, Suzuki, K., Kadota, A., Camilo, A.S., 2004. Stochastic generation of inflow scenarios to be used by optimal reservoir operation models. Annual Journal of Hydraulic Engineering, 48, 451-456.

Charles, S., Bates, B., Hughes, J., 1999. A spatio-temporal model for downscaling precipitation occurrence and amounts. Journal of Geophysical research, 104, 31657-31669

Gentle, J.E., 2003. Random number generation and Monte Carlo methods. 2nd edition, Springer, New York.

Hisashi, T., 2004. Computational methods in statistics and econometrics. Marcel Dekker, New York.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481.

Kim, B.S., Kim, H.S., Seoh, B.H., 2004. Streamflow simulation and skewness preservation based on the bootstrapped stochastic models. Stochastic Environmental Research and Risk Assessment, 18, 386-400.

Maidment, D.R., 1993. Handbook of hydrology. McGraw-Hill, New York.

Matlab.b, 2008. Statistical toolbox, User′s guide. The MathWorks, Inc.

Mendes, A.G, et al., 2007. Generation of multivariate monthly synthetic water streamflows through multiplicative PARMA model. Universities Power Engineering Conference, 42nd International, 539-544.

Ochoa-Rivera, J.C., García-Bartual, R. andreu, J., 2002. Multivariate synthetic streamflow generation using a hybrid model based on artificial neural networks. Hydrology and Earth System Sciences, 6(4), 641-654.

Phien, H.N., Khan, M.A., 1981. Comparison of two autoregressive models for monthly stream flow generation. Journal of the American Water Resources Association, 17(6), 1035-1041.

Phien, H.N., Ruksasilp, W., 1981. A review of single-site models for monthly streamflow generation. Journal of Hydrology, 52, 1-12.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

Raman, H., Sunilkumar, N., 1995. Multivariate modeling of water resources time series using artificial neural networks. Hydrological Sciences Journal, 40(2), 145-163.

Sarma, A.K., Ahmed, J.A., 2002. Stochastic model for synthetic streamflow generation. Paper presented in the international conference recent trends in probability and statistics: Theory and applications.

Sarma, A.K., Ahmed, J.A., 2004. Stochastic model for synthetic streamflow generation. Statistical thinking: Method and application, Capital Publishing Co., 253-258

Tableman, M., Kim, J.S., 2004. Survival analysis using S: Analysis of time-to-event data. Chapman & Hall/CRC.

Thomas, H.A., Burden, R.P., 1963. Operations research in water quality management. Harvard Water Resources Group.

Thomas, H.A., Fiering, M.B., 1962. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In: Maass et al. Design of water resources systems. Harvard University Press, Cambridge, Mass.

Wilson, E.B., Hilferty, M.M., 1931. The distribution of chi-square. Proceedings of the National Academy of Sciences of the United States of America, 17, 684-688.

Zucchini, W., Guttorp, P., 1991. A hidden Markov model for space-time precipitation. Water Resources Research 27, 1917-1923.

Chapter 6

Prediction of the travel time of the reservoirs' releases along the Ruhr and the Lenne

6.1 Introduction

Travel time estimation of the reservoirs' releases is very important to aid the users in approximating the time that water may become available to them. In this chapter, historical flow data (15 minute time series) are used to estimate the travel time of the released flow from the Bigge, Sorpe, Möhne and Henne reservoirs to some downstream gauges. This procedure was first used by Budach (1993) to determine the travel time of the reservoirs' releases along the Ruhr and Lenne.

The estimated travel time values and the corresponding discharges at the release and downstream gauges are used to build the nonlinear regression (NLR), multiple linear regression (MLR), ANFIS and BPNN models. The models are applied to predict the travel time of the reservoirs' releases along the Ruhr and Lenne Rivers. The performances of the ANFIS, BPNN and MLR models are compared to select the best one. Also, the HEC-RAS model is employed to predict the travel time along the upper and middle reaches of the Ruhr.

Fluorescent dye was used to determine the time of travel time along the Ruhr (Morgenschweis and Nusch, 1990). Jobson (1997) used soluble tracer method to predict travel time and dispersion in rivers. He concluded that the travel time of the leading edge averages 89 % of the travel time of the peak concentration. Hydraulic characteristics can also be used to predict travel time. A study that provides guidance on extrapolating travel-time information using wave speed and hydraulic characteristics from one within bank discharge to another was done by Jobson (2001). Samuels et al. (2001) developed a GIS-based tool

Table 6.1: *Data of the main reservoirs in the Ruhr River basin*

Reservoir	Release gauge	Downstream gauge	Case study	Time series

Bigge	Ahausen	Altena
		Hagen-Hohenlimburg
		Rönkhausen
			case 1
			case 2
			case 3
				1992 - 2007

Sorpe	Langscheid	Bachum	case 4	1992 - 2007
Möhne	Günne	Bachum	case 5	1993 - 2007

Henne	Meschede-Henne	Oeventrop
		Meschede-Rhur
			case 6
			case 7
				1992 - 2007

to calculate the time of travel (based on real-time stream flow measurements), decay and dispersion of a pollutant introduced into surface water. Abida et al. (2005) routed the released flows from the Sidi Salem Dam Reservoir on the Medjerda River, Tunisia, using hydrologic and hydraulic flood routing techniques. They used the hydrologic flood routing method of Muskingum and numerical model for hydraulic flood routing. The numerical model was based on the complete numerical solution of Saint-Venant equations.

Different artificial intelligence methods have been applied successfully in many water resources problems. Some of these works are introduced in chapter 3.

6.2 Travel time estimation

A Matlab graphical user interface (Fliesszeit GUI) is created to determine the jump points (jump point means increase in the reservoir release) at the release gauge for each reservoir using the 15 minutes historical flow time series $\{X\}$. The times of travel from release gauge to the gauges downstream it are estimated for each determined jump point (if possible).

The main reservoirs, discharge stations and the studied reaches (7 case studies) in the Ruhr River basin are shown in figure 6.1. Table 6.1 gives a list of the release gauges and the corresponding downstream gauges for the Bigge, Henne, Möhne and Sorpe reservoirs. The table lists also the length of the time series which are used to estimate the travel time along each reach. The first and most important step in estimating the time of travel is to detect the jump points at the release gauge (T_{Ri}) and the corresponding jump points at the downstream gauge (T_{Di}), i = 1, 2, ..., n where n is the number of the total observed jump points. Figure 6.2 shows the different variables at each upstream jump point T_{Ri} and at the corresponding downstream one T_{Di}. The variables Q_R and Q_D (see figure 6.2) can be defined as:

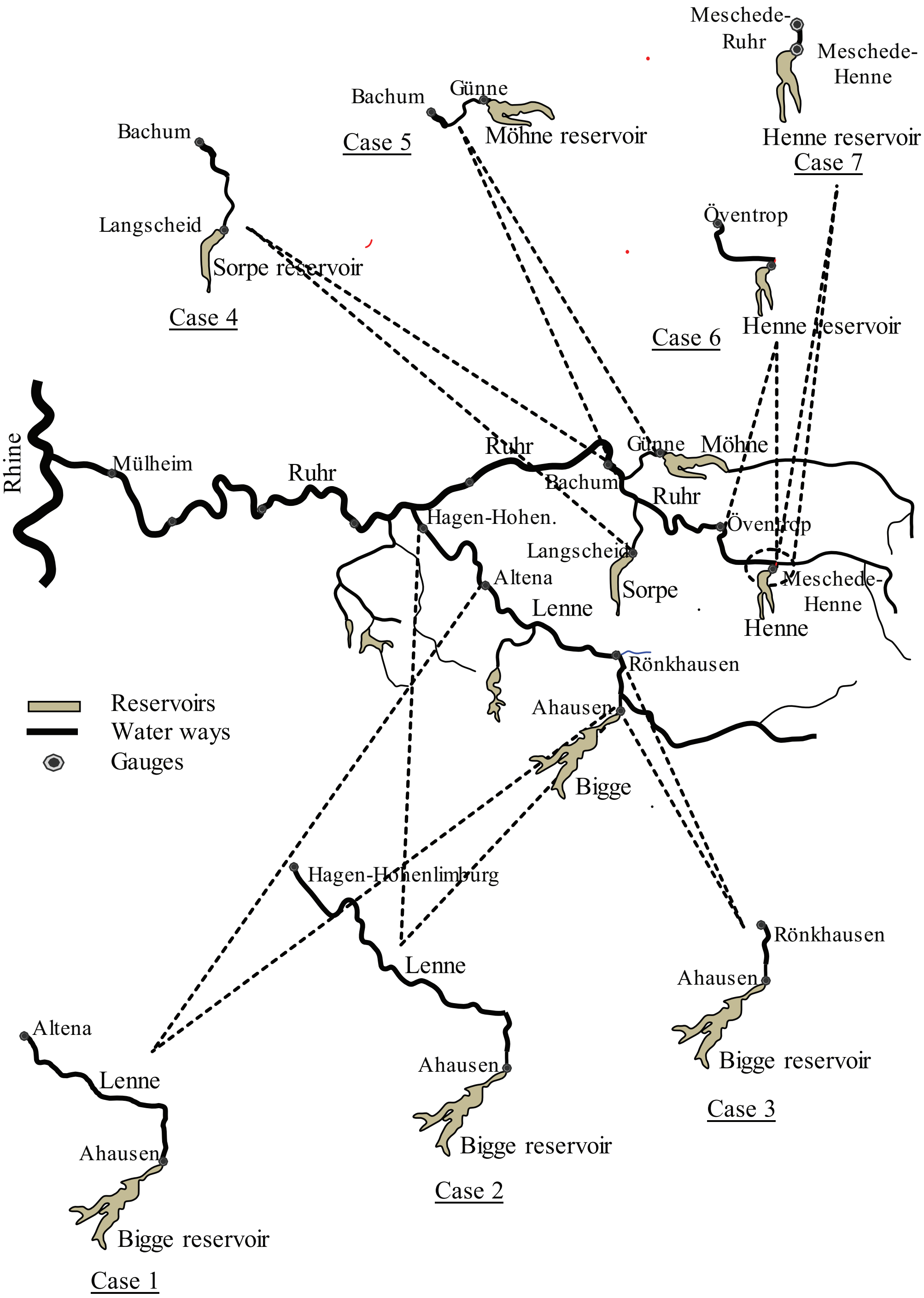


Figure 6.1: *List of the release and downstream gauges*

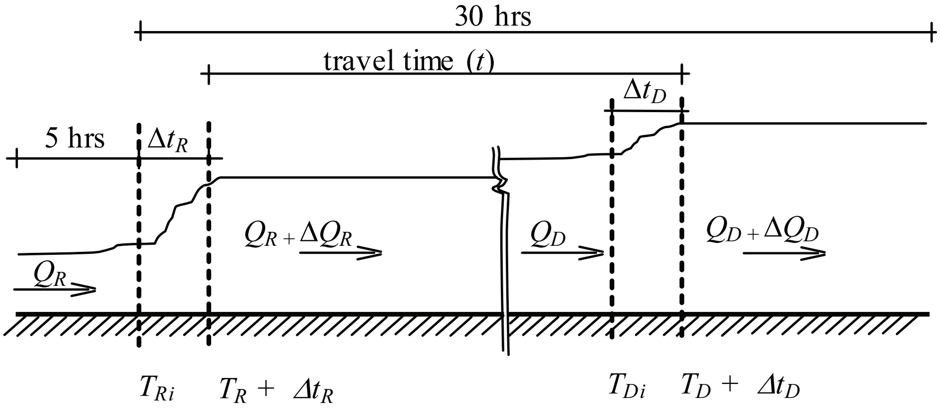


Figure 6.2: *Determination of the jump points TRi at the release gauge and TRi at the downstream gauge*

- Q_R in m^3/s is the reservoir release just before T_{Ri} and Δt_R is the time which the flow takes to increase from Q_R to $Q_R + \Delta Q_R$.

- Q_D in m^3/s is the discharge at the downstream gauge just before T_{Di} and Δt_D is the time which the flow takes to increase from Q_D to $Q_D + \Delta Q_D$.

After determination of the jump points T_{Ri} and the corresponding T_{Di}, a time series (t_{s140}) with 20 observations before and 120 observations after the jump point T_{Ri} respectively is extracted from $\{X\}$ for each jump point. The length of this series (t_{s140}) is checked to cover an interval of 35 hrs which is more than the maximum observed Δt_R+ t + Δt_D (see figure 6.2). Second, the hydrographs at the release gauge and at the corresponding downstream gauges are plotted for each jump point using the corresponding t_{s140}. The plotted hydrographs are used to estimate the travel time corresponding to each jump point. Figures 6.3, 6.4, 6.5 and 6.6 show how to estimate travel time from the hydrographs. The values of the estimated travel time and the corresponding downstream flow for each case study are divided into classes as given in table 6.2 which can be used to guess travel time value by knowing Q_D.

6.3 Nonlinear regression analysis

The correlation coefficients between the travel time (t) and Q_R, ΔQ_R, Q_D and ΔQ_D for all case studies are estimated and the results are given in table 6.3. Values of the correlation coefficient between the estimated travel time t and Q_R, ΔQ_R, Q_D and ΔQ_D for case study

2 (see table 6.1) are plotted in figure 6.7. According to table 6.3 it is obvious that the correlation between travel time (t) and downstream flow (Q_D) have the maximum values for case studies 1 through 6.

The NLR models are built to detect the relation between travel time (t) from the reservoirs to the downstream gauges and the flow at the downstream gauge (Q_D). The results of the nonlinear regression analysis are shown in figure 6.8. The nonlinear regression equation corresponds to each case study is also presented. This figure can be used to predict the expected travel time (t) by knowing the downstream discharge (Q_D).

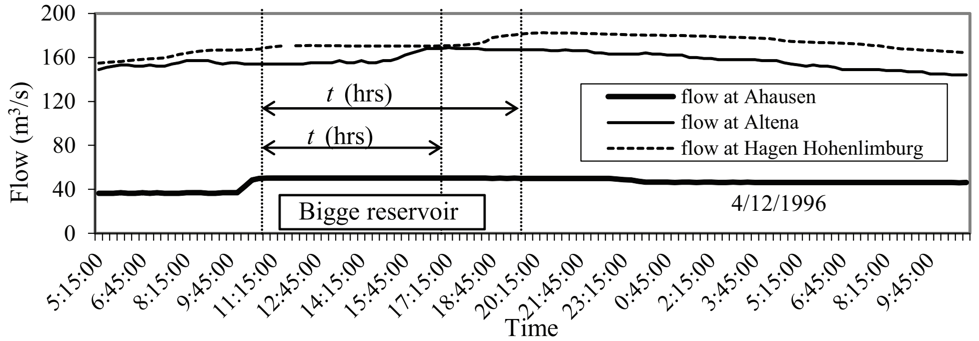


Figure 6.3: *Hydrographs at Ahausen, Altena and Hagen-Hohenlimburg (4/12/1996)*

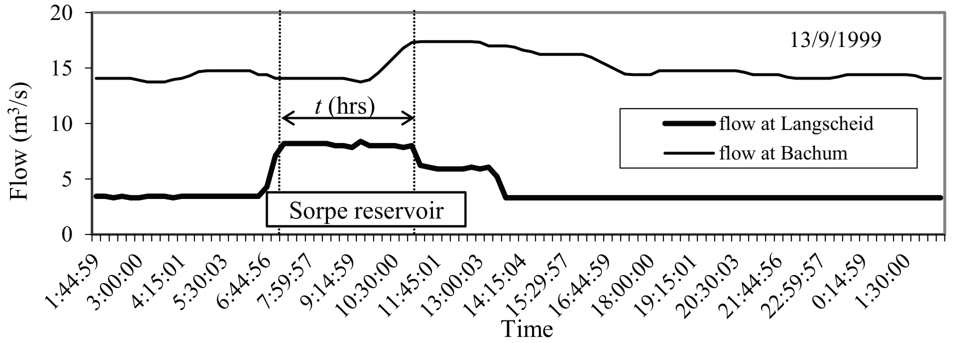


Figure 6.4: *Hydrographs at Langscheid and Bachum (13/9/1999)*

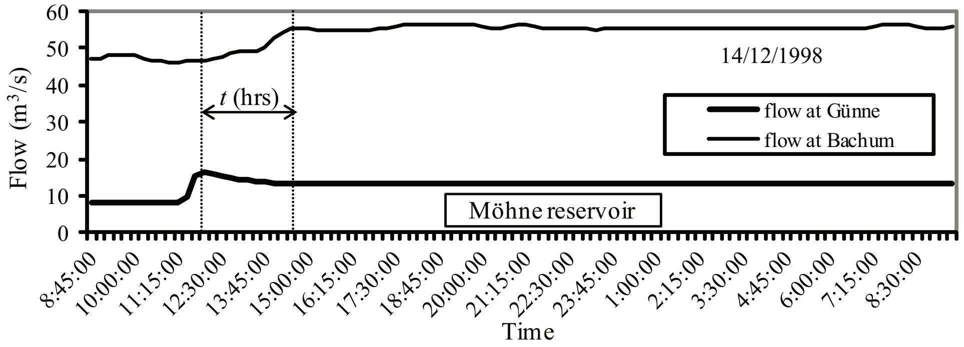


Figure 6.5: *Hydrographs at Günne and Bachum (14/12/1998)*

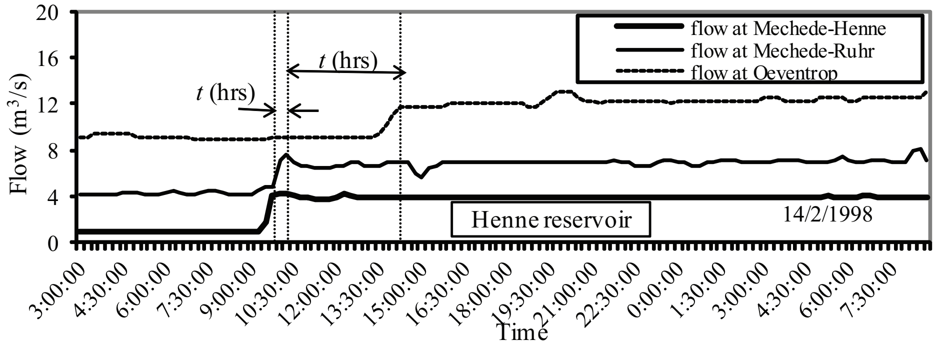


Figure 6.6: *Hydrographs at Meschede-Henne, Meschede-Ruhr and Oeventrop (14/2/1998)*

