UNSW SYDNEY

The Rescaled adjusted range in hydrologic design. June 1980.

Author:
Lee, M. O.

Publication details:
Report No. UNSW Water Research Laboratory Report No. 157
0858242974 (ISBN)

Publication Date:
1980

DOI:
https://doi.org/10.4225/53/579843bccf9e2

License:
https://creativecommons.org/licenses/by-nc-nd/3.0/au/
Link to license to see what you are allowed to do with this resource.

Downloaded from http://hdl.handle.net/1959.4/36167 in https://unsworks.unsw.edu.au on 2024-04-25

The quality of this digital copy is an accurate reproduction of the original print copy

628.105
S
Sect 1

THE UNIVERSITY OF NEW SOUTH WALES
water research laboratory
Manly Vale N.S.W. Australia

REFERENCE ONLY
Not For Loan

RESEARCH REPORT NO. 157

THE RESCALED ADJUSTED RANGE IN HYDROLOGIC DESIGN

by

M.O. Lee

June, 1980

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF CIVIL ENGINEERING

THE RESCALED ADJUSTED RANGE IN HYDROLOGIC DESIGN

by

M.O. LEE

https://doi.org/10.4225/53/579843bccf9e2

Water Research Laboratory

Report No.15

June, 1980

UNIVERSITY OF NEW SOUTH WALES
WATER REFERENCE
LIBRARY
WATER RESEARCH FOUNDATION

BIBLIOGRAPHIC DATA SHEET	1. REPORT No. 157	2. I.S.B.N. 0-85824-297-4

3. TITLE AND SUBTITLE
The Rescaled Adjusted Range in Hydrologic Design

4. REPORT DATE
June 1980

5. AUTHOR(S)
M. O. Lee

6. SPONSORING ORGANIZATION
School of Civil Engineering, University of New South Wales.

7. SUPPLEMENTARY NOTES

8. ABSTRACT

The findings of H.E. Hurst 30 years ago have had a profound effect on developments in data synthesis and reservoir analysis over the intervening period. His principal finding, which has become known as the Hurst effect, is now recognised as a feature of time series which should be considered when modelling such series. Much research effort has been spent on development and testing of complicated models which do attempt to model the Hurst effect, often at the expense of more significant characteristics. This work has been carried out in many cases because of a presumption that the simple data generation models do not reproduce Hurst's findings about natural series. This report identifies some areas of possible misunderstanding of the Hurst phenomenon and results are given to support two underlying hypotheses. The first is that the rescaled adjusted range statistic associated with Hurst's work is capable of effectively measuring characteristics of practical importance. The second relates to the finding that, on the basis of testing, simple generation models are suitable for many natural time series.

9. DISTRIBUTION STATEMENT

10. KEY WORDS
Data Generation, Hurst Effect, Stochastic Models, Storage Design, Hydrology, Rescaled Adjusted Range.

11. CLASSIFICATION	12. NUMBER OF PAGES 144	13. PRICE $20
Unclassified		

SUMMARY

This study examines the usefulness of the rescaled adjusted range statistic as an aid in the analysis of hydrologic time series and the construction of synthetic data generation models.

A definition is given of the rescaled adjusted range and other related statistics in terms of the 'residual mass curve' approach to reservoir storage design. A more rigorous definition is presented together with analytical results available in the literature for range statistics in theoretical processes.

The pioneering work by Hurst (1951) in the use of the rescaled adjusted range in the analysis of hydrologic time series is reviewed. Perceptions of the 'Hurst Phenomenon' are discussed and various estimators of the Hurst exponent that have been proposed in the literature are described. It is pointed out that the erroneous comparison of sample estimates of the expected value of the Hurst exponent with its theoretical asymptotic value pervades much of the literature on the 'Hurst Phenomenon'.

Sampling experiments with computer generated data sequences are carried out and show large sampling variation in the Hurst coefficient and the rescaled adjusted range. Difficulties in determining the underlying Hurst exponent from sample series are illustrated by examination of synthetic and real data series.

The structure of 'short-memory' autoregressive and moving average stochastic process models are examined in detail and model identification and fitting procedures discussed. Useful properties of the rescaled adjusted range in these types of theoretical processes are identified and an analogy is drawn between the rescaled adjusted range function as a function of sub-series length and the autocorrelation function as a function of lag interval.

The comparison of observed and theoretical rescaled adjusted range functions is proposed as a design method for examining the adequacy of a stochastic model for reservoir storage design purposes in particular. Comparisons are made between observed and theoretical functions for many Australian and overseas hydrologic data series and series generated by appropriately identified models. In most cases the observed functions fall within an approximate 95% confidence region surrounding the theoretical function. Such comparisons discriminate between various model structures proposed for a given data series.

In conclusion it is pointed out that a stochastic model should produce series showing realistic values of the rescaled adjusted range as a pre-requisite for application in the reservoir storage design process. Behaviour of the Hurst exponent which is not reproducible by 'short-memory' models in some cases may be an indication of a 'Hurst Phenomenon' in very long data series. Consideration of such an effect may not have much relevance to hydrological design.

ACKNOWLEDGEMENTS

The author would like to express his gratitude to the following staff of the School of Civil Engineering at the University of New South Wales, Assoc. Professor D.H. Pilgrim, Dr I. Cordery, Mr D.T. Howell and Mr D.G. Doran for their assistance and encouragement in this work.

The author is particularly indebted to Mr D.G. Doran and Dr I. Cordery for many stimulating discussions and helpful suggestions.

Details of computer programs written for this study are given in the appendix to this report. Considerable additional analysis was carried out, however, using a time-series analysis program package developed within the Department of Water Engineering of the School of Civil Engineering at the University of New South Wales. This program package was of great assistance in the study.

The onerous task of typing the drafts and final copy of this report fell to the author's wife, Julie, to whom he is deeply grateful.

TABLE OF CONTENTS

LIST OF SYMBOLS

a_t	Uncorrelated random variate sampled at time t
C ( )	Covariance of argument ( ) - (Sen - 1977a)
E [ ]	Expected value of argument
FH(i)	Estimator of the Hurst exponent (Wallis and Matalas - 1970)
GH(i)	Estimator of the Hurst exponent (Wallis and Matalas - 1970)
H	Estimator of the Hurst exponent (Mandelbrot and Wallis - 1969d)
h	Hurst exponent (Hurst - 1951)
h(n)	Local Hurst exponent (Anis and Lloyd - 1976)
i	Time interval number
K	Hurst coefficient (Hurst - 1951)
k	Lag interval
M_n	Adjusted surplus (Anis and Lloyd - 1976)
m_n	Adjusted deficit (Anis and Lloyd - 1976)
N,n	Series or sub-series length
p	Number of autoregressive terms in stochastic process or model
Q_i, Q_t	Series of annual river flows, general time series
q_t	Values of a general time series (standardised values in Chapters 9 and 10)
q_t^*	Values of a general time series
q	Number of moving average terms in a stochastic process or model
R_n^*	Adjusted range
R_n^{**}	Rescaled adjusted range
r_k	Estimates of autocorrelation coefficient at lag k
S_r	Partial sum (Anis and Lloyd - 1976)
nS_r^*	Adjusted partial sum (Anis and Lloyd - 1976)

nS_r^{**}	Rescaled adjusted partial sum (Anis and Lloyd - 1976)
$S(\phi_i)$	Sum of squares function leading to maximum likelihood estimate of ϕ_i (Nelson - 1973)
SH	Estimator of the Hurst exponent (Siddiqui - 1976)
s, s_z, s_n	Sample standard deviation of time series values
s_a^2	Variance of random independent variate a_t
s_z^2	Variance of time-series values z_t
$s(q_t^*)$	Standard deviation of time series values q_t^*
$s(R_n^{**})$	Standard deviation of R_n^{**} values
t	Time interval number
$V(\ )$	Variance of argument ( ) - (Sen - 1977a)
v_t	Identically distributed standardised independent random variate sampled at time t
$\{X_t\}$	Time series of values x_t
$\bar{x}_n$	Sample mean of time series values x_t
x_i, x_t	Time series values
YH	Estimator of the Hurst exponent (Gomide - 1975)
Z_t	Time series values formed of deviations from the process mean
$\bar{z}$	Mean of time series values z_t
γ_v	Skewness of independent random variate v_t
γ_z	Skewness of time series of values z_t
γ_k	Autocovariance at lag k
θ_i	Moving average parameter for stochastic process or model
ϕ_i	Autoregressive parameter for stochastic process or model
ϕ_{kk}	Partial autocorrelation coefficient at lag k

ρ_k	Autocorrelation coefficient at lag k
σ	Population standard deviation of normal variate (Sen - 1977a)
Γ	Gamma function

CHAPTER 1: AN INTRODUCTION

1.1 Introduction

The engineering hydrologist is a practitioner in the art of using the past to gather information about the future. He is called on to forecast future risks of floods and droughts and to construct likely future sequences of rainfall and streamflow. The important business of designing and operating systems for the exploitation, and hopefully the protection, of the accessible parts of the hydrological cycle depends for its success on such information. The often competing pressures on our water resources continue to grow and therefore the challenge of adequate resource management demands improvement in the engineering hydrologist's ability to provide information about the future.

The only rational way to investigate the future is to learn as much as possible of what the past has to teach and put that information to careful use. The main difficulties are that the past historical record never seems long enough for the task and that the future will be quite different from the past in any case. A saving grace is the apparent order underlying the natural phenomena with which the hydrologist is involved.

This study relates to the attempts by many investigators to analyse the underlying order in series of observations of rainfall, streamflow and other geophysical phenomena and also the use of such information to construct plausible future sequences of events.

1.2 Time-series and Stochastic Processes

In this study attention will be focussed on time-series of observed phenomena particularly streamflow. The instantaneous rate of flow in a stream is a continuous variable with time, at least while flow is occurring. To facilitate analysis, the continuous streamflow record is broken up into equally spaced segments of time such as a day, month or year. The total volume of flow over each segment of time is regarded as a discrete quantity or event and the succession of these events at equally spaced time intervals is regarded as a time-series of streamflow. Monthly and annual time-series of streamflow, rainfall and other geophysical phenomena are of interest in this study.

Natural phenomena such as rainfall and streamflow show a great deal of variation and the extremes of drought and flood are a common experience, particularly in the Australian situation. Within this variability however, the concept of an average value of rainfall or streamflow is commonly accepted. Experience tells us that rainfall and streamflow appear to fluctuate about an average level which does not seem to change greatly in the long term.

To some extent it is convenient to think of the rainfall or streamflow time series as the output from some unknown mechanism or process. Time-series which appear to have a constant mean level and a constant average variability about that mean level over a long period of time are often referred to as being the product of a stationary process of more strictly a weakly stationary process.

Non-stationarity may take on different forms such as trend or periodicity. The natural phenomenon being observed may be undergoing systematic change as the result of natural processes such as the gradual silting up of a river or man's activities as in the case

of higher flood flows due to increasing urbanisation of a catchment. Observations affected in this way will contain underlying trends. Natural phenomena may have an underlying systematic variation which is repeated each period of a day, lunar month or year. Such periodicity is usually evident in monthly streamflow series where mean values of streamflow, for example, may be lower in summer months than winter months. Such a time series of monthly streamflows would not be regarded as the product of a stationary process. However the series of all the January flows in particular, taken as an entity, might be considered to share a common mean and variance and therefore might be regarded as the product of a stationary process.

Annual streamflow and rainfall series are often regarded as stationary or weakly stationary as they are not subject to any obvious periodicities. Whether or not such an assumption is valid, or whether random shifts in climate occur or hidden periodicities exist, is a puzzle which has occupied the attention of many investigators. This question has obvious implications for hydrological design and will be touched on later in this report.

1.3 Reservoir Storage Design

Because of the variability of rainfall and streamflow the problem of storage is of great interest. Intuitively it would seem that the more variable the rainfall or streamflow 'process', the greater the amount of storage required to supply some fixed water demand at a given reliability. It would seem also that for the same rainfall or streamflow process and for a given desired reliability, the higher the required demand the greater the amount of storage required.

A frequent task for the engineering hydrologist is the

estimation of the amount of storage required at a stream site to provide a particular level of water supply. One approach to this problem is often referred to as the critical period approach. In this method the reservoir is assumed to have operated throughout the time-span of the available historical record. The storage selected is such that satisfactory levels of supply would have been maintained over the historical period with the reservoir just reaching emptiness once in the period. The reservoir system therefore would survive the critical period in the historical record. Additional storage is sometimes added as a factor of safety. The rationale for this traditional approach is not that the historical record will be exactly reproduced in future, but that the approach provides an acceptable, if somewhat arbitrary, basis for design which makes use of some of the information available in the historical record.

The drawback of the critical period approach to storage design is that very little indication is gained as to the risk of failure to deliver the desired water supply. This drawback is particularly significant for economic decision making, as estimates of the risk to supply are essential for optimising the storage size with respect to irrigation benefits for example. It is this coupling of hydrologic analysis to economic decision making that has encouraged the search for different approaches to reservoir storage design.

1.4 Synthetic Data Generation

The streamflow series itself was earlier described as the product of a process. It is a tempting thought that the process might be discovered and the handle of the mechanism cranked to turn out future sequences of flows. Stream flow is of course the product of a very complex physical process which

consists of a random input of rainfall which is modified by the catchment to produce the streamflow output. A study of the catchment's response to rainfall may lead to confidence in the ability to predict streamflow knowing what rainfall occurred. However, if the future sequence of streamflows is to be predicted, then the rainfall inputs themselves have to be predicted which is not possible.

Given the impossibility of predicting the future sequence of rainfall because of its random or stochastic nature, a second best approach is to artificially construct a set of rainfall sequences, each sequence being equally likely to occur in the future. These sequences can be converted by the established non-random or deterministic rainfall-streamflow relationship into equally likely future series of streamflows. The artificially constructed or synthetic rainfall series can be generated by 'drawing numbers from a hat' so to speak. The 'numbers' in the hat would have to be such that a random sampling of them would generate synthetic rainfall sequences statistically indistinguishable from the real historical series. The underlying assumption would be that the past and the future both 'obey' the same set of statistical rules.

An alternative approach is to consider the streamflow series as the product of an unknown stochastic process in the same way as the rainfall series was considered previously. The process can then be represented by a numerical model which uses some form of random sampling as an input and which generates synthetic streamflow sequences directly.

1.5 Storage Design Using Synthetic Data

If realistic synthetic streamflow series can be generated then storage design can proceed using the criterion of risk of failure

to supply rather than the survival of the system during the historical critical period. A trial storage volume can be assumed, the reservoir operation simulated over many synthetic sequences and the frequency of failure to supply noted. In this way the storage design procedure is able to sample the many different patterns of sequences of high and low flows which are possible in the future.

Storage design methods have been proposed which rely on synthetic data generated by process models. The process models in turn rely on random sampling experiments sometimes referred to as "Monte Carlo" techniques. Hazen (1914) was the first to apply such a method to the problem of storage design and was followed by Sudler (1927), Barnes (1954), Fiering (1961) and many others since.

An alternative approach follows on from the work of Moran (1954). In this approach the theory of stochastic processes is used to develop stochastic equations relating probabilities of inflow and release for a given storage. From these probability statements the risk of failure to supply can be obtained. The reader is referred to Doran (1975) for a comprehensive review of this field. In general this approach to storage design is more elegant mathematically but less flexible in relation to complex reservoir systems than that using synthetic data.

1.6 The Development of Synthetic Data Generation Techniques

This study relates to the task of constructing synthetic data generation models which fully use the statistical information to be found in the available record. The aim of using such models is to produce sets of realistic data each of which is statistically indistinguishable from the historic sequence. Given the assumption of the stationarity of the real physical process, it is assumed that

the synthetic data sets are all equally likely to occur in the future. It should be noted that such models are operational devices in that no attempt is made to simulate real physical processes. Instead the historical sequence is regarded as a set of numbers ordered in time and as such is a sample realisation of a theoretical stochastic process. This stochastic process relates to the real physical process only in its ability to produce series of numbers appearing to have the same statistical character as the observed series. This particular field of hydrology relating to synthetic data generation has been referred to as operational hydrology, (Fiering - 1967). A more common and perhaps more satisfactory term is synthetic hydrology.

The search for techniques to construct adequate synthetic data generation models and to test their adequacy has led to a vast literature. A starting point for much of this work was the study by Hurst (1951) who examined many streamflow, rainfall and other geophysical series related to hydrological phenomena.

Hurst used in his studies a statistic which is referred to in this report as the rescaled adjusted range. For streamflows and other natural phenomena where storage has relevance the rescaled adjusted range statistic is closely related to the 'storage' character of the series; that is, the extent to which storage must be provided to assure a required supply. Hurst found a discrepancy between the behaviour of the rescaled adjusted range in real data series and in series of numbers resulting from such simple random independent processes as coin tossing and card drawing experiments. He concluded that real data series have complexities which simple random processes do not emulate and he was interested in the implication of this result for storage design.

After more than a quarter of a century the difference in

character which Hurst observed between series derived from theoretical stochastic processes and real series is still the subject of research by many investigators. The discrepancy has come to be known as the Hurst Phenomenon and the attempts to explain it have led to many insights into the statistical nature of theoretical and real data series and also to controversy over the adequacy of different types of data generation models.

Since the advent of high-speed electronic computers there has been a great deal of development of synthetic data generation methods and time-series analysis techniques within the discipline of hydrology as well as within other disciplines such as econometrics. The work of Box and Jenkins (1970) stands out as a landmark and provides a unifying treatment of the subject.

1.7 Features and Aims of this Study

This report presents a general study of the rescaled adjusted range statistic and its significance in time-series analysis and synthetic data generation. It includes a review of the Hurst Phenomenon, not as a primary aim but as part of the overall story of how the rescaled adjusted range statistic has been used in 'synthetic hydrology' and how the statistic behaves in theoretical processes and real data series.

Following the review of the Hurst Phenomenon, synthetic data generation models are discussed and various properties of the rescaled adjusted range in theoretical processes examined. This part of the study provides the justification of a proposed method in which the rescaled adjusted range statistic is used to assist in time-series analysis and data generation model building.

The study concludes with applications of the proposed

method of comparison of theoretical and observed rescaled adjusted range values. The examples used are drawn from Australian and overseas streamflow, rainfall and other geophysical data series. The ability of various data generation models to adequately preserve the 'storage' character of the series being modelled is examined.

CHAPTER 2: THE RESCALED ADJUSTED RANGE AND OTHER RELATED STATISTICS

2.1 Introduction

The statistic that H. E. Hurst (1951) used to examine many geophysical time series was the range of the accumulated sums of the residuals from the sample mean, divided by the sample standard deviation. This statistic has come to be known as the rescaled adjusted range. Since Hurst's work, the rescaled adjusted range and other related statistics such as the crude or population range have received much attention in the literature. In the following sections these range statistics will be illustrated by reference to the close analogy they hold with the familiar 'residual mass curve' technique used in the critical period approach to reservoir storage design. A more rigorous definition of the rescaled adjusted range will follow and a graphical method due to Hurst (1951) for conveniently evaluating the statistic will be described.

2.2 The 'Residual Mass Curve' Storage Design Analogy

The following simple example of the residual mass curve technique is given in order to define the terms used in this discussion and to illustrate how the range statistics mentioned above encapsulate the storage nature of a time series.

Suppose that we have the following series of observations, perhaps annual inflows to a reservoir:

TABLE 2.1

TIME SERIES FOR ILLUSTRATION

Time i	1	2	3	4	5	6	7	8
Value x_i	1	2	2	1	11	1	7	7

Although i varies from 1 to 8, consider only the first six terms so that i varies from 1 to n = 6 where n is the sub-series length. The mean inflow $\bar{x}_n$ is seen to be 3. The residuals from this mean value can be accumulated as shown in Table 2.2 and plotted as in Figure 2.1

TABLE 2.2

ACCUMULATING THE RESIDUALS

x_i	$(x_i - \bar{x}_n)$	$\sum_{i=1}^{n} (x_i - \bar{x}_n)$
		0
1	-2	-2
2	-1	-3
2	-1	-4
1	-2	-6
11	+8	+2
1	-2	0

For the sub-series with n = 6, the largest negative value of the accumulated residuals from the sub-series mean defines the <u>adjusted deficit</u> $m_n = -6$, the largest positive value, the <u>adjusted surplus</u> $M_n = 2$ and their difference $M_n - m_n$, the <u>adjusted range</u> $R_n^* = 8$. The term 'adjusted' was introduced by Feller (1951) to differentiate these statistics from similar statistics defined in terms of the non-varying population mean or its estimate.

