COMPARISON OF STOCHASTIC MODELS FOR MONTHLY STREAMFLOW SEQUENCES

A Thesis Presented to the School
of Graduate Studies
Addis Ababa University

In Partial Fulfillment of the
Requirements for the Degree of Master of Science
in Statistics

By

EMMANUEL GABREYOHANNES

June, 1994

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

COMPARISON OF STOCHASIC MODELS FOR MONTHLY STREAMFLOW SEQUENCES

By

Emanuel Gebre Yohannes
Department of Statistics
Faculty of Science

Approved by the Examining Board

Prof. Asmerom Kidane
Chairman, Dept. Graduate Committee

Dr. Abebe Tessera
Advisor

Dr. Igor Litvin
Ext. Examiner

Dr. Eshetu Wencheko
Examiner

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my advisor Dr. Abebe Tessera for suggesting the paper and for his generous advice and encouragement at all stages of the work.

I am so much grateful to the staff of the Ethiopian Valleys Development Studies Authority for their kindness in making their data available to me. Special thanks are due to Ato Geremew Mekonen for providing me various materials and for his helpful comments.

My heart felt appreciation goes to Ato Eyob Woldeyohannes for his unlimited assistance in the course of the work and for printing the manuscript.

Finally, I would like to thank all my friends and staff members of the department of statistics who, one way or another, contributed to the realization of this work.

ABSTRACT

In this research, statistical work in stochastic modelling of hydrologic time series is discussed. The theme is the comparison of stationarity and non-stationarity approaches for stochastic simulation of monthly streamflow sequences. The paper first introduces various concepts related to stochastic modelling; the history of stochastic simulation and applications of stochastic modelling in hydrology; and the step-by-step modelling procedures. Then the paper describes how synthetic streamflow sequences can be generated using these procedures. As a particular case, synthetic streamflow sequences were generated for two streams in Ethiopia using both stationarity and non-stationarity approaches. At last, based on these synthetic flows, an attempt was made to compare the performance of the two approaches with some comments and recommendations.

CONTENTS

CONTENTS - (Continued)

CHAPTER ONE

INTRODUCTION

The first question that naturally arises in regard to stochastic modelling of hydrologic time series (such as annual rainfall, monthly streamflow, etc.) is as to why we should apply statistics in hydrologic analysis. The answer to this question lies in the fact that the planning and designing of water resources projects need information on different hydrologic events that are not governed by the known physical and chemical laws, but are governed by the laws of chance or are stochastic in nature. For example, streamflow in any given river varies from day to day, from month to month and from year to year. The fact is that no one can exactly predict the discharge of a certain river at a particular point in any particular month, say January 1996. Since the exact amount of flow cannot be predicted, the analyst has to rely on the statistical and probabilistic analyses of past observed (historic) streamflow data to determine the risk involved in designing the structure. But, due to different reasons, information on streamflow data at the point of interest for a good length of time is rarely available. Thus, quite often one is confronted with the problem of analyzing and designing water resources projects with the available short historical hydrologic data. But basing analysis on a limited length of historical (recorded) data is found to be inadequate. For instance, suppose we have for analysis a 50-year length of streamflow data and that the overall purpose of the analysis is to determine the size and future performance of a storage reservoir to be constructed near the streamflow measurement site. It has been a common procedure to use the longest run length or largest run sum to determine the probability of occurrence of critical droughts (probability of the reservoir going dry) ; the range to determine the storage capacity of the reservoir; and frequency analysis to estimate the frequency of flood, or in general, various sample statistics are used to determine the probability of the reservoir to meet long term storage requirements. But it is quite likely that the river, and hence the streamflows, might exhibit much larger pseudo-cyclic changes than the 50-year maximum allowed for by the historical record due to variations in the energy radiated by the sun or atmospheric circulations. As a consequence, the analysis based on the 50-year historic data may lead to economic wastage or failure. As a remedy for this, statisticians and hydrologists have embarked upon formulating models which would be able to reproduce, by simulation, longer sequences of flows that are statistically indistinguishable from the relevant historical sequence. The techniques and procedures for finding such models is what we call stochastic modelling.

The stochastic modelling of streamflow sequences is motivated by the desire to produce synthetic sequences of flows to route through a water resource system design; the idea being to test it out under a variety of conditions, and with longer sequences of flows than historically available. The implication is that longer sequences will contain more extreme events than observed and thus give a more convincing test of the system.

There are several models that can be applied for modelling hydrologic time series; some are simple and others are complex. As for complex models, often there arise difficulties in identifying them and in estimating their parameters. Besides this, they involve too many parameters. Therefore, as a general rule, it is wise to start with a simple model and keep it simple. The simplest and most widely used models are the autoregressive (AR) models. These models have been extensively used in hydrology and water resources since the early 1960's for modelling annual and periodic hydrologic time series such as annual rainfall, monthly streamflow, etc. The historical development of AR models in hydrology may be divided into two periods: the 1960 decade initiated mainly by the work of Thomas and Fiering (1962) and Yevjevich (1963), and the 1970 decade motivated by the publication of the book of Box and Jenkins (1970). Since then, more refined methodologies for improving the estimates of the parameters of the model, for verifying or checking the assumptions of the model and for selecting among competing models have been developed.

Though various models are available for stochastic modelling of monthly streamflow sequences, as stated above, the AR models are the simplest and widely used models and hence in this study we will restrict ourselves to these.

Generally two approaches are developed for stochastic modelling of hydrologic time series [Seth & Chandra, 1977]: the stationarity and non-stationarity approaches. The stationarity approach is based upon decomposition of the non-stationary stochastic process into deterministic and stochastic components; quantifying and removing the deterministic part; and then representing the stochastic component by an autoregressive model of order one, two or three. In the second approach, the non-stationary stochastic process is treated as it is, i.e., without decomposing it into its various components.

The overall objective of this study is to compare the performance of these two approaches for stochastic modelling of monthly streamflow sequences for two streams in Ethiopia : Rivers Megecha and Gudla. The inferred best approach then may be used to generate synthetic streamflow sequences for a longer period so as to be used as inputs for some water resources project such as storage reservoir or dam which may be constructed along the two rivers in the future.

In this research, the second chapter describes various concepts related to stochastic models and modelling together with the historic development of stochastic simulation in hydrology; some common applications of stochastic models in hydrology; the various properties and description of autoregressive (AR) models; and methods of parameter estimation,model identification and selection among competing models. Chapter three covers the step-by-step modelling procedures for monthly streamflow sequences for Rivers Megecha and Gudla using both the stationarity and non-stationarity approaches. Chapter four deals with the generation of synthetic samples and comparison of the two approaches based on these samples. And the last chapter presents some recommendations and limitations and other comments extending or relating to presentations in the previous chapters.

CHAPTER TWO

STOCHASTIC MODELS AND MODELLING IN HYDROLOGY

2.1 What is stochastic modelling ?

Let T denote the set of time points at which measurements are made and let X(t) be the observation made at time t. The set of observations $\{X(t), t \varepsilon T\}$ is called a *time series*. A time series may also be defined as a collection of observations made sequentially in time. The basic idea of the statistical theory of analysis of a time series $\{X(t), T \varepsilon t\}$ is to regard the time series as an observation made on a family of random variables $\{X(t), t \varepsilon T\}$; that is, for each t in T, the observation X(t) is an observed value of a random variable. Such a family of random variables $\{X(t), t \varepsilon T\}$ is called a *stochastic process* [Parzen, 1965]. When $T = \{ 0, \pm 1, \pm 2, ... \}$, the stochastic process is said to be discrete and when $T = \{ t: -\infty < t < \infty \}$, it is said to be continuous. An important class of stochastic process are those which are stationary. A stochastic process is said to be *stationary* if the joint distribution of $X(t_1),...,X(t_n)$ is the same as the joint distribution of $X(t_1 + \tau),..., X(t_n + \tau)$ for all $t_1,...,t_n$, τ. A stochastic process is stationary in the mean or *first-order stationary* if

$$E\{X(1)\} = E\{X(2)\} = ... = E\{X(t)\} = E(X) = \mu$$

The process is stationary in the covariance when the covariance depends only on the time lag k and not on the position t. That is, $Cov\{ X(t) , X(t-k) \} = Cov(k)$ regardless of t. A stochastic process is said to be *second order stationary* or *weak stationary* if it is stationary in the mean and covariance. If the other statistical properties besides the mean and covariance do not depend on time, then the stochastic process is *strong stationary*. But if shifting the time origin by an amount say τ affects the joint distributions, then the process is *non-stationary*.

A mathematical model representing a stochastic process is called a *stochastic model*. In order to analyze a time series $\{X(t), t \varepsilon T\}$, we would like to find a mathematical model for $\{X(t), t \varepsilon T\}$, which is completely specified except for the values of certain parameters which one proceeds to estimate on the basis of an observed sample. The techniques and procedures for finding such a model is called *stochastic modelling*.

Stochastic modelling is a process which can be simple or complex depending on the characteristics of the available sample series, on the type of model to use and on the selected techniques of modelling. For instance, series with statistical characteristics that do not vary with time usually lead to models and modelling techniques which are simpler than those of series with time varying characteristics. There are various types of stochastic models which can be used to represent a time series. Some are more complex than others. For a particular type of model, there are various techniques for estimating the parameters of the model and for testing how good the model is. Also in this case, some are more complex than others.

In general, stochastic modelling can be organized in the following stages [Box & Jenkins, 1970]:

1) the selection of the type of model ,
2) the identification of the form of the model ,
3) the estimation of the model parameters , and
4) the diagnostic check of the model.

The first stage refers to selecting a type of model among the various types of models available to the analyst. Once the type of model is selected, the next stage is to identify the form or order of the model. For instance, if autoregressive(AR) models were selected in the first stage, then we need to identify the order of the AR model, say order one (one AR coefficient or one parameter), order two, etc. The third stage is to estimate the parameters of the model identified in stage two and some checks are made on the conditions to be met by the estimated parameters. The final stage of the modelling is to make some diagnostic checks to verify how good the model is.

Models are built to reproduce or to resemble the main statistical characteristics of the historical time series. Such reproduction or resemblance is understood to be in the statistical sense. It doesn't mean that a generated series based on the model has to give exactly the same statistical characteristics as shown by the historical record. This brings up the questions of what statistical characteristics are to be reproduced by the model and how these characteristics should be interpreted or understood.

Unfortunately, there are no unique and easy answers to the above questions. First of all, the true or population statistical characteristics of hydrologic series is never known, because what is observed or measured is only a finite (sample) number of years(N), and as a result the characteristics derived from such samples are only estimates of the true (unknown) characteristics. If, for example, instead of N years of observations, a different number of years N' either smaller or larger than N were observed, then those estimates based on a sample of N' years would be different from those based on N years. That is, the sample estimates are random variables.

The main statistical characteristics of most time series are the mean, standard deviation, skewness and autocorrelation. Usually the mean and standard deviation can be reproduced by a given model with a high degree of certainty. But, this is not so for the skewness and autocorrelation. In particular, it is difficult to reproduce them in small samples. Their interpretation often decides the type of the model to be used as well as its form [Salas *et al*, 1980].

2.2 History of synthetic generation of streamflows: beginnings of stochastic simulation.

The inadequacy of basing analysis on a limited length of historical data has been appreciated for many decades. Hazen(1914) followed the work of Rippl(1883) by studying the variability of flows in rivers and its effect on the requisite storage capacities of reservoirs. He made a joint study of the streamflow records of 25 - 45 years in length of 14 rivers around New York by adding up these streamflow records so as to get a longer flow sequence. He used Gaussian probability paper which he originated and mass diagrams, after initial standardization with respect to sample means. A drawback in this particular approach was that the effective total length of historical record was much less than the sum of the lengths of the records of the individual rivers; clearly because of geographical proximity and meteorological similarities. Thus the flow sequences did not give a comprehensive crossection of possible events and combination of events from which a sound hydrological assessment could be made.

These shortcomings were recognized by Sudler(1927) who extended the work of Hazen by the inclusion of an element of chance. Discrete values of annual streamflow from a 50-year sequence were transferred one to a card of a pack of fifty. The pack was shuffled to form new sequences of data. This approach, though it was an advancement, had the serious limitation that the flows were collectively the same from one synthetic sequence to the next; thus, for instance, moments and extreme values were always exactly the same - obviously unrealistic of future sequences. There was also the fact that the method ignored possible dependency in the streamflow sequences.

Perhaps the first real stochastic simulation was by Barnes(1954); in designing a reservoir for the Upper Yarra river in Australia, he used a table of mutually independent standardized Gaussian variables to generate a 1,000 year sequences of synthetic annual data. The mean and standard deviation were estimated from the 29-year historical record. This was a reasonable approach since the assumption of independence and Gaussian distribution of flows were not in conflict with the record. The idea of reproducing particular historical aspects of the data in a simulation model is central to much present-day generation of synthetic flows.

2.3 Stochastic models in hydrology and their applications

2.3.1 Stochastic models in hydrology

Though early studies by Hazen and Sudler (and then Barnes) showed the feasibility of using statistics and probability theory for hydrologic simulation, it was not until the beginning of the 1960's that the formal development of stochastic modelling began with the introduction and application of autoregressive(Markov) models to seasonal and annual hydrologic time series. Since then a great deal of work has been done and published. Research on hydrologic time series has been aimed at studying the main statistical characteristics, providing physical justification to some stochastic models; developing new and/or alternative models; improving the estimates of model parameters; developing new or improving existing modelling procedures; and improving tests of goodness of fit.

Several stochastic models have been proposed in the past for modelling hydrologic time series. These include : autoregressive models(AR); fractional gaussian noise models(FGN); autoregressive moving average models(ARMA); broken-line models(BL); model of intermittent processes; disaggregation models; Markov mixture models; ARMA-Markov models; and general mixture models. Although each model has its own merit and some of them can be successfully applied in operational hydrology, they do have limitations: they may not be able to reproduce long (or short) term dependence; there may be difficulty in estimating parameters; they may involve too many parameters and others.

Although the development of time series modelling in hydrology has reached some degree of sophistication, unfortunately most time series modelling in practice is still generally based on the simple methods. This usually involves the selection of an AR model in advance mainly because:

1) the AR form has an intuitive type of time dependence; i.e., the value of a variable at the present time depends on the values at previous times ,and
2) they are the simplest models to use.

a) Description of AR models

Let $y_{v,\tau}$ be the observation for the year v and time τ (say months, weeks, etc) of a hydrologic time series with mean μ_τ and variance σ^2_τ , where τ=1,...,w and w is the total number of time intervals within the year; v=1,...,N and N is the total number of years of observations.The general pth order AR linear model is [Box & Jenkins, 1970]:

$$y_{v,\tau}=\mu_\tau+\sigma_\tau z_{v,\tau} \quad \text{and}$$

$$z_{v,\tau}=\sum_{j=1}^{p} \phi_j z_{v,\tau-j}+\sigma_\epsilon \xi_{v,\tau}$$

where ϕ_j are the constant (non-periodic) AR coefficients and σ_ϵ is a non-periodic standard deviation which gives $\xi_{v,\tau}$ as a second order stationary and independent variable with mean zero and variance one.

The variable $z_{v,\tau}$ may also be represented by an AR model with periodic coefficients as [Salas *et al*, 1980]:

$$z_{v,\tau}=\sum_{j=1}^{p} \phi_{j,\tau} z_{v,\tau-j}+\sigma_{\epsilon\tau}\xi_{v,\tau}$$

where $\phi_{j,\tau}$ is the jth periodic AR coefficient at time τ, $\sigma_{\epsilon\tau}$ is also a periodic standard deviation and $\xi_{v,\tau}$ is as defined above. The above model is non-stationary because both the autoregression coefficients and the standard deviation are periodic. This fact makes the model more complicated. However, if the size of the historical record is small, AR models with constant coefficients yield equally good results.

b) Properties of AR models

$$E(y_{v,\tau})=\mu_\tau$$

$$E(z_{v,\tau})=E(\xi_{v,\tau})=0$$

$$VAR(y_{v,\tau})=\sigma_\tau^2$$

and

$$VAR(z_{v,\tau})=VAR(\xi_{v,\tau})=1$$

for τ = 1,...,w. μ_τ, σ_τ^2 will be estimated from historical data using appropriate methods (see section 2.4).