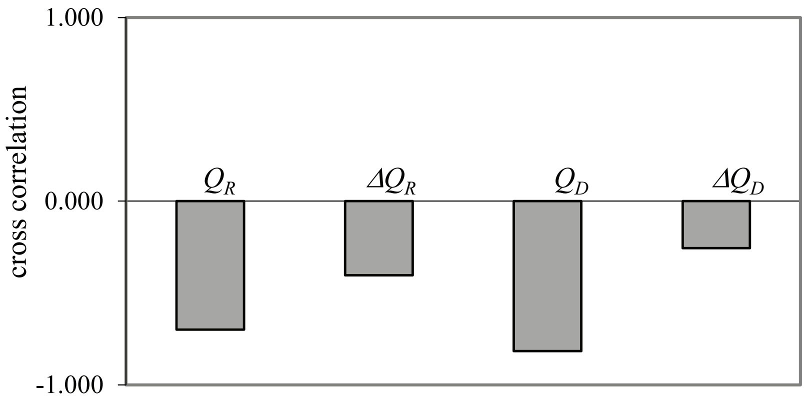


Figure 6.7: *Cross correlation between travel time t and ΔQ_R, Q_D and ΔQ_D (case study 2)*

Table 6.2: *Expected travel time values for different values of the downstream discharges (Q_D)*

Reservoir	Case study	Q_D (m^3/s)	t (hrs)	
			Range	Average
Bigge	case 1	< 20	8.0 -11.0	(10)
		20 - 40	6.75 - 9.75	(9)
		40 - 60	6.25 - 9.25	(7.75)
		60 - 80	5.75 - 9.00	(7.25)
		80 - 100	5.25 - 8.50	(6.5)
		100 - 160	4.75 - 7.75	(6.25)
		> 160	4.0 -7.50	(5.5)
	case 2	< 20	13.0-17.0	(15.5)
		20 - 40	10.5-14.25	(12.5)
		40 - 60	9.25-11.0	(11.25)
		60 - 80	8.50-12	(10)
		80 - 100	8-11.5	(9.5)
		100 - 140	7.50-11	(8.75)
		> 140	6.5-10	(8)
	case 3	< 20	2.5-3.5	(2.75)
		20 -30	3-Feb	(2.5)
		30 - 50	1.75-2.75	(2.25)
		50 - 80	1.5-2.5	(2)
		> 80	1.25-2.25	(1.75)
Sorpe	case 4	< 20	3-4.75	(4)
		20 - 30	2.25-4	(3.25)
		30 - 50	2.-3.75	(2.75)
		50 - 70	1.5-3.5	(2.5)
		> 70	1.25-3.25	(2.25)
Möhne	case 5	< 20	3.75-6.25	(6.25)
		20 - 40	2.75-5	(3.75)
		40 - 60	2.25-4.50	(3.5)
		60 - 80	1.75-4.25	(3.25)
		80 - 100	1.75-4	(3)
		100 - 160	1.25-3.5	(2.5)
		> 160	1-3.25	(2)
Henne	case 6	< 10	4..25 - 6.75	(5.5)
		10 - 20	3.75 - 6.0	(4.5)
		20 - 30	3.0 - 5.25	(4.25)
		30 - 40	2.75 - 5.0	(4)
		60 - 70	2.0 - 4.5	(3)
		> 80	1.0 - 4.0	(2.75)
	case 7	t = (0.25 - 1.0) hour		

Table 6.3: *Cross correlation between the travel time (t) and Q_R, ΔQ_R, Q_D and ΔQ_D*

Case study	Flow reach	Q_R	ΔQ_R	Q_D	Q_D
case 1	flow from gauge Ahausen to gauge Altena	-0.699	-0.322	-0.757	-0.319
case 2	flow from gauge Ahausen to gauge Hagen-Hohenlimburg	-0.7	-0.403	-0.817	-0.257
case 3	flow from gauge Ahausen to gauge Rönkhausen	-0.58	-0.041	-0.638	0.102
case 4	flow from gauge Langscheid to gauge Bachum	0.11	0.036	-0.819	0.356
case 5	flow from gauge Günne to gauge Bachum	-0.612	-0.278	-0.65	-0.239
case 6	flow from gauge Meschede-Henne to gauge Oeventrop	-0.113	0.062	-0.837	0.409
case 7	flow from gauge Meschede-Henne to gauge Meschede-Ruhr	-0.039	-0.07	-0.046	-0.046

6.4 Travel time estimation using multivariate models

The observed travel time values from gauge Ahausen to gauge Hagen-Hohenlimburg (case study 2) and the corresponding Q_R, ΔQ_R and Q_D are simulated using three multivariate models. These models are:

1. The adaptive neuro-fuzzy inference system (ANFIS).
2. The backpropagation neural network (BPNN).
3. The multiple linear regression (MLR).

A real challenge for using these models is the limited amount of the estimated travel time data. In the next subsection, we will discuss how to improve the generalization ability of these models in case of limited amount of data.

6.4.1 Model generalization

In chapter 3, we used an early stopping procedure to prevent overfitting and to improve the generalization ability of the BPNN and ANFIS models. To use an early stopping procedure, we need enough data which is not available in the estimated travel time data. Statisticians developed several techniques for dealing with limited amounts of data. These techniques can be also used by the neural networks designers to choose the neural network architecture, the number of hidden neurons, select input variables, training parameters (Priddy and Keller, 2005). The most common techniques are the cross-validation methods (the k-fold and leave-one-out cross validation methods) and the jackknife and bootstrap resampling.

The cross-validation is a technique for estimating the generalization error to improve the generalization ability of the models. It is the most frequently used method for small number of observations (Hu and Hwang, 2002). In this method, the available data is to be divided into two subsets: the training set and the validation or test set. We used the k-fold cross validation (KFCV) and leave-one-out cross validation (LOOCV) to select the best ANFIS, BPNN and MLR models. We assumed the model with the minimum average relative error percentage ($AREP$) to be the best model.

k-fold cross-validation
The algorithm for the k-fold cross-validation is as follows (Good, 2006):

1. Split the original data into k subsets of equal size.

2. Use a single subset for testing the model (validation data) and the remaining k -1 subsets to train the model (training data).

3. Repeat the cross-validation process k times to ensure that each of the k subsets is used exactly once as the validation data.

4. Take the average of the k results from the folds to produce a single estimation for the $AREP$.

Leave one-out cross-validation
The Leave-one-out cross-validation (LOOCV) involves using a single observation from the original sample as the validation data and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

6.4.2 Adaptive neuro-fuzzy inference system (ANFIS)

ANFIS is discussed in details in chapter 3. Ten ANFIS models are trained, five of them have ΔQ_R and Q_D as input variables and the other models have Q_R, ΔQ_R and Q_D as input variables (see table 6.4). Table 6.4 gives the number of membership functions for each input variable. Figure 6.9 shows ANFIS architecture of the input–output system for the ANFIS model, A5.

The ANFIS models are trained using two input-output data sets as follows:

data1: the determined travel time data are arranged according to date.
data2: the determined travel time data are arranged randomly.

As mentioned before, we used the KFCV and LOOCV methods to estimate the generalization error of the ANFIS, BPNN and MLR models. Three values of k are used ($k = 4$, 8 and 16) in the KFCV and the average $AREP$ values correspond to each k are estimated. The average values of the estimated $AREP$ using the KFCV method for data1 and data2 are plotted in figures 6.10 and 6.11 respectively. The average values of $AREP$ for data1 are estimated using the LOOCV method and the results are shown in figure 6.12 (the average values of the $AREP$ are equal in both data1 and data2).

Table 6.4: *The number of membership functions corresponds to Q_R, ΔQ_R and Q_D in the ANFIS models*

	Group A					Group B				
	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5
QR	-	-	-	-	-	2	3	4	5	6
ΔQR	2	2	2	2	2	2	2	2	2	2
QD	7	6	5	4	3	3	3	3	3	3

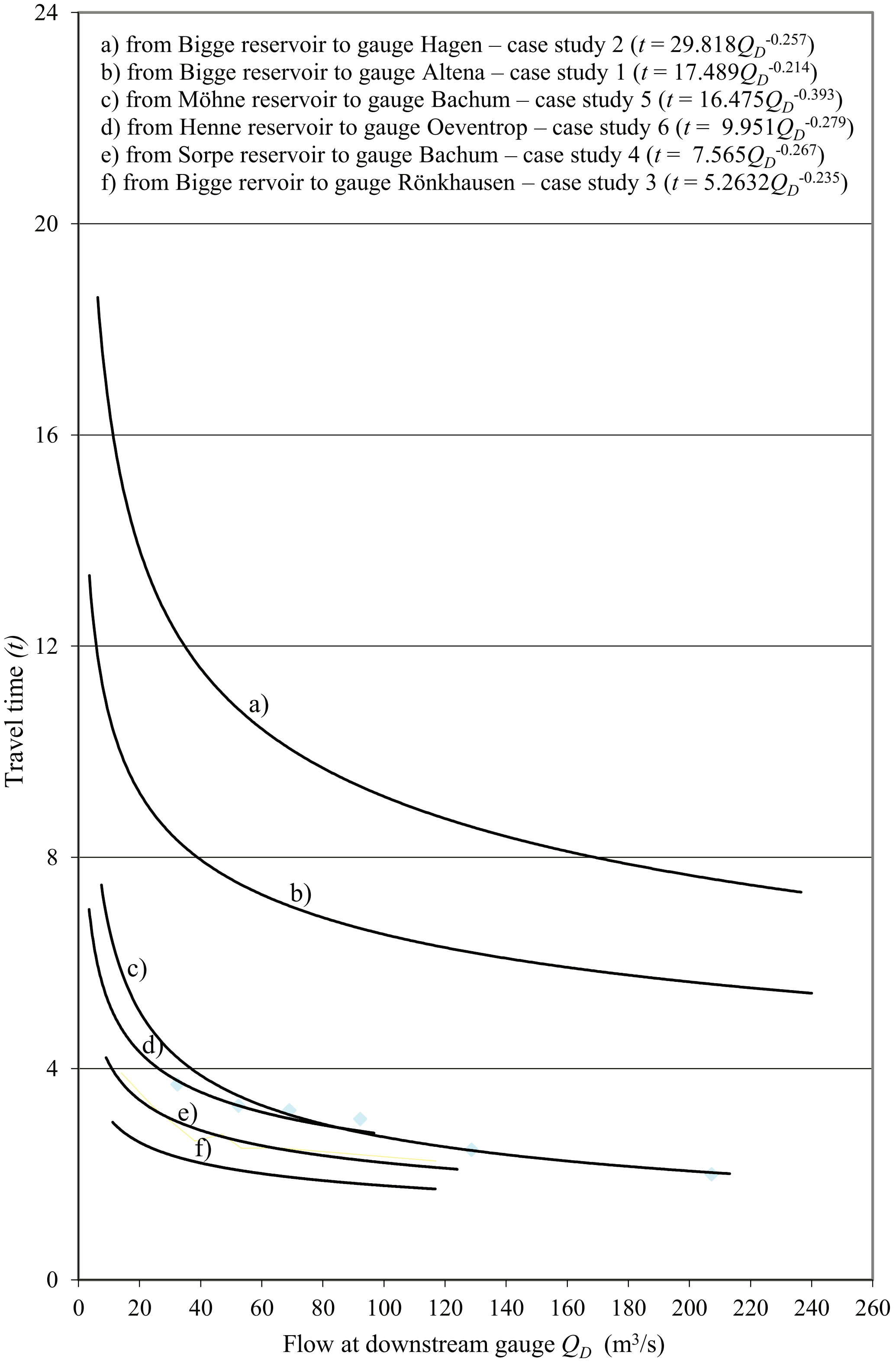


Figure 6.8: *The relation between the travel time (t) and downstream flow (Q_D) for all case studies*

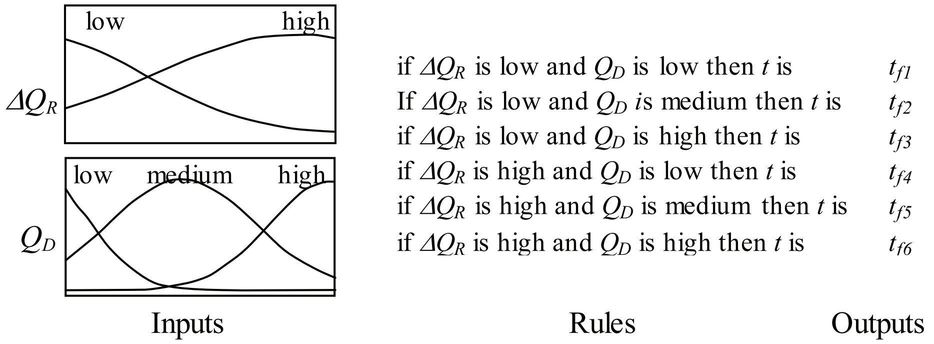

low
high
ΔQ_R
low
medium
high
Q_D
if ΔQ_R is low and Q_D is low then t is
If ΔQ_R is low and Q_D is medium then t is
if ΔQ_R is low and Q_D is high then t is
if ΔQ_R is high and Q_D is low then t is
if ΔQ_R is high and Q_D is medium then t is
if ΔQ_R is high and Q_D is high then t is
t_{f1}
t_{f2}
t_{f3}
t_{f4}
t_{f5}
t_{f6}
Inputs
Rules
Outputs