If the whole available series with i varying from 1 to n = 8 is considered, the residuals may be accumulated against the mean $\bar{x}_n = 4$, giving $M_n = 0$, $m_n = -10$ and $R_n^* = 10$. M_n, m_n and R_n^* are seen to have values varying with n, the number of terms of the available series considered in the analysis.

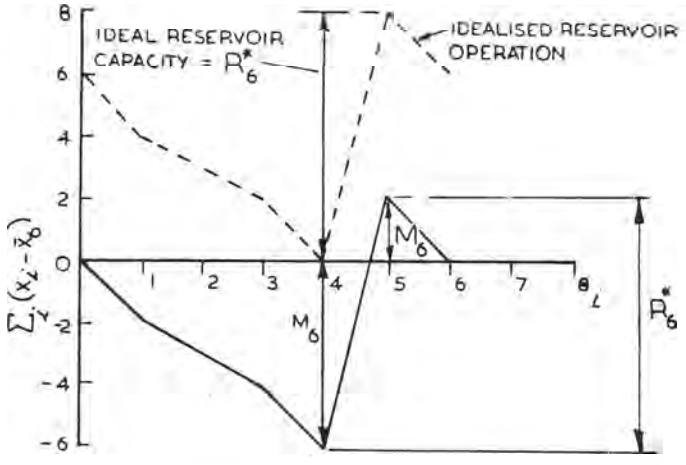

FIGURE 2·1
RESIDUAL MASS CURVE

There is an important difference between these adjusted statistics and the crude or unadjusted surplus, deficit and range. If the mean $\bar{x}_{n=8} = 4$ is considered to be the estimate of the population mean, then the crude and adjusted statistics at n = 8 are equal in value. However, to consider the variation of the crude range with n, would require the recalculation of the range at smaller n values in terms of the fixed mean value of $\bar{x}_{n=8} = 4$.

The residual mass diagram (Figure 2.1) illustrates that the adjusted range is the required storage size for an idealised reservoir operation in which the reservoir starts and finishes with a storage equal to the absolute value of the maximum deficit and continuously delivers the mean inflow.

To allow comparison of the adjusted range values determined for different time series a non-dimensional form of the adjusted range is obtained by dividing by the standard deviation, s, of the sub-series.

The resulting rescaled adjusted range, R_n^*/s, will be denoted by R_n^{**}. One feature of the residual mass curve approach to reservoir storage sizing is that as the length of record included in the analysis is increased, the amount of storage required to meet the fixed demand will tend to increase also. It will be seen later that values of R_n^* and R_n^{**} tend to increase with increasing sub-series length n.

Another feature of storage design is that the required storage size tends to be relatively greater when flow events in the series occur in 'clusters' of high and low events. The tendency of low flows to follow low flows and high flows to follow high flows is referred to as autocorrelation. It will be seen later that the presence of autocorrelation in a series also leads to relatively higher values of R_n^* and R_n^{**}.

2.3 A Formal Definition

Anis and Lloyd (1976) give the following formal definition of the rescaled adjusted range.

Consider a time series

$$\{x_t\} \quad (t = 1, \quad 2\ldots) = x_1, \quad x_2 \ldots , \qquad \text{---(2.1)}$$

the n-term mean

$$\bar{X}_n = (X_1 + \ldots + X_n) / n , \qquad \text{---(2.2)}$$

and standard deviation

$$s_n = n^{-\frac{1}{2}} \sqrt{\left\{\sum (X_r - \bar{X}_n)^2\right\}} , \qquad \text{---(2.3)}$$

the partial sums

$$S_r = X_1 + \ldots + X_r (r = 1, 2, \ldots), \qquad \text{---(2.4)}$$

the adjusted partial sums

$$nS_r^* = S_r - r\bar{x}_n \quad (r = 1, \ldots, n), \qquad \text{---(2.5)}$$

the rescaled adjusted partial sums

$$nS_r^{**} = nS_r^* / s_n \quad (r = 1, \ldots n). \qquad \text{---(2.6)}$$

The rescaled adjusted range is defined as

$$R_n^{**} = \underset{1 \leq r \leq n}{\text{Max}} (nS_r^{**}) - \underset{1 \leq r \leq n}{\text{Min}} (nS_r^{**}) \qquad \text{---(2.7)}$$

A corresponding definition is given for the adjusted range R_n^* so that

$$R_n^{**} = R_n^* / s_n \qquad \text{---(2.8)}$$

2.4 A Graphical Method for Determining the Rescaled Range

Figure 2.2 illustrates a convenient graphical method for determining the rescaled range. The method is due to Hurst (1951). The series of annual flows, Q_i, in the Brisbane River at Savages Crossing (1910 to 1951) is analysed. Instead of accumulating residuals from a mean value as was the case in section 2.2 an arbitrary base is selected for convenience. In this case residuals from a base of 500 are calculated and summed and the partial sums plotted.

The value of the rescaled range for n = 20 for example, can be evaluated for the sub-series containing the years 1 to 20. The line AB is drawn from the origin to the point on the residual mass curve corresponding to i = 20. The sum of the largest deviations (CD + EF) above and below this line gives the rescaled range R_{20}^* which in this case is approximately 3,500. If the result is divided by the standard deviation of the sub-series Q_i, i = 1, 20 (in

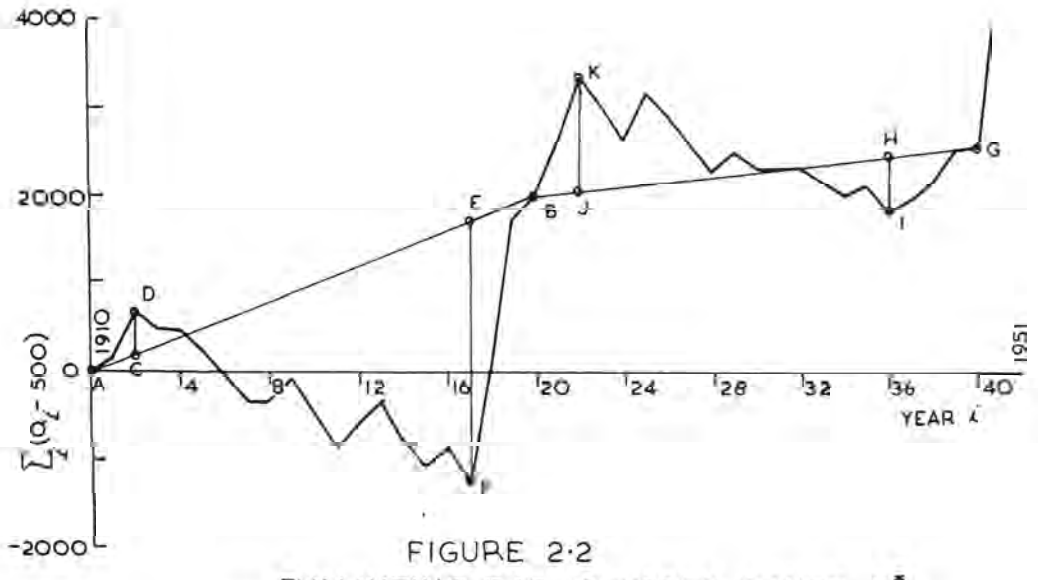

FIGURE 2·2
EVALUATING THE ADJUSTED RANGE R_N^*
BRISBANE RIVER AT SAVAGES CROSSING
ANNUAL FLOWS

this case 564) the value of the rescaled adjusted range R_{20}^{**} of 6.2 is obtained.

Another independent evaluation of R_{20}^{*} can be obtained by examining the sub-series Q_i, i = 21, 40. Again a line BG is drawn and a value of R_{20}^{*} = 2,000 determined from the sum of KJ and HI. A value of R_{20}^{**} = 6.4 is obtained by dividing the value R_{20}^{*} by the standard deviation of the sub-series Q_i, i = 21, 40 (in this case 310).

The two values of R_{20}^{**} obtained above can be regarded as independent estimates of the expected value of R_{20}^{**} for an underlying stochastic process producing the Q_i. In general the available series may be subdivided into non-overlapping sub-series of length n to obtain independent estimates of R_n^{**}. Of course where n is greater than half the series length then only one independent estimate of R_n^{**} is available.

The range statistics described in this chapter have been the subject of considerable investigation. In the next chapter some theoretical properties of these statistics are discussed.

CHAPTER 3: ANALYTICAL RESULTS FOR RANGE STATISTICS OF SOME THEORETICAL PROCESSES

3.1 Introduction

In the previous chapter the definition of the range statistics was illustrated by crude example using some data series of finite length.

The underlying stochastic process in the case of the Brisbane River flow data analysed in the previous chapter is of course unknown. It is possible however to consider data series for which the underlying stochastic process is known. For example, computers are commonly used to generate numbers which are apparently randomly selected from a normal distribution. Many sequences of such numbers can be analysed to determine the mean value of the rescaled adjusted range (R_n^{**}) at a particular sub-series length n. In this context the mean value is an approximation to the true expected value of R_n^{**} for the process. The values of R_n^{**} obtained from each of the sequences will show sampling variance about the mean and this variance will be an estimate of the true variance of R_n^{**} for the process.

A useful feature of the range statistics described in Chapter 2 is that statisticians have been able to provide closed-form expressions for their expected values in the case of some theoretical processes. Attention has centred mainly on simple independent random processes but recently theoretical results have become available for some dependant processes. These analytical expressions complement the knowledge to be gained by the alternative approach of computer simulation experiments.

3.2 Expressions Valid Asymptotically - Independent Variates

Hurst (1951) derived an expression for the expected value of the adjusted range in terms of the series length n for an independent normal variate. He used a combinatorial argument regarding the theoretical result of a coin tossing experiment. The expression is as follows:

$$E\left[R_n^*\right]/s = \sqrt{\left[\frac{\pi n}{2}\right]} \doteq 1.2533\sqrt{n} \qquad ---(3.1)$$

Feller (1951), using a different approach, proved that the above expression is valid asymptotically (i.e. its accuracy increases as n becomes large) and applies for any identically distributed independent random variate. Feller (1951) also derived the following expression for the asymptotic variance of the rescaled range for such variates.

$$\text{Var}\left[R_n^*\right]/s = \left(\frac{\pi^2}{6} - \frac{\pi}{2}\right)n \qquad ---(3.2)$$

McLeod and Hipel (1978a) point out that due to a standard convergence theorem in probability theory, for large n, $E\left[R_n^*\right]/s = E\left[R_n^*/s\right] = E\left[R_n^{**}\right]$. The above expressions, (3.1) and (3.2), can therefore be considered as asymptotic expressions for the rescaled adjusted range for an identically distributed independent random variate.

3.3 Exact Expected Values - Independent Variates

Anis and Lloyd (1953) determined an expression for the exact expected value of the crude or unadjusted range for the standard (zero mean, unit variance) independent normal variate. Solari and

Anis (1957) determined an expression for the exact expected value of the adjusted (but not rescaled) range. However, it was not until Anis and Lloyd (1976) that an expression for the exact expected value of the rescaled adjusted range of an independent normal variate became available. The expression they derived is as follows:

$$E\left[R_n^{**}\right] = \frac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n\right)} \cdot \sum_{r=1}^{n-1} \sqrt{\frac{n-r}{r}} \qquad ---(3.3)$$

where Γ represents the Gamma function.

Sen (1974) independently obtained the above expression. However, Anis and Lloyd (1976) cast doubts on the mathematical validity of Sen's derivation referring to it as "conjecture".

Anis and Lloyd (1977) derived exact explicit formulae for the distribution of the rescaled adjusted range of an independent normal variate for the cases $n \leq 4$. They surmised however that the problem of deriving such formulae for general values of n is of "unmanageable complexity".

3.4 Expressions for Dependent Variates

Sen (1977a) presented an expression for the exact expected value of the rescaled adjusted range which it is claimed applies for any normal stationary process either independent or dependent. The expression is as follows:

$$E\left[R_n^{**}\right] = \frac{2\,(n)^{\frac{1}{2}}}{(\pi)^{\frac{1}{2}}\sigma(n-1)} \cdot \frac{\Gamma\left[(n+1)/2\right]}{\Gamma(n/2)}$$

$$\cdot \sum_{k=1}^{n} \left[V(\bar{x}_k) - 2.C\,(\bar{x}_k, \bar{x}_n) + V(\bar{x}_n)\right]^{\frac{1}{2}} \qquad (3.4)$$

V( ) and C( ) are respectively the variance and co variance of the arguments, σ is the population standard deviation of the underlying normal distribution function, and $\bar{x}_k$ and $\bar{x}_n$ are the sample means of the sequence of observations up to the k^{th} and the n^{th} time points respectively. The expression (3.4) reduces to(3.3) for the independent normal process.

Sen (1977c) gives expressions for $V(\bar{x}_k)$, $V(\bar{x}_n)$ and $C(\bar{x}_k, \bar{x}_n)$ for various dependent processes. These can be substituted in (3.4) but the resulting closed-form expressions are very large. The value of $E\left[R_n^{**}\right]$ is shown to depend only on n and the lag-one autocorrelation coefficient (ρ_1) for a lag-one Markov process. Sen found good agreement between the analytical expressions and results from computer simulation experiments.

Siddiqui (1976) obtained a general expression for the asymptotic value of $E\left[R_n^{**}\right]$ for any ARMA process having a normally distributed random component. ARMA processes will be described in detail later in this report. They are a class of dependent 'short memory' processes made up of autoregressive and moving average terms. The expression obtained by Siddiqui is as follows:

$$E\left[R_n^{**}\right] = a' n^{0.5} \qquad \text{---(3.5)}$$

where $a' = 1.2533\, \gamma_0^{-0.5}(1 - \sum_{i=1}^{q} \theta_i) / (1 - \sum_{i=1}^{p} \emptyset_i)$ and γ_0 is a theoretical autocovariance function at lag 0 evaluated using an algorithm given by McLeod (1975). θ_i and $\emptyset_i$ are the moving average and autoregressive parameters respectively. q and p are the number of such parameter terms included in the process.

3.5 Comparison of Asymptotic and Exact Expressions for an Independent Normal Variate

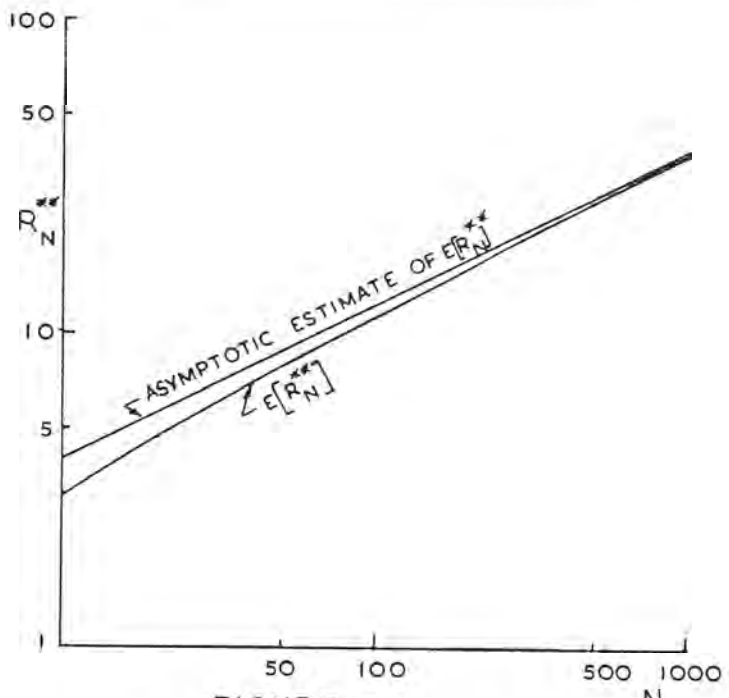

FIGURE 3·1
CONVERGENCE WITH N OF EXACT AND ASYMPTOTIC EXPRESSIONS FOR $E[R_N^{**}]$ - INDEPENDENT NORMAL VARIATE

Figure 3.1 shows the asymptotic and exact expected values of the rescaled adjusted range for an independent normal variate. The values are derived from expressions (3.1) and (3.3). The convergence of the exact result to the asymptotic value is quite slow as is shown in Table 3.1.

TABLE 3.1

INDEPENDENT NORMAL VARIATES

n	$\frac{\text{Asymptotic value } R_n^{**}}{\text{Exact expected value } R_n^{**}}$
20	1.26
50	1.134
100	1.094
200	1.066
500	1.039
1000	1.027

CHAPTER 4: THE OBSERVATIONS OF H. E. HURST

4.1 Introduction

The rescaled adjusted range is closely related to reservoir storage capacity determined by the 'residual mass curve' method due to Rippl (1883). Although this method has been in use for a very long time, interest in the rescaled adjusted range as a general tool for time series analysis stems from the work of Hurst (1951).

Hurst was primarily interested in computing the storage required in the Great Lakes of the Nile Basin to provide adequate regulation of Nile River flows. In his (1951) paper he pointed out the uncertainty in estimates of storage requirements computed from a single historical record. Large variations in storage requirements were observed between that obtained from the whole available record, and those obtained assuming various portions of the same record were all that was available to the designer.

Hurst attempted to overcome the problem of uncertainty by resorting to a theoretical approach. He noted that many natural phenomena have frequency distributions which are approximately normal if the order of occurrence is disregarded. He therefore sought to obtain a theoretical expression for storage requirement for a random process involving sampling from a normal distribution. The form in which storage requirement was expressed was the adjusted range R_n^* which is, as previously defined, the minimum storage required to maintain a constant discharge equal to the mean inflow.

In this chapter a detailed review is carried out of Hurst's (1951) paper with a view to highlighting aspects of his work which are important in later discussion.

4.2 Derivation of an Expression for the Expected Value of the Adjusted Range

As mentioned in Chapter 3, Hurst derived the expression (3.1) for the expected value of the adjusted range. The expression was derived using a combinatorial argument regarding the theoretical result of a coin tossing experiment. It involved the assumption of large n and the use of Stirling's approximation for factorial n. The expression is therefore an asymptotic result and is not true for small n.

Hurst tested the expression by a series of experiments: (a) ten coins tossed 1,000 times, (b) probability cards cut 1,000 times and (c) a sequence of 1,000 numbers derived from bond serial numbers published in newspapers. Each trial was repeated 1,000 times, certainly a prodigious amount of work. He found for his n = 1,000 sequences a mean value of $\frac{R^*}{s\sqrt{n}}$ of 1.22, close to the theoretical value of 1.25. He also observed considerable variation in individual values of $\frac{R^*}{s\sqrt{n}}$. The standard deviation of the mean values derived from 30 sets of 100 observations was 0.32.

4.3 Evaluation of the Rescaled Adjusted Range in Geophysical Time Series

Hurst carried out an extensive investigation of 75 different observed annual series which included river and lake levels and flows, rainfall, temperature and pressure means, annual growth of tree rings, mud varve thicknesses, sunspot numbers and wheat prices. The longest series examined was 4,000 years of mud varve thicknesses from Lake Saki in the Crimea - (mud varve thicknesses are believed to be related to annual inflows). He also examined a 1,040 year long record of annual high flood levels at the Roda Gauge on the Nile

River. In all, 690 values of the rescaled adjusted range were calculated for data series of various lengths n.

Related phenomena were treated as a group with sub-series selected so as to give a number of samples of the same length n. The mean R_n^{**} value for that n value was calculated.

4.4 Observed Behaviour of the Rescaled Adjusted Range with Time Series Length

Hurst found that log-log plots of the averaged R_n^{**} values versus n showed approximately linear relationships over the range of $n(35 \leq n \leq 2\ 000)$ considered. He simplified the process of fitting a straight line to the points. The assumption was made that, as the theoretical value of the rescaled adjusted range is unity for n = 2, one end of the line should pass through the point $R_n^{**} = 1, \quad n = 2$.

The equations of the lines then had the form:

$$\overline{\log R_n^{**}} = K\left[\overline{\log(n)} - \log(2)\right] \qquad ---(4.1)$$

where $\overline{\log R_n^{**}}$ is the mean of the logarithms of the R_n^{**} values and $\overline{\log(n)}$ is the mean of the logarithms of the series lengths n.

The Hurst coefficient K was then defined as the slope

$$K = \overline{\log R_n^{**}} \Big/ \left[\overline{\log(n)} - \log(2)\right] \qquad ---(4.2)$$

K derived in this way, is an expression for a slope on the log-log plot of the averaged $\log R_n^{**}$ and log n values. Values of K for each of the group of related phenomena were determined using expression (4.2). Hurst's results are summarised in Table 4.1.