The parameters $\phi_1, \phi_2, \ldots, \phi_p$ are estimated by solving

$$\rho_k = \sum_{j=1}^{p} \phi_j \rho_{k-j} \quad k>0$$

where the population correlation coefficients ρ_j are replaced by the sample correlation coefficients r_j of the standardized series $Z_{v,\tau}$ and ϕ_j are replaced by the estimates $\hat{\phi}_j$;that is,

$$r_k = \hat{\phi}_1 r_{k-1} + \hat{\phi}_2 r_{k-2} + --- + \hat{\phi}_p r_{k-p} \quad ,k>0$$

In particular, for an AR(1) model,

$$\hat{\phi}_1 = r_1$$

Similarly, for an AR(2) model,

$$\hat{\phi}_1 = \frac{r_1(1-r_2)}{1-r_1^2} \quad , \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1-r_1^2}$$

The parameter σ_ϵ^2 is estimated as

$$\hat{\sigma}_\epsilon^2 = \frac{N\hat{\sigma}^2 (1 - \sum_{j=1}^{p} \hat{\phi}_j r_j)}{N-p}$$

where $\hat{\sigma}^2 = Var(z_{v,\tau}) \simeq 1$

2.3.2 Model applications

What good are stochastic models of hydrologic time series ? What use can they be put to ? When and where are they most useful ? When are they a waste of time ?

These questions, ofcourse, cannot be answered once for all cases. There are always exceptions to any rule and there are no general rules to be followed. However, some of the common applications of stochastic models are the following [Salas *et al*, 1980]:

a) Reservoir sizing studies

Perhaps the most obvious application of stochastic models is the sizing of storage reservoirs where there is a carryover from one year to the next. Reservoir sizing is the determination of a size of reservoir which is required to produce a certain dependable yield. Sizing traditionally has been done with historical data in which case the most critical (adverse) period is used for sizing. This implies an assumed unknown level of risk for failure in the future and the resulting sizes may be far from the desired optimum with a resultant economic waste. The variability of water supply is the hardest to accurately quantify and, therefore, the use of stochastically generated data usually is the best alternative.

b) Reservoir operation studies

A reservoir operation study is the determination of the best manner to operate a reservoir. As before, if the sequencing of data is important, then the use of stochastically generated data will result in an improved analysis. A better assessment of the operation plan will be made with a more accurate estimation of the probabilities associated with various levels of failure.

c) Basin-wide studies

Much more complex than the single reservoir problem is the study of a complete river basin with many data sites and many reservoirs. While sizing may still be of concern, more often basin studies deal with the operational strategy of a system of reservoirs. The determination of the best joint operational strategy for all the reservoirs within an entire basin can be studied using stochastically generated data.

2.4 Statement of the problem and methods of solving the problem

2.4.1 Statement of the problem

In this study, the statistical behaviour of monthly streamflow time series(sequences) and the modelling of such series will be discussed. The main objective of the study is to stochastically model the monthly streamflow sequences for some streams in Ethiopia using various approaches; to use the models for generation of synthetic flows ; and based on the resulting generated data, to compare the performance of each approach and make inference as to which method should be applied.

2.4.2 Methods

Here, two different approaches will be used in the stochastic modelling of monthly streamflow sequences :

I. Stationarity approach
II. Non-stationarity approach

I. STATIONARITY APPROACH

A typical hydrologic time series may be decomposed into three main components [Seth & Chandra, 1977]:

1) Deterministic trend component T_t, which represents smooth motion of the series over a long period of time,

2) Deterministic periodic component P_t, which represents a seasonal or cyclic component, and

3) Stationary stochastic component S_t, which may consist of dependent and independent (random) components.

For purposes of analysis, these components are isolated and the time series element X_t of monthly streamflow sequences is represented by an additive model of deterministic and stochastic components :

$$X_t = T_t + P_t + S_t$$

The removal of trend component, if significant, makes the given time series homogeneous and then the removal of periodic component makes the stochastic process stationary.

TREND COMPONENT

Several characteristics of time series such as the mean, standard deviation and serial correlations may be affected whenever a trend and/or a positive or negative jump (slippage) are produced in hydrologic series by non-homogeneity. The identification or detection, description and removal of such non-homogeneity are important aspects of time series analysis.

Generally trend is assumed to be found only in the mean for hydrologic series. The synthesis of streamflow data requires detection and separation of trend component from recorded sequences. However, if the same or similar trend is not likely to occur in the future, then it is not considered as part of the model in order to avoid bias in generated sequences [Seth & Chandra, 1977].

The identification and description of the characteristics of changes in hydrologic time series are based on [*Salas et al*, 1980]:

1) fitting a trend function and testing that its parameters are significantly different from zero; and

2) testing that the basic statistical characteristics of subseries of the sample series are statistically different among themselves.

The trend analysis assumes a monotonic function expanded in power series form as

$$X_t = b_0 + b_1 t + b_2 t^2 + --- + b_m t^m$$

where $b_0, b_1, ---, b_m$ are the parameters to be estimated. Only when any of the parameters $b_1, b_2, ---, b_m$ are found to be significantly different from zero, with b_0 a constant or zero, that a linear or non-linear trend becomes a characteristic of the series.

The second technique divides the historic time series into two or more subseries and the main statistical characteristics are estimated for each subseries. Then tests are performed to check whether or not these statistical characteristics are significantly different among themselves.

PERIODIC COMPONENT

Periodic hydrologic time series such as seasonal, monthly, weekly and daily series, in most cases known in nature, have significant periodic behaviour in the mean, standard deviation and skewness. In addition to these periodicities, they show a time correlation structure which may be constant or periodic. Consequently, such time dependence may be represented by say, AR models with constant or periodic coefficients. In order to obtain second order stationarity in the stochastic component, it is necessary to identify and remove periodicity in the mean and standard deviation.

Let us assume that $y_{v,\tau}$ represents monthly streamflow time series where v is the year and τ is the month. The mathematical model for the standardized homogeneous sequence $y_{v,\tau}$ can generally be written as

$$y_{v,\tau} = \mu_\tau + \sigma_\tau z_{v,\tau}$$

where μ_τ and σ_τ are the periodic mean and periodic standard deviation, respectively, and $z_{v,\tau}$ represents the value of the stochastic component in the year v and for the month τ, τ=1,...,12. This removes the periodicity in the mean and standard deviation and gives a second order stationary stochastic component. Since the monthly population parameters μ_τ

and σ_τ are not known, they must be estimated from historical data. Two procedures can be followed for obtaining these estimates [Mutreja, 1986]:

1) The non-parametric approach

Here the sample mean for the month τ is determined by

$$\hat{\mu}_\tau = \bar{y}_\tau = \frac{1}{N}\sum_{\nu=1}^{N} y_{\nu,\tau}$$

where $\tau = 1,\ldots,12$ and N is the total number of years of available data. And an estimate of σ_τ is given by

$$\hat{\sigma}_\tau = S_\tau = \sqrt{\frac{1}{N-1}\sum_{\nu=1}^{N} (y_{\nu,\tau} - \bar{y}_\tau)^2} \quad \tau = 1, \ldots, 12$$

And the standardized homogeneous sequence $y_{\nu,\tau}$ becomes

$$z_{\nu,\tau} = \frac{y_{\nu,\tau} - \bar{y}_\tau}{S_\tau} \quad \nu = 1, \ldots, N; \tau = 1, \ldots, 12$$

2) The parametric approach

In the non-parametric approach, the number of statistics used to describe the periodic parameters is twenty four and hence to reduce the number needed for the mathematical description of the periodicity, the parametric method is adopted.

Let us consider that u_τ represents a statistical characteristic of a periodic hydrologic time series $y_{\nu,\tau}$ (such as the periodic mean $\bar{y}_\tau$ and the periodic standard deviation S_τ). Assume also that u_τ is a sample estimate of the unknown population parameter denoted by ν_τ . The parametric or Fourier series representation of u_τ , denoted in general as $\hat{\nu}_\tau$, can be obtained by

$\hat{v}_\tau = \bar{u} + \sum_{j=1}^{h} (A_j \cos(\frac{2\pi j\tau}{w}) + B_j \sin(\frac{2\pi j\tau}{w}))$ $\tau=1,...,12$ where $\bar{u}$ is the mean of u_τ , A_j, B_j are the Fourier series coefficients, j is the harmonic and h is the total number of harmonics ; h = w/2 or h = (w-1)/2 according to whether w is even or odd, respectively. For instance, for monthly series w=12 and so h=6.

The mean $\bar{u}$ and the Fourier coefficients A_j and B_j are determined by

$$\bar{u} = \frac{1}{w}\sum_{\tau=1}^{w} u_\tau$$

$$A_j = \frac{2}{w}\sum_{\tau=1}^{w} u_\tau \cos(\frac{2\pi j\tau}{w}) \quad j=1,...,h$$

and

$$B_j = \frac{2}{w}\sum_{\tau=1}^{w} u_\tau \sin(\frac{2\pi j\tau}{w}) \quad j=1,...,h$$

When w is even, the last coefficients A_h and B_h are given by

$$A_h = \frac{1}{w}\sum_{\tau=1}^{w} u_\tau \cos(\frac{2\pi h\tau}{w})$$

and

$$B_h = 0$$

In the above equation, $\hat{v}_\tau$ is determined by considering all the harmonics j=1,...,h. But in practice, a smaller number of harmonics $h^* < h$ is used; that is, $\hat{v}_\tau$ should be computed with only those harmonics which significantly contribute to the variability of U_τ. The correlogram analysis, the periodogram analysis and others may be used to identify the number of significant harmonics h^*.

When dealing with daily or weekly series (or any other series with a relatively small time interval), the application of Fourier analysis is found to give better estimates of parameters. But often in practice for monthly, bimonthly, quarterly or similar series, the Fourier analysis is not applied [Salas *et al*, 1980].

STOCHASTIC COMPONENT - DEPENDENCE MODEL

The variable $z_{v,\tau}$ obtained by removing the periodicity in the mean and/or standard deviation is only approximately a second order stationary stochastic time series. The dependence is generally approximated by the first, second or third order AR linear models for short hydrologic series. As described in section 2.3, the general pth order AR linear model is

$$z_{v,\tau} = \sum_{j=1}^{p} \phi_j z_{v,\tau-j} + \sigma_\epsilon \xi_{v,\tau}$$

where ϕ_j represents estimates of constant (non-periodic) AR coefficients, σ_ϵ is a non-periodic standard deviation which gives $\xi_{v,\tau}$ as a second order stationary and independent variable with mean zero and variance one.

The parameters ϕ_j and σ_ϵ are estimated from the available sample data. For the selection of the order of the AR model, one approach is to plot the expected correlograms for the AR(1), AR(2) and AR(3) models and that of the historical series and to choose the one whose correlogram is closer to the correlogram of the historical record [Salas *et al*, 1980]. Another approach is to use the so-called Akaike Information Criterion (AIC) [Akaike, 1974]. For comparing competing AR(p) models, compute

$$AIC(p) = N \ln(\hat{\sigma}_\epsilon^2) + 2p$$

where N is the sample size and $\hat{\sigma}_\epsilon^2$ is an estimate of the residual variance, and choose the model which gives the minimum AIC. The interpretation of this criteria is as follows.

As the order of the model increases (i.e. more explanatory variables), we expect a reduction of the residual variance. But, the larger the order, the greater is the number of estimated parameters and hence the more unreliable the model becomes. Therefore, a model which minimizes the sum of these two effects is desirable.

INDEPENDENT STOCHASTIC COMPONENT

The independent stochastic component $\xi_{\nu,\tau}$ is obtained after removing the dependence structure from the standardized residual component $z_{\nu,\tau}$ using the AR model of particular order selected. The $\xi_{\nu,\tau}$ of the process thus computed is a random variable with mean zero and variance one. Once $\xi_{\nu,\tau}$ is accepted as a stationary independent variable, a theoretical best fit distribution is found from normal, three parameter Lognormal or three parameter Gamma distribution, using Chi-square criteria. Sometimes Kolmogorov-Smirnov test is also used. The performance of stationarity approach have been found to depend on the best fit distribution for the random component and generation of random numbers having that probability density function [Seth & Chandra, 1977]. However, since most probability theory and statistical techniques applied to hydrologic time series analysis are developed assuming the variables are normally distributed, it is often necessary to transform the original skewed series into normal before carrying out the statistical analysis of interest. In such cases, the generation of a synthetic time series will start with the generation of independent normal variables with mean zero and variance one, and then adding the time dependence structure as well as periodic components, whichever is necessary [Salas *et al*,1980].

At last, a word of warning may be desirable against unjustified attempts at detection and removal of trend component when analyzing a time series by the stationarity approach. That is, slow oscillatory movements (low frequency random effects - a considerable rise or decline of a time series once in a decade, or once in a quarter of a century, etc) might be treated or regarded as trends and as a result, the inference from such 'trend' in a particular period to the effect in the future might be quite false [Kendall & Stuart, 1968]. Therefore, the verification process of a trend should ascertain whether the history of data collection (various sources of systematic errors); the history of river basin developments (the abstraction of water from a river or the construction of a reservoir affecting a lake level); or the other natural factors (changes in nature by natural disruptive, evolutive or sudden processes) may support the final acceptance of the inferred trend. In short, the detection and removal of a trend is more reliable if changes are substantiated by both statistical tests and physical or historical evidence and justification [Salas *et al*, 1980].

II. NON-STATIONARITY APPROACH

Many investigators have considered the hydrologic time series as a non-stationary process entirely. It is similar to Markov chain model in which the variate at the t-th time is comprised of a component linearly related to that at (t-1)th time and a random component. For relating the streamflow during any month to that during the month immediately preceding it, this method uses serial correlation coefficients of monthly streamflow computed from the record using linear regression analysis. The following recursion equation [Thomas and Fiering,1962] can be written to represent this model :

$$Q_{i+1} = \overline{Q}_{j+1} + B_j(Q_i - \overline{Q}_j) + Z_i S_{j+1}[1 - r_j^2]^{\frac{1}{2}}$$

where Q_i and Q_{i+1} are flows during i-th and (i+1)th months, respectively, counted from the start of the generated sequence; $\overline{Q}_j$ and $\overline{Q}_{j+1}$ are the mean monthly flows during j-th and (j+1)th months , respectively, within a repetitive annual cycle of 12 months; S_{j+1} is the standard deviation of monthly flows during (j+1)th month; B_j is the regression coefficient for estimating the flow in the (j+1)th month from the flow in the j-th month; r_j is the correlation coefficient between the flows of j-th and (j+1)th months and Z_i is an independent variable with mean zero and variance 1.

The square root and logarithmic transformation of the original data have been recommended by Thomas and Fiering to avoid negative values of generated flows and also to narrow down the gap between high flows and low flows.

As in the case of stationarity approach, this method also requires generation of random numbers of appropriate probability distribution. But since most frequency curves of hydrologic variables are asymmetrically distributed, or are bound by zero (they are positively valued variables), it is often necessary to transform these variables to normal before carrying out the statistical analysis of interest.

Generation of normally distributed independent random numbers

Whenever historical flows are available, a sequence of future flows can be synthesised by consulting a table of random numbers, or by generating using a digital computer or by similar approaches.

The first step in synthetic generation of flows is the generation of uniformly distributed independent random numbers. In our case, this was done by using RANDOM, a built-in routine available in the Turbo Pascal compiler, version 4.0. The second step is the transformation of these random numbers into those having the required fitted distribution of the observed streamflow data. In this study, because of various reasons mentioned above, the observed data were transformed into normal before carrying out the desired analysis. Therefore, in our case, we need to transform the generated uniformly distributed independent random numbers into normal. This was done by using the Polar Marsaglia method. The procedure is as follows. Let U_1 and U_2 be two independent uniform(0,1) random numbers. Let

$$V_1 = 2U_1 - 1$$

$$V_2 = 2U_2 - 1$$

and

$$q_2 = V_1^2 + V_2^2$$

If $q_2 < 1$, then

$$R_1 = V_1 q$$

and

$$R_2 = V_2 q$$

will be a pair of independent standard normal random numbers, where

$$q = [-2 \ln (q_2) / q_2]^{1/2}$$

Using the above steps, normally distributed independent random numbers of length as required can be generated.

CHAPTER THREE

MODELLING AND GENERATION OF MONTHLY STREAMFLOW SEQUENCES

3.1 Modelling of Megecha riverflow time series and generation of synthetic flows

3.1.1 A stationarity approach

Megecha river is located in the Ghibie - Omo basin. It has a total drainage area of 286 square kilometres. Monthly runoff data (in million cubic meters) recorded at a gauging station near Gubrie for 12 years (1977 - 1988) are used for analysis.