Figure 6.9: *The ANFIS architecture of the input-output system (model A5)*

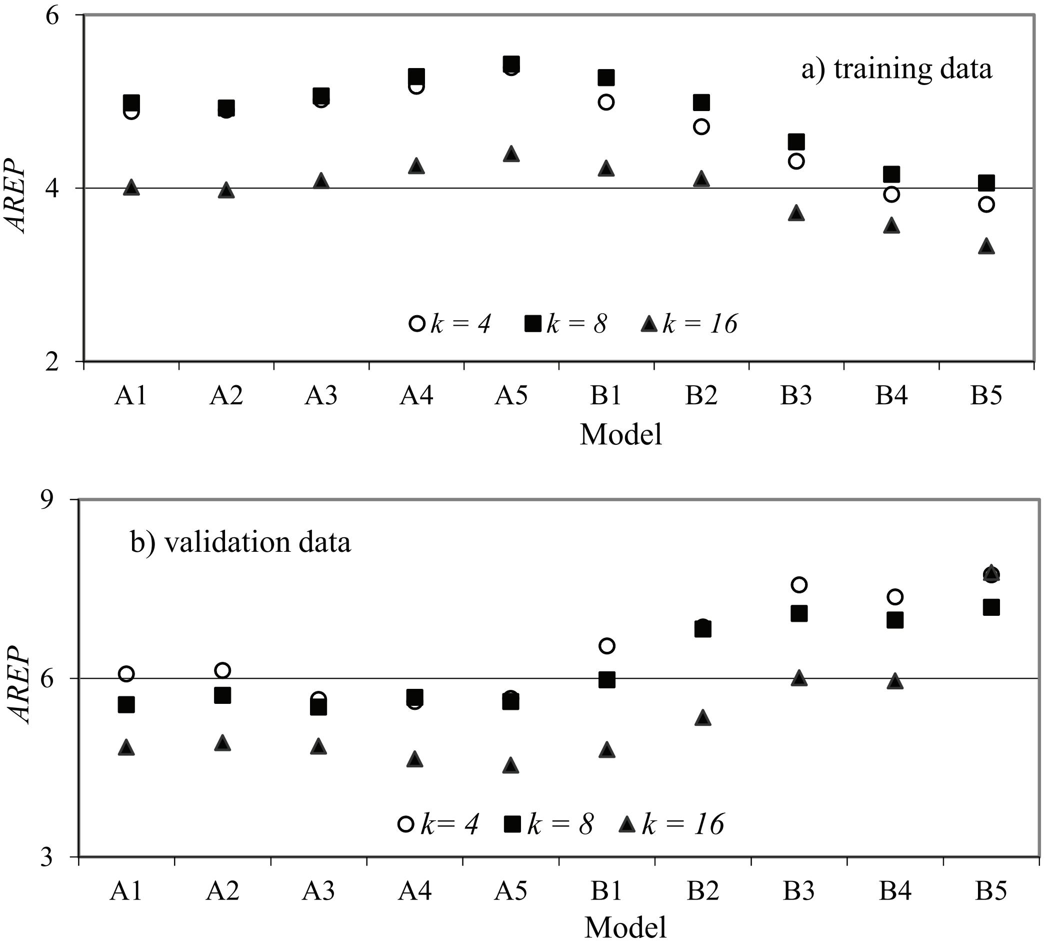


Figure 6.10: *The values of the AREP for the ANFIS models (data1 - KFCV)*

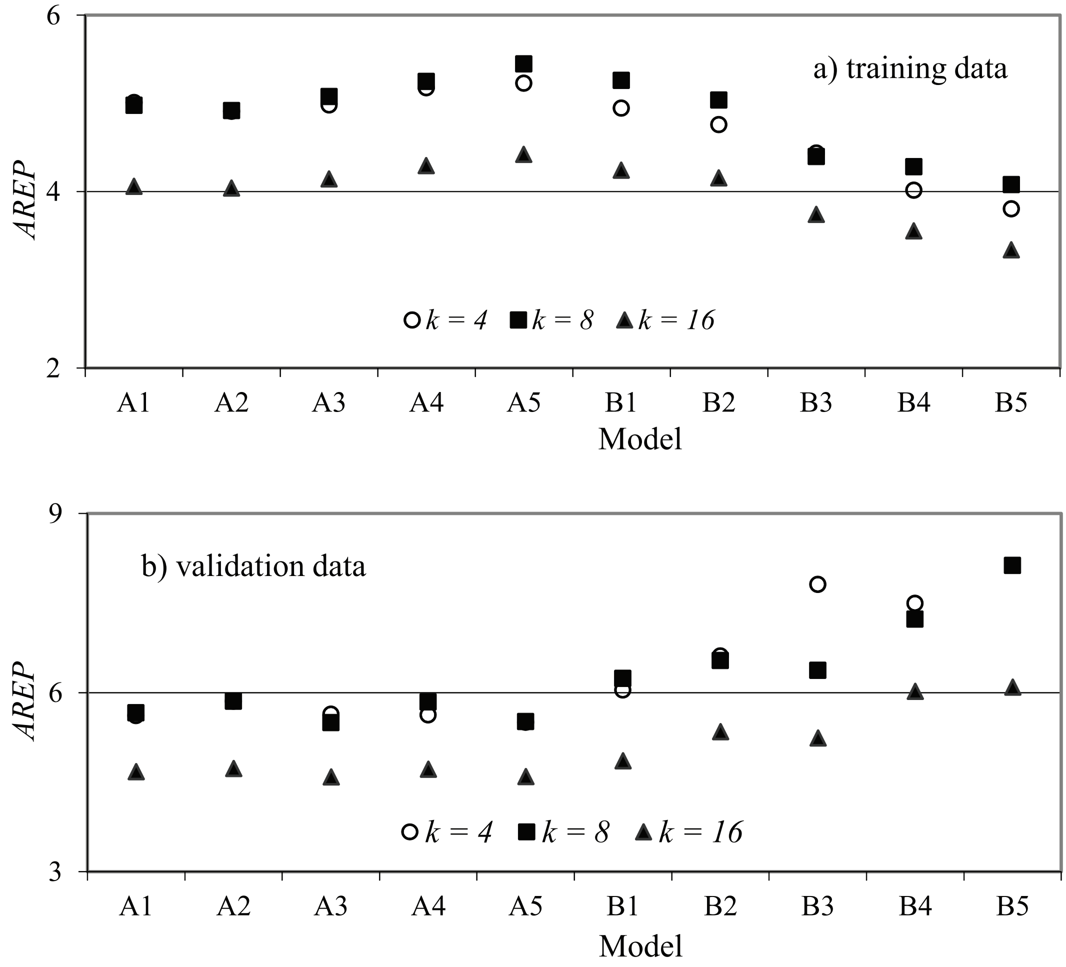


Figure 6.11: *The values of the AREP for the ANFIS models (data2 - KFCV)*

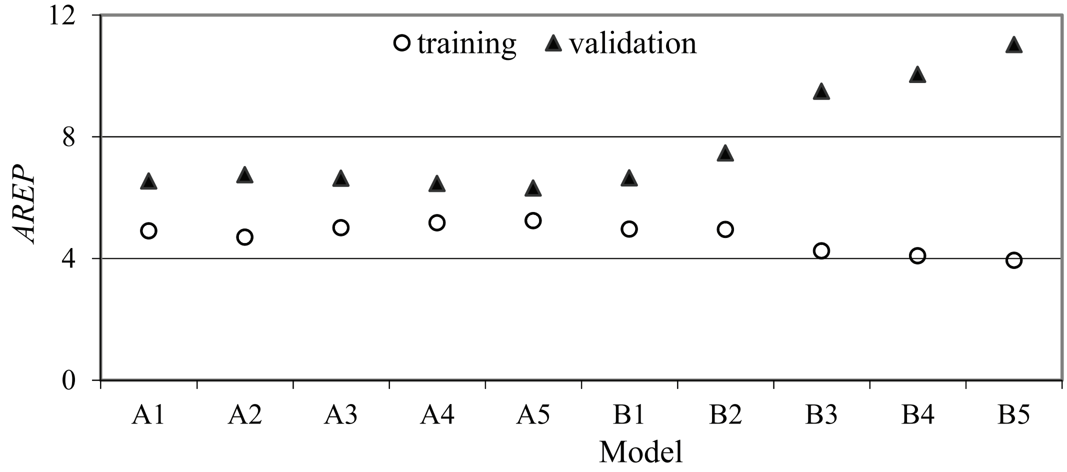


Figure 6.12: *The values of the AREP for the ANFIS models (data1 & data2 - LOOCV)*

Removing of the outliers

In statistics, an outlier is an observation that is numerically distant from the rest of the data. The models which are derived from data sets that include outliers will often be misleading. To know the effect of the outliers on the performances of the models, we removed the outliers and then retrained the ANFIS models. The following procedure is assumed to detect and remove the outliers:

1. The ANFIS model A5 is trained with the total observations (data1) and the predicted travel time values are denoted as $\hat{t}$.

2. The deviation of the estimated travel time ($\hat{t}_i$) from the observed one (t_i) is $\Delta t_i = \left| t_i - \hat{t}_i \right|$, where $i = 1, 2, \ldots, n$ where n is the number of observed travel time values.

3. The average deviation is$\Delta t_{av} = \frac{1}{n} \sum_{i=1}^{n} \left| t_i - \hat{t}_i \right| = 0.51$. The observations which have $\Delta t \geq 2\Delta t_{av}$,are assumed to be outliers and deleted.

The outliers are deleted from data1 and data2 to get data3 and data4 respectively. The ANFIS models (table 6.4) are retrained using the data3 and data4 and the average *AREP* are estimated using the KFCV and LOOCV methods. Figures 6.13 to 6.14 show the average estimated values of the *AREP* for data3 and data4 using the KFCV method. Figure 6.15 shows the average estimated values of the *AREP* for data3 using the LOOCV method (the average values of the *AREP* are equal in both data3 and data4).

Analysis of the results of ANFIS models for travel time prediction

We used the estimated *AREP* in the validation data sets to test the performances of the models. As given in table 6.4, we divided the ANFIS models into two groups according to the number of input variables. The ANFIS models, group A, with two input variables (ΔQ_R and Q_D) and the models, group B, with three input variables (Q_R, ΔQ_R and Q_D).

ANFIS models (group A)

Figures 6.10 through 6.15 show that there is no significant difference in the performances of the models, group A, except for $k = 4$ (the KFCV method). For $k = 4$, the performance of the ANFIS models improves by decreasing the number of membership functions corresponding to Q_D from 7 (model A1) to 2 (model A5). This improvement in the performances is due to the decrease in the total number of parameters (linear and nonlinear parameters) in the ANFIS model with respect to the size of the training data set (Anuradha et al., 2008).

ANFIS models (group B)

Figures 6.10 through 6.15 also show that the ANFIS models, group B, have performances worse than those of the models, group A, because the models, group B, have total number of parameters more than those in the models, group A.

From the above results we recommend A5 to be the best ANFIS model that can be used with limited observations. The ANFIS models, group B, may produce good performances with more observations.

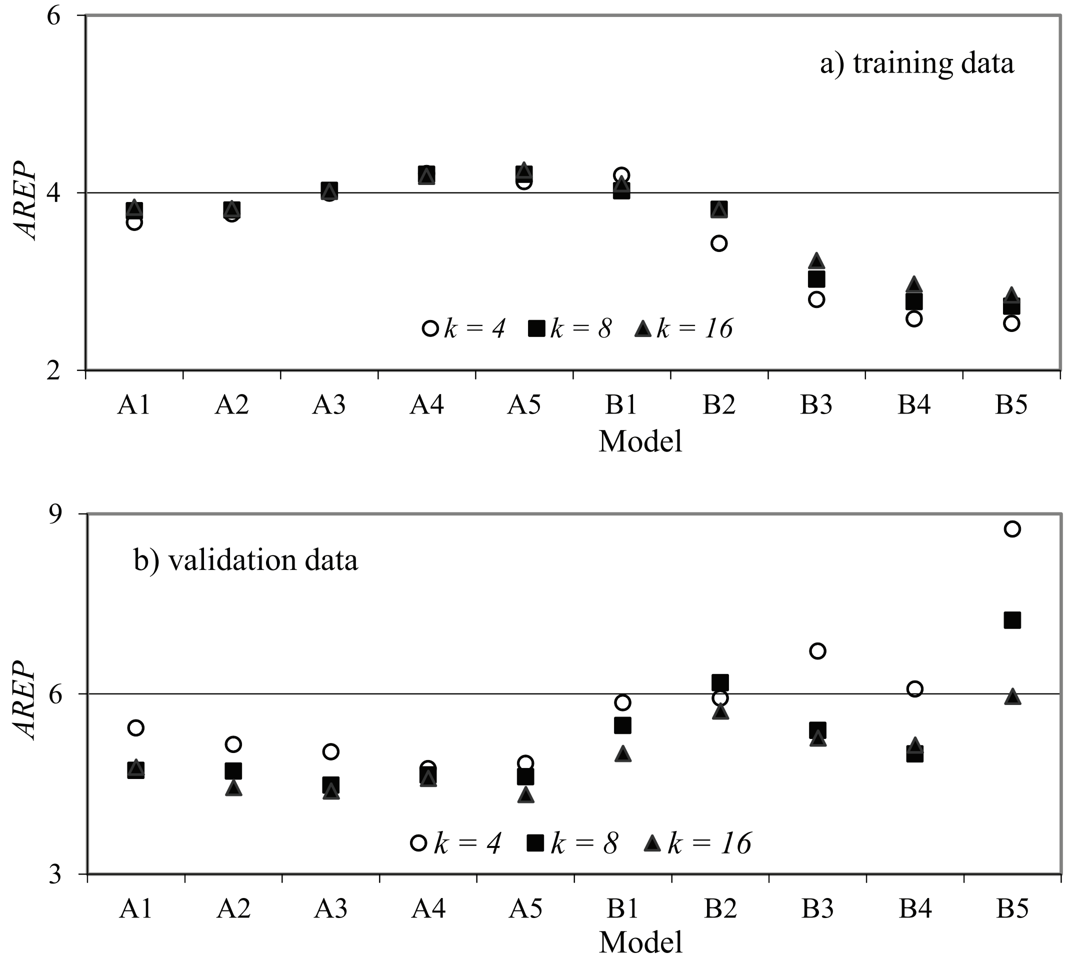


Figure 6.13: *The values of the AREP for the ANFIS models (data3 - KFCV)*

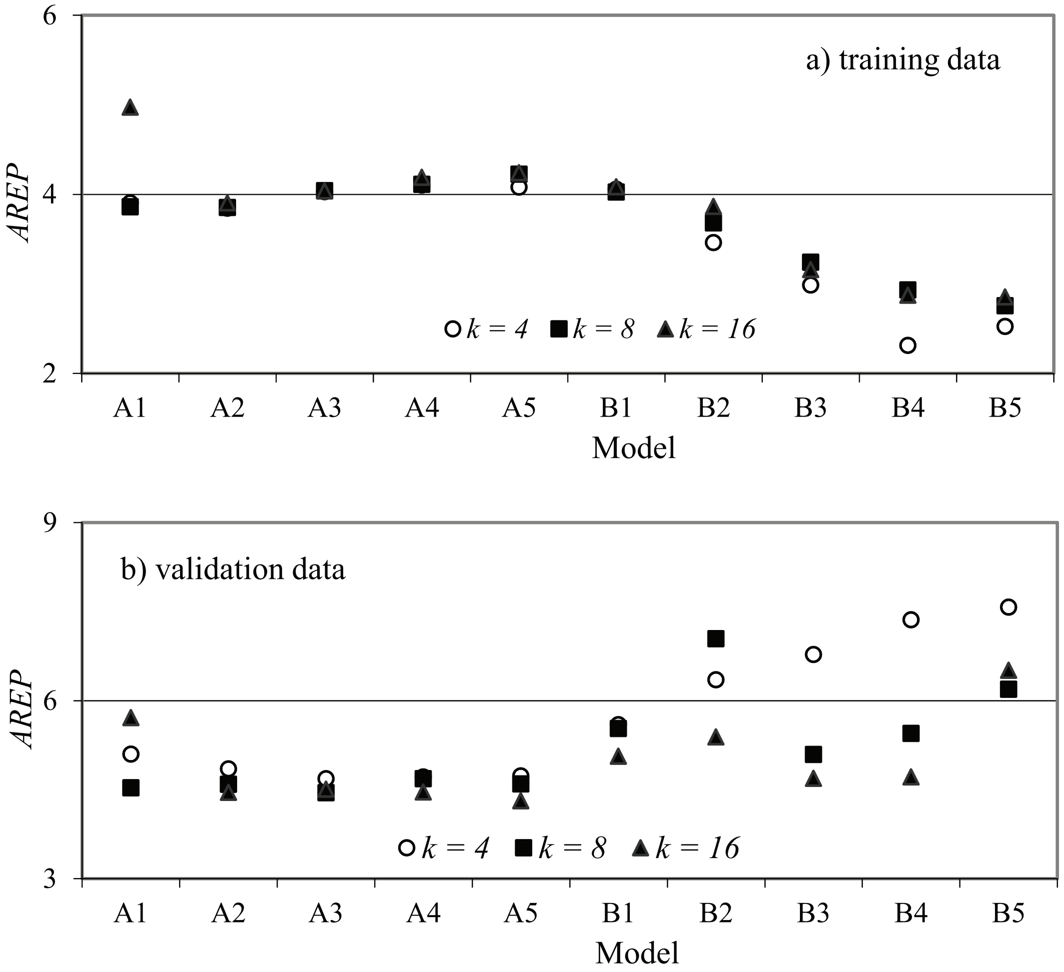


Figure 6.14: *The values of the AREP for the ANFIS models (data4 - KFCV)*

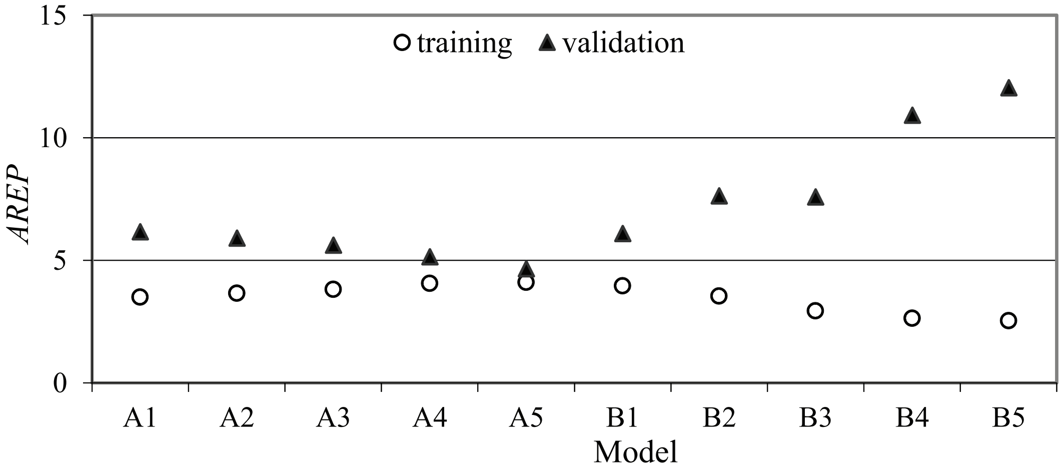


Figure 6.15: *TThe values of the AREP for the ANFIS models (data3 - LOOCV)*

6.4.3 Backpropagation neural network (BPNN)

Two-layer multilayer perceptron (MLP) backpropagation neural network (BPNN) is applied to simulate case study 2 (see chapter 3 for more details about BPNN). In the present study, the BPNN models are trained and simulated using Matlab 7.5 developed by the Math Works Inc, Natick, Massachusetts.

