EVALUATION AND MODELING OF STREAMFLOW DATA: ENTROPY METHOD, AUTOREGRESSIVE MODELS WITH ASYMMETRIC INNOVATIONS AND ARTIFICIAL NEURAL NETWORKS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

NERMİN ŞARLAK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

IN

CIVIL ENGINEERING

JUNE 2005

Approval of the Graduate School of Natural and Applied Sciences

Prof. Dr. Canan ÖZGEN
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Erdal ÇOKÇA
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. A. Ünal ŞORMAN
Supervisor

Examining Committee Members

Prof. Dr. Uygur ŞENDİL (METU,CE) _____

Prof. Dr. A. Ünal ŞORMAN (METU,CE) _____

Prof. Dr. A. Melih YANMAZ (METU,CE) _____

Prof. Dr. Erol KESKİN (Süleyman Demirel Unv.) _____

Assoc. Prof. Dr. Ayşen AKKAYA (METU, STAT) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Nermin ŞARLAK

Signature :

ABSTRACT

EVALUATION AND MODELING OF STREAMFLOW DATA: ENTROPY METHOD, AUTOREGRESSIVE MODELS WITH ASYMMETRIC INNOVATIONS AND ARTIFICIAL NEURAL NETWORKS

Şarlak, Nermin

Ph.D., Department of Civil Engineering

Supervisor: Prof. Dr. A. Ünal ŞORMAN

June 2005, 171 pages

In the first part of this study, two entropy methods under different distribution assumptions are examined on a network of stream gauging stations located in Kızılırmak Basin to rank the stations according to their level of importance. The stations are ranked by using two different entropy methods under different distributions. Thus, showing the effect of the distribution type on both entropy methods is aimed.

In the second part of this study, autoregressive models with asymmetric innovations and an artificial neural network model are introduced. Autoregressive models (AR) which have been developed in hydrology are based on several assumptions. The normality assumption for the innovations of AR models is investigated in this study. The main reason of making this assumption in the autoregressive models established is the difficulties faced in finding the model parameters under the distributions other than the normal distributions. From this point of view, introduction of the modified maximum likelihood procedure developed by Tiku et. al. (1996) in estimation of the autoregressive model parameters having non-normally distributed residual series, in the area of hydrology has been aimed. It is also important to consider how the autoregressive model parameters having skewed distributions could be estimated.

Besides these autoregressive models, the artificial neural network (ANN) model was also constructed for annual and monthly hydrologic time series due to its advantages such as no statistical distribution and no linearity assumptions.

The models considered are applied to annual and monthly streamflow data obtained from five streamflow gauging stations in Kızılırmak Basin. It is shown that AR(1) model with Weibull innovations provides best solutions for annual series and AR(1) model with generalized logistic innovations provides best solution for monthly as compared with the results of artificial neural network models.

Keywords: Entropy method, Autoregressive model, Asymmetric innovations, Modified maximum likelihood, Artificial neural network

ÖZ

AKIM VERİLERİNİN DEĞERLENDİRİLMESİ VE MODELLENMESİ: ENTROPİ METODU, SİMETRİK OLMAYAN HATA TERİMLİ OTOREGRESSİF MODELLER VE YAPAY SİNİR AĞLARI

Şarlak, Nermin
Doktora, İnşaat Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. A. Ünal ŞORMAN

Haziran 2005, 171 sayfa

Çalışmanın ilk kısmında, iki ayrı Entropi yöntemi farklı dağılım varsayımları altında Kızılırmak havzasında yer alan akım gözlem ağındaki istasyonları önem seviyelerine göre sıralamak için irdelenmiştir. Farklı dağılımlar için Yöntem 1 ve Yöntem 2'den istasyon sıralamaları elde edilmiştir. Böylece, dağılım tiplerinin her iki yöntem üzerindeki etkilerinin gösterilmesi amaçlanmıştır.

Çalışmanın ikinci kısmında, simetrik olmayan hata terimli otoregressif modeller ve yapay sinir ağları tanıtılmıştır. Hidrolojik zaman serilerini

modellemek için hidrolojide geliştirilen otoregressif modeller (AR) çeşitli varsayımlara dayanmaktadır. Bu çalışmada, otoregressif modellerin dayandığı başlıca varsayım olan artık serilerin normal dağıldığı varsayımı irdelenmiştir. Kurulan otoregressif modellerde bu varsayımın yapılmasının başlıca nedeni normal dağılım dışındaki dağılımlarda da model parametrelerinin bulunmasında karşılaşılan zorluklardır. Bu bakımdan, normal dağılıma uymayan artık serilere sahip otoregressif modellerin parametrelerini tahmin etmede Tiku ve diğerleri (1996) tarafından geliştirilen "uyarlanmış en çok olabilirlik" yönteminin hidroloji alanına tanıtımı amaçlanmıştır. Çarpık dağılımlara sahip otoregressif modellerin parametrelerinin nasıl tahmin edilebileceğinin gözönüne alınması da önemlidir.

Otoregressif modellerin yanısıra istatistiksel dağılım ve lineer ilişki varsayımları içermeyen yapay sinir ağları modeli (YSA) de yıllık ve aylık hidrolojik zaman serileri için kurulmuştur.

Dikkate alınan modeller, Kızılırmak havzasındaki beş akım gözlem istasyonları yıllık ve aylık veri setlerine uygulanmıştır. Yapay sinir ağları modelleri sonuçları ile kıyaslandığında yıllık akım verileri için Weibull hata terimli AR(1) modeli en iyi sonucu verirken, aylık akım verileri için genel lojistik hata terimli AR(1) modeli en iyi sonucu vermiştir.

Anahtar kelimeler: Entropi yöntemi, Otoregressif model, Simetrik olmayan hata terimleri, Uyarlanmış en çok olabilirlik, Yapay sinir ağı

TO MY PARENTS

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Dr. A. Ünal Şorman for providing me with the opportunity to undertake the projects and his guidance, advice, criticism throughout the research.

I would like to thank the thesis follow-up committee members Prof Dr. Uygur Şendil and Assoc. Prof. Dr. Ayşen Akkaya. I appreciate the valuable discussions of Dr. Ayşen Akkaya. Her knowledge and teaching have improved the thesis greatly. I acknowledge Dr. Uygur Şendil's encouragement, criticism and editing.

I would like to thank also the jury members for their participation and criticism.

The research work reported in this thesis is supported by the Scientific and Technical Research Council of Turkey under grant: İÇTAG I-843. The scholarship provided by the Scientific and Technical Research Council of Turkey is gratefully acknowledged.

I would like to thank Celil Ekici who edited my thesis.

The continual support of my friends encouraged me in my work. To all my friends in my life, thanks for their laughs, support, patience and memories. Thank you for always being near me.

My special thanks go to my family for their encouragement and support in all my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURES

LIST OF SYMBOLS

a	:	Theormodynamic probability
AR(1)	:	Autoregressive model of order 1
ARMA	:	Autoregressive moving average
ANN	:	Artificial neural network
b	:	Shape parameter for generalized logistic distribution
C	:	Covariance matrix
DSI	:	State Hydraulic Works
EIEI	:	Electrical Power Resources Survey and Development Administration
f(x)	:	Probability density function
FGN	:	Fractional Gaussian noise models
g_1	:	Skewness coefficient for sample
GAR(1)	:	Gamma autoregressive model of order 1
GL	:	Generalized logistic distribution
H_j	:	Input to the j. hidden node
H(X)	:	Marginal entropy
H(X,Y)	:	Joint entropy
H(X/Y)	:	Conditional entropy
HO_j	:	Hidden node output
k	:	Shape parameter for gamma distribution
KHGD	:	General Directarate of Rural Services
L	:	Likelihood function

ℓnL	:	Log-likelihood function
LS	:	Least squares
m	:	Station number
ML	:	Maximum likelihood
MML	:	Modified maximum likelihood
MOM	:	Method of moment
n; N	:	Sample size
N(m)	:	Net information
p	:	Shape parameter for Weibull distribution
		lag of the model
PE	:	Processing elenment
p(x)		Probability distribution function
Q-Q	:	Quantile-quantile plot
r_1	:	Lag-one autocorrelation coefficient for sample
R(m)	:	Information received
R_{we}	:	Correlation coefficient
RMSE	:	Root mean square error
R-R	:	Rainfall-runoff
RWES	:	Critical values of Shapiro-Wilk test
s	:	Boltzman coefficient
s^2	:	Variance for sample
s_τ	:	Periodic standard deviation
S	:	Entropy
S(m)	:	Information transmitted
SAC-SMA	:	Sacromento soil moisture accounting
SE	:	Sum of squares
T(X,Y)	:	Transinformation
U_j	:	Independent identically distributed with uniform distribution

w	:	Interconnection weight
$\bar{x}$	:	Mean for sample
x_i	:	Streamflow during time interval
x_{max}	:	Maximum value of the flows
x_s	:	Standardized series
$x_{v,\tau}$	:	Monthly time series
x_τ	:	Periodic mean
$y_{v,\tau}$	:	Deseasonalized monthly series
t	:	Time
t_i	:	Expected values of the ith standardized order statistics
z_i	:	Intractable terms
α	:	Scale parameter
β_1	:	Skewness coefficient
β_2	:	Kurtosis coefficient
ε_i	:	Independent random variables or error terms or residuals
ϕ	:	Autoregression coefficient
Γ	:	Gamma function
Ψ	:	Digamma function
γ	:	Skewness coefficient for population
$\hat{\gamma}$	:	Unbiased estimator for the skewness
γ_ε	:	Skewness coefficient for residual
ζ	:	Independent variable
λ	:	Location parameter
μ	:	Mean for population
ρ_1	:	Lag-one autocorrelation coefficient for population
σ^2	:	Variance for population

PART I

EVALUATION OF STREAMFLOW DATA BY USING ENTROPY METHOD

CHAPTER 1

INTRODUCTION

1.1 GENERAL INFORMATION

Entropy concept can be best explained by spontaneous processes. A spontaneous process is a physical change that occurs by itself. It requires no continuining outside agency to make it happen. For example, a rock at the top of a hill rols down. Heat flows from a hot object to a cold one. An iron object rusts in moist air. These processes occur spontaneously, or naturally, without requring an outside force or agency. They continue until equilibrium is reached. If these processes happen in the opposite direction, they would be nonspontaneous. Such that the rolling of a rock uphill by itself is not a natural process; it is nonspontaneous. The rock could be moved to the top of the hill, but work would have to be expended. Heat

can be made to flow from a cold to a hot object, however a heat pump is needed. Rust can be converted to iron, but the process requires chemical reactions used in the manufacture of iron from its ore (iron oxide) (Ebbing and Gammon, 1999).

The second law of thermodynamics provides a way to answer questions about spontaneity of a reaction. The second law is expressed in terms of quantity called as entropy. Entropy is a thermodynamic quantity that is a measure of molecular disorder, or molecular randomness. As a system becomes disordered, the positions of the molecules become less predictable and the entropy increases. For example, while the entropy of a substance is lowest in the solid phase, it is highest in the gas phase.

The molecules of a substance in solid phase continually oscillate, creating an uncertainty about their positions. These oscillating molecules become completely motionless when absolute temperature is zero. There is no uncertainty about the state of the molecules at that instant. Therefore, from a microscopic point of view, the entropy of a system increases whenever the molecular randomness or uncertainty of a system increases (Çengel , 1997).

Boltzmann gave a new definition for entropy concept by analyzing microscopic states of a thermodynamic system (McMurry and Fay, 2001). Boltzmann's definition is related to the total number of possible microscopic states of that system. This relation is expressed as (Çengel , 1997):

$$S = s\ \ell n\, a \qquad (1.1)$$

where S is the entropy, s is the Boltzmann constant and a is the thermodynamic probability.

Shannon (1948) adopted Boltzmann's definition by probability distribution. He described that entropy is the amount of uncertainty in any probability distribution. Thus, entropy concept can be used as a measure of uncertainty and indirectly as a measure of information in probabilistic terms. He considered the transmission of signals through a communication channel as to be a stochastic process. He expressed the concept of information as "entropy" since his mathematical formula for the concept is similar to the entropy function defined in statistical mechanics.

According to Shannon, information is accomplished only when there is uncertainty about an event. This uncertainty points out the presence of alternative results the event may assume and the action of making selections among them. Alternatives with a high probability of occurrence convey little information and vice versa. So, the probability of occurrence of a certain alternative is the measure of uncertainty or the degree of expectedness of a sign, symbol or number. It is this uncertainty that Shannon refers to as "entropy".

When a signal is sent in a communication process, it assumes a certain value among the original series of alternatives; its uncertainty is reduced, thus it brings information as much as its uncertainty is removed. So, the information gained is indirectly measured as the amount of reduction of uncertainty or of entropy. According to Shannon, signals must have a "surprise value" to create information. If this is not valid, signals convey no information.

Sampling data in hydrology is basically a way of communicating with the natural system which is uncertain prior to the making of any observation. Each collected sample actually represents a signal from the natural system. Redundant information does not help us to reduce the uncertainty further; it only increases the costs of obtaining data. On the basis of this analogy, a methodology based on the entropy concept of information theory has been developed for the evaluation of hydrological data networks (Özkul, 1996). Therefore, entropy is a measure of the degree of uncertainty of random hydrological processes. Since the reduction of uncertainty by means of making observations is equal to the amount of information gained, the entropy criterion indirectly measures the information content of a given series of data. Once the statistical structure of a process is known, its entropy can be computed and expressed in specific units (bits, napier or decibel) (Harmancıoğlu, 1981).

1.2 LITERATURE SURVEY

Amorocho, J. and Espildora, B., (1973), considered that the entropy concept, as defined by Shannon, gave satisfactory results in the comparison between various mathematical models developed for the same hydrological process and in the selection of the most appropriate model.

The concept of a hydrologic network as a communication channel which is designed for transmitting hydrologic information was introduced in Caselton, W. F. and Husain T., (1980).

Harmancıoğlu, N.B. and Yevjevich, V., (1987), carried out studies using entropy method on monthly observed data of a highly polluted river basin.

They used entropy-based measures in their study to evaluate the goodness of information transfer by regression. The results of their study have basically revealed that the association between most of the water quality variables is insignificant.

Husain, T., (1989), presented a simple methodology, using the entropy concept, to estimate regional hydrologic uncertainty and information at both gauged and ungauged grids in a basin. The computation formulas of single and joint entropy terms depend on single and multivariable probability density functions were derived for the gamma distributions.

Yang, Y. and Burn, D. H., (1994), developed a new methodology for data collection network design. The approach employed a measure of the information flow between gauging stations in the network which was referred as the directional information transfer. Non-parametric estimation was used to approximate the multivariate probability density functions which were required in the entropy calculations. The directional information transfer was found useful in a network study to measure the association between gauging stations.

Harmancıoğlu, N.B., Fistikoglu, O. and Özkul, S., (2003), discussed an entropy-based approach for the assessment of combined spatial/temporal frequencies of monitoring networks. The results were demonstrated in the case of water quality data observed along the Mississippi River in Louisiana. The authors emphasized that the entropy method used in this study was best to utilize different techniques in combination and to investigate network features from different perspectives before a final decision was made for network assessment and redesign.

1.3 SCOPE OF THIS STUDY

In recent years, it is realized that the reserves of the world's water resources with regard to both quantity and quality are limited. In order to enhance the economic situation of a developing country like Turkey, the water resources systems should be planned much more effectively. Water resources engineers should determine the amount of existing water potential from the observed values of hydrological variables measured at different points in time and space. The duty of the engineer is to extract the maximum amount of information conveyed through these data which are used in the design and operation of water resources systems. In Turkey, these measurements are being conducted by mainly Electrical Power Resources Survey and Development Administration (EIEI), State Hydraulic Works (DSI) and General Directorate of Rural Services (KHGD). The gauging stations of EIEI are mainly located on the main rivers of large catchments whereas the ones of DSI are generally installed on the main streams and their tributaries. Some stations operate for relatively short time and they are closed as soon as their functions are over.

The above mentioned three governmental agencies collect data from their stream gauging stations. These data are published yearly for corresponding water year. Although there are a lot of collected hydrologic data useful information conveyed by these data are insufficient, which makes the data redundant or unuseful. Therefore, to prevent the collection of unnecessary data, it is necessary to monitor the performance of the existing networks with respect to cost-efficiency. The result of such an evaluation should then lead to redesign stream-gauging network for assuring an optimal network.

Study of literature survey indicated that entropy method gives reliable means for evaluating the performance of the existing stream-gauging networks. According to these studies, entropy method can be used to select appropriate stations so as to avoid redundant information. That is, entropy method is suitable for use in the design of sampling stations. The concept of entropy methods for univariate, bivariate and multivariate cases are introduced in **Chapter 2**.

In this study, two entropy methods are presented to design stream-gauging network according to the importance of the information level of stations. The first entropy method is based on the combination of stations with the least transinformation and was developed with normal and lognormal distributions. The second method is based on ranking the stations and applying normal, log-normal and gamma distributions.

The aim of this study is to investigate the effect of distribution types on the two entropy methods. To achieve this objective, these entropy methods were applied under different distributions for the annual observations of five runoff stations in the Kızılırmak basin. The results are given in **Chapter 3**. It was found out that rating positions of the selected stations were changed for each distribution type. This indicates that the designer should be very careful in selecting the type of distribution that he (or she) will use in the calculations.

CHAPTER 2

ENTROPY METHODS

2.1 INTRODUCTION

The methodology based on entropy principle is considered in this study due to the advantages of the method in network design problems. In the following sections, the brief introduction is presented to describe the entropy principle. Therefore, the derived mathematical formulations for univariate, bivariate, multivariate and continuous processes are given from the principles of information theory.

2.2 ENTROPY CONCEPT FOR UNIVARIATE CASE

Marginal entropy is the measure of the total amount of uncertainty or it is the indirect measure of the total amount of information content of a single process, X. According to information theory, the amount of uncertainty reduced is equal the amount of information gained. Thus H(X) is delimitated as "marginal entropy" of X. The entropy of a discrete random variable X with N elementary events of probability defined in information

theory is calculated using the appropriate distribution function from Eq. (2.1).

$$H(X) = s\sum_{n=1}^{N} p(x_n)\ell og[1/p(x_n)] \tag{2.1}$$

where the probability $p(x_n)$ is based on the empirical frequency of variable X. The entropy concept of H(X) is defined in (bits) for logarithms to the base 2, in (napier) to the base e, or in (decibel) to the base 10. The value, s can be taken as 1 (Harmancıoğlu, 1981). The probabilities $p(x_n)$ can be approximated as:

$$p(x_n) = f(x_n)\Delta x \tag{2.1a}$$

in which Δx intervals are chosen to be sufficiently small and f(x) is the density function of any distribution (Ang and Tang, 1975).

2.3 ENTROPY CONCEPT FOR BIVARIATE CASE

When two random processes X and Y occur at the same time, the total entropy or the total amount of information conveyed by these independent random variables is equal to the sum of their marginal entropies:

$$H(X,Y)=H(X)+H(Y) \tag{2.2}$$

When significant stochastic dependence exists between variables X and Y, the total entropy is less than the total entropy of Eq. (2.2):

$$H(X,Y) = s\sum_{n=1}^{N}\sum_{n=1}^{N} p(x_n, y_n)\ell og[1/p(x_n, y_n)] \quad (2.3)$$

which $p(x_n,y_n)$ is a probability of an outcome for X and Y. The total entropy is also expressed as (Harmancıoglu, 1981):

$$H(X,Y) = H(X) + H(Y/X) \quad (2.4)$$

$$H(X,Y) = H(Y) + H(X/Y) \quad (2.5)$$

in which H(X/Y) or H(Y/X) is the conditional entropy. The concept of "conditional entropy" has to be introduced as a function of conditional probabilities of X and Y with respect to each other:

$$H(X/Y) = -s\sum_{n=1}^{N} p(x_n, y_n)\ell ogp(x_n/y_n) \quad (2.6)$$

$$H(Y/X) = -s\sum_{n=1}^{N} p(x_n, y_n)\ell ogp(y_n/x_n) \quad (2.7)$$

where $p(x_n/y_n)$ or $p(y_n/x_n)$ (n=1,2,...,N) defines the conditional probabilities of the values x_n and y_n. The conditional entropy H(X/Y) defines the amount of uncertainty that still remains in X, even if Y is known, and the same amount of information can be gained by observing X.

Therefore, the total entropy H(X,Y) of dependent X and Y will be less than the total entropy if the processes were independent:

$$H(X,Y) < H(X) + H(Y) \quad (2.8)$$

Transinformation is another entropy measure which measures the redundant or mutual information between X and Y. It is described as the difference between the total entropy and joint entropy of dependent X and Y (Harmancıoglu et. al., 2003).

$$T(X,Y)=H(X)+H(Y)-H(X,Y) \tag{2.9}$$

Since transinformation represents the amount of information that is repeated in X and Y, the total uncertainty is reduced in the amount of T(X,Y) which is common to both processes when stochastic dependence exists between X and Y. In other words, transinformation defines the amount of uncertainity that can be reduced in one of the processes when the outcomes of the other processes are known.

By replacing the term H(X,Y) in Eq. (2.9) with its definition given in Eq. (2.4) or (2.5), transinformation can be formulated as:

$$T(X, Y) =H(Y) – H(Y/X) \tag{2.10}$$

$$T(X, Y) =H(X) – H(X/Y) \tag{2.11}$$

Transinformation and the other concepts of the entropy, always assumes positive values:

$$T(X, Y) \geq 0 \tag{2.12}$$

When two variables have no information in common, that is when they are independent of each other; it is obvious that transinformation equals to zero (Harmancıoğlu, 1981; Özkul, 1996).

2.4 ENTROPY CONCEPT FOR CONTINUOUS CASE

The formulation of entropy concept introduced in the previous sections for any random variable may be applied to hydrological variables which are basically stochastic in nature. However hydrological process is generally represented by a continuous random variable and the probability distribution function of the variable is assumed to be known. Therefore, the summation procedures of Eq. (2.1) are usually replaced by integrals:

$$H(X) = \int_{-\infty}^{\infty} f(x_n) \log[1/f(x_n)] dx \tag{2.13}$$

Similarly the total entropy of X and Y and the conditional entropy of X with respect to Y can be expressed for continues case as:

$$H(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_n, y_n) \log[1/f(x_n, y_n)] dxdy \qquad \text{and} \tag{2.14}$$

$$H(X/Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_n, y_n) \log f(x_n / y_n) dxdy\,. \tag{2.15}$$

2.5 PROBABILITY DISTRIBUTIONS

Among a number of probability density functions, only those which have bivariate density function in literature are described in subsequent sections.

2.5.1 Normal Probability Distribution

The normal probability density function of a random variable X is obtained as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{2.16}$$

where μ is the mean and σ is the standard deviation of the sample being equal to population descriptions. Eq. (2.16) is usually symbolized by N(μ,σ) (Ang and Tang, 1975).

The joint probability function of the normal distribution is known stated as:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-q/2} \quad , \quad -\infty < x < \infty \, , -\infty < y < \infty \tag{2.17}$$

where

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right] \quad (2.17a)$$

where X is $N(\mu_x,\sigma_x^2)$, Y is $N(\mu_y,\sigma_y^2)$ and ρ is the correlation coefficient of X and Y.

2.5.2 Gamma Distribution

The probability density function of gamma distribution is:

$$f(x,\alpha,k) = \frac{1}{\alpha\Gamma(k)}\left(\frac{x}{\alpha}\right)^{k-1} e^{-x/\alpha} \quad (2.18)$$

where, α is the scale parameter; k is the shape parameter, and $\Gamma(k)$ is the gamma function which is defined as (Husain, 1989):

$$\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx\,. \quad (2.19)$$

2.6 ENTROPY CONCEPT FOR MULTIVARIATE CASE

There are two methods in order to design network problems in multivariate case. The first method, Method 1, is proposed by Harmancıoğlu (1981). The objective of this method is to minimize the transinformation by an appropriate choice of the number of monitoring stations by stochastic approach in spatial orientation in order to design network stations. The combination of stations with the least transinformation reflects the variability of the quality variable along the river without producing redundant information. Such an approach

foresees the monitoring of a variable at points where it is the most variable or the most uncertain. Accordingly, existing sampling sites can be sorted in the order of decreasing uncertainty or decreasing informativeness. Thus, the first station is the one where the highest uncertainty occurs about the variable. The following stations serve to reduce this uncertainty further so that the last station brings the least amount of information.

On the other hand, a new concept of entropy which is emphasized as a Method 2 was developed for normal and log-normal distributions by Markus et al., 2003. The entropy approach is applied for information theory to evaluate stations through their information transmission to and from other stations.

The following procedures for both Method 1 and Method 2 are applied to select the best combination of stations for multivariate case.

Method 1:

The stochastic dependence between two processes causes their marginal entropies and the total entropy to decrease. The same is true for more than two variables which are stochastically dependent to each other.

For the multivariate case, the total entropy of M stochastically independent variables X_m (m=1,…,M) is:

$$H(X_1, X_2, ..., X_M) = \sum_{m=1}^{M} H(X_M) \qquad (2.20)$$

If significant stochastic dependence occurs between the variables, the total entropy has to be expressed in terms of conditional entropies added to the marginal entropy of one of the variables (Özkul, 1996):

$$H(X_1, X_2, ..., X_M) = H(X_1) + \sum_{m=2}^{M} H(X_m | X_1, ..., X_{m-1}) \quad (2.21)$$

As it is mentioned, entropy is a function of probability distribution of a process. Therefore, the multivariate joint and conditional probability distribution functions of M variables should be determined to compute the related entropies:

$$H(X_1, X_2, ..., X_M) = -\int_{-\infty}^{\infty} .. \int_{-\infty}^{\infty} f(x_1, ..., x_M).\ell ogf(x_1, ..., x_M) dx_1 dx_2 ... dx_M \quad (2.22)$$

$$H(X_m | X_1, ..., X_{m-1}) = -\int_{-\infty}^{\infty} .. \int_{-\infty}^{\infty} f(x_1, ..., x_M).\ell ogf(x_m | x_1, ..., x_{m-1}) dx_1 dx_2 ... dx_m . \quad (2.23)$$

For a single process, the marginal entropy defined by Eq. (2.1) represents the total uncertainty of the variable without removing the effect of any serial dependence. Nevertheless, if the i^{th} value of variable X or x_i is significantly correlated to values x_{i-k}, k being the time lag, knowledge on these previous values x_{i-k} will make it possible to predict the value of x_i. In this case, the marginal entropy of X reduces (Harmancıoğlu, 1981).

The next step in the computation of total, marginal or conditional entropies is to determine the type of probability distribution function which best fits the analyzed process. Harmancıoğlu (1981) proposed the multivariate normal or log-normal probability distribution functions because of

simplicity in the mathematical computations. If a multivariate normal distribution is assumed, the joint entropy of X is obtained using Eq. (2.24) (Harmancıoğlu, 1981):

$$H(X) = (M/2)\ell n 2\pi + (1/2)\ell n|C| + M/2 \tag{2.24}$$

where M is the number of variables and $|C|$ is the determinant of the covariance matrix C. Eq. (2.24) gives a single value for the entropy of M variables and the unit of entropy is napier since logarithms are taken to the base e. If logarithms of observed values are taken, the same procedure can be applied for log-normal distribution.

In the above formula, the covariance matrix C involves the cross covariances, C_{ij} of M different variables:

$$C = \begin{bmatrix} C_{11} & C_{12} & . & . & . & C_{1M} \\ C_{21} & C_{22} & . & . & . & C_{2M} \\ . & . & . & . & . & . \\ C_{M1} & C_{M2} & . & . & . & C_{MM} \end{bmatrix}_{(M*M)} \tag{2.25}$$

For a single variable, covariance matrix includes the autocovariances as a measure of the serial dependence within the process. When both the serial and cross covariance are considered, the matrix includes both the auto and cross covariances (Harmancıoğlu, 1981 and Özkul, 1996):

$$C = \begin{bmatrix} C_{11}(0) & \dots & C_{11}(K) & \dots & C_{1M}(0) & \dots & C_{1M}(-K) \\ C_{11}(K) & \dots & C_{11}(0) & \dots & C_{1M}(K) & \dots & C_{1M}(0) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ C_{M1}(0) & \dots & C_{M1}(-K) & \dots & C_{MM}(0) & \dots & C_{MM}(K) \\ C_{M1}(K) & \dots & C_{M1}(0) & \dots & C_{MM}(K) & \dots & C_{MM}(0) \end{bmatrix}_{[M*(K+1)][M*(K+1)]} \quad (2.26)$$

The covariance matrix given in Eq. (2.26), is a nonsingular, positive-definite and symmetric matrix.

Eq. (2.24) can also be used in the calculation of conditional entropies as the difference between joint entropies for three variables:

$$H(X|Y,Z) = H(X,Y,Z) - H(Y,Z) \quad (2.27)$$

Consequently in Method 1, existing stations in a basin are listed in the order of priority. The benefits for each combination of sampling sites are measured in terms of the least transinformation or the highest conditional entropy produced by that combination. In a manner of this, addition or elimination of new stations lead to decrease or increase in transinformation and conditional entropies (Özkul, 1996).

Method 2:

Marcus et al. (2003), defined a new concept of entropy. This is a fractional reduction of entropy of X by R(X, Y):

$$R(X,Y) = \frac{T(X,Y)}{H(X)} \quad (2.28)$$

which also can be viewed as a reduction of uncertainty of X if Y is known, or information received by X from Y. Similarly, the information sent (transmitted) from X to Y is defined as:

$$S(X,Y) = \frac{T(X,Y)}{H(Y)}. \tag{2.29}$$

Eqs. (2.28) and (2.29) which describe the relationship between two variables, X and Y are adapted to the network of stream gauges (Markus et al., 2003). Using these equations information received and sent at station m is defined as:

$$R(m) = R(X(m),\ \hat{X}(m)) \tag{2.30}$$

$$S(m) = S(X(m),\ \hat{X}(m)) \tag{2.31}$$

where X(m) represents the data at site m. The quantity, $\hat{X}(m)$ at station m is obtained by multiple linear regression as

$$\hat{X}(m) = a(m) + \sum_{j=1}^{M-1} Y_j(m).b_j(m) \tag{2.32}$$

where $Y_j(m)$ is a matrix of data from all other stations, a(m) and b(m) parameters of the multiple regression between site m and all other sites and M is the number of stations. As the relations between data at different sites are found to be linear or close to linear, this assumption of linearity is deemed appropriate.

In this method the concept of entropy is used to determine the stations with the highest amounts of S(m) and R(m). If R(m) is large relative to other stations, it indicates that the station denoted as m receives a lot of information. On the other hand, stations sending more information, having larger S(m), are considered to be more valuable and to remain active. Finally, the net information transfer, N(m), is defined as the difference between S(m) and R(m):

$$N(m) = S(m) - R(m) \tag{2.33}$$

Stations with positive N(m) are considered to be more valuable in regional analysis. If the number of stations in the network is to be reduced, such a station is more likely to be retained in the network than a station with a negative N(m) (Markus et al.,2003).

Method 2 can be applied for normal, log-normal, and gamma distributions. The marginal and joint entropy terms are calculated using the Eqs. (2.1) and (2.3). The computation of these terms for normal and log-normal distributions is straightforward. Appropriate probability density functions of the distributions are then incorporated to the related equations. In the case of gamma distribution, the parameters; α and k can be calculated using the derived method of moment estimators:

i. The expected value of X: $\mu_1(X) = \alpha k$ (2.34a)

ii. The variance of X: $\mu_2(X) = \alpha^2 k$ (2.34b)

The marginal entropy of the gamma probability density function was defined as (Husain, 1989):

$$H(X) = -(\lambda-1)\,\psi(k) + \Gamma(k+1)/\,\Gamma(k) + \ell n\ (\alpha\Gamma(k)) \tag{2.35}$$

where $\psi(k)$ is the digamma function:

$$\psi(k) = \partial / \partial k\left(\ell n \Gamma(k)\right). \tag{2.36}$$

Due to limitations in the derivation of bivariate gamma distribution functions and complexities in their mathematical computations, application of bivariate gamma distribution function is very limited. However, the bivariate gamma distribution, as proposed by Husain (1989), can be transformed to normalized variates z and w. The information transmission relationship is defined by the Eq. (2.37a) and (2.37b):

$$1\Big/\sqrt{2\pi}\int_{-\infty}^{z} e^{-0.5t^2}\,dt = \int_{0}^{X} f(t;\alpha_x, k_x)dt \tag{2.37a}$$

$$1\Big/\sqrt{2\pi}\int_{-\infty}^{w} e^{-0.5t^2}\,dt = \int_{0}^{Y} f(t;\alpha_y, k_y)dt \tag{2.37b}$$

In the above expressions, X and Y are variables with univariate gamma distributions with parameters (α_x, k_x) and (α_y, k_y), respectively, z and w are normalized variates of X and Y, respectively, with a mean of zero and standard deviation of unity. If ρ_{zw} is the correlation coefficient between z and w, then the information transmitted by variable Y about X, i.e., T(X,Y) or by variable X about Y, i.e., T(Y, X) is simplified as (Husain, 1989):

$$T(X, Y) = T(Y, X) = -\frac{1}{2}\ell n\left(1-\rho_{zw}^2\right). \tag{2.38}$$

CHAPTER 3

CASE STUDY ON KIZILIRMAK BASIN

3.1 INTRODUCTION

Streamflow gauging stations located on the Kızılırmak River are selected as the data set used in this study for application of the entropy methods, since more water monitoring stations exist with longer length of historical records. In fact Kızılırmak River is the longest river in Turkey. It has a length of 1355 km and the basin covers an area of 122 277 km^2. The river originates around Kızıldağ region in Sivas's sub-province, Zara and flows through Sivas around Kayseri, Nevşehir, and Kırşehir Provinces up to the reservoir of Hirfanlı Dam. After passing Kesikköprü and Kapulukaya Dams and Kırıkkale Province, it joins with the tributaries Acı Creek which collects the water of Çankırı Region, Delice River coming from Yozgat region and the Devres Creek at the north. Meanwhile it makes a wide curvature toward the north when it flows in the direction of the west. After joining with the Gökırmak Creek, it enters the reservoir of Altınkaya Dam. The river also feeds the reservoir of Derbent Dam which is located at the downstream of Altınkaya Dam. Eventually Kızılırmak reaches the Black Sea at the north of Bafra.

The Kızılırmak River is extremely rich with respect to soil and water resources. In order to perform measurements, DSI and EIE have constructed a number of gauging stations on the main river and the tributaries.

Figure 3.1 shows the locations of the selected streamflow gauging stations. As seen from the Figure 3.1, the EIE 1501 is located at the upstream of Hirfanlı, Kesikköprü and Kapulukaya Dams. The other station, EIE 1541 is located on the tributary (Delice) of the main river, which is the most important stream of Kızılırmak River. The remaining stations, EIE 1503, EIE 1528 and EIE 1536 are located at the downstream of these dams (Figure 3.1).

The above mentioned gauging stations within the Kızılırmak Basin have been monitored since 1955 by EIE. The available records at the existing five runoff stations cover a period of 41 years between the years 1955 and 1995 (EIEI water year books).