I. Preliminary analysis

The monthly time series, denoted by $y_{v,\tau}$, v=1,...,12 and τ=1,..,12 , are plotted as shown in Fig.1. It shows that during the summer (July, August, and September) the flows are generally higher than during the rest of the year; this situation repeats itself every year in a periodic manner.

To check whether the flows are trend free or not, a linear trend function is fitted. The trend function is :

$$y_t = \beta_0 + \beta_1 t = 9.479272 - 0.006181t$$

where y_t is the flow in month t , t=1,2,...,12(12)=144. We say a linear trend exists if the slope is significantly different from zero.Here the standard error of β_1 is $SE(\beta_1)$ =0.02673 and

$$T = \frac{\beta_1}{SE(\beta_1)} = -0.231$$

Since the sample size (N=144) is large, the sampling distribution of T is approximately standard normal. For α = 0.05, the $(1-\alpha/2)$th quantile of the standard normal distribution is 1.96. Since the absolute value of T is less than 1.96, the hypothesis that the slope is not significantly different from zero is accepted. Thus, we can assume that the flows are trend free.

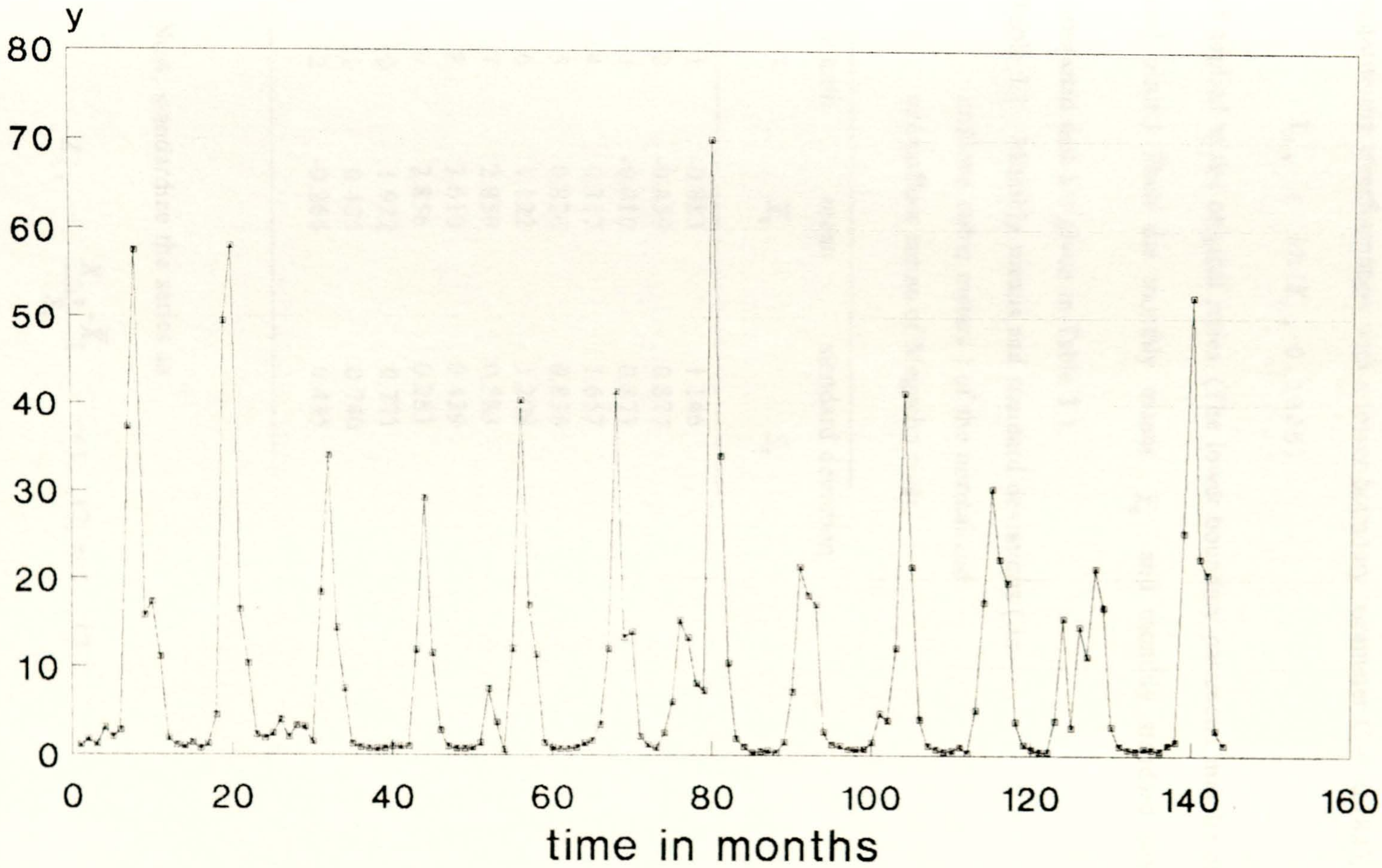

Figure 1: Monthly streamflow series of
Megecha river for the period 1977-1988

II. Parameter estimation and model identification

In order to avoid negative values of generated flows and to make the data normal, a logarithmic transformation with a lower boundary parameter C = 0.1825, i.e.,

$$X_{v,\tau} = \ln(Y_{v,\tau} - 0.1825)$$

is applied to the original series. (The lower boundary parameter may be obtained by trial and error.) Then the monthly means $\overline{x}_\tau$ and monthly standard deviations S_τ are computed and are given in Table 3.1.

Table 3.1 : Monthly means and standard deviations (in millions cubic meters) of the normalized streamflow series of Megecha river.

month τ	mean $\overline{x}_\tau$	standard deviation S_τ
1	-0.883	1.146
2	-0.459	0.877
3	-0.010	0.873
4	0.317	1.657
5	0.826	0.836
6	1.122	1.209
7	2.859	0.583
8	3.610	0.439
9	2.856	0.283
10	1.922	0.771
11	0.405	0.740
12	-0.266	0.435

Now, standardize the series as :

$$Z_{v,\tau} = \frac{X_{v,\tau} - \overline{X}_\tau}{S_\tau} \quad , v=1,...,12;\ \tau=1,...,12$$

This standardized series has approximately mean zero and variance one. Since this series is free of periodicity, we can write it as Z_t ,t=1,..,12(12)=144. Then the correlogram of Z_t , $r_k(Z)$, and the expected correlograms for the AR(1), AR(2) and AR(3) models are computed.

1) AR(1) model

The model is

$$z_t = \phi_1 z_{t-1} + \varepsilon_t \qquad (\varepsilon_t = \sigma_\varepsilon \xi_t)$$

An estimate of the AR coefficient ϕ_1 is

$$\hat{\phi}_1 = 0.4549.$$

The autocorrelation function (ACF) is

$$r_k = \hat{\phi}_1^k \quad , \quad k>0$$

An estimate of the error standard deviation is

$$\hat{\sigma}_\varepsilon = 0.8936519.$$

2) AR(2) model

The model is

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \varepsilon_t$$

The estimates of the AR coefficients ϕ_1 and ϕ_2 are

$\hat{\phi}_1 = 0.4437$ and $\hat{\phi}_2 = 0.0246.$

The ACF is

$$r_k = \hat{\phi}_1 r_{k-1} + \hat{\phi}_2 r_{k-2} \quad , \quad k>0$$

An estimate of the error standard deviation is

$$\hat{\sigma}_\varepsilon = 0.8965210$$

3) AR(3) model

The model is

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \phi_3 z_{t-3} + \varepsilon_t$$

By solving the ACF (system of linear equations)

$$r_k = \phi_1 r_{k-1} + \phi_2 r_{k-2} + \phi_3 r_{k-3} \quad , \quad k>0$$

we get $\hat{\phi}_1 = 0.4415$, $\hat{\phi}_2 = -0.0252$ and $\hat{\phi}_3 = 0.1107$.

An estimate of the error standard deviation is

$$\hat{\sigma}_\varepsilon = 0.8941214.$$

The computed values of r_k ,k=1,2,..,10 , for the three models are given in Table 3.2 below.

Table 3.2 : Expected correlograms for AR(1),AR(2) and AR(3) models.

k	AR(1)	AR(2)	AR(3)
1	0.4549	0.4549	0.4551
2	0.2069	0.2265	0.2261
3	0.0941	0.1117	0.1990
4	0.0428	0.0551	0.1326
5	0.0195	0.0272	0.0785
6	0.0089	0.0134	0.0534
7	0.0040	0.0066	0.0363
8	0.0018	0.0033	0.0234
9	0.0008	0.0016	0.0153
10	0.0004	0.0008	0.0102

Now, based on these values of r_k ,k=1,2,..,10 , of the AR(1), AR(2) and AR(3) models, the expected correlograms are plotted together with the sample (historical) correlogram as shown in Figures 2(a), 2(b), 2(c) and 2(d).

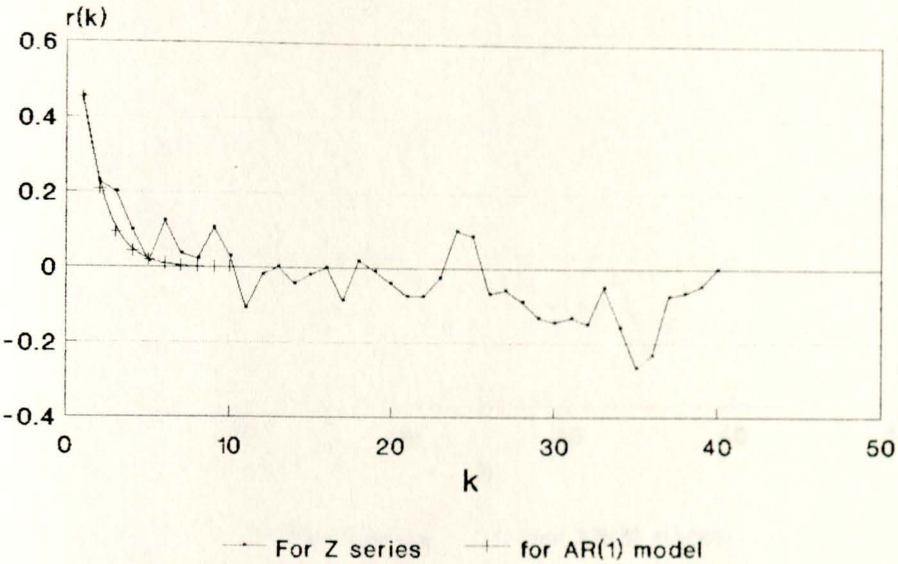

Figure 2(a) : Correlogram of the standardized series and expected correlogram for the AR(1) model.

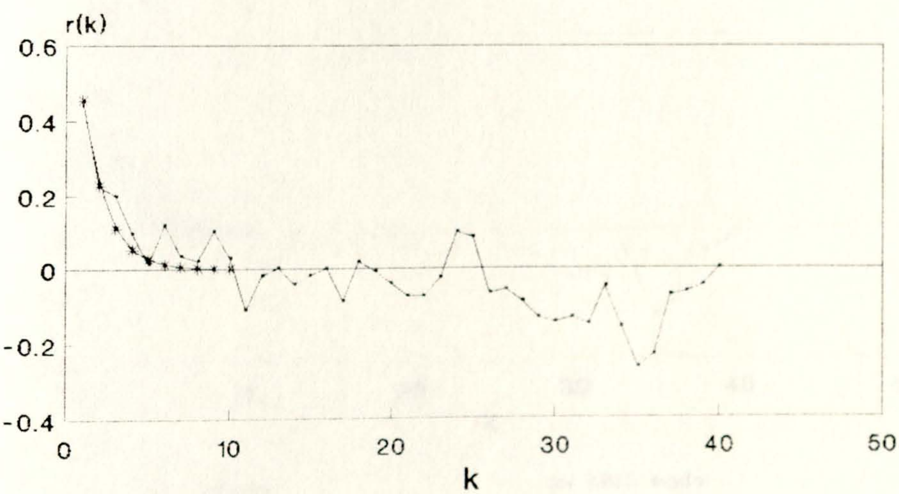

Figure 2(b) : Correlogram of the standardized series and expected correlogram for the AR(2) model.

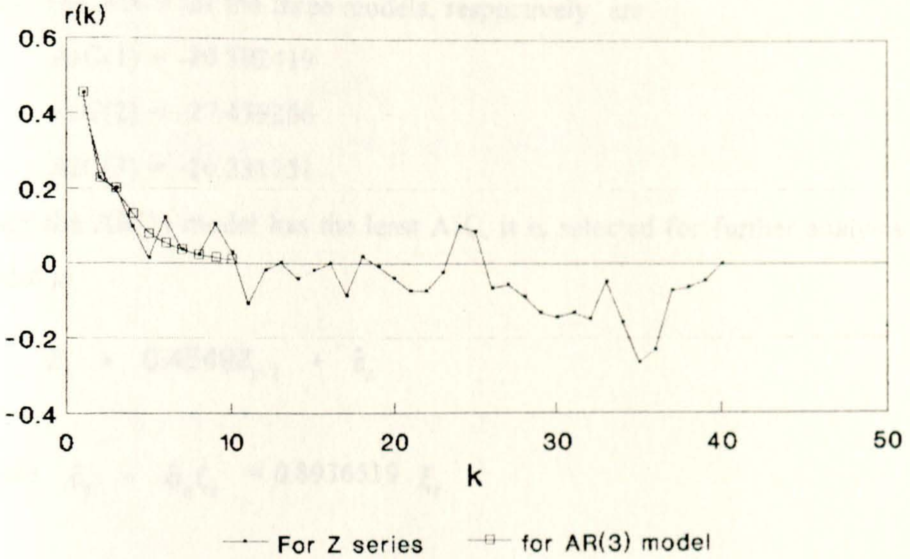

Figure 2(c) : Correlogram of the standardized series and expected correlogram for the AR(3) model.

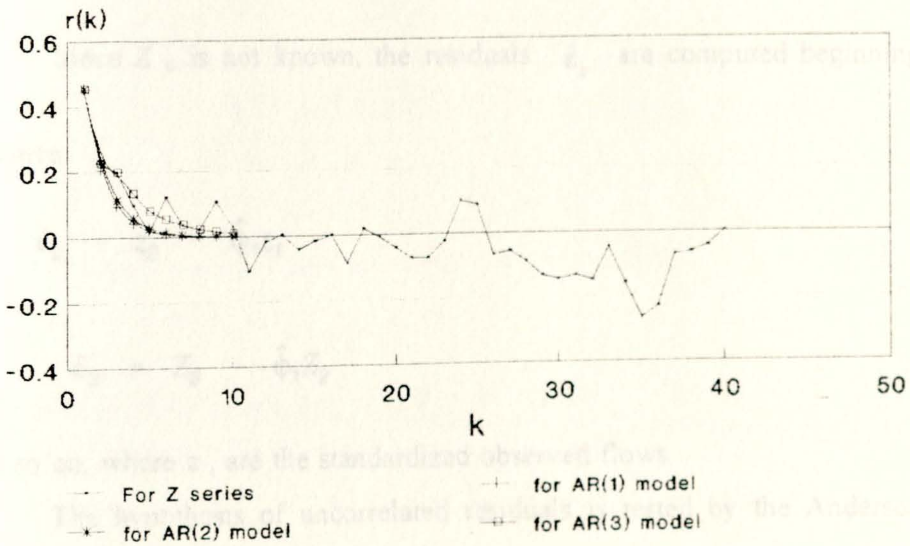

Figure 2(d) : Correlogram of the standardized series and expected correlograms for AR(1), AR(2) and AR(3) models.

Visual inspection of the plots doesn't help to decide which expected correlogram is closer to the sample correlogram. Therefore the so-called Akaike Information Criterion (AIC) is used. The AIC's for the three models, respectively, are

AIC(1) = -30.382419

AIC(2) = -27.459266

AIC(3) = -26.231151

Since the AR(1) model has the least AIC, it is selected for further analysis. The selected model is

$$Z_t = 0.4549 Z_{t-1} + \hat{\varepsilon}_t$$

where $\hat{\varepsilon}_t = \hat{\sigma}_\varepsilon \xi_t = 0.8936519\ \xi_t$.