TABLE 4.1

VALUES OF K DETERMINED BY HURST (1951)
FOR GROUPS OF RELATED PHENOMENA

Phenomena	K
River Statistics	0.75
Rainfall	0.70
Temperature & Pressure	0.70
Tree Rings	0.80
Varves, Lake Saki	0.69
Varves, Canada & Norway	0.77
Sunspot Nos. & Wheat Prices	0.69

Values of K were then determined for each of the 690 individual data points using the expression:

$$K = \log R_n^{**} / \log (n/2) \qquad \text{---(4.3)}$$

UNIVERSITY OF NEW SOUTH WALES
WATER RESOURCES LIBRARY
WATER RESEARCH FOUNDATION OF AUSTRALIA

which is now familiar as the definition of the Hurst Coefficient. The mean value of K was found to be 0.73 and standard deviation 0.09. Individual values varied over the range 0.46 to 0.96.

When the series longer than 200 years were disregarded in order to give weight to the more precise measurements relating to river flows, rainfall and temperature, the mean value of K was found to be 0.72. This led to the following expression for R_n^{**}:

$$R_n^{**} = \left(\frac{n}{2}\right)^{0.72} = 0.61\, n^{0.72} \qquad \text{---(4.4)}$$

4.5 Comparison of Observed and Theoretical Behaviour

Hurst drew a comparison between the theoretically derived expression (3.1),

$$\frac{R_n^*}{s} = 1.25\, n^{0.5}$$

and the empirical relationship (4.4),

$$R_n^{**} = 0.61\, n^{0.72}$$

He noted that the higher exponent in expression (4.4) indicated a more rapid growth of the rescaled adjusted range with series length in real data than theory would predict. This apparent discrepancy between theory and observation has come to be known as the 'Hurst Phenomenon'.

In summary, Hurst drew attention to the tendency of natural data to occur in groups of high and low values even though the data may have a normal frequency distribution when the order of occurrence is not considered. As a consequence of the additional complexity of real data, the theoretical expression (3.1) for the adjusted range understates the storage requirements of real data series.

It should be noted that a distinction has been drawn in this discussion between $\frac{R_n^*}{s}$ and R_n^{**} in relation to expression (3.1) which expresses an expected value. It is not generally true that

$$\frac{E\left[R_n^*\right]}{s} = E\left[R_n^{**}\right]$$

and in fact such a relationship is only valid asymptotically (See section 3.2). Expression (3.1) is also only valid asymptotically but it is not clear that Hurst was aware of this point.

In this chapter various aspects of the contents of Hurst's (1951) paper have been highlighted. The following chapter will present some comments on Hurst's observations and method of analysis with a view to clarifying aspects of the 'Hurst Phenomenon' and also as a means of progressing towards the use of the rescaled adjusted range as a time series analysis tool.

CHAPTER 5: SOME COMMENTS ON HURST'S OBSERVATIONS AND METHOD

5.1 Introduction

Hurst's remarkable work (Hurst 1951) has led to a large literature which continues unabated more than a quarter of a century later. Consideration of the 'Hurst Phenomenon' described in the previous chapter (section 4.5) has led investigators to the formulation of new and more complex stochastic models, and to such important questions as to what extent historical data series of hydrological phenomena can be regarded as statistically stationary over a period of time. Much attention has been focussed on the rescaled adjusted range statistic itself both as regards its properties for theoretical processes and its behaviour in real data series.

In this chapter attention will be centred on aspects of Hurst's method which have important consequences for interpreting the 'Hurst Phenomenon'.

5.2 The Asymptotic Nature of Hurst's Expression for the Adjusted Range

Yevjevich (1972) makes the following statement regarding the work of Feller (1951): "Feller used other means to develop the asymptotic mean and variance of the adjusted range and not the expected value of the range as Hurst thought." Whether or not Hurst realised the asymptotic nature of his expression is not clear but the assumption of large n in its derivation is quite clearly stated. The combinatorial argument upon which it is based depends upon Stirling's approximation for factorial n. Hurst proceeded to check the expression by means of experimental series of length n = 1,000

involving coins, cards and published bond numbers and found good agreement at this large value of N. The mean value of $R_n^{**} / \sqrt{n}$ for the experiments was found to be 1.22 as against the theoretical value of 1.25. It is interesting to note that for the independent normal variate with n = 1,000, the exact expected value $R_n^{**} / \sqrt{n}$ is 1.217 from Anis and Lloyd (1976).

5.3 Comparison of Observed and Theoretical Asymptotic Values

In his paper Hurst does not discuss the range of values of n for which the expression (3.1)

$$\frac{R_n^*}{s} = 1.25\, n^{0.5}$$

is applicable. He does however compare the exponent of 0.5 with the exponent of 0.72 in the empirical relationship (4.4)

$$R_n^{**} = 0.61\, n^{0.72}$$

determined from series varying in length from n = 30 to 200.

The theoretical result (3.3) derived by Anis and Lloyd (1976) for the exact expected value of R_n^{**} for an independent normal variate, allows an evaluation of the effect of Hurst's comparison of observed and asymptotic behaviour. They defined a local Hurst exponent h(n), where

$$h(n) = \partial(\log E\left[R^{**}\right]) / \partial(\log n) \qquad ---(5.1)$$

h(n) can be approximated as

$$h(n) = \frac{\log E\left[R_{n+1}^{**}\right] - \log E\left[R_{n-1}^{**}\right]}{\log (n+1) - \log (n-1)} \qquad ---(5.2)$$

h(n) was evaluated using values obtained from expression (3.3). Table 5.1 shows that h(n) is significantly higher than 0.5 for an independent normal variate when evaluated at small to medium values of n.

TABLE 5.1

THEORETICAL VALUES OF THE LOCAL HURST EXPONENT h(n) FOR AN INDEPENDENT NORMAL VARIATE

(Anis and Lloyd - 1976)

n	h(n)
5	.6762
40	.5672
100	.5429
200	.5315
500	.5202

It is of interest to see what results Hurst would have obtained if all the series he examined were composed of values derived from independent normal variates. This can be done by using values of R_n^{**} derived from expression (3.4). The analysis of the Lake Saki mud varves is reproduced in Table 5.2 using this assumption. (See Table 7 of Hurst (1951).)

With the advantage of hindsight and the availability of the expression for the exact expected value of R_n^{**}, it becomes apparent that the comparison of exponents that Hurst made is not a valid one. The reason for the failure of K in Table 5.1 to equal 0.5 lies not only in the fact that R_n^{**} varies with $n^{0.5}$ only asymptotically, but more significantly in the definition of K itself which will be discussed further.

TABLE 5.2

HURST'S ANALYSIS OF LAKE SAKI MUD VARVE DATA REPRODUCED ASSUMING AN INDEPENDENT NORMAL VARIATE

No. of cases	N years	R_n^{**} (independent normal variate)	$\log_{10} N$	$\log_{10} R_n^{**}$	$\left[\frac{\log_{10} R_n^{**}}{\log_{10} \frac{N}{2}}\right]^K$
40	50	7.81	1.70	.89	.62
40	100	11.45	2.00	1.06	.62
20	200	16.62	2.30	1.22	.61
8	500	26.90	2.70	1.43	.60
4	1000	38.50	3.00	1.59	.59
2	2000	55.90	3.30	1.75	.58
mean of 114 cases			2.06	1.08	.61

(mean K = .69 for historical data)

5.4 The Definition of Hurst's Coefficient K

Hurst was interested in comparing the rate of growth of the rescaled adjusted range with n, for observed series, with that predicted by theory. As such he was primarily concerned with values of h and hence slopes on log-log plots.

A least squares fit to log R_n^{**} versus log n data yields an equation of the form:

$$\log (R_n^{**}) = \log a + b \log n \qquad \text{---(5.3)}$$

This indicates the relationship

$$R_n^{**} = a (n)^b \qquad \text{---(5.4)}$$

which is an expression containing two constants or parameters (a, b) to be determined in the fitting process.

As described in section 4.4, Hurst chose a simpler approach forcing the straight line fit through the point $R_n^{**} = 1$, $n = 1$ ($\log R_n^{**} = 0$, $\log 2$). The other end of the line was the centre of gravity of the observed data points i.e. some point $(\overline{\log R_n^{**}})$, $(\overline{\log n})$. This gave expression (4.2)

$$\overline{\log R_n^{**}} = K\,(\overline{\log n} - \log 2) \qquad (4.2)$$

Hurst used $\overline{\log R_n^{**}}$, $\log R_n^{**}$ and $\overline{\log n}$, $\log n$ interchangeably in deriving expression (4.3) from individual data points. This led to the definition of K given by expression (4.3)

$$K = \log R_n^{**} / \log (n/2) \qquad (4.3)$$

It is important to note here the dual character of K. In expression (4.2) it is clearly a slope estimator. However in expression (4.3) it is both a single point estimate of slope and a logarithmic transformation of a single point value of R_n^{**}. It will be seen later that two main schools of thought about the 'Hurst Phenomenon' diverge essentially from this point. One school follows the idea of K being a slope and the other assumes K to be a transformation indicating the magnitude of R_n^{**}.

5.5 Hurst's Coefficient K as an Estimator of Slope

The limitations of K as an estimator of the slope of a plot of $\log R_n^{**}$ versus $\log n$ are shown by Figure 5.1.

The $\log R_n^{**}$ versus $\log n$ plot for such a 'short memory' theoretical process as an independent normal variate is in fact curvilinear. If we estimate the slope empirically by K at, for example,

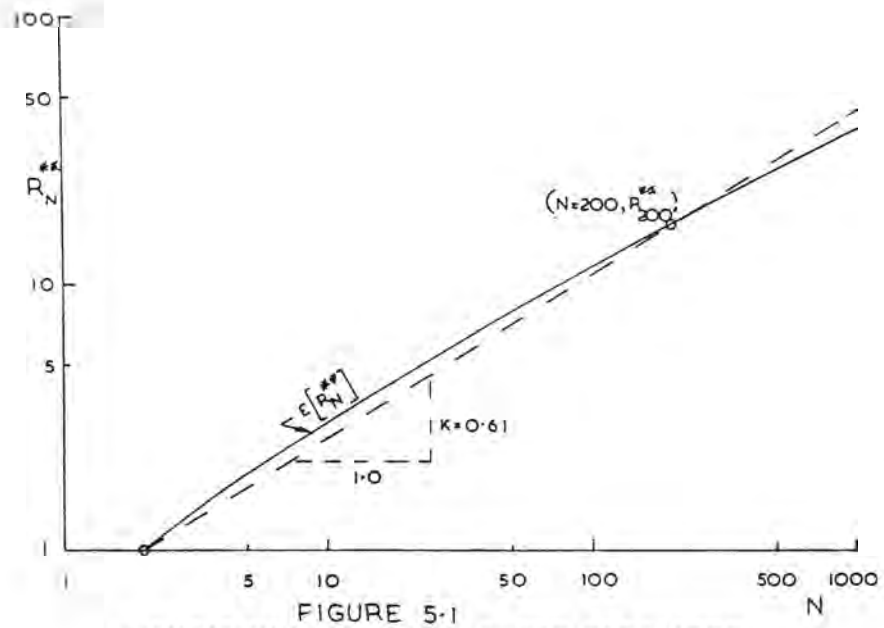

FIGURE 5·1
LIMITATIONS OF K AS A SLOPE ESTIMATOR
- INDEPENDENT NORMAL VARIATE

n = 200 is seen to be greater than the true slope due to the curvature in the log-log plot. The plotted values are derived from expression (3.3) given by Anis and Lloyd (1976).

At n = 200, k, the slope of the chord between the points (R_n^{**} = 1, n = 2) and (R_{200}^{**}, n = 200), is seen from Figure 5.1 to be 0.61. The values of K at other values of n for this case of an independent normal variate are shown in Table 5.3. The expected values of R_n^{**} were obtained from expression (3.3) and the corresponding value of K from Hurst's definition, expression (4.3).

The limitations of K as a slope are to be seen in Table 5.3 where very slow convergence to 0.5 for an independent normal variate is apparent. As a starting point for looking for some unexplained behaviour in an observed series of even very long length, the values of K given in Table 5.3 should be considered.

It is obviously incorrect to assume the presence of some 'Hurst Phenomenon' or 'Hurst Effect' in real data simply on the basis

TABLE 5.3

VALUES OF THE HURST COEFFICIENT K FOR AN INDEPENDENT NORMAL VARIATE

n	$E\left[R_n^{**}\right]$	K
5	1.9274	.716
40	6.8895	.644
100	11.4533	.623
200	16.6214	.610
1000	38.4969	.587

of observed K values being higher than 0.5. It will be seen later in this report that processes which possess autocorrelation exhibit higher values of K than those shown in Table 5.3.

It remains to be said that the preceding comments on Hurst's work raise points which are more relevant to a discussion of the literature that follows his paper than to the paper itself. Hurst's work is an outstanding contribution. His main point, that natural time series have characteristics which lead to the requirement of more storage for the same yield than simple independent random processes, is of course a valid one. The question as to the nature of this difference between observation and theory remains both important and topical.

CHAPTER 6: INDICATORS OF THE HURST PHENOMENON

6.1 Introduction

Hurst's observations showed that for a wide sampling of natural time series the rescaled adjusted range varied with series length n as

$$R_n^{**} \propto n^h \qquad ---(6.1)$$

The exponent h has become known as the 'generalised Hurst coefficient' or 'Hurst exponent'. Hurst estimated it as K and found an average value of 0.73 over many observed natural series. The tendency for estimates of h obtained from observed series to be larger than the theoretical value of 0.5 has come to be known as the 'Hurst Phenomenon'.

Two contributing factors to the discrepancy in exponents have been discussed already:

(i) The comparison of the theoretical asymptotic behaviour of the rescaled adjusted range with that observed in series of finite length.

(ii) The bias inherent in Hurst's estimator K.

In this chapter some of the alternative estimators of h proposed in the literature will be described.

6.2 A Perception of the Hurst Phenomenon

Given an appreciation of the above mentioned factors (i) and (ii), the Hurst Phenomenon may be perceived as the failure of the behaviour of the estimated expected value of the rescaled adjusted range in long observed series to show the theoretical asymptotic

behaviour. Mandelbrot and Wallis (1969) take this view and reject many theoretical processes on the basis that for these processes there is an asymptotic convergence of the exponent h towards a value of 0.5. The processes which are rejected are Gaussian in nature and include identically distributed independent variates and the general class of 'short memory' autoregressive and moving average processes.

Mandelbrot and Wallis proposed an alternative theoretical process called 'Fractional Gaussian Noise' which is capable of preserving a specified value of h, between zero and one, for any value of series length n. McLeod and Hipel (1978a) give a concise account of Fractional Gaussian Noise and its approximations.

A feature of Fractional Gaussian Noise is that such a process is specified completely by its mean, variance and constant Hurst exponent h. The critical importance of h in this context led Mandelbrot and Wallis (1969d) to propose an alternative estimation procedure to the Hurst coefficient K.

6.3 Alternatives to the Hurst Coefficient K - Slopes Estimated from Many Data Points

Mandelbrot and Wallis (1969d) proposed a graphical procedure which they call a 'pox diagram', an example of which, taken from their paper, is shown as Figure 6.1. R_n^{**} is evaluated for a standardised set of subseries lengths n = 3,4,5,7,10,20,40,70,100, 200,700,1000,2000,4000,7000 and 9000 and $n \leq T$, where T is the total length of the series to be analysed. For each value of n a maximum of 14 evaluations of R_n^{**} are made using sub-series obtained by moving the starting point progressively along the record. The sub-series overlap and hence the values of R_n^{**} obtained are not independent. At each value of n the various R_n^{**} values are plotted on a log-log

diagram. Mean values of R_n^{**} are marked for each set and the log-log slope estimated by a straight line fitted by eye. Mandelbrot and Wallis designated this indicator of the Hurst Phenomenon as H.

Wallis and Matalas (1970) proposed refinements to the above procedure and defined estimatore FH(i) and GH(i). In this approach slope estimates are obtained by a least squares regression on the mean R_n^{**} values. The regression is carried out through the mean values at each sub-series length to avoid the bias that would occur if all individual values were included in the regression, there being fewer values available as the sub-series length increases.

The designation (i) in FH(i) and GH(i) refers to the minimum sub-series length included in the regression. The latter is specified in an attempt to avoid bias being introduced by the pronounced curvature in the log-log plot at small n. This curvature is apparent for example in Figure 5.1. The designations F and G

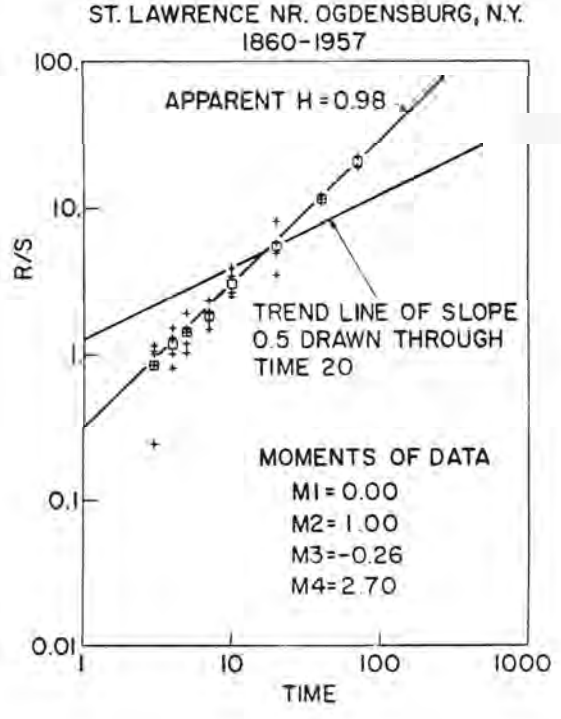

FIGURE 6·1
"POX" DIAGRAM - AFTER MANDELBROT AND WALLIS (1969d)

refer to different schemes for dividing the series to be analysed into sub-series. In both cases the subseries overlap.

Wallis and Matalas (1970) carried out computer experiments to investigate the small sample properties of K and GH(10) for independent variables and lag-one Markov processes and found that K has greater bias and GH(10) greater variance.

6.4 Alternatives to the Hurst Coefficient K - Slopes Estimated from A Single Data Point

Gomide (1975) proposed another estimator of the slope h which will be referred to here as YH. He used, as a basis for the definition, expression (3.1) due to Hurst (1951) and Feller (1951) which gives the asymptotic value of $E\left[R_n^{**} \right]$ for an identically distributed independent random variate. Gomide proposed the following definition of YH

$$R_n^{**} = 1.2533\ n^{YH} \qquad \text{---(6.2)}$$

which by taking logarithms may be restated as

$$YH = (\log R_n^{**} - \log 1.2533) / (\log n) \qquad \text{---(6.3)}$$

Siddiqui (1976) proposed another estimator of the slope h which will be referred to here as SH. He also based the definition of the estimator on an expression for the asymptotic value of $E\left[R_n^{**} \right]$, in this case expression (3.5). This expression is generally applicable to ARMA processes having normally distributed random components. Siddiqui defined SH as follows:

$$R_n^{**} = a' \, n^{SH} \qquad ---(6.4)$$

where a' is a constant, the value of which depends upon the assumed underlying ARMA process as previously discussed in section 3.4 of this report. By taking logarithms, expression 6.4 may be restated as

$$SH = (\log R_n^{**} - \log a') / (\log n) \qquad ---(6.5)$$

Both Gomide's YH and Siddiqui's SH estimators give slope values derived from values of R_n^{**} and n at a single point as does Hurst's estimator K.

Hipel and McLeod (1978a) examined values of K, YH and SH obtained from 23 geophysical series of lengths varying from 96 to 1164 years. The results of their study are shown in Table 6.1

TABLE 6.1

VALUES OF SLOPE ESTIMATORS K, YH & SH FROM
23 GEOPHYSICAL TIME SERIES
(McLeod & Hipel 1978a)

	K	YH	SH
Mean	0.701	0.660	0.577
Standard deviation	0.084	0.131	0.078

Table 6.1 shows that the YH and SH values are lower than K. Mean values of YH and SH are within 1 or 2 standard deviations of 0.5, the theoretical asymptotic value of each of the estimators for 'short memory' processes. McLeod and Hipel suggest that it can be argued

that the 'Hurst Phenomenon' is not significant for the YH and SH statistics.

The variety of approaches that have been proposed for estimating the slope parameter h in expression (6.1) is an indication of the difficulties involved. These problems will be further illustrated in the next chapter where some sampling experiments carried out in this study are described.