Since the three streamflow stations, EIE1503, EIE 1528 and EIE 1536 have been affected by existing dams, records from these stations are converted to natural streamflow characteristics. The streamflow gauging station (i.e. EIE 1501) located on the upstream of these dams is used as the reference to obtain the natural values (unaffected form) of the downstream streamflow characteristics at these three stations. For this reason, seasonal correlations of the streamflow values, which has been observed before the construction of the existing structures, are obtained both for the affected and the unaffected gauging stations. Using these seasonal correlation coefficients,

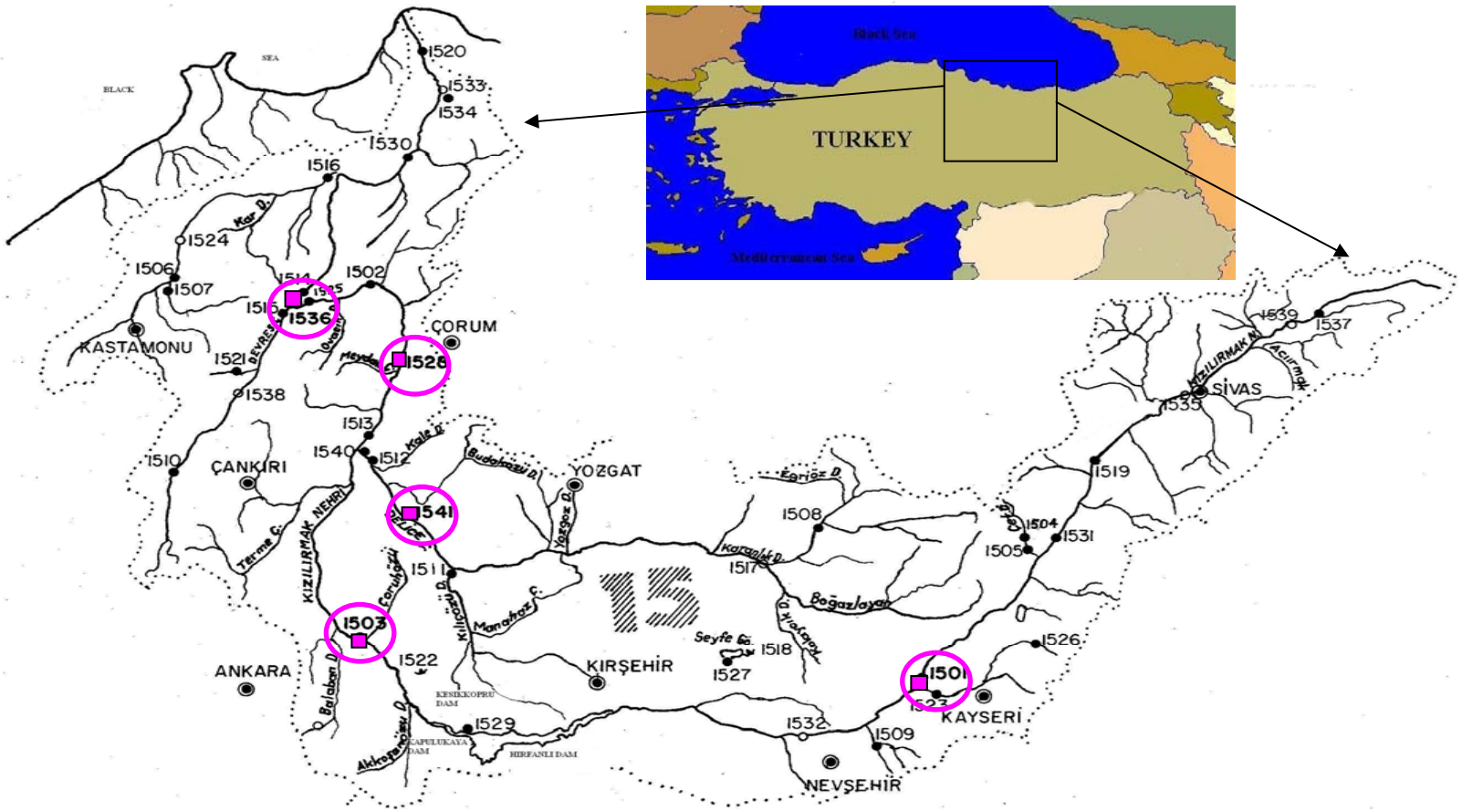

Figure 3.1 Kızılırmak Basin and Location of the Streamflow Gauging Stations

streamflow values of the affected gauging stations are adjusted by regression equation in order to obtain the natural records. The inflow and outflow data of the dam reservoirs are collected for processing from The State Hydraulic Works (DSI). It is used to obtain the change in storage of the reservoirs (±ΔS) of the dams so that the corrected streamflow values can be checked.

After obtaining (±ΔS), these values are compared with the differences of affected and natural streamflows in order to find out whether there is a correlation between them. The correlation results obtained from statistical analysis are given in Table 3.1. As shown in this table, the affected flows converted into natural flows are assumed to be acceptable.

Table 3.1. Comparasion with the Differences of Affected and Natural Streamflows

Station no.	The Name of the Hydraulic Structures	Correlation coefficient, R^2
1503	Hirfanlı, Kesikköprü and Kapulukaya	0.85
1528	Hirfanlı, Kesikköprü and Kapulukaya	0.83
1536	Hirfanlı, Kesikköprü and Kapulukaya	0.84

3.2 CASE STUDY FOR ENTROPY METHODS

3.2.1 Method 1 for Normal and Log-normal Distributions

The methodology described for Method 1 in Section 2.6 is applied to the annual streamflow values of five stream gauging stations in Kızılırmak River according to the normal and log-normal distributions. Although the period of observation varied for each station, a common period of 41 years between 1955 and 1995 is considered for all stations. The procedure of Method 1 is summarized below to select the best combination of stations based on minimum transinformation principle of normal distribution. Taking the logarithm of data set, the same procedure is also followed to log-normal distribution case.

i. Since five stream gauging stations are considered in this study, the data set for each station is represented by X_m where m (m=1,…,M) represents the station number.

ii. The marginal entropy H (X_m) of the variable for each station is computed first by using Eq. (2.24) where M is replaced by 1. As it can be seen from Table 3.2, the marginal entropy value of EIE 1536 streamflow gauging station is greater than the marginal entropy value of the other stations, that means, the highest uncertainty occurred about the variable at this location. Hence station EIE 1536 is selected as the first priority station, X_1, to continue its observations.

iii. Next, the selected station EIE 1536 is coupled with every other station in the network to select the pair that leads to the least transinformation. EIE 1541 station that fulfills this condition is marked as the second priority location X_2. A pair of stations is selected that has the highest joint entropy and the least transinformation. Accordingly, these stations produce the highest amount of information when they operate together.

iv. As the third step, the conditional entropies and transinformations of the (X_1, X_2) pair with every other station in the network are computed to select a triple station with the least transinformation.

v. The same procedure is continued by considering successive combinations of 4 and 5 stations and selecting the combination that produces the least transinformation and minimum redundant information.

vi. The stations are ranked according to their priority orders. For example, higher rank (r=5) represents the first priority order which means that this station is necessary for this network to remove the uncertainty.

The priority orders of the selected stations using Method 1 for normal and log-normal distributions are presented in Table 3.2 and Table 3.3, respectively. As it is seen in Table 3.2 station EIE 1536 is the most important station to continue its observations with rank 5, and station EIE 1503 is the least important station. On the other hand in Table 3.3 for log-normal distribution, station EIE 1541 becomes the most important station.

Table 3.2. Selection of Sampling Stations for Normal Distribution

Station no.	Station Added	Marginal Entropy	Joint Entropy	Conditional Entropy	Transinfor-mation	Rank (r)
	(M)	(napier)	(napier)	(napier)	(napier)	-
1536	1	8.74	8.74	-	-	5
1541	2	5.88	12.82	6.94	1.795	4
1501	3	7.56	18.06	10.50	2.319	3
1528	4	8.47	22.89	14.42	3.630	2
1503	5	7.97	25.83	17.85	5.039	1

When each distribution type of annual streamflow series are assumed to be normally distributed, EIE 1536 which is the most downstream station located on the main river is selected as the most priority station using Method 1. Moreover the applied procedure selects the most upstream station on the tributary as the second station in the priority list. Accordingly, EIE 1536 and EIE 1541 constitute the pair with the least amount of redundant information. The third location is station EIE 1501 which is the most upstream station on the main river. As it can be seen in Table 3.2, the joint entropy and transinformation values increase contributing the other station at the network. The percentages of redundant information varied with the addition of each new station to the combination. Planner can decide discontinued stations in the network according to amount of transinformation which is determined beforehand.

Table 3.3. Selection of Sampling Stations for Log-normal Distribution

Station no.	Station Added	Marginal Entropy	Joint Entropy	Conditional Entropy	Transinfor-mation	Rank (r)
	(M)	(napier)	(napier)	(napier)	(napier)	-
1541	1	-2.08	-2.08	-	-	5
1501	2	-2.58	-5.79	-3.22	1.14	4
1536	3	-3.06	-10.40	-7.34	1.54	3
1503	4	-2.88	-17.12	-14.24	3.84	2
1528	5	-3.08	-30.10	-27.03	9.91	1

The results of Method 1 under log normal distribution which is similar to normal distribution for computations are given in Table 3.3 for this network. It is observed from Table 3.3 that EIE 1541 and EIE 1501 stations are selected as the first and the second priority stations. As was expected, due to their locations, being in the most downstream of the basin, stations EIE 1536 and EIE 1528 have produced high redundant information. Therefore, looking at Table 3.3 it would be logical to close station EIE 1528 as it is ranked 1, and keep station EIE 1536 which is ranked 3.

However, the amount of transinformation values show differences under normal and log normal distributions. This emphasizes that the selection of an appropriate distribution type is very important before any analysis. On the other hand, rank of stations which is retained in the network are reasonable under both normal and log normal distributions.

Figure 3.2 shows station ranking using annual discharge time series for normal and log normal distributions.

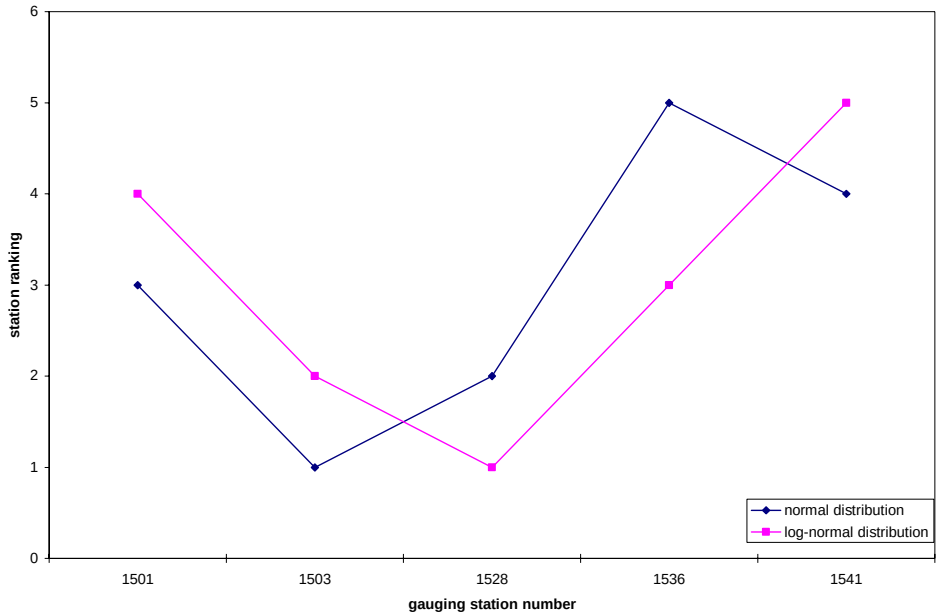

Figure 3.2. Station ranking based on minimum transinformation for normal and log-normal distributions using annual discharges

3.2.2 Method 2 for Normal and Log-normal Distributions

The Method 2, which is based on entropy principle, is also applied for Kızılırmak Basin using Eq. (2.1) for marginal entropy, Eq. (2.3) for joint entropy and Eq. (2.10) or Eq. (2.11) for transinformation. The normal probability density function given by Eqs. (2.16) and (2.17) are substituted in mathematical formulations of these measures which are based on entropy. Later Eqs. (2.28) and (2.30) are used to compute the total information received by a station m; R (m). Eqs. (2.29) and (2.31) are then used to compute the total information sent by a station m; S (m). Finally, Eq. (2.33) is used to compute the total net information transfer for station m; N (m).

As it has been mentioned above, if logarithms of observed values are taken, the same procedure can be used for log-normal distribution. The information transfer parameters S (m), R (m) and N (m) as well as the station ranks, based on these parameters, are shown in Table 3.4 and Figure 3.3 for normal and in Table 3.5 and in Figure 3.3 for log normal distributions.

The stations having the lowest rank (r =1 or r =2) are less important in the information transfer process. It is not necessary to continue observation at these stations. On the other hand the stations having higher ranks (r =4 or r =5) should be retained in the network according to this method.

Table 3.4. Station Ranking According to Information Transmitted, S(m) Information Receieved, R(m) and Net Information, N(m) for Normal Distribution

Station no.	Information transfer			Rank (r)		
	Send	Received	Net	Send	Received	Net
	S(m)	R(m)	N(m)	S(m)	R(m)	N(m)
	(napier)	(napier)	(napier)	-	-	-
1501	0.9229	0.9228	0.0002	2	2	2
1503	0.9251	0.9238	0.0013	3	3	5
1541	0.9181	0.9178	0.0004	1	1	4
1528	1.2347	1.2350	-0.0003	5	5	1
1536	1.0896	1.0894	0.0002	4	4	3

Table 3.5. Station Ranking According to Information Transmitted, S(m) Information Receieved, R(m) and Net Information, N(m) for Log- normal Distribution

Station no.	Information transfer			Rank (r)		
	Send	Received	Net	Send	Received	Net
	S(m)	R(m)	N(m)	S(m)	R(m)	N(m)
	(napier)	(napier)	(napier)	-	-	-
1501	-96.05	-96.91	0.139	5	5	4
1503	-358.21	-358.57	-0.358	3	3	2
1541	-118.05	-117.50	0.551	4	4	5
1528	-2320.40	-2320.92	-0.524	1	1	1
1536	-1892.09	-1892.38	-0.287	2	2	3

While EIE1536 station is selected as a first priority station by Method 1 under normal distribution assumption, the same station is selected as the third prior station which must be remained in the streamflow gauging network by Method 2 according to net information. Similarly EIE 1503 station which is chosen as the last priority station by Method 1 with normal distribution is selected as the most important station by Method 2 with normal distribution.

Nonetheless, it is observed that while the results of the two methods are changed under normal distribution assumption, the results for both methods are the same under log-normal distribution.

3.2.3 Method 2 for Gamma Distribution

The Method 2 for gamma distribution is applied using Eq. (2.35) for marginal entropy and Eq. (2.38) for transinformation. The rest of the

procedure of Method 2 for gamma distribution is exactly same as it is for the other distributions. Finally, the information transfer parameters S (m), R (m) and N (m) are shown in Table 3.6 and Figure 3.3 for gamma distribution.

Table 3.6. Station Ranking According to Information Transmitted, S(m) Information Receieved, R(m) and Net Information, N(m) for Gamma Distribution

Station no.	Information transfer			Rank (r)		
	Send	Received	Net	Send	Received	Net
	S(m)	R(m)	N(m)	S(m)	R(m)	N(m)
	(napier)	(napier)	(napier)	-	-	-
1501	0.6349	0.6350	-0.000121	1	1	1
1503	0.6903	0.6902	0.000098	2	2	5
1541	0.7200	0.7201	-0.000096	3	3	2
1528	1.0479	1.0479	0.000006	5	5	3
1536	0.9935	0.9935	0.000025	4	4	4

According to the method 2 which is obtained from gamma distribution, EIE 1503 station is the most important station in this network to be kept. Moreover EIE 1501 station which is the most upstream station is selected as the least important station for the information transfer process.

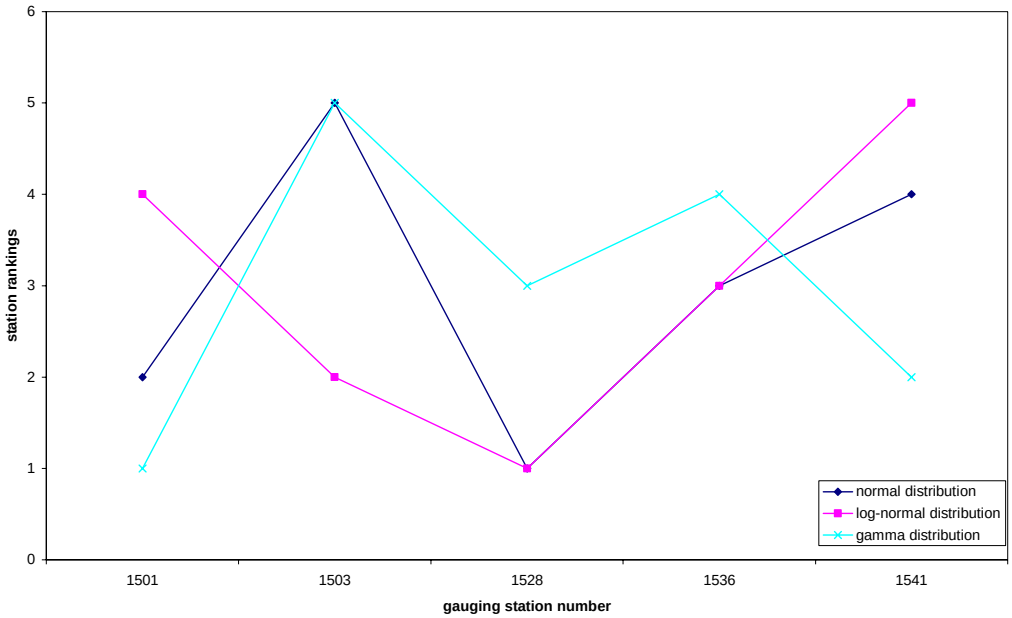

Figure 3.3 Station ranking based on N(m) for normal, log-normal and gamma distributions using annual discharges

As it was mentioned earlier, stations ranking vary for both methods under different distributions. However only for log-normal distribution the stations ranking are similar. Nevertheless, it is observed that these ordering under log normal distribution are quite reasonable if it is compared with the results of the other distributions. Since the most upstream stations located at different tributaries (EIE 1501 and EIE 1541) have the highest uncertainty, these stations are selected as the second and the third priority stations for this network as it can be expected.

CHAPTER 4

SUMMARY AND CONCLUSIONS

In this part of the study, two entropy methods are applied on a network of gauging stations located in the Kızılırmak Basin. The aim of this study is to rank the stations according to the amount of their information contribution. Once the first priority station is selected according to the marginal entropy (as explained in Section 3.2), then the next station to be combined is selected according to the minimum transinformation. The first station selected in this method is the station having the highest priority, which means that this station must be retained in the network.

The stations are also ranked by Method 2. As also used in the literature the station with the lowest rank is the least information contributing station, hence the station with low rank could be discontinued. Higher ranks indicate the stations that should be retained in the network.

For Method 1, it is required to apply the multivariate density function, whereas for Method 2 the bivariate density function of distribution is found to be adequate. Although the mathematical development of entropy is easily made for skewed distributions in bivariate cases, the development

becomes much more difficult when multivariate distributions are considered.

Because of the difficulties involved in the mathematical development of multivariate distributions, Method 1 is applied for just normal and log-normal distributions, while Method 2 is applied for normal, log-normal and gamma distributions.

In order to demonstrate the effect of the distribution type on each entropy method, ranking of stations obtained by using Method 1 and Method 2 with different distributions are summarized in Table 4.1. As it is seen from Table 4.1, the importance level of each station on the existing stream gauging network is changed for different distribution types. Rankings obtained by normal distribution showed anomalies in Method 1 and Method 2, whereas rankings obtained by log-normal distribution were consistent in the two methods.

As a result, determination of appropriate distribution for streamflow series is an important point to rely on the results which can be obtained from entropy methods. For example it is obvious that planning stream gauging network system under normal distribution for data set having right skewed distribution may not be reliable.

From this point of view, the selection of appropriate distributions for variables is the crucial part of any issue in which the optimum network system is to be planned with entropy methods. For that reason, the author emphasized that the distribution types for data series should be determined properly before applying two entropy methods.

Table 4.1. Ranking of Stations Using Method 1 and Method 2 with Different Distributions

Station no.	Entropy Method 1 Ranking based on min. transinformation		Entropy Method 2 Ranking based on net information transfer, N(i)		
	Distribution types		Distribution types		
(i)	Normal	Log-normal	Normal	Log-normal	Gamma
1501	3	4	2	4	1
1503	1	2	5	2	5
1541	4	5	4	5	2
1528	2	1	1	1	3
1536	5	3	3	3	4

PART II

MODELING OF STREAMFLOW DATA USING AUTOREGRESSIVE MODELS WITH ASYMMETRIC INNOVATIONS AND ARTIFICIAL NEURAL NETWORKS

CHAPTER 1

INTRODUCTION

1.1 GENERAL INFORMATION

Modeling streamflow data has become important to the water resources planner, since it allows him to evaluate proposed system designs more thoroughly. Early civil engineers have realized that the flow patterns of different streams vary considerably and that the history of flows in a particular stream provides a very valuable clue to the future behavior of that stream. If the flow of this year is low, it is likely, although not certain, that the flow of next year will also be lower than average. Similarly, high flows tend to follow high flows. Thus the history of a stream provides valuable information about probable future flows. In reality, a set of historical or synthetic flows for a stream is a sequence of numbers or

values produced by a random process in a succession of time intervals; such a sequence is called a time series. Several time series models such as autoregressive models (AR), fractional gaussian noise models (FGN), autoregressive moving-average models (ARMA), broken-line models, shot-noise models, model of intermittent processes, disaggregation, ARMA– Markov models and general mixture models have been proposed for modeling hydrologic data (Salas et. al., 1980).

Traditionally, AR models have been the most widely used statistical method for modeling hydrologic process, since the autoregressive form has an intuitive type of time dependence which the value of a variable at the present time depends on the values at previous times and they are the simplest models to use.

The autoregressive time series models are typically of the following form,

$$x_i = \sum_{j=1}^{p} \phi_j x_{i-j} + \varepsilon_i \quad , \qquad i=1,2,\ldots,n \tag{1.1}$$

where ϕ_j are autoregression coefficients and p is the lag of the model. In Eq. (1.1) while $\phi_j x_{i-j}$ is a deterministic part of the model, ε_i is the random component of the generating scheme.

In previous hydrological time series models, the innovations or residuals, ε_i, are generally assumed to be normally distributed N (0, σ^2). Classical autoregressive and moving average models require transformation of the original series into normal. That is to say that, if the hydrologic variable, x_i is not normal, an appropriate method is used to transform it to normal or near-normal. Thus, the distribution of ε_i becomes normal. Taking the

logarithm of the series is one of the transformation methods. Logarithmic transformation makes positively skewed sample series symmetrical or slightly negatively skewed.

In fact both of the annual and monthly streamflow series are generally positively skewed. In the past few years, AR type models with skewed marginal distributions have been developed and applied to hydrologic time series. These models are more realistic than previous ones.

In this part of the thesis, autoregressive models (AR(1)) with asymmetric innovations represented by gamma, generalized logistic (GL) and weibull distributions are introduced in **Chapter 2**. Then, alternatively the artificial neural network (ANN) model has been proposed **in Chapter 2** for modeling hydrologic time series data, since it does not require any assumption about linearity and statistical distribution. The above models are applied for the annual and monthly observations of five runoff stations in the Kızılırmak basin in **Chapter 3**.

CHAPTER 2

AUTOREGRESSIVE MODELS WITH NON-NORMAL INNOVATIONS AND ARTIFICIAL NEURAL NETWORK

2.1 HISTORICAL REVIEW

In modeling the streamflow time series, the asymmetry of the marginal distribution creates some problems. The problem of skewed streamflows or residuals has been handled by a number of methods. A widely used technique is to use transformations to render a series close to normal as said earlier (Box and Jenkins, 1976). Another approach is to find the statistical properties of the residuals. In this approach, the variable transformation is not required as do the classical models. For this aim, firstly Wilson-Hilferty transformations which were only valid for gamma distribution were used by Thomas and Fiering (1962), McMahon and Miller (1971) and O'Connel (1974). Although the coefficient of skewness can be reproduced by this approach, the underlying variable is not gamma.

During recent years, a number of non-Gausian models with AR-type correlation structure have been proposed. The simplest of this kind of

models corresponds to the so-called exponential autoregressive (EAR(1)) model (Gaver and Lewis, 1980). They showed that there was an innovation process $\{\varepsilon_i\}$ such that the sequence of random variables $\{x_i\}$ generated by the linear additive first order autoregressive scheme $x_i=\phi x_{i-1}+\varepsilon_i$ were marginally distributed as gamma variables if $0\leq\phi\leq1$. They claimed that this first order autoregressive gamma sequence was useful for modeling a wide range of observed phenomena. Properties of sums of random variables from this process were also studied.

A new exponential autoregressive model, NEAR(1) was presented by Lawrance and Lewis (1981). They used the NEAR(1) model to accommodate uniform marginal distribution by using an exponential transformation.

Obeysekera and Yevjevich (1985) reported a procedure for generation of samples of an autoregressive scheme that had an exact gamma marginal distribution with given mean, variance and skewness. They gave proper modifications in Gaver and Lewis (1980) method to produce the gamma marginal distribution with given mean, variance and skewness.

Fernandez and Salas (1986) developed and tested a new class of time series models capable of reproducing the covariance structure normally found in periodic streamflow time series under non-Gaussian marginal distribution. Specially the models could be either linear or non-linear or a combination of both and assume a gamma marginal and a lag-one autoregressive correlation structure. Five series of weekly streamflow were used for applications and comparisons of the proposed models. The results showed that the new class of gamma models compared favorably with respect to

the normal models in reproducing the basic statistics usually analyzed for streamflow simulation.

Sim (1987), considered a time series model which can be used for simulating stationary river flow sequences with high skewness and the long-term correlation structure of an ARMA(1,1). The model parameters under gamma distribution were estimated by using the method of moments. 100 sequences of monthly streamflows were simulated from this model. He observed that the simulated data bear a close resemblance to the historical data in terms of the autocorrelation structure, skewness, mean, and standard deviation.

Fernandez and Salas (1990) studied the gamma autoregressive models for streamflow simulation also. Since moment estimators are biased for dependent and non-normal variables, they emphasized that some kind of correction is needed to make them unbiased. A procedure to obtain unbiased estimators of parameters for a stationary, first order gamma-autoregressive model, capable of reproducing the mean, variance and skewness structure of the available historical streamflow data was presented in this study. Applications of the proposed procedure to annual streamflow series of several rivers were done. They claimed that the GAR(1) model was an attractive alternative for synthetic streamflow simulation.

Lawrance and Lewis (1990) introduced the idea of reversed pth- order autoregressive residuals and developed some of their properties. The use of reversed residuals was illustrated on a series of deseasonalized montly river flow data in which it was shown that there was non-linear first order autoregressive dependence.

Cığızoğlu and Bayazıt (1998), used GAR(1) model to determine the statistical run and range parameter values of the annual flow series and applied to the 10 Turkish rivers. They found out that the analyses based on the GAR(1) model emphasized that the bias adjustments was very important to obtain the run and range parameters.

A more interesting generalization of the preceding above models, at least from the hydrologic point of view, consists of using a marginal gamma distribution. In these studies, model parameters were estimated by using the method of moment (MOM). The MOM procedure does not have general properties as good as the maximum likelihood method (ML) which is another parameter estimation technique. However, ML solution is not always possible for whole distribution types. In recent years, Tiku et. al. (1996) have proposed the modified maximum likelihood (MML) procedure to estimate the model parameters under non-Gausian distributions.

Tiku et.al. (1996) considered AR(p) models in time series with non-normal innovations represented by a member of a wide family of symmetric distributions (Student's t type). Since the ML estimators are intractable, they derived the MML estimators of the parameters and showed that they were remarkably efficient, robust and powerful.

The first order autoregressive model, AR(1) with asymmetric innovations of the gamma type considered in Tiku et al., (1999a) (the generalized logistic case was also briefly discussed). The same model has been considered in Tiku et al., (2000) with symmetric non-normal innovations of the Student's t type. The simple regression model with first order autoregressive errors with asymmetric innovations was considered in Akkaya and Tiku, (2001a). The model in Tiku et. al. (1999a) is a special case

of the model in Akkaya and Tiku, (2001a) with no regression component. Turker, (2002) extended the methodology under non-normality to various independent sources of information and developed robust and efficient statistics for testing whether parameter vector remains the same from one source to another. Akkaya and Tiku (2005) also studied AR(1) models under short-tailedness and inliers.

There are also numerous studies related to the application of ANN to various problems frequently encountered in water resources. In time series analysis, stochastic models are fitted one more of the time series describing the system for purposes which include forecasting, generating synthetic sequences for use in simulation studies, and investigating and modeling the underlying characteristics of the system under study. Most of the monthly time series modeling procedures fall within the framework of multivariate autoregressive moving average (ARMA) models (Raman and Sunilkumar, 1995).

ANNs have been successfully applied in a number of diverse fields including water resources. In order to optimally fit an AR and ARMA type models to a time series, the data must be stationary and follow a normal distribution (Hipel, 1986). Lorrai and Sechi (1995) verified the possibility of utilizing ANNs to predict rainfall-runoff (R-R) when only information about the variation of the basic input variables, namely rainfall and temperature, is available. Cheng and Noguchi (1996) obtained better results modeling the R-R process with ANNs using previous rainfall, soil moisture deficits, and runoff values as model inputs, when compared with that from a R-R model. Smith and Eli (1995) applied ANNs to convert remotely sensed, spatially distributed rainfall patterns into rainfall rates, and hence into runoff for a given river basin.

Hsu et al. (1995) showed that a non-linear ANN model provided better representation of the R-R relationship of the medium-sized Leaf River basin near Collins, Mississippi, than the linear ARMAX (autoregressive moving average with exogenous inputs) time series approach or the conceptual SAC-SMA (Sacromento soil moisture accounting) model (Sorooshian et. al., 1993). In the hydrological forecasting context, recent experiments have reported that ANNs may offer a promising alternative for rainfall-runoff modeling (Zhu and Fujita, 1994; Smith and Eli, 1995; Hsu et al., 1995; Yapo et al., 1996; Shamseldin, 1997; Sajikumar and Thandaveswara, 1999; Tokar and Johnson, 1999, Tokar and Markus, 2000, Rajurkar et al., 2004), streamflow prediction (Kang et al., 1993; Karunanithi et al., 1994; Thirumalaiah and Deo, 1998; Clair and Ehrman, 1998; Zealand et al., 1999; Campolo et al., 1999; Chang and Chen, 2001; Sivakumar et al., 2002; Kişi, 2003; Castellona-Mendez et al., 2004, Moradkhani et. al., 2004), reservoir inflow forecasting (Saad et al., 1996; Jain et al., 1999), prediction of water quality parameters (Maier and Dandy, 1996), regional drought analysis (Shin and Salas, 2000), real time forecasting (Kitanidis and Bras, 1980; Thirumalaiah and Deo, 2000) and estimating evapotranspiration (Kumar et al., 2002).

2.2 PARAMETER ESTIMATION METHODS FOR AUTOREGRESSIVE MODELS

One of the main objectives of mathematical statistics is to estimate reliable model parameters through estimation methods. If the estimated parameters obtained from a sample are good, then it will be possible to extract wider information on synthetic data. The most commonly used

parameter estimation methods in autoregressive models are given below in order of increasing efficiency:

1. Method of Moments (MOM)
2. Least- Squares Method (LS)
3. Maximum Likelihood Method (ML)
4. Modified Maximum Likelihood Method (MML)

2.2.1 Method of Moments (MOM)

The method of moments (MOM) is a natural and relatively easy parameter estimation method. This method relates the derived moments to the parameters of the distribution. The relation can simply be that the mean is equal to the first moment about origin, variance is the second central moment and the coefficient of skewness is the third central moment divided by the second central moment with a power of 3/2. However, MOM estimates are usually inferior in quality and generally are not as efficient as the ML estimates, especially for distributions with large number of parameters (three or more), because higher order moments are more likely to be highly biased in relatively small samples (Haan, 1977). In general, moment estimators are inefficient. They are efficient only for normal and near-normal populations.

2.2.2 Least- Squares Method (LS)

If no distributional assumption is made about the random errors of a linear model then the LS methodology can be used in estimating the parameters of this model. That means least square method does not utilize the prior

information about the distribution of the data. This method states that the sum of squares of all deviations from the AR(1) model is to be minimized:

$$\text{minimize } SE = \sum_{i=1}^{n} \left(x_i - \phi x_{i-1} - \lambda \right)^2 = \sum_{i=1}^{n} \varepsilon_i^{\,2} \tag{2.1}$$

To obtain the minimum sum of squares, Eq.(2.1) is partially differentiated with respect to the best estimates of model parameters as λ, ϕ and α:

$$\frac{\partial \sum_{i=}^{n} \varepsilon_i^2}{\partial \lambda} = 0 \tag{2.2a}$$

$$\frac{\partial \sum_{i=}^{n} \varepsilon_i^2}{\partial \phi} = 0 \quad \text{and} \tag{2.2b}$$

$$\frac{\partial \sum_{i=}^{n} \varepsilon_i^2}{\partial \alpha} = 0 . \tag{2.2c}$$

As it is known if the series has normal distribution, the least-square estimators are equal to maximum likelihood estimators.

2.2.3 Maximum Likelihood Method (ML)

The maximum likelihood (ML) method is considered the most efficient method since it provides the smallest sampling variance of the estimated parameters. Maximum likelihood estimation begins with writing a mathematical expression known as the likelihood function of the sample

data. The likelihood expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the maximum likelihood estimators. The likelihood function is defined as a function of the unknown model parameters:

$$L = \prod_{i=1}^{n} p(\varepsilon_i, \phi, \lambda, \alpha) \tag{2.3}$$

where $p(\varepsilon_i, \phi, \lambda, \alpha)$ is a probability function of ε with λ, ϕ and α being a model parameters to be estimated.

To estimate λ, ϕ and α logarithmic L should be maximized by differentiating it with respect to each parameter and equating to zero:

$$\frac{\partial \ln L}{\partial \lambda} = 0 \tag{2.4a}$$

$$\frac{\partial \ln L}{\partial \phi} = 0 \tag{2.4b}$$

$$\frac{\partial \ln L}{\partial \alpha} = 0 \tag{2.4c}$$

which are called the maximum likelihood functions (Ang and Tang, 1975).

2.2.4 Modified Maximum Likelihood Method (MML)

The error terms have a non-normal distribution with respect to the real life situation. However LS, used to estimate unknown parameters in the model, are inefficient with non-normal cases (Tiku et.al., 1999a). Similarly

another well known procedure, which is ML procedure, becomes also unsuccessful in numerous situations since explicit solutions from the likelihood equations cannot be obtained and iterative methods have to be used. There are some difficulties in solving likelihood function iteratively. These difficulties are in general (Tiku et al., 1999a):

i. The iteration may converge very slowly,
ii. They may converge to local maximum values,
iii. They may not converge at all.

If data contain outliers, the iterations with likelihood equations might never converge. In addition the successes of the iteration methods depend on the shape of the function. For instance if function is concave, Newton-Raphson method has poor result. Steepest Ascent method may give more reliable solution than the former; however, it converges very slowly (Hamilton 1994). Thereby estimators based on iterative methods are not optimal. At this point it is obvious that a robust estimator technique while the error terms are not normally distributed is needed. Hence, the modified maximum likelihood (MML) method has been developed by Tiku (1967) and applied to some non-normal time series models. This method is based on linearization of intractable terms of the log-likelihood function for a location scale family of innovations. These intractable terms are linearized by using first-order Taylor series expansion.

This procedure can be summarized as below:

i. Express ML equations in terms of the ordered variates,
ii. Replace the intractable terms by their Taylor series expansion,
iii. Solve the resulting equations to get the MML estimators.

In fact, published time series applications of this method in the literature have shown that the estimated parameters using this method are highly efficient, robust and superior to the least squares (LS) estimators. Although this methodology applies to any location-scale distribution, three distributions are considered in this study as gamma, weibull and generalized logistic.

The procedures of three parameter estimation methods under non-normal distributions are given in detail in the following sections.