In section 6.4.2, the ANFIS models are trained using different input variables and different number of membership functions. Ten BPNN models which have the same input variables are trained using different number of neurons in the hidden layer. We used the Levenberg-Marquardt as training algorithm and tan-sigmoid and linear functions as activation functions for the hidden layer neurons and for the output one respectively. Table 6.5 lists the input variables and the number of neurons in the hidden layer for each model. The architecture of BPNN model, D5 for travel time forecasting is shown in figure 6.16. The BPNN models are trained using data1, data2, data3 and data4 and the average $AREP$ are estimated. Figures 6.17, 6.18, 6.19 and 6.20 display the estimated average values of the $AREP$ (using the KFCV method) in data1, data2, data3 and data4 respectively. The estimated average values of the $AREP$ in data1 and data3 using the LOOCV method are plotted in figures 6.21 and 6.22 respectively.

Analysis of the results of BPNN models for travel time prediction

We have two groups of the BPNN models as given in table 6.5. The BPNN models, group C, with two input variables (ΔQ_R and Q_D) and BPNN models, group D, with three input variables (Q_R, ΔQ_R and Q_D). From figures 6.17 through 6.22, it is clear that there is no significant difference between the performances of the models, group C and the models, group D, with respect to the values of the estimated $AREP$ in the validation data sets. The figures also show that the performance of the model C5 is slightly better than the performances of the other models.

Table 6.5: *The number of neurons in the hidden layer for the BPNN models*

Model	Group C					Group D				
	C1	C2	C3	C4	C5	D1	D2	D3	D4	D5
Inputs	ΔQ_R and Q_D					Q_R, ΔQ_D and Q_D				
No. of neurons	6	5	4	3	2	2	3	4	5	6

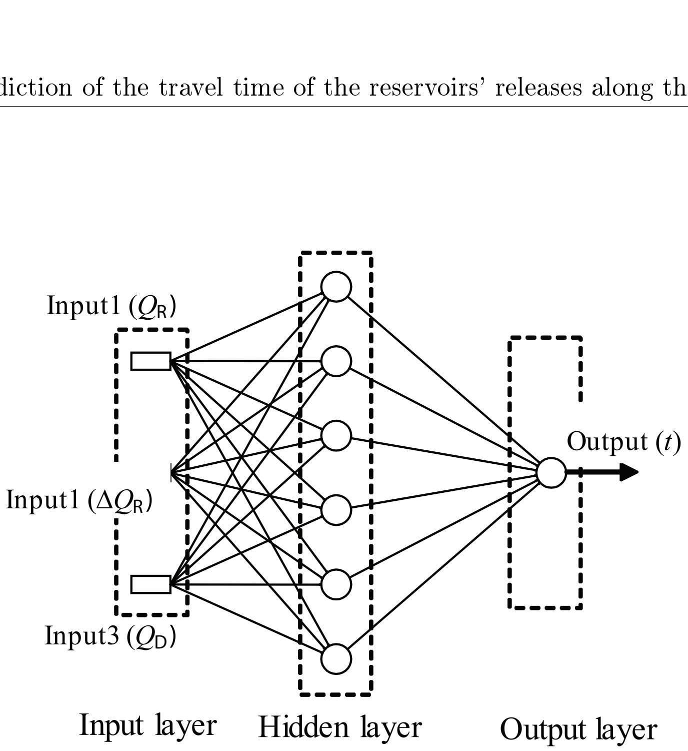


Figure 6.16: *Architecture of the Backpropagation Neural Network (BPNN) for travel time prediction (model, D5)*

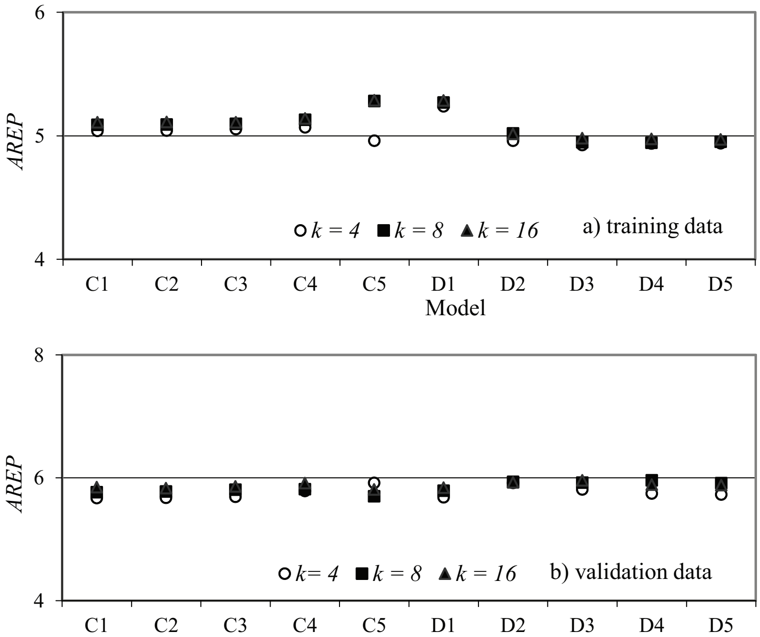


Figure 6.17: *The values of the AREP for the BPNN models (data1 - KFCV)*

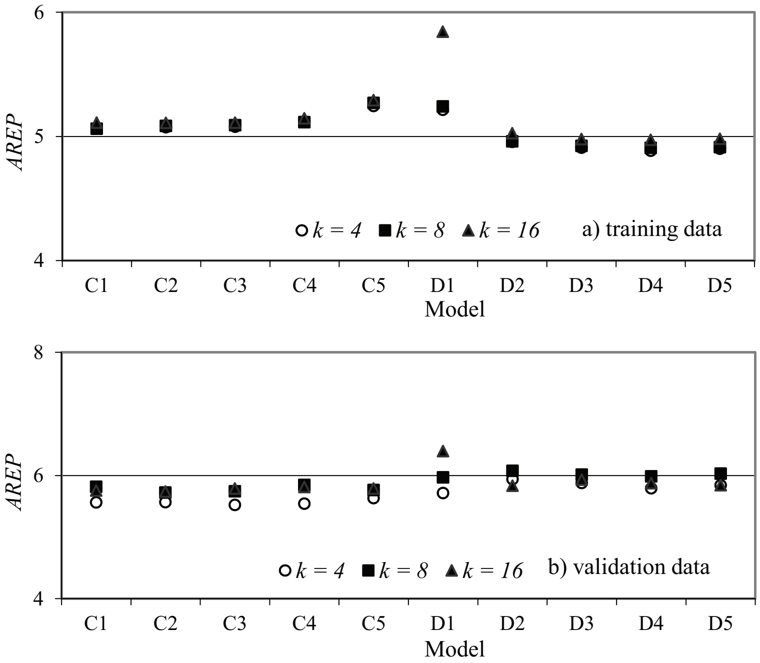


Figure 6.18: *The values of the AREP for the BPNN models (data2 - KFCV)*

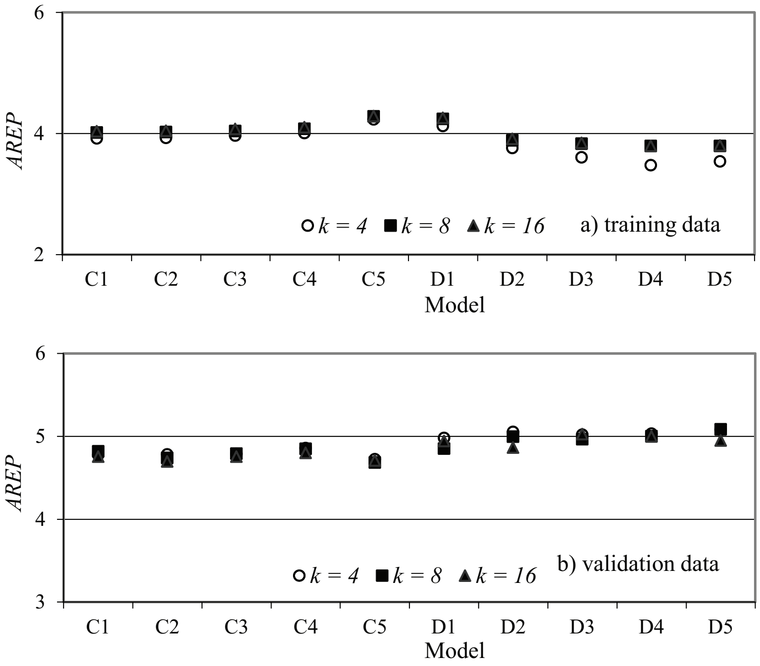


Figure 6.19: *The values of the AREP for the BPNN models (data3 - KFCV)*

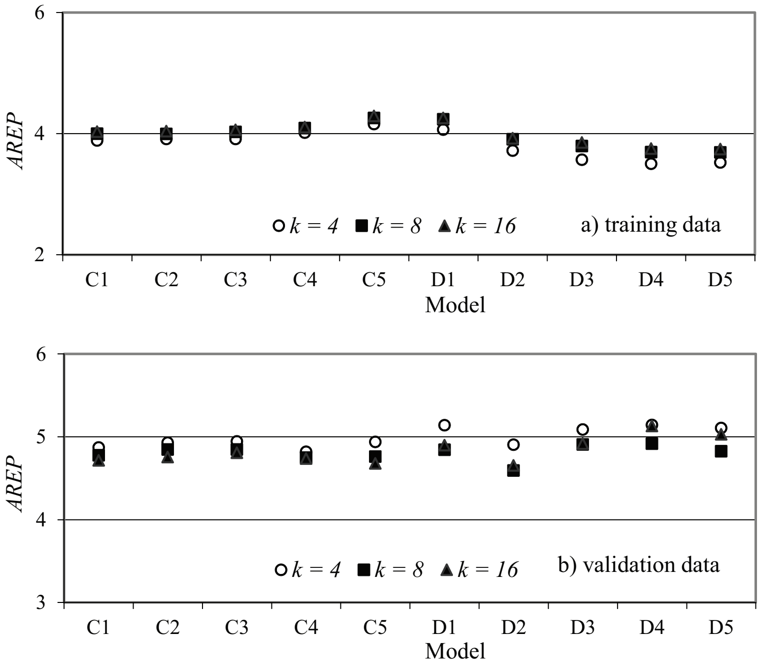


Figure 6.20: *The values of the AREP for the BPNN models (data4 - KFCV)*

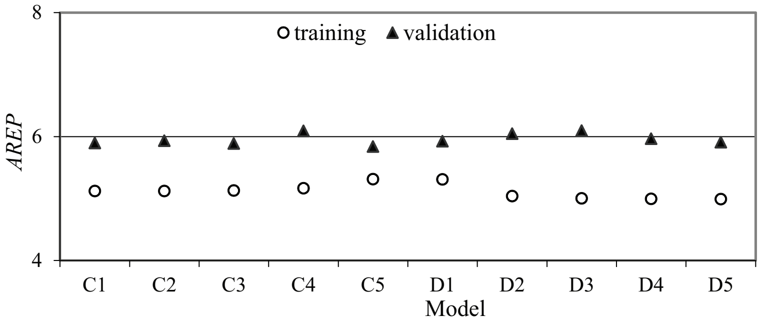


Figure 6.21: *The values of AREP for BPNN models (data1 - LOOCV)*

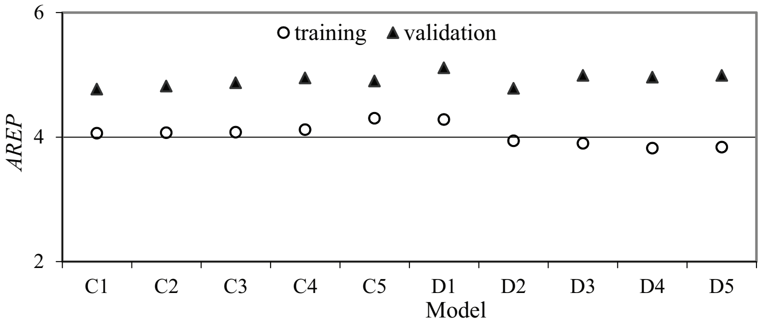


Figure 6.22: *The values of AREP for BPNN models (data3 - LOOCV)*

6.4.4 Multiple linear regression (MLR)

Two MLR models (E and F) with different input variables are assumed for travel time prediction, the model E with two input variables (ΔQ_R and Q_D) and the model F with three input variables (Q_R, ΔQ_R and Q_D). The general forms of the multiple linear regression MLR models E and F are given in equations, 6.1 and 6.2 respectively.

$$\hat{t} = a + b \times \Delta Q_R + c \times Q_D \tag{6.1}$$

$$\hat{t} = d + e \times Q_R + f \times \Delta Q_R + g \times Q_D \tag{6.2}$$

$\hat{t}$ is the predicted travel time and a, b, c, d, e, f and g are the parameters of the multiple linear regression.

The predicted travel time values are compared with the observed one for both training and validation data sets and the corresponding *AREP* values are estimated. The average values of the estimated *AREP* in data1, data2, data3 and data4 are given in table 6.6. Table 6.6 shows that the model E is slightly superior to the model F according to the average *AREP* in the validation data using LOOCV method. Comparing the performances of the MLR models with those of the ANFIS and BPNN models, it is clear that both the ANFIS model (A5) and the BPNN model (C5) outperform the MLR model (E).

Table 6.6: *The average values of the AREP for the MLR models*

Model	Training data				Validation data			
	KFCV			LOOCV	KFCV			LOOCV
	$k = 4$	$k = 8$	$k = 16$		$k = 4$	$k = 8$	$k = 16$	
data1								
E	7.57	7.59	7.6	7.6	7.81	7.76	7.82	7.89
F	7.53	7.81	7.59	7.76	7.6	7.82	7.6	7.97
data2								
E	7.58	7.81	7.59	7.6	7.6	7.82	7.6	7.89
F	7.51	7.54	7.56	7.76	7.89	8.04	7.97	7.97
data3								
E	6.27	6.26	6.27	6.29	6.43	6.58	6.66	6.56
F	6.27	6.25	6.27	6.29	6.44	6.72	6.79	6.66
data4								
E	6.27	6.3	6.29	6.29	6.43	6.3	6.35	6.56
F	6.25	6.3	6.29	6.29	6.65	6.37	6.39	6.66

Table 6.7 gives the values of the average the *AREP* in validation data set for the models A5 and model C5. The values of the *AREP* indicate that the ANFIS model A5 is superior to the BPNN model C5 except for data1 (LOOCV) and for data3 (KFCV with $k = 4$). With respect to the estimated values of the *AREP*, we select the ANFIS model (A5) to be the best model for travel time prediction.