Since $|\hat{\phi}_1| < 1$ we can conclude that the estimated parameter $\hat{\phi}_1$ complies with the stationarity condition.

III. Tests of goodness of fit and generation of synthetic flows

Since Z_0 is not known, the residuals $\hat{\varepsilon}_t$ are computed beginning at t=2. For instance

$$\hat{\varepsilon}_2 = z_2 - \hat{\phi}_1 z_1$$

$$\hat{\varepsilon}_3 = Z_3 - \hat{\phi}_1 Z_2$$

and so on, where z_t are the standardized observed flows.

The hypothesis of uncorrelated residuals is tested by the Anderson test of the correlogram [Anderson, 1941]. The tolerance band around r_k at tolerance level 5% on each side is given by:

$$r_k(5\%) = \frac{-1 \pm 1.96\sqrt{N-K-1}}{N-K}$$

where N is the total number of observations and k is the maximum lag considered. Using the above formula the 95% lower and upper probability limits are computed, respectively, as -0.2009 and 0.1817. The correlogram of the residuals together with the 95% limits is plotted as shown in Fig.3. From the figure we can see that all values lie inside the limits. Therefore, the hypothesis of uncorrelated residuals is accepted.

The hypothesis of normality of the residuals is tested by the skewness test of normality. In our case, the skewness coefficient is $\hat{\gamma}$ =0.019. For a normal population, the (1-α) probability limits on the population skewness coefficient γ may be defined by [Snedecor & Cochran, 1967]:

$$\left(-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{6}{N}}\ ,\ Z_{1-\frac{\alpha}{2}}\sqrt{\frac{6}{N}}\right)$$

For α=0.05, the limits are (-0.3998,0.3998). Since $\hat{\gamma}$ lies in this interval, the hypothesis of normality of the residuals is accepted.

Now, standardize the residuals as

$$\xi_t = \frac{\hat{\varepsilon}_t}{\hat{\sigma}_\varepsilon} = \frac{\hat{\varepsilon}_t}{0.8936519}$$

This standardized series is approximately standard normal.

Here our model is

$$\hat{Z}_t = 0.4549\hat{Z}_{t-1} + 0.8936519\xi_t \qquad (3.1)$$

First generate standard normal random numbers ξ_t and then assuming Z_0 =0, use equation (3.1) to generate the variable $\hat{Z}_t$. To begin the generation we assumed Z_0 =0 and as a result the first generated values of $\hat{Z}_t$ are biased. To avoid such bias, the

procedure is [Fiering and

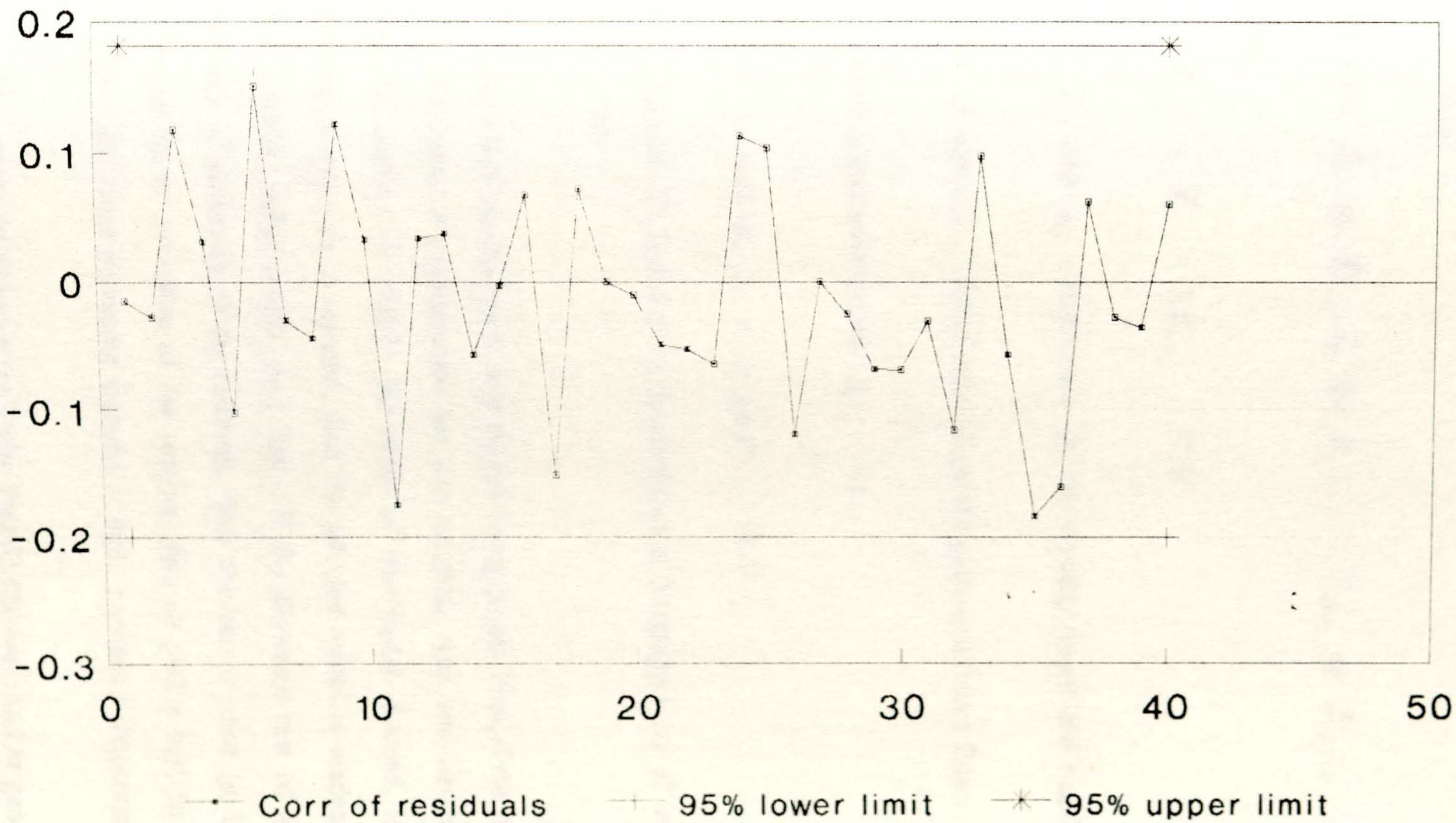

Figure 3 : Correlogram of the residuals together with the 95% lower and upper limits.

Jackson, 1971] to drop the first 50 generated values of $\hat{z}_t$ and then to re-initialize

$\hat{z}_t$ as $\hat{z}_{t+50}$.

Then write $\hat{z}_1$ *as* $\hat{z}_{1,1}, \hat{z}_2$ *as* $\hat{z}_{1,2}, \ldots, \hat{z}_{144}$ *as* $\hat{z}_{12,12}$,...

so that

$$\hat{X}_{\nu,\tau} = \overline{X}_\tau + S_\tau \hat{z}_{\nu,\tau} \qquad (3.2)$$

where $\overline{X}_\tau$ and s_τ ,respectively, are the monthly means and standard deviations of the normalized series $X_{\nu,\tau}$. And finally to get the generated future flows $\hat{y}_{\nu,\tau}$ apply an anti-logarithmic transformation on $\hat{x}_{\nu,\tau}$; i.e.,

$$\hat{Y}_{\nu,\tau} = \exp(\hat{x}_{\nu,\tau}) + 0.1825 \qquad (3.3)$$

Using equations (3.2) and (3.3), future flows of Megecha river of length as required can be generated.

Here it is worthwhile to note the following points. First, if the stationarity conditions of the estimated AR parameters are not satisfied, then one has to either change the estimation method or change the order of the model. Second, if the hypothesis of uncorrelated residuals is rejected, then the selected model is inadequate and one has to select a higher order model. And lastly, if the skewness test of normality rejects the hypothesis of normality of the residuals, then one has to either go back and try another transformation to normality of the original data or find a best fit distribution for the residuals (from three parameter Gamma or three parameter lognormal).

The computer program (in Turbo Pascal) that was used to generate synthetic flows by the stationarity approach is given in Appendix A.

3.1.2 A non-stationarity approach

Here, a mathematical model developed by Thomas and Fiering (1962) will be used. The method assumes that the streamflow is made up of two components: the deterministic and the random. As described in the previous section, the model for unit time interval (one month in our case) to generate flows is:

$$Q_{i+1} = \overline{Q}_{j+1} + B_j(Q_i-\overline{Q}_j) + Z_i S_{j+1}(1-r_j^2)^{\frac{1}{2}} \tag{3.4}$$

i=0,1,...,12N-1;j=1,2,...,12 and N is the total number of record years to be generated.

The monthly streamflow observations, y_i, of Megecha river are found out to be skewed. A simple logarithmic transformation

$x_i = \mathrm{Ln}(y_i - 0.1825)$

makes the observations normal with mean 1.032 and standard deviation 1.658. And hence

$a_i = (x_i - 1.032)/1.658$

is standard normal.

Using this standardized series, the values of $\overline{Q}_j$, S_j, r_j and B_j are computed for j=1,...,12. These values are given in Table 3.3.

By substituting these estimates, Equation (3.4) is recursively used to get values of Q_{i+1}, i=0,1,...,12N-1 where N is the total number of record years to be generated, and then the inverse (backward) transformation

$y_{i+1} = \exp(1.658\, Q_{i+1} + 1.032) + 0.1825$

is applied to obtain the generated future flows.

The computer program (in Turbo Pascal) that was used to generate synthetic flows by the non-stationarity model of Thomas and Fiering is given in Appendix B.

Table 3.3 : Monthly means, standard deviations, correlation and regression coefficients of the standardized normalized streamflow sequence of Megecha river.

month	mean	standard deviation	correlation coefficient	regression coefficient
j	$\bar{Q}_j$	S_j	r_j	B_j
1	-1.1552	0.6911	0.6704	0.513343
2	-0.8995	0.5290	0.4553	0.453061
3	-0.6283	0.5266	0.7825	1.486542
4	-0.4301	1.0003	0.4803	0.242317
5	-0.1245	0.5045	0.3321	0.480008
6	0.0544	0.7295	0.1073	0.051722
7	1.1017	0.3519	0.0604	0.045420
8	1.5552	0.2646	0.3906	0.251766
9	1.1002	0.1705	0.2036	0.555120
10	0.5367	0.4649	0.7533	0.723420
11	-0.3782	0.4465	0.8289	0.486585
12	-0.7829	0.2621	0.1691	0.445710

3.2 Modelling of Gudla riverflow time series and generation of synthetic flows

3.2.1 A stationarity approach

River Gudla is located in the western part of Ethiopia in the Blue Nile basin. It is a tributary to Abay river with a total drainage area of 242 Sq.Kms. Monthly run off data (in million cubic meters) for 33 years (1960-1992) recorded at a gauging station near Dembecha are used for analysis.

I. Preliminary Analysis

The monthly streamflow sequence $Y_{v,\tau}$,v=1,...,33; τ=1,...,12 is plotted as shown in Fig.4(a) and 4(b). From the figures it can be seen that during the months of July, August and September the flows are higher than during the rest of the year.

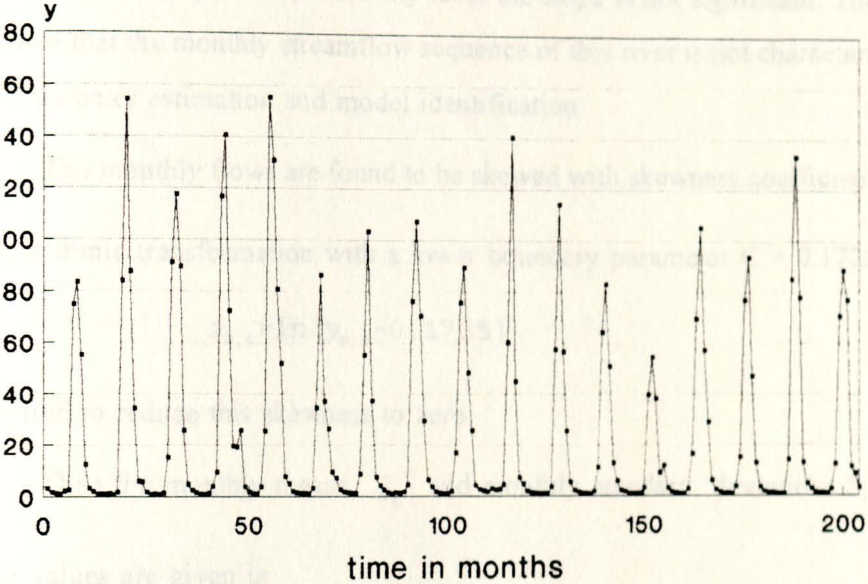

Figure 4(a):Time series of monthly streamflow of Gudla river for the period 1960 - 1976.

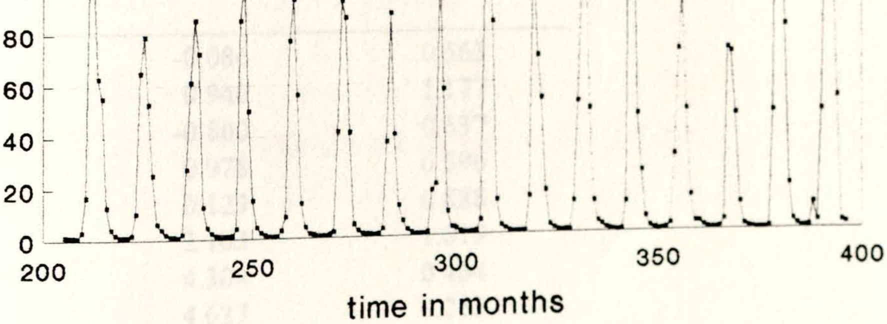

Figure 4(b):Time series of monthly streamflow of Gudla river for the period 1977-1992.

For monthly streamflow data of Gudla river a linear trend function

$$Y_t = 22.83926 + 0.009617t$$

is fitted. At the 95 percent probability level the slope is not significant. Therefore, we can conclude that the monthly streamflow sequence of this river is not characterized by a trend.

II. Parameter estimation and model identification

The monthly flows are found to be skewed with skewness coefficient $\gamma = 1.705$.

A logarithmic transformation with a lower boundary parameter C = 0.1725, i.e.,

$$x_{v,\tau}=\ln(y_{v,\tau}-0.1725)$$

is applied to reduce this skewness to zero.

Then the monthly means $\bar{x}_\tau$ and monthly standard deviations S_t are computed.

These values are given in

Table 3.4. Using these values standardize the data as

$$z_{v,\tau}=\frac{x_{v,\tau}-\bar{x}_\tau}{s_\tau}, \quad \tau=1,...,12 \text{ and } v=1,...,33$$

Table 3.4 : Monthly means and standard deviations of the normalized streamflow sequence of river Gudla.

month τ	mean $\bar{x}_\tau$	standard deviation S_τ
1	-0.086	0.563
2	0.940	1.177
3	-0.800	0.637
4	0.978	0.896
5	0.123	0.888
6	2.163	1.019
7	4.304	0.464
8	4.627	0.253
9	4.065	0.289
10	2.861	0.575
11	1.522	0.515
12	0.796	0.650

Now write $z_{v,\tau}$ as z_t ,t=1,...,12(33)=396 and determine the correlogram $r_k(z)$ of z_t.

Also compute the expected correlograms for AR(1), AR(2) and AR(3) models.

1) AR(1) model

The model is

$$z_t=\phi_1 z_{t-1}+\varepsilon_t$$

with estimates $\hat{\phi}_1$ = 0.5566 and $\hat{\sigma}_\varepsilon$ = 0.831841.

2) AR(2) model

The model is

$$z_t=\phi_1 z_{t-1}+\phi_2 z_{t-2}+\varepsilon_t$$

with estimates $\hat{\phi}_1$ = 0.5447, $\hat{\phi}_2$ = 0.0214 and

$\hat{\sigma}_\varepsilon$ = 0.8327058.

3) AR(3) model

The model is

$$z_t=\phi_1 z_{t-1}+\phi_2 z_{t-2}+\phi_3 z_{t-3}+\varepsilon_t$$

with estimates $\hat{\phi}_1$ = 0.5455, $\hat{\phi}_2$ = 0.0309, $\hat{\phi}_3$ = -0.0175 and $\hat{\sigma}_\varepsilon$ = 0.8334994.