2.3 GAMMA AUTOREGRESSIVE MODELS

The first order gamma autoregressive model with the usual structure of an additive process is:

$$x_i = \lambda + \phi x_{i-1} + \varepsilon_i \tag{2.5}$$

where λ is the location parameter, x_i is the streamflow during time interval t, ϕ is the autoregression coefficient, and ε_i is the independent random variable or error terms. The only difference with the well-known AR(1) model lies on the fact that x_i or ε_i has a marginal distribution given by a three-parameter gamma density function:

$$f_x(x) = \frac{(x)^{k-1} \exp\left[-\left(\frac{x}{\alpha}\right)\right]}{\alpha^k \Gamma(k)} \tag{2.6}$$

where α and k are the scale and shape parameters, respectively and $\Gamma(.)$ is the gamma function.

2.3.1 Parameter Estimation Procedure with Method of Moments

In this procedure, x_i in Eq. (2.5) is assumed to be gamma distributed. The model which estimate the model parameters by using method of moment procedure is known in literature as GAR(1) model. Fernandez and Salas (1990) derived the following relationships between the parameters of model and the population moments of the underlying variable x_i using the method of moments (MOM) procedure:

$$\mu = \lambda + k\alpha \tag{2.7}$$

$$\sigma^2 = k\alpha^2 \tag{2.8}$$

$$\gamma = \frac{2}{\sqrt{k}} \tag{2.9}$$

$$\rho_1 = \phi \tag{2.10}$$

where μ is the mean, σ^2 is the variance, γ is the coefficient of skewness, and ρ_1 is the lag- one autocorrelation coefficient. These population moments can be estimated based on the sample data $x_1, x_2, \ldots, x_n$, using the following relationships:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.11}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.12}$$

$$g_1 = \frac{n}{(n-1)(n-2)s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 \tag{2.13}$$

$$r_1 = \frac{1}{(n-1)s^2}\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x}) \tag{2.14}$$

where $\bar{x}$ is the estimator of μ, s is the estimator of σ, g_1 is the estimator of γ, r_1 is the estimator of ρ and n is the sample size.

Since these estimators are biased for dependent and non-normal variables, some kind of correction needs to be made before using them in the system of Eqs. (2.11) - (2.14) to estimate the parameters of the GAR(1) model. Estimator of the expected value (μ) can be used without a correction factor (Fernandez and Salas,1990). These corrections are:

Correction for r_1

Wallis and O'Connell (1972) suggested the following correction to obtain an unbiased estimator of ρ_1 for an autoregressive AR(1) model:

$$\hat{\rho}_1 = \frac{r_1 n + 1}{n - 4} \tag{2.15}$$

where r_1 is the lag-one autocorrelation coefficient.

Correction for s^2

If the series is uncorrelated, the estimator of the variance of the process obtained from Eq. (2.12) is unbiased. For correlated streamflow series, an unbiased estimator of the variance can be obtained from:

$$\hat{\sigma}^2 = \frac{n-1}{n-K} s^2 \tag{2.16}$$

in which

$$K = \frac{\left[n(1-\hat{\rho}_1^2) - 2\hat{\rho}_1(1-\hat{\rho}_1^n\right]}{\left[n(1-\hat{\rho}_1)^2\right]} \tag{2.16a}$$

and s^2 and $\hat{\rho}_1$ are given by Eqs. (2.12) and (2.15), respectively.

Correction for g

The unbiased estimator of the skewness can be obtained from:

$$\hat{\gamma} = \frac{\hat{\gamma}_0}{f} \tag{2.17}$$

where

$$\hat{\gamma}_o = \frac{L g_1 \left[A + B\left(\frac{L^2}{n}\right) g_1^2 \right]}{\sqrt{n}} \tag{2.17a}$$

g_1 is given by Eq. (2.13).

$$L = \frac{(n-2)}{\sqrt{(n-1)}} \tag{2.17b}$$

$$A = 1 + 6.51n^{-1} + 20.2n^{-2} \tag{2.17c}$$

$$B = 1.48n^{-1} + 6.77n^{-2} \tag{2.17d}$$

$$f = \left(1 - 3.12\rho_1^{3.7} n^{-0.49}\right) \tag{2.17e}$$

It may be observed that, f=1 if ρ_1=0 for independent values.

Therefore, the general procedure to estimate the parameters of a stationary GAR(1) model to be used in simulation studies based on available sample series should be as follows:

i. Values of the mean; $\bar{x}$, variance; s^2, coefficient of skewness; g_1, and lag-one autocorrelation coefficient; r_1 are obtained from Eqs. (2.11)-(2.14), respectively.
ii. The unbiased autocorrelation coefficient, $\hat{\rho}_1$, variance, $\hat{\sigma}^2$ and skewness, $\hat{\gamma}$ are calculated from Eqs. (2.15)-(2.17), respectively
iii. The set of model parameters; α, k, λ and ϕ are determined by using Eqs. (2.7)-(2.10), respectively (Fernandez and Salas, 1990).

2.3.2 Parameter Estimation Procedure with Modified Maximum Likelihood

While the marginal distribution of x_i in Eq. (2.5) is assumed to be gamma under GAR(1) model, the marginal distribution of error terms ε_i is assumed to be gamma for AR(1) model in time series with asymmetric innovations. This model is generally called as AR(1) model with gamma innovations. Therefore the gamma density function (k>2) is represented by

$$G(k,\sigma): f_{\varepsilon}(\varepsilon) = \frac{\varepsilon^{k-1}\exp\left[-\left(\frac{\varepsilon}{\alpha}\right)\right]}{\alpha^{k}\Gamma(k)} \qquad 0 < \varepsilon < \infty. \tag{2.18}$$

This family represents positively skewed distributions with kurtosis $\beta_2 \geq 3$. In fact, $\beta_2=3+1.5\beta_1$; $\beta_2 = \mu_4/\mu_2^2$ and $\beta_1 = \mu_3^2/\mu_2^3$ (Akkaya and Tiku, 2001b).

The shape parameter k is not known. Here, the likelihood function of gamma distribution is:

$$L \propto \frac{1}{\alpha^n}\exp\left\{-\frac{1}{\alpha}\sum_{i=1}^{n}(x_i-\phi x_{i-1}-\lambda)\right\}\left\{\prod_{i=1}^{n}\left(\frac{x_i-\phi x_{i-1}-\lambda}{\alpha}\right)^{k-1}\right\}. \quad (2.19)$$

The ℓn-likelihood function is

$$\ell nL = Const - n\ell n\alpha - \sum_{i=1}^{n} z_i + (k-1)\sum_{i=1}^{n}\ell n z_i \quad (2.20)$$

where $z_i = \varepsilon_i (= x_i - \phi x_{i-1} - \lambda)/\alpha$. Differentiating Eq. (2.20) with respect to λ, ϕ and α and setting the derivatives equal to zero gives the following equations:

$$\frac{\partial \ell nL}{\partial \lambda} = \frac{1}{\alpha} + (k-1)\sum_{i=1}^{n} -\frac{1}{\alpha} z_i^{-1} = 0, \quad (2.21)$$

$$\frac{\partial \ell nL}{\partial \phi} = \frac{1}{\alpha}\sum_{i=1}^{n} x_{i-1} - \frac{k-1}{\alpha}\sum_{i=1}^{n} x_{i-1} z_i^{-1} = 0 \quad \text{and} \quad (2.22)$$

$$\frac{\partial \ell nL}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha}\sum_{i=1}^{n} z_i - \frac{k-1}{\alpha}\sum_{i=1}^{n} z_i z_i^{-1} = 0. \quad (2.23)$$

These equations are functions in terms of z_i^{-1} and they have no explicit solutions. Thus, they have to be solved by iterative methods which can be

problematic. The MML method which has explicit solutions is used here. To obtain the explicit solution, we order ε_i (for a given ϕ) in order of increasing magnitude and ordered ε- values by $\varepsilon_{(i)}=(x_{[i]}-\phi x_{[i]-1}-\lambda)$, $1\leq i \leq n$. It may be noted that $(x_{[i]}, x_{[i]-1})$ is that pair of (x_i, x_{i-1}) observations which corresponds to the ordered $\varepsilon_{(i)}$. Therefore, $x_{[i]}$ are the concomitants of $\varepsilon_{(i)}$ (Tiku et al.,1999a). Hence the Eqs. (2.21)- (2.23) become,

$$\frac{\partial \ell nL}{\partial \lambda} = \frac{1}{\alpha} + (k-1)\sum_{i=1}^{n} - \frac{1}{\alpha} z_{(i)}^{-1} = 0\,, \tag{2.24}$$

$$\frac{\partial \ell nL}{\partial \phi} = \frac{1}{\alpha}\sum_{i=1}^{n} x_{[i]-1} - \frac{k-1}{\alpha}\sum_{i=1}^{n} x_{[i]-1} z_{(i)}^{-1} = 0 \text{ and} \tag{2.25}$$

$$\frac{\partial \ell nL}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha}\sum_{i=1}^{n} z_{(i)} - \frac{k-1}{\alpha}\sum_{i=1}^{n} z_{(i)} z_{(i)}^{-1} = 0\,. \tag{2.26}$$

Modified likelihood equations are obtained by linearizing the intractable terms in likelihood equations. First two terms of a Taylor series expansion is used to linearize the intractable term $g(z_{(i)}) = z_{(i)}^{-1}$ (Tiku et.al., 1996):

$$g(z_{(i)}) \cong g(t_{(i)}) + \left[z_{(i)} - t_{(i)}\right]\left\{\frac{\partial g(z)}{\partial z}\right\}_{z=t_{(i)}}. \tag{2.27}$$

Thus,

$$z_{(i)}^{-1} \approx \nu_i - \beta_i z_{(i)} \tag{2.28}$$

where $\nu_i = 2\, t_{(i)}^{-1}$, $\beta_i = t_{(i)}^{-2}$ and $t_{(i)} = E\{z_{(i)}\}$ $(1\leq i \leq n)$.

The coefficients $t_{(i)}$ may be obtained from:

$$\frac{1}{\Gamma(k)}\int_0^{t_{(i)}} e^{-z}z^{k-1}dz = \frac{i}{n+1} \qquad (1\le i \le n). \tag{2.29}$$

Eq. (2.28) is incorporated in Eqs. (2.24) - (2.26). Then by solving the resulting equations the following MML estimators (Akkaya and Tiku, 2001b) are obtained.

$$\hat{\lambda} = \bar{\kappa}_{(.)} - \frac{\Delta}{m}\hat{\alpha}\,, \tag{2.30}$$

$$\hat{\phi} = K - D\hat{\alpha} \quad \text{and} \tag{2.31}$$

$$\hat{\alpha} = \frac{-B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}} \tag{2.32}$$

where $\bar{\kappa}_{(.)} = \frac{1}{m}\sum_{i=1}^{n}\beta_i\kappa_{(i)}$ $\qquad \kappa_{(i)} = x_{(i)} - \hat{\phi}x_{(i)-1}$ $\qquad m = \sum_{i=1}^{n}\beta_i$

$$\beta_i = \frac{1}{t_{(i)}^2} \qquad \Delta = \sum_{i=1}^{n}\Delta_i \qquad \Delta_i = \nu_i - \frac{1}{k-1}$$

$\nu_i = \frac{2}{t_{(i)}}$ and $t_i = E\{z_{(i)}\}$,

$$K = \frac{\sum_{i=1}^{n}\beta_i x_{(i)}x_{(i)-1} - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)\left(\sum_{i=1}^{n}\beta_i x_{(i)}\right)}{\sum_{i=1}^{n}\beta_i x_{(i)-1}^2 - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)^2}$$

$$D = \frac{\sum_{i=1}^{n}\left\{\Delta_i - \left(\frac{\Delta}{m}\right)\beta_i\right\}x_{(i)-1}}{\sum_{i=1}^{n}\beta_i x_{(i)-1}^2 - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)^2}$$

and

$$B = (k-1)\sum_{i=1}^{n}\Delta_i(\kappa_{(i)} - \overline{\kappa}_{(.)}),$$

$$C = (k-1)\sum_{i=1}^{n}\beta_i(\kappa_{(i)} - \overline{\kappa}_{(.)})^2 .$$

Computations: To start ordering of ε_i, it is required to estimate initial value of the autoregression parameter of the model. The Least- square (LS) methodology which involves minimizing the error in the sum of squares was proposed for estimating the initial parameters. The LS estimators are:

$$\hat{\lambda}_0 = \overline{\kappa} - \hat{\alpha}k, \tag{2.33}$$

$$\hat{\phi}_0 = \frac{n\left(\sum_{i=1}^{n}x_{i-1}x_i\right)}{n\sum_{i=1}^{n}x_{i-1}^2 - \left(\sum_{i=1}^{n}x_{i-1}\right)^2} - \frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}x_{i-1}\right)}{n\sum_{i=1}^{n}x_{i-1}^2 - \left(\sum_{i=1}^{n}x_{i-1}\right)^2}, \tag{2.34}$$

and $\hat{\alpha}_0 = \sqrt{\dfrac{\sum_{i=1}^{n} \{\kappa_i - \bar{\kappa}\}^2}{(n-2)k}}$ (2.35)

where $\kappa_{(i)} = x_{(i)} - \hat{\phi} x_{(i)-1}$ and $\bar{\kappa} = \dfrac{\sum_{i=1}^{n} \kappa_i}{n}$.

Using the initial concomitants in $(x_{[i]} - \hat{\phi}_0 x_{[i]-1} - \hat{\lambda}_0)$, the MML estimates $\hat{\lambda}$ and $\hat{\alpha}$ are calculated from Eq. (2.30) and (2.32), respectively with $\hat{\phi} = \hat{\phi}_0$. The MML estimate $\hat{\phi}$ is then calculated from Eq. (2.31). A second iteration is carried out with $\hat{\lambda}_0$, $\hat{\alpha}_0$ and $\hat{\phi}_0$ replaced by $\hat{\lambda}$, $\hat{\alpha}$ and $\hat{\phi}$, respectively. This is done a number of times till the estimates stabilize. In all our computations, no more than three iterations are needed for the estimates to stabilize (Akkaya and Tiku, 2001a).

In this procedure, since the shape parameter k is not known, it is chosen initially by either using Q-Q (quantile - quantile probability) plots or considering the coefficient of skewness and kurtosis for error terms (Tiku et. al., 1996). Q-Q plots graphically compare the distribution of a given variable to the desired distributions (represented by a straight line). These plots are constructed by plotting the ordered residuals against $t_{(i)}$ $(1 \le i \le n)$ which are calculated by using the Eq. (2.29). Q-Q plots are obtained for different shape parameter values. The shape parameter value which gives closest to a straight line pattern can be chosen as an initial shape parameter. The straight line shows what our data would look like, if it were perfectly the desired distributed. To determine the shape parameter of distribution, the coefficient of skewness and kurtosis from the residual for comparing the theoretical coefficient of skewness and kurtosis can be

obtained also. The shape parameter which closes the theoretical values can be selected as an initial shape parameter.

A more formal method is to calculate the likelihood function for a series of values of k. The value that maximizes L (or lnL) is chosen as the exact (or nearest) value of k.

2.4 AUTOREGRESSIVE MODELS WITH WEIBULL INNOVATIONS

The time series AR(1) model given by Eq. (2.5), ε_i have a Weibull distribution. The Weibull distribution with shape (p) and scale (α) parameters has the density function (p>0):

$$W(p,\sigma): f(\varepsilon) = \frac{p}{\alpha}\left(\frac{\varepsilon}{\alpha}\right)^{p-1} e^{-\left(\frac{\varepsilon}{\alpha}\right)^{p}} \quad 0<\varepsilon<\infty. \tag{2.36}$$

To have an idea about the nature of the Weibull W(p,σ), some of the theoretical values of its skewness μ_3^2/μ_2^3 and kurtosis μ_4/μ_2^2 are given in Table 2.1 (Tiku and Akkaya, 2004).

Table 2.1 The Theoretical Skewness and Kurtosis Values for Weibull Distribution for Some Shape Parameters

b(shape parameter)	1.5	2	2.5	3	4	6
β_1(coeff. of skewness)	1.064	0.631	0.358	0.168	-0.087	-0.158
β_2 (kurtosis)	4.365	3.246	2.858	2.705	2.752	2.538

2.4.1 Parameter Estimation Procedure with Modified Maximum Likelihood

When ε_i's are iid (identically and independent distributed) and have Weibull distribution, the likelihood function is:

$$L \propto \alpha^n p^n \prod_{i=1}^{n} z_i^{p-1} e^{-\sum_{i=1}^{n} z_i^p}, \qquad z_i=\varepsilon_i/\alpha, \tag{2.37}$$

and

$$\ell nL = \text{const} - n\ell n\alpha - \sum_{i=1}^{n} z_i^{\,p} + (p-1)\sum_{i=1}^{n} \ell n z_i; \tag{2.38}$$

$z_i = (1/\alpha)(x_i - \phi x_{i-1} - \lambda)$. Differentiating Eq. (2.38) with respect to λ, ϕ and α and setting the derivatives equal to zero gives the following equations (Akkaya and Tiku, 2005):

$$\frac{\partial \ell nL}{\partial \lambda} = (p-1)\left[-\frac{1}{\alpha}\sum_{i=1}^{n} z_{(i)}^{-1} + \frac{p}{\alpha}\sum_{i=1}^{n} z_{(i)}^{p-1}\right] = 0, \tag{2.39}$$

$$\frac{\partial \ell nL}{\partial \phi} = \frac{-(p-1)}{\alpha}\sum_{i=1}^{n} x_{[i]-1} z_{(i)}^{-1} + \frac{p}{\alpha}\sum_{i=1}^{n} x_{[i]-1} z_{(i)}^{p-1} = 0 \text{ and} \tag{2.40}$$

$$\frac{\partial \ell nL}{\partial \alpha} = -\frac{n}{\alpha} - \frac{(p-1)}{\alpha}\sum_{i=1}^{n} z_{(i)} z_{(i)-1} + \frac{p}{\alpha}\sum_{i=1}^{n} z_{(i)}^{p-1} z_{(i)} = 0. \tag{2.41}$$

Modified likelihood equations are obtained by linearizing the intractable terms, $z_{(i)}^{-1}$ and $z_{(i)}^{p-1}$ in likelihood equations using the first two terms of a Taylor series expansion:

$$z_{(i)}^{-1} \approx \nu_{i0} - \beta_{i0} z_{(i)} \text{ and } \quad z_{(i)}^{p-1} \approx \nu_{i0}^{*} + \beta_{i0}^{*} z_{(i)} \tag{2.42}$$

where,

$$\nu_{i0} = 2t_{(i)}^{-1}, \quad \beta_{i0} = t_{(i)}^{-2},$$

$$\nu_{i0}^{*} = (2-p)t_{(i)}^{p-1} \text{ and } \beta_{i0}^{*} = (p-1)t_{(i)}^{p-2} \qquad (1 \le i \le n). \tag{2.43}$$

Therefore,

$$\nu_i = (p-1)\nu_{i0} - p\nu_{i0}^{*} \text{ and } \beta_i = (p-1)\beta_{i0} + p\beta_{i0}^{*}. \tag{2.44}$$

The coefficients t_i are determined by the equations $F(t_{(i)};b)=q_i$, $q_i=i/(n+1)$ where F(.) is the cumulative distribution function of weibull:

$$t_{(i)} = (-\ell n(1-q_i))^{(1/p)} \quad 1 \le i \le n. \tag{2.45}$$

Then putting these new terms on the likelihood equations and solving the resulting equations simultaneously, the following MML estimators of $\hat{\lambda}$, $\hat{\phi}$ and $\hat{\alpha}$ are obtained:

$$\hat{\lambda} = \frac{1}{m}\sum_{i=1}^{n} \beta_i \kappa_{(i)} - \frac{\Delta}{m}\hat{\alpha}, \tag{2.46}$$

$$\hat{\phi} = K - D\hat{\alpha} \tag{2.47}$$

$$\text{and } \hat{\alpha} = \frac{-B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}} \tag{2.48}$$

where $\kappa_{(i)} = x_{(i)} - \hat{\phi} x_{(i)-1}$

$$\bar{\kappa}_{(.)} = \frac{1}{m}\sum_{i=1}^{n}\beta_i \kappa_{(i)}\,, \qquad m = \sum_{i=1}^{n}\beta_i\,,$$

$$\Delta = \sum_{i=1}^{n}\Delta_i\,, \qquad \Delta_i = \nu_i\,,$$

$$K = \frac{\sum_{i=1}^{n}\beta_i x_{(i)} x_{(i)-1} - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)\left(\sum_{i=1}^{n}\beta_i x_{(i)}\right)}{\sum_{i=1}^{n}\beta_i x_{(i)-1}^2 - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)^2},$$

$$D = \frac{\sum_{i=1}^{n}\left\{-\nu_i + \left(\frac{\Delta}{m}\right)\beta_i\right\} x_{(i)-1}}{\sum_{i=1}^{n}\beta_i x_{(i)-1}^2 - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)^2} \text{ and}$$

$$B = \sum_{i=1}^{n}\nu_i(\kappa_{(i)} - \bar{\kappa}_{(.)})\,,$$

$$C = \sum_{i=1}^{n}\beta_i(\kappa_{(i)} - \bar{\kappa}_{(.)})^2\,.$$

LS estimators of Weibull distribution:

$$\hat{\phi}_0 = \frac{n\left(\sum_{i=1}^{n} x_{i-1}x_i\right)}{n\sum_{i=1}^{n} x_{i-1}^2 - \left(\sum_{i=1}^{n} x_{i-1}\right)^2} - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_{i-1}\right)}{n\sum_{i=1}^{n} x_{i-1}^2 - \left(\sum_{i=1}^{n} x_{i-1}\right)^2} \qquad (2.49)$$

and

$$\hat{\alpha}_0 = \sqrt{\frac{\sum_{i=1}^{n}\left\{\kappa_i - \bar{\kappa}\right\}^2}{(n-2)\left\{\Gamma\left(1+\frac{2}{p}\right) - \Gamma^2\left(1+\frac{1}{p}\right)\right\}}} \tag{2.50}$$

where $\kappa_{(i)} = x_{(i)} - \hat{\phi}x_{(i)-1}$ and $\bar{\kappa} = \frac{\sum_{i=1}^{n}\kappa_i}{n}$

2.5 AUTOREGRESSIVE MODELS WITH GENERALIZED LOGISTIC INNOVATIONS

Consider the time series AR(1) model given by Eq. (2.5), ε_i have a generalized logistic (GL) distribution. The GL distribution with shape (b) and scale (α) parameters has the density function:

$$GL(b,\sigma): f(\varepsilon) = \frac{b}{\alpha}\frac{e^{-\varepsilon/\alpha}}{\left(1+e^{-\varepsilon/\alpha}\right)^{b+1}} \qquad -\infty < \varepsilon < \infty \quad (b>0). \tag{2.51}$$

This distribution is negatively skewed for b < 1, and positively skewed for b > 1. For b = 1, it reduces to the logistic which is a symmetric distribution. The theoretical coefficient of skewness and kurtosis according to different shape parameters are summarized in Table 2.2 (Tiku and Akkaya, 2004).

Table 2.2 The Theoretical Skewness and Kurtosis Values for Generalized Logistic Distribution for Some Shape Parameters

b (shape parameter)	0.50	1	2	4	6	8
β_1 (coeff. of skewness)	-0.86	0	0.58	0.87	0.96	1.05
β_2 (kurtosis)	5.40	4.20	4.33	4.76	4.95	5.20

2.5.1 Parameter Estimation Procedure with Modified Maximum Likelihood

When ε_i's are iid and have generalized logistic distribution, the likelihood function is:

$$L \propto \left(\frac{b}{\alpha}\right)^n e^{-\sum_{i=1}^{n} z_i} \prod_{i=1}^{n} \left(1+\exp(-z_i)\right)^{-(b+1)}, \qquad z_i=\varepsilon_i/\alpha, \tag{2.52}$$

and

$$\ell nL = const + n\ell nb - n\ell n\alpha - \sum_{i=1}^{n} z_i - (b+1)\sum_{i=1}^{n} \ell n(1+e^{-z_i}); \tag{2.53}$$

$z_i = (1/\alpha)(x_i - \phi x_{i-1} - \lambda)$. Differentiating Eq. (2.53) with respect to λ, ϕ and α and setting the derivatives equal to zero gives the following equations:

$$\frac{\partial \ell nL}{\partial \lambda} = \frac{1}{\alpha} - \frac{b+1}{\alpha}\sum_{i=1}^{n} \frac{e^{-z_{(i)}}}{\left(1+e^{-z_{(i)}}\right)} = 0, \tag{2.54}$$

$$\frac{\partial \ell nL}{\partial \phi} = \frac{1}{\alpha}\sum_{i=1}^{n} x_{[i]-1} - \frac{b+1}{\alpha}\sum_{i=1}^{n} x_{[i]-1} \frac{e^{-z_{(i)}}}{\left(1+e^{-z_{(i)}}\right)} = 0 \quad \text{and} \tag{2.55}$$

$$\frac{\partial \ell nL}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha}\sum_{i=1}^{n} z_{(i)} - \frac{b+1}{\alpha}\sum_{i=1}^{n} z_{(i)} \frac{e^{-z_{(i)}}}{\left(1+e^{-z_{(i)}}\right)} = 0 . \tag{2.56}$$

Modified likelihood equations are obtained by linearizing the intractable term in likelihood equations using the first two terms of a Taylor series expansion:

$$\frac{e^{-z_{(i)}}}{\left(1+e^{-z_{(i)}}\right)} = \nu_i - \beta_i z_{(i)} \tag{2.57}$$

where,

$$\beta_i = \frac{e^{t_{(i)}}}{(1+e^{t_{(i)}})^2} \text{ and } \qquad \nu_i = \frac{1}{1+e^{t_{(i)}}} + \beta_i t_{(i)} \qquad (1 \le i \le n). \tag{2.58}$$

The coefficients t_i are determined by the equations $F(t_{(i)};b)=q_i$, $q_i=i/(n+1)$ where F(.) is the cumulative distribution function of generalized logistic:

$$t_{(i)} = -\ell n(q_i^{-1/b} - 1) \quad 1 \le i \le n. \tag{2.59}$$

Then putting these new terms on the likelihood equations and solving the resulting equations simultaneously, the following MML estimators of $\hat{\lambda}$, $\hat{\phi}$ and $\hat{\alpha}$ are obtained (Akkaya and Tiku, 2001b):

$$\hat{\lambda} = \bar{\kappa}_{(.)} - \frac{\Delta}{m}\hat{\alpha} , \tag{2.60}$$

$$\hat{\phi} = K - D\hat{\alpha} \tag{2.61}$$

and $\hat{\alpha} = \dfrac{-B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}}$ (2.62)

where $\bar{\kappa}_{(.)} = \dfrac{1}{m}\sum_{i=1}^{n}\beta_i \kappa_{(i)}$, $\kappa_{(i)} = x_{(i)} - \hat{\phi} x_{(i)-1}$,

$m = \sum_{i=1}^{n}\beta_i$, $\Delta = \sum_{i=1}^{n}\Delta_i$, $\Delta_i = \nu_i - \dfrac{1}{b+1}$ and

$$K = \frac{\sum_{i=1}^{n}\beta_i x_{(i)} x_{(i)-1} - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)\left(\sum_{i=1}^{n}\beta_i x_{(i)}\right)}{\sum_{i=1}^{n}\beta_i x_{(i)-1}^2 - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)^2},$$

$$D = \frac{\sum_{i=1}^{n}\left\{\Delta_i - \left(\frac{\Delta}{m}\right)\beta_i\right\} x_{(i)-1}}{\sum_{i=1}^{n}\beta_i x_{(i)-1}^2 - \frac{1}{m}\left(\sum_{i=1}^{n}\beta_i x_{(i)-1}\right)^2},$$

$$B = (b+1)\sum_{i=1}^{n}\Delta_i(\kappa_{(i)} - \bar{\kappa}_{(.)}),$$

$$C = (b+1)\sum_{i=1}^{n}\beta_i(\kappa_{(i)} - \bar{\kappa}_{(.)})^2 .$$

LS estimators of generalized logistic distribution:

$$\hat{\lambda}_0 = \bar{\kappa} - \hat{\alpha}[\psi(b) - \psi(1)], \quad (2.63)$$

$$\hat{\phi}_0 = \frac{n\left(\sum_{i=1}^{n} x_{i-1}x_i\right)}{n\sum_{i=1}^{n} x_{i-1}^2 - \left(\sum_{i=1}^{n} x_{i-1}\right)^2} - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_{i-1}\right)}{n\sum_{i=1}^{n} x_{i-1}^2 - \left(\sum_{i=1}^{n} x_{i-1}\right)^2}, \tag{2.64}$$

and

$$\hat{\alpha}_0 = \sqrt{\frac{\sum_{i=1}^{n}\left\{\kappa_i - \overline{\kappa}\right\}^2}{(n-2)\left[\psi'(b) + \psi'(1)\right]}} \tag{2.65}$$

where $\kappa_{(i)} = x_{(i)} - \hat{\phi}x_{(i)-1}$ and $\overline{\kappa} = \frac{\sum_{i=1}^{n}\kappa_i}{n}$.

The computation procedures of MML estimators for weibull and generalized logistic distributions are exactly the same way as for the procedure of the gamma distribution.

2.6 GENERATING RANDOM COMPONENTS OF SKEWED HYDROLOGIC VARIABLES FOR AR(1) MODEL WITH ASYMMETRIC INNOVATIONS

After model parameters are estimated, the generation of samples of stochastic processes of a given autocorrelation structure and marginal distribution is a current problem in stochastic hydrology. As mentioned earlier, the innovations, ε_i, are generally assumed to be normally distributed N (0, σ^2) in hydrological time series model because of the regenerative property of the normal distribution under additive linear models. By means of this property, the addition of normal variates generates other Gaussian variables. This property is not valid for the other

distributions. This creates problems during identification stage. For example, even if x_i are distributed approximately as gamma with mean $\bar{x}$, variance σ^2, and skewness γ_x, the skewness of the random component, $\gamma_{\varepsilon,}$ is different from the skewness of the flows or vice versa. This is due to the sum of gamma variates, not nearly as attractive as those of normal variates, are not necessarily gamma (Fiering and Jackson, 1971). The reason of this situation is explained in infinite moving-average form of ε_i by Hamilton (1994).

If the series are assumed to be gamma distribution as GAR(1) model, this situation must be considered for generation of synthetic events. Since if the model parameters are estimated by means of method of moment procedure used in the GAR(1) model, this must simply require substitution of sample moments $\bar{x}$, r_1 and s for population moments μ, ρ and σ. Therefore keep in mind that these statistics are random variables and can be expected to vary when estimated from different samples (Bras and Rodriguez- Iturbe, 1985). Because of sampling uncertainty, there is no assurance that the moments of the underlying populations are preserved. To preserve the coefficient of skewness of historical time series in simulated time series under gamma distribution, there are several alternatives which have been proposed to overcome this problem using gamma distribution.

In this part of the study, the procedures of the generation of autoregressive synthetic series that preserve the first three moments of the recorded flows under gamma distribution is introduced and the papers related with gamma distribution are briefly reviewed.

One method using the gamma distribution was proposed by Wilson and Hilferty (1931). In this method, the skewness of the random component was modified. Let ζ_i be normally distributed with a mean of zero and a variance of one. Then the modified random sampling variate ε_i was defined by Wilson and Hilferty (1931) as:

$$\varepsilon_i = \frac{2}{\gamma_\varepsilon}\left(1 + \frac{\gamma_\varepsilon \zeta_i}{6} - \frac{\gamma_\varepsilon^{\,2}}{36}\right)^3 - \frac{2}{\gamma_\varepsilon} \tag{2.66}$$

where the skewness of ε, (γ_ε), is related to the estimate of the skewness of x, (γ_x), by

$$\gamma_\varepsilon = \gamma_x \frac{(1-\phi^3)}{(1-\phi^2)^{3/2}} \tag{2.67}$$

Eq. (2.66) was distributed approximately as gamma with a mean of zero, variance of one, and coefficient of skewness, γ_ε. Moreover, its use preserved the third moment of the recorded flows.

McMahon and Miller (1971) stated that simulation model using the Wilson-Hilferty approximation can not reproduce the observed statistical moments of highly skewed hydrologic variables. They claimed that the Wilson-Hilferty approximation was adequate only when the skewness was less than about 4.0. Thus, Kirby (1974) proposed the modified Wilson-Hilferty transformation that preserved the first three moments of the Pearson Type III distribution (gamma distribution with three parameters). This transformation was,

$$\varepsilon_i^{\text{modified}} = A\{\max\left[H, 1 - \left(\frac{\gamma_\varepsilon}{6}\right)^2 + \left(\frac{\gamma_\varepsilon}{6}\right)\zeta_i\right]^3 - B\} \tag{2.68}$$

The corresponding value of H was; $H = \{B - ((2/\gamma_\varepsilon)/A)\}^{1/3}$. The other parameters were given a tabular form up to γ_ε=9.75 (Kirby, 1972).

Obeysekera and Yevjevich (1985) showed the limitations of the modified Wilson Hilferty transformation. Acording to this study, although this information preserved the first three moments, it can deviate significantly from the gamma distribution it was intended to reproduce. They found out that this deviation can be significiantly for large skewness values for which $\varepsilon_i^{\text{modified}}$ was suggested to be applicable (skewness of < 9.0).

Another method which generated AR(1) samples with an exact gamma marginal distribution was proposed by Lawrance and Lewis (1981). This method was used by both Obeysekera and Yevjevich, (1985) and Fernandez and Salas (1990). In this method, the random component, ε_i can be obtained by using two alternative approaches namely integer and non-integer values of shape parameter, k:

i) for integer values of k, ε of Eq (2.5) was given by (Gaver and Lewis,1980):

$$\varepsilon = \frac{\lambda(1-\phi)}{\beta} + \sum_{j=1}^{k} \eta_j \tag{2.69}$$

where η_j = 0, with probability ϕ; η_j = exp (α), with probability ($1-\phi$); and $\exp(\alpha)$ = an exponentially distributed random variable

with expected value $1/\alpha$. This approach is valid for coefficient of skewness less than or equal to 2.0.

ii) for non-integer values of k, based on the shot-noise process, ε can be obtained by:

$$\varepsilon = \lambda(1-\phi)+ \eta \tag{2.70}$$

where

$$\eta = 0 \qquad \text{if } M = 0 \tag{2.71}$$

and

$$\eta = \sum_{j=1}^{M} Y_j (\phi)^{U_j} \qquad \text{if } M > 0 \tag{2.72}$$

where M is an integer random variable with Poisson distribution of mean value $-k\ell n(\phi)$. The set (U_j) is independent, identically distributed, random variables with uniform distribution in (0,1). The variable (Y_j) is also independent, identically distributed; random variables with exponential distribution of mean value $1/\alpha$ (Fernandez and Salas,1990). The above schemes are valid only for positive autoregression coefficient (Obeysekera and Yevjevich, 1985).

If the modified maximum likelihood procedure is used for any location-scale distribution, there is no need to preserve the skewness value of the historical data to generate the random variate. In addition, the model parameters are estimated from residuals in this method. Since the population moments are not used to estimate the model parameters, the

inverse transform method is used to generate a random variate for any distribution due to this property of this method. Let F^{-1} denote the inverse of the cumulative distribution function, F of any distribution. Then, an algorithm for generating a random variate having distribution function F is as follows:

i. Generate random variate U~U(0,1)
ii. Return random variable = $F^{-1}(U)$

where U is pseudorandom number uniformly distributed on the interval [0, 1]. F^{-1} is obtained by setting U= F and solving for random variable. Thus, to generate the desired random variate, U~U(0,1) is generated firstly and then the random variate is obtained by using $F^{-1}(U)$ of related distribution. The inverse cumulative density functions to generate the desired random variate are given below for gamma, weibull and generalized logistic distributions. These techniques will be used to obtain generated and forecasting streamflow data for relating distribution in Chapter 3.