Table 6.7: *The average values of the AREP for the models A5 and C5*

Model	data1				data2			data3				data4		
	KFCV			LOOCV	KFCV			KFCV			LOOCV	KFCV		
	k=4	k=8	k=16		k=4	k=8	k=16	k=4	k=8	k=16		k=4	k=8	k=16
A5	5.66	5.61	4.54	6.32	5.5	5.51	4.59	4.84	4.62	4.32	4.65	4.73	4.6	4.31
C5	5.91	5.7	5.81	5.84	5.63	5.76	5.79	4.72	4.69	4.71	4.9	4.94	4.76	4.68

6.5 HEC-RAS for travel time prediction

6.5.1 Introduction to HEC-RAS

HEC-RAS is a software system that is designed to perform three one-dimensional hydraulic analysis components (HEC, 2006):

1. Steady flow river calculations.

2. Unsteady flow simulation.

3. Movable boundary sediment transport computations.

Barkau (1992) developed an Unsteady NETwork (UNET) model, to simulate one-dimensional unsteady flow through a full NETwork of open channels. HEC (1997) modified UNET to perform the unsteady flow computations in HEC-RAS by solving the full dynamic Saint-Venant equations using an implicit finite difference method. Saint-Venant equations are derived from the equations of conservation of mass (continuity equation) and conservation of momentum (momentum equation). The equations of continuity and momentum can be defined as (Haestad Methods, 2003):

The continuity equation

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial T} - q_l = 0 \tag{6.3}$$

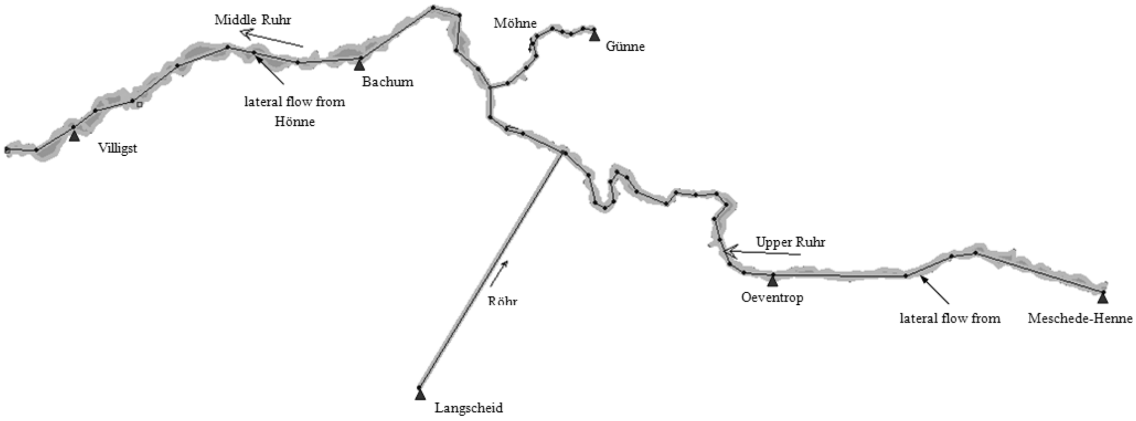


Figure 6.23: *The Geometric data of the HEC-RAS model*

The momentum equation

$$\frac{\partial Q}{\partial T} + \frac{\partial QV}{\partial l} + gA\frac{\partial z}{\partial l} - gA(S_o - S_f) = 0 \tag{6.4}$$

where Q is the flow rate, A is the cross-sectional area, T is the time, l is the distance along the channel, q_l is the lateral inflow per unit length, g is the acceleration gravity, $\partial z/\partial l$ is the water surface slope, $S_o$is the bed slope and $S_f$is the frictional slope. Equation (6.4) consists of five terms which known as local acceleration, connective acceleration, pressure force, gravity force and friction force terms.

6.5.2 Description of the HEC-RAS geometry model

The upper and middle reaches of the Ruhr are simulated with HEC-RAS to predict the travel time from the release gauges to the downstream gauges. The upper reach flows from km 179370 (Meschede-Henne) to Km 137525 (mouth of the Möhne) and the middle one flows from Km 137525 to Km 93006. The main tributaries which flow into these reaches of the Ruhr are the Wenne, Röhr, Möhne and Hönne Rivers. The flows from Wenne and Hönne Rivers are entered into HEC-RAS as lateral flows. Figure 6.23 shows the geometric data of the HEC-RAS model. Table 6.8 gives the main stations along the upper and middle reaches of the Ruhr and along the Röhr and Möhne Rivers.

6.5.3 Flow scenarios

We assumed seven flow scenarios to estimate the travel time of the releases from the Sorpe, Möhne and Henne reservoirs along the Ruhr using HEC-RAS. Table 6.9 lists the values of the flow at the release gauges and the flow from the tributaries for each scenario. The table gives also, the predicted values of travel time from the release gauge to the downstream gauges for each scenario. Figure 6.24 shows the hydrographs at the release and the downstream gauges for the scenario Meschede-Henne_3.

Table 6.8: *The main flow stations along the upper and middle reaches of the Ruhr River*

Station	River Kilometer
-along the Ruhr River	
Meschede-Henne	179370
The mouth of the Wenne	175046
Oeventrop	159500
The mouth of the Röhr	143200
The mouth of the Möhne	137525
Bachum	133830
The mouth of the Hönne	112904
Villigst	100150
-along the Möhne River	
Günne	11366
-along the Röhr River	
Langscheid	28900

6.5.4 Comparison between HEC-RAS and NLR for travel time prediction

From tables 6.1 and 6.9, it is obvious that reach 1 for the Henne reservoir (table 6.9) corresponds case study 6 (table 6.1) and reach 2 for the Sorpe and Möhne reservoirs (table 6.9) correspond case studies 4 and 5 (table 6.1) respectively. The predicted travel times from the HEC-RAS simulation and the NLR models are plotted versus the values of the flow at the downstream gauges (Q_D) as shown in figures 6.25, 6.26 and 6.27 for case studies 6, 4 and 5 respectively. Figures 6.25 and 6.27 show a moderate agreement between the predicted travel time using the HEC-RAS and that using the NLR model for case studies 6 and 5 respectively. Figure 6.26 shows that for case study 4, the predicted values of t using the HEC-RAS simulation deviate from that by the NLR model by 37 % for Q_D = 20 m^3/s and deviate by large amount for $Q_D <$ 20 m^3/s (more than 100 %).

Table 6.9: *List of the flow at the release gauges and the flow from tributaries and the predicted travel time values for each scenario*

Scenario	Flow at different gauges along the Ruhr (m^3/s)					Travel time (hrs)		
	Meschede-Henne	Wenne[1]	Langscheid	Günne	Hönne[1]	reach 1[2]	reach 2[2]	reach 3[2]
Henne reservoir								
Meschede-Henne_1	3_8[3]	1.7	1.5	2	1.7	6.3	13.75	23.45
Meschede-Henne_2	10_15	6	5	7	6	6.15	8.25	15.25
Meschede-Henne_3	20_25	12	10	14	12	3.35	6.95	11.23
Meschede-Henne_4	30_35	17	15	20	17	2.7	5.75	10.35
Meschede-Henne_5	40_45	24	20	28	24	2.75	5.35	10.1
Meschede-Henne_6	60_65	36	30	42	36	2.6	5.35	10
Meschede-Henne_7	80_85	48	40	56	48	2.1	4.85	9.65
Sorpe reservoir								
Langscheid_1	3	1.7	1.5_6.50	2	1.7	—-	10.7	20.95
Langscheid_2	7.2	4.1	3.6_8.60	4.8	4.1	—-	6.4	13.8
Langscheid_3	10	6	5_10	7	6	—-	3.8	10.6
Langscheid_4	20	12	10_15	14	12	—-	3.25	9.9
Langscheid_5	30	17	15_20	20	17	—-	3.35	8.25
Langscheid_6	40	24	20_25	28	24	—-	3.08	7.95
Langscheid_7	60	36	30_35	42	36	—-	2.75	7.75
Möhne reservoir								
Günne_1	3	1.7	1.5	2_7	1.7	—-	8.15	19.15
Günne_2	10	6	5	7_12	6	—-	4.5	11.35
Günne_3	20	12	10	14_19	12	—-	3.85	8.95
Günne_4	30	17	15	20_25	17	—-	3.5	8.25
Günne_5	40	24	20	28_33	24	—-	3.05	7.6
Günne_6	60	36	30	42_47	36	—-	2.6	7.25
Günne_7	80	48	40	56_61	48	—-	2.35	7.05

1 the flows from Wenne and Hönne are entered into the HEC-RAS simulations as lateral flows

2 reach 1, reach 2 and reach 3 are the reaches from the release gauges to gauges Oeventrop, Bachum and Villigst respecktively.

3 the flow is 3 m^3/s and it will be increase to reach 8 m^3/s in 1 hour then decrease to its original value in 1 hour.

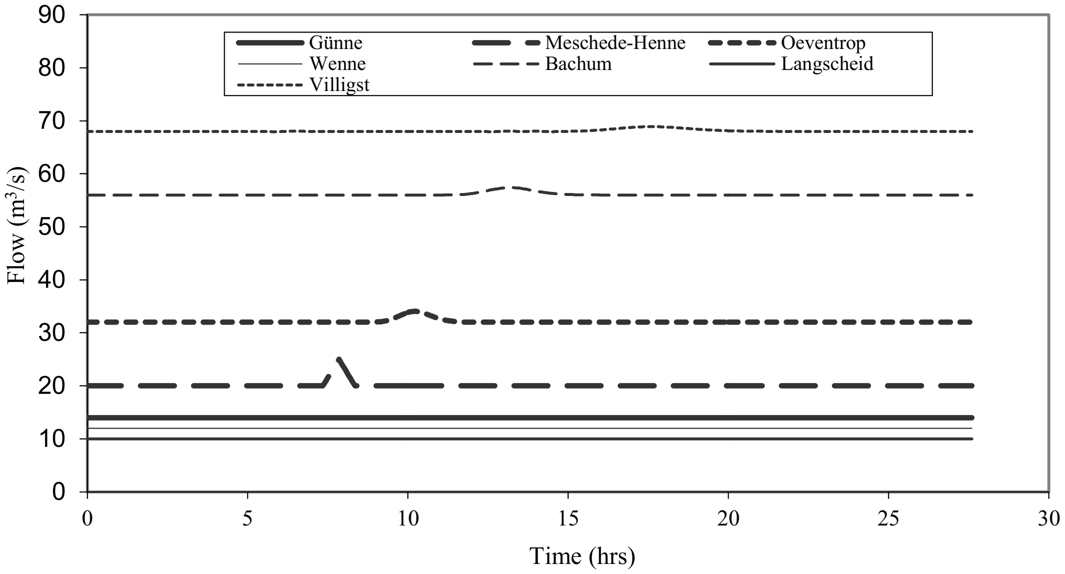


Figure 6.24: *The flow at the release gauges and the predicted flow (from the HEC-RAS simulation) at the downstream gauges (scenario Meschede-Henne_3)*

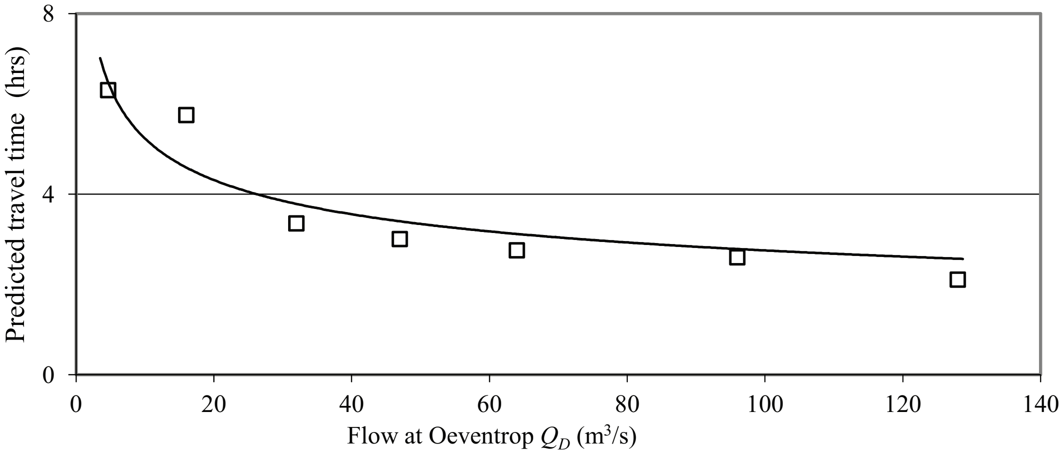


Figure 6.25: *Travel time from Henne reservoir to the gauge Oeventrop versus* Q_D

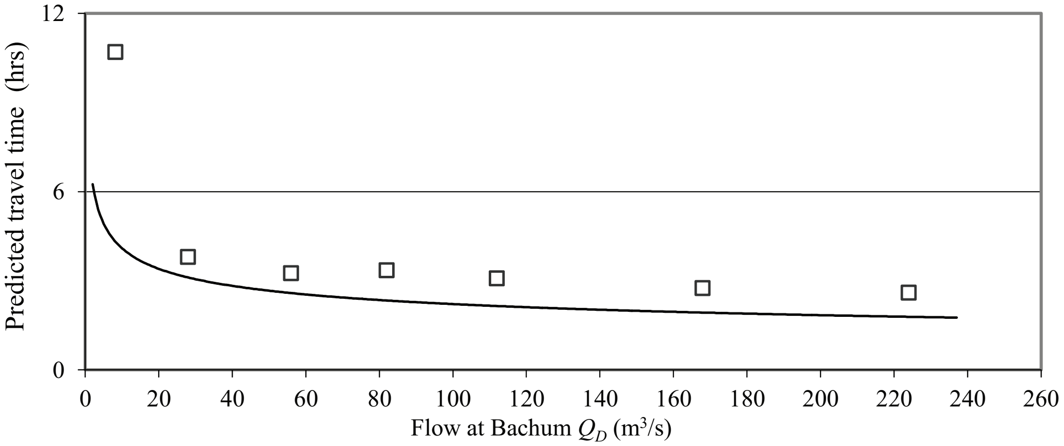


Figure 6.26: *Travel time from Sorpe reservoir to the gauge Bachum versus* Q_D

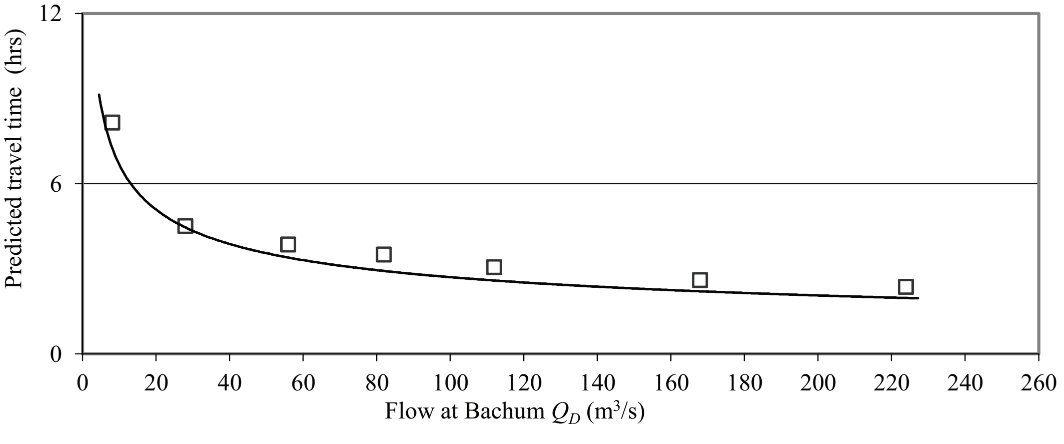


Figure 6.27: *Travel time from Möhne reservoir to the gauge Bachum versus* Q_D

6.6 Conclusions

Historical flow data (15 minute time series) are used to estimate the travel time of the reservoirs releases along the Ruhr and the Lenne Rivers. Using the estimated travel time values, the NLR models are built to detect a relation between travel time (t) from the reservoirs to some downstream gauges and the flow at each downstream gauge (Q_D). The estimated travel time values along the reach from gauge Ahausen to gauge Hagen-Hohenlimburg (case study 2) are simulated using the ANFIS, BPNN and MLR models. With respect to the estimated AREP, we suggested A5, C5 and E as the best ANFIS, BPNN and MLR models respectively with clear superiority of the A5 and C5 over E. The performances of the models A5 and C5 are compared based on the average values of the AREP and the ANFIS model (A5) is selected as the best model for travel time prediction. HEC-RAS is used to simulate the upper and middle reaches of the Ruhr River. Seven flow scenarios are assumed to estimate the travel time of the releases along the Ruhr using the HEC-RAS simulation. The results of the NLR models for case studies 4, 5 and 6 are compared with those of the HEC-RAS simulation and the results of the comparison show a moderate agreement for case studies 6 and 5. For case study 4, the predicted values of travel time using the HEC-RAS simulation deviate from that by the NLR model by 37 % for $Q_D \geq 20$ m^3/s and by large amount for $Q_D < 20$ m^3/s.

6.7 References

Abida, H., Ellouze, M., Mahjoub, M.R., 2005. Flood routing of regulated flows in Medjerda River, Tunisia. Journal of Hydroinformatics, 73, 209-216.

Anuradha, B., et al., 2008. Classification of cardiac signals using time domain methods. Journal of Engineering and Applied Sciences, 3(3), 7-12.

Barkau, R.L., 1992. UNET, One-dimensional unsteady flow through a full network of open channels. Computer Program, St. Louis, MO.

Budach, I., 1993. Ermittlung von Fliesszeiten in Ruhr und Lenne in Zusammenhang mit Abgabenänderungen aus den Ruhrtalsperren. Dipl. -Arbeit, Technische Univerität Dresden, Institut für Hydrologie und Meteorologie.

Good, P.I., 2006. Resampling methods: A practical guide to data analysis. 3rd edition, Birkhäuser.

Haestad Methods, Dyhouse, G., Hatchett, J., Benn, J., 2003. Floodplain modeling using Hec-Ras. Haestad Press, Waterbury, CT.

HEC, 1997. UNET, One-dimensional unsteady flow through a full network of open channels, User's manual. Hydrologic Enginering Center, U.S. Army Corps of Engineers, Davis, CA.

HEC, 2006. HEC-RAS, River analysis system user's manual. Version 4.0 Beta, Hydrologic Enginering Center, U.S. Army Corps of Engineers, Davis, CA.

Hu,Y.H., Hwang, J-N., 2002. Handbook of neural network signal processing. CRC Press.

Jobson, H.E., 1997. Predicting travel time and dispersions in rivers and streams. Journal of Hydraulic Engineering, 123(11), 971-978.