In Table 3.5 below, the correlogram of the historical monthly flows of river Gudla is given. By using the estimates of the AR coefficients for the AR(1), AR(2) and AR(3) models, the expected correlograms are computed and are given in

Table 3.6.

Table 3.5 : Correlogram of the historical monthly streamflow sequence of Gudla river.

k	$r_k(z)$	k	$r_k(z)$
1	0.557	21	-0.024
2	0.325	22	0.027
3	0.177	23	0.066
4	0.067	24	0.094
5	-0.001	25	0.145
6	0.016	26	0.081
7	0.010	27	0.104
8	0.061	28	0.081
9	0.067	29	-0.019
10	0.132	30	-0.062
11	0.136	31	-0.028
12	0.113	32	-0.003
13	0.091	33	0.034
14	0.072	34	0.038
15	0.085	35	0.064
16	0.022	36	0.069
17	0.015	37	0.009
18	-0.013	38	-0.015
19	-0.059	39	0.031
20	-0.051	40	0.027

Table 3.6 : Expected correlograms for the AR(1), AR(2) and AR(3) models.

k	AR(1)	AR(2)	AR(3)
1	0.5566	0.5566	0.5570
2	0.3098	0.3245	0.3250
3	0.1724	0.1887	0.1770
4	0.0960	0.1097	0.0969
5	0.0534	0.0638	0.0526
6	0.0297	0.0371	0.0286
7	0.0165	0.0216	0.0155
8	0.0092	0.0125	0.0084
9	0.0051	0.0073	0.0046
10	0.0029	0.0042	0.0025

For identification purposes, the expected correlogram of each of the above models is plotted together with the historical correlogram as shown in Figures 5(a), 5(b), 5(c) and 5(d). It is really hard to identify the model whose correlogram is closer to the sample correlogram by visual inspection of the plots. Therefore, the Akaike Information Criterion (AIC) is used for this purpose. The AIC's for the AR(1), AR(2) and AR(3) models, respectively, are

AIC(1) = -143.818

AIC(2) = -140.995

AIC(3) = -138.241

Thus, the AR(1) model has the least AIC and is selected for further analysis. The selected model is

$$z_t = 0.5566 z_{t-1} + \hat{\varepsilon}_t$$

where $\hat{\varepsilon}_t = \hat{\sigma}_\varepsilon \xi_t = 0.831841 \xi_t$.

Since $|\hat{\phi}_1|$ = 0.5566 < 1, the stationarity condition for the AR(1) model is satisfied.

III. Tests of goodness of fit and generation of synthetic flows.

The residuals $\hat{\varepsilon}_t$ are computed beginning at t=2 since z_0 is not known. For instance,

$$\hat{\varepsilon}_2 = z_2 - \hat{\phi}_1 z_1$$

$$\hat{\varepsilon}_3 = z_3 - \hat{\phi}_1 z_2$$,and so on.

To test whether the residuals are correlated or not, Anderson's test of the correlogram at 5 percent tolerance level is used. The computed 95 percent lower and upper probability limits, respectively, are -0.1065 and 0.1009. The correlogram of the residuals together with these limits is plotted as shown in Figure 6. Here only one value lies outside the limits (a maximum of (1-0.95)40 = 2 is allowed) and , therefore, we can conclude that the residuals for the AR(1) model are uncorrelated.

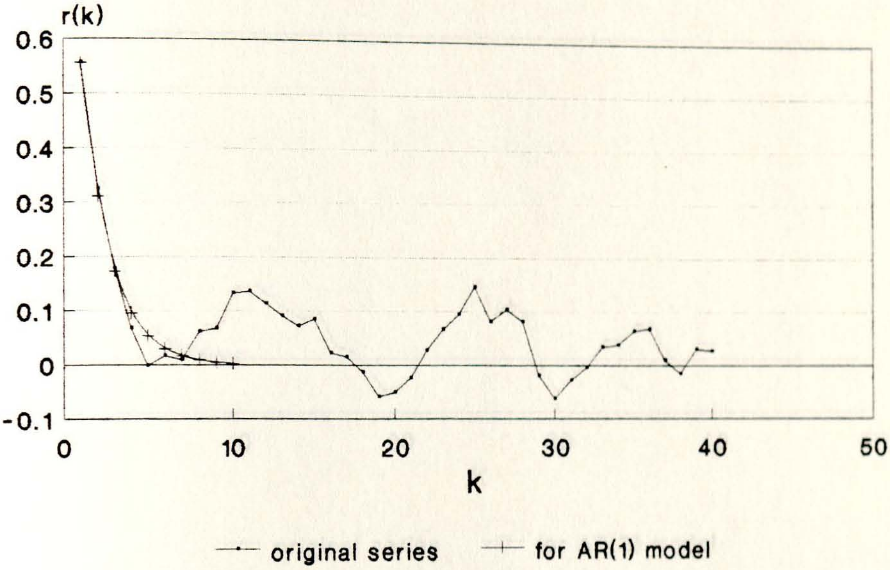

Figure 5(a) :Correlogram of the standardized historical series and expected correlogram for the AR(1) model.

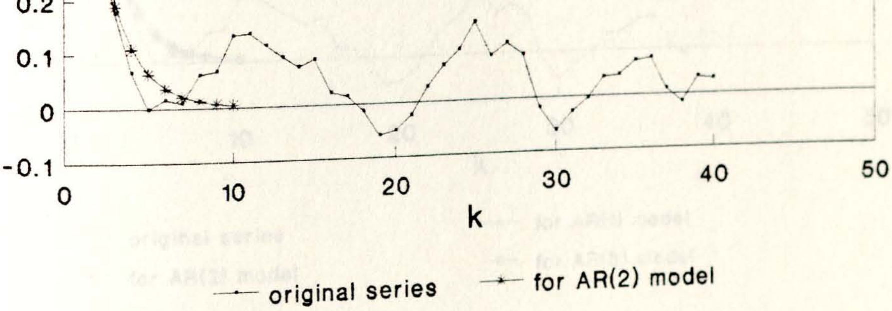

Figure 5(b) :Correlogram of the standardized historical series and expected correlogram for the AR(2) model.

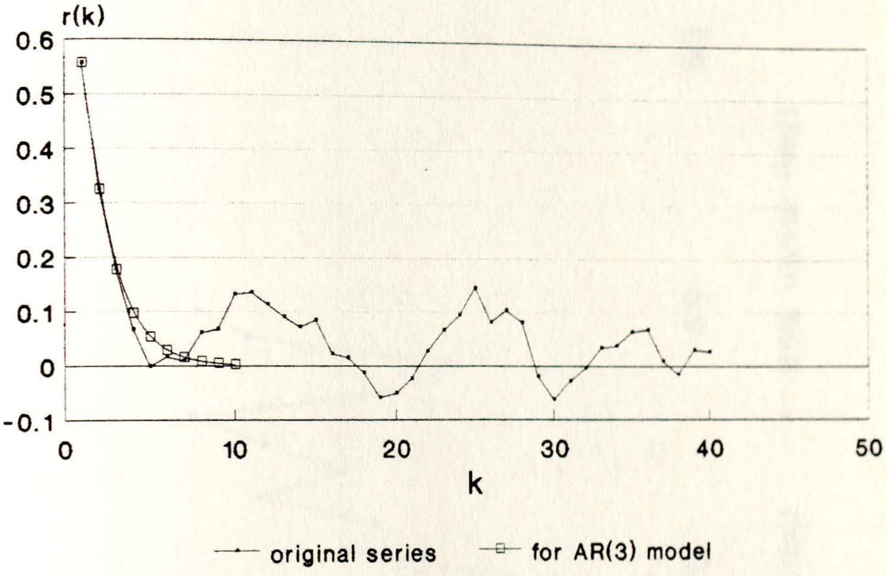

Figure 5(c) :Correlogram of the standardized historical series and expected correlogram for the AR(3) model.

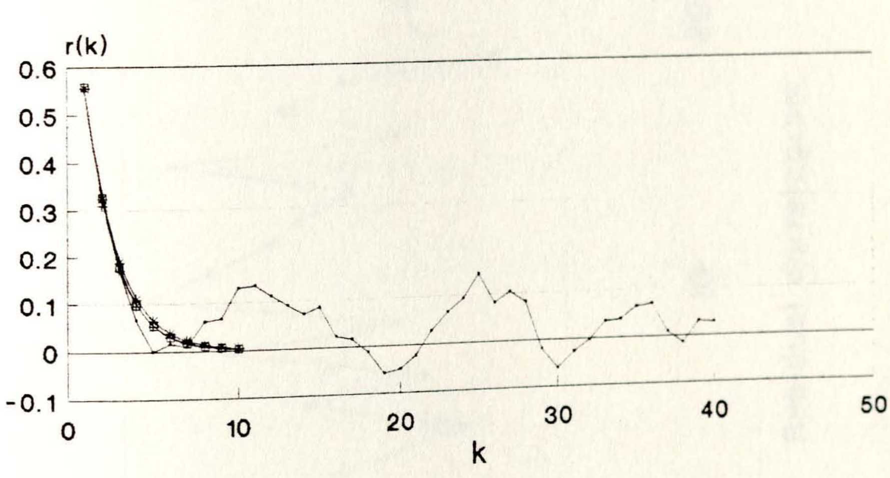

Figure 5(d) :Correlogram of the standardized historical series and expected correlograms for the three models.

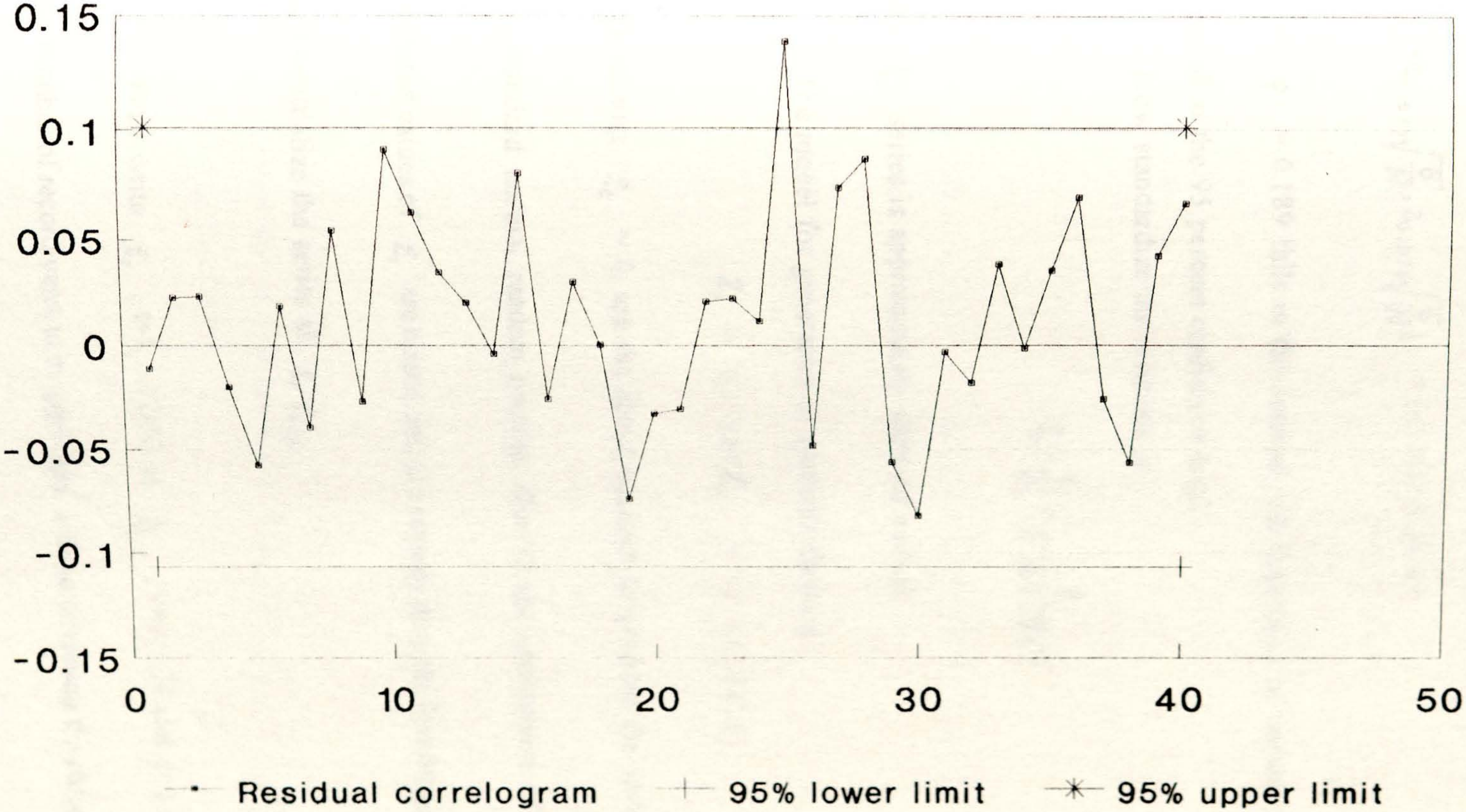

Figure 6 : correlogram of the error terms (residuals) with the 95% lower and upper tolerance limits.

The coefficient of skewness of the residuals is $\hat{\gamma}$ = 0.189. and the 95 percent confidence interval for the population skewness coefficient is

$$(-z_{0.975}\sqrt{\frac{6}{N}}, z_{0.975}\sqrt{\frac{6}{N}}) = (-0.2411, 0.2411).$$

Since $\hat{\gamma}$ = 0.189 falls in this interval, the hypothesis of normality of the residuals is accepted at the 95 percent confidence level.

Now standardize the residuals as

$$\hat{\xi}_t = \frac{\hat{\varepsilon}_t}{\hat{\sigma}_\varepsilon} = \frac{\hat{\varepsilon}_t}{0.831841}$$

This $\hat{\xi}_t$ series is approximately standard normal.

The model for generation of synthetic flows is

$$\hat{z}_t = 0.5566\hat{z}_{t-1} + 0.831841\xi_t$$

By assuming $\hat{z}_0$ = 0, use the above equation to generate the variable $\hat{z}_t$ where ξ_t is a standard normal random number. Due to the assumption of $\hat{z}_0$ = 0, the first generated values of $\hat{z}_t$ are biased, and as a remedy drop the first 50 generated values of $\hat{z}_t$ and re-initialize the series as $\hat{z}_t = \hat{z}_{t+50}$.

Then write $\hat{z}_t$, t=1,...,12(N) as $\hat{z}_{v,\tau}$, v=1,...,N and τ=1,...,12, where N is the total number of record years to be generated, without disturbing the chronological order, so that

$$\hat{x}_{v,\tau} = \bar{x}_\tau + s_\tau \hat{z}_{v,\tau}$$

where $\overline{x}_\tau$ and s_τ are the monthly means and standard deviations of the normalized series $x_{v,\tau}$. And finally, to obtain the skewed generated synthetic flows $\hat{y}_{v,\tau}$ of Gudla river, apply the inverse of the normalization transformation, i.e.;

$$\hat{y}_{v,\tau} = \exp(\hat{x}_{v,\tau}) + 0.1725$$

3.2.2 A non-stationarity approach

Let y_i, i=1,...,396, denote the streamflow sequence of Gudla river. As stated in the previous section, the flows are skewed. Therefore, a logarithmic transformation

$$x_i = \ln(y_i - 0.1725)$$

is applied to make the flows normal with mean 1.471 and standard deviation 2.138. And finally, the sequence is standardized as

$$a_i = (x_i - 1.471)/2.138$$

This sequence a_i is approximately standard normal. Then determine the monthly means $\overline{Q}_j$, the monthly standard deviations S_j, and the regression coefficients B_j and the correlation coefficients r_j of successive months, j=1,...,12.These values are given in Table 3.7.

Then insert these values in the model :

$$Q_{i+1} = \overline{Q}_{j+1} + B_j(Q_i - \overline{Q}_j) + Z_i S_{j+1}\sqrt{1 - r_j^2}$$

where z_i is a sequence of standard normal random numbers. Finally, apply the inverse of the standardization and normalization transformations on Q_{i+1} to get the generated flows, i.e., apply the transformation :

$$\hat{y}_{i+1} = \exp(2.138 Q_{i+1} + 1.471) + 0.1725$$

i=0,...,12N-1, where N is the number of record years to be generated.