Gamma Innovation:

The inverse cumulative distribution function of gamma distribution is:

$$F^{-1}(U) = \zeta_t = \frac{1}{\hat{\alpha}^k \Gamma(k)} \int_0^{U_i} e^{-t} t^{k-1} dt \qquad (2.73)$$

Weibull Innovation:

The inverse cumulative distribution function of Weibull distribution is:

$$F^{-1}(U) = \zeta_{ti} = \hat{\alpha}(-\ln(1-U_i)^{1/p}) \qquad (2.74)$$

Generalized Logistic Innovation:

The inverse cumulative distribution function of GL is:

$$F^{-1}(U) = \zeta_t = -\hat{\alpha}\ln\left[U_i^{-1/b} - 1\right]. \quad (2.75)$$

2.7 ARTIFICIAL NEURAL NETWORK

As it has been mentioned before, the above techniques for time series analysis assume linear relationships among variables. In the real world, however, temporal variations in data are difficult to analyze and predict accurately. Artificial neural networks (ANN) which are suited to complex nonlinear models be used for the analysis of real world temporal data. Besides the linearity assumption, to use the time series model in the literature, the distribution types to be fitted to given data series must be determined before any analysis. However, there is no need to make any statistical distribution assumption for ANN model. Therefore, it is not necessary to know whether a feature set is of a normal or gamma or generalized logistic distributions to estimate the model parameters for this model.

An ANN is an information processing technique of artificial intelligence which is inspired by the biological brain model and formed by numerous interconnected neurons. It is a simplified mathematical representation of this biological neural network.

2.7.1 Methodology of Artificial Neural Network

In this chapter, the information about neural networks is not given in detail. However, only main concepts have been introduced.

Artificial neural networks (ANN) are parallel computing systems whose original development was based on the structure and function of brain as said earlier (Imrie et al., 2000). In an ANN, the unit is known as the processing element (PE). Each PE receives inputs through connections with other elements and multiplies every input by its interconnection weight, sums the product, and then passes the sum through a transfer function to produce its result. A neural network is a combination of a number of processing elements organized in layers. Typically there are two layers connected to the external environment, an input layer where the data is presented to the network and output layer that holds the response of the network to the given input. Layers in between are called hidden layers which provide the nonlinearity relationships for the network. Each layer is made up of several nodes. The pattern of connectivity and the number of processing units in each layer may vary within some constraints. No communication is permitted between the processing units within a layer, but the processing units in each layer may send their output to the processing units in the succeeding layers. The neurons in each layer are connected to the neurons in a subsequent layer by a weight w, which may be adjusted during training. A three layer which has input, output and hidden layers, feed-forward ANN is shown in Figure 2.1. The data passing through the connections from one neuron to another are multiplied by weights. When these are modified, the data transferred through the network changes; consequently, the network output also changes.

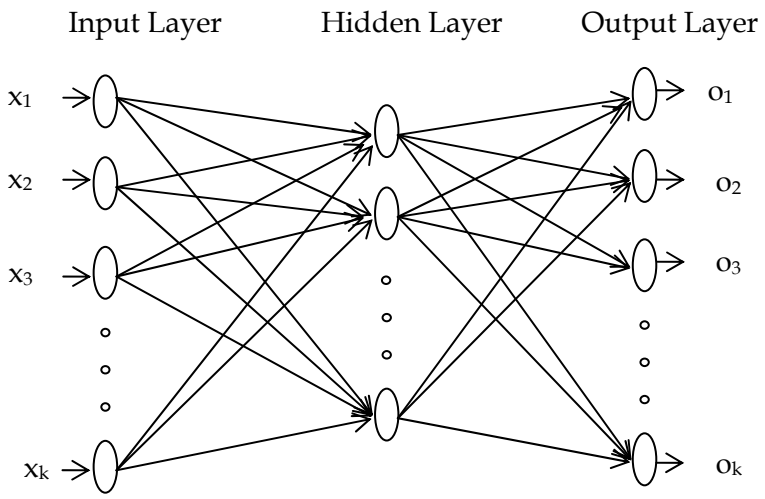

Figure 2.1 A Three-layer ANN Architecture

There are several activation functions that can be used in ANN such as step, sigmoid, threshold, linear, etc. The sigmoid function to give the neural network output values is commonly used as an activation function in ANN.

The input values, x_i are multiplied by the first interconnection weights w_{ij}, j=1,…,h at the hidden nodes, and the products are summed over the index, i, and become the inputs to the hidden layers i.e.;

$$H_j = \sum_{i=1}^{k} w_{ij} x_i \quad j=1,\ldots,h \tag{2.76}$$

where H_j is the input to the jth hidden node, w_{ij} is the connection weight from the ith processing element (PE) to the jth PE. Each hidden node is transformed through a sigmoid function to produce a hidden node output, HO_j and is defined as mathematically (Raman and Sunılkumar, 1995):

$$HO_j = f(H_j) = \frac{1}{1+e^{-(H_j)}}. \qquad (2.77)$$

where H_j is the input to the node, $f(H_j)$ is the node output. The output, HO_j serves as the input to succeeding layer and this produce is continued until the output layer is reached. This is referred to as forward activation flow. The input to the m output nodes, IO_n, is expressed as:

$$IO_n = \sum_{i=1}^{h} w_{jn} HO_i \quad n=1,\ldots,m \qquad (2.78)$$

These input values are processed using the sigmoid function until obtaining the output values O_n. A portrayal of a unit and its function is given in Figure 2.2.

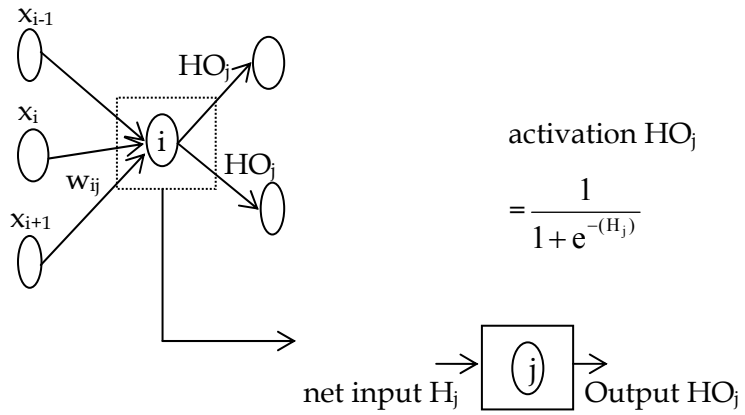

Figure 2.2 Portrayals of a Unit and Its Function

The subsequent weight adoption or learning process is accomplished by the back propagation learning algorithm. The O_n at the output layer will

not be the same as the target value, T_n. For each input pattern, the root mean square error, RMSE for the pth input pattern is calculated by using Eq. (2.79):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{m}(T_n - O_n)^2}\ . \tag{2.79}$$

The aim of the back propagation learning algorithm is to minimize iteratively the average squared error between values of the output, O_n, at the output layer and the correct value, T_n, provided by teaching (Dikmen, 2001).

CHAPTER 3

CASE STUDY ON KIZILIRMAK BASIN

3.1 INTRODUCTION

As a part of this study, Kızılırmak River Basin is taken into account. The annual and monthly data sets are chosen from five streamflow gauging stations located on Kızılırmak River.

The streamflow series from Kızılırmak Basin consist of time-dependent variables. Therefore, time series models are to be used in the examination of data. The autoregressive time series are the most commonly used models in hydrology. GAR(1) and AR(1) models with asymmetric innovations are proposed in the literature for the series not normally distributed. In contrast to GAR(1), AR(1) model with asymmetric innovations is a more advantageous model considering the varsity of its applications to different distributions. Although there are studies showing the application of GAR(1) model to hydrological data, the application of AR(1) model with asymmetric innovations is not yet examined for non-normal hydrological time series. In this chapter, the results from GAR(1) model are compared with the results from AR(1) model with asymmetric

innovations and the advantages of AR(1) model are examined. AR(1) model with asymmetric innovations is found to be more advantageous since it can be used with different distribution whereas GAR(1) model can be used with only gamma distribution.

Synthetic series are required to be generated during the planning phase of dams. The choice of model in the representation of investigated series is important for the reliability of the synthetic streamflow series to be generated. The selected model leads to the generation of synthetic series which affects the dimensions of the dam, and in turn affects the stability of the dam with some social and economical consequences. To this end, the reliable synthetic series generation from observed data is one of the main research issues in hydrology since 1950s.

In order to construct an appropriate model for the present data sets it was necessary to identify the distribution types of the error terms. Weibull distribution is determined as a good fit for the annual data sets. On the other hand, the errors of the monthly data sets are found to be best represented by the generalized logistic distribution. Next, the model parameters for annual and monthly data are needed to be estimated by MML method for selected distributions. Then, the synthetic generation and forecasting series are obtained with estimated model parameters set using AR(1) model and compared with real observed data set for the respective stations. Forecasting is done for only station EIE 1501 to find out the reliability of the generated model. The parameters of the model were found by using the first 40 years of data and the forecasting series was calculated for comparing with the extra seven years of observed data.

Moreover, ANN models are developed for annual and monthly series. The results from ANN models are compared with the results from AR(1) with asymmetric innovation models for one streamflow gauging station (EIE 1501).

3.2 GAR(1) AND AR(1) MODEL WITH GAMMA INNOVATIONS FOR ANNUAL STREAMFLOW DATA

Gamma autoregressive model, GAR(1), was developed and applied to annual streamflow data for the case that the gamma distribution representing streamflow data better than the normal distribution. However, in the previous studies on GAR(1) model such as Fernandez and Salas (1990), the Gamma distribution was studied only as an assumption. The identification of the distribution type was not based on any statistical investigation on real data. The previous study used the method of moments procedure in estimating the model parameters.

In this study, however, AR(1) model with gamma innovations is set up as presented by Tiku et. al., (1996). Model parameters under gamma distribution assumption are estimated by means of modified maximum likelihood method, rather than using method of moments.

It is found in this study that the gamma likelihood function cannot be maximized for all model parameters for the data set examined. However, maximum likelihood procedure is based on maximization of the likelihood function by all of the model parameters. Therefore, we can say that the errors of these data sets cannot be represented with Gamma distribution.

Since both residual series of the annual data and monthly data do not fit to the Gamma distributions, the statistical results corresponding to the gamma distribution methodology were not included here.

3.3 AR(1) MODEL WITH WEIBULL INNOVATIONS FOR ANNUAL STREAMFLOW DATA

The autoregressive models are constructed using 40 years of streamflow data series for four gauging stations located on the main river. Although, 41 years of data (1955- 1995) are available for five stations, the autoregressive models are based on 40 years of data for EIE 1501, EIE 1503, EIE 1528 and EIE 1536 gauging stations. The monitored data for EIE 1541 streamflow gauging station show drastically decreasing trend in the last two years. In order to avoid some measuring error or inconsistent factors, autoregressive model is set up using 38 years of data for the this station. The starting data for each station is used as an initial value for estimating the MML parameters.

In the implementation of AR(1) model, first, the normality of historical series and error terms (residuals) are tested and the coefficient of skewness and kurtosis are calculated. The residuals are obtained firstly from Eq. (2.5) by using the least squares (LS) estimators of autoregressive coefficient. The coefficient of skewness and kurtosis for historical series and residuals are presented in Table 3.1.

Table 3.1. The Coefficient of Skewness and Kurtosis Values of Historical Series and Residuals at Various Runoff Stations

Stations	**Coefficient of Skewness**		**Kurtosis**	
	Historical series	residuals	Historical series	residuals
EIE1501	0.70	0.63	3.09	2.99
EIE1503	0.85	0.78	3.37	3.25
EIE1541	0.48	0.98	2.77	3.18
EIE1528	0.72	0.61	3.15	3.05
EIE1536	0.72	0.60	3.14	3.05

As seen in Table 3.1, the coefficients of skewness and kurtosis for each series are far from the theoretical skewness and kurtosis values of a normal distribution, zero and three respectively. Therefore, these values clearly indicate the non-normalities of both distributions of historical series and residuals.

Jarque-Bera statistics (Bowman and Shenton, 1975) can be used to test the normality. This method measures the difference of the skewness and kurtosis of the series with those from the normal distribution. The statistic is computed as

$$\text{Jarque} - \text{Bera} = \frac{\text{N} - \text{u}}{6}\left(\text{skewness}^2 + \frac{(\text{kurtosis} - 3)^2}{4}\right) \tag{3.1}$$

where u represents the number of estimated coefficients used to create the series. Under the null hypothesis of a normal distribution, the Jarque-Bera statistic is distributed as chi-square statistic with 2 degrees of freedom (one for skewness one for kurtosis). The calculated Jarque-Bera test statistics for

each runoff data set are shown in Table 3.2. The critical value at 99.95 % level for 2 degrees of freedom is 0.103.

Table 3.2 Jarque-Bera Statistic Values for Each Runoff Data Set

Stations	Jarque-Bera statistic for historic series	Jarque-Bera statistic for residuals	Critical value
EIE1501	3.12	2.51	< 0.103
EIE1503	4.79	3.95	< 0.103
EIE1541	1.54	6.14	< 0.103
EIE1528	3.32	2.36	< 0.103
EIE1536	3.31	2.28	< 0.103

These values are not smaller than the critical value. Therefore, according to Jarque-Bera statistic the null hypothesis of normal distribution is rejected indicating that residuals are not from a normal distribution.

After verifying the non-normality and right-skewed distribution with the above procedures, Weibull distributions are decided to be used to set the AR(1) model in order to obtain the synthetic series for each station. Furthermore the coefficient of skewness and kurtosis values obtained from observed data are similar to theoretical values of Weibull distribution. To further examine whether the error distributions fit to the Weibull distributions or not, Q-Q plots and goodness of fit tests are used. Q-Q plots for each station are constructed for different shape parameter values.

Only one of these Q-Q plots for the station EIE1501 is represented in Figure 3.1 for the shape parameter, p=1.5. The other Q-Q plots for other stations are given in Appendix A1.

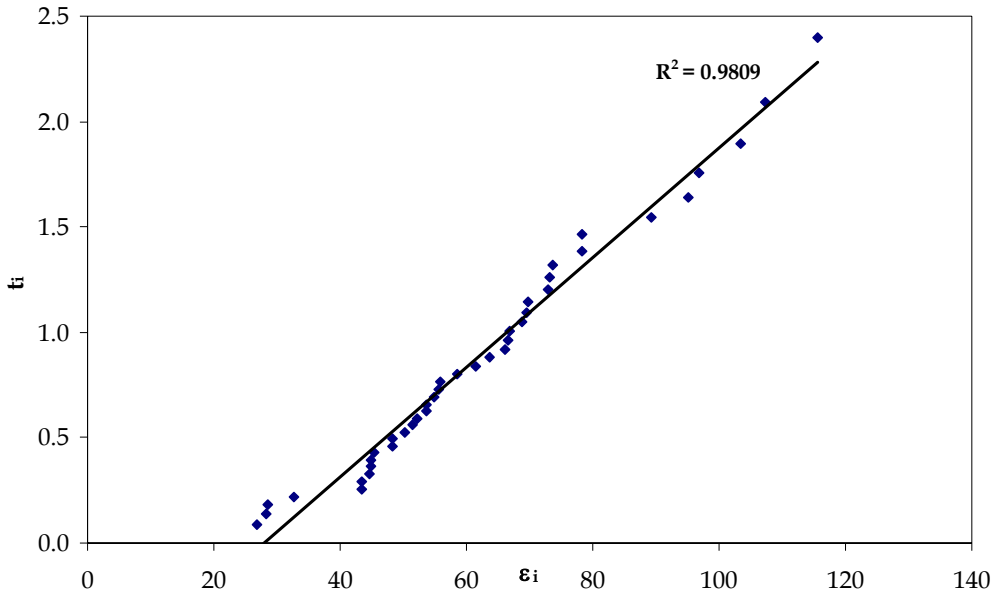

Figure 3.1 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovations for EIE 1501 (p=1.5)

Q-Q plots pointed out that the Weibull distribution is suitable to represent streamflow data. On the other hand, the goodness of fit tests are used to decide whether the residuals come from a population represented by Weibull distribution or not. There are two goodness of fit tests for Weibull distribution. The first was developed by Tiku and Singh (1981). Their test statistics is presented here in Eq.(3.2). The second test was developed by Evans et. al. (1989). Their results showed that the Anderson- Darling statistics is the most sensitive to the lack of fit to a two-parameter Weibull distribution, and the correlation statistics of the Shapiro-Wilk type (see Eq. (3.4)) is the most sensitive to departures from a three-parameter Weibull distribution. Because of the fact that the number of parameters are three in our model of the residuals for each station, the correlation statistics of the Shapiro-Wilk type are used to test the goodness of fit to the Weibull distribution. We here introduce in detail the two goodness of fit test statistics mentioned above.

Testing the Weibull Distribution by Tiku test

Tiku's goodness of fit statistics is defined as (Tiku and Singh, 1981):

$$Z^* = 2\sum_{i=1}^{n-2}(n-1-i)G_i \Big/ (n-2)\sum_{i=1}^{n-1} G_i \ . \qquad (3.2)$$

By using Z^* values, the fitness to the Weibull distribution can be tested as follows:

i. From the n number of observations $Y_i = \ell n\varepsilon_i$, i=1,…,n, calculate the statistics Z^*, with $G_i = (Y_{i+1} - Y_i)/\mu_{i+1:n} - \mu_{i:n})$, where $\mu_{r:n}$ is the expected value of the rth order statistic in a random sample size n. Before using this test, residuals are ordered in ascending magnitudes.

ii. The null distribution of Z^* does not depend on the shape parameter.

iii. To calculate the statistic, the values of the expected values $\mu_{r:n}$ are needed. These values are obtained from Eq. (3.3).

$$\mu_{r:n} = \log(-\log(1 - \frac{(i-0.50)}{(n+0.25)}) \quad (i=1,\ldots,n). \qquad (3.3)$$

iv. Substituting the G_i values in Eq. (3.3), Z^* values for each station are obtained and represented with tabulated values in Table 3.2. When these values are compared with the tabulated lower and upper limits with a certain level of confidence, the null

hypotheses are accepted at 5% significance level for all stations. The tabulated values were given in Tiku and Singh (1981).

The results of goodness of fit test using Tiku Statistics are shown in Table 3.3. Stations EIE 1501 and EIE 1541 pass the test while station EIE 1503 fails the test. Stations EIE 1528 and EIE 1536 pass the test just on the lower limit.

Table 3.3 Goodness of Fit Test for Weibull Distribution Using Tiku Test

Stations	Computed values Z^*	Critical Value, Z^*		Overall decision
		95% confidence level lower limit	95% confidence level upper limit	
EIE 1501	0.87	0.84	1.16	Accepted
EIE 1503	0.81	0.84	1.16	Rejected
EIE 1541	0.93	0.81	1.10	Accepted
EIE 1528	0.84	0.84	1.16	Accepted
EIE 1536	0.84	0.84	1.16	Accepted

Shapiro-Wilk type correlation statistic

Evans et. al. (1989) modified a simplified form of Shapiro- Wilk statistics to develop a goodness of fit test for Weibull distribution. Shapiro-Wilk statistics had been used earlier to test the goodness of fit of normal distribution. Modified goodness of fit test for Weibull distribution is given below:

i. Let $\varepsilon_{(1)}, \varepsilon_{(2)},\ldots, \varepsilon_{(n)}$ denote an ordered residual sample of size n from the population of interest. If a variable ε has the three

parameter Weibull distribution, the variable $Y\varepsilon = \ell n\, \varepsilon$ has an extreme value distribution (Evans et. al., 1989). Thus for three parameter Weibull distributions, the correlation type statistic R_{we}^2 is:

$$R_{we}^{\ 2} = \left\{ \frac{\sum_{i=1}^{n} \left(\ell n\varepsilon_i - mean_{\ell n\varepsilon_i} \right) \ell nm_{w,i}}{\left[\sum_{i=1}^{n} \left(\ell n\varepsilon_i - mean_{\ell n\varepsilon} \right)^2 \sum_{i=1}^{n} \left(\ell nm_{w,i} - mean_{\ell nm_{w,i}} \right) \right]^{1/2}} \right\}^2 \qquad (3.4)$$

and

$$\ell nm_{w,i} = \ell n \left\{ \left[-\ell n \left(1 - \frac{i - 0.3175}{n + 0.365} \right) \right]^{1/p} \right\} \qquad (3.5)$$

where p is the shape parameter of this distribution.

ii. Tests are carried out to check the goodness of each fit using the techniques of correlation statistics of the Shapiro-Wilk type under confidence levels of 90%, 95% and 99%. The overall decisions for fits according to selected shape parameter value are considered to be acceptable if any one of the test results is acceptable. Critical values of this test at the 0.10, 0.05 and 0.01 significance levels by adding the shape parameter are:

$$(R^2_{WES})_{0.10} = \{0.994111418 - (1.81407/n) + (12.38547217/n^2) - 0.00705129 + [0.003971786(shape)] - [0.000508929(shape)(shape)]\}^2 \qquad (3.6)$$

$$(R^2_{WES})_{0.05} = \{0.99229032 - (2.24194/n) + (16.33414042/n^2) - 0.00551925 + [0.00348(shape)] - [0.000492187(shape)(shape)]\}^2 \qquad (3.7)$$

$$(R^2_{WES})_{0.01}=\{0.98757887-(3.37283/n)+(26.99680370/n^2)+0.001807429 + [0.0006810714(shape)]-[0.000299107(shape)(shape)]\}^2 \quad (3.8)$$

iii. The hypothesis that a three parameter Weibull fits the data if R^2_{we} is less than the critical value is rejected. The assumed theoretical distribution is accepted otherwise.

The results of goodness of fit test using Shapiro-Wilk statistics are shown in Table 3.4.

Table 3.4 Goodness of Fit Test for Weibull Distribution Using Shapiro-Wilk Test

Stations	Shape parameter	Computed value R^2_{we}	Critical Value, (R^2_{WES})			Overall decision
			90%	95%	99%	
EIE 1501	1.8	0.95510	0.9119	0.8942	0.8504	Accepted
EIE 1503	1.5	0.93784	0.9106	0.8931	0.8506	Accepted
EIE 1541	1.5	0.96927	0.9077	0.8896	0.8458	Accepted
EIE 1528	1.8	0.95722	0.9119	0.8942	0.8504	Accepted
EIE 1536	1.8	0.95862	0.9119	0.8942	0.8504	Accepted

According to the results presented in Table 3.3 and Table 3.4, the null hypothesis is accepted indicating that error terms come from Weibull distribution. As a result, the Weibull distribution is proven to be a good fit to represent the residuals of annual data sets of each station.

3.3.1 MML Estimators for the Model Parameters

The shape parameter is not known in this procedure. Q-Q plots can also be used to determine the initial shape parameter. The closest values to the

straight line are chosen as the initial shape parameters. The theoretical coefficient of skewness and kurtosis values for Weibull distribution represented in Table 2.1 are compared to the computed values to select the initial shape parameter values as a second procedure. According to this, it is found out that the shape parameter (p) values for each station can be selected between 1.5 and 2.2. As seen in the Q-Q plots in Figure 3.1 and in the Appendix A1, these values are 1.8 for EIE1501, EIE1528 and EIE1536 and 1.5 for EIE1503 and EIE1541 stream gauging stations. The goodness of fit test for Weibull using Shapiro-Wilk test support these conclusions.

After determining the distribution types according to Q-Q plots and goodness of fit test and shape parameters for each runoff station, the MML method is used to estimate the model parameters at the respective stations in Kızılırmak Basin.

To apply MML procedure, all the least square estimators are needed to order variates $\varepsilon_{(i)}$ ($1 \le i \le n$). There are two kinds of least square estimators used here. The first one is denoted by $\hat{\phi}_0$ and represents the autoregression coefficient and the other is $\hat{\alpha}_0$, representing scale parameter for AR(1) model. They are computed using Eqs. (2.49)–(2.50) for each stream gauging station and presented in Table 3.5 in the second and the third columns.

Using these initial values, the ordered variates $\varepsilon_{(i)}$ ($1 \le i \le n$) and the corresponding concomitants ($x_{[i]}$ and $x_{[i]-1}$) are obtained. From these values, MML estimators are calculated and then LS estimators are replaced with MML estimators. This procedure is repeated until the estimates stabilize. The corresponding MML estimates shown in Table 3.5 (the forth and seventh columns) are obtained explicitly from Eqs. (2.46)-(2.48). A Fortran program given in Appendix B is prepared and run for these computations.

Table 3.5. Estimated Parameters from AR(1) Model with Weibull Innovations for Each Gauging Station

Stations	**LS Estimators**		**MML Estimators**			
	Autoreg. ($\hat{\phi}_0$)	**Scale** (α_0)	**Shape** (p)	**Location** (λ)	**Autoreg.** ($\hat{\phi}$)	**Scale** (α)
EIE 1501	0.1019	42.63	1.8	32.04	0.0043	42.33
EIE 1503	0.0954	43.84	1.5	56.32	0.0014	45.26
EIE 1541	0.5364	13.86	1.5	7.88	0.1333	15.52
EIE1528	0.1965	66.93	1.8	76.98	0.0134	66.83
EIE1536	0.1971	76.52	1.8	87.04	0.0134	76.50

Next, the aim is to find out whether the initially specified shape parameters are correct or not. It needs to be checked whether the likelihood function for Weibull distribution is maximized or not by other estimated model parameters, which are λ, $\hat{\phi}$ and $\hat{\alpha}$. As known, maximum likelihood procedure is based on maximization of the likelihood function by all of the parameters.

The log-likelihood functions for Weibull distribution are given in Figures 3.2-3.6. Therefore it can be said that the assumptions mentioned earlier for shape parameters are correct.

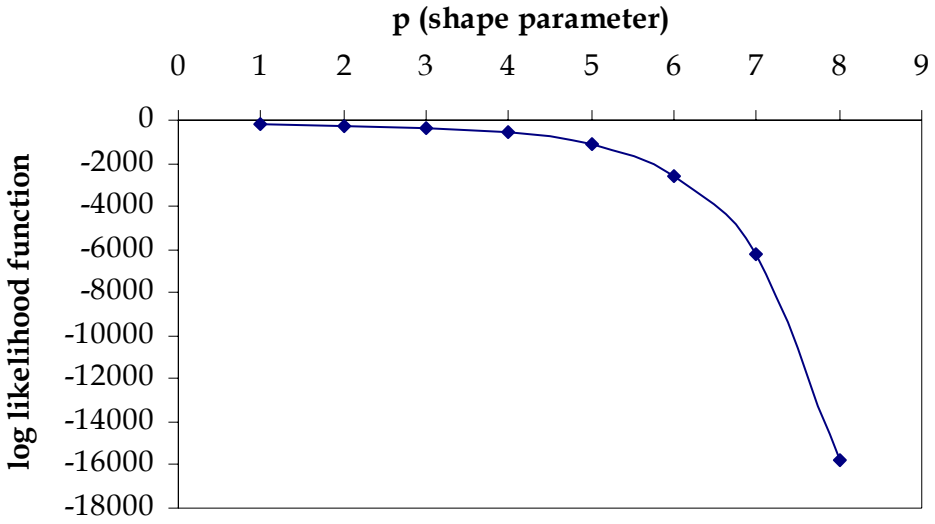

Figure 3.2 Weibull Log-likelihood Function with respect to the Different Shape Parameters for EIE 1501

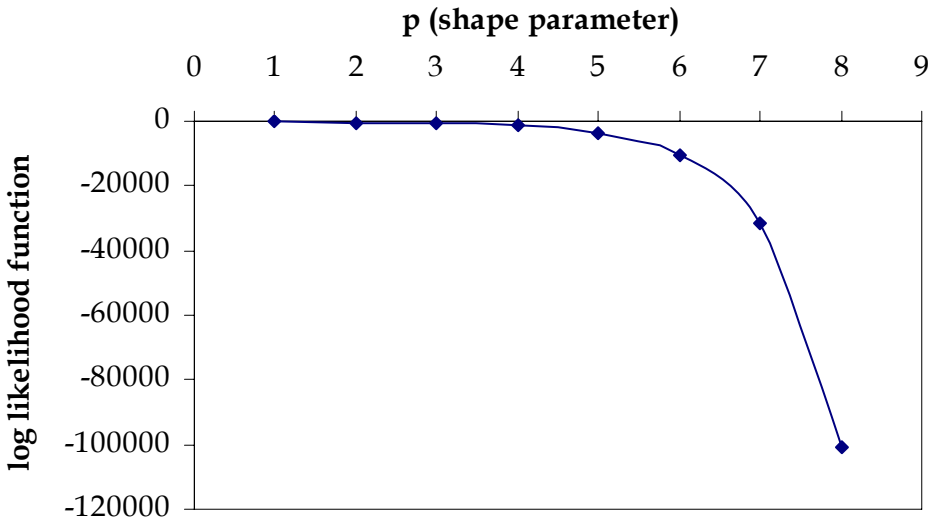

Figure 3.3 Weibull Log-likelihood Function with respect to the Different Shape Parameters for EIE 1503

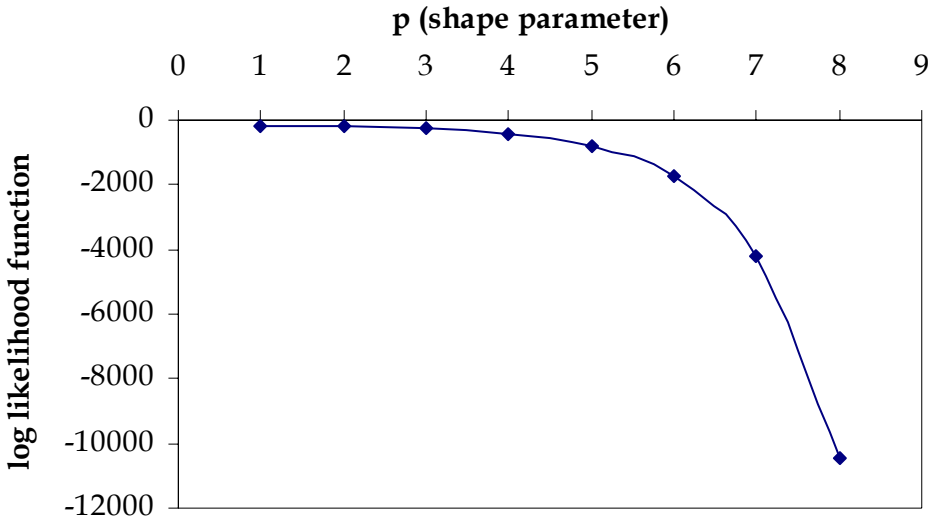

Figure 3.4 Weibull Log-likelihood Function with respect to the Different Shape Parameters for EIE 1541

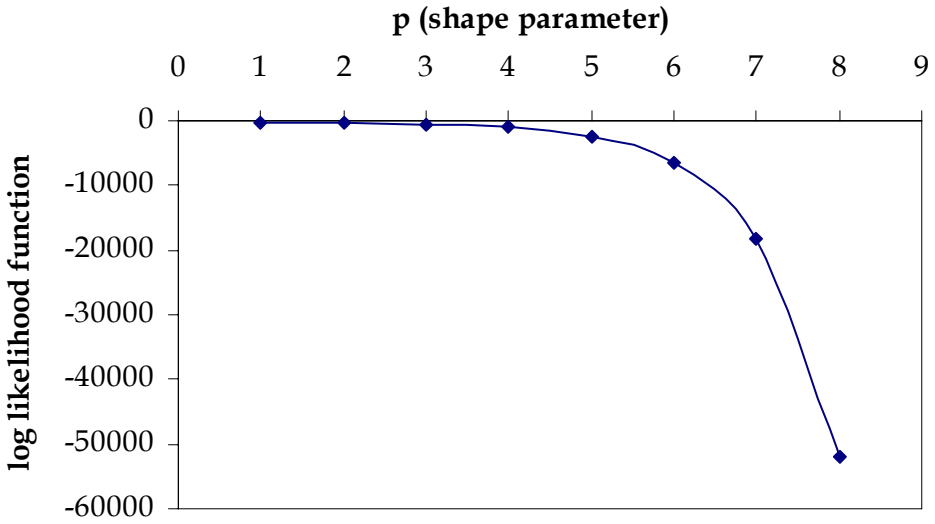

Figure 3.5 Weibull Log-likelihood Function with respect to the Different Shape Parameters for EIE 1528

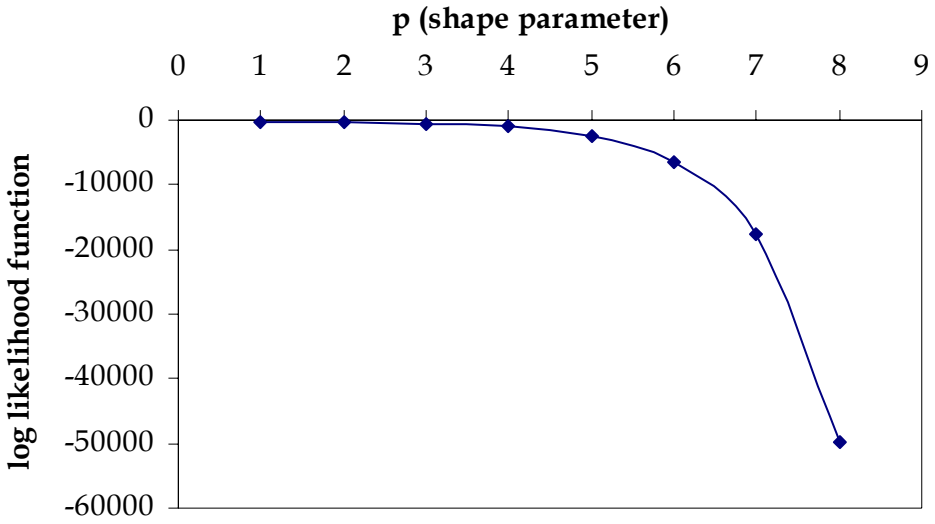

Figure 3.6 Weibull Log-likelihood Function with respect to the Different Shape Parameters for EIE 1536

As a result of MML procedure, all of the model parameters (p, λ, ϕ and α) are estimated. In addition, the estimated modeling parameters are unbiased, efficient and robust from the statistical point of view as seen in the previous studies such as Akkaya and Tiku, (2001a), Tiku and Akkaya (2004) and Akkaya and Tiku (2005) etc. Since the aim of the study is to introduce and apply an alternative modeling procedure in hydrology, and since the previous studies showed that the model parameters are unbiased, efficient and robust, therefore, the present study did not elaborate on these matters.

Consequently, the model parameters are estimated explicitly by using the MML method. The AR(1) model can now be constructed with the estimated parameters.

3.3.2 Generation Annual Data Using AR(1) Model with Weibull Innovations

Once the parameters of the AR(1) model are determined, the autoregressive model are used in order to generate synthetic annual series. The first value in annual data is used as an initial value, x_0 for simulation part. The value of ζ_1 which is an independent Weibull variable is randomly selected from a Weibull distribution by using Eq. (2.74). With the starting data, x_0 and ε_1 ($=\alpha\zeta_1$) values, Eq. (3.9) yields a value for x_1, the first synthetic event by using estimated parameters; $\hat{\phi}$, α and λ

$$x_1 = \lambda + \hat{\phi}x_0 + \alpha\zeta_1 . \tag{3.9}$$

The summation of the x_1 and random number, ε_2 ($=\alpha\zeta_2$) has given us a new value for x_2 as the second synthetic event. This procedure is repeated n times for each station. The mean and standard deviation of generated series from AR(1) models are calculated in order to compare with the values of historical series for each stream gauging station.

Following the above procedure, a sufficiently large number of synthetic series each having the length of the historical series, which is 1000, are decided to be generated. For each one of the generated synthetic series, the mean and standard deviation are determined. In fact, the synthetically generated series is expected to conserve some of the statistical properties of the historical data because of the stationarity condition of autoregressive models. This implies that mean is not a function of time. Similarly, the variance and correlation coefficient will also be independent of time (Bras and Rodriguez-Iturbe, 1985).