CHAPTER 7: SOME SAMPLING EXPERIMENTS WITH AN INDEPENDENT NORMAL VARIATE

7.1 Introduction

In this chapter the results are presented of some sampling experiments involving the standard normal independent variate. Sequences of values of the variate were produced by computer simulation and were examined to determine values of the rescaled adjusted range statistic.

The results presented here show the considerable variation in individual values of R_n^{**} obtained from realisations of theoretical processes.

7.2 Variation in the Slope of the Log-log Plot

Figure 7.1 shows a log-log plot of R_n^{**} versus n from a sequence of 900 values generated using the FORTRAN routine GENRATE written for this study (See Appendix). Each plotted point represents the calculated value of R_n^{**} for the single sub-series of length n extending from the start of the synthesised sequence. That is, each point represents a single sampling of R_n^{**} for each value of n. There is considerable scatter on the plot which illustrates some of the problems in estimating a Hurst exponent. In this case a least squares fit through the plotted points indicates a slope of .60 .

Figure 7.2 shows a plot resulting from an approach similar to the GH procedure described by Wallis and Matalas (1970). The same synthesised sequence of 900 independent normal variables is divided into as many non-overlapping sub-series as are available for each nominated sub-series length n. For example, the value of R_n^{**} plotted for n = 10 is the mean of 90 values and for n = 100 is the mean of 9 values. In each case the first or only sub-series of

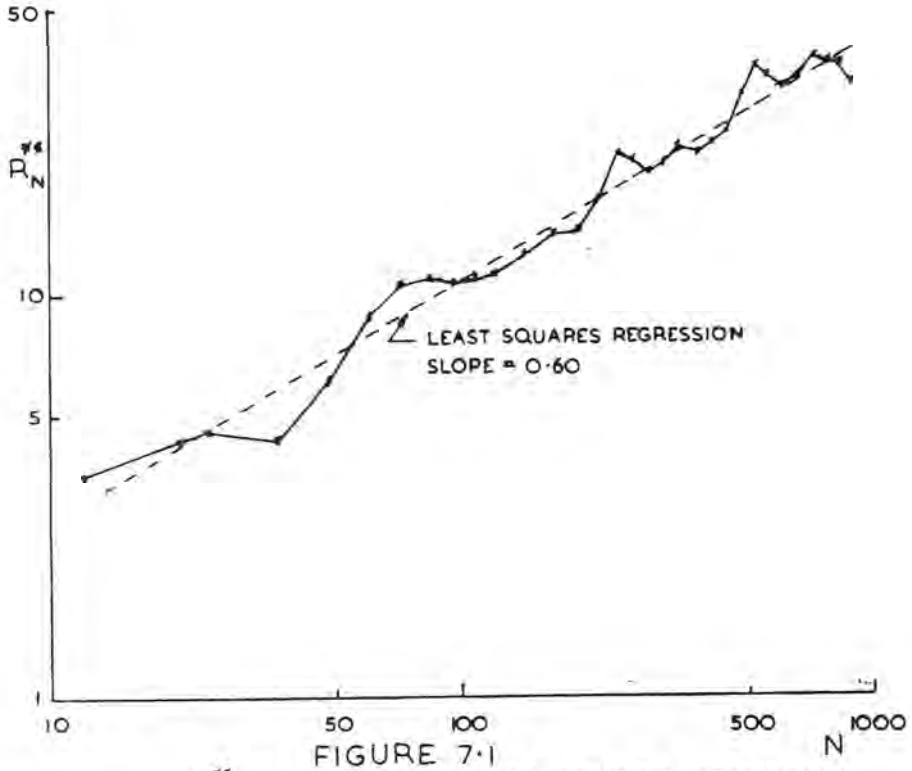

FIGURE 7·1
R_N^{**} FROM 900 VALUES OF THE STANDARD NORMAL VARIATE - SINGLE VALUES OF R_N^{**} PLOTTED

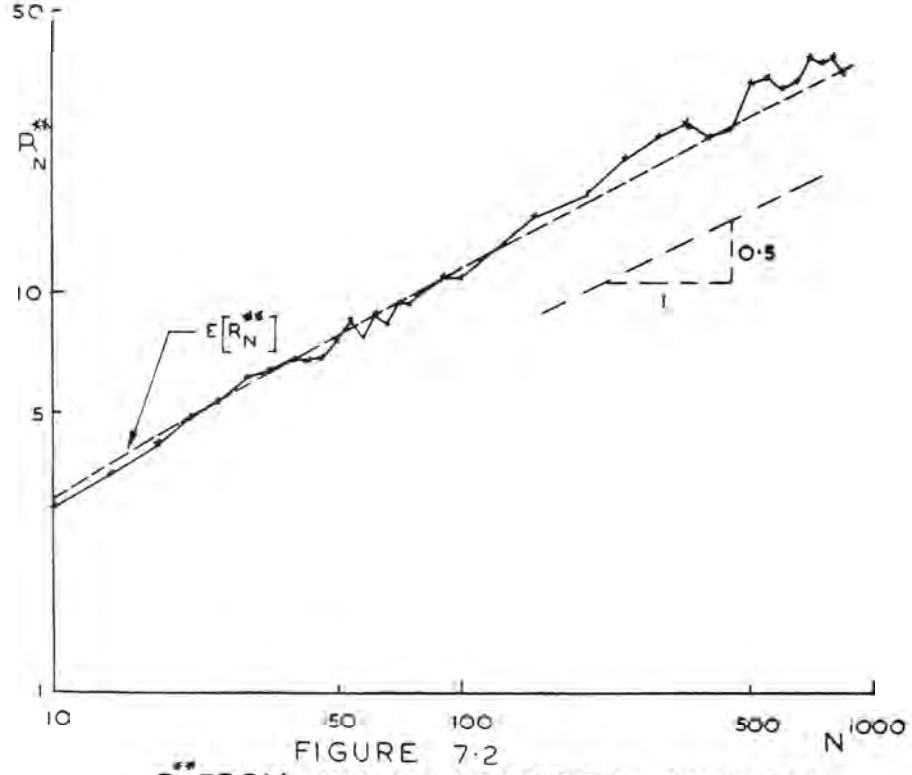

FIGURE 7·2
R_N^{**} FROM 900 VALUES OF THE STANDARD NORMAL VARIATE - MEAN VALUES OF R_N^{**} PLOTTED

length n starts at the beginning of the overall series. The procedure used here differs from that used by Wallis and Matalas (1970) in that only non-overlapping sub-series are considered.

Exact expected values obtained from expression (3.3) as given by Anis and Lloyd (1976) are also plotted in Figure 7.2. The increase in scatter about the exact expected values can be seen as n becomes larger and the plotted points represent averages from small samples. The difficulties of estimation of the true underlying slope from the plotted points is apparent. In comparison with Figure 7.1 the analysis of all available non-overlapping sub-series i gives, where n is small enough, more than one independent sample of R_n^{**}. The resulting mean values of R_n^{**} plotted in Figure 7.2 are more representative of the 'true' expected values of R_n^{**}. Hence, for smaller n, a better view is gained of the underlying relationship between R_n^{**} and n. The rescaled adjusted range analysis was carried out using the FORTRAN routine RANGE written for this study (See Appendix).

7.3 Variation in Rescaled Adjusted Range Values

In Chapter 5 it was pointed out that two factors contribute to the Hurst exponent h being greater than 0.5 when estimated from an observed series. Firstly the comparison of observed values with the value of 0.5 determined from an asymptotic expression is erroneous, and secondly, the estimator K gives higher values than the true slope of the log R_n^{**} versus log n plot. Figures 7.1 and 7.2 indicate a further problem, that of the considerable variation in sample R_n^{**} values.

Table (7.1) shows some results from an analysis of 12 sequences of 70 values of the standard normal variate. The table

TABLE 7.1

12 SEQUENCES OF THE STANDARD INDEPENDENT NORMAL VARIATE

Sequence length = 70

Sequence	R_n^{**} at n=30	R_n^{**} at n=70	K at n=30	K at n=70	Least sqs. slope H from n=10,70
1	5.009	10.523	.59	.66	.51
2	3.156	6.485	.42	.53	.39
3	8.560	15.262	.79	.77	.68
4	4.559	11.036	.56	.68	.57
5	4.150	6.292	.53	.52	.45
6	8.255	9.266	.78	.63	.53
7	4.938	16.452	.59	.79	.66
8	4.121	7.818	.52	.58	.47
9	3.812	9.770	.49	.64	.53
10	6.398	6.506	.69	.53	.49
11	4.615	9.019	.56	.62	.48
12	5.921	6.665	.66	.53	.45
Average	5.29	9.59	.60	.62	.52

presents a small sample of values but illustrates an important point. The variability of R_n^{**} can lead to very high values of K in individual realisations of the underlying process even in the case of a process consisting of independent random variables. This is illustrated by an estimate of K of 0.79 for n = 70 in Table (7.1). The values of H, the slope of the least squares fit to the log-log plot, for points between n = 10 and n = 70, also have considerable variation but their mean lies closer to the asymptotic exponent h = 0.5. This is in accord with the results obtained by Wallis and Matalas (1970).

The matter of the variance of R_n^{**} and K for some theoretical processes will be examined in detail in a later chapter of this report.

CHAPTER 8: THE LOG-LOG SLOPE OF THE RESCALED ADJUSTED RANGE PLOT FOR SOME OBSERVED AND SYNTHETIC SERIES

8.1 Introduction

As discussed in Chapter 6, Mandelbrot and Wallis (1968) in their approach to the Hurst Phenomenon, concentrated their attention on the slope of the log-log relationship between the rescaled adjusted range R_n^{**} and subseries length n. They developed a Fractional Gaussian Noise process which has the implication of infinite memory and which is based on the concept of 'self similarity'. A consequence of the self similarity concept is that the expected value of the Hurst exponent is constant over all time intervals.

A Fractional Gaussian Noise process is defined by its mean, variance and Hurst coefficient and consequently the problem of obtaining the most appropriate estimate of the Hurst coefficient is of vital concern. The argument for this type of process model relies heavily upon the linearity of the log R_n^{**} versus log n plot of observed series. It is claimed that the failure of observed slopes to approach the value of 0.5, which is the asymptotic slope for the 'short memory' process models commonly used in synthetic hydrology, is proof of the superiority of Fractional Gaussian Noise models. Mandelbrot and Wallis go further than this claim and propose 'infinite memory' processes as a physical reality.

It should be noted that this approach of concentrating attention on the slope of the log-log plot does not directly concern itself with the absolute size of R_n^{**} for particular finite values of series length. However, the magnitude of R_n^{**} has a direct implication for reservoir storage design. The magnitude of R_n^{**} is the storage

size, expressed in standard deviation units, of an 'ideal' reservoir. This ideal reservoir has the minimum storage required to maintain a constant discharge equal to the mean inflow over the duration of the series while starting and finishing the series with the same storage.

Many series of the same length can be sampled from a theoretical stochastic process. The ideal reservoir size for each series will show considerable variation but the mean of the sizes is an estimate of the true expected ideal size for the process and for the series length nominated.

It may be that a designer wishes to examine the usefulness of a process as a stochastic model of some real data series, say with a view to the computer generation of synthetic sequences for storage design. The expected ideal reservoir size for the process at particular series lengths may be compared with the best estimates obtained from the real data series. While it is of interest to observe whether the ideal storage size increases with series length at a similar rate in both cases, it is of perhaps more crucial importance to observe whether the process values are much higher or lower than those obtained from the real data.

It is important to note that K, YH and SH, the estimators of Hurst's coefficient h, discussed in Chapter 6, are in fact transformations of the magnitude of the ideal reservoir size (R_n^{**}) at a particular series length. The other estimators discussed in Chapter 6, that is, H, GH and FH, are true slopes depending upon values of R_n^{**} at many different series lengths and hence give no indication of the magnitude of R_n^{**}.

In this chapter the log R_n^{**} versus log n relationships of some observed and synthesised data series will be examined.

8.2 The Rescaled Adjusted Range for Some Synthetic Series

Figure 8.1 shows the behaviour of the rescaled adjusted range for series sampled from three theoretical 'short memory' process models. The processes are the independent normal variate, a lag-one Markov process with lag-one autocorrelation coefficient of 0.3 , and a mixed Autoregressive-Moving Average (ARMA (1,1)) process. The autoregressive and moving average parameters of the latter process are $\phi_1 = 0.9$ $\theta_1 = 0.7$ respectively. All processes have a theoretical mean of zero and unit variance.

The terms used above in relation to the theoretical processes as well as their structure will be fully discussed in Chapter 9. For the time being they can be regarded as being representative of short memory processes or 'models' that have been used in hydrology to generate sequences of synthetic data.

A rescaled adjusted range analysis has been carried out on a single realisation of 900 values of each process using the procedure described in Chapter 7, that is, each plotted point represents the mean value determined from as many non-overlapping subseries as are available for the selected value of n. The single sets of 900 values have been selected to allow a valid visual comparison between results for realisations of the processes and results for some observed series with length n in the range 800 to 1,000.

Figure 8.1 shows how increasing complexity of the process autocorrelation increases the slope of the plot particularly in the range of n less than about 200. The size of R_n^{**} is also increased with the increase generally larger at greater subseries length. With a sample of this size there are considerable fluctuations in the plot but visually one gets the impression of the plot 'rolling over' or converging to a slope of about 0.5. The problem of determining

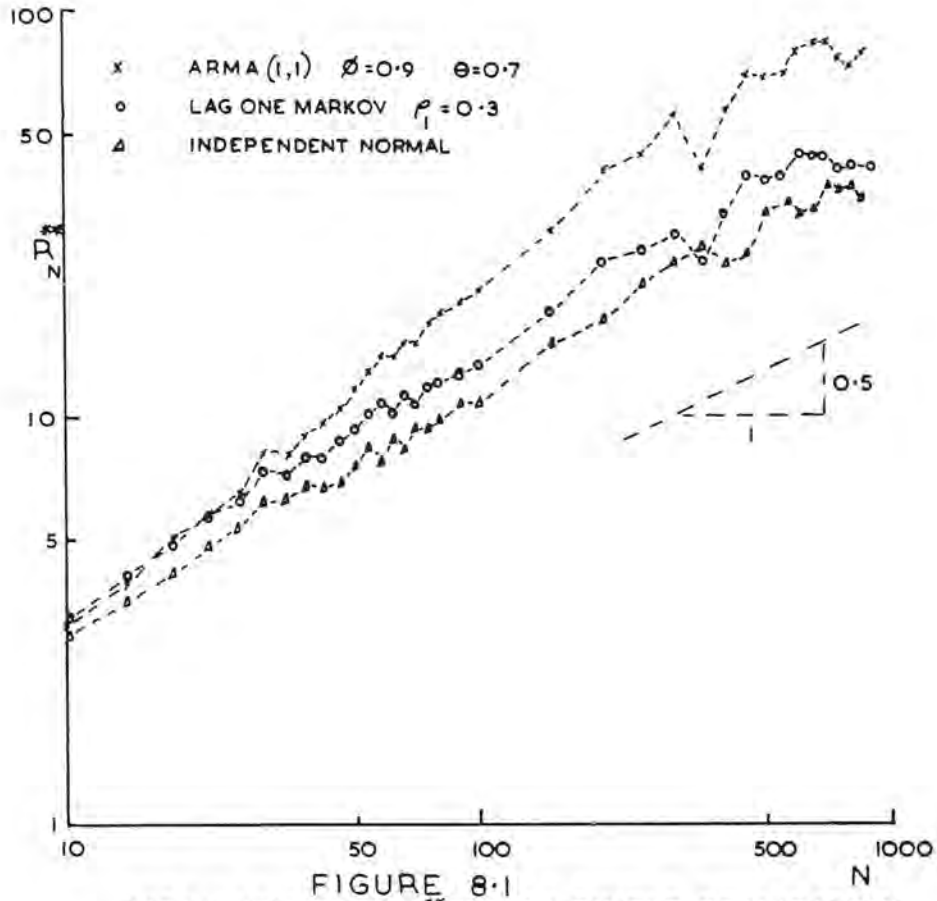

FIGURE 8·1
MEAN VALUES OF R_N^{**} FROM A SINGLE REALISATION OF 900 VALUES FROM EACH OF THREE THEORETICAL PROCESSES

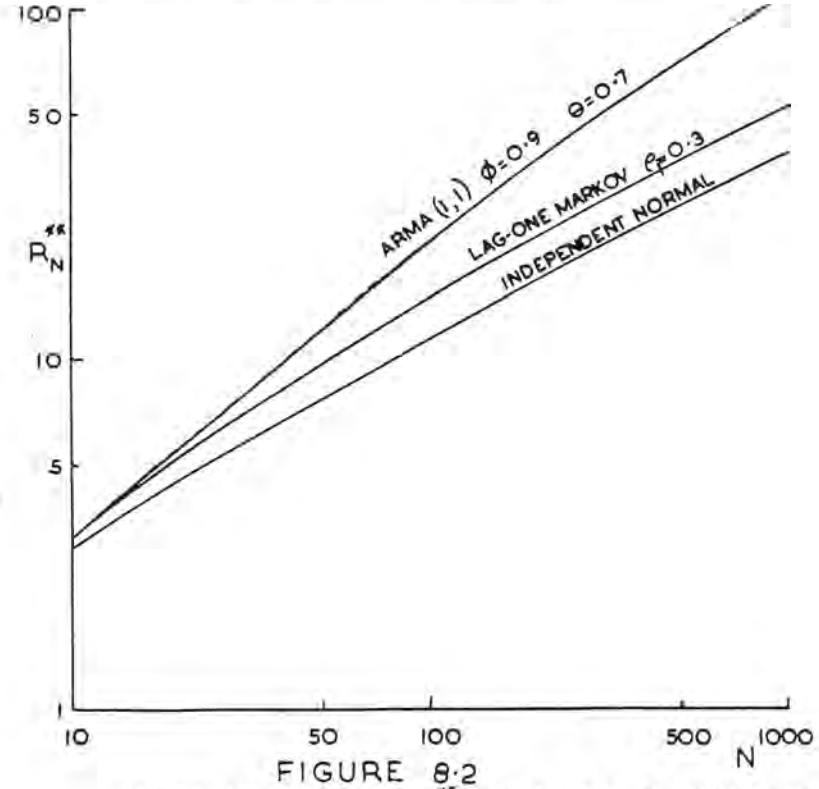

FIGURE 8·2
MEAN VALUES OF R_N^{**} FROM 150 REALISATIONS OF LENGTH N FROM EACH OF THREE THEORETICAL PROCESSES

the slope at large values of n is made difficult by the large fluctuations in R_n^{**}. These are due to the availability of only one independent value when n is greater than half the series length. The impression is gained also that more complex autocorrelation causes the convergence to a slope of 0.5 to occur at larger n. Figure 8.2 confirms these impressions with plots of mean values of R_n^{**} calculated from 150 independent realisations, of length n, of the three processes.

8.3 The Rescaled Adjusted Range for some Observed Series

Figures 8.3 to 8.5 show the results of some of the rescaled adjusted range analyses carried out in this study. Results are presented here for five long data series as follows:

(i) Mud varve thicknesses at Lake Saki in the Crimea. The source for this series is Shostakovitsch (1934). The data analysed covers the period 2290 B.C. to 1889 A.D., a continuous record of 4,180 years. The mud varves are believed to be correlated with annual lake inflows.

(ii) Tree Ring Index - Finland. This series is due to Siren (1961) and the data is reproduced in Lamb (1977). The series covers the period 1181 to 1960, a total length of 700 years. The data is believed to be correlated with summer mean temperature.

(iii) Standardised monthly flow volumes for the Snowy River at Jindabyne N.S.W. (1905-1977). Flows in the post-Snowy Mountains Scheme period have been corrected for regulation on the basis of careful operational water balances.