Table 3.7 : Monthly means, standard deviations, regression and correlation coefficients of the standardized normalized series a_i

month	mean	std.dev.	corr.coe.	regr.coe.
j	$\overline{Q}_j$	S_j	r_j	B_j
1	-0.728	0.263	0.7767	1.623561
2	-1.127	0.551	0.7875	0.426318
3	-1.062	0.298	0.5658	0.795559
4	-1.145	0.419	0.2032	0.201343
5	-0.630	0.415	0.4589	0.526827
6	0.324	0.477	0.4451	0.202733
7	1.325	0.217	0.4863	0.264944
8	1.476	0.118	0.5469	0.624314
9	1.213	0.135	0.5209	1.037894
10	0.650	0.269	0.2904	0.260209
11	0.024	0.241	0.7889	0.995751
12	-0.316	0.304	0.1367	0.118352

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Testing the adequacy of models

A modeller is usually interested in finding a model which can reproduce the historical statistical characteristics considered relevant. Usually the main statistical characteristics that the generated data is required to preserve are the historical mean, standard deviation, skewness coefficient and correlogram.

In order to compare the statistical characteristics of the generated data with those of the historical data, first based on the selected model generate say m samples of length the same as that of the historical sample and then determine the mean $\hat{\mu}(i)$ standard deviation $\hat{\sigma}(i)$, skewness coefficient $\hat{\gamma}(i)$ and correlogram $r_k(i)$, (k=1,...,40) for each generated sequence, i=1,...,m. Let u(i) represent any of these statistical characteristics. Compute the sample mean and standard deviation of the u(i)'s, respectively, as

$$\overline{u} = \frac{1}{m}\sum_{i=1}^{m} u(i)$$

$$s(u) = \sqrt{\frac{1}{(m-1)}\sum_{i=1}^{m} (u(i) - \overline{u})^2}$$

Then calculate the confidence interval

$$[\ \overline{u} - c\, S(u)\ ,\ \overline{u} + c\, S(u)\ ]$$

Usually c is taken as 1, 1.5 or 2 depending on how strict the test is made. Alternatively, c could be taken as the standard normal variate corresponding to a specified significance level such as 1.64 for ten percent level [Salas *et al*, 1980]. Finally check whether the historical statistical value falls inside its respective interval or not. If the historical statistical value doesn't fall inside the interval, then the analyst has to either change the parameters of the model or perhaps change the model itself. But if all statistical values considered relevant fall in their respective confidence intervals, then we say the model is adequate.

4.2 Comparison of stationarity and non-stationarity approaches for generation of synthetic streamflows

4.2.1 Megecha River

As discussed in the previous chapter, the model selected for generation of synthetic streamflows for river Megecha by the stationarity approach is the AR(1) model. This model was used to generate 38 samples each of length 12 years. The general statistical properties of the generated and historical samples are given in Table 4.1.

Confidence intervals, each one standard deviation away from the mean, were constructed for the mean, standard deviation and coefficient of skewness for checking the adequacy of the model. These confidence intervals are

(8.597 , 10.435) for the mean;

(12.665 , 17.655) for the standard deviation; and

(2.016 , 3.948) for the skewness coefficient.

Since each historical statistical characteristic lies in its respective confidence interval, we can conclude that the model is adequate.

In order to compare the performance of the stationarity and non-stationarity approaches for generation of synthetic monthly streamflow sequences for river Megecha, 38 samples of length equal to that of the historical sample, that is, of length 12 years, were also generated by the non-stationarity model of Thomas and Fiering. The main statistical properties of these samples and of the original sample are given in Table 4.2.

Table 4.1 : General statistical properties of the historical sample and of samples generated by the stationarity approach for river Megecha monthly flow sequence.

	mean	std.dev.	skewness
Historic sample	9.086	13.319	2.263
Generated samples			
1	10.374	18.761	3.588
2	8.276	13.070	2.611
3	8.198	13.553	4.168
4	10.514	17.267	2.890
5	9.607	13.913	2.328
6	8.654	12.676	2.122
7	8.381	12.471	2.339
8	9.504	15.123	2.426
9	9.521	17.633	4.909
10	9.444	16.391	3.579
11	8.674	13.697	2.836
12	8.083	11.960	2.997
13	9.130	13.180	1.988
14	9.589	14.478	3.263
15	9.476	13.562	2.159
16	9.081	13.046	1.992
17	9.155	12.498	2.084
18	8.085	11.912	2.453
19	10.681	17.931	3.416
20	10.677	16.614	2.753
21	7.804	11.697	2.305
22	9.817	15.613	3.139
23	10.042	17.856	3.828
24	9.214	13.574	2.579
25	10.042	16.785	3.132
26	11.197	22.840	7.233
27	9.849	17.273	3.131
28	11.736	20.779	3.500
29	9.220	14.240	2.740
30	9.430	14.170	2.720
31	9.040	13.580	2.220
32	10.460	14.320	2.020
33	10.250	17.080	2.900
34	9.010	14.290	2.340
35	9.930	16.140	3.430
36	9.420	15.120	2.940
37	11.020	16.450	2.630
38	9.040	14.530	3.640
Mean of 38 values	9.516	15.160	2.982

Table 4.2 : Main statistical properties of the historic sample and of samples generated by the non- stationarity approach for monthly flow series of Megecha river.

	mean	std.dev.	skewness
Historic sample	9.086	13.319	2.263
Generated samples			
1	10.559	13.225	2.102
2	8.798	11.525	2.915
3	9.555	14.593	2.810
4	9.466	13.505	2.213
5	10.200	13.854	2.242
6	9.721	15.517	3.838
7	10.073	12.036	1.756
8	9.166	12.158	2.217
9	9.013	13.959	3.487
10	9.481	12.918	2.724
11	9.017	12.739	2.243
12	9.857	14.469	2.308
13	9.524	12.566	1.949
14	8.959	13.654	2.580
15	10.102	14.013	2.287
16	10.006	14.785	2.956
17	10.300	14.861	2.188
18	9.268	13.594	2.637
19	8.362	11.054	2.074
20	10.588	14.574	2.289
21	10.260	13.769	2.054
22	9.575	14.838	2.735
23	10.144	13.049	1.705
24	10.481	16.869	3.836
25	8.675	11.303	1.868
26	9.697	15.306	2.516
27	9.156	14.496	3.481
28	7.917	10.573	2.165
29	8.280	12.150	2.190
30	9.040	13.430	2.120
31	8.500	13.440	2.380
32	7.960	13.840	2.710
33	8.080	13.180	2.700
34	8.750	15.040	3.500
35	8.240	13.380	2.850
36	9.520	14.630	2.240
37	8.510	12.430	2.040
38	8.590	13.490	2.260
Mean of 38 values	9.300	13.548	2.504

As in the case of stationarity approach, based on the samples generated by the non-stationarity model, confidence intervals for the mean, standard deviation and skewness were computed for checking the adequacy of the model.

The confidence intervals are

(8.527 , 10.073) for the mean ;

(12.222 , 14.874) for the standard deviation ; and

(1.959 , 3.049) for the coe. of skewness .

Each historical statistical characteristic falls inside its respective confidence interval. Hence, we can say that the model is adequate and that the main historical statistical characteristics are preserved.

Though both approaches provide adequate models for generation of synthetic monthly streamflow sequences for Megecha river, the average values of the standard deviation and coefficient of skewness of samples generated by the stationarity approach show a considerable upward bias as can be seen from Table 4.1. And from Fig.7 it can be observed that the correlogram of samples generated by the non-stationarity approach is closer to the historical correlogram than that of stationarily generated samples. A current practice in generating new samples of hydrologic time series is to preserve, exactly (which is less likely) or very closely, those estimates of the historic sample which have the smallest sampling variances such as the mean and standard deviation [Salas *et al*, 1980]. But in our case, the mean value of the standard deviations of samples generated by the stationarity approach is considerably larger than the historic standard deviation. Therefore, generating using the non-stationarity approach is somewhat better in this case.

4.2.2 Gudla River

As in the case of Megecha river, the model that is used to generate synthetic flows for Gudla river by the stationarity approach is the autoregressive model of order one (AR(1)). By using this model 38 samples each of length 33 years were generated. The general statistical properties of the original and generated samples are given in Table 4.3.

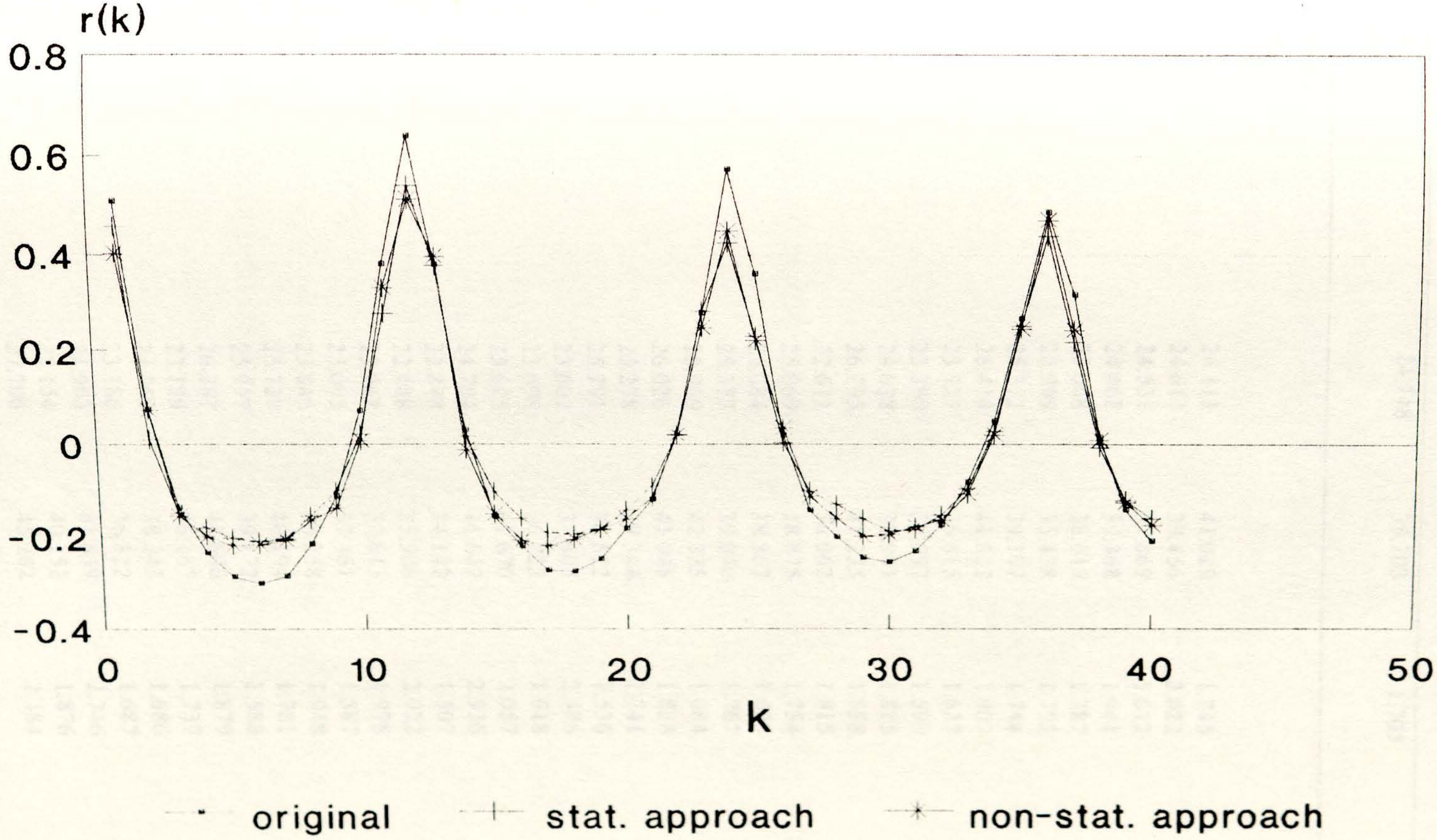

Figure 7 : Correlograms of the original
and generated samples of river Megecha.

Table 4.3 : General statistical properties of the historical sample and of samples generated by the stationarity approach for Gudla river.

	mean	std. dev.	skewness
Historic sample	24.748	38.203	1.705
Generated samples			
1	26.111	41.029	1.745
2	24.614	39.456	2.022
3	24.531	37.869	1.612
4	24.462	39.368	1.941
5	25.068	38.619	1.787
6	22.999	35.438	1.752
7	25.372	39.197	1.744
8	26.414	41.823	1.909
9	23.727	36.815	1.627
10	23.390	36.097	1.590
11	24.638	38.931	1.815
12	26.795	41.235	1.658
13	22.612	36.062	1.815
14	25.069	38.875	1.754
15	25.084	38.867	1.693
16	25.357	39.040	1.767
17	25.969	42.325	1.084
18	26.026	41.045	1.828
19	25.278	39.344	1.741
20	25.791	38.522	1.536
21	23.602	37.805	2.476
22	23.698	36.223	1.618
23	25.625	41.670	2.057
24	24.794	41.647	2.926
25	22.848	34.142	1.507
26	22.448	35.906	2.022
27	23.962	37.613	1.976
28	25.093	37.461	1.587
29	23.845	38.258	1.948
30	25.729	40.194	1.781
31	25.917	39.172	1.588
32	26.287	41.606	1.879
33	23.199	36.167	1.737
34	24.759	38.343	1.686
35	23.186	36.322	1.687
36	24.903	41.839	2.746
37	25.154	40.152	1.876
38	27.706	45.282	2.484
mean of 38 values	24.791	38.941	1.842

In order to check whether the model is adequate or not, confidence intervals for the mean, standard deviation and skewness are constructed by taking c=1. These confidence intervals are

(23.549 , 26.033) for the mean;

(36.599 , 41.283) for the standard deviation; and

(1.503 , 2.181) for the coefficient of skewness.

Since each historical statistical characteristic falls inside its respective confidence interval, we can conclude that the model is adequate and that the main historical characteristics are preserved.

Following the procedure developed in the previous chapter, the non-stationarity model of Thomas and Fiering was applied and 38 samples were generated. The main statistical properties of these samples are given in Table 4.4.

For checking the adequacy of the model, confidence intervals for the mean, standard deviation and skewness were constructed by taking c=1. The confidence intervals are

(24.615 , 27.315) for the mean ;

(37.219 , 41.429) for the standard deviation ; and

(1.551 , 2.183) for the coefficient of skewness .

Here each historical statistical characteristic falls inside its respective confidence interval. Therefore, we can conclude that the model is adequate. But as can be seen from Table 4.4, the average values of the mean and standard deviation of samples generated by the non-stationarity approach show a considerable upward bias. Furthermore, as can be seen from Fig.8, the correlogram of the samples generated by the stationarity approach is much closer to the historical correlogram than that of non-stationarily generated samples. Therefore, we can conclude that the stationarity approach is better in this case.