However, this is only true (especially for the case of correlation coefficients) when the method of moments is used for estimating the parameters of the model. Nevertheless, when the estimation method of maximum likelihood is used, the model correlogram will not necessarily be the same as the historical correlogram. In conclusion, one can not say in general that the AR(1) model "preserves" the first serial correlation coefficient (Salas et.al., 1980). For this aim, Table 3.6 shows the mean values of the computed moments (mean and standard deviation) based on the 1,000 series for each stream gauging station.

Table 3.6 Mean Values of Moments Derived from Generated Series Based on AR(1) Model with Weibull Innovations

Stations	Generated Synthetic Series		Historical Series	
	Mean	Std. dev.	Mean	Std. dev.
EIE 1501	69.89	21.57	69.21	21.63
EIE 1503	97.22	27.75	96.08	26.65
EIE 1541	25.14	9.54	24.71	9.75
EIE 1528	138.08	34.10	136.91	34.46
EIE 1536	157.44	39.25	155.63	39.41

The first two moments (mean and standard deviation) computed from the synthetic data are compared with the respective moment values of the historical series using relative errors. The relative error values in percentages (Table 3.7) are obtained by subtracting the historical moment value from generated moment value and dividing by generated moment value, multiplying by 100.

Table 3.7 Relative Errors between Generated and Historical Moment Values for Annual Data

Stations	Relative errors	
	Mean (%)	Std. dev. (%)
EIE 1501	0.97	0.27
EIE 1503	1.17	-3.96
EIE 1541	1.71	-2.20
EIE 1528	0.85	-1.06
EIE 1536	1.15	-0.40

It is seen that the relative error values given as percentages for AR(1) model are relatively low. There is only one error value for standard deviation higher than 3 percent on one station. This error also is quite small.

These results show that AR(1) model with Weibull innovations satisfactorily preserve the mean and standard deviation of a historical series when the model parameters are estimated by using MML procedure.

3.3.3 Forecasting Annual Data Using AR(1) Model with Weibull Innovations

The second important application in time series model is in forecasting or predicting future events. For this purpose, we use the seven years of annual data observed at EIE 1501 station after 1995 which was not used to set up the AR(1) model with Weibull innovations. These seven years are utilized to see how the model forecast the future annual runoff data. The annual 1995 data is utilized as initial value for synthetic forecasting data.

Then the same procedure explained in section 3.3.2 is followed to generate data. Forecasting and observed annual data series for EIE 1501 station are plotted in Figure 3.7 for comparison.

As seen in Fig. 3.7, AR(1) model with Weibull innovations based before 1995 data forecast closely the observed hydrograph ordinates after 1995.

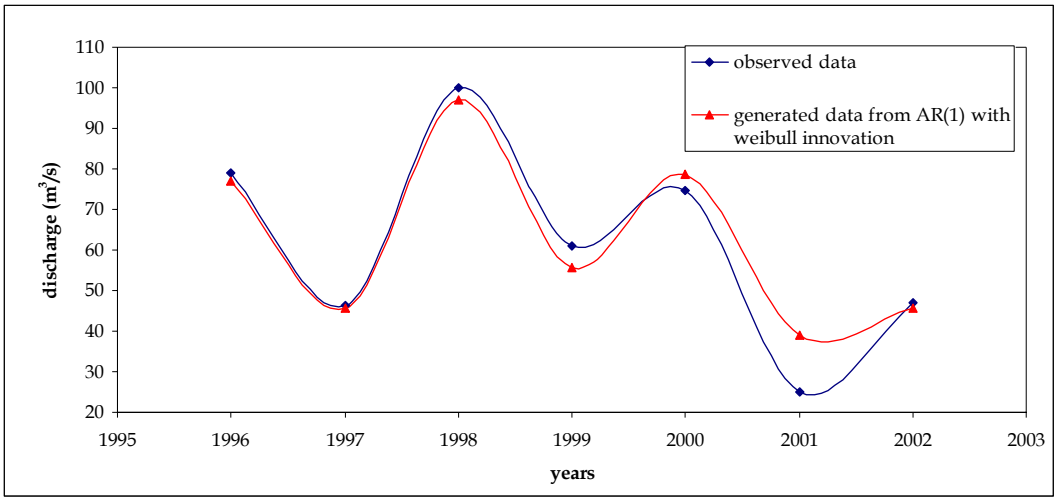

Figure 3.7 Forecasted Annual Data Series and Observed Annual Data Series for EIE 1501 Stream Gauging Station

The AR(1) model with Weibull innovations predicted peak flow is 97.03 m^3/s with an underestimation of 3.0%. The predicted lowest flow is 38.86 m^3/s with an overestimation of 54.5%. Although there is only one value which is overestimated, model performance can be evaluated as good according to the differences between observed and AR(1) modeled with Weibull innovations.

To evaluate the overall forecasting performance for seven years, the root mean square error (RMSE) term is used. RMSE between forecasting and observed values is calculated as 6.00 from Eq. (2.79). This test shows a

good forecasting performance. This test is chosen to illustrate the difference between forecasting and observed rather than the error, because there is a significant uncertainty in the measured parameter. As a part of conclusion, the Table 3.6 and Figure 3.7 show that AR(1) with Weibull innovations model estimates are consistent with the observed values with small root mean square error terms.

The results suggest that proposed model may provide a superior alternative to the AR(1) model for developing simulation and forecasting models in situations that do not require modeling of the internal structure of the basin.

3.4 AR(1) MODEL FOR MONTHLY STREAMFLOW DATA

Periodic hydrological data such as seasonal, monthly, weekly and daily series generally have periodicity in the mean and in the standard deviation (Salas et. al., 1980).

In this study, 480 (40 years*12) monthly streamflow data for each station are used.

Monthly time series denoted as $x_{v,\tau}$, v=1,...,n and τ=1,...,w, where n is the total number of years of data and w is the number of time intervals within a year (w=12 for monthly data) are plotted for each station and shown in Figures 3.8-3.12. These figures show that streamflows during the spring and early part of the summer are generally higher than the rest of the year; this situation repeats itself every year in a periodic manner.

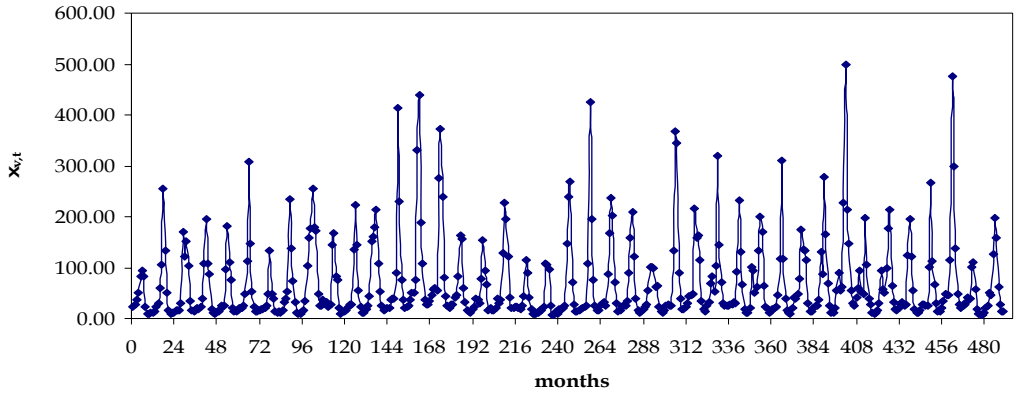

Figure 3.8 Time Series of Monthly Streamflow of EIE 1501 for the Period of 1956-1995.

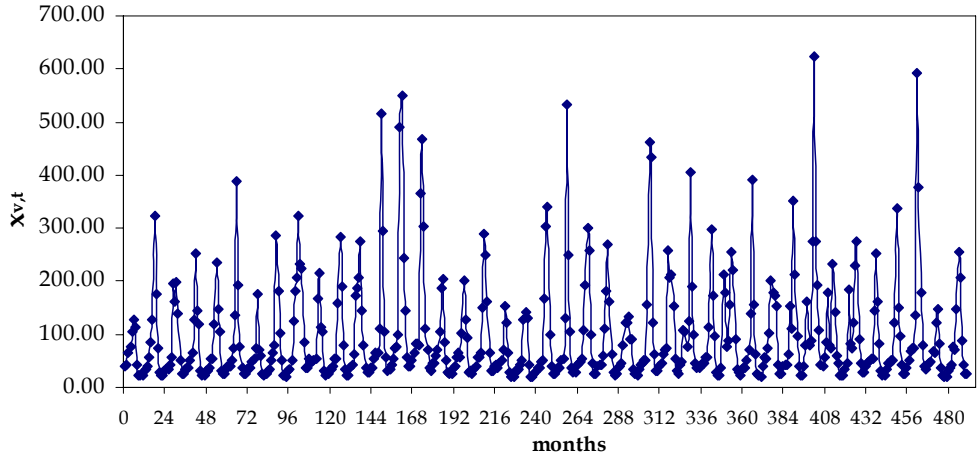

Figure 3.9 Time Series of Monthly Streamflow of EIE 1503 for the Period of 1956-1995.

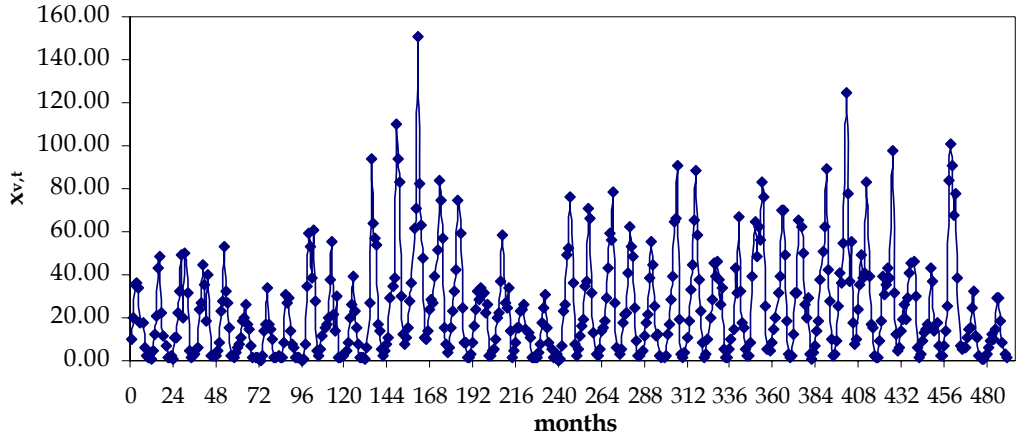

Figure 3.10 Time Series of Monthly Streamflow of EIE 1541 for the Period of 1956-1995.

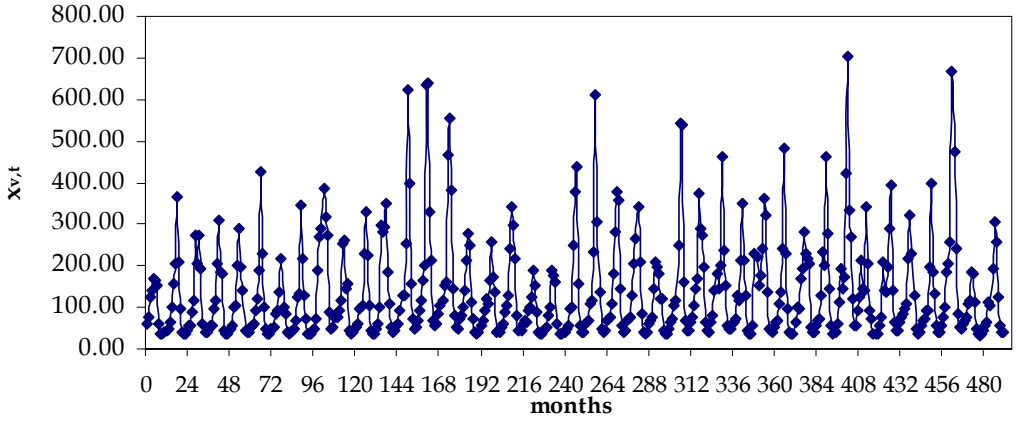

Figure 3.11 Time Series of Monthly Streamflow of EIE 1528 for the Period of 1956-1995.

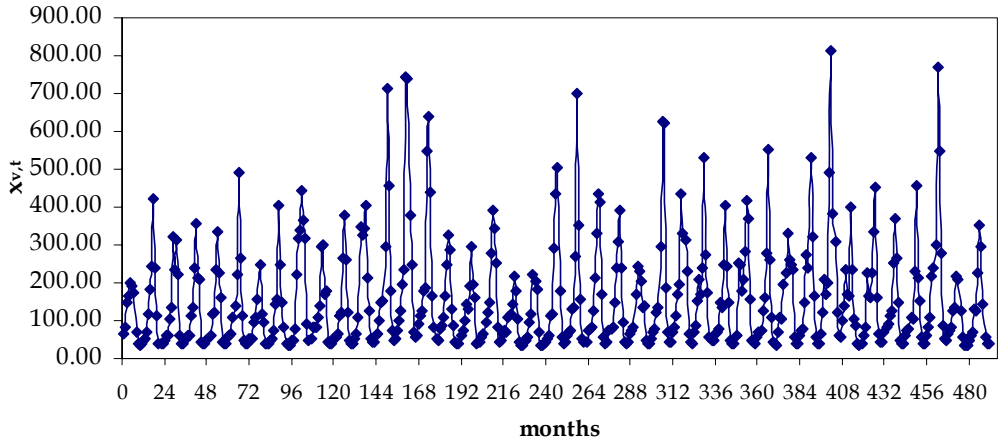

Figure 3.12 Time Series of Monthly Streamflow of EIE 1536 for the Period of 1956-1995.

Periodicities of the streamflow series need to be eliminated to be able to use deseasonalized models. For this purpose, periodic mean and periodic standard deviation are estimated by computing the sample mean $\overline{x}_\tau$ and the sample standard deviation s_τ for each time interval $\tau=1,...,w$ respectively.

In this way, the streamflow series are deseasonalized by the author first by subtracting the monthly mean value and then dividing it by seasonal standard deviations to fit to the AR(1) model as calculated below;

$$y_{v,\tau} = \frac{x_{v,\tau} - \overline{x}_\tau}{s_\tau} \qquad v=1,...,N \ , \quad \tau=1,...,w. \tag{3.10}$$

where

$$\overline{x}_\tau = \frac{1}{N}\sum_{v=1}^{N} x_{v,\tau} \qquad \tau=1,...,w \qquad \text{and} \qquad s_\tau = \left[\frac{1}{(N-1)}\sum_{v=1}^{N}\left(x_{v,\tau} - \overline{x}_\tau\right)^2\right]^{1/2} \qquad \tau=1,\ldots,w.$$

As a result, the periodicities of the original series are removed. Afterwards, the coefficient of skewness and kurtosis values for deseasonalized monthly data and residuals are computed and given in Table 3.8 to have an idea about the distribution of residuals.

Table 3.8. The Coefficient of Skewness and Kurtosis Values of Deseasonalized Monthly Historical Series and Residuals

Stations	**Coefficient of Skewness**		**Kurtosis**	
	Historical series	Residuals	Historical series	Residuals
EIE1501	1.33	1.60	5.19	9.09
EIE1503	1.81	2.19	7.45	13.00
EIE1541	1.18	1.08	4.42	5.62
EIE1528	1.51	1.81	6.07	10.70
EIE1536	1.52	1.82	6.11	10.79

The above tabulated values indicate non-normality of the residual distribution. Using again Jarque-Bera test, normality assumption is rejected as seen in Table 3.9. From tables critical value at 5% level for 2 degrees of freedom is 5.99.

Table 3.9. Jarque-Bera Statistic Values for Each Runoff Data Set

Stations	Jarque-Bera statistic for historic series	Jarque-Bera statistic for residuals	Critical value
EIE1501	241.89	964.31	< 5.99
EIE1503	670.48	2428.38	< 5.99
EIE1541	154.56	234.92	< 5.99
EIE1528	377.86	1475.04	< 5.99
EIE1536	385.37	1506.40	< 5.99

The residual distribution is not normal and positively right skewed. Therefore, the generalized logistic distribution is selected to set AR(1) model for monthly series.

To further examine whether the error distributions fit to the generalized logistic distribution or not, Q-Q plots are used by employing Eq. 2.29. Q-Q plots for each station are constructed for different shape parameter values.

Only one of these Q-Q plots for the station EIE1501 is represented in Figure 3.13 for the shape parameter, b=8. The other Q-Q plots for other stations were given in Appendix A2. Q-Q plots pointed out that the Generalized logistic distribution is viable.

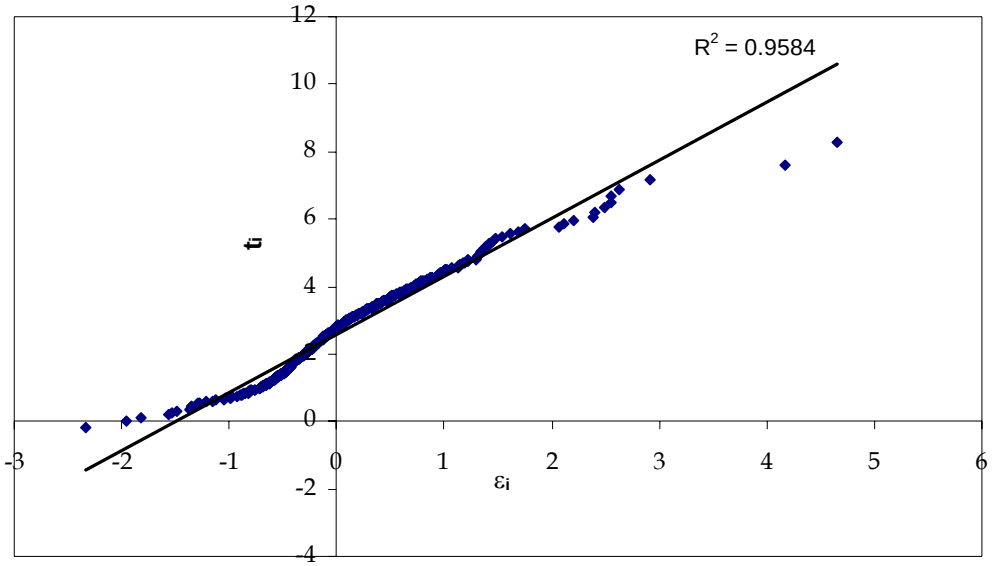

Figure 3.13 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovations for EIE 1501 (b=8)

Examining the Q-Q plots, the three largest data are suspected to be outliers for all stations except for EIE 1541 station. Hence these three data are removed from the data set as outliers.

Q-Q plots are replotted for the rest of data. As seen in Figures 3.14 and A24-A30 at the Appendix A3, the resulting correlation coefficients are higher than the previous values. This situation supports that three removed data are outliers.

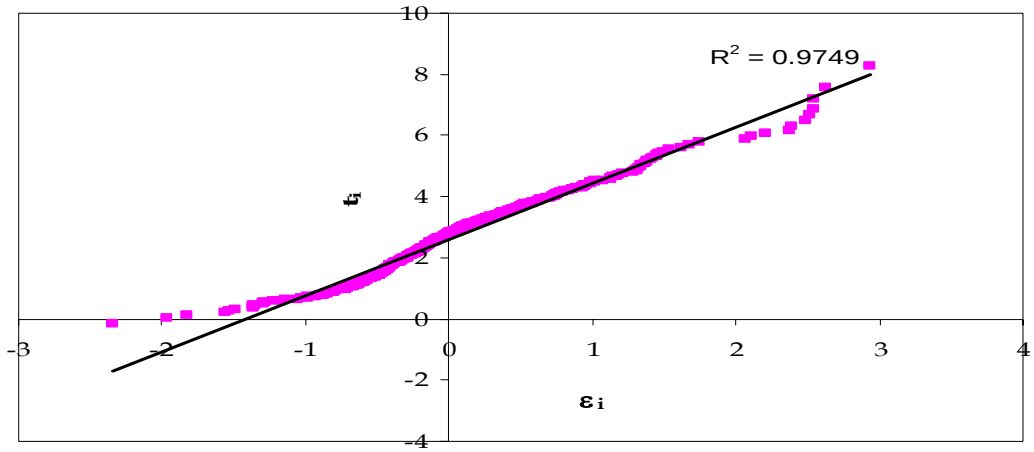

Figure 3.14 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovations for EIE 1501 (b=8) Without Outliers

The coefficient of skewness and kurtosis values are given in Table 3.10 for the reduced data set. As compared with the values in Table 3.8, the coefficient of skewness and kurtosis values are reduced.

Table 3.10. The Coefficient of Skewness and Kurtosis Values of Residuals without Outlier Data

Stations	Coefficient of Skewness for residuals	Kurtosis for residuals
EIE1501	1.04	5.50
EIE1503	1.33	6.41
EIE1528	1.02	5.19
EIE1536	1.02	5.21

According to Jarque- Bera test, this residual also do not come from normal distribution. The results of this test are given in Table 3.11.

Table 3.11. Jarque-Bera Statistic Values for Each Runoff Data Set

Stations	Jarque-Bera statistic for residuals	Critical value
EIE1501	214.61	< 5.99
EIE1503	379.53	< 5.99
EIE1528	181.77	< 5.99
EIE1536	183.55	< 5.99

Therefore, it can be accepted that residual series of the monthly data fit better to the generalized logistic distribution (GL).

3.4.1 MML Estimators for the Model Parameters

The procedure of the AR(1) model application for GL distribution is exactly the same procedure as in Weibull distribution. The obvious alteration is to use formulas for GL distribution. Similarly, the model parameters are estimated by MML procedure. The initial shape parameter values can be selected between 8 and 10 values when the theoretical

coefficient of skewness and kurtosis for generalized logistic distribution given in Table 2.2 are compared to the computed values. These values are chosen as 8 for EIE1501, EIE1541, EIE1528 and EIE1536 and 10 for EIE1503 according to Q-Q plots for each station. All of the model parameters are presented in Table 3.12 with LS estimators.

Table 3.12. Estimated Parameters from AR(1) Model with Generalized Logistic Innovations for Each Gauging Station

Stations	**LS Estimators**			**MML Estimators**			
	(λ_0)	**($\hat{\phi}_0$)**	**(α_0)**	**(b)**	**(λ)**	**($\hat{\phi}$)**	**(α)**
EIE 1501	-1.4476	0.6575	0.5544	8	-1.3420	0.5257	0.5142
EIE 1503	-1.5942	0.6304	0.5591	10	-1.4187	0.4630	0.4940
EIE1541	-1.3930	0.7240	0.5365	8	-1.2788	0.5878	0.4898
EIE1528	-1.3804	0.6880	0.5285	8	-1.2696	0.5387	0.4828
EIE1536	-1.3814	0.6872	0.5289	8	-1.2697	0.5375	0.4827

It is found that all of the estimated parameters maximize the log-likelihood function with initial shape parameter under generalized logistic distribution for each streamflow station. The log-likelihood functions are given for generalized logistic distribution in the Figures 3.15-3.19.

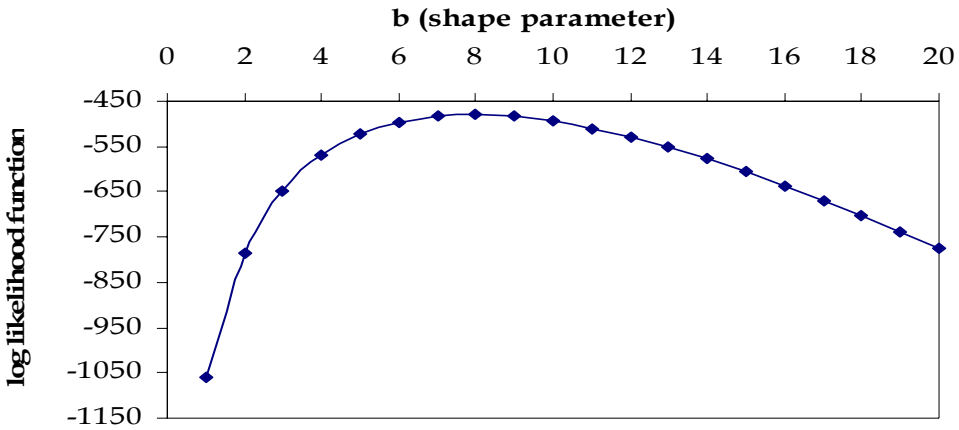

Figure 3.15 Generalized Logistic Log-likelihood Function with respect to the Different Shape Parameters for EIE 1501

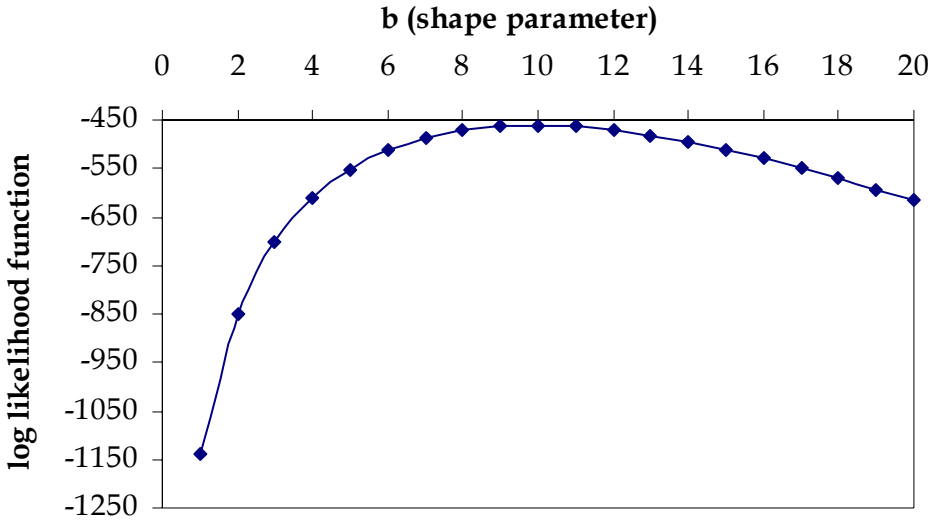

Figure 3.16 Generalized Logistic Log-likelihood Function with respect to the Different Shape Parameters for EIE 1503

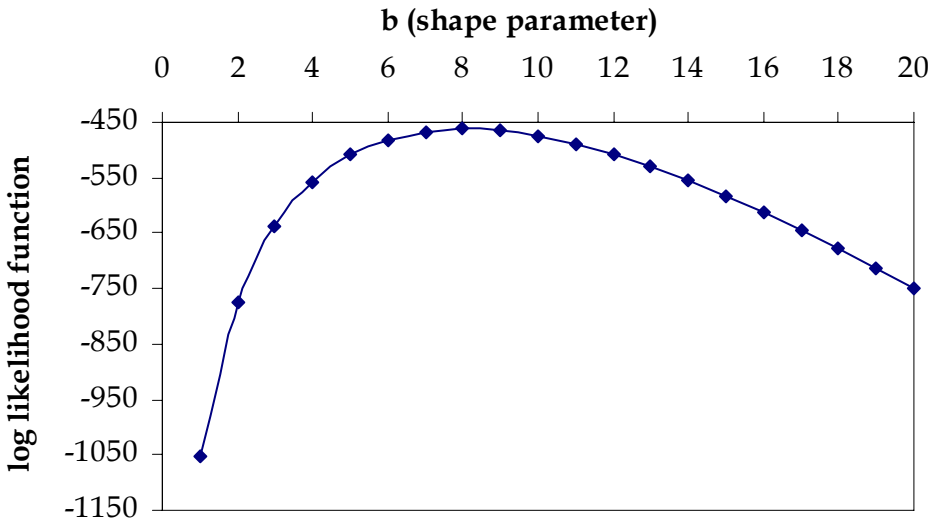

Figure 3.17 Generalized Logistic Log-likelihood Function with respect to the Different Shape Parameters for EIE 1541

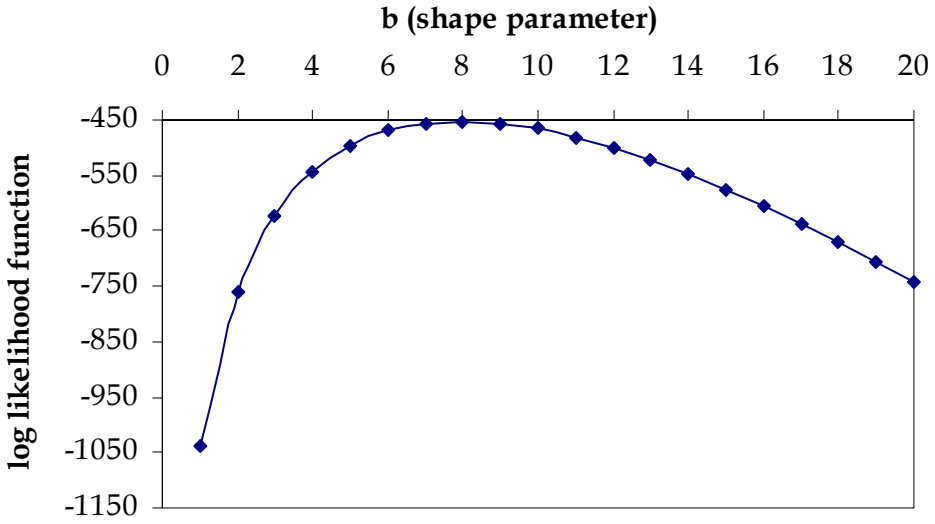

Figure 3.18 Generalized Logistic Log-likelihood Function with respect to the Different Shape Parameters for EIE 1528

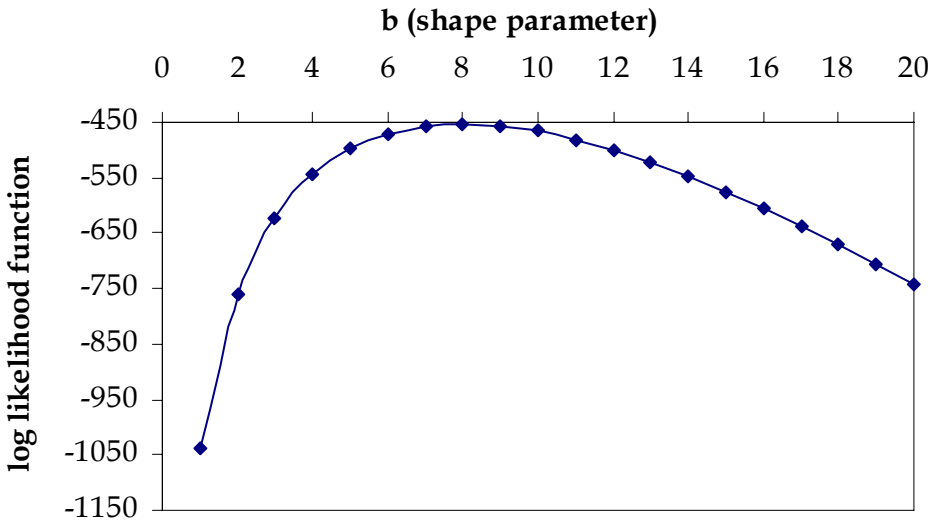

Figure 3.19 Generalized Logistic Log-likelihood Function with respect to the Different Shape Parameters for EIE 1536

3.4.2 Generation Monthly Data Using AR(1) Models with Generalized Logistic Innovations

Having estimated the parameters of the model, the synthetic periodic time series are now generated by substituting the estimated parameters into the Eq. (3.11). As a result, the original seasonal series $x_{v,\tau}$ is obtained from:

$$x_{v,\tau} = \overline{x}_{\tau} + y_{v,\tau} s_{\tau} \tag{3.11}$$

The computations $y_{v,\tau}$ can be done either using constant autoregression coefficients or periodic autoregression coefficients. The constant autoregression coefficients are used here in constructing the synthetic time series by the following formula:

$$y_{v,\tau} = \hat{\phi} y_{v,\tau-1} + \hat{\alpha}\zeta_{v,\tau} \tag{3.12}$$

where $\zeta_{v,\tau}$ is an independent generalized logistic variable computed by Eq. (2.75). Thus, Eq. (3.11) and Eq. (3.12) are used to generate the synthetic periodic series $x_{v,\tau}$. The actual generating procedure is practically the same as in the case of generating annual time series section (3.3.2).

As done before, 1000 series are generated from the AR(1) model with generalized logistic innovations, each having the length of the historical series. The averages of the generated series are presented together with the historical values in the Table 3.13.

Table 3.13 Mean Values of Moments Derived from Generated Monthly Series Based on AR(1) Model with Generalized Logistic Innovations

Stations	Generated Synthetic Series		Historical Series	
	Mean	Std. dev.	Mean	Std. dev.
EIE 1501	68.39	74.23	68.58	78.92
EIE 1503	93.69	87.85	95.31	96.35
EIE 1541	23.92	21.34	23.89	23.43
EIE 1528	133.82	107.8	136.01	116.97
EIE 1536	152.15	126.43	154.59	136.813

The relative error values are also computed for the monthly data set in order to show the reliability of this model as seen in the Table 3.14.

Table 3.14 Relative Errors between Generated and Historical Moment Values for Monthly Data

Stations	Relative errors	
	Mean (%)	Std. dev. (%)
EIE 1501	-0.27	-6.32
EIE 1503	-1.73	-9.68
EIE 1541	0.12	-9.79
EIE 1528	-1.64	-8.51
EIE 1536	-1.60	-8.21

The Table 3.14 indicates that the developed models tend to produce a more conservative decision since the moment values from the model are close to historical values.

3.4.3 Forecasting Monthly Data Using AR(1) Models with Generalized Logistic Innovations

For EIE 1501 station, we still have 84 monthly data observed after 1995. This data is not used in setting up the model but used to assess the forecasting performance of the proposed methodology. Similar procedure used for the case of annual data will be followed now forecasting the observed monthly data set.

Forecasting monthly data series and the observed monthly data series are plotted together against the time axis for comparison in Figure 3.20.

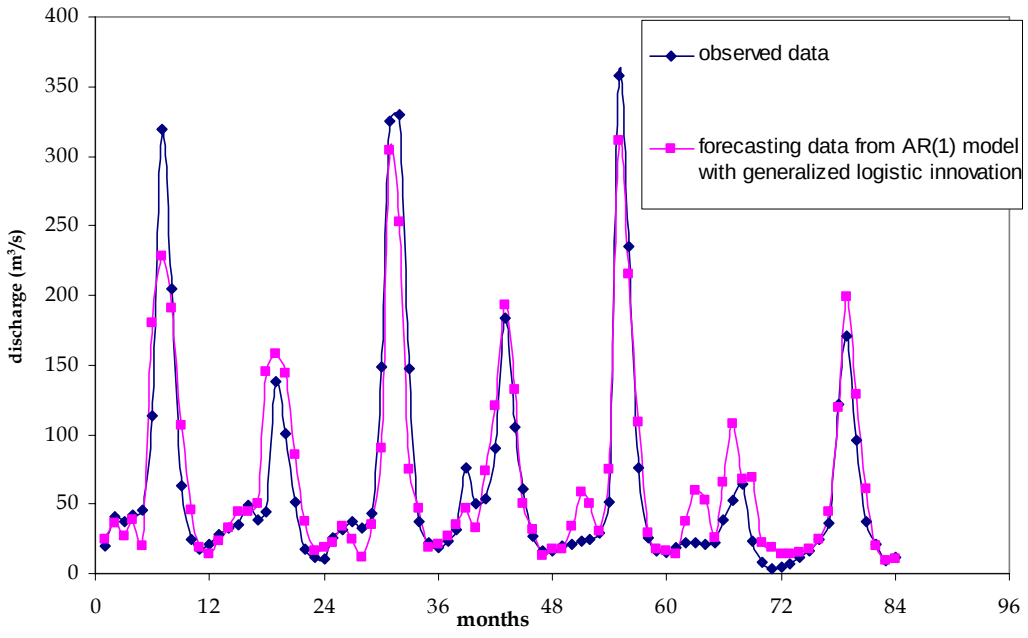

Figure 3.20 Forecasted Monthly Data Series and Observed Monthly Data Series for EIE 1501 Stream Gauging Station

Figure 3.20 shows the performance of AR(1) model with generalized logistic innovations to be a very good fit to the data. This model predicted

the first peak flow as 222.96 m^3/s, while the observed value was 318.8 m^3/s, hence an underestimation of 28.53%. Therefore, AR(1) model with generalized logistic innovations approaches yield reasonably good forecasts for 1-month ahead forecast. The root mean square error (RMSE) term is determined as 28.46 from Eq. (2.79) for the forecasted series. This RMSE value can be considered to indicate a good fit.