(iv) Standardised monthly flow volumes for the Macquarie

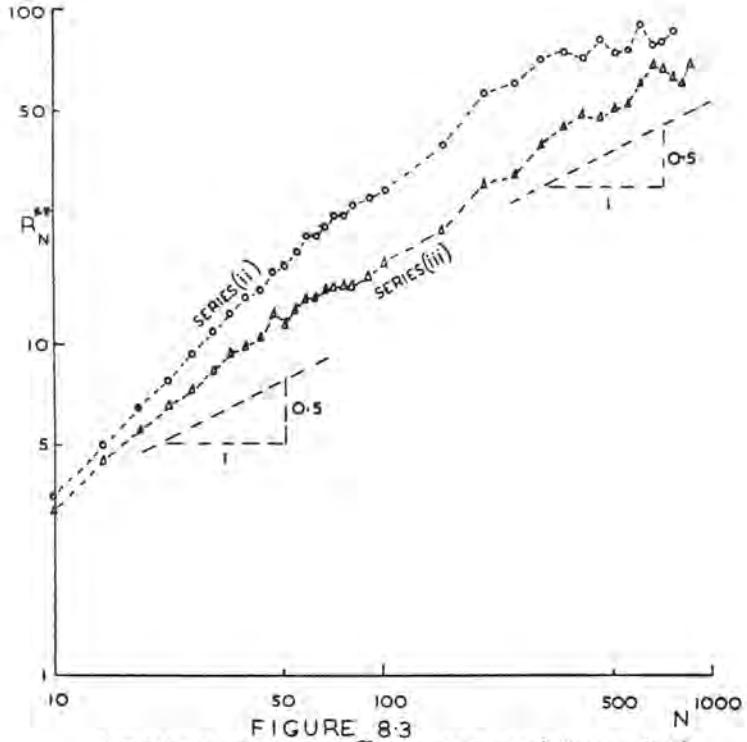

FIGURE 8·3
MEAN VALUES OF R_N^{**} FOR SERIES (ii) AND (iii)
TREE RING INDEX AND SNOWY RIVER FLOWS

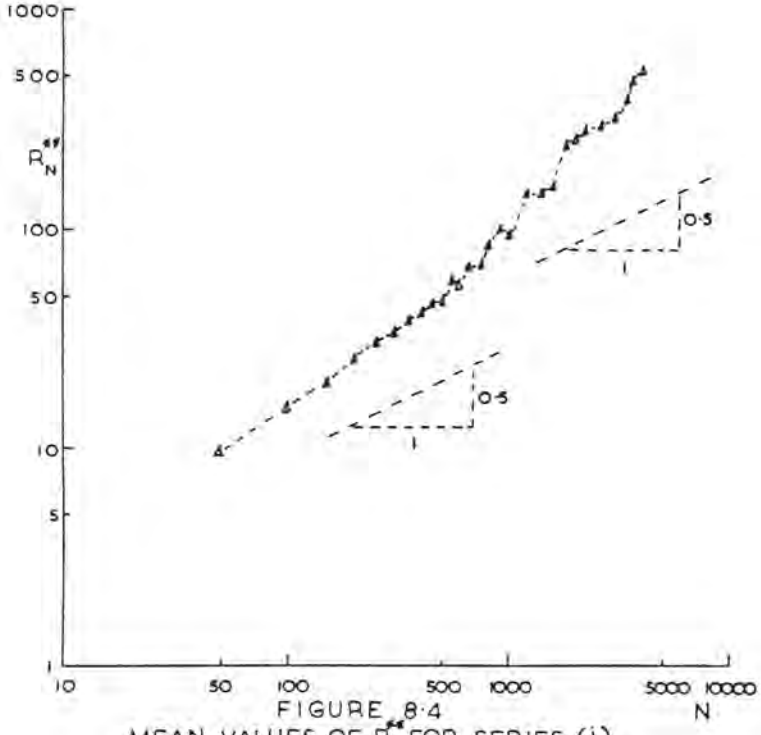

FIGURE 8·4
MEAN VALUES OF R_N^{**} FOR SERIES (i)
LAKE SAKI MUD VARVES

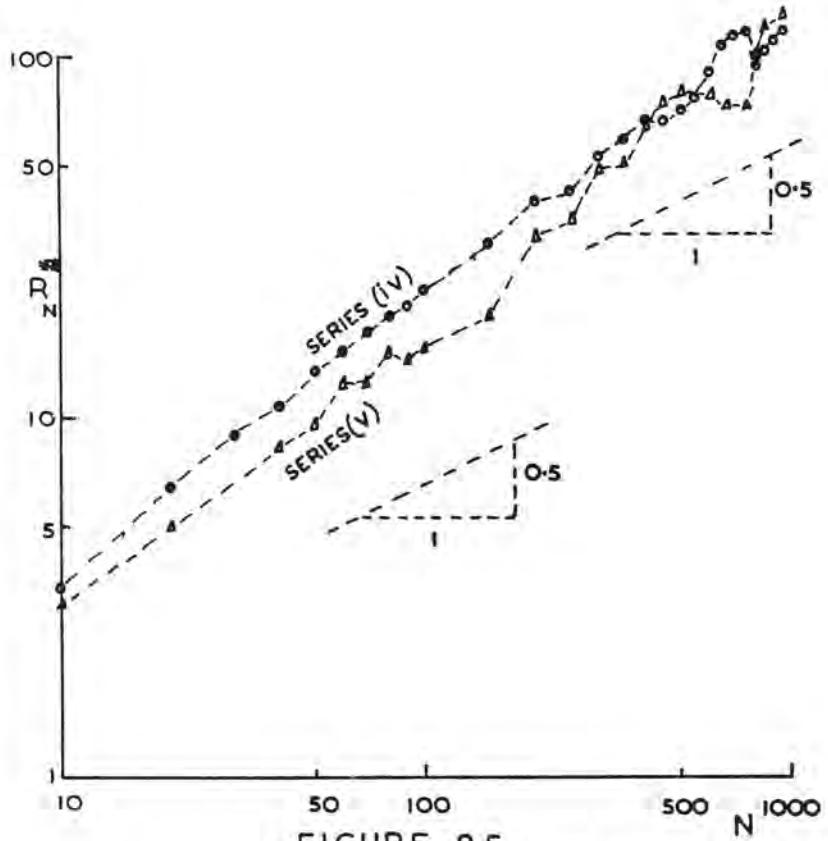

FIGURE 8·5
MEAN VALUES OF R_N^{**} FOR SERIES (iv) AND (v)
MACQUARIE AND KIEWA RIVER FLOWS

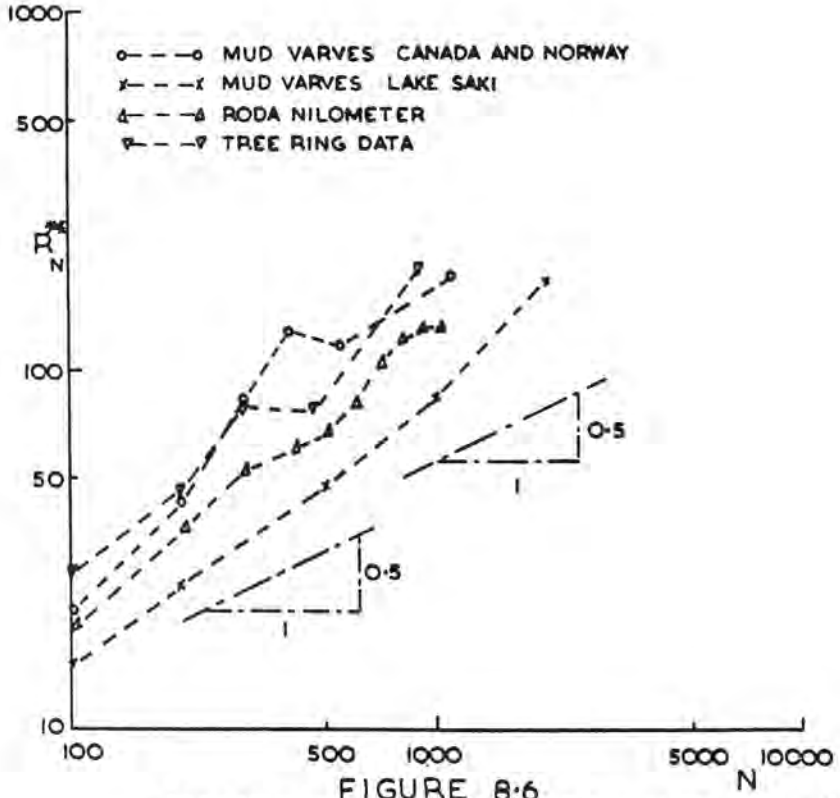

FIGURE 8·6
MEAN VALUES OF R_N^{**} - FROM HURST (1951)

River at Burrendong N.S.W. (1886-1964).

(v) Standardised monthly flow volumes for the Kiewa River at Kiewa VIC. (1893-1970).

The monthly flow series have been rendered stationary as far as their means and variance are concerned by the commonly used process of subtracting monthly means from the data and dividing by monthly standard deviations. This 'standardising' process is an efficient way of removing deterministic periodicities or cyclicities in data.

The results of the analysis of series (ii) and (iii) are shown in Figure 8.3. The two plots give the visual impression of their slope decreasing with n. This result, particularly for the tree ring series, is surprising in the light of the extensive analysis presented by Mandelbrot and Wallis (1969d) who make the following statement:

"Were the records in question generated by a random process such that observations far removed in time can be considered independent, (R_n^{**}) would become asymptotically proportional to ($n^{0.5}$), which means that Hurst's law would have to 'break' for large enough lags. No such break has been observed. Thus for practical purposes, geophysical records must be considered to have an 'infinite' span of statistical interdependence."

It should be pointed out again that the procedure for plotting R_n^{**} values used here differs from that used by Mandelbrot and Wallis (1969d). In this study, when the sub-series length n is greater than half the total available series leaving only one independent sub-series, one determination of R_n^{**} is made. In each case the sub-series starting at the beginning of the record is used.

In their work Mandelbrot and Wallis make several determinations of R_n^{**} by sliding the start of the sub-series forward in the record from the beginning. The resulting values of R_n^{**} are highly correlated as they result from overlapping series and hence exhibit only a small amount of variation.

The analysis of series (i), (iv) and (v) is shown in Figures 8.4 and 8.5. For these series no 'rolling over' of the plot is apparent. A feature of the Lake Saki plot (Fig. 8.4) is the distinct 'break' to a greater slope at about n = 600. Mandelbrot and Wallis (1969d) state that such a break in slope is characteristic of a strongly periodic series. Figure 8.6 shows some of the results obtained by Hurst (1951) which also show no strong indication of convergence to a slope of 0.5.

8.4 Further Comments on the Work of Mandelbrot and Wallis

The weight of the evidence available points to the apparent failure of slopes of log R_n^{**} versus log n plots of long observed series to converge to 0.5. Mandelbrot and Wallis (1969d) carried out a comprehensive analysis of about 70 observed data series of which only five have an apparent Hurst Coefficient H of 0.5. Their analysis is however open to the following comments.

Many of the series analysed by Mandelbrot and Wallis are too short for the theoretical asymptotic behaviour to develop. As an example the reader is referred to Figure 6.1 presented earlier in this report. Figure 6.1 is taken from Mandelbrot and Wallis (1969d) and shows results for the series of annual flows in the St. Lawrence River at Ogdensburg, New York (1860-1957). As part of this study the same data was obtained from the acknowledged source, Yevjevich (1963), and a rescaled adjusted range analysis of it was carried out.

The results are shown in Figure 8.7.

Figure 8.7 also shows a plot of estimated expected values of R_n^{**} derived from a stochastic model. The model is a constrained three-lag autoregressive process. The autoregressive parameter values $\emptyset_1$, $\emptyset_2$ and $\emptyset_3$, are shown on the figure. Such stochastic models will be discussed in detail in the next chapter of this report.

The three-lag autoregressive model was used to generate, by computer, 500 synthetic sequences of length n. The model structure and parameter values were proposed for the St. Lawrence River data by McLeod et al (1977).

In Figure 6.1 Mandelbrot and Wallis contrast the slope of their 'pox' diagram with a line of slope 0.5, the theoretical asymptotic slope for the general class of Gaussian 'short memory' processes. They use the apparent discrepancy in slopes as evidence of the non-Gaussian nature of real data. The slope estimator H is seen to have the high value of 0.98.

Figure 8.7 shows the irrelevance of the comparison of slopes made by Mandelbrot and Wallis. In fact the 'short memory' model generates sequences giving mean values of R_n^{**} which are quite close to the observed values. At the available series length (n = 98 years) of the observed data it is seen that the asymptotic slope of the theoretical process has not yet developed.

Further comments on the work of Mandelbrot and Wallis and the 'Hurst Phenomenon' in general will be made later in this report. The next chapter will deal with the structure of various 'short memory' stochastic processes and how models based on these processes can be fitted to observed data series. The discussion therein will allow further comparisons of theoretical and observed rescaled adjusted range values.

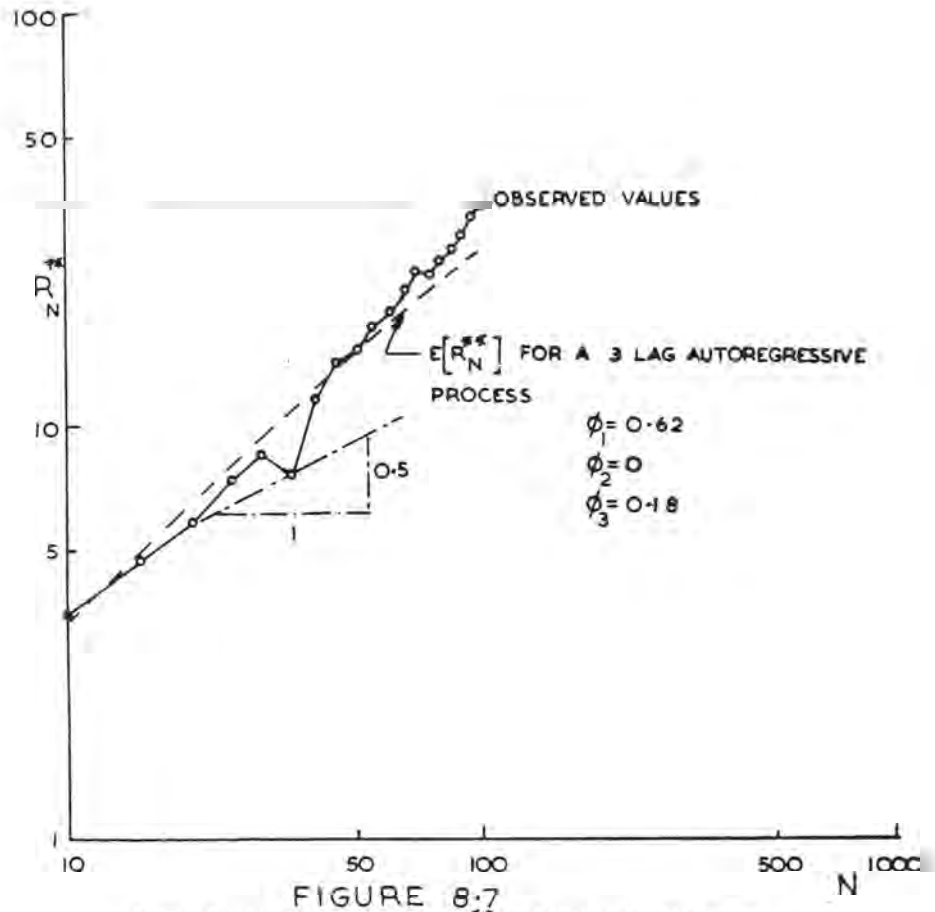

FIGURE 8.7
MEAN VALUES OF R_N^{**} - ANNUAL FLOWS
ST LAWRENCE RIVER AT OGDENSBURG
(1860 - 1957)

CHAPTER 9: SYNTHETIC DATA GENERATION MODELS AND FITTING PROCEDURES

9.1 Introduction

In summary, the areas of discussion covered and the findings in this report so far are as follows:

(i) The definition of the rescaled adjusted range was presented showing the analogy with the residual mass curve approach to storage design. Next, the available analytical results for the rescaled adjusted range and related range statistics for some theoretical processes were presented. The discussion of Hurst's work which followed showed that he found the rescaled adjusted range statistic (R_n^{**}) could be used to reveal complexities of real data series. These complexities he typified with K, the slope of the log-log plot of the rescaled adjusted range versus the series length n.

(ii) A close look at Hurst's methodology and the definition of K, showed K to have a dual character, that of a slope and that of a logarithmic transformation of R_n^{**} at a single point.

Later chapters discussed the approach of some investigators, notably Mandelbrot and Wallis, who concentrate on the slope nature of K replacing it with more refined slope estimators. They see the behaviour of the R_n^{**} slope in long observed series as proof of the 'self-similar' nature of real data - a theory which carries an implication of infinite memory. Fractional Gaussian Noise models are proposed which approximate such an infinite memory process and produce linear log-log slopes. Some reservations were expressed in

the report about this approach and it was pointed out that perhaps a more pressing concern for the hydrologist is the question of whether or not familiar 'short memory' models preserve the observed magnitudes of the rescaled adjusted range.

The remaining part of this report will concentrate on the question of the ability of 'short memory' models to preserve the observed magnitudes of R_n^{**}. Observed data series will be analysed and sample R_n^{**} values compared with those expected from theoretical stochastic processes. It will be shown that such a comparison provides a powerful means of process model evaluation. However, before such an exposition can proceed, it is necessary to examine the structure of various 'short memory' process models and the properties of the R_n^{**} statistic. This will identify the detail in which data generation models must be specified in order to allow valid comparisons between observed and theoretical R_n^{**} values.

The discussion which follows, of autoregressive (Markov) and general autoregressive-moving average (ARMA) models, relies heavily on the general exposition of Box and Jenkins (1970) and because of this will retain most of their notation. Box and Jenkins provide a lucid and unifying treatment of data generation methods which have developed in a rather ad hoc fashion within the field of Hydrology. It seems also that hydrologists do not as yet generally perceive of the familiar Markov models as being members of a wider family of stationary autoregressive-moving average (ARMA) processes.

9.2 Autoregressive Models - General

A general autoregressive process can be described as

$$Z_t = \emptyset_1 Z_{t-1} + -- + \emptyset_p Z_{t-p} + a_t \qquad ---(9.1)$$

where $Z_t = z_t - \bar{z}$ is the deviation of the process from its mean $\bar{z}$. Z_t is therefore the weighted sum of the p past deviations and a random shock a_t, where a_t is a realisation of an uncorrelated random variate with mean zero and constant variance s_a^2. The $\emptyset_i$ are weighting parameters. These processes will be designated as AR(p).

The above is a generalisation of a concept introduced by the Russian mathematician Markov (1856-1922). The term 'Markov model' is commonly applied in the hydrological literature to auto-regressive models, sometimes with the connotation of p = 1; i.e. 'lag-one Markov' models.

Expression (9.1) may be multiplied throughout by Z_{t-k} to obtain

$$Z_{t-k} Z_t = \emptyset_1 Z_{t-k} Z_{t-1} + \emptyset_2 Z_{t-k} Z_{t-2} + ---$$

$$+ \emptyset_p Z_{t-k} Z_{t-p} + Z_{t-k} a_t \qquad ---(9.2)$$

and taking expected values in expression (9.2) gives

$$\gamma_k = \emptyset_1 \gamma_{k-1} + \emptyset_2 \gamma_{k-2} + -- + \emptyset_p \gamma_{k-p} \quad k > 0 \qquad ---(9.3)$$

noting that γ_k is the autocovariance at lag k, that is, $\gamma_k = E\left[(z_t - \bar{z}) (z_{t-k} - \bar{z}) \right]$ and that $E\left[Z_{t-k} a_t \right]$ vanishes when $k > 0$ as Z_{t-k} can only involve the shocks a_j up to the time t-k, which are uncorrelated with a_t.

Dividing expression (9.3) by the variance $\gamma_0 = s_z^2$ gives

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + .. + \phi_p \rho_{k-p} \qquad k > 0 \qquad \text{---(9.4)}$$

where ρ_k = lag k autocorrelation = $\frac{\gamma_k}{\gamma_0}$. Substituting k=1, 2...p in expression (9.4) gives a set of linear equations for $\phi_1, \phi_2 \ldots \phi_p$ in terms of $\rho_1, \rho_2, \rho_3, \ldots \rho_p$ as follows, (noting that $\rho_{-i} = \rho_i$)

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2 \rho_1 + \ldots + \phi_p \rho_{p-1} \\ \rho_2 &= \phi_1 \rho_1 + \phi_2 + \ldots + \phi_p \rho_{p-2} \\ \rho_p &= \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \ldots + \phi_p \end{aligned} \qquad \text{---(9.5)}$$

The expressions (9.5) are known as the Yule-Walker equations and form the basis of the traditional hydrological approach to model fitting. Theoretical autocorrelations ρ_k are replaced by the estimated autocorrelations r_k for an assumed order of process p. The equations may then be solved for $\phi_{1,2 \ldots p}$.