Table 4.4 : Main statistical properties of the historic sample and of samples generated by the non-stationarity approach for Gudla riverflow time series.

	mean	std. dev.	skewness
Historic sample	24.748	38.203	1.705
Generated samples			
1	24.294	38.722	1.701
2	26.569	39.931	1.877
3	25.967	40.110	1.056
4	22.659	34.330	1.965
5	25.718	38.181	1.794
6	27.987	41.330	1.694
7	27.414	40.941	2.114
8	27.212	40.851	2.079
9	25.998	37.449	1.758
10	24.892	35.574	1.596
11	28.496	42.880	1.933
12	27.842	41.433	1.844
13	25.835	36.883	1.716
14	25.679	40.488	2.433
15	25.405	36.613	1.738
16	29.601	43.796	1.865
17	25.748	38.346	1.872
18	27.300	40.671	1.884
19	27.289	41.345	1.832
20	26.826	39.851	1.954
21	25.880	36.482	1.773
22	25.527	38.165	1.856
23	26.477	38.685	1.681
24	25.123	37.353	1.825
25	25.223	37.503	1.711
26	26.299	38.848	1.858
27	26.367	39.221	1.808
28	25.468	41.880	3.328
29	25.076	38.328	1.709
30	25.181	38.577	1.744
31	26.747	41.329	1.667
32	23.620	37.406	1.842
33	25.480	41.114	1.877
34	26.061	41.696	2.084
35	23.988	36.951	1.652
36	24.798	40.061	1.976
37	24.859	40.238	1.963
38	25.750	40.742	1.902
Mean of 38 values	25.965	39.324	1.867

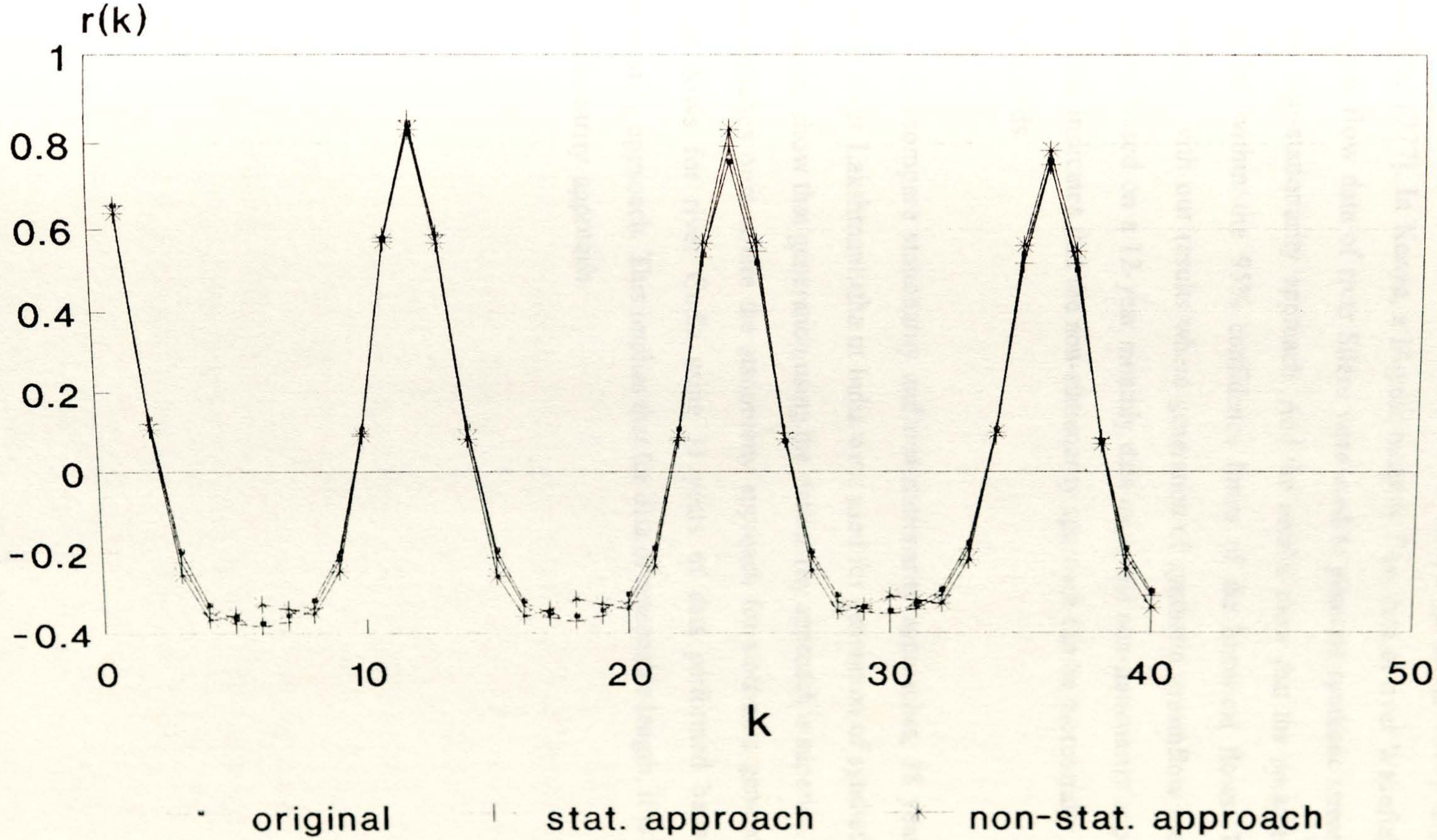

Figure 8 : Correlograms of the original and generated samples of river Gudla.

4.3 Comparison of results with similar studies

Similar studies were conducted in Kenya [Charnia, 1973] and India [Seth & Chandra, 1977]. In Kenya, a 14-year monthly flow data of river Wakefield and, in India, a 15-year flow data of river Sileru were used to generate synthetic streamflow sequences by the non-stationarity approach. And the results show that the parameters of synthetic flows lie within the 95% confidence limits of the historical flows. This is in direct agreement with our results where generation of synthetic streamflow sequences for river Megecha based on a 12-year monthly data using the non-stationarity approach performed well. This indicates that the non-stationarity approach can be successfully applied for short term records.

To compare stationarity and non-stationarity approaches, 38 years of streamflow data of river Lakshmantirtha in India were used for generation of synthetic sequences and the results show that generation using the stationarity approach is superior. Again this result also matches ours where the stationarity approach for stochastic generation of synthetic streamflows for river Gudla using 33 years of data performed better than the non-stationarity approach. This implies that for data of reasonable length, it is preferable to use the stationarity approach.

CHAPTER FIVE
CONCLUSIONS

In this study, it was attempted to describe the steps involved in modelling hydrologic time series, in general, and monthly streamflow sequences,in particular. An ever-present goal has been to compare stationarity and non-stationarity approaches for generation of synthetic flows so as to establish the performance of each model in planning and designing water resource systems. As a particular case, these two approaches were applied for modelling the streamflow data of rivers Megecha and Gudla. In both cases, to make the data normal and to avoid negative values of generated flows, a logarithmic transformation was applied to the original data. And then synthetic samples of length equal to the corresponding historical sample were generated and the statistics of the generated samples were compared with the corresponding historical statistics.

In case of river Megecha, both the stationarity and non-stationarity approaches provide adequate models for generation of synthetic streamflow sequences. But the average standard deviation and skewness of the samples generated by the stationarity approach show a considerable upward bias. Besides this, the correlogram of the samples generated by the non-stationarity approach is closer to the historical correlogram than that of stationarily generated samples. Therefore, we can conclude that the non-stationarity approach is superior to the stationarity approach in this case.

Also in the case of river Gudla, both approaches provide adequate models. But since the average values of the mean and standard deviation of samples generated by the non-stationarity approach show a considerable upward bias, we can say that those estimates of the historic sample with minimum sampling variance (i.e. the mean and standard deviation) are not closely preserved. The various statistical properties of samples generated by the stationarity approach, however, are in good agreement with their corresponding historical statistical properties and, therefore, we can conclude that this approach is superior.

Generally, it seems that for short term data, the performance of the non-stationarity approach is better than the stationarity approach in generating sequences of streamflow. But for data of considerable length, the stationarity approach give better results with less number of parameters as compared to the non-stationarity approach.

One question that may arise here is that when do we say a given hydrologic data is of short term or of considerable length. Unfortunately, there is no definite answer for this since what is long for one purpose may be short for another. Therefore, one has to take much care when selecting among the two approaches. But in situations where it is hard to say whether a given data is of short term or long term, it is advisable to use the stationarity approach because it involves a minimum number of parameters as compared to the non-stationarity approach. This is known as the principle of parsimony of parameters [Salas *et al*, 1980].

In this study, the models considered have been limited to those which are of most use and simple, i.e., the autoregressive (AR) models. But it should be stressed that if AR models fail to preserve the main historical statistical characteristics, then this defeats the basic objective of generating new samples and, therefore, one has to consider alternative models.

Another factor which is important to consider in practical applications of AR models is related to the modelling of skewed variables. Though it is desirable to work with series transformed to normal (because the estimation of parameters and goodness of fit tests are more efficient than for skewed variables), biases in the mean, standard deviation and serial correlations of the original series may occur when synthetic samples are generated with the AR model based on these transformed series. If such biases are considerable, then the analyst may avoid such transformation.

At last, it must be mentioned that the intrinsic value of the data generation method itself may be questioned. Clearly it cannot yield more information than is in the historical record; but longer sequences of synthetic flows, having statistical characteristics indistinguishable from that of the historical record, are used as inputs to a water resource system so as to judge its overall value. In this respect, a limitation of the data generation method which is not commonly accepted is that each simulation should be no longer than the historical record. If longer sequences are used, then unsubstantiated but sensible extrapolation is being invoked [Lawrance & Kottegoda, 1977].

APPENDIX A

```
program stationary ;
{ This program can be used to generate synthetic streamflow sequences of length as one
wishes by the stationarity approach. Before executing this program replace c1 by the
available number of years of record. }
const nyears = c1; k=40;
type  type1=array[1..nyears,1..12] of real;
      type2=array[-2..3000] of real;
      type3=array[1..1000] of real;
      type4=array[1..k] of real;
      type5=array[1..12] of real;
      type6=array[-2..10] of real;
var   y,z,genz,ycap : type1;
      rn,zcap : type2;
      zp,e : type3;
      r,err : type4;
      mean,stdev : type5;
      ar1,ar2,ar3 : type6;
      year,month,j,h,m,i,t,order,no,s,d,w :integer;
      sum,ss,sum1,sum2,sum3,ss1,meanz:real;
 ersum1,ersum2,ersum3,erss1,meaner,sigmae3,sm;real;
rand,v1,v2,q,q2,phicap11,phicap21,phicap22:real;
      phicap31,phicap32,phicap33,c2,sigmae2,sigmae1 :real;
      f1,f2,f3,f4,f5 : text;
      fname1,fname2,fname3,fname4,fname5 : string;
begin
   Writeln( 'Enter the file containing the data.') ;
   readln(fname1);
   assign(f1,fname1);
   reset(f1);
   for year:=1 to nyears do
     begin
        for month:=1 to 12 do
          begin
            readln(f1,y[year,month]);
          end;
     end;
     close(f1);
{ Perform a logarithmic transformation ( with lower boundary parameter c2 ) to make the
original data normal and to avoid negative values of generated streamflows. The lower
boundary parameter c2 may be obtained by trial and error.}
     Writeln( ' Enter the lower boundary parameter. ');
     Readln( c2 );
     for year:=1 to nyears do
       begin
         for month:=1 to 12 do
```

```
      begin
        y[year,month]:= ln( y[year,month] - c2 );
      end;
    end;
{ Calculate the monthly means and standard deviations}
    for month:=1 to 12 do
     begin
       sum:=0;
       ss:=0;
       for year :=1 to nyears do
        begin
          sum:=sum + y[year,month];
          ss:=ss + sqr(y[year,month]);
        end;
        mean[month]:=sum/nyears;
        stdev[month] := sqrt((ss -
         nyears*sqr(mean[month]))/(nyears - 1));
     end;
{ Standardize the data.}
   for year:=1 to nyears do
    begin
       for month:=1 to 12 do
          begin
  z[year,month]:=(y[year,month] - mean[month])/stdev[month];
          end;
    end;
{ Write z[1,1] as z[1], z[1,2] as z[2], z[1,3] as z[3],..., z[2,1] as z[13], z[2,2] as z[14], and
so on.}
    for year :=1 to nyears do
      begin
        m:= year - 1;
        for i:=(12*m)+1  to 12*(m+1) do
          begin
            month:=i mod 12;
            if month=0 then month:=12;
            zp[i]:=z[year,month];
          end;
      end;
{ Compute the correlogram of the standardized series.}
   sum1:=0;
   ss1:=0;
   for i:=1 to 12*nyears do
      begin
        sum1:=sum1 + zp[i];
        ss1:=ss1 + sqr(zp[i]);
      end;
      meanz:=sum1/12*nyears;
      sum2:=ss1 - 12*nyears *sqr(meanz);
```

```
for t:=1 to k do
     begin
       sum3:=0;
        for i:=1 to (12*nyears - t) do
         begin
     sum3:=sum3 + (zp[i] - meanz)*(zp[t+i] - meanz);
          end;
          r[t]:=sum3/sum2;
     end;
{ Compute the estimates of the AR coefficients and residual variances and expected
correlograms for AR(1),AR(2) and AR(3) models.}
{ For the AR(1) model }
     phicap11:=r[1];
     for h:=1 to 10 do
     begin
       ar1[h]:= exp(h*ln(phicap11));
     end;      sigmae1:=sqrt((12*nyears*(1-sqr(phicap11)))/(12*nyears-1));
{ For the AR(2) model }
     phicap21:=(r[1]*(1-r[2]))/(1-sqr(r[1]));
     phicap22:=(r[2]-sqr(r[1]))/(1-sqr(r[1]));
     for h:=1 to 10 do
     begin
        ar2[0]:=1;
        ar2[-1]:=r[1];
        ar2[h]:=phicap21*ar2[h-1] + phicap22*ar2[h-2];
     end;
     w:=12*nyears;
sigmae2:=(w*(1+phicap22)*(sqr(1-phicap22)  - sqr(phicap21)))/((w-2)*(1-phicap22));
       sigmae2:=sqrt(sigmae2);
{ For the AR(3) model. Here a simplified form for the estimates of the AR coefficients is
not available; so solve the ACF to get them. }
       Writeln(' Enter the three AR coefficients.');
       Readln(phicap31,phicap32,phicap33);
       for h:=1 to 10 do
       begin
         ar3[0]:=1;
         ar3[-1]:= r[1];
         ar3[-2]:= r[2];
         ar3[h]:= phicap31*ar3[h-1] + phicap32*ar3[h-2]  + phicap33*ar3[h-3];
       end;
       sm:= phicap31*r[1] + phicap32*r[2] +  phicap33*r[3];
       sigmae3:=sqrt((w*(1-sm))/(w-3));
{ Write the correlogram of the standardized series and expected correlograms for AR(1),
AR(2) and AR(3) models to the file. Plot the correlograms and choose the best model.}
      Writeln( ' Enter the file to store the correlogram  of the standardized series and
expected   correlograms for the three models.');
      Readln( fname2);
      assign(f2,fname2);
```

```
      Rewrite(f2);
   Writeln(f2,'Correlogram of the standardized series');
      For t:=1 to k do
         begin
            Writeln( f2, t ,'  ',r[t]:7:4 );
         end;
Writeln(f2,'AR(1) model',' ','AR(2) model',' ','AR(3) model');
      For h:=1 to 10 do
      begin
 Writeln(f2,' ',ar1[h]:7:4,' ',ar2[h]:7:4,' ',ar3[h]:7:4 );
      end;
      close(f2);
{ Calculate the error terms.}
      writeln('Enter the order of the selected model- 1, 2 or 3? ');
      readln(order);
      if order=1 then
      begin
         for i:= order+1 to 12*nyears do
           begin
             e[i]:= zp[i] - phicap11*zp[i-1]
           end;
      end;
      if order=2 then
      begin
         for i:= order+1 to 12*nyears do
           begin
             e[i]:= zp[i] - phicap21*zp[i-1] - phicap22*zp[i-2];
           end;
      end;
      if order=3 then
      begin
         for i:= order+1 to 12*nyears do
           begin
             e[i]:=zp[i] - phicap31*zp[i-1] - phicap32*zp[i-2] - phicap33*zp[i-3];
           end;
      end;
{ Write the error terms to the file and test for normality. If the hypothesis of normality is
rejected, then go back and try another normalization transformation on the original data.}
      writeln('Enter the file to store the error terms');
      readln(fname3);
      assign(f3,fname3);
      rewrite(f3);
      For i:= order+1 to 12*nyears do
        begin
          writeln(f3,e[i]:7:4);
        end;
      close(f3);
{ Compute the correlogram of the residuals (error terms). }
```

```
ersum1:=0;
erss1:=0;
for i:= order+1 to 12*nyears do
  begin
    ersum1:=ersum1 + e[i];
    erss1:=erss1 + sqr(e[i]);
  end;
    meaner:=ersum1/(12*nyears - order);
ersum2:=erss1 - (12*nyears - order)*sqr(meaner);
    for t:=1 to k do
      begin
        ersum3:=0;
        for i:= order+1 to (12*nyears - t) do
          begin
ersum3:=ersum3 + (e[i] - meaner)*(e[t+i] - meaner);
          end;
          err[t]:=ersum3/ersum2;
      end;
{ Write the error correlogram to the file and test for independence. If the hypothesis of
uncorrelated residuals is rejected, then go back and change the order of the model. }
  Writeln(' Enter the file to store the error correlogram. ');
      Readln(fname4);
      assign(f4,fname4);
      Rewrite(f4);
      for t:=1 to k do
        begin
          writeln(f4,t,' ', err[t]:7:4);
        end;
      close(f4);
{ Generate N(0,1) random numbers.}
      writeln('Enter the number of N(0,1) random  numbers to  generate.');
      readln(no);
      randomize;
      rand:=random;
      s:=1;i:=1;
      repeat
        v1:=2*random - 1;
        v2:=2*random - 1 ;
        q2:=sqr(v1) + sqr(v2);
        if q2 < 1 then
          begin
            q:=sqrt((-2*ln(q2))/q2);
            rn[i]:=v1*q;
            rn[i+1]:=v2*q;
            i:=i+2;
          end;
        s:=s+1;
      until (s > no/2) and (i > no);
```

```
{ Generate standardized flows.}
        zcap[-2]:=0.0;
        zcap[-1]:=0.0;
        zcap[0]:=0.0;
        for i:=1 to no do
          begin
            If order=1 then
            begin
      zcap[i]:=phicap11*zcap[i-1] + sigmae1*rn[i];
            end;
            If order=2 then
            begin
               zcap[i]:=phicap21*zcap[i-1] + phicap22*zcap[i-2] + sigmae2*rn[i];
            end;
           If order=3 then
           begin
  zcap[i]:=phicap31*zcap[i-1]+phicap32*zcap[i-2]+phicap33*zcap[i-3]+sigmae3*rn[i];
                end;
          end;
{ Re-initialize the generated standardized flows.}
      for i:=1 to (no - 50) do
        begin
          zcap[i]:=zcap[i+50];
        end;
{ Write zcap[1] as zcap[1,1], zcap[2] as zcap[1,2],..., zcap[13] as zcap[2,1], zcap[14] as
zcap[2,2],..., and then generate future flows.}
   writeln('Enter the file to store the generated flows.');
        readln(fname5);
        assign(f5,fname5);
        rewrite(f5);
 Writeln('Enter the number of years of data to generate ');
        Readln(d);
        for year:=1 to d do
          begin
            m:=year-1;
            for i:= (12*m)+1 to 12*(m+1) do
            begin
              month:= i mod 12;
              if month=0 then month:=12;
              genz[year,month]:=zcap[i];
ycap[year,month]:=mean[month] + stdev[month]*genz[year,month];
          ycap[year,month]:= exp(ycap[year,month]) + c2;
     writeln(f5,ycap[year,month]:7:3);
            end;
          end;
        close(f5);
        end.
```