Therefore, as a result, this model offers a reasonable alternative for univariate modeling of water resources time series.

3.5 APPLICATION OF ARTIFICIAL NEURAL NETWORK

Using the main concepts of ANN model defined in Section 2.7, following procedures are some of the major principles that have been considered during model development;

- Three-layer FNN used in this study contains two hidden layers. Since the ANN model are used to obtain the annual and monthly simulation streamflow values at time i, the previous flow values at time i-1 are defined as input values of the input layer. The streamflow values at time i are delimited as output values of the output layer. In other words one previous annual or monthly flow are used to simulate the next yearly or monthly mean flows (one time ahead simulating).

- The selection of number of hidden nodes appropriate for the particular application is the most difficult task with ANN model. Since there is no theory yet to tell how many hidden nodes are

needed to approximate any given function, the common trial and error method to select the number of hidden nodes is used. The optimum number of hidden nodes is found as two with trial and error methods. It is observed that an addition of a further node(s) do not provide any improvement.

- The sigmoid function Eq. (2.77) is used as the hidden nodes and the output node activation function. Since the validation of sigmoid function is bounded between 0 and 1, the input data are standardized and scaled to fall in range [0.1, 0.9] before applying the ANN. After the data presented to the model, "MinMax" tables are formed by the software according to sigmoid function selection. If a tanh transfer function is used, the data should be transformed in a bipolar range (eg. -1 and 1).The river flow x_i is standardized by the following formula:

$$x_s = (x/1.24x_{max}) + 0.1 \tag{3.13}$$

where x_s= standardized flow; and x_{max}= maximum of the flow values. After standardizing the flow values, the network should be trained by using a set of the learning cases. In the literature, it is proposed that 75% of the training set is used for training or learning. In this application, 40 annual data and 491 monthly data are selected separately for training and 7 annual data and 84 monthly data are used for testing case to obtain the synthetic series from ANN model.

- Learning rate, which is the rate at which weights are modified during backpropagation, should be high enough to speed up learning. However, although high learning rate means rapid

learning, it may push training towards a local minimum or may cause oscillation. On the other hand, small learning rate requires a longer time to reach a minimum value and one could be training forever.

- To solve this problem, a momentum factor is proposed to be applied which is multiplied by the previous weight change so that while learning rate is kept low, changes are still rapid. It is proposed that by choosing different learning rate coefficients for each layer, one can optimize the network performance.

  Backpropagation requires that learning rates approaching to zero should be used but as it has mentioned previously, smaller rates increases the training time. To decrease the initial learning rates, learning coefficient ratio is used in neuralware program. Learning coefficient is reduced from the initial learning coefficient by an amount corresponding to learning coefficient ratio until training time. Hence, initially high learning rates are selected as 0.3 for hidden layer and 0.15 for output layer. However after a training time it is observed that the learning rates are 0.00001 for hidden layer and 0.00117 for output layer for EIE 1501.

- Training time is application specific and also depends on performance level expected from the network. Testing phase is one way of determining how well the network had learned. Thus, root mean square error (RMSE) is used to measure how well the network performance. This method is measured and monitored performance of the model during both training and testing. RMSE of the error is calculated using Eq. (2.79). Smaller the error represents the better

performance. However, RMSE do not always mean superior performance because of "overtraining". Overtraining could be detected when RMS error of training is high but testing error is substantially lower than this. If this situation occurs, the training data set is changed. Then the training set is used in testing stage to evaluate the accuracy of the ANN models. After construction of the initial ANN model, the testing set is introduced to model to test the performance of the ANN model. Obtained data from the testing stage are determined as forecasting data from the model.

Considering the above criteria, forecasting annual and monthly data series are obtained by using neural ware packet program. Firstly, data set are standardized by using Eq. (3.13). Secondly, input file and test file are constituted for using the neural ware program. After determining the learning coefficient and momentum coefficient, the neural ware program is run in order to obtain forecasting data.

The outputs of test file are accepted as forecasting data set. After that, obtained standardized series are converted to annual and monthly data set by using Eq. (3.13).

Seasonal effects are also removed from monthly data set before using it in ANN model.

Forecasted series and observed annual and monthly data series for EIE 1501 station are plotted together against the time axis, separately in Figures 3.21-3.22.

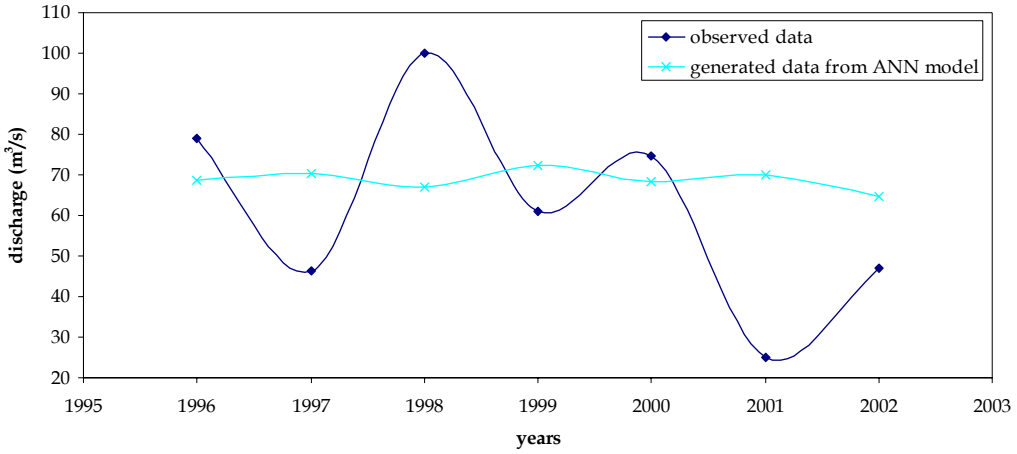

Figure 3.21 Forecasted Annual Data Series and Observed Annual Data Series for EIE 1501 Stream Gauging Station from ANN Model

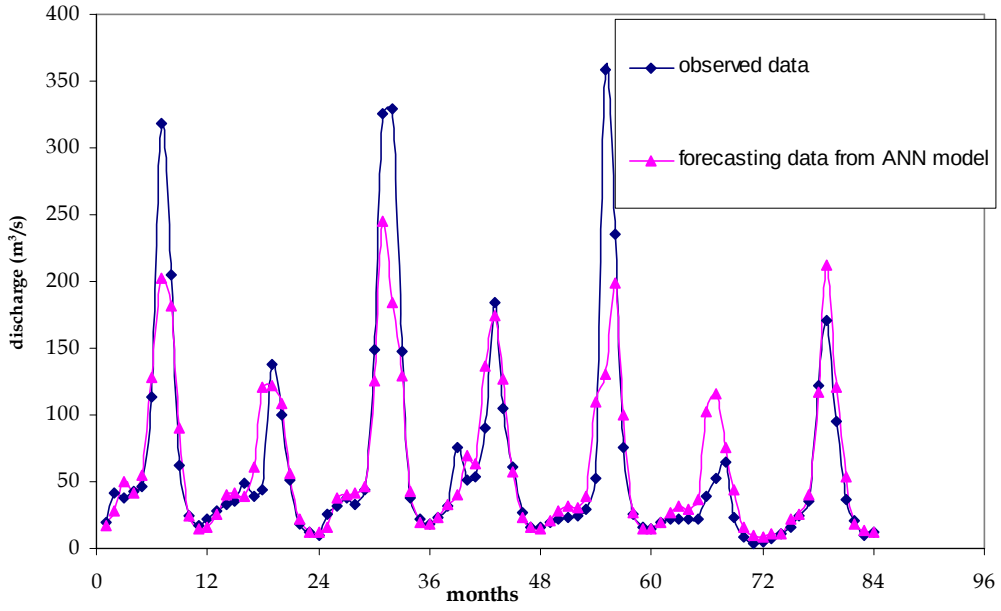

Figure 3.22 Forecasted Monthly Data Series and Observed Monthly Data Series for EIE 1501 Stream Gauging Station from ANN Model

The Figures 3.21 and 3.22 represent the plots of the observed and the computed annual and monthly data set obtained from ANN model. The root mean square error (RMSE) terms for ANN model were as 24.66 for

annual series and 38.58 for monthly series. Both RMSE values for ANN model are higher than the RMSE values of previous models. For the given data set, AR(1) model with asymmetric innovations for both annual and monthly series fit better than ANN model for 1-time ahead forecast.

To further assess the ANN model, it can be seen from these figures that ANN estimates for monthly data better than the annual data. This result may not mean that ANN works better for monthly data. This situation may be resulted from the nature of annual data set taken. More explicitly in the annual observed values, the correlation coefficient between the input and output values for the system is found to be around 0.1. We further anticipate that if this correlation value were higher, ANN could provide as better results for annual values.

On the other hand, the fact is that the correlation coefficients between the previous and the successor streamflow values in monthly case are naturally expected to be higher with respect to the annual correlations. As a result, the ANN model for monthly data naturally better captures the monthly extreme streamflow values than the annual extreme values.

Looking at Figs. 3.21 and 3.22 the forecasting results from ANN procedure are not satisfactory as there is a big difference between the forecast and the observed data.

CHAPTER 4

SUMMARY AND CONCLUSIONS

Autoregressive models are stochastic in nature. They have been utilized in hydrology for a long time. In autoregressive models, all influential physical parameters are not required in modeling the occurrences of the events. Applying time series approach to only one variable enables the stochastic model for forecasting the past and future events. Some assumptions are required to be made to express and model the hydrological events to be resolved.

In most research, the autoregressive models are based upon one major assumption that the residual series distribute normally. The main reason of making this assumption is the difficulties faced in finding the model parameters in the distributions when the distribution is not normal. In this study, the non-normality case has been investigated thoroughly.

There are various parameter estimation methods utilized in finding the model parameters. As of one these methods, maximum likelihood procedure is more unbiased and more efficient than the other procedures in finding the model parameters. However, if the distributions are not

normal, model parameters can not be found out exactly by maximum likelihood method. Iterative solution is required. Still problems arise in iterative methods such as the drawback of convergence of the model parameters to their real values. Maximum likelihood procedure as shown by Tiku et al. can be modified in a way that iterative method is not needed.

In this study, the modified maximum likelihood procedure is introduced into the area of hydrology to estimate the autoregressive model parameters of non-normally distributed series. We examine necessary formulations for the autoregressive model parameters with skew distributions such as gamma, Weibull and generalized logistic distributions.

Furthermore, this study looks into artificial neural network models as another stochastic model utilized in the generation of synthetic series without requiring normality assumption and compares with the autoregressive models.

After introducing these models in hydrology, consequently, they are applied for the annual and monthly streamflow data of five streamflow gauging stations located in Kızılırmak basin.

The conclusions of this study may be summarized as follows:

- Normality assumption in autoregressive models is not an appropriate assumption either for annual or monthly hydrological data utilized in this study. The fact that the data are quite skewed result in questioning how reliably forecasting the streamflow values can be done by using the models with normality assumption.

- In sum, for reliability which is an essential issue in the dam planning, the skewness of the streamflow series needs to be seriously taken into account in setting up their corresponding models.

- GAR(1) model was generally proposed to represent non-normally distributed streamflow data. The observed data from Kızılırmak Basin is not normally distributed and moreover, do not come from gamma distribution. Therefore, GAR(1) model does not function as a viable alternative to study the Kızılırmak Basin data. In such cases, AR(1) model with asymmetric innovations is proved to work better to represent streamflow data series.

- It is found out that while the error terms obtained from annual data set for selected streamflow gauging stations in Kızılırmak basin are fit to Weibull distribution by means of goodness of the fit tests, the error terms for monthly series are fit to generalized logistic distribution.

- It is shown that all of the estimated model parameters under Weibull distribution for each annual streamflow series and generalized logistic distribution for each monthly streamflow series maximize the likelihood function with initial shape parameter. As known, likelihood function is based on the assumption of maximization of the likelihood function by all parameters.

- The model parameters pertaining to the aforementioned distributions can be determined by means of MML method in an

unbiased, efficient and robust manner (Akkaya and Tiku, 2001; 2005).

- When the moment values are compared for the synthetic series determined in generation section with the ones of the historical data, it is obtained that they are quite close to each other.

- In forecasting section, studying the observed data not utilized in setting up of the model parameters, both visual (Figure 4.7 and 4.20) and statistical procedures are applied for comparison of synthetic and historical series. The root mean square error values for AR(1) model are calculated for the purpose of statistical comparison. The error amount is found to be low. This result points out that the deviation of the minimum and maximum streamflow values are low. In reality, the root mean square error amount is not expected to come out as zero. Therefore, the water structure systems to be constructed by making use of this system will be reasonably acceptable.

- It is found out that the values determined by means of AR(1) models are better than the ANN model in especially the annual streamflow data.

- As a final conclusion, it can be stated that AR(1) model with asymmetric innovations constructed by making use of MML method is pretty good in determining the synthetic and forecasting streamflow values, and is a unprecedented and reliable method in the area of stochastic hydrology.

SUGGESTED FUTURE STUDIES

In this thesis, two entropy methods under non-normal distribution, AR(1) model with asymmetric innovations and ANN model were applied to annual and monthly streamflow series in Kızılırmak basin. The following studies may also be suggested as future research topics:

- Entropy methods can also be applied for unaffected natural series for selected basins using more number of streamflow gauging stations after determining appropriate distribution(s) for streamflow series.

- The formulations of bivariate density functions for some distributions (generalized logistic etc.) can be developed and applied.

- AR(1) model with asymmetric innovations may be applied to daily unaffected streamflow series (natural) and compared with ANN model.

- The sensibility analysis can be made to evaluate the model performance. For this aim, the error values are plotted together against the time or forecasting series axis.

- The AR(1) models considered here include only one variable. They need to be extended to more than one variable. In fact, this extension will be of enormous interest from a hydrological point of view.

- AR(p) and ARMA(p,q) models with asymmetric innovations can be applied to annual and monthly streamflow series.

- Annual, monthly or daily streamflow data can be forecasted considering with physical parameters (precipitation, evaporation and etc.) by using ANN model.

- When the streamflow values for two stations located upstream have exceeded a threshold level, the streamflow values at the downstream station exceeding a threshold level can be computed by using ANN model and by using bivariate and multivariate conditional probability approach. Therefore the flood discharges of the downstream station in case flood occurs in the runoff stations located at two different upstream branches can be determined.

REFERENCES

Akkaya, A.D. and Tiku, M.L. (2001a), "Estimating Parameters in Autoregressive Models in Non-normal Situations: Asymmetric Innovations." Communications in Statistics Theory and Methods, 30 (3), 517-536.

Akkaya, A.D. and Tiku, M.L. (2001b), "Corrigendum: Time Series Models with Asymmetric Innovations." Communications in Statistics Theory and Methods, 30 (10), 2227-2230.

Akkaya, A.D. and Tiku, M.L. (2005), "Time Series AR(1) Model for Short Tail Distribution." Statistics, 39(2), 117-132.

Amorocho, J. and Espildora, B. (1973), "Entropy in the Assessment of Uncertainity of Hydrologic Systems and Models." Water Resources Research, 9(6), 1511-1522.

Ang, A.H.S and Tang, W.H. (1975), "Probability Concepts in Engineering Planning and Design." Volume 1, John Wiley & Sons.

Bowman, K. O and Shenton, L. R. (1975), "Omnibus test contours for departures from normality based on pb1 and b2", Biometrika, 62, 243-250.

Box, G. E. P., and Jenkins, G. M. (1976), "Time Series Analysis: Forecasting and Control." Holden-Day, San Francisco, Calif.

Brass, R.L. and Rodriguez-Iturbe I. (1985), "Random Functions and Hydrology." Addison- Wesley Publishing Company.

Campolo, M., Andreussi, P. and Soldati, A. (1999), "River Flood Forecasting with a Neural Network Model." Water Resources Research, 35(4), 1191-1197.

Caselton, W. F. and Husain T. (1980), "Hydrologic Networks: Information Transinformation." Journal of the Water Resources Planning and Management Division, ASCE, 106, WR2, July, 503-529.

Castellano-Mendez, M., Gonzalez-Mateiga, W., Febrero-Bande, M., Prada-Sanchez, J. M. and Lozano-Caldero, R. (2004), "Modeling of the Monthly and Daily Behaviour of the Runoff of the Xallas River Using Box-Jenkins and Neural Networks Methods." Journal of Hydrology, Elsevier, 296, 38-58.

Cheng, X. and Noguchi, M. (1996), "Rainfall-runoff Modelling by Neural Network Approach." In: Proc. Int. Conf. Water Reour. Environ. Res., vol. II, October 29-31, Kyoto, Japan.

Chang, F-J and Chen, Y-C. (2001), "A Counterpropagation Fuzzy-Neural Network Modeling Approach to Real Time Streamflow Prediction." Journal of Hydrology, 245, 153-164.

Cığızoğlu, K. and Bayazıt, M. (1998), "Application of Gamma Autoregressive Model to Analysis of Dry Periods." Journal of Hydrologic Engineering, 3(3), 218-221.

Clair, T. A. and Ehrman, J.M. (1998), "Using Neural Networks to Assess the Influence of Changing Seasonal Climates in Modifying Discharge, Dissolved Organic Carbon, and Nitrogen Export in Eastern Canadian Rivers." Water Resources Research, 34(3), 447-455.

Çengel, Y.A. (1997), "Introduction to Thermodynamics and Heat Transfer." The McGraw-Hill Companies, Inc, New York.

Dikmen, İ. (2001), "Strategic Decision Making in Construction Companies: an Artificial Neural Network Based Decision Support System for International Market Selection." PhD. Thesis, Metu, Ankara, Turkey.

Ebbing, D.D. and Gammon, S. D., (1999), "General Chemistry." Houghton Mifflin Company, Boston, New York, Seventh Eddition.

Evans, J. W., Johnson, R. A. and Green, D. W. (1989), "Two and Three Parameter Weibull Goodness of Fit Tests." Research Paper FPL-RP-493. Madison, WI: US. Department of Agriculture, Forest Service, Forest Products Laboratory, 27p.

Fernandez, B. and Salas, J.D. (1986), "Periodic Gamma Autoregressive Processes for Operational Hydrology." Water Resources Research, 22(10), 1385-1396.

Fernandez, B. and Salas, J.D. (1990), "Gamma Autoregressive Models for Stream- Flow Simulation." Journal of Hydraulic Engineering, 116(11), 1403-1414.

Fiering, M.B. and Jackson, B.B. (1971), "Synthetic Stremflows." Water Resour. Monogr. Ser., vol.1, 98 p., AGU, Washington,D.C.

Gaver, D. P. and Lewis P.A.W. (1980), "First Order Autoregressive Gamma Sequences and Point Process." Adv. Appl. Prob., 12(3), 727-745.

Haan, C.T. (1977), "Statistical Methods in Hydrology." Iowa State University Press.

Hamilton, J.D. (1994), "Time Series Analysis." Princeton University Press. Princeton, New Jersey.

Harmancıoğlu N. (1981), "Measuring the Information Content of Hydrological Process by the Entropy Concept", Ege Üniversitesi, İnşaat Fakültesi Dergisi, Atatürk'ün 100. Doğum Yılı Özel Sayısı. 13-40, İzmir.

Harmancıoğlu, N.B., Yevjevich, V. and Obeysekera, J.T.B. (1985), "Measures of Information Transfer between Variables.", Proceedings of Fourth International Hydrology Symposium Multivariate Analysis of Hydrologic Process. Forth Collins, Colorado State University, 481-499.

Harmancıoğlu, N.B. and Yevjevich, V. (1987), "Transfer of Hydrologic Information Among River Points." Journal of Hydrology, Amsterdam: Elsevier, 91, 103-118.

Harmancıoğlu, N.B., Fistikoglu, O. and Özkul, S. (2003), "Integrated Technologies for Enviromental Monitoring and Information Production." Kluwer Academic Publishers, 496pp.

Hipel, K. W. (1986), "Time Series Analysis in Perspective.", Water Resources Bulletin, 21 (4), 609-623.

Hsu, K., Gupta, H.V. and Sorooshian, S. (1995), "Artificial Neural Network Modelling of the Rainfall-runoff Process." Water Resources Research, 31 (10), 2517-2530.

Husain, T. (1989), "Hydrologic Uncertainity Measure and Network Design", Water Resources Bulletin, 25, 527-534.

Imrie, C. E., Durucan, S. and Korre, A. (2000), "River Flow Prediction Using Artificial Neural Networks: Generation Beyond the Calibration Range." Journal of Hydrology, 233, 138-153.

Jain, S.K., Das, D. and Srivastava, D.K. (1999), "Application of ANN for Reservoir Inflow Prediction and Operation." Journal of Water Resources Planning and Management, ASCE, 125(5), 263-271.

Kang, K.W., Park, C.Y. and Kim, J.H. (1993), "Neural Network and its Application to Rainfall-runoff Forecasting." Korean J. Hydroscience, 4, 1-9.

Karunanithi, N., Grenney, W.J., Whitley, D. and Bovee, K. (1994), "Neural Networks for River Flow Prediction." J. Comp. Civil Engineering ASCE, 8(2), 201-220.

Kirby, W. (1972), "Computer –oriented Wilson Hilferty Transformation that Preserves the First Three Moments and Lower Bound of the Pearson Type 3 Distribution." Water Resources Research, 8(5), 1251-1254.

Kişi, Ö. (2003), "River Flow Modeling Using Artificial Neural Networks." Journal of Hydrologic Engineering, ASCE, Technical Notes, 9(1), 60-63.

Kitanidis, P.K. and Bras, R.L. (1980), "Adaptive Filtering through Detection of Isolated Transient Errors in Rainfall-runoff Models." Water Resources Research, 16(4), 740-748.

Kumar, M., Raghuwanshi, N.S., Singh, R., Wallender W.W. and Pruitt, W.O. (2002), "Estimating Evapotranspiration Using Artificial Neural Network." Journal of Irrigation and Drainage Engineering, 128 (4), 224-233.

Lawrance, A.J. and Lewis, P.A.W. (1981), "A New Autoregressive Time Series Model in Exponential Variables [NEAR(1)]." Adv. Appl. Prob., 13(4), 826-845.

Lawrance, A.J. (1982), "The Innovation Distribution of a Gamma Distributed Autoregressive Process." Scand J. Statist, 9, 234-236.

Lawrance, A.J. and Lewis, P.A.W. (1990), "Reversed Residuals in Autoregressive Time Series Analysis." Journal of Time Series Analysis, 13 (3), 253-266.

Lorrai, M., Sechi, G. M. (1995), "Neural Nets for Modeling Rainfall-Runoff Transformations." Water Resources Man. 9, 299-313.

Maier, H. R., Dandy, G.C. (1996), "The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters." Water Resources Research, 32 (4), 1013-1022.

Markus M., Knapp H. V. and Tasker G.D. (2003), "Entropy and Generalized Least Square Methods in Assessment of the Regional Value of Streamgages.", Journal of Hydrology 283, 107-121.

Matalas N.C. (1967), "Mathematical Assessment of Synthetic Hydrology." Water Resources Research, 3(4), 937-945.

McMahon, T. A. and Miller, A.J. (1971), "Application of the Thomas and Fiering Model to Skewed Hydrologic Data." Water Resources Research, 7(5), 1338-1340.

McMurry, J. and Fay, R. C. (2001), "Chemistry." Prentice – Hall, Inc. Upper Saddle River, New Jersey.

Moradkhani, H., Hsu, K., Gupta, H. V. and Sorooshian, S. (2004), "Improved Streamflow Forecasting Using Self-Organizing Radial Basis Function Artificial Neural Networks." Journal of Hydrology, Elsevier, 295, 246-262.

Obeysekera, J.T.B. and Yevjevich, V. (1985), "A Note on Simulation of Samples of Gamma – Autoregressive Variables." Water Resources Research, 21(10), 1569-1572.

O' Connell, P. E. (1974), "Stochastic Modeling of Long-term Persistence in Streamflow Sequences.", Rep. 1974-2, 284 pp., Hydrol. Sect. Dep. of Civ. Eng., Imperial College, London.

Özkul S. (1996), "Space / time Design of Water Quality Monitoring Networks by the Entropy Method.", Ph. D thesis, Dokuz Eylül University, Faculty of Civil Engineering, İzmir, 169 pages.

Rajurkar, M. P., Kothyari, U. C., and Chaube, U. C. (2004), "Modeling of the Daily Rainfall-Runoff Relationship with Artificial Neural Network." Journal of Hydrology, Elsevier, 285, 96-113.

Raman, H. and Sunilkumar, N. (1995), "Multivariate Modeling of Water Resources Time Series Using Artificial Neural Networks." Hydrologic Sciences, 40 (2), 145-163.

Saad, M., Bigras, P., Turgeon, A. and Duquette, R. (1996), "Fuzzy Learning Decomposition for the Scheduling of Hydroelectric Power Systems." Water Resources Research, 32(1), 179-186.

Sajikumar, N. and Thandaveswara, B. S. (1999), "A Non-linear Rainfall-Runoff Model Using an Artificial Neural Network." Journal of Hydrology, 216, 32-55.

Salas, J.D., Delleur J.W., Yevjevich V. and Lane W.L. (1980), "Applied Modeling of Hydrologic Time Series." Water Resources Publications.

Shamseldin, A.Y. (1997), "Application of Neural Network Technique to Rainfall- Runoff Modelling." Journal of Hydrology, 199, 272-294.

Shannon, C. E. (1948a), "A Mathematical Theory of Communications, I and II." Bell System Tech. Journal, 27, 379-423.

Shannon, C. E. (1948b), "A Mathematical Theory of Communications, III and IV." Bell System Tech. Journal, 27, 379-423.

Shin H.S. and Salas, J.D. (2000), "Regional Drought Analysis Based on Neural Networks." Journal of Hydraulic Engineering, ASCE, 5(2), 145-155.

Sim C. H. (1987), "A Mixed Gamma ARMA(1,1) Model for River Flow Time Series." Water Resources Research, 23 (1), 32-36.

Sivakumar, B., Jayawardena, A.W., and Fernendo, T.M.K.G. (2002), "River Flow Forecasting: Use of Phase Space Reconstruction and Artificial Neural Networks Approach." Journal of Hydrology, 265, 225-245.

Smith, J. and Eli, R.N. (1995), "Neural Network Models of Rainfall-Runoff Process" Journal of Water Resources Planning Management, 499-508.

Sorooshian, S., Daun, Q., Gupta, V.K. (1993), "Calibration of Rainfall-runoff Models: Application of Global Optimization to the Sacramento Soil Moisture Accounting Model." Water Resources Research, 29(4), 1185-1194.

Thirumalaiah, K. and Deo, M. C. (2000), "Hydrological Forecasting Using Neural Networks." Journal of Hydrologic Engineering, ASCE, 5(2), 180-189.

Thomas, H. A., Jr., and Fiering, M. B. (1962), "Mathematical Synthesis of Streamflow Sequences for Analysis of River Basins by Simulation, in the design of water resources system, pp. 459-493, edited by A. Maas, et. al." Harward University Press, Cambridge, Mass.

Tiku, M. L. (1967), "Estimating the mean and standard deviation from a censored normal sample." Biometrica 54: 155-165.

Tiku, M.L. and Singh, M. (1981), "Testing the Two Parameter Weibull Distribution." Communications in Statistics, Part A- Theory and Methods. 10: 907-917.

Tiku, M.L., Wong, W.K., Vaughan, D.C. and Bian, G. (1996), "Time Series Models in Non-Normal Situations: Symmetric Innovations." Journal of Time Series Analysis, 21(5), 571-596.

Tiku, M.L., Wong, W.K. and Bian, G. (1999a), "Time Series Models with Asymmetric Innovations." Communications in Statistics Theory and Methods, 28(6), 1331-1360.

Tiku, M.L., Wong, W.K. Vaughan, D.C. and Bian, G. (2000), "Time Series Models in Non-normal Situations: Symmetric Innovations." Journal Time Series Analysis, 21(5), 571-596.

Tiku, M.L. and Akkaya A.D. (2004), "Robust Estimation and Hypothesis Testing." New Age International Publishers (P): New Delhi.

Tokar, A. S. and Johnson, P. A. (1999), "Rainfall- Runoff Modeling Using Artificial Neural Networks." Journal of Hydrologic Engineering, ASCE, 5(2), 156-161.

Tokar, A. S. and Markus, M. (2000), "Precipitation- Runoff Modeling Using Artificial Neural Networks and Conceptual Models." Journal of Hydrologic Engineering, ASCE, 4(3), 232-239.

Türker, Ö. (2002), "Autoregressive Models: Statistical Inference and Applications." PhD. Thesis, Metu, Ankara, Turkey.

Wallis, J. R. and O'Connell, P. E. (1972), "Small Sample Estimation of r_1." Water Resources Research, 8(3), 707-712.

Wilson, E.B. and Hilferty, M.M. (1931), "Distribution of Chi-square." Proc. Nat. Acad. Sci., 17, 684-688.

Yang, Y. and Burn, D. H. (1994), "An Entropy Approach to Data Collection Network Design." Journal of Hydrology, 157, 307-324.

Yapo, P., Gupta, V.K. and Sorooshian, S. (1996), "Automatic Calibration of Conceptual Rainfall-runoff Models: Sensitivity to Calibration Data." Journal of Hydrology, 181, 23-48.

Zealand, C.M., Burn, D.H. and Simonovic, S.P. (1999), "Short-term Streamflow Forecasting Using Artificial Neural Networks." Journal of Hydrology, 214, 32-48.

Zhu, M.L., Fujita, M., (1994), "Comparisons between Fuzzy Reasoning and Neural Network Methods to Forecast Runoff Discharge." Journal of Hydroscience Hydraulic Engineering, 12(2), 131-141.