The variance of the process is derived as follows. On taking expected values in expression (9.2) with k = 0, $E[Z_{t-k} a_t] = E[a_t^2] = s_a^2$ since the only part of Z_t which will be correlated with a_t is the most recent shock a_t. Therefore for k = 0,

$$\gamma_0 = \phi_1 \gamma_{-1} + \phi_2 \gamma_{-2} + \ldots + \phi_p \gamma_{-p} + s_a^2,$$

on dividing through by $\gamma_0 = s_z^2$ and substituting $\gamma_k = \gamma_{-k}$ the variance s_z^2 may be written

$$s_z^2 = \frac{s_a^2}{1 - \rho_1 \phi_1 - \rho_2 \phi_2 - \ldots - \rho_p \phi_p} \qquad \text{---(9.6)}$$

9.3 Lag-one Markov Processes

The familiar lag-one Markov (AR(1)) process may be written as

$$Z_t = \phi_1 Z_{t-1} + a_t \qquad \text{---(9.7)}$$

and for p = 1, expression (9.4) becomes

$$\rho_k = \phi_1 \rho_{k-1} \qquad k > 0 \qquad \text{---(9.8)}$$

which with $\rho_0 = 1$ gives

$$\rho_k = (\phi_1)^k \qquad k \geq 0 \qquad \text{---(9.9)}$$

Equation (9.9) gives the theoretical expression for the AR(1) autocorrelation function (correlogram) indicating that the function decays exponentially to zero when ϕ_1 is positive.

Putting k = 1 in expression (9.9) gives

$$\rho_1 = \phi_1 \qquad \text{---(9.10)}$$

Expression (9.6) gives the variance of the process as

$$s_z^2 = \frac{s_a^2}{1 - \rho_1 \phi_1} = \frac{s_a^2}{1 - \rho_1^2} \qquad \text{---(9.11)}$$

The lag-one Markov model proposed by Brittan (1961) is now evident. Equation (9.7) can be expressed as

$$(z_t - \bar{z}) \quad \phi_1 (z_{t-1} - \bar{z}) + s_a v_t \qquad \text{---(9.12)}$$

where v_t is an identically distributed independent random variate of zero mean and unit variance. Now from expression (9.11)

$$s_a = \sqrt{s_z^2 (1 - \rho_1^2)} = s_z \sqrt{1 - \rho_1^2} \qquad ---(9.13)$$

Substituting (9.10) and (9.13) in (9.12) gives

$$z_t - \bar{z} = \rho_1 (z_{t-1} - \bar{z}) + s_z \sqrt{1 - \rho_1^2} \; v_t \qquad ---(9.14)$$

and if process parameters $\bar{z}$, s_z, ρ_1 are replaced by estimates $\bar{z}', s_z', r$ then

$$z_t - \bar{z}' = r_1 (z_{t-1} - \bar{z}') + s_z' \sqrt{1 - r_1^2} \; v_t \qquad ---(9.15)$$

As proposed by Brittan, v_t was a normal independent variate of zero mean and unit variance. The derivation shows that expression (9.15) preserves the process mean and variance for any identically distributed independent variate of zero mean and unit variance.

Fiering (1967) presents the following analysis to show that for skewed series the skewness of the random component and the skewness of the values generated by the process model are related through the correlation structure of the process.

Consider standardised values q_t such that

$$q_t = \frac{z_t - \bar{z}}{s_z} = \frac{Z_t}{s_z} \qquad ---(9.16)$$

Now dividing expression (9.14) by s_z gives

$$q_t = \rho_1 q_{t-1} + \sqrt{1 - \rho_1^2} \; v_t \qquad ---(9.17)$$

and by definition

$$E\left[q_t \right] = E\left[q_{t-1} \right] = 0$$

$$E\left[q_t^2 \right] = E\left[q_{t-1}^2 \right] = 1$$

$$E\left[q_t^3 \right] = E\left[q_{t-1}^3 \right] = \gamma_z$$

$$E\left[v_t \right] = 0$$

$$E\left[v_t^2 \right] = 1$$

$$E\left[v_t^3 \right] = \gamma_v \qquad \text{---(9.18)}$$

where γ_z and γ_v are the series and random component skewnesses respectively.

By cubing expression (9.17)

$$q_t^3 = \rho_1^3 q_{t-1}^3 + 3v_t\sqrt{1-\rho_1^2}\,(\rho_1 q_{t-1})^2 + v_t^3 (1-\rho_1^2)\rho_1 q_{t-1} + v_t^3\sqrt{(1-\rho_1^2)}^3 \qquad \text{---(9.19)}$$

and by taking expectations

$$E\left[q_t^3 \right] = \gamma_z = \rho_1^3 E\left[q_{t-1}^3 \right] + \sqrt{(1-\rho_1^2)}^3 E\left[v_t \right]$$

$$= \rho_1^3 \gamma_z + \sqrt{(1-\rho_1^2)}^3 \gamma_v$$

and

$$\gamma_v = \frac{\gamma_z (1 - \rho_1^3)}{\sqrt{(1 - \rho_1^2)^3}} \qquad ---(9.20)$$

The skewness γ_v of the random component of the process is seen therefore to be modified by serial correlation as in expression (9.20) to give a different skewness γ_z of the process values.

If the series skewness is estimated as γ'_z, then γ'_v can be calculated from (9.20). The following transformation due to Wilson and Hilferty (1931) can be used to convert a normal independent variate n_t with zero mean and unit variance to a random variate v_t which is distributed like the gamma distribution with zero mean, unit variance and skewness γ'_v.

$$v_t = \frac{2}{\gamma'_v}\left[1 + \frac{\gamma'_v n_t}{6} - \frac{\gamma'^2_v}{36}\right]^3 - \frac{2}{\gamma'_v} \qquad ---(9.21)$$

This transformation has been shown by McMahon and Miller (1971) to become unstable for values of γ'_v greater than about two and for such values a modification due to Kirby (1972) should be used.

Other approaches have been taken to generating skewed series. One method is to transform the observed series by taking logarithms. Expression (9.15) is used with moments of the transformed log series and the resulting generated log values are re-transformed by taking anti-logs. Other normalising transformations can be used. There is an inherent drawback in this approach in that moments of the transformed, rather than the original, series are maintained.

Another approach, due to Matalas (1967), assumes a three-parameter log-normal distribution and transforms the parameters $\bar{z}'$, s', r, in expression (9.15) while sampling v_t from a normal distribution.

From the preceding discussion it can be seen that this approach masks the clear functional relationships between the terms in (9.15) and also lacks generality as the transformation of the parameters depends on the assumption of an underlying three-parameter log-normal distribution. On the other hand, expression (9.20) which preserves the third moment is free from an assumption regarding distribution type and hence v_t may be drawn from a range of distributions. The role of a transformation such as that of Wilson-Hilferty (1931) is to provide an algorithm for converting the more easily sampled normal variate to one having the desired distribution of v_t.

9.4 Multi-lag Markov Processes

The approach of 9.2 can be extended to processes containing more autoregressive terms. This is illustrated by the following relationships for a two-lag Markov process. Again following Box and Jenkins (1970), such a process can be written as

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + a_t \qquad \text{---(9.22)}$$

which now has two autoregressive parameters ϕ_1, ϕ_2. The Yule-Walker equations (9.5) become for p = 2

$$\rho_1 = \phi_1 + \phi_2 \rho_1$$

$$\rho_2 = \phi_1 \rho_1 + \phi_2 \qquad \text{---(9.23)}$$

Equation (9.23) can be solved to give

$$\phi_1 = \frac{\rho_1 (1 - \rho_2)}{1 - \rho_2}$$

$$\phi_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \qquad \text{---(9.24)}$$

From expression (9.6) the variance of the process is given by

$$s_z^2 = \frac{s_a^2}{1 - \rho_1 \phi_1 - \rho_2 \phi_2} \qquad \text{---(9.25)}$$

In practice ρ_1 and ρ_2 are replaced by estimates r_1 and r_2.

9.5 ARMA (Autoregressive Moving Average Models)

Box and Jenkins (1970) describe a wider class of stationary linear stochastic processes which they label autoregressive moving average (ARMA). These processes have the general form

$$Z_t = \phi_1 Z_{t-1} + .. + \phi_p Z_{t-p} + a_t - \theta_1 a_t -..- \theta_q a_{t-q} \qquad (9.26)$$

Z_t is seen here to be the sum of a random shock and weighted sum of previous values of Z and previous values of random shocks a_i. The a_i are as before identically distributed with zero mean and constant variance s_a^2. The above process is described as ARMA (p,q).

If all the θ_i are zero the process is pure autoregressive of order p, i.e. AR(p) or ARMA(p,0). If all the ϕ_i are zero, then the process is pure moving average of order q, i.e. MA(q) or ARMA(0,q).

The theoretical autocorrelation function (correlogram) of pure MA processes truncates after q lags while that of AR processes attenuates. The mixing of AR and MA terms provides a flexible modelling tool for preserving observed autocorrelation functions.

The ARMA (1,1) process has received considerable attention in the hydrology literature and can be written as

$$Z_t = \phi_1 Z_{t-1} + a_t - \theta_1 a_{t-1} \qquad ---(9.27)$$

Such a process can be seen as a simple extension of the Markov lag-one case. The process 'remembers' not only the previous value but the previous random disturbance.

The following relationships are obtained for the ARMA (1,1) process by an approach similar to that for the pure AR process.

$$\rho_1 = \frac{(1 - \phi_1 \theta_1)(\phi_1 - \theta_1)}{1 + \theta_1^2 - 2\phi_1\theta_1} \qquad ---(9.28)$$

$$\rho_2 = \phi_1 \rho_1$$

Values of ϕ_1 and θ_1 have the following limitations

$$-1 < \theta < +1 \qquad ---(9.29)$$

$$-1 < \phi < +1$$

which meet stationarity and 'invertability' conditions defined by Box and Jenkins (1970).

The parameters ϕ_1, θ_1 of an ARMA (1,1) process may be estimated by replacing ρ_1, ρ_2 with estimates r_1, r_2 in expression (9.28).

The series and random component variances are related as follows:

$$s_z^2 = \frac{1 + \theta_1^2 - 2\phi_1\theta_1}{1 - \phi_1^2} \, s_a^2 \qquad ---(9.30)$$

Srikanthan and McMahon (1977) present an expression for the random component skewness of an ARMA (1,1) process.

$$\gamma_v = \left[\left[\frac{1 + \theta_1^2 - 2\theta_1\phi_1}{1 - \phi_1^2}\right]^{\frac{3}{2}}\left[\frac{1 - \theta^3 + 3\phi_1\theta_1^2 - 3\phi_1^2\theta_1}{1 - \phi_1^3}\right]\right]\gamma_z \quad (9.31)$$

The reader is referred to Nelson (1973) who gives a quite readable account of ARMA processes and model fitting.

9.6 Model Fitting

The hydrological approach to determining data generation model parameters, such as ϕ_1, ϕ_2 in an AR(2) model, is generally to solve the Yule-Walker equations (9.5). The appropriate number p of autoregressive terms is assumed and sample autocorrelation estimates (e.g. r_1, r_2) inserted in the equations. The variance s_a^2 of the random component is estimated from the sample series variance and a relationship such as expression (9.30). This approach which relies on sample moments and autocorrelations will be referred to as the 'moment-estimation' procedure.

The moment-estimation procedure is heavily dependent on the estimated values r_1, r_2 .... r_p. These sample autocorrelations have considerable uncertainty about them as is shown in Figure 9.1 where the sample autocorrelations calculated from a set of 100 values of an independent normal variate are shown. The theoretical value of the autocorrelation coefficient is of course zero at all lags for an independent process. The point is further illustrated by Figure 9.2, which shows a sequence of 100 values generated by a lag-one Markov process with a theoretical lag-one autocorrelation coefficient of 0.15.

Box and Jenkins (1970) present an alternative to the moment-

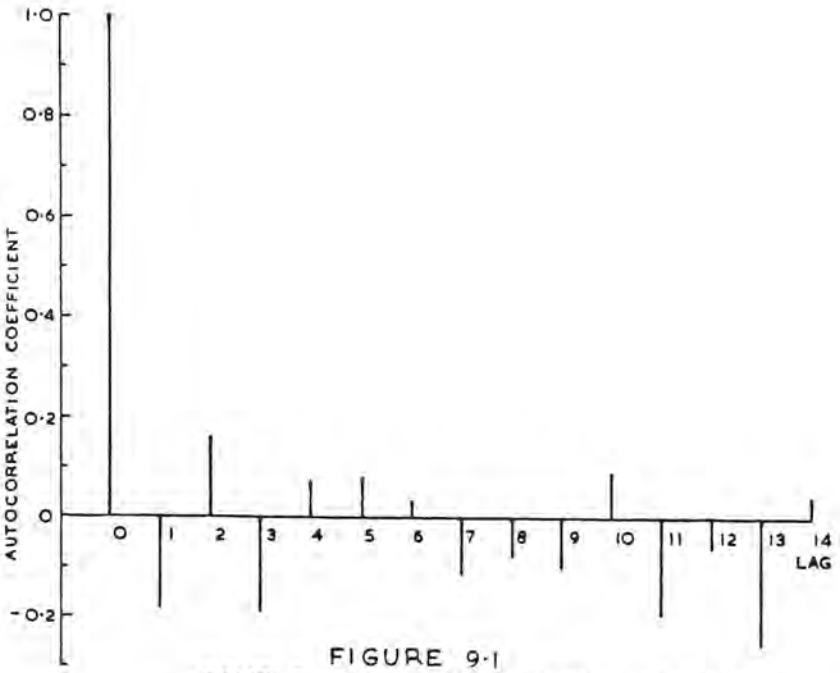

FIGURE 9·1
SAMPLE AUTOCORRELATION FUNCTION - 100 VALUES OF AN INDEPENDENT NORMAL VARIATE

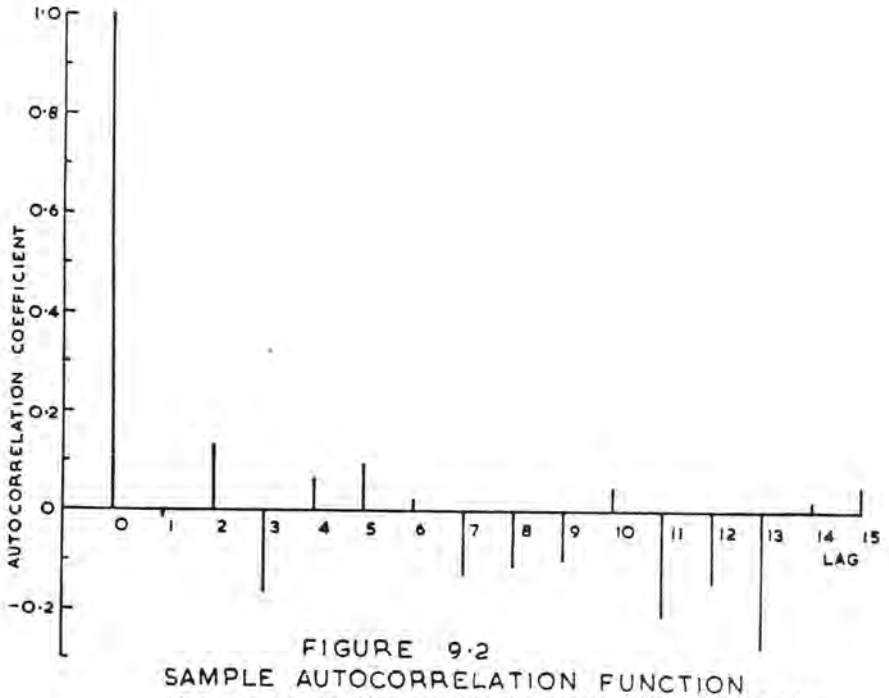

FIGURE 9·2
SAMPLE AUTOCORRELATION FUNCTION
100 VALUES OF A LAG-ONE MARKOV PROCESS
$\rho_1 = 0{\cdot}15$

estimation type approach to model fitting. They propose a systematic method of analysis consisting of three distinct phases

(i) Identification

(ii) Estimation

(iii) Diagnostic Checks

In phase (i) the model structure is inferred from the sample autocorrelation function (correlogram) and the sample partial autocorrelation function. These sample functions are compared with the theoretical functions for AR, MA or ARMA models. The theoretical partial autocorrelation function comes from the successive solution of the Yule-Walker equations (9.6) for an increasing number of lags k. The value of the partial autocorrelation function at lag k is ϕ_{kk}, the last parameter of a pure autoregressive process if it was of the order p = k. For a pure autoregressive process of lag p the theoretical partial autocorrelation function truncates at p. Figures 9.3 and 9.4 show autocorrelation and partial autocorrelation functions from a sample of 900 values of a lag-one Markov process with $\rho_1 = 0.5$.

McLeod and Hipel (1977) propose two other functions, the 'inverse autocorrelation function' and 'inverse partial autocorrelation function', as an aid to model identification.

In phase (ii) the parameters of the proposed model are estimated by an approximate maximum likelihood method in which a 'sum-of-squares' function is minimised. For example the AR(1) model given by expression (9.7) can be written as

$$a_t = Z_t - \phi_1 Z_{t-1} \qquad \text{---(9.32)}$$

and if the Z_t are replaced by a set of observed values Z'_t then

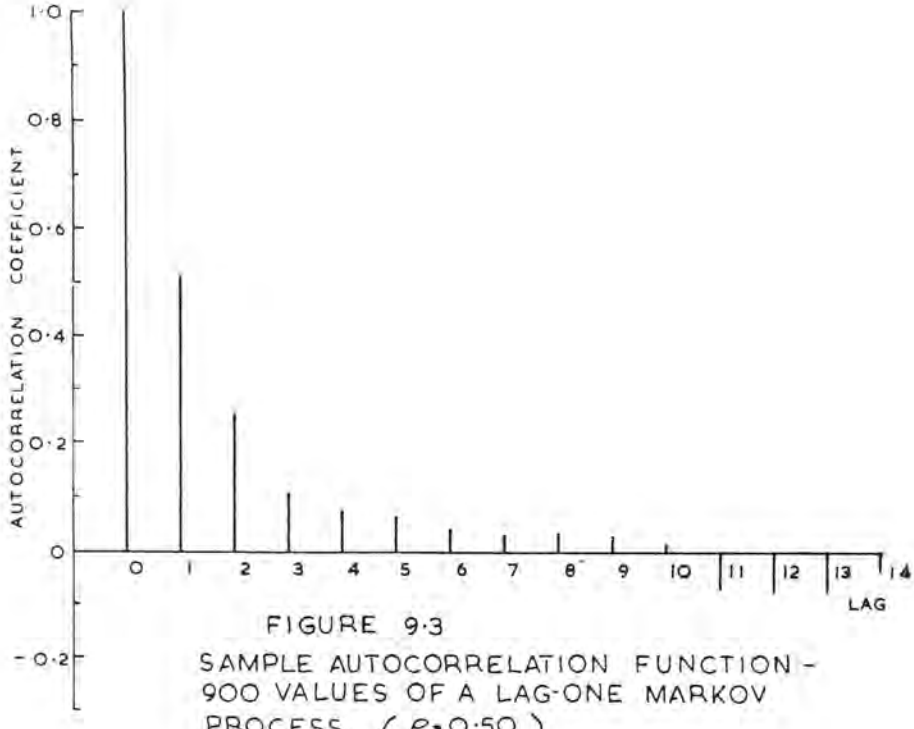

FIGURE 9·3
SAMPLE AUTOCORRELATION FUNCTION - 900 VALUES OF A LAG-ONE MARKOV PROCESS ($\rho_1 = 0.50$)

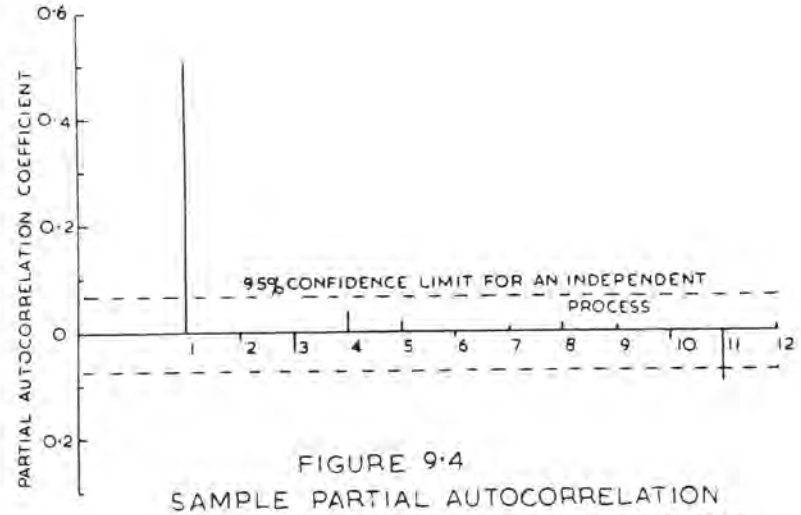

FIGURE 9·4
SAMPLE PARTIAL AUTOCORRELATION FUNCTION - 900 VALUES OF A LAG ONE MARKOV PROCESS ($\rho_1 = 0.50$)

$$a'_t = Z'_t - \phi_1 Z'_{t-1} \qquad \text{---(9.33)}$$

If some starting value Z'_0 is nominated then the set of residuals for the assumed value of ϕ_1, i.e. $a'(\phi_1)_t$, can be calculated from

$$a'(\phi_1)_1 = Z'_1 - \phi_1 Z'_0$$

$$a'(\phi_1)_2 = Z'_2 - \phi_1 Z'_0 \quad \text{etc.} \qquad \text{---(9.34)}$$

In this case the sum of squares function is

$$S(\phi_1) = \sum_{t=1}^{T} \left[Z'_t - \phi_1 Z'_{t-1} \right]^2 = \sum_{t=1}^{T} \left[a'(\phi_1)_t \right]^2 \qquad (9.35)$$

Minimising $S(\phi_1)$ leads to an approximate maximum likelihood estimate ϕ_1. The maximum likelihood estimate of the variance of the residuals s_a^2 is obtained from

$$s_a^2 = \frac{S(\phi'_1)}{T} \qquad \text{---(9.36)}$$

The procedure of minimising the sum of squares function involves multivariate search in the case of a proposed process with more than one autocorrelation parameter e.g. ARMA (1,1). Starting estimates for the search are usually determined by the moment-estimation procedure. (Nelson - 1973)

The maximum likelihood estimates of parameters have been shown to be insensitive to lack of normality in the residuals (McLeod, 1974). McLeod and Hipel (1977) discuss refinements to the maximum likelihood estimating procedure.