Jobson, H.E., 2001. Estimating the variation of travel time in rivers by use of wave speed and hydraulic characteristics. Journal of Hydraulic Engineering, 127(11), 911-918.

Morgenschweis, G., Nusch, E.A., 1990. Fliesszeitmessungen in der Ruhr bei Niedrigwasserabfluss. Sonderdruck aus: Jahresbericht Ruhrwassermege.

Priddy, K.L., Keller, P.E., 2005. Artificial neural networks: An introduction. SPIE, Tutorial Texts In Optical Engineering, TT68.

Samuels, W.B., Amstutz, D.E., Bryant, P., 2001. Using network analyst to calculate pollutant travel times.
http://proceedings.esri.com/library/userconf/proc01/professional/papers/pap172/p172.htm.

Chapter 7

Conclusions and future work

7.1 Conclusions

The thesis employs stochastic analysis and data-driven models into the Bigge, Henne, Möhne and Sorpe reservoirs in the Ruhr River basin. The inflow processes were investigated for seasonality, trend, long memory and stationarity at several characteristic timescales (e.g., one day; 10 days; one month; 3 months; ...). A clear seasonality is detected for different statistical characteristics (mean, standard deviation, skewness and season-to-season correlation). The linear regression, Mann-Kendall and seasonal Mann-Kendall tests are used to test for the presence of trend in the inflow time series at 5 % significance level. A downward trend is detected in the inflow time series of the Sorpe reservoir at different timescales. The 10-days, monthly, 3-months and 6-months inflow times of Sorpe reservoir are test against shift (occurred at 1970) and the detected shifts are removed. These seasonal time series (after removing the detected shifts) are tested against trend using seasonal Mann-Kendall test and a trend is detected only in the 10-days inflow time series. The ADF and PP tests are used to test the stationarity of the inflow time series and the results show that all log-transformed standardized time series are found to be stationary at 5 % significance level.

The BPNN, ANFIS, ARMA and ARFIMA models are developed for daily inflow forecasting. These models are divided into two groups, univariate models, group M1-1 and group M1-2 (the simulation models are BPNN, ANFIS, ARMA and ARFIMA) and multivariate models, group M2 (the simulation models are BPNN and ANFIS). The estimated values of the efficiency criteria of the univariate models show that all models have similar performance but the BPNN and ANFIS models are slightly better however. The multivariate models, group M2, are found to outperform the univariate models, group M1-2, in respect of all used efficiency criteria.

The BPNN and ANFIS and GLM models are used for filling in the missing values in the inflow time series. The estimated values of the *rmse* show that the BPNN model outperforms the other models. The BPNN model is applied to extend the daily inflow into the Bigge reservoir in the period from 1/11/1960 to 31/10/1965.

We developed four models (the T-F, Gamma T-F, MC and PHMM) for generating monthly inflow. The statistical parameters of the generated data are compared to those of the observed one and the results show that except the Gamma T-F each model preserves at least two statistical parameters. The Q-Q plot and the survivor function plot are used for visual validation of the models. The results of the Q-Q plot show superiority of the MC and PHMM models, however the PHMM outperforms the other models according to the results of the survivor function plot. The MC model is applied to detect the expected consecutive 5 years that have minimum total inflow.

Historical flow data (15 minute time series) are used to estimate the travel time of the released flow from the release gauges to some downstream gauges. Using the estimated travel time values, the ANFIS, BPNN and MLR are trained to predict travel time. The ANFIS model (A5) is proposed as the best model for travel time prediction.

HEC-RAS is applied to simulate the upper and middle reaches of the Ruhr River and the results of the HEC-RAS simulation are compared with those of the NLR model and the results of the comparison show a moderate agreement for case studies 6 and 5.

7.2 Future work

For efficient planning and management of water resources in the Ruhr River basin, it is recommended that further research be undertaken in the following areas:

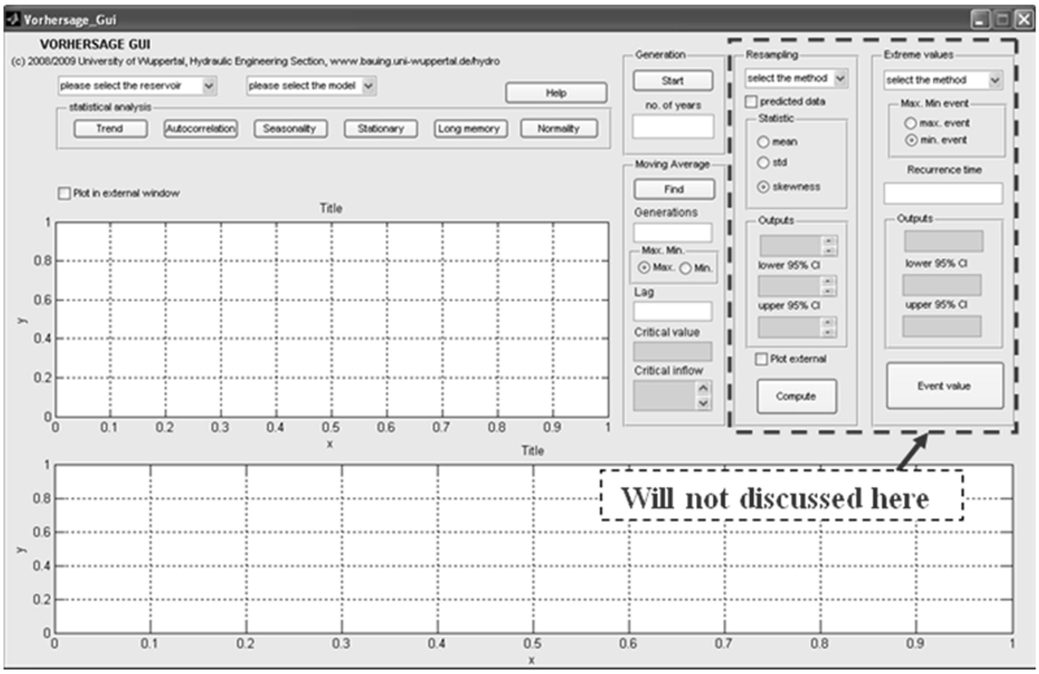
- Improving the performance of the PHMM in preserving month-to-month correlation.
- Developing a multi-site inflow forecasting model to forecast daily inflow into multiple reservoir systems.
- Synthetic generation of the seasonal inflow into multiple reservoir systems using a multivariate, seasonal inflow generation model.

Appendix A

Graphical user interfaces

A.1 Vorhersage GUI

Vorhersage GUI is a graphical user interface which is developed by using Matlab. Figure A.1 shows a general overview of the Vorhersage GUI. This GUI can be used for statistical analysis of the inflow time series at different timescales, for monthly inflow generation and for many other applications. Some of these applications will not be discussed here.



VORHERSAGE GUI

(c) 2008/2009 University of Wuppertal, Hydraulic Engineering Section, www.bauing.uni-wuppertal.de/hydro

Will not discussed here

Figure A.1: *General overview of the Vorhersage GUI*

A.1.1 Statistical analysis

Different statistical analysis can be applied using Vorhersage GUI. In the following is a list of these statistical analyses:

- Trend analysis.
- Autocorrelation analysis.
- Seasonality analysis.
- Stationarity analysis.
- Long memory analysis.

Select the reservoir and then press the button corresponding to the statistical analysis that you want to start. Each statistical analysis can be done for different time scales (daily, 10-days, monthly ...). Figure A.2 shows the seasonal variation in the standard deviation for monthly inflow time series of the Bigge reservoir.

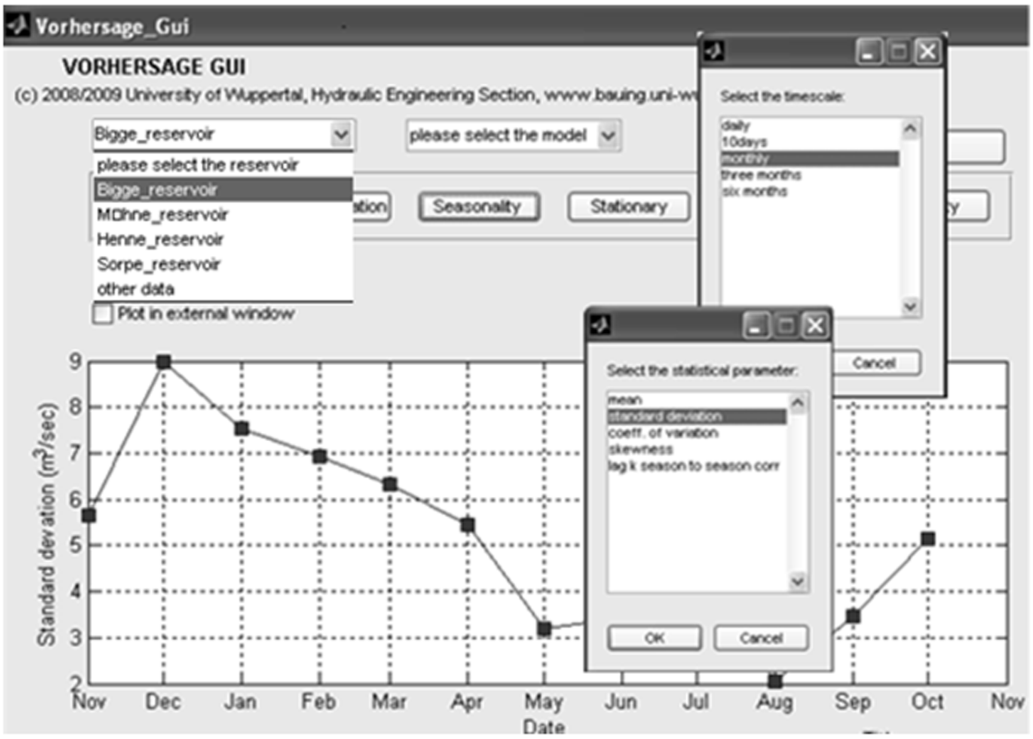


Figure A.2: *Vorhersage GUI for seasonality analysis*

A.1.2 Monthly inflow generation

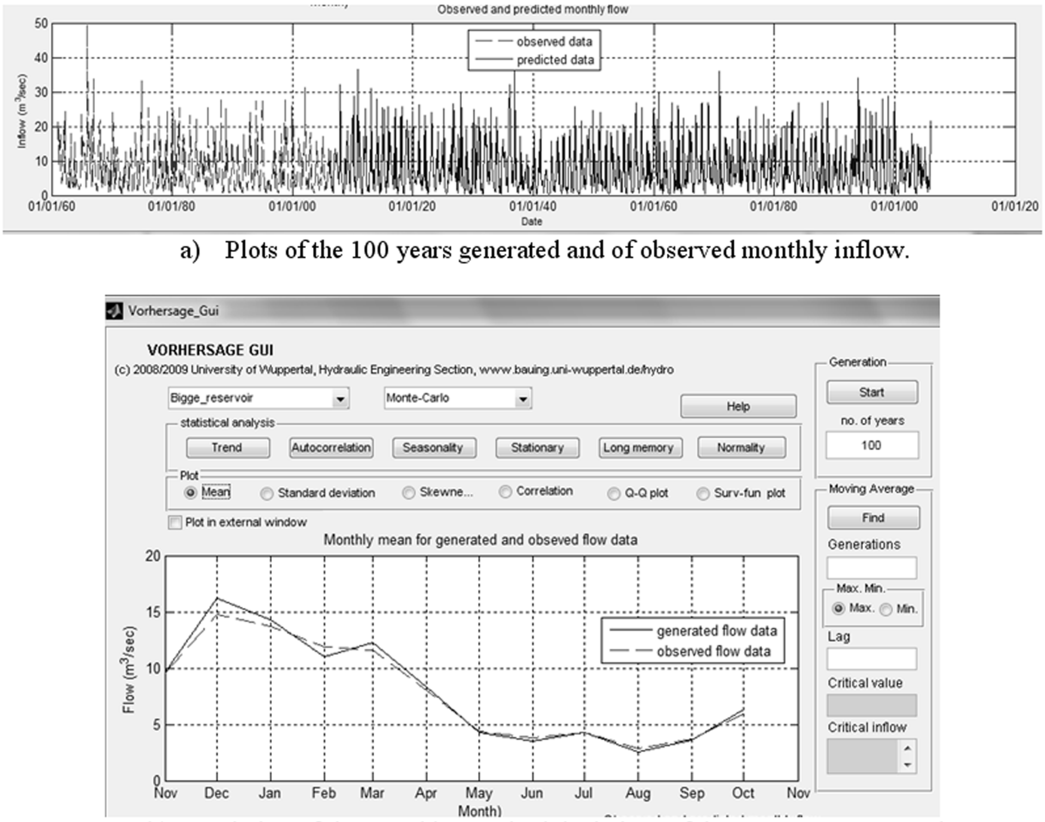
One of the other applications of the Vorhersage GUI is to generate the monthly inflow into the Bigge, Henen, Möhne and Sorpe reservoirs. Four models can be implemented using this GUI:

- Thomas-Fiering model (T-F).
- Monte-Carlo model (MC).
- Gamma Thomas-Fiering model (Gamma T-F).
- Periodic hidden Markov model (PHMM).

The following steps have to be followed to generate the monthly inflow into the selected reservoir (see figure A.3):

1. Select the reservoir and the model from the corresponding Popup Menus.
2. Insert the number of years to which you want to generate the monthly inflow.
3. Press Start Button.
4. Once this Button is pressed, the generated inflow data will be plotted against dates. Also, the statistics of the generated and the observed inflow data can be plotted against months.

Figure A.3.a shows the 100 years generated monthly inflow into the Bigge reservoir. It shows also a comparison between the standard deviations of the 100 years generated monthly inflow and the observed one (see figure A.3.b).



a) Plots of the 100 years generated and of observed monthly inflow.

b) Variation of the monthly standard deviations of the 100 years generated monthly inflow and of the observed one.

Figure A.3: *Vorhersage GUI for monthly inflow generation*

A.1.3 Critical moving sum

Vorhersage GUI can be also used to find the consecutive n years which have the critical total inflow (minimum or maximum). The following steps have to be followed to find the consecutive n years which have the critical sum (see figure A.4):

1. Select the reservoir and the model.

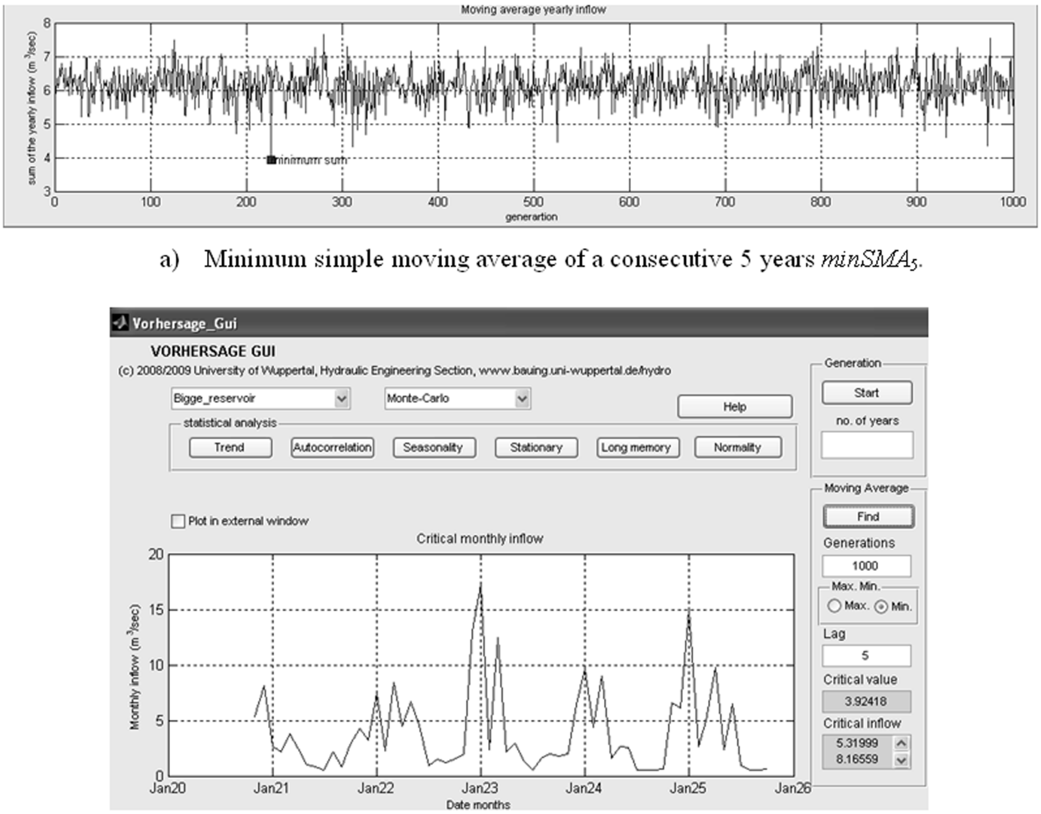
2. Insert the number of generations.