APPENDIX A1

Q-Q PLOTS FOR AR(1) MODEL WITH WEIBULL INNOVATION

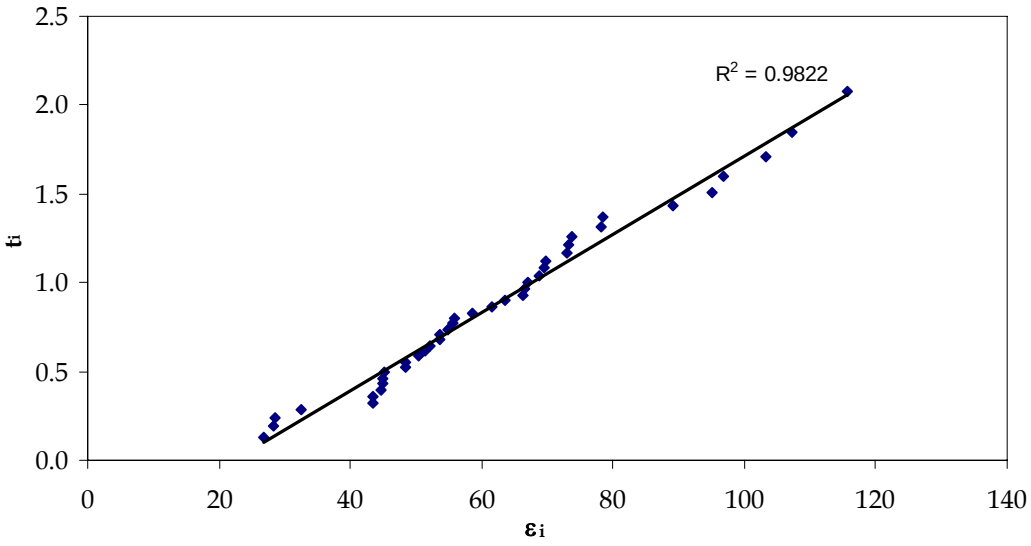

Figure A.1 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1501 (p=1.8)

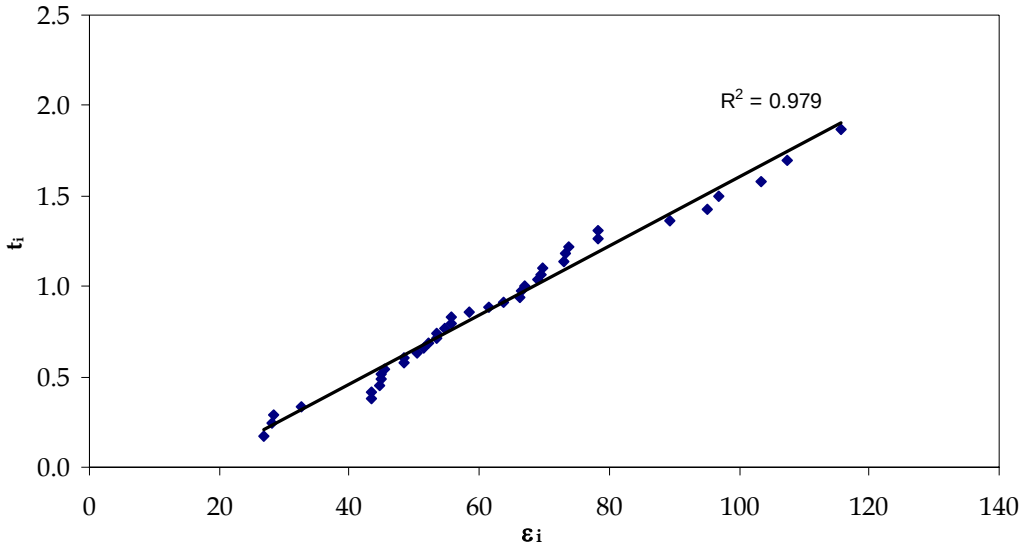

Figure A.2 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1501 (p=2.1)

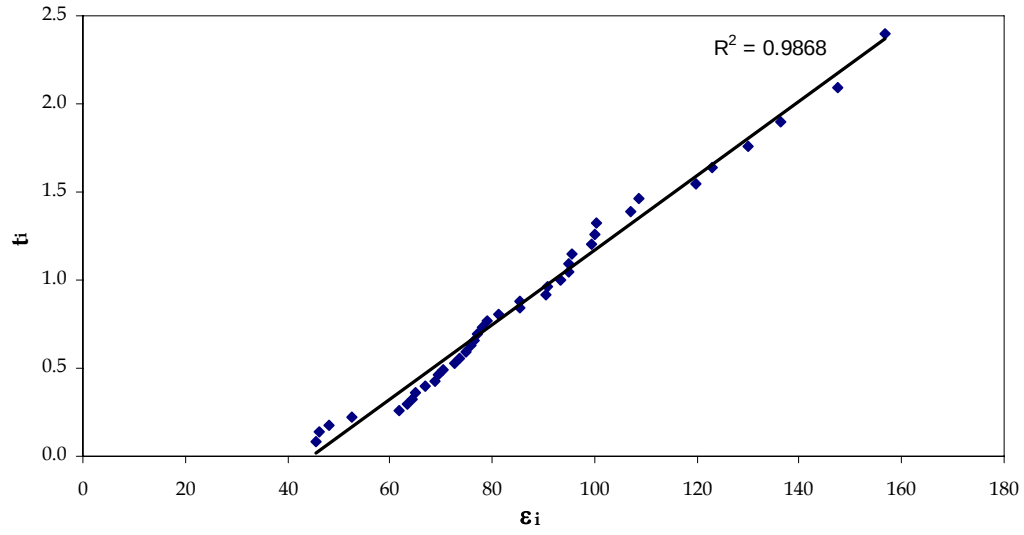

Figure A.3 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1503 (p=1.5)

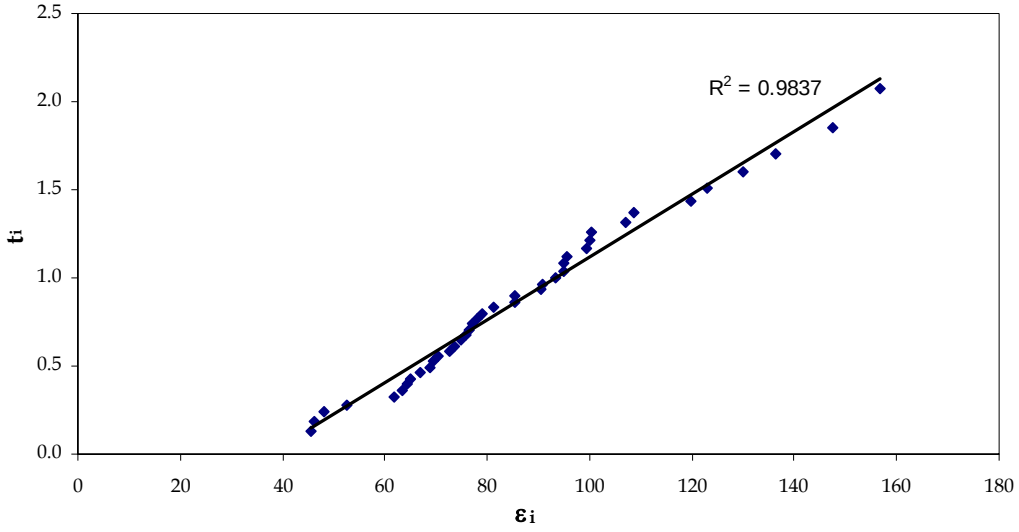

Figure A.4 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1503 (p=1.8)

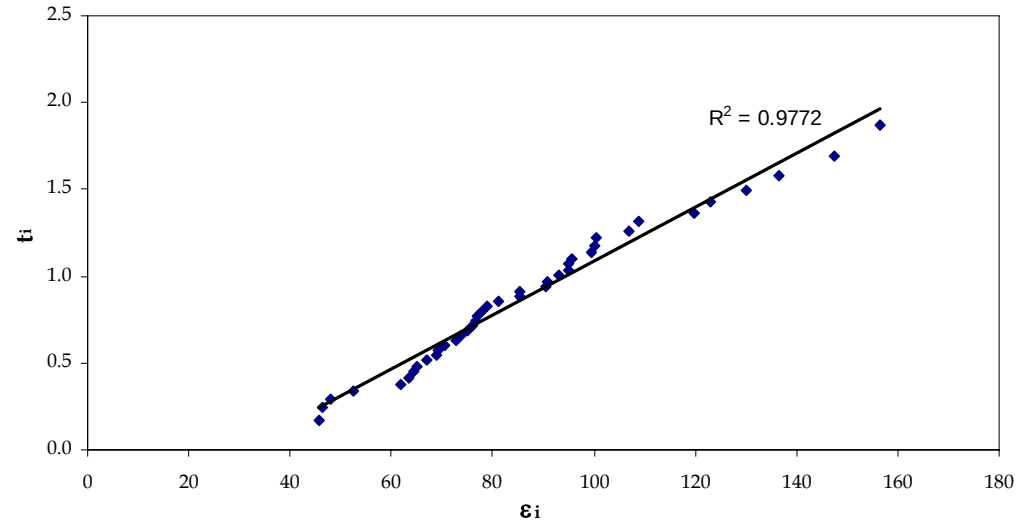

Figure A.5 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1503 (p=2.1)

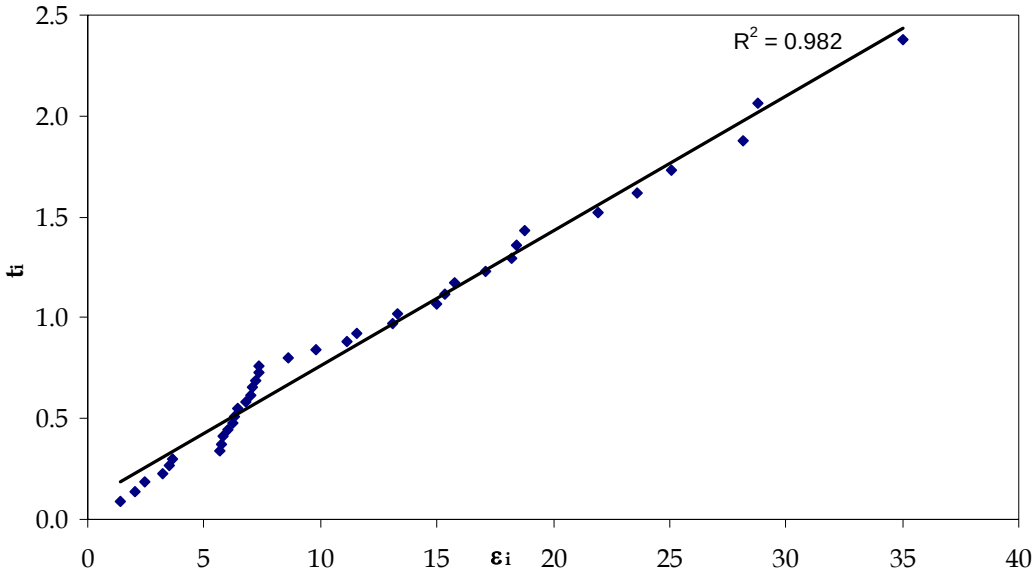

Figure A.6 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1541 (p=1.5)

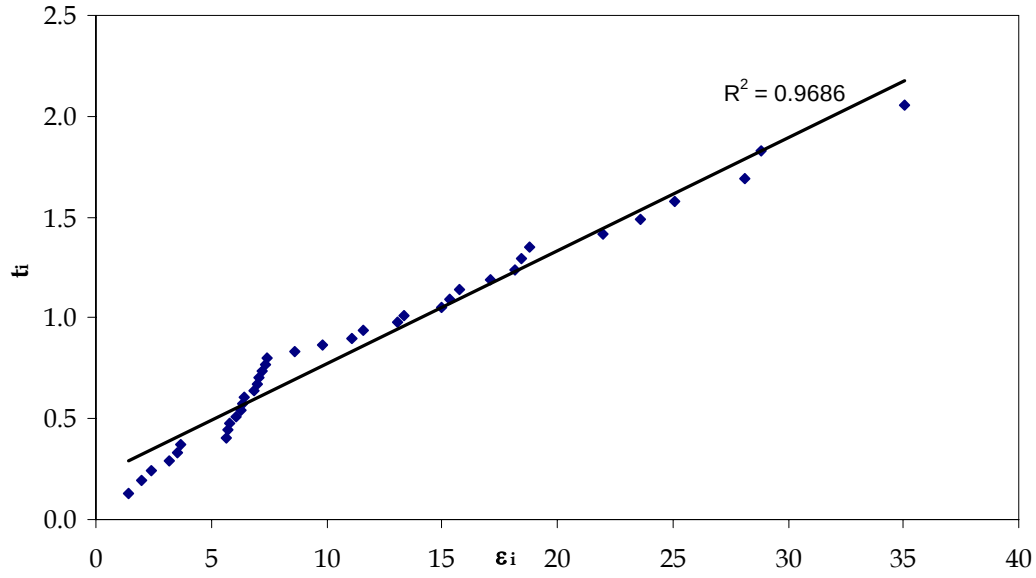

Figure A.7 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1541 (p=1.8)

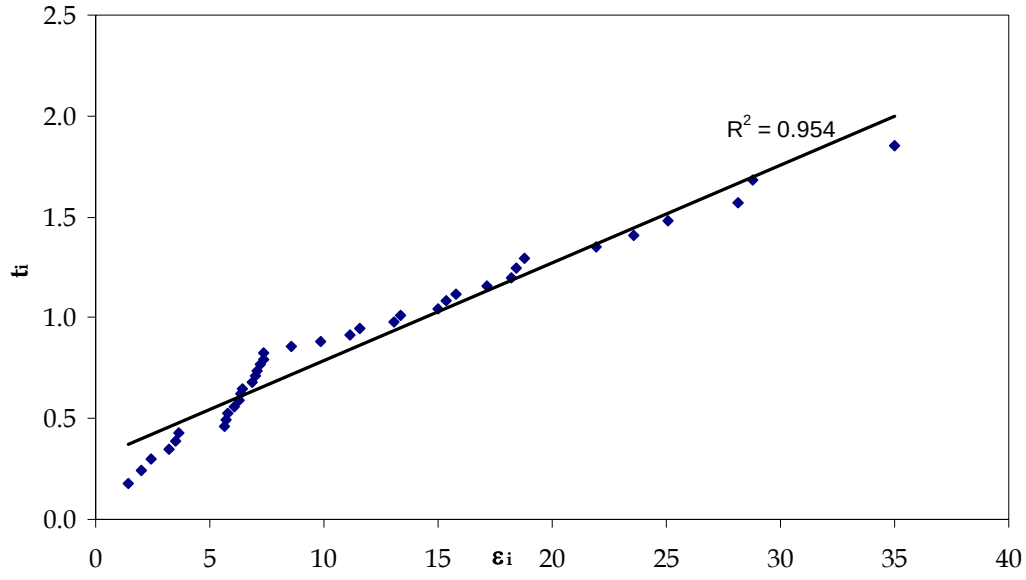

Figure A.8 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1541 (p=2.1)

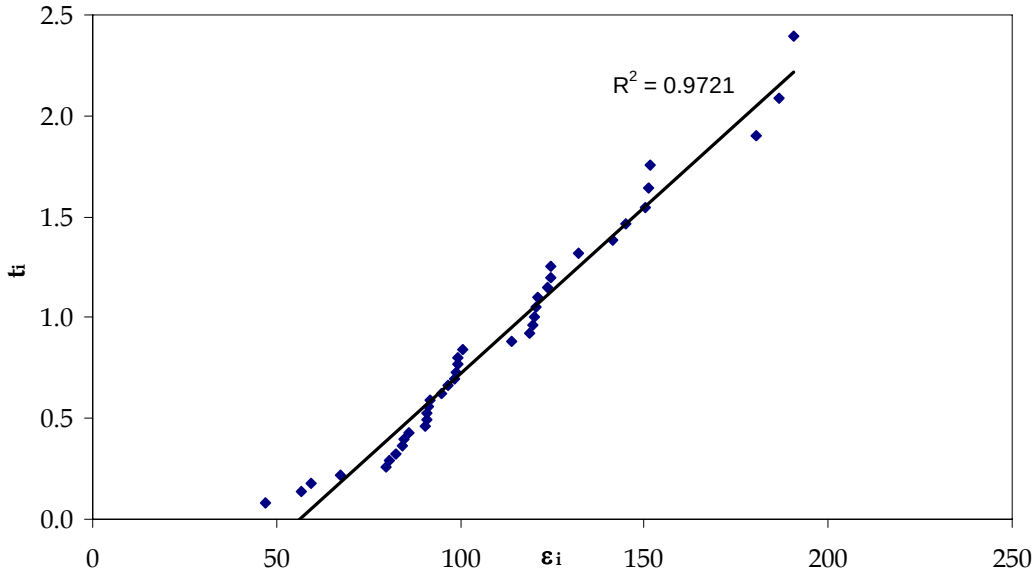

Figure A.9 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1528 (p=1.5)

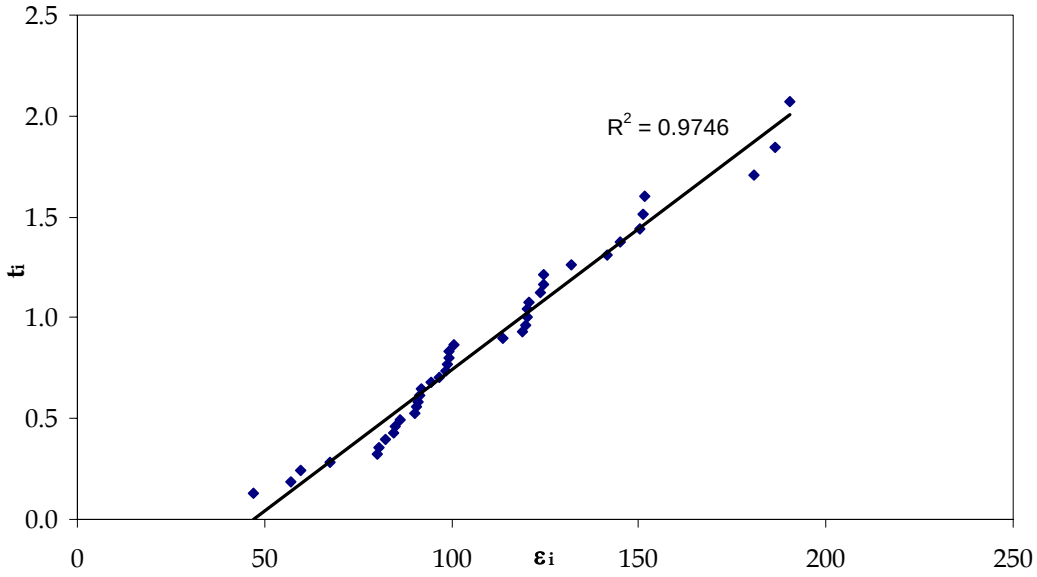

Figure A.10 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1528 (p=1.8)

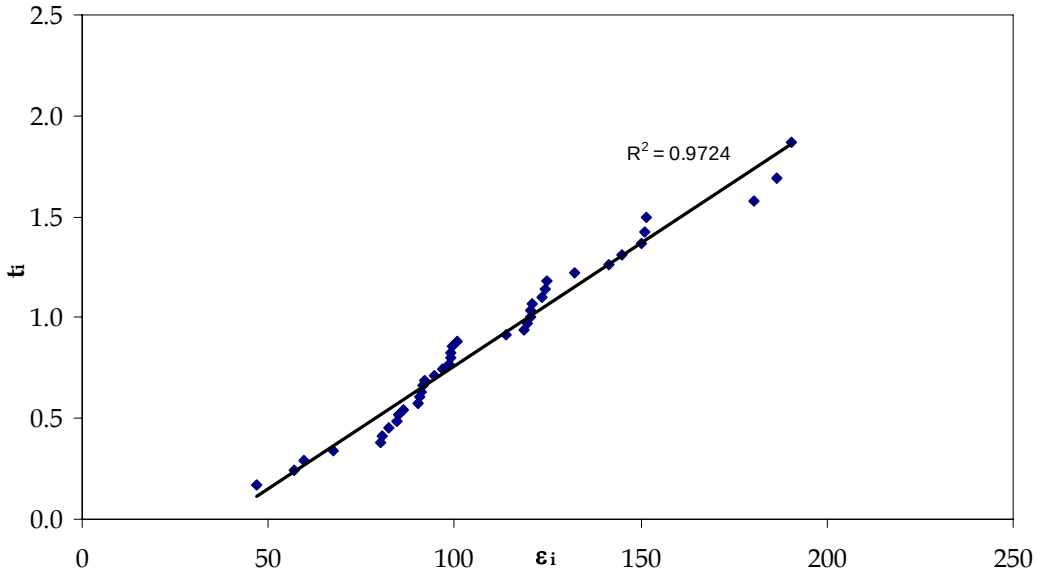

Figure A.11 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1528 (p=2.1)

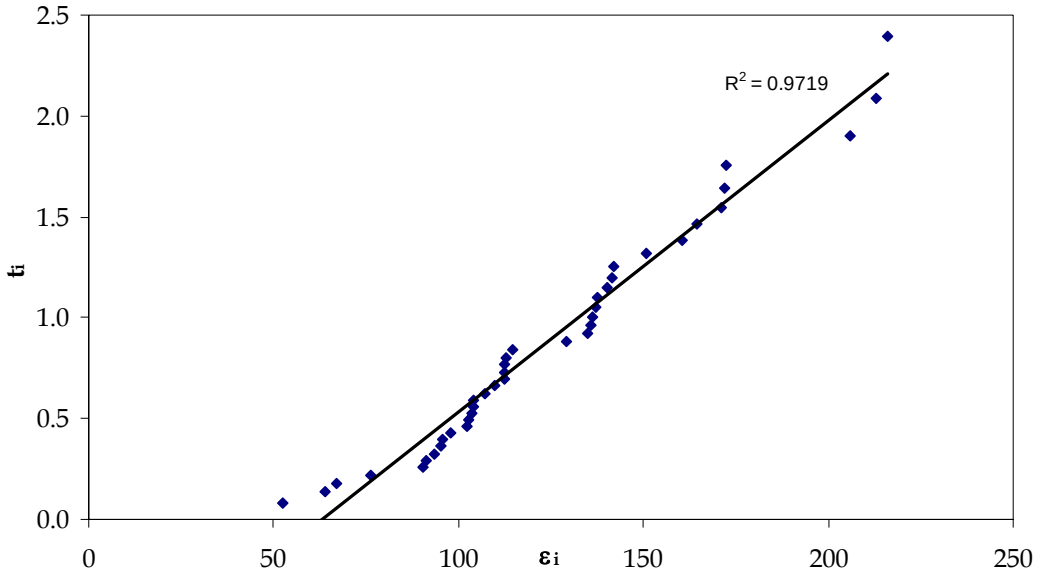

Figure A.12 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1536 (p=1.5)

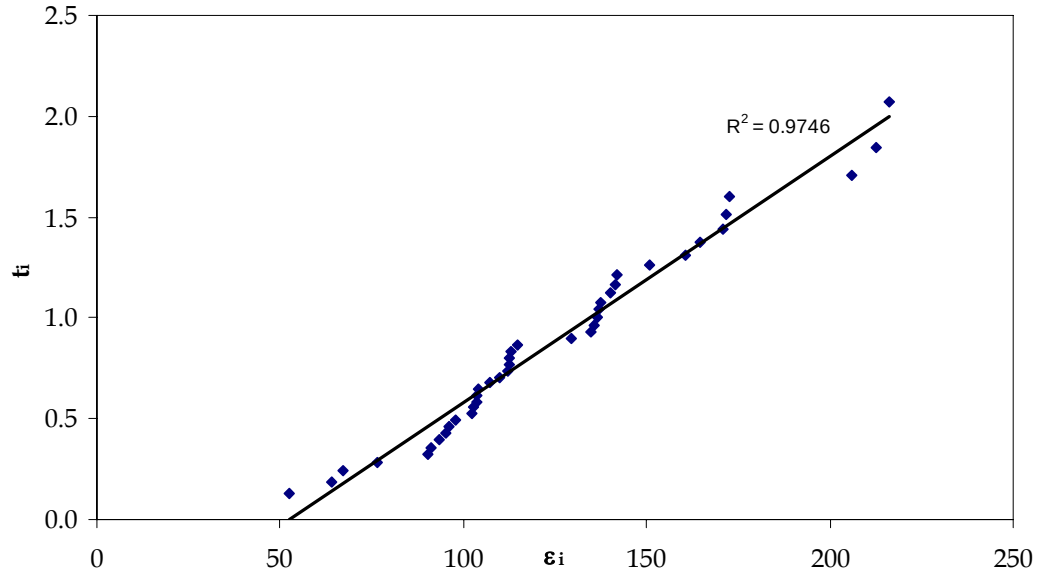

Figure A.13 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1536 (p=1.8)

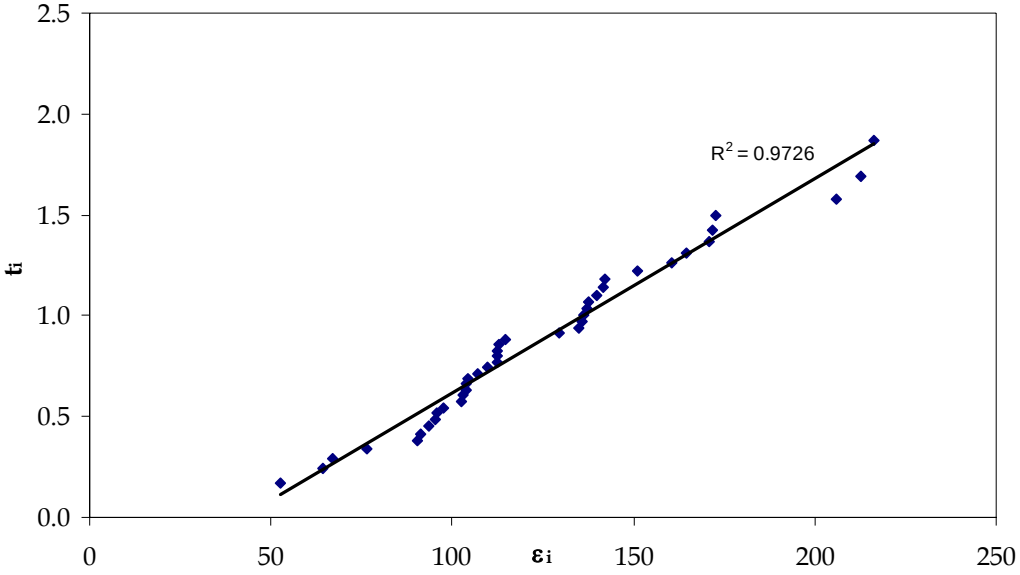

Figure A.14 Q-Q Plot of Residuals from AR(1) Model with Weibull Innovation for EIE 1536 (p=2.1)

APPENDIX A2

Q-Q PLOTS FOR AR(1) MODEL WITH GENERALIZED LOGISTIC INNOVATION

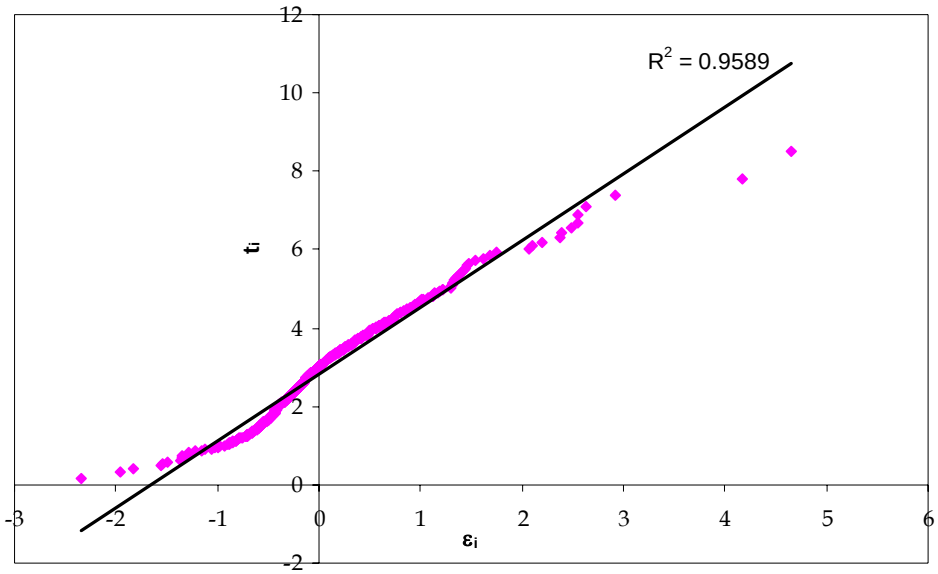

Figure A.15 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1501 (b=10)

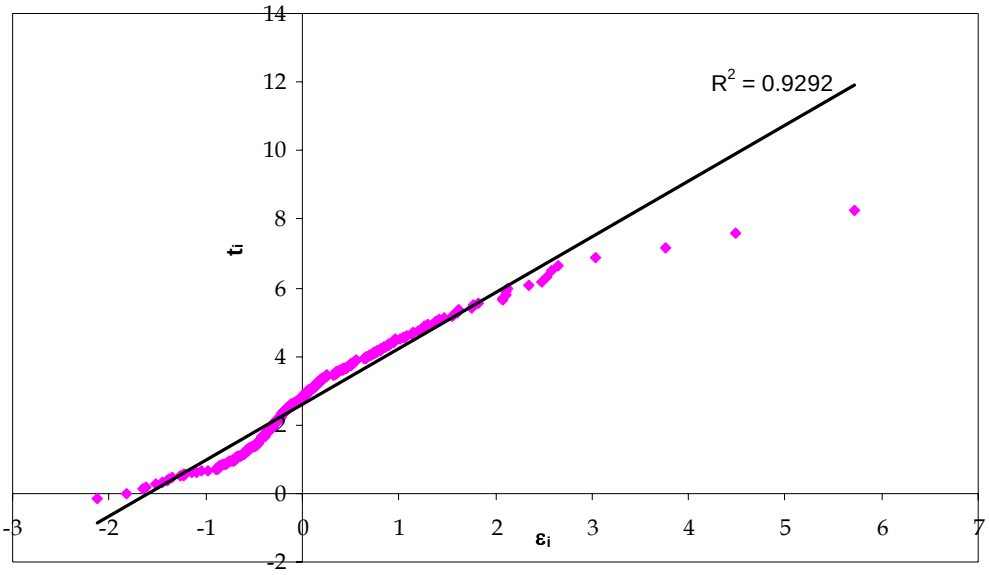

Figure A.16 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1503 (b=8)

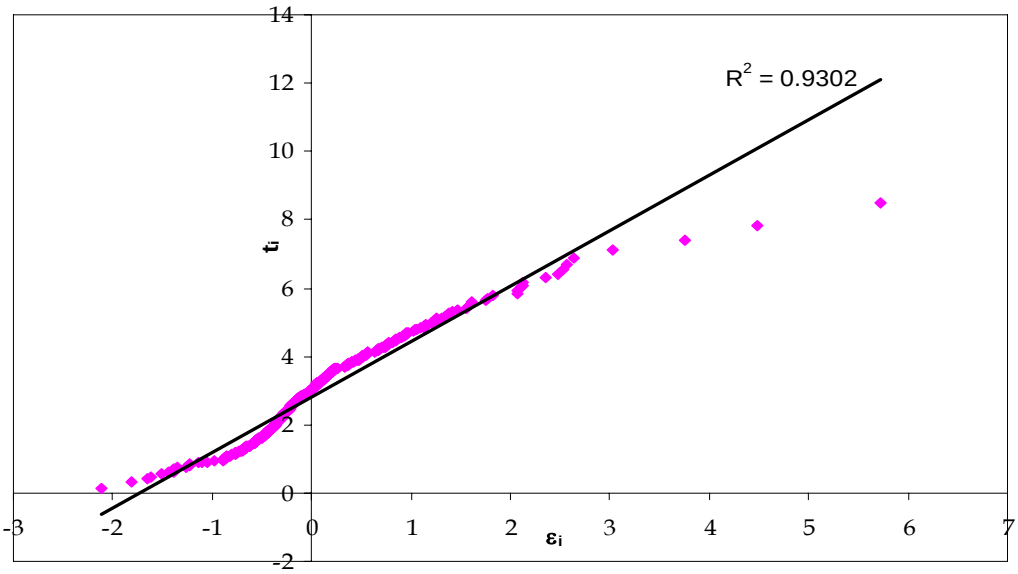

Figure A.17 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1503 (b=10)

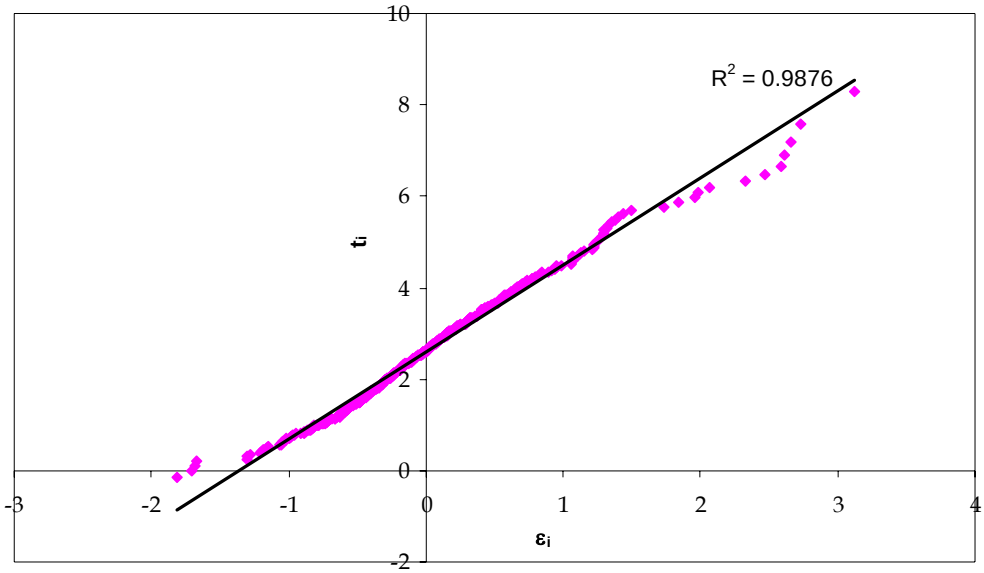

Figure A.18 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1541 (b=8)

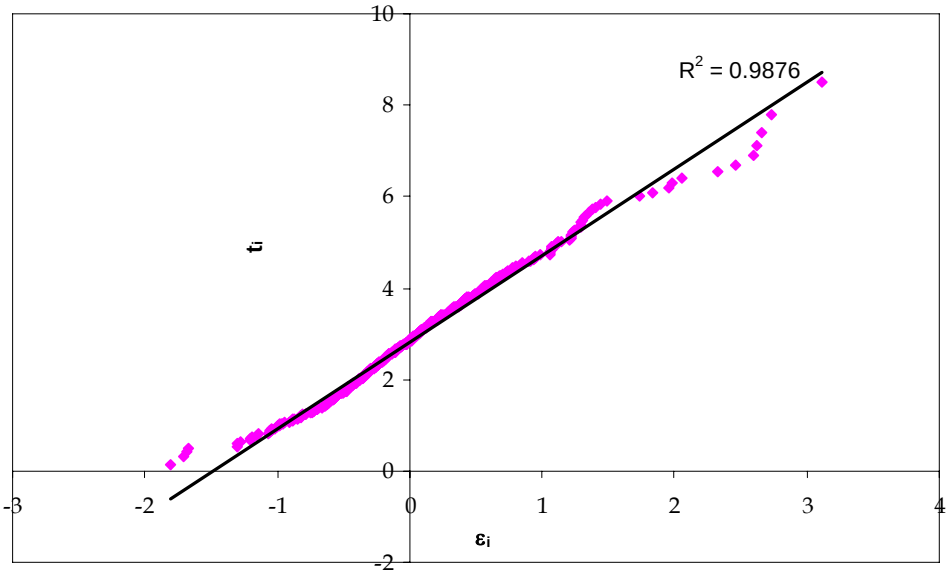

Figure A.19 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1541 (b=10)

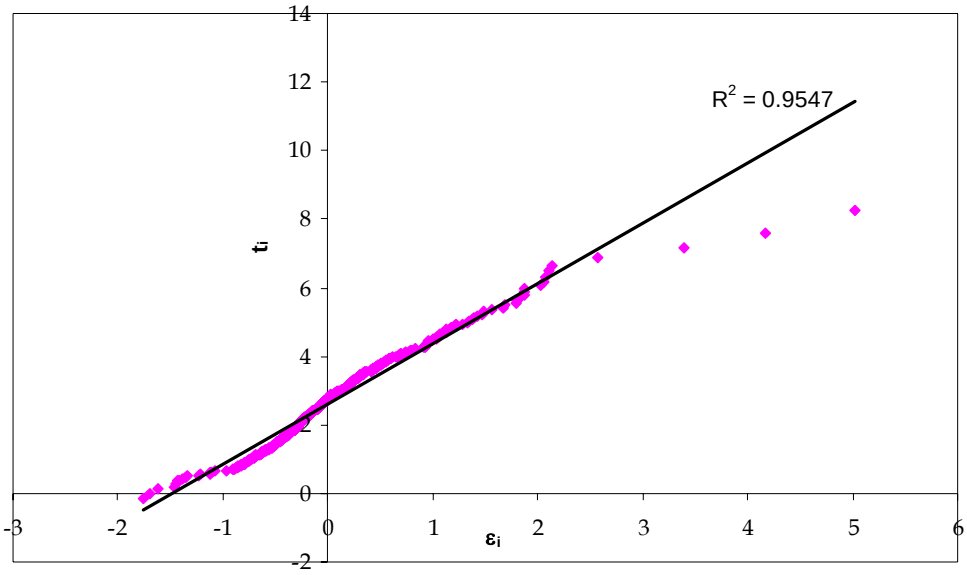

Figure A.20 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1528 (b=8)

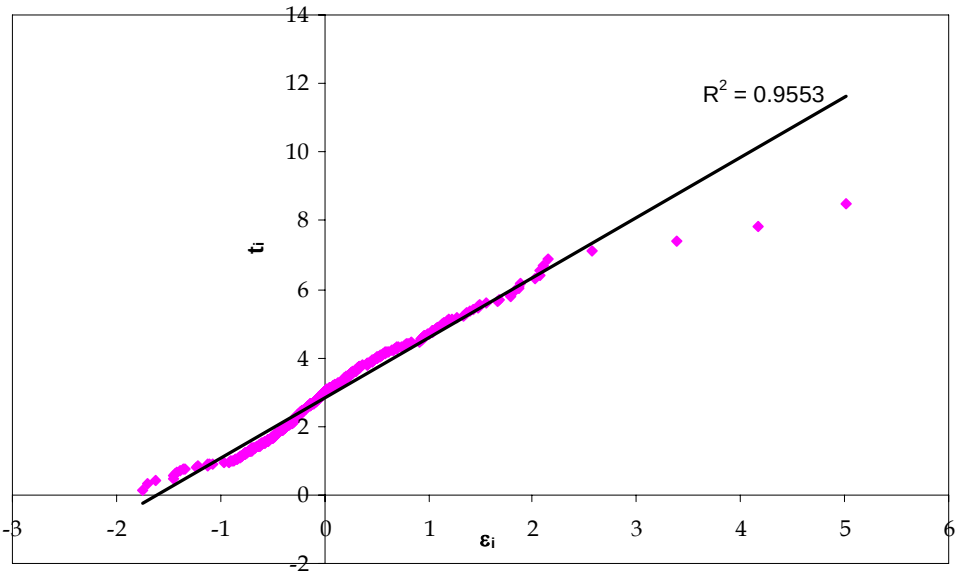

Figure A.21 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1528 (b=10)

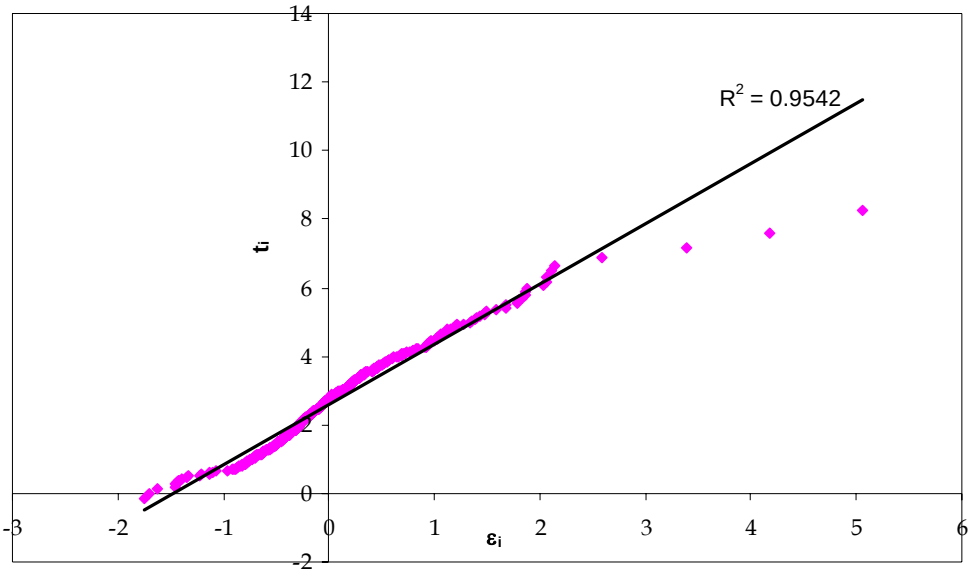

Figure A.22 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1536 (b=8)