Phase (iii) involves the examination of the residuals a'_t determined using the estimated model parameters. The residuals are checked for the assumptions of independence, normality and constant variance. The most important assumption is the independence of the residuals for which several statistical tests are available (McLeod and Hipel 1977). The latter two features are less important and it is suggested that transformations of the original data may rectify problems in this area. If the residuals prove to be not independent, then the model structure should be revised by removing or adding AR or MA terms and the analysis repeated.

In this chapter the structure of various 'short-memory' stochastic process models has been examined in some detail. In the next chapter attention will be turned to the effect on the rescaled adjusted range, in series produced by such process models, of the model structures and parameters.

CHAPTER 10: SOME PROPERTIES OF THE RESCALED ADJUSTED RANGE OF THEORETICAL PROCESSES

10.1 Introduction

The present point of interest is the ability of 'short memory' data generation models to preserve the size of the rescaled adjusted range observed in real data series. This question can be approached by fitting models to observed series and comparing the theoretical and observed rescaled adjusted range.

The structure of a general class of 'short memory' models was discussed in Chapter 9. The next logical step is to see in how much detail a proposed model has to be specified if the comparison of theoretical and observed rescaled adjusted range values is to be valid. For instance, is the rescaled adjusted range affected by the assumed distribution of the random component of the model? If the latter were the case, then attention would have to be given to correctly inferring the marginal distribution of the observed series. Investigators have in fact examined skewness as a possible explanation of the 'Hurst Phenomenon'.

A FORTRAN routine, DRSGE, was written to assist in this investigation (see appendix). DRSGE calculates means and standard deviations of R_n^{**} and K at nominated values of series length n and for a nominated number of independent sequences. The programme simulates AR(p) and ARMA (1,1) processes. In the discussion that follows, means and standard deviations have been sampled from a maximum of 500 sequences in order to keep computing time to within reasonable limits.

10.2 The Effect of Process Mean and Variance on the Rescaled Adjusted Range

Given a series of values z_t, $t = 1,2....n$, the adjusted range for series length n is formed from the series $z_t - \bar{z}$ where $\bar{z}$ is the mean value of the z_t, that is

$$\bar{z} = \left(\sum_{t=1}^{n} z_t \right) / n \qquad ---(10.1)$$

The rescaled adjusted range (R_n^{**}) is formed from the series

$$q_t = \frac{z_t - \bar{z}}{s_z} \quad t = 1, 2....n \qquad ---(10.2)$$

where

$$s_z = \left[\frac{\sum_{t=1}^{n} (z_t - \bar{z})^2}{(n - 1)} \right]^{\frac{1}{2}} \qquad ---(10.3)$$

R_n^{**} is obtained directly from the series

$$Q_t = \sum_{i=1}^{t} q_i \quad t = 1, 2....n \qquad ---(10.4)$$

Since $\bar{z}$ and s_z are the mean and standard deviation of the sub-series z_t, $t = 1, 1....n$, the q_t are standardised estimates of z_t and therefore q_t itself will have zero mean and unit variance.

The standardisation inherent in R_n^{**} means, loosely speaking, that the statistic is independent of the underlying process mean and variance. This is a useful property when comparing observed values of R_n^{**} with theoretical values derived by some proposed process as is shown below.

For example, taking the lag-one Markov process specified by

expression (9.14)

$$z_t - \bar{z} = \rho_1 (z_{t-1} - \bar{z}) + s_z\sqrt{1 - \rho_1^2}\, v_t$$

and dividing both sides by s_z gives

$$\frac{z_t - \bar{z}}{s_z} = \rho_1 \frac{(z_{t-1} - \bar{z})}{s_z} + \sqrt{1 - \rho_1^2}\, v_t$$

that is

$$q_t = \rho_1 q_{t-1} + \sqrt{1 - \rho_1^2}\, v_t \qquad \text{---(10.5)}$$

Expression (10.5) is a lag-one Markov process producing standardised values q_t.

Because R_n^{**} is independent of the process mean and variance, expected values of R_n^{**} and its higher moments derived from series generated by (9.14) and (10.5) will be identical.

A further simplification can be made by noting from the discussion in section 9.3 that the role of the term $\sqrt{1 - \rho_1^2}$ in expression (10.5) is to scale the random component (v_t), which has unit variance, so that the process variance equals unity. Dividing (10.5) by $\sqrt{1 - \rho_1^2}$ gives

$$q_t^* = \rho_1 q_{t-1}^* + v_t \qquad \text{---(10.6)}$$

where $q_t^* = q_t / \sqrt{1 - \rho_1^2}$

Now for a lag-one Markov model expression (9.11) gives the relationship between the variance of the process s_z^2 and the variance of the

random component s_a^2 as follows

$$s_z^2 = \frac{s_a^2}{1 - \rho_1^2}$$

Substituting $s_a^2 = 1$ in expression (9.11) gives, for the series q_t^*, the variance

$$s^2(q_t^*) = \frac{1}{1 - \rho_1^2} \qquad ---(10.7)$$

Therefore the variate q_t^* defined in expression (10.6), while retaining a zero mean, has a variance of $\frac{1}{1 - \rho_1^2}$ which is finite and greater than unity for $0 < \rho_1 < 1$ and $-1 < \rho_1 < 0$.

One may wish to test the validity of assuming that an observed series can be modelled by a lag-one Markov process. If interest lies in the ability of such a process to reproduce observed R_n^{**} values, then it is sufficient to examine the values produced by the simplified model form of expression (10.6). This result applies generally and hence process models used for this purpose may be simplified into a form which produces a variate having zero mean and some finite variance greater than unity.

It is now necessary to examine the effect of the distribution of the random component (v_t) on the rescaled adjusted range.

10.3 The Effect of Skewness on the Expected Value of the Rescaled Adjusted Range

The skewed nature of real data was seen as one possible explanation of the 'Hurst Phenomenon', a question which has received considerable attention in the literature.

Feller (1951) showed rigorously that the asymptotic value

of the rescaled adjusted range is the same for any independent variate having finite variance. Langbein (1956) argued that skewness should not affect the expected value of the rescaled adjusted range and hence could not provide an explanation of the Hurst Phenomenon. Matalas and Huzen (1967) examined this question by computer experiment involving lag-one Markov processes of varying ρ_1. They found that the expected values of R_n^{**} at varying series lengths remained virtually unchanged when the distribution of the random component was changed from normal to log normal. Values of process skewness up to 2.0 were considered.

Mandelbrot and Wallis (1969c) reported on the distribution-free nature of the expected value of the rescaled adjusted range and declared it to be a very robust statistic. Their results however are subject to the computer programming error reported by Taqqu (1970). McLeod and Hipel (1978a) showed that expected values of the rescaled adjusted range do not vary substantially from one distribution to another. They examined independent variates with the normal, gamma, stable and cauchy distributions.

Tables 10.1, 10.2 and 10.3 present the results of some computer simulation experiments carried out using the FORTRAN routine DSRGE developed for this study (see Appendix). Mean values of R_n^{**} are shown for a lag-one Markov process and an ARMA (1,1) process. The mean values were obtained from 500 independent series generated by the process model. Results are shown for the case of a normally distributed random component and for the random component distributed like gamma. Skewnesses of the random component were selected using expressions (9.20) and (9.31) to give series skewness values of 2.0, 4.0 and 6.0. Kirby's (1972) modification of the Wilson-Hilferty transformation was used to generate the skewed random component values.

TABLE 10.1

ESTIMATED VALUES OF $E[R_n^{**}]$ FOR A LAG-ONE MARKOV PROCESS ($\rho_1 = 0.5$) - Sample Size 500

Series length n	$E[R_n^{**}]$			
	Normal	$\gamma_z = 2.0$	$\gamma_z = 4.0$	$\gamma_z = 6.0$
20	5.88	5.89	5.97	5.98
40	9.64	9.75	9.63	9.69
60	12.67	12.85	12.71	12.67
80	15.49	15.41	15.24	15.18
100	17.65	17.63	17.41	17.46

The values shown in Tables 10.1 and 10.3 at each value of n differ from each other by less than one or two standard errors of estimate. The approximate value of the standard error of estimate is, for example, 0.15 in the case of n = 100, $\gamma_z = 6.0$ and the lag-one Markov process. The differences between values show also, no consistent trends with changes in n or γ_z.

Table 10.2 shows a similar analysis for series lengths up to n = 1000. In this case sampling is limited to 150 independent sequences. The differences in values are again not significant for the smaller sample size. The approximate standard error of estimate for n = 1000 and $\gamma_z = 4.0$ is 1.23.

The results shown in Tables 10.1, 10.2 and 10.3 confirm that the expected value of the rescaled adjusted range is quite unaffected by process skewness.

TABLE 10.2

ESTIMATED VALUES OF $E\left[R_n^{**} \right]$ FOR A LAG-ONE MARKOV PROCESS

($\rho_1 = 0.5$) - Sample Size 150

Series length n	$E\left[R_n^{**} \right]$	
	Normal	$\gamma_z = 4.0$
200	26.24	26.28
400	38.58	39.41
600	47.58	49.05
800	55.13	58.10
1000	63.39	64.47

TABLE 10.3

ESTIMATED VALUE OF $E\left[R_n^{**} \right]$ FOR AN ARMA (1,1) PROCESS

($\phi_1 = .9$, $\theta_1 = .7$) - Sample Size 500

Series length n	$E\left[R_n^{**} \right]$	
	Normal	$\gamma_z = 4.0$
20	5.37	5.44
40	9.96	9.81
60	14.25	14.00
80	18.16	17.84
100	21.79	21.65

10.4 The Effect of Skewness on the Standard Deviation of the Rescaled Adjusted Range

It will be shown in the next chapter that plots of expected values of R_n^{**} and confidence limits obtained from the standard deviations of R_n^{**} i.e. $s(R_n^{**})$ provide a powerful tool for model evaluation. To the writer's knowledge the effect of process skewness on $s(R_n^{**})$ has not been discussed in detail in the literature. Strictly speaking, $s(R_n^{**})$ refers to estimated values of the standard deviation. Exact values have not been derived for other than the case of an independent normal variate with n = 3 and n = 4. (Anis and Lloyd - 1977).

The problem of a general expression for the exact value of the standard deviation of R_n^{**} for larger values of n or for dependent variates is one of great difficulty and has not yet been solved. (E. H. Lloyd 1978 - private communication).

Values of $s(R_n^{**})$ determined by computer simulation for independent processes distributed like normal, gamma, stable and cauchy, may be inferred from Table 5 of McLeod and Hipel (1978). These authors present estimates of the expected values $E\left[R_n^{**} \right]$ together with standard errors of estimates of $E\left[R_n^{**} \right]$. These standard errors can be converted to $s(R_n^{**})$ by multiplying by $\sqrt{N} = 100$ where N is the sample size of 10^4. The inferred values of $s(R_n^{**})$ from this source indicate that there is little variation between the different distributions. A slight reduction in $s(R_n^{**})$ is apparent in the gamma case compared with the normal distribution.

Tables 10.4, 10.5, 10.6 and 10.7 show the results of the computer simulation experiments carried out in this study to determine values of $s(R_n^{**})$. The skewed random components are distributed like gamma and are obtained as described in Section 10.3.

TABLE 10.4

ESTIMATED VALUES OF $s(R_n^{**})$ FOR AN INDEPENDENT VARIATE

Sample Size 500

Series length n	$s(R_n^{**})$				
	Normal (McLeod & Hipel)	Normal	$\gamma_z = 2.0$	$\gamma_z = 4.0$	$\gamma_z = 6.0$
20	1.00	.99	.96	.84	.75
40	1.58	1.56	1.53	1.32	1.25
60	1.98	1.84	1.87	1.78	1.63
80	2.33	2.35	2.21	2.18	1.90
100	2.62 (Sample of 10^4)	2.74	2.60	2.50	2.25

TABLE 10.5

ESTIMATED VALUES OF $s(R_n^{**})$ FOR A LAG-ONE MARKOV PROCESS

($\rho_1 = 0.5$) - Sample Size 500

Series length n	$s(R_n^{**})$			
	Normal	$\gamma_z = 2.0$	$\gamma_z = 4.0$	$\gamma_z = 6.0$
20	1.19	1.10	0.99	0.86
40	2.14	1.97	1.82	1.59
60	2.83	2.81	2.59	2.29
80	3.50	3.45	3.07	2.88
100	4.09	3.92	3.39	3.37

TABLE 10.6

ESTIMATED VALUES OF $s(R_n^{**})$ FOR A LAG-ONE MARKOV PROCESS

($\rho_1 = 0.5$) - Sample Size 150

Series length n	$s(R_n^{**})$	
	Normal	$\gamma_z = 4.0$
200	5.80	5.39
400	8.43	9.13
600	10.73	11.59
800	12.69	13.48
1000	13.65	15.03

TABLE 10.7

ESTIMATED VALUES OF $s(R_n^{**})$ FOR AN ARMA (1,1) PROCESS

($\phi_1 = .9$, $\theta_1 = .7$) - Sample Size 500

Series length n	$s(R_n^{**})$	
	Normal	$\gamma_z = 4.0$
20	1.22	1.13
40	2.48	2.24
60	3.80	3.45
80	4.98	4.56
100	6.01	5.30

Standard errors of the estimates of $s(R_n^{**})$ in Tables 10.4, 10.5, 10.6 and 10.7 have not been calculated and hence the significance of the differences between values for the same series length can not be determined. There are however some apparent trends. In Tables 10.4, 10.5 and 10.7, there is an apparent reduction in $s(R_n^{**})$ as the skewness increases. There is an apparent opposite trend in the results shown in Table 10.6 which may result however, from sampling error due to the smaller sample size.

This study of the effect of process skewness on $s(R_n^{**})$ has certainly not been an exhaustive one but it does appear that the standard deviation of the rescaled adjusted range is only slightly affected by process skewness. As an example, the coefficient of variation, which is the ratio of the standard deviation to the expected value, changes from 0.23 to 0.19 for the lag-one Markov case with n = 100 and skewness changing from zero to 6.0.

10.5 Distribution of the Rescaled Adjusted Range

Mandelbrot and Wallis (1969b p.253) state that the distribution of the rescaled adjusted range is 'markedly skew'. Their Figure 11, showing results obtained by computer simulation, indicates a median value of R_n^{**} of about 11 at n = 100 for the normal independent variate. The authors compare this with a value of 12.5 obtained from Feller's (1951) asymptotic expression and conclude that the difference between the mean and median is indicative of high skewness. The comparison is however not valid as the exact expected value from Anis and Lloyd (1976) is 11.45. In fact this tendency to compare observed values of the rescaled adjusted range or Hurst's coefficient K with expected values obtained from expressions which are only asymptotically valid, pervades much of the literature

relating to the 'Hurst Phenomenon'. Hurst (1951) himself made this error.

Wallis and O'Connel refer to the above-mentioned Mandelbrot and Wallis conclusion stating - "The distribution of R/s (i.e. R_n^{**}) is known to be highly skewed when n is small." They go on to state that in using the rescaled adjusted range statistic as a test of statistical independence it is insufficient to know its expected value and standard deviation. They suggest that a knowledge of the whole empirical distribution is required and then proceed to produce this with an extensive computer simulation experiment. Sen(1977a) also refers to Mandelbrot and Wallis (1969b) and states that the distribution of the rescaled adjusted range is highly skewed. Hipel and McLeod (1978a - microfiche version) present the full empirical cumulative distribution functions for the rescaled adjusted range for lag-one Markov models with ρ_1 varying from 0.1 to 0.9 and various series length.

Anis and Lloyd (1977) have obtained the exact distribution of R_n^{**} for the case of an independent normal variate with series lengths n = 3 and n = 4. The distributions they obtained were complex and highly skewed. It has been found in this study however, that in all cases examined using computer simulation and with n greater than 20, the distribution of R_n^{**} is only moderately skewed. The skewness is generally insufficient to invalidate making statistical inferences on the basis of a knowledge of expected values and standard deviations. Table 10.8 shows the calculated skewness of R_n^{**} for several different processes.

The values of the skewness of R_n^{**} shown in Table 10.8 exhibit some variation due to the relatively small sample size of 500. However they suffice to show the moderate degree of skewness

TABLE 10.8

ESTIMATED SKEWNESS OF R_n^{**} FOR VARIOUS PROCESSES

Sample Size 500

Series length n	Estimated Skewness R_n^{**} - $\gamma(R_n^{**})$				
	Independent normal	Lag-one Markov $\rho_1 = 0.5$	ARMA(1,1) $\phi_1 = .82$ $\theta_1 = .40$	Independent gamma $\gamma_z = 4.0$	Skewed lag-one $\rho_1 = 0.5$ $\gamma_z = 2.0$
20	.50	.01	-.25	.53	-.15
40	.50	.22	.01	.56	.03
60	.49	.27	.15	.59	.28
80	.66	.33	.18	.68	.38
100	.68	.50	.26	.41	.28

evident. It appears that as the complexity of the correlation structure of the process increases, the skewness tends to decrease.

The following check was carried out on the loss of accuracy likely to result from using the assumption of normality when making inferences about the rescaled adjusted range. Hipel and McLeod (1978a - microfiche version) present tables derived from a sample size of 10^4, giving the full empirical cumulative distribution functions of the rescaled adjusted range for lag-one Markov processes with normally distributed random components. They present values for processes with ρ_1 ranging from 0.1 to 0.9 and series length n up to 200. For the case of $\rho_1 = 0.5$, values of R_n^{**} at the 84.13 percentile and 15.87 percentile were interpolated from the table. These percentiles correspond to the mean plus or minus one standard deviation on the assumption of the normality of R_n^{**}. The interpolated values were

TABLE 10.9

TESTING THE ERROR AT ONE STANDARD DEVIATION IN ASSUMING R_n^{**} TO BE NORMALLY DISTRIBUTED - Lag-one Markov Process ($\rho_1 = 0.5$)

Series length n	Interpolated from Hipel & McLeod ('78a)		Computer simulation Sample size of 500	
	84.13 percentile R_n^{**}	15.87 percentile R_n^{**}	mean + std. deviation R_n^{**}	mean - std. deviation R_n^{**}
20	7.34	4.71	7.50	4.67
40	12.14	7.63	11.86	7.46
60	15.87	9.88	15.62	9.97
80	19.15	11.78	18.81	11.77
100	21.79	13.50	21.62	13.39

then compared with the experimental results obtained in this study from a much smaller sample of 500. Table 10.9 shows the results obtained.

A similar exercise was carried out comparing the values of R_n^{**} at the 97.72 and 2.28 percentiles with values of R_n^{**} at plus or minus two standard deviations from the mean. The results are shown in Table 10.10.

Table 10.11 shows the differences between mean and median values of the rescaled adjusted range from three lag-one Markov processes. The values are taken from McLeod and Hipel (1978) and Hipel and McLeod (1978a).