3. Give the number of the consecutive years (is denoted by, Lag) for which you want to find the critical value.

4. Press Find Button.

5. A Dialog Box will appear and ask about the number of years to be generated each time.

A plot which shows the values of the critical sum vs. generations will appear as shown in figure A.4.a. The monthly inflow during the critical period is plotted in figure A.4.b.



VORHERSAGE GUI
(c) 2008/2009 University of Wuppertal, Hydraulic Engineering Section, www.bauing.uni-wuppertal.de/hydro

Bigge_reservoir
Monte-Carlo

□ Plot in external window

Generations: 1000
Max. Min.: ○ Max. ● Min.
Lag: 5
Critical value: 3.92418
Critical inflow: 5.31999 8.16559

a) Minimum simple moving average of a consecutive 5 years *minSMA*$_5$.

b) Monthly inflow during critical consecutive 5 years.

Figure A.4: *Vorhersage GUI for determination of the critical moving sum*

A.2 Fehlende Daten GUI

Fehlende Daten GUI can be used estimate the missing data in the daily inflow time series using BPNN, ANFIS or GLM models. It can be also used to re-train these models using different input variables and training parameters. The main features of the Fehlende Daten GUI are summarized as follows (see figure A.5):

1. to select the gauge at which you want to estimate the missing data.

2. to select the model which you want to use.

3. to insert the values of the input variables at which the missing data to be estimated.

4. another method to insert the value of the input variables is to load them as *.xls* file. Select the corresponding Check Box to activate this feature.

5. press Start Button to estimate the missing data.

6. to display the estimated data.

7. to display the performance of the used model in terms of the correlation coefficient between the estimated data and the observed one.

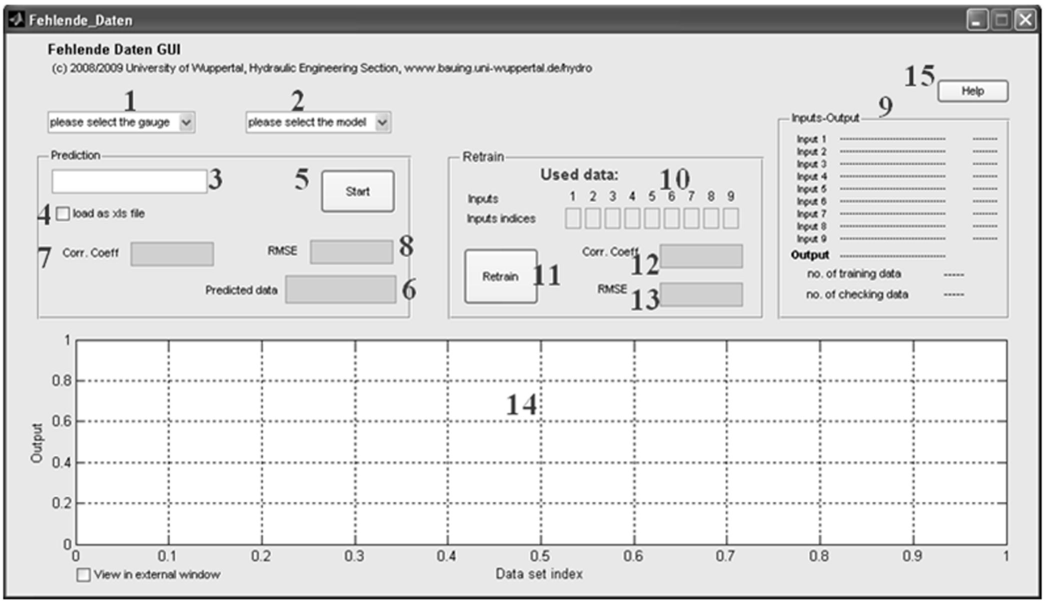


Figure A.5: *General overview of the Fehelend Daten GUI*

8. to display the performance of the used model in terms of the root mean square error of the estimated data.

9. to display the input variables which are used to train the model.

10. to select the input variables to train the model. Input index of nonzero value means the corresponding input variable will be used to train the model otherwise it will not be used to train the model.

11. press Train Button to re-train the model with different input variables or different parameters.

12. to display the correlation coefficient between the estimated data using the re-trained model and the observed data.

13. to display the root mean square error of the estimated data using the re-trained model.

14. to display the estimated and the observed data vs. date and

15. press Help Button to open a file in PDF format. This file describes the different features of the GUI.

A.3 Zuflussprognose GUI

Zuflussprognose GUI is a graphical user interface which is used to train the different models for daily inflow forecasting. This GUI can be also used to forecast the daily inflow using these models. In the following are the main features of the Zuflussprognose GUI (see figure A.6): Features 1 through 15 are the same as in the Fehlende Daten GUI.

16. to select the models group.

17. press this Button to apply the diagnostic checking to the ARMA and ARFIMA model.

18. to plot the autocorrelation and the partial autocorrelation function of the selected inflow process.

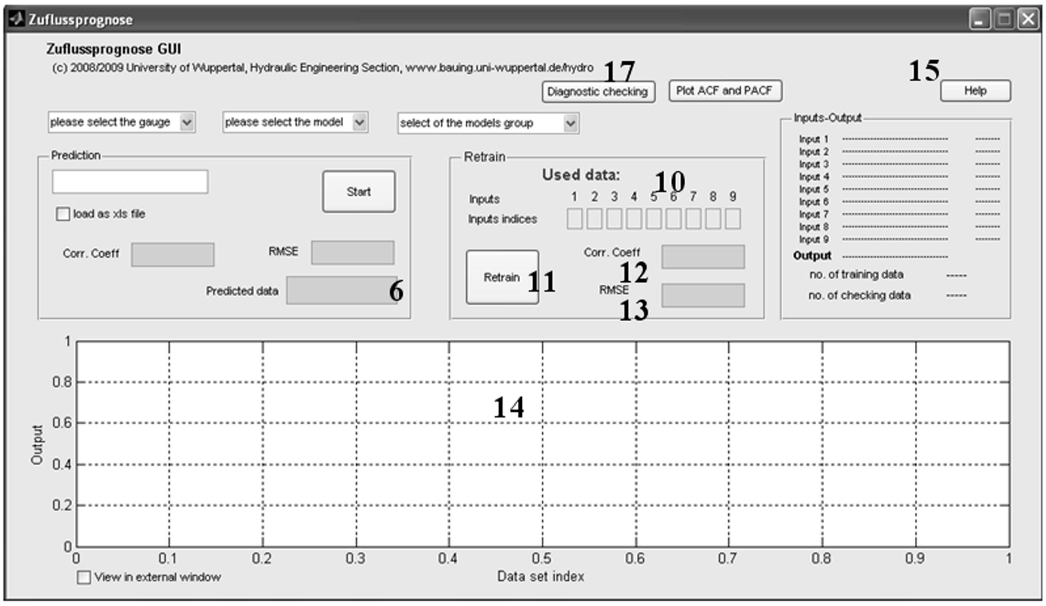


Figure A.6: *General overview of the Zuflussprognose GUI*

A.4 Fliesszeit GUI

Fliesszeit is a Graphical User Interface which can be used to (see figure A.7):

1. Prediction of the travel time (t).

2. Updating of training models.

3. Determination of jump points.

A.4.1 Prediction of the travel time

To predict the travel time (t) and the increase in the downstream flow (ΔQ_d), the following steps should to be followed (see figure A.8):

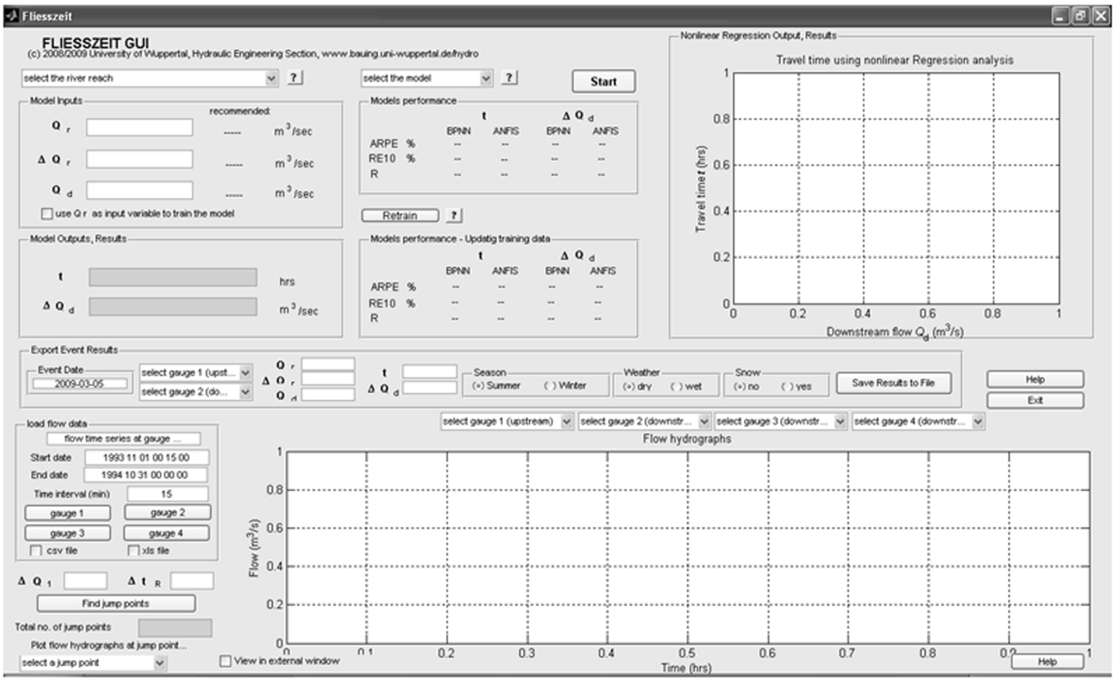
1. Choose the river reach.

2. Choose the model.

3. Enter the values of Q_r (if it will be used as input variable), ΔQ_r and Q_d. The recommended range for each input variable will appear next to the corresponding Edit Text Box.

4. Press Start Button.

5. The predicted values of t and ΔQ_d will appear in the corresponding Text Boxes; also a figure which shows the relation between t and Q_d will appear. This figure shows the relation between t and Q_d using nonlinear regression analysis.



FLIESSZEIT GUI
(c) 2008/2009 University of Wuppertal, Hydraulic Engineering Section, www.bauing.uni-wuppertal.de/hydro

Model Inputs

		recommended:	
Q_r		-----	m^3/sec
ΔQ_r		-----	m^3/sec
Q_d		-----	m^3/sec

□ use Q r as input variable to train the model

Models performance

		t		ΔQ_d	
		BPNN	ANFIS	BPNN	ANFIS
ARPE	%	--	--	--	--
RE10	%	--	--	--	--
R		--	--	--	--

Model Outputs, Results

t hrs

ΔQ_d m^3/sec

Models performance - Updatig training data

		t		ΔQ_d	
		BPNN	ANFIS	BPNN	ANFIS
ARPE	%	--	--	--	--
RE10	%	--	--	--	--
R		--	--	--	--

Export Event Results

Event Date 2009-03-05

select gauge 1 (upst...

select gauge 2 (do...

Q_r ΔQ_r Q_d t ΔQ_d

Season (•) Summer ( ) Winter

Weather (•) dry ( ) wet

Snow (•) no ( ) yes

load flow data

flow time series at gauge ...

Start date 1993 11 01 00 15 00

End date 1994 10 31 00 00 00

Time interval (min) 15

□ csv file □ xls file

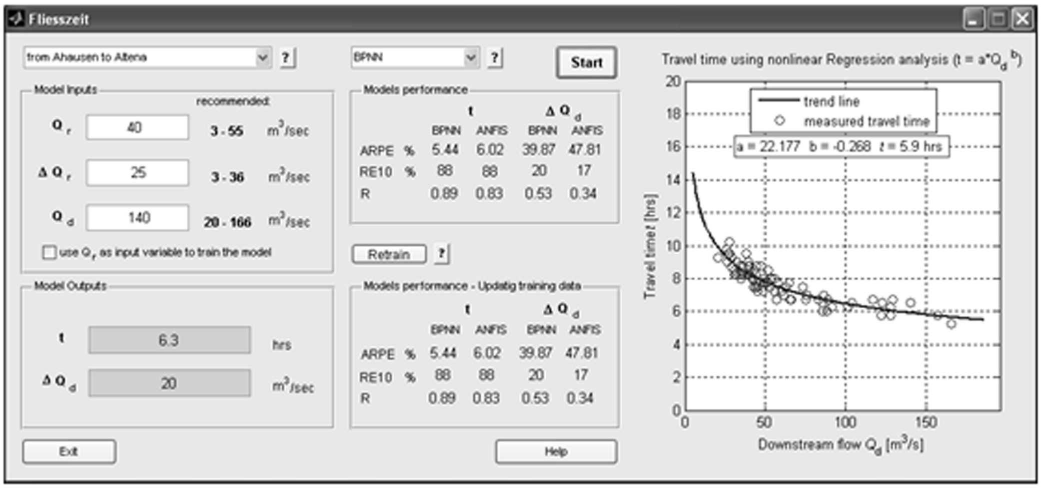
ΔQ_1 Δt_R

Total no. of jump points

Plot flow hydrographs at jump point...

□ View in external window

Figure A.7: *General overview of the Fliesszeit GUI*



from Ahausen to Altena

BPNN

Model Inputs

		recommended:	
Q_r	40	**3 - 55**	m^3/sec
ΔQ_r	25	**3 - 36**	m^3/sec
Q_d	140	**20 - 166**	m^3/sec

□ use Q_r as input variable to train the model

Models performance

		t		ΔQ_d	
		BPNN	ANFIS	BPNN	ANFIS
ARPE	%	5.44	6.02	39.87	47.81
RE10	%	88	88	20	17
R		0.89	0.83	0.53	0.34

Model Outputs

t 6.3 hrs

ΔQ_d 20 m^3/sec

Models performance - Updatig training data

		t		ΔQ_d	
		BPNN	ANFIS	BPNN	ANFIS
ARPE	%	5.44	6.02	39.87	47.81
RE10	%	88	88	20	17
R		0.89	0.83	0.53	0.34

a = 22.177 b = -0.268 t = 5.9 hrs

Figure A.8: *Prediction of the travel time*

A.4.2 Updating of training models

Using Fliesszeit GUI the decision maker can add training samples and try different training parameters. If the decision maker evaluates the performance of the model and decides that the new training samples or/and the new training parameters improve the performance of the model, he can update the training data (see figure A.9).

To update the training models (add training samples and try different training parameters), the following steps should to be followed:

1. Choose the river reach.

2. Press Retrain Button.

3. A question dialog will appear and asks about the password.

4. After entering the password another question dialog will appear and asks if you want to add new training sets or not. Press yes to add new training sets or no to continue. If you press yes an Input Dialog will appear and asks about the values of Qr, ΔQ_r, Q_d, ΔQ_d and t.

5. A Question Dialog will appear and asks if you want to replace the existing training sets with the new or not, press yes to replace the existing training sets or no to add the new one.

6. Then you have the ability to try different training parameters:

BPNN model

- number of neurons in the hidden layer.
- maximum number of epochs.

ANFIS model

- number of membership functions corresponding to the input variables Q_r, ΔQ_r and Q_d.
- maximum number of epochs.

7. After entering the training parameters press ok to start the training process (it can take few minutes).

8. After finishing the training process a Question Dialog will appear and asks if you want to update the training model. Compare between the efficiency criteria in the Models performance Group and that in the Models performance-Updating training data Group (see figure A.10).

9. If the performance of the model is improved press ok to update the training model otherwise press no will delete the new training sets and return to the previous model.

Models performances
Three efficiency criteria (*ARPE*, *RPE10* and *R*) are used to measure the performance each model where:

ARPE is the average relative percentage error.

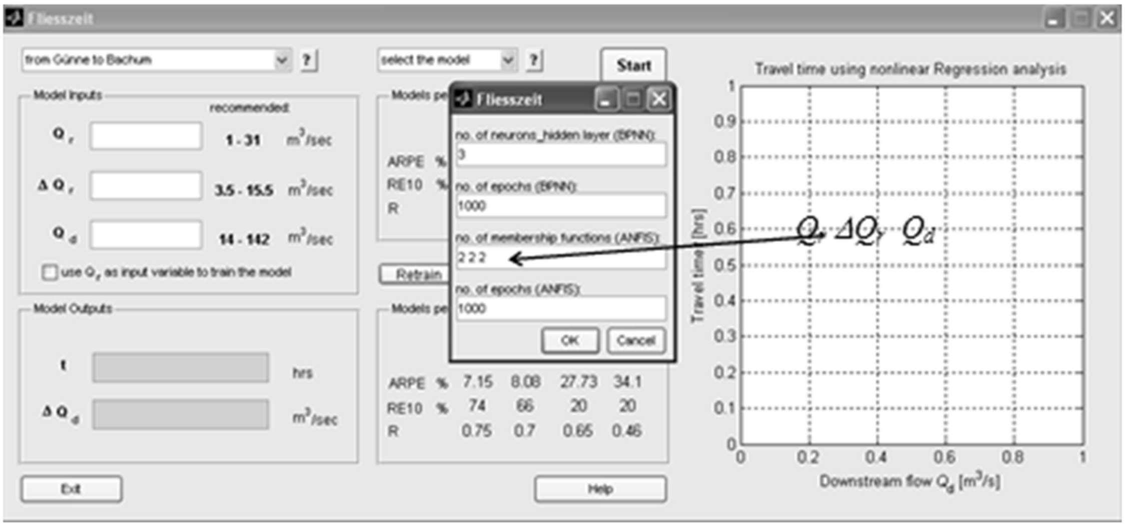
RE10 is the percentage of the observations which have *RPE* less than 10, for example if the total number of observations is 50 and the number of observations which have $RPE \leq 10 = 43$ then $RE10 = 43/50 = 0.86$).