APPENDIX B

```
program non_stationary ;
{ This program can be used to generate synthetic streamflow
  sequences of length as one wishes by the non-stationary model
  of Thomas and Fiering. Before executing this program, replace
  c1 by the available number of  years of record. }
const nyears = c1;
type type1=array[1..nyears,1..13] of real;
     type2=array[0..4500] of real;
     type3=array[1..13] of real;
var y, ycap : type1;
    z, Q : type2;
    B,r,Qbar,S : type3;
    year,j,no,i,d,t,m,n : integer;
    sum,ss,rand,v1,v2,q3,q2,c2,w,sumt,sst,meant,sdt : real;
    sumreg,ssreg1,ssreg2,c3 :real;
    f1,f2 : text;
    fname1,fname2 : string;
begin
  Writeln( 'Enter the file containing the data.') ;
  readln(fname1);
  assign(f1,fname1);
  reset(f1);
  for year:=1 to nyears do
     begin
        for j:=1 to 12 do
          begin
            readln(f1,y[year,j]);
          end;
        end;
  close(f1);
{ Perform a logarithmic transformation (with a lower boundary parameter c2) on the
original series to make it normal (i.e. to reduce the skewness closer to zero). }
     Writeln( ' Enter the lower boundary parameter. ');
     Readln( c2 );
     for year:=1 to nyears do
       begin
         for j:=1 to 12 do
           begin
              y[year,j]:= ln( y[year,j] - c2 );
           end;
       end;
{ Calculate the overall mean and standard deviation of the
  transformed data. }
       sumt:=0; sst:=0;
       For year := 1 to nyears do
        begin
```

```
          For j := 1 to 12 do
            begin
                sumt:=sumt + y[year,j];
                sst:=sst + sqr(y[year,j]);
              end;
        end;
        w:= 12*nyears;
        meant:= sumt/w;
        sdt:=sqrt((sst - w*sqr(meant))/(w - 1));
  { Standardize the normalized data }
       for year:=1 to nyears do
       begin
         for j:=1 to 12 do
         begin
           y[year,j]:=(y[year,j] - meant)/sdt;
         end;
       end;
  { Calculate the monthly means and monthly standard deviations of the standardized normalized
  data. }
       for j:=1 to 12 do
        begin
         sum:=0;
         ss:=0;
         for year :=1 to nyears do
          begin
          sum:=sum + y[year,j];
            ss:=ss + sqr(y[year,j]);
          end;
          Qbar[j]:=sum/nyears;
     S[j] := sqrt((ss - nyears*sqr(Qbar[j]))/(nyears - 1));
       end;
  { Compute the monthly regression coefficients }
      For j:= 1 to 12 do
        begin
          sumreg:=0;ssreg1:=0;ssreg2:=0;
          For year:=1 to nyears do
            begin
          y[year,13]:=y[year,1];Qbar[13]:=Qbar[1];
          sumreg:=sumreg+(y[year,j]-Qbar[j])*(y[year,j+1]-Qbar[j+1]);
            ssreg1:=ssreg1+sqr(y[year,j] - Qbar[j]);
    ssreg2:=ssreg2 + sqr(y[year,j+1] - Qbar[j+1]);
            end;
          r[j]:=sumreg/sqrt( ssreg1*ssreg2);
        end;
  { Compute the monthly correlation coefficients. }
       For j:=1 to 12 do
         begin
           S[13]:=S[1];
           B[j]:=( r[j]*S[j+1] )/S[j];
         end;
  { Generate N(0,1) random numbers.}
```

```
randomize;
rand:=random;
no:=3000;
t:=1;i:=1;
repeat
  v1:=2*random-1;
  v2:=2*random-1 ;
  q2:=sqr(v1) + sqr(v2);
  if q2 < 1 then
    begin
      q3:=sqrt((-2*ln(q2))/q2);
      Z[i]:=v1*q3;
      Z[i+1]:=v2*q3;
      i:=i+2;
    end;
  t:=t+1;
until (t > no/2) and (i > no);
{ Generate synthetic flows }
Writeln(' Enter the streamflow measurement corresponding to the last month of historical data.');
Readln(c3);
for i:=0 to no do
begin
  Q[0]:= ( ln(c3 - c2) - meant )/sdt;
  j:= (i+1) mod 12;
  if j=0 then j:=12;
Q[i+1]:=Qbar[j+1] + B[j]*(Q[i]-Qbar[j])  + Z[i]*S[j+1]*sqrt(1-sqr(r[j]));
end;
writeln('Enter the number of years of data to generate ');
readln(n);
for year:=1 to n do
begin
  m:=year - 1;
for i:= 12*m  to  12*(m+1) -1 do
begin
  j:=(i+1) mod 12;
  if j=0 then j:=12;
ycap[year,j]:= exp( Q[i+1]*sdt + meant ) + c2 ;
end; end;
{ Write the generated synthetic streamflows to the file. }
Writeln( 'Enter the file to store the generated flows.');
Readln(fname2);
Assign(f2,fname2);
Rewrite(f2);
For year:=1 to n do
  begin
    For j:=1 to 12 do
      begin
        Writeln( f2,ycap[year,j]:7:3);
      end;
  end;
close(f2); end.
```

APPENDIX C

Monthly streamflow data of River Megecha
(1977 - 1988)

year	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1977	1.000	1.660	1.160	2.950	1.970	2.780	37.060	57.480	15.790	17.340	11.110	1.810
1978	1.079	0.878	1.338	0.711	1.135	4.472	49.360	57.910	16.440	10.330	2.218	1.856
1979	2.291	3.970	2.014	3.291	3.130	1.429	18.428	33.890	14.310	7.398	1.217	0.828
1980	0.692	0.612	0.694	0.956	0.910	0.937	11.839	29.070	11.460	2.793	0.932	0.712
1981	0.660	0.750	1.270	7.490	3.650	0.470	11.970	40.190	16.960	11.280	1.320	0.720
1982	0.700	0.690	0.800	1.290	1.610	3.400	11.940	41.250	13.320	13.900	2.140	1.160
1983	0.763	2.420	5.940	15.200	13.250	8.070	7.260	69.880	33.930	10.420	1.840	0.954
1984	0.200	0.450	0.500	0.330	1.430	7.220	21.310	18.110	17.030	2.570	1.190	0.990
1985	0.670	0.600	0.630	1.390	4.640	3.910	12.070	41.130	21.320	4.040	1.060	0.710
1986	0.434	0.463	0.977	0.280	5.070	17.370	30.190	22.280	19.540	3.720	1.130	0.690
1987	0.442	0.401	3.970	15.470	3.150	14.560	11.200	21.180	16.830	3.230	1.100	0.690
1988	0.579	0.730	0.691	0.587	1.236	1.605	25.326	52.320	22.420	20.630	2.818	1.175

APPENDIX D

Monthly streamflow data of River Gudla
(1960 - 1992)

year	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1960	2.930	1.360	0.990	0.610	1.990	2.450	74.370	83.240	54.710	12.030	3.020	1.590
1961	0.780	0.660	0.640	0.940	0.510	1.380	83.700	154.000	87.210	34.120	5.540	5.260
1962	1.330	0.490	0.680	0.390	1.010	14.710	90.780	117.270	89.320	25.300	4.260	2.220
1963	1.240	0.600	0.510	0.850	3.670	9.140	116.510	140.430	72.090	19.070	18.700	25.360
1964	3.460	1.540	0.780	0.780	3.460	25.120	154.900	130.600	80.400	51.600	7.440	4.730
1965	1.660	1.110	0.900	0.980	0.620	3.010	34.270	86.120	36.790	26.170	9.110	6.890
1966	2.320	1.500	1.240	0.861	1.050	8.360	54.900	103.200	36.900	7.040	3.910	2.050
1967	1.180	0.740	1.190	0.820	1.530	4.030	75.910	107.180	70.360	29.150	6.230	3.440
1968	1.390	0.960	0.780	0.760	1.640	16.440	75.460	89.380	48.090	25.280	3.940	2.380
1969	1.950	1.390	2.750	2.210	2.380	4.090	59.800	140.300	44.580	6.880	4.480	1.760
1970	0.920	0.580	0.640	0.680	0.670	3.090	57.300	114.100	56.150	25.170	5.150	2.020
1971	1.190	0.660	0.650	0.460	1.410	10.700	63.820	82.970	50.430	12.490	5.520	2.040
1972	1.390	0.760	0.590	0.570	0.650	2.860	39.410	54.160	38.060	8.600	11.450	2.740
1973	1.180	0.575	0.455	0.372	2.520	15.900	69.280	105.060	56.950	28.520	5.960	2.220
1974	1.110	0.627	0.755	0.363	3.278	14.830	76.700	93.400	46.720	12.500	3.140	1.390
1975	0.810	0.920	0.557	0.433	0.440	13.800	84.800	132.900	77.770	12.420	4.340	2.300
1976	0.910	0.702	0.639	0.318	2.410	12.084	70.340	88.200	76.740	11.110	8.420	2.860
1977	0.946	0.610	0.773	0.390	2.930	16.450	114.490	99.320	62.560	54.900	12.590	3.790
1978	1.705	0.754	0.770	0.586	1.377	9.940	64.720	78.600	52.420	24.820	5.910	3.740
1979	1.720	0.750	0.665	0.452	1.708	27.100	60.910	84.980	71.980	12.320	4.080	1.880
1980	0.860	0.780	0.589	0.776	0.956	3.530	84.890	101.300	49.460	14.680	4.070	2.000
1981	0.892	0.518	0.526	0.453	2.680	8.090	76.810	92.880	55.840	13.190	3.980	1.270
1982	0.900	0.490	0.520	0.360	1.010	1.620	40.940	92.140	85.670	40.580	4.580	2.200
1983	0.920	0.640	0.680	0.420	0.740	2.570	36.660	86.980	39.680	11.860	5.700	2.200
1984	0.920	0.460	0.400	0.310	0.850	17.480	20.000	105.920	56.880	8.900	2.680	2.240
1985	0.910	0.470	0.440	0.620	0.920	5.240	96.080	126.840	83.200	12.940	2.830	1.600
1986	0.696	0.340	0.532	0.547	0.494	14.160	113.900	69.520	52.870	16.520	2.920	1.400
1987	0.670	0.410	0.728	0.460	12.000	51.350	101.200	108.600	48.250	12.130	4.010	2.640
1988	1.140	0.584	0.391	0.266	0.517	11.560	172.900	147.000	46.220	23.540	5.340	1.610
1989	0.577	0.250	0.463	0.558	2.430	29.830	71.310	94.990	47.610	13.550	3.170	3.080
1990	1.520	0.716	0.492	0.428	0.628	3.630	71.446	69.940	45.730	10.660	1.700	0.732
1991	0.395	0.174	0.217	0.247	0.657	46.690	164.500	143.500	80.450	17.810	3.430	1.380
1992	0.539	0.310	0.672	10.120	3.000	46.870	123.900	159.300	106.200	52.010	2.630	1.980

REFERENCES

1.Akaike, H., 1974, A New look at the Statistical ModelIdentification, IEEE Trans. on Automatic Control, AS-19, 6, pp.716-723.

2.Anderson,R.L., 1941, Distribution of the serial CorrelationCoefficients, Annals of Math. Statistics, Vol.8, no.1, pp.1-13.

3.Box, G.E.P. and Jenkins, G., 1970, Time Series Analysis :Forecasting and Control, San Francisco, Holden-Day.

4.Charnia, S.H.,1973, Extension of Runoff Records for Small Catchments in Semi-Arid Regions, In Design of Water Resources Projects with Inadequate Data, Proceedingsof the Madrid Symposium, Vol.1, pp.377-391.

5.Chatfield, C., 1975, The Analysis of Time Series : Theory and Practice, Chapman and Hall, London.

6.Cox, D.R. and Miller, H.D., 1965, The Theory of Stochastic Processes, Methuen, London.

7.Hurst,H.E., 1951, Long Term Storage Capacity of Reservoirs, Trans. Amer. Soc. Civil Engrs., Vol.116, pp.770-798.

8.Kendall, M.G. and Stuart, H., 1968, The Advanced Theory of Statistics, Vol.3, Design and Analysis and Time series, Second Edition, Hafner, New York.

9.Lawrance, A.J. and Kottegoda, N.T., 1977, Stochastic Modelling of Riverflow Time Series, Journal of the Royal Statistical Society, Series A, Vol. 140, Part 1, pp. 1-47.

10.Morgan, B.J.T., 1984, Elements of Simulation, Chapman and Hall,U.K.

11.Mutreja,K.N., 1986, Applied Hydrology, Tata McGraw-Hill, New Delhi.

12.Parzen, E., 1965, Stochastic Processes, Holden-Day, San Francisco, California.

13.Salas, J.D., Delleur, J.W., Yevjevich, V. and Lane,W.L., 1980,Applied Modelling of Hydrologic Time Series, Water Resources Publications, Littleton, Colorado, USA.

14.Seth, S.M. and Chandra, S., 1977, Stochastic Modelling of Monthly Streamflow Sequences, In Stochastic Processes inWater Resources Engineering -Proceedings of the Second International IAHR Symposiumon Stochastic Hydraulics, Water Resources Publications, Colorado, pp.18/1-21.

15.Snedecor, G.W. and Cochran, W.G., 1967, Statistical Methods, The Iowa State University Press, Iowa, USA.

16.Svanidze, H.G., 1972, Mathematical Modelling of Hydrologic Series, Water resources Publications, Fort collins, Colorado

17.Thomas,H.A. and Fiering,M.B., 1962, Mathematical Synthesis of Streamflow sequences for the Analysis of River Basins by Simulation, In Design of Water Resources Systems, pp.459-493

18.Yevjevich, V., 1972a, Probability and Statistics in Hydrology, Water Resources Publications, Fort Collins, Colorado,USA.

19.Yevjevich, V., 1972b, Stochastic Processes in Hydrology, Water Resources Publications, Fort Collins, Colorado, USA.

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been submitted for a degree to any other university. All sources of materials used for the thesis have been duly acknowledged.

Name. Emmanuel Gabreyohannes

Signature.

Date of submission. June 1994

This thesis has been submitted for examination with my approval as university advisor.

Name. Dr. Abebe Tessera

Signature.