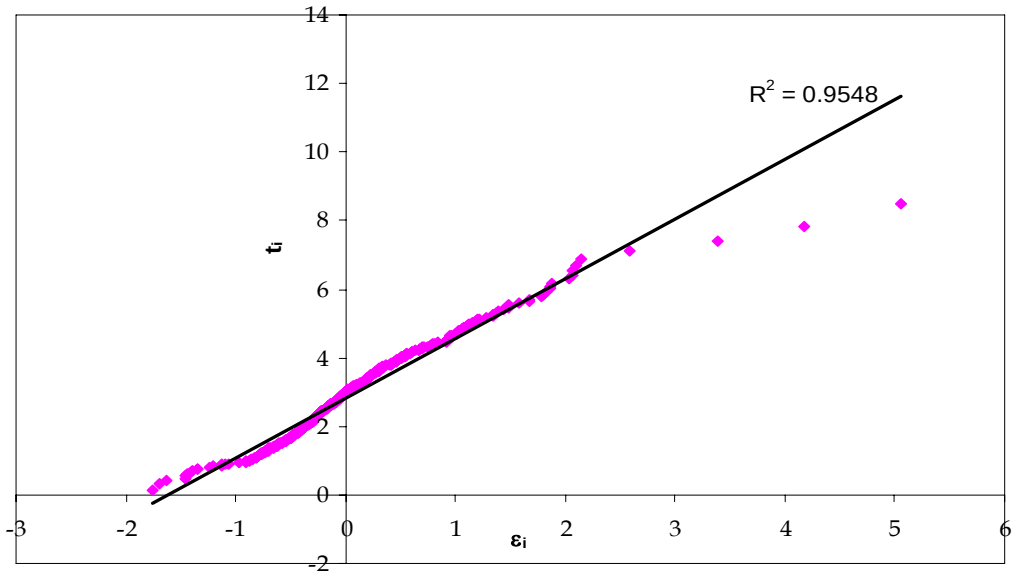

Figure A.23 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for EIE 1536 (b=10)

APPENDIX A3

Q-Q PLOTS FOR AR(1) MODEL WITH GENERALIZED LOGISTIC INNOVATION WITHOUT OUTLIER

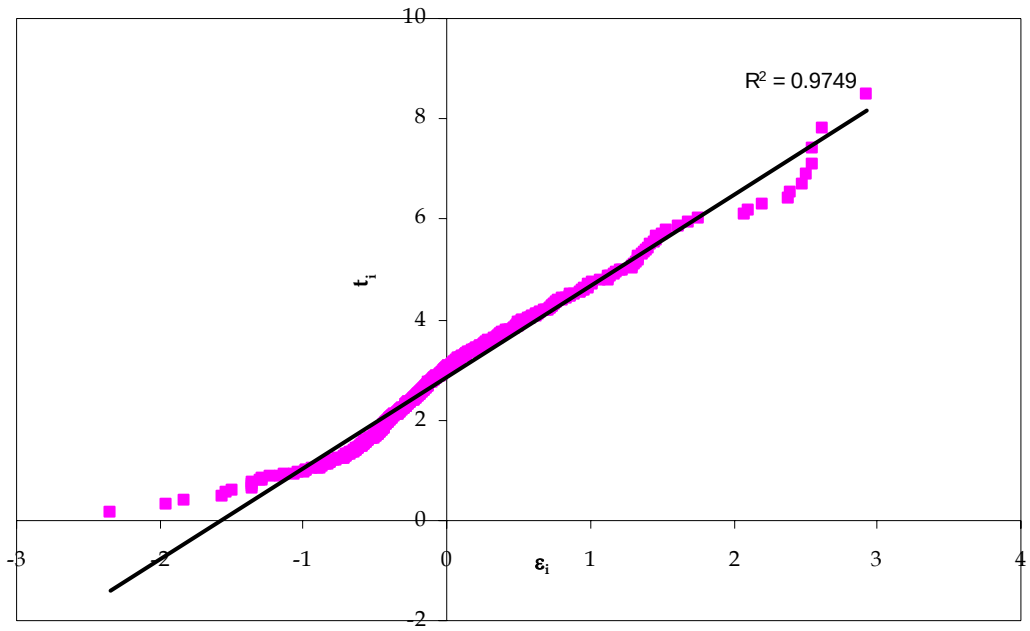

Figure A.24 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation without outlier for EIE 1501 (b=10)

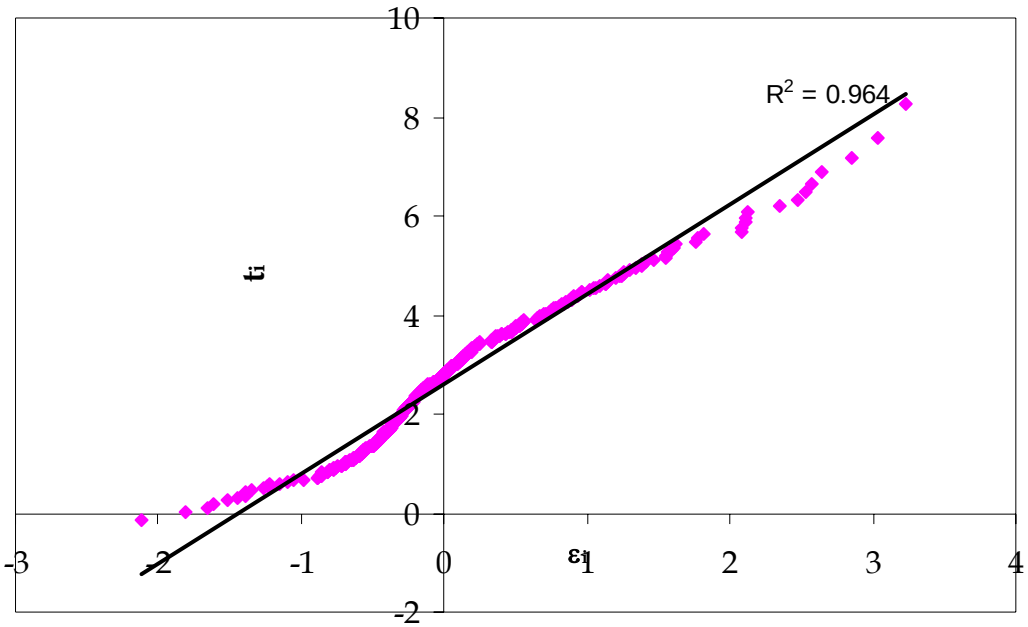

Figure A.25 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation without outlier for EIE 1503 (b=8)

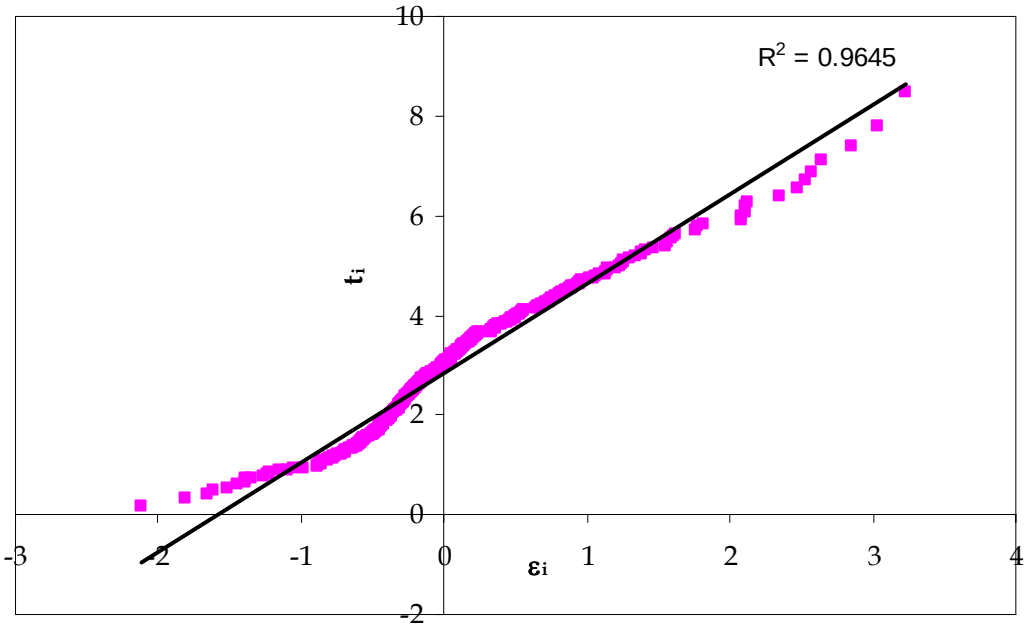

Figure A.26 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation without outlier for EIE 1503 (b=10)

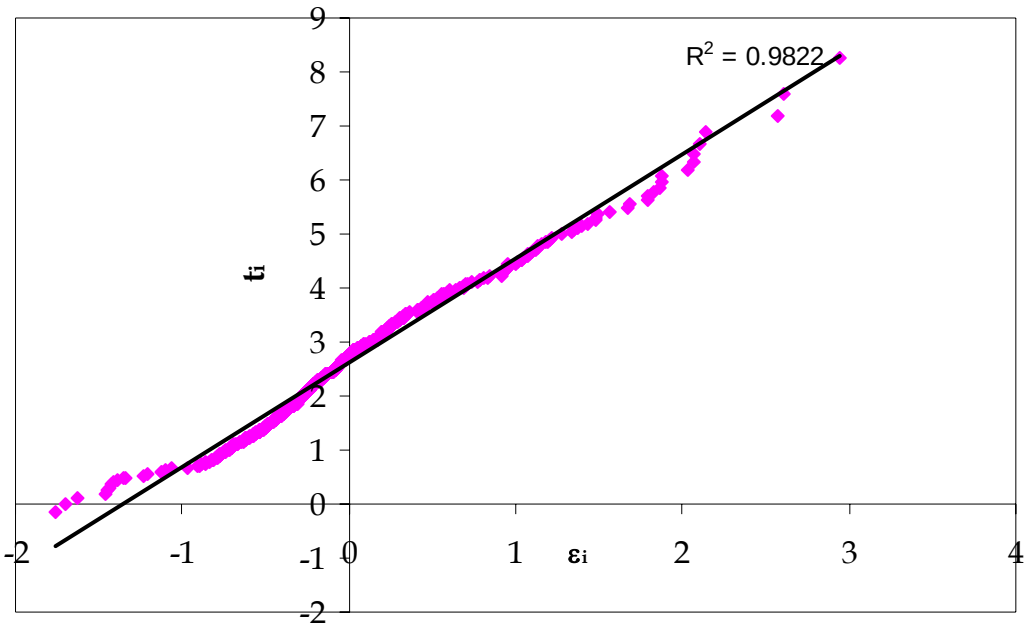

Figure A.27 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation without outlier for EIE 1528 (b=8)

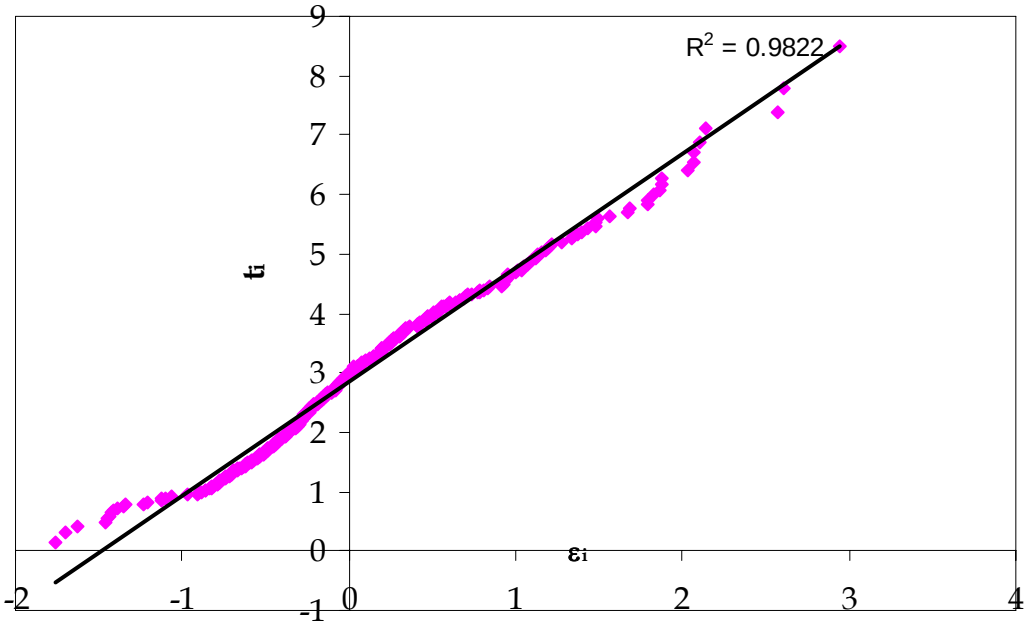

Figure A.28 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation for without outlier EIE 1528 (b=10)

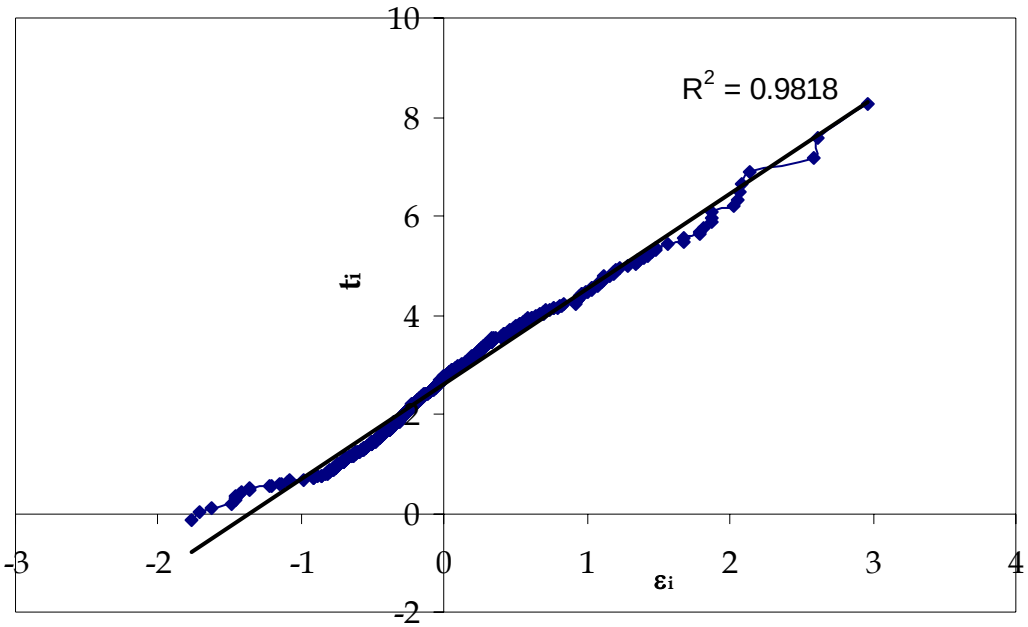

Figure A.29 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation without outlier for EIE 1536 (b=8)

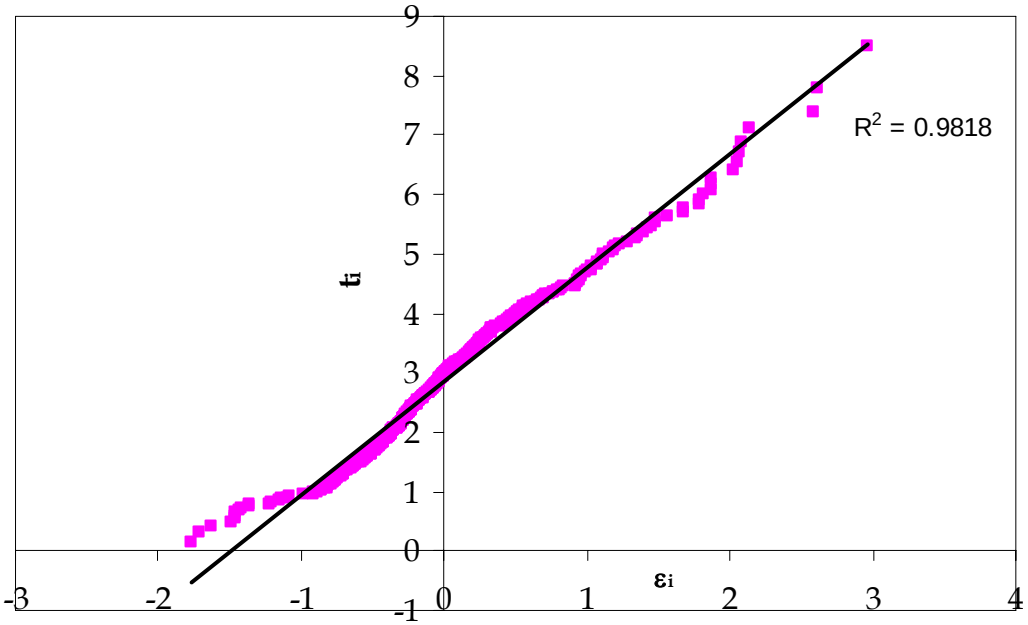

Figure A.30 Q-Q Plot of Residuals from AR(1) Model with Generalized Logistic Innovation without outlier for EIE 1536 (b=10)

APPENDIX B

COMPUTER PROGRAM FOR THE PARAMETER ESTIMATION OF AR(1) MODEL WITH WEIBULL INNOVATION FROM MML PROCEDURE

```
USE numerical_libraries
INTEGER  I,NOUT,J,N,M
REAL  P, VALUE1, VALUE2, MBET, DEL, YORT, Y1ORT, AB2, AB3, FI, SIG, K, D, BF, C
REAL  Y(1000), S(1000), T(1000), BET1(1000), ALF2(1000), BET2(1000), ALF(1000), BET(1000), W(1000), Y11(1000), ZEN(1000), SIG2(1000), FI2(1000), MU2(1000), Z(1000), AB4(1000), AP1(1000), LIK(1000),VG(1000), VGORT

OPEN (1,FILE='ANNUALDATA.txt')
OPEN (2,FILE='FISIGMU.txt')
OPEN (3,FILE='LSFI.txt')
OPEN (7,FILE='LIK.txt')
N=number of data
M=(number of data -1)
```

```
P=shape parameter
VALUE1 = GAMMA(1.+(2./P))
VALUE2 = GAMMA(1.+(1./P))
DO 1 I=1,N
READ (1,1000,END=33) Y(I)
1 CONTINUE
33 CLOSE(1)

!calculation of alpha and beta
SUM74=0; SUM75=0
DO 2 I=2,N
S(I)=(I-1.)/N
T(I)=(-ALOG(1.- S(I)))**(1.0/P)
ALF1(I)=2./T(I)
BET1(I)= 1./(T(I)**2.)
ALF2(I)=(2.-P)*(T(I)**(P-1.))
BET2(I)=(P-1.)*(T(I)**(P-2.))
ALF(I)=(P-1.)*ALF1(I)-P*ALF2(I)
BET(I)=(P-1.)*BET1(I)+P*BET2(I)
SUM74=SUM74+ALF(I)
SUM75=SUM75+BET(I)
2 CONTINUE
DEL=SUM74
MBET=SUM75

! calculation of LS estimators
AB1=0;AB2=0;AB3=0;AB4=0
DO 3 I=2,N
AB1=AB1+Y(I)*Y(I-1)
AB2=AB2+Y(I)
```

```
AB3=AB3+(Y(I-1))**2
AB4=AB4+Y(I-1)
3 CONTINUE
FI= (M*AB1-AB2*AB4)/(M*AB3-AB4**2)
TG=0.
DO 4 I=2,N
VG(I)=Y(I)-FI*Y(I-1)
TG=TG+VG(I)
4 CONTINUE
VGORT=TG/M
TS=0.
DO 5 I=2,N
TS=TS+(VG(I)-VGORT)**2
5 CONTINUE
SIG=SQRT(TS/((M-2.)*(VALUE1-(VALUE2**2))))
WRITE(3,*) 'FI:',FI, 'SIG:',SIG
```

!calculation of concomitant

```
DO 6 L=1,5
        DO 7 I=2,N
        W(I)= (Y(I)-FI*Y(I-1))
        7 CONTINUE
        DO 8 I=2,N
        Y11(I)=Y(I)
        ZEN(I)= Y(I-1)
        8 CONTINUE
        DO 9 I=2,N
        DO 10 J=I+1,N
        IF (W(I).GT.W(J)) THEN
```

```
Q= W(I);W(I)=W(J);W(J)=Q
Q=Y11(I);Y11(I)=Y11(J);Y11(J)=Q
Q=ZEN(I);ZEN(I)=ZEN(J); ZEN(J)=Q
ENDIF
10 CONTINUE
9 CONTINUE
```

!calculation of MU, FI and SIGMA

```
AS1=0;AS2=0;AS3=0;AS76=0;AS77=0;AS78=0;AS79=0
DO 11 I=2,N
AS1= AS1+ (BET(I)*(Y11(I)-FI*ZEN(I)))
AS76=AS76+(BET(I)*Y11(I)*ZEN(I))
AS77=AS77+(BET(I)*Y11(I))
AS78=AS78+(BET(I)*ZEN(I))
AS2= AS2+ (BET(I)*(ZEN(I)**2))
AS3= AS3+((((DEL*BET(I))/MBET)-ALF(I))*ZEN(I))
11 CONTINUE
K= (AS76-((1/MBET)*AS77*AS78))/(AS2-((1/MBET)*(AS78**2)))
D= AS3/(AS2-(1/MBET)*(AS78**2))
AS4=0;AS5=0
DO 12 I=2,N
AS4=AS4+ALF(I)*((Y11(I)-FI*ZEN(I))-((1/MBET)*AS1))
AS5=AS5+ BET(I)*((Y11(I)-FI*ZEN(I))-((1/MBET)*AS1))**2
12 CONTINUE
BF=AS4
C= AS5
```

! calculation of MML estimators

```
SIG2(L)= (-BF+SQRT(BF**2+4*M*C))/(2*SQRT((M)*(M-1)))
```

```
FI2(L)= K-(D*SIG2(L))
MU2(L)=((1/MBET)*(AS1-SIG2(L)*DEL))
FI=FI2(L)
SIG=SIG2(L)
MU=MU2(L)
6 CONTINUE
WRITE(2,*) 'FI:',FI, 'SIG:',SIG, 'MU:',MU
```

!calculation of likelihood function

```
AB4(P)=0;AP1(P)=0
DO 13 P=1,10
DO 14 I=2,N
Z(I)= (Y11(I)-FI*ZEN(I))/SIG
AB4(P)=AB4(P)+(Z(I)**P)
AP1(P)=AP1(P)+ALOG(Z(I))
14 CONTINUE
13 CONTINUE

DO 15 P=1,10
LIK(P)=-M*ALOG(SIG)+((P-1)*AP1(P))-AB4(P)
WRITE (7,*) LIK(P)
15 CONTINUE

1000 FORMAT(F8.8)
2000 FORMAT (I8)
STOP
END
```

APPENDIX C

COMPUTER PROGRAM FOR GENERATION OF AR(1) MODEL WITH WEIBULL INNOVATION

```
USE numerical_libraries
INTEGER    I, NOUT,J,SR,NE, L
REAL      YS1, F, ALF, MU, BET1, RNUNF, B, ORT, VAR, SKEW
REAL      EPS(1000,1000),    Y(1000,1000),    ORTX(1000),    VARX(1000),
SKEWX(1000)

OPEN (1,FILE='EPS.txt')
OPEN (2,FILE='moment.txt')
SR=  NUMBER OF DATA

! generation of Y values
NE=1000
        DO 1 J=1,NE
        YS1=first value of the data
        DO 2 K=1,SR
        CALL UMACH (2, NOUT)
DO 3 L=1,1000
```

```
          F= MML ESTIMATOR OF FI
          ALF= MML ESTIMATOR OF ALF
          MU= MML ESTIMATOR OF LA
          B=SHAPE PARAMETER
          EPS(J,K)=ALF*(-ALOG(1.- RNUNF()))**(1.0/B)
3 CONTINUE
Y(J,K)=F*YS1+EPS(J,K)+MU
YS1=Y(J,K)
WRITE(1,*) Y(J,K)
2 CONTINUE

! calculation of moment value for generated series
SUM1=0
DO 4 K=1,SR
SUM1=SUM1+Y(J,K)
4 CONTINUE
ORTX(J)= SUM1/SR

TOP1=0
DO 5 K=1,SR
TOP1 = TOP1+(Y(J,K)-ORTX(J))**2
5 CONTINUE
VARX(J)= TOP1/(SR-1)

CAR=0
DO 6 K=1,SR
CAR= CAR+ (Y(J,K)-ORTX(J))**3
6 CONTINUE
SKEWX(J) = (CAR*SR)/((SR-1)*(SR-2)*VARX(J)**1.5)
```

```
1 CONTINUE
SUM2=0;SUM3=0; SUM5=0
DO 7 J=1,NE
SUM2=SUM2+ORTX(J)
SUM3=SUM3+VARX(J)
SUM5=SUM5+SKEWX(J)
7 CONTINUE
ORT=SUM2/NE
VAR=SUM3/NE
SKEW=SUM5/NE

WRITE(2,*) 'ORT:',ORT, 'VAR:',VAR,'SKEW:',SKEW

1000 FORMAT(F5.2)
2000 FORMAT (I2)
STOP
END
```

APPENDIX D

COMPUTER PROGRAM FOR THE PARAMETER ESTIMATION OF AR(1) MODEL WITH GENERALIZED LOGISTIC INNOVATION FROM MML PROCEDURE

```
USE numerical_libraries
INTEGER    I, NOUT, J, N, M
REAL       MUL, B, MBET, DEL, FI, VGORT, PSISB, PSIBIRS, PSIB, PSIBIR,
SIG, NU, VORT, K, D, BF, C
REAL       Y(1000), P(1000), T(1000), ALF(1000), BET(1000), VG(1000),
W(1000), Y11(1000), ZEN(1000), V(1000), SIG2(1000), FI2(1000), NU2(1000),
Z(1000), LIK(1000)

OPEN (1,FILE='monthly.txt')
OPEN (2,FILE='FISIGHATA.txt')
OPEN (4,FILE='LSFI.txt')
OPEN (6,FILE='LIK.txt')
N=NUMBER OF DATA
M= (NUMBER OF DATA-1)
MUL=2.*SQRT(M*(M-1.))
```

```
B=SHAPE PARAMETER
DO 1 I=1,N
READ (1,1000,END=33) Y(I)
1 CONTINUE
33 CLOSE(1)
```

!calculation of alpha and beta

```
AT2=0;AT3=0
DO 2 I=1,M
P(I)=I/(M+1.)
T(I)=-ALOG(P(I)**(-1./B)-1.)
ALF(I)=(1+EXP(T(I))+T(I)*EXP(T(I)))/(1+EXP(T(I)))**2
BET(I)= EXP(T(I))/(1+EXP(T(I)))**2
AT2=AT2+BET(I)
AT3=AT3+(ALF(I)-(1./(B+1.)))
2 CONTINUE
MBET=AT2
DEL=AT3
```

! calculation of LS estimators

```
AB1=0;AB2=0;AB3=0;AB4=0
DO 3 I=2,N
AB1=AB1+Y(I)*Y(I-1)
AB2=AB2+Y(I)
AB3=AB3+(Y(I-1))**2
AB4=AB4+Y(I-1)
3 CONTINUE
FI= (M*AB1-AB2*AB4)/(M*AB3-AB4**2)
TG=0.
```

```
DO 4 I=2,N
VG(I)=Y(I)-FI*Y(I-1)
TG=TG+VG(I)
4 CONTINUE
VGORT=TG/M
TS=0.
DO 5 I=2,N
TS=TS+(VG(I)-VGORT)**2
5 CONTINUE
PSISB=  (1./B)+(1./(2.*(B**2.)))+(1./(6.*(B**3.)))-
(1./(30.*(B**5.)))+(1./(42.*(B**7.)))-(1./(30.*(B**9.)))
PSIBIRS= (1./1.)+(1./(2.*(1.**2.)))+(1./(6.*(1.**3.)))-
(1./(30.*(1.**5.)))+(1./(42.*(1.**7.)))-(1./(30.*(1.**9.)))
PSIB= ALOG(B)-(1./(2.*B))-(1./(12.*(B**2)))+(1./(120.*(B**4)))-
(1./(252.*(B**6)))
PSIBIR=ALOG(1.)-(1./(2.*1.))-(1./(12.*(1.**2)))+(1./(120.*(1.**4)))-
(1./(252.*(1.**6)))
SIG=SQRT(TS/((M-2.)*(PSIBS+PSIBIRS)))
NU=VGORT-(SIG*(PSIB-PSIBIR))
WRITE(4,*) 'FI:',FI,'SIG:',SIG,'NU:',NU
```

!calculation of concomitant

```
DO 6 L=1,5
        DO 7 I=2,N
        W(I)= ((Y(I)-FI*Y(I-1))-NU)
        7 CONTINUE
        DO 8 I=2,N
        Y11(I)=Y(I)
        ZEN(I)= Y(I-1)
```

```
8 CONTINUE
DO 9 I=2,N
DO 10 J=I+1,N
IF (W(I).GT.W(J)) THEN
Q= W(I);W(I)=W(J);W(J)=Q
Q=Y11(I);Y11(I)=Y11(J);Y11(J)=Q
Q=ZEN(I);ZEN(I)=ZEN(J); ZEN(J)=Q
ENDIF
10 CONTINUE
9 CONTINUE
```

! calculation of MML estimators

```
AT1=0;AS1=0;AS2=0;AS3=0;AS4=0;AS5=0
DO 11 I=2,N
V(I)=Y11(I)-(FI*ZEN(I))
AT1=AT1+BET(I-1)*V(I)
11 CONTINUE
VORT=AT1/MBET
DO 12 I=2,N
AS1= AS1+ (BET(I-1)*Y11(I)*ZEN(I))
AS2= AS2+ (BET(I-1)*ZEN(I))
AS3= AS3+ (BET(I-1)*Y11(I))
AS4= AS4+ (BET(I-1)*ZEN(I)**2)
AS5= AS5+((ALF(I-1)-(1./(1.+ B)))-(DEL/MBET)*BET(I-1))*ZEN(I)
12 CONTINUE
K= (AS1-((1./MBET)*AS2*AS3))/(AS4-((1./MBET)*(AS2**2)))
D= AS5/(AS4-((1./MBET)*(AS2**2)))
AS6=0;AS7=0
DO 13 I=2,N
```

```
AS6=AS6+(ALF(I-1)-(1./(1.+ B)))*(V(I)-VORT)
AS7=AS7+ BET(I-1)*(V(I)-VORT)**2
13 CONTINUE
BF=(B+1.)*AS6
C=(B+1.)*AS7
SIG2(L)= (-BF+SQRT(BF**2+4*M*C))/(MUL)
FI2(L)= K-(D*SIG2(L))
NU2(L)=VORT-(SIG2(L)*(DEL/MBET))
FI=FI2(L)
SIG=SIG2(L)
NU=NU2(L)
6 CONTINUE
```

!calculation of likelihood function

```
AB4=0;AP1=0
DO 14 I=2,N
Z(I)= (Y(I)-FI*Y(I-1)-NU)/SIG
AB4=AB4+Z(I)
AP1=AP1+ALOG(1+EXP(-Z(I)))
14 CONTINUE
DO 15 C=1,20
LIK(C)=M*ALOG(C)-M*ALOG(SIG)-AB4-((C+1)*AP1)
WRITE (6,*) LIK(C)
15 CONTINUE
WRITE(2,*) 'FI:',FI,'SIG:',SIG,'NU:',NU
1000 FORMAT(F8.8)
2000 FORMAT (I8)
STOP
END
```

APPENDIX E

COMPUTER PROGRAM FOR GENERATION OF AR(1) MODEL WITH GENERALIZED LOGISTIC INNOVATION

```
USE numerical_libraries
INTEGER    I,  J, SR, NE, L
REAL     YS1, F, ALF, LA, BET1, BET2, RNUNF, ORT, VAR, SKEW
REAL     XM(1000), XS(1000), EPS(1000,1000), Y(1000,1000), SY(1000,1000),
ORTX(1000), VARX(1000), SKEWX(1000)
OPEN (2,FILE='Y.txt')
OPEN (3,FILE='mean.txt')
OPEN (4,FILE='stddev.txt')
OPEN (5,FILE='moment.txt')
SR=NUMBER OF DATA
NE=1000
DO 1 I=1,SR
READ (3,1000,END=33) XM(I)
READ (4,1000,END=34) XS(I)
1 CONTINUE
33 CLOSE(3)
34 CLOSE(4)
```

! generation of Y values

```
DO 2 J=1,NE
        YS1= first value of the deaseasonalized monthly data
        DO 3 K=1,SR
       CALL UMACH (2, NOUT)
       DO 4 L=1,NE
        F= MML ESTIMATOR OF FI
       ALF= MML ESTIMATOR OF ALF
        LA= MML ESTIMATOR OF LA
        BET1=SHAPE PARAMETER
        BET2=1./BET1
        EPS(J,K)=-ALF*ALOG((1/(RNUNF()))**BET2)-1)
4 CONTINUE

Y(J,K)=F*YS1+EPS(J,K)+LA
SY(J,K)=Y(J,K)*XS(K)+XM(K)
YS1=Y(J,K)
3 CONTINUE
```

! calculation of moment value for generated series

```
SUM1=0
DO 5 K=1,SR
SUM1=SUM1+SY(J,K)
5 CONTINUE
ORTX(J)= SUM1/SR
TOP1=0
DO 6 K=1,SR
TOP1 = TOP1+(SY(J,K)-ORTX(J))**2
```

```
6 CONTINUE
VARX(J)= TOP1/(SR-1)

CAR=0
DO 7 K=1,SR
CAR= CAR+ (SY(J,K)-ORTX(J))**3
7 CONTINUE
SKEWX(J) = (CAR*SR)/((SR-1)*(SR-2)*VARX(J)**1.5)
2 CONTINUE

SUM2=0;SUM3=0; SUM5=0
DO 8 J=1,NE
SUM2=SUM2+ORTX(J)
SUM3=SUM3+VARX(J)
SUM5=SUM5+SKEWX(J)
8 CONTINUE

ORT=SUM2/NE
VAR=SUM3/NE
SKEW=SUM5/NE

WRITE(5,*) 'ORT:',ORT, 'VAR:',VAR,'SKEW:',SKEW

1000 FORMAT(F5.2)
2000 FORMAT (I2)
STOP
END
```

VITA

The author was born in Konya, on the 11 of April 1974. After the completion her elemantary, secondary, and high school education in Konya, she started her university education in 1991 at the Selçuk University, Konya. She received her B.S. and M.S. degrees in Civil Engineering department from the same university in June 1995 and September 1999, respectively. She worked as a research assistant in the same university from 1996 to 1999. After that she joined the Ph.D studies at Middle East Technical University (METU).

She has been working in METU as a research assistant since 1999. She is funded by the Turkish Council of Higher Education. After completing her Ph.D studies, she is expected to take part in the teaching staff at the Inonu University, Malatya. Her main areas of interest are statistical hydrology and water reources systems.