R is the average value of the correlation between the observed values and the predicted one.

A.4.3 Estimation of travel time

Other application of the Fliesszeit GUI is to estimate the travel time between two gauges using the flow time series at each gauge. The following two steps can be used to estimate the travel time using Fliesszeit GUI (see figure A.11):

1. Load flow data
Fliesszeit GUI has the ability to load the flow time series from Edit Text Box, as *.csv* file or as *.xls* file. One of these methods can be used to load the flow data at the release gauge (gauage1) and at the downstream gauges (gauage2, gauage3, gauage4).



from Günne to Bachum

select the model

Model Inputs (recommended):

Q_r 1 - 31 m³/sec

ΔQ_r 3.5 - 15.5 m³/sec

Q_d 14 - 142 m³/sec

□ use Q_r as input variable to train the model

Model Outputs:

t hrs

ΔQ_d m³/sec

Fliesszeit

no. of neurons_hidden layer (BPNN): 3

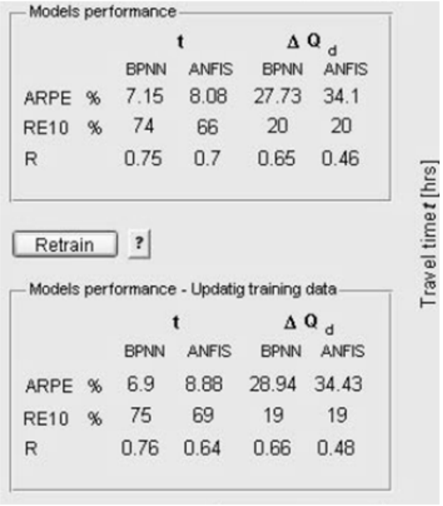
no. of epochs (BPNN): 1000

no. of membership functions (ANFIS): 2 2 2

no. of epochs (ANFIS): 1000

ARPE	%	7.15	8.08	27.73	34.1
RE10	%	74	66	20	20
R		0.75	0.7	0.65	0.46

Figure A.9: *Updating of the training models*



Models performance

		t		ΔQ_d	
		BPNN	ANFIS	BPNN	ANFIS
ARPE	%	7.15	8.08	27.73	34.1
RE10	%	74	66	20	20
R		0.75	0.7	0.65	0.46

Models performance - Updatig training data

		t		ΔQ_d	
		BPNN	ANFIS	BPNN	ANFIS
ARPE	%	6.9	8.88	28.94	34.43
RE10	%	75	69	19	19
R		0.76	0.64	0.66	0.48

Figure A.10: *Performance of the models*

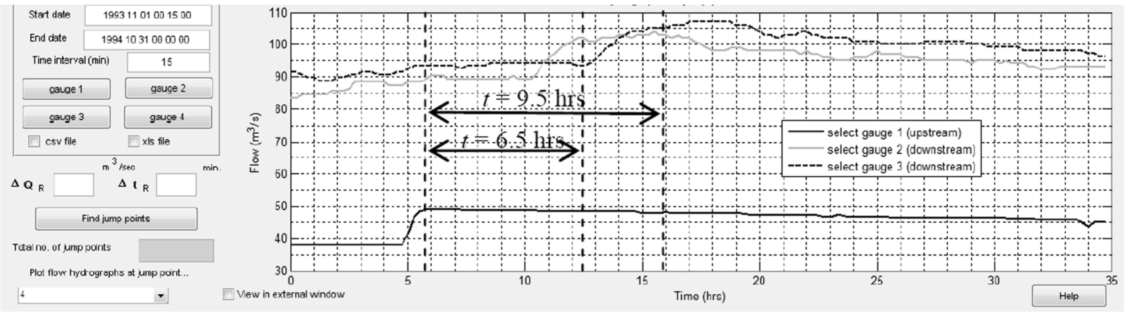
2. Determination of the jump points

Determination of the jump points is the next step after loading the flow data. The following two parameters are required to determine the jump point at the release gauge (see figure A.11):

ΔQ_R the increase in the flow at gauge 1 (release gauge)

Δt_R the assumed rise time at gauge 1 (different values can assumed 45min, 60min, 75min ...)

After entering the values of the required parameters press on the Push Button (Find jump points). The total number of the determined jump points will appear. The flow hydrographs at each jump point can be drawn using the Popup Menu (Plot flow hydrographs at jump point). This plot can be used to estimate the travel time at this jump point manually (if possible) as shown in figure A.11.



Start date 1993 11 01 00 15 00
End date 1994 10 31 00 00 00
Time interval (min) 15
□ csv file □ xls file
Plot flow hydrographs at jump point... 4
□ View in external window

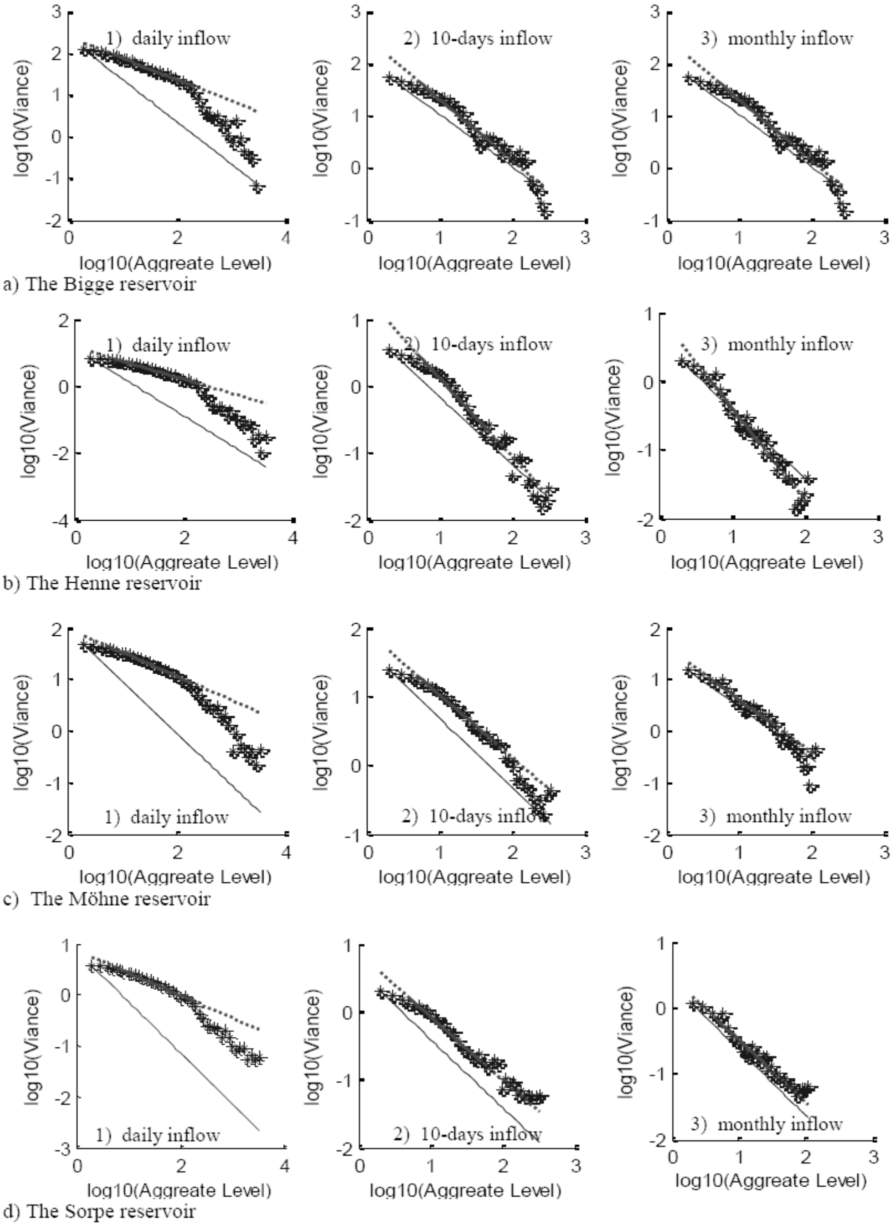
Figure A.11: *Estimation of the travel time*

Appendix B

Graphical presentation of the long memory detection methods

B.1 Aggregated variance method

Hurst parameter (H) is estimated as $H = 0.5\,e + 1$,where e is slope of the trend line.



a) The Bigge reservoir

b) The Henne reservoir

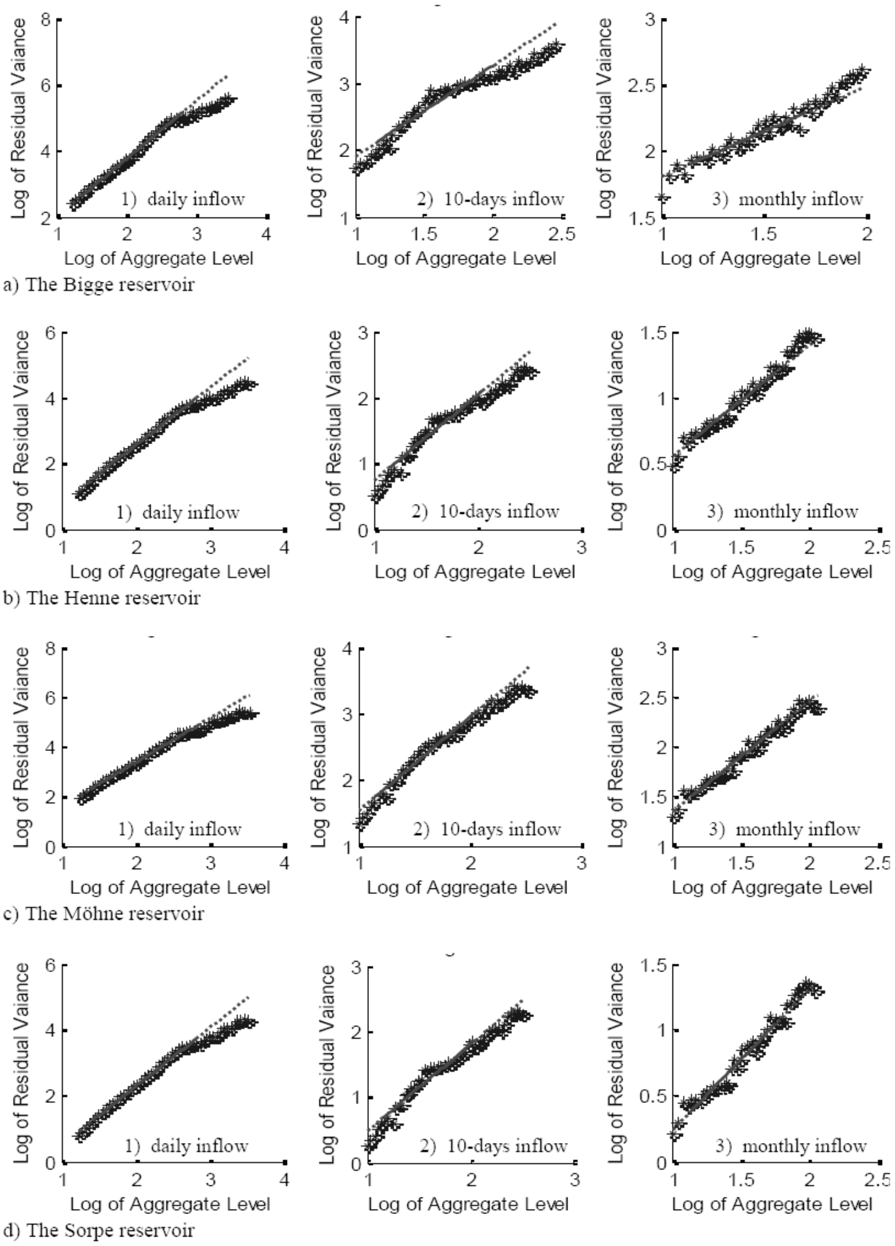
c) The Möhne reservoir

d) The Sorpe reservoir

Figure B.1: *Logarithmic plot of the variance versus the aggregate level (the aggregate variance method)*

B.2 Residuals of regression method

Hurst parameter (H) is estimated as $H = 0.5\,e$,where e is slope of the trend line.



a) The Bigge reservoir

b) The Henne reservoir

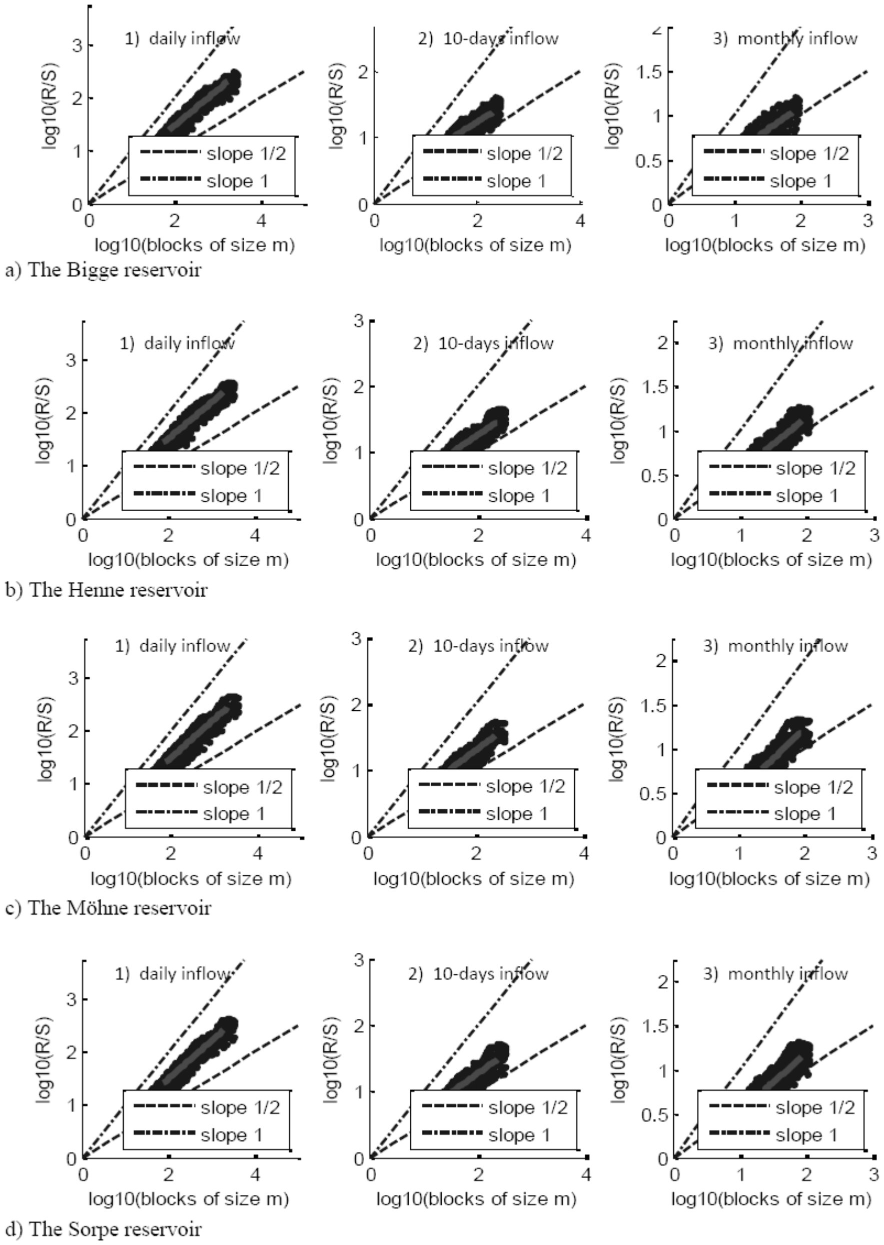
c) The Möhne reservoir

d) The Sorpe reservoir

Figure B.2: *Logarithmic plot of the residual variance versus the aggregate level (the residuals of regression method)*

B.3 The rescaled adjusted range (R/S) method

Hurst parameter (H) is estimated as $H = e$,where e is slope of the trend line.



a) The Bigge reservoir

b) The Henne reservoir

c) The Möhne reservoir

d) The Sorpe reservoir

Figure B.3: *Logarithmic plot of the R/S variance versus the size of block m (the R/S method)*