A visual impression of the near-normality of the distribution of R_n^{**} is obtained from Figure 10.1. This figure shows a computer line-printer plot of relative frequencies for 1000 values of R_n^{**} at n = 50 for an independent normal variate. The slight positive

TABLE 10.10

TESTING THE ERROR AT TWO STANDARD DEVIATIONS IN ASSUMING R_n^{**} TO BE NORMALLY DISTRIBUTED - Lag-one Markov Process ($\rho_1 = 0.5$)

Series length n	Interpolated from Hipel & McLeod ('78a)		Computer simulation Sample size of 500	
	97.22 percentile R_n^{**}	2.28 percentile R_n^{**}	mean + 2x std deviation R_n^{**}	mean -2x std deviation R_n^{**}
20	8.26	3.66	8.20	3.50
40	14.17	6.01	14.06	5.26
60	18.86	7.79	18.43	7.15
80	23.06	9.29	22.33	8.25
100	26.47	10.71	25.74	9.28

TABLE 10.11

MEAN AND MEDIAN VALUES OF R_n^{**} FOR THREE LAG-ONE MARKOV PROCESSES

Values obtained from McLeod & Hipel (1978) and Hipel & McLeod (1978a)

Sample Size 10,000

Series length n	$\rho_1 = 0.3$		$\rho_1 = 0.5$		$\rho_1 = 0.7$	
	mean	median	mean	median	mean	median
20	5.43	5.36	6.04	6.05	6.73	6.84
40	8.50	8.35	9.87	9.76	11.66	11.68
60	10.87	10.66	12.85	12.65	15.63	15.58
80	12.92	12.65	15.43	15.18	19.12	18.92
100	14.62	14.31	17.60	17.29	22.07	21.82

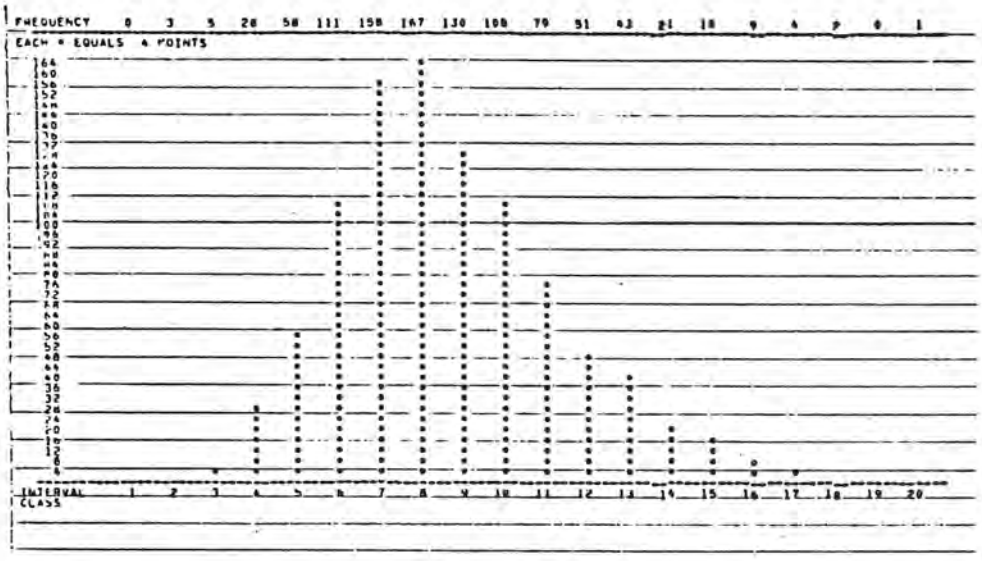

FIGURE 10·1
RELATIVE FREQUENCY DIAGRAM
1000 VALUES OF R_N^{**} AT N=50 FOR
AN INDEPENDENT NORMAL VARIATE

skewness is apparent.

The preceding discussion shows that it should be generally sufficient for practical purposes to assume that R_n^{**} is normally distributed for the purpose of making statistical inferences about its preservation by a particular process model. This assumption leads to a considerable saving in computing effort and assists in the graphical presentation of results of analysis. Inferences may be drawn on the basis of standard deviations calculated from smaller samples than that required to adequately define a complete empirical distribution.

10.6 Useful Properties of the Rescaled Adjusted Range

The discussion up to now has shown the rescaled adjusted range statistic to have a great deal of promise as a time series analysis tool for the hydrologist. Its useful properties can be

summarised as follows:

(i) It provides a measure of the 'storage' character of a series.

(ii) The sample value at various values of sub-series length n can be readily determined from an observed series.

(iii) The exact expected value has been obtained in analytical form for some theoretical processes.

(iv) Means and standard deviations or complete empirical distributions can be obtained to a desired accuracy by computer simulation for specified theoretical processes.

(v) Its distribution may be considered to be approximately normal for medium and large n.

(vi) Expected values and standard deviations are independent of the assumed mean level and variance of the theoretical process.

(vii) Expected values are apparently independent of the marginal distribution of the theoretical process.

(viii) Standard deviations are only slightly affected by changes in the marginal distribution of the theoretical process.

In the next chapter it will be shown that comparison of observed values of R_n^{**} with those produced by theoretical processes provides a powerful method of checking the adequacy of proposed process models. It will also be seen that properly identified 'short memory' processes are generally capable of preserving observed values of R_n^{**} up to large values of n. This latter feature and sampling variability go much of the way to accounting for the claimed unexplained behaviour or 'Hurst Phenomenon' in observed data series.

CHAPTER 11: MODELLING THE RESCALED ADJUSTED RANGE IN SOME AUSTRALIAN AND OVERSEAS DATA

11.1 Introduction

In Chapter 10 it was seen that the rescaled adjusted range statistic in theoretical processes has properties which should make it a useful tool for investigating the performance of data generation models. In particular the available evidence points to mean values and standard deviations of the rescaled adjusted range being sensitive only to the autocorrelation structure of the process. The statistic has considerable significance for the hydrologist due to the close analogy with residual mass curve storage analysis.

In this chapter observed hydrologic data series from Australian and overseas locations will be analysed and inferences made about underlying autocorrelation structures on the basis of sample autocorrelation functions and partial autocorrelation functions. Moment estimation type procedures as described in Chapter 9 will be used to fit data generation models. The rescaled adjusted range values obtained from these models will be compared with those obtained from the observed series.

It should be noted once again that the rescaled adjusted range values obtained from the model being examined are sensitive only to the model's autocorrelation structure. To investigate the adequacy of the model as far as the rescaled adjusted range is concerned, it is sufficient to specify the autoregressive and moving-average term parameters. The estimation of parameters relating to the mean, variance and skewness of the process is therefore not required. In this study the models used assume a process mean of zero and a normally distributed random component with a variance of

unity.

11.2 Annual Flows in the St. Lawrence and Niger Rivers - Models Proposed in the Literature

Carlson, MacCormick and Watts (1970) examined the annual series of flows of the St. Lawrence River at Ogdensburg, New York for the period 1860-1957. This series is given by Yevjevich (1963). The authors used the sample autocorrelation function to identify the series as a lag-one autoregressive process and using the maximum likelihood approach to model fitting proposed the following equation as the best model

$$Z_t = 0.69\ Z_{t-1} + a_t \qquad \text{---(11.1)}$$

where Z_t, Z_{t-1} are series values expressed as deviations from the process mean at times t, t-1, and a_t is a realisation of an uncorrelated random variate with a mean of zero and constant variance.

McLeod, Hipel and Lennox (1977) examined the same series of flows. They used the sample autocorrelation function and partial autocorrelation function to identify the process structure as well as two additional functions, the inverse autocorrelation and inverse partial autocorrelation functions. These authors suggest the following constrained three-lag autoregressive model as being superior to that proposed by Carlson et al.

$$Z_t = 0.62\ Z_{t-1} + 0.18\ Z_{t-3} + a_t \qquad \text{---(11.2)}$$

McLeod et al showed that their model is superior to equation (11.1) on the grounds of the comparison of parameter estimates with

their standard errors, the likelihood ratio test (Hipel, McLeod and Lennox - 1979) and the Akaike information criterion (Akaike 1974).

In this study expression (11.1) was replaced by the following

$$q_t = 0.69\, q_{t-1} + v_t \qquad \text{---(11.3)}$$

and expression (11.2) replaced by

$$q_t = 0.62\, q_{t-1} + 0.18\, q_{t-3} + v_t \qquad \text{---(11.4)}$$

where q_t, q_{t-1} and q_{t-3} are values of the process at times t, t-1, t-3 and v_t is a value sampled from the standard normal variate at time t. A differentiation has been made between process values Z_t and q_t because the random components in each process have different variance. The processes represented by Z_t and q_t will have different variance but the same rescaled adjusted range properties. (See section 10.2 of this report for discussion.)

Expression (11.3) and (11.4) were used to generate 500 sequences of various lengths n from which estimates $\bar{R}_n^{**}$, $s(R_n^{**})$ were made of the mean value of the rescaled adjusted range and its standard deviation respectively at each n value. A FORTRAN routine DSRGE (see appendix) was written to perform this task. Figures 11.1 and 11.2 show curves drawn through the mean values of the rescaled adjusted range and the mean plus and minus two standard deviations. The area between the two outlying curves corresponds to a 95% confidence region assuming that R_n^{**} is approximately normally distributed (See section 10.4 for discussion).

Also plotted on Figures 11.1 and 11.2 are R_n^{**}

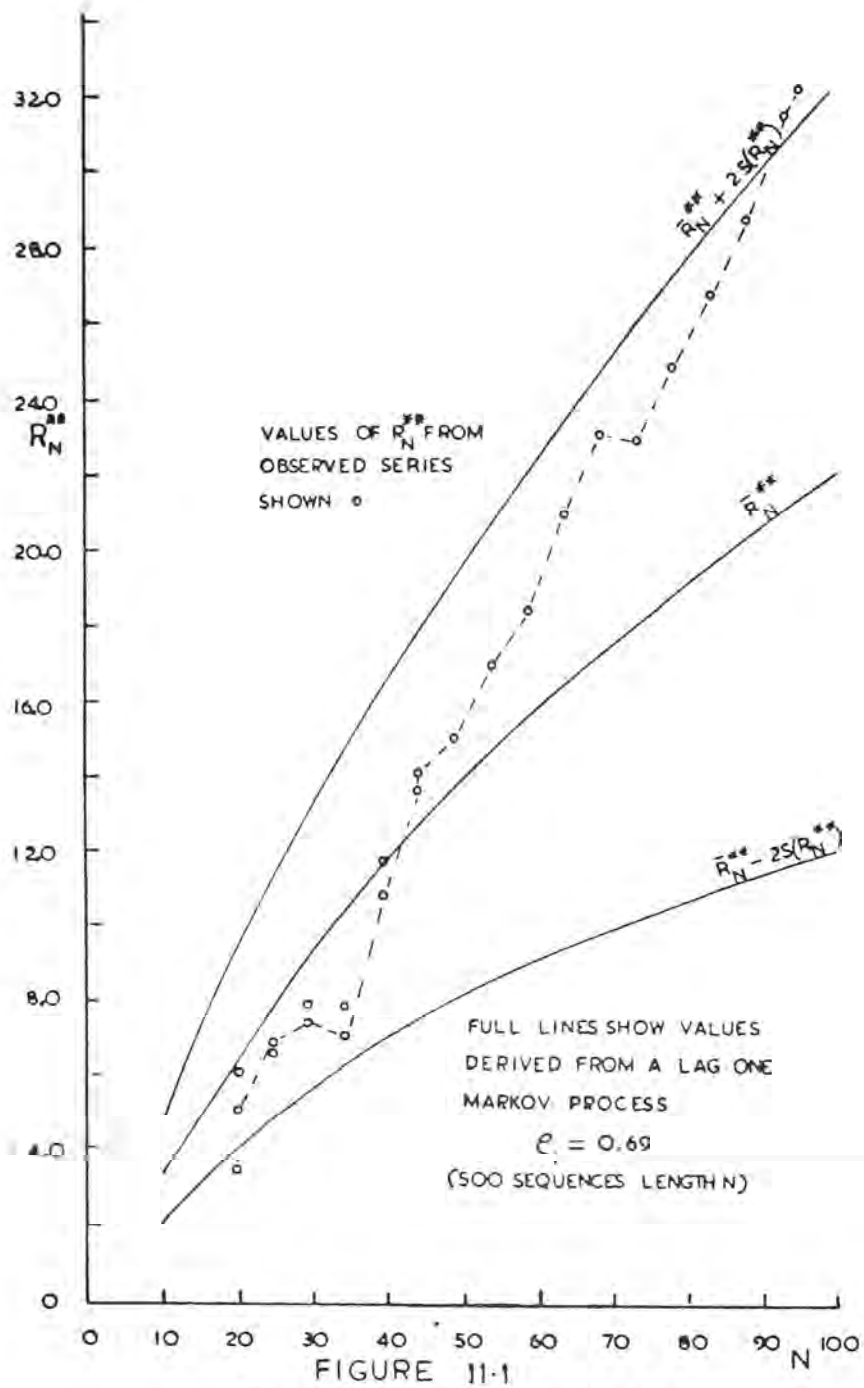

FIGURE 11·1

ST LAWRENCE RIVER (OGDENSBURG)
ANNUAL FLOWS 1860–1957

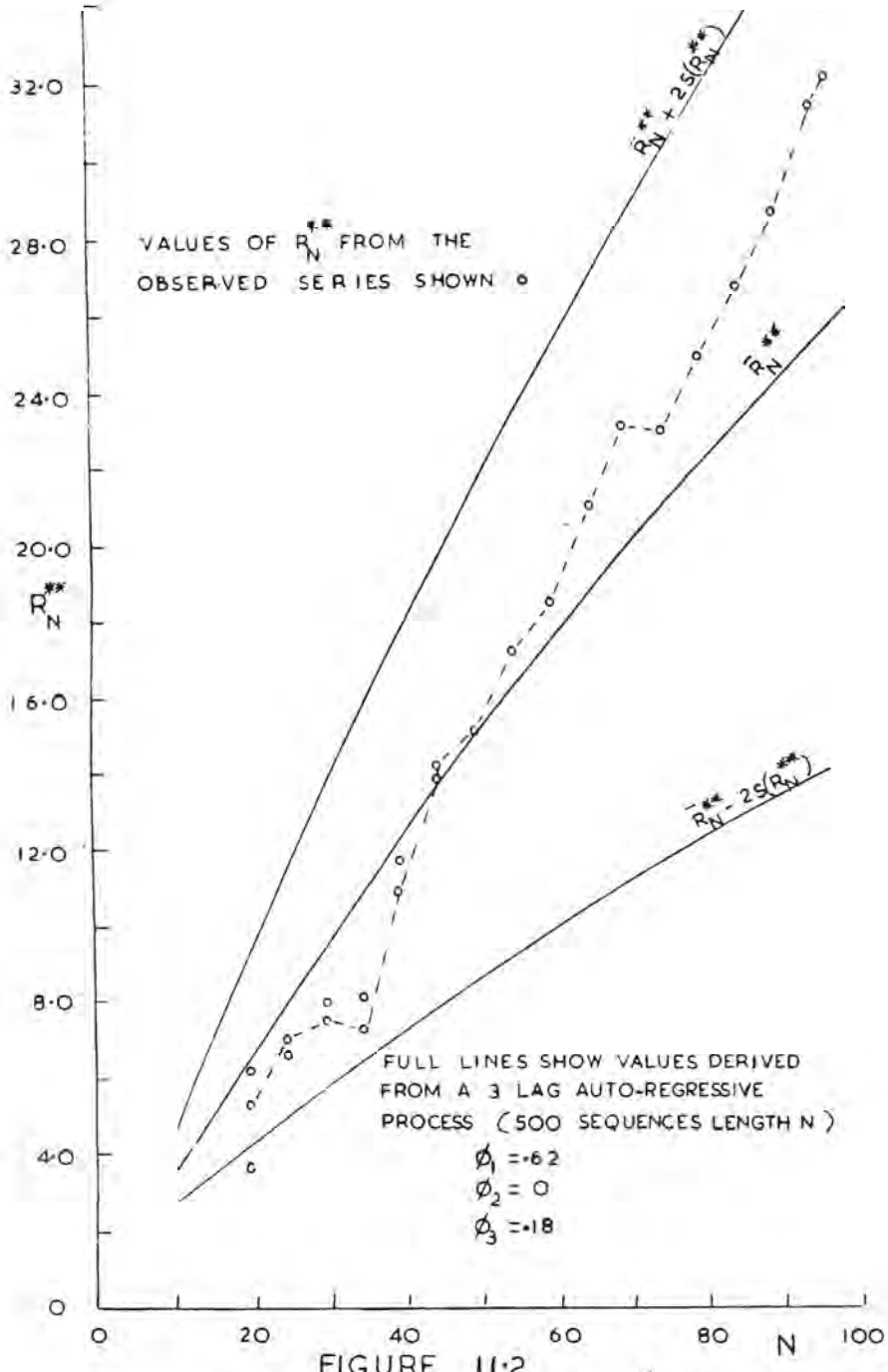

FIGURE 11·2
ST LAWRENCE RIVER (OGDENSBURG)
ANNUAL FLOWS 1860-1957

values obtained from an analysis of the St. Lawrence River annual flow series. The data used was the series given by Yevjevich (1963); the source acknowledged by McLeod et al and Carlson et al. The FORTRAN routine RANGE (see appendix) was used for the analysis.

As discussed in section 7.2 of this report the method used in this study to evaluate the rescaled adjusted range in an observed data series is as follows. A sub-series length n is nominated. The observed series is subdivided into as many adjacent non-overlapping sub-series as the length of the record permits and with the first sub-series commencing at the start of the record. R_n^{**} is then evaluated for each sub-series. Where more than one sub-series is available from the record, the various values of R_n^{**} obtained can be regarded as independent estimates of the true expected value of R_n^{**} for the underlying stochastic process. A sample function of the rescaled adjusted range can be developed from the observed series by repeating the above procedure for various values of n. A necessary feature of such a function is the decreasing number of independent samplings of R_n^{**} as the value of n increases. For n greater than one half of the observed series length there is of course only one independent sample of R_n^{**} available.

The sample rescaled adjusted range function for the St. Lawrence River data was plotted in Figures 11.1 and 11.2. The values obtained from sub-series commencing at the beginning of the series are connected with a dotted line as an aid to visual interpretation. As might be expected there appears to be a considerable amount of dependence between the R_n^{**} values at adjacent values of n. This is particularly so for the values of R_n^{**} joined by the dotted line. Adjacent values of R_n^{**} so marked are derived from sub-series which include common terms up to the smaller value of n.

A comparison of Figures 11.1 and 11.2 shows that the claimed superiority of the three-lag model proposed by McLeod et al is reflected in an apparent improvement in the ability of the model to generate series which have values of the rescaled adjusted range which agree with those obtained from the historical data series. The word apparent has been used because in the area of greatest interest where the value of n approaches the series length, there is only one independent estimate of the rescaled adjusted range available. It is not possible therefore to infer in a strict statistical sense that the rescaled adjusted range is better preserved by either model. One gains the visual impression however that the constrained three-lag process models the observed rescaled adjusted range values more successfully. It will be seen on carrying out a similar analysis of other series and models, that when components are added to a model so that the model better preserves the auto-correlation structure of the historical data series, there is a consistently better fit between the theoretical confidence region of the rescaled adjusted range values and those values obtained from the historical data.

Carlson, MacCormick and Watts (1970) examined the series of annual flows in the Niger River at Koulicoro (1906-1957). This data is also obtained from Yevjevich (1963). The authors identified a number of possible models, among them a lag-one autoregressive model:

$$q_t = 0.55\, q_{t-1} + v_t \qquad \text{---(11.5)}$$

and a mixed autoregressive moving average model of order one (ARMA (1,1,)):

$$q_t = 0.82\, q_{t-1} + v_t - 0.40\, v_{t-1} \qquad \text{---(11.6)}$$

The authors showed that equation (11.6) is the superior model on the basis of a least squares fit criterion.

Figures 11.3 and 11.4 compare theoretical and observed rescaled adjusted range values and again there is some apparent improvement in the modelling of the rescaled adjusted range by the claimed superior model.

It should be noted that the above comparison between theoretical and observed values of the rescaled adjusted range is a comparison of the fit between the sampling spread of observed values and a theoretical confidence region. The comparisons between observed and theoretical values of R_n^{**} or Hurst's coefficient K that have been made in the literature to date have, it would seem, only been made on the basis of mean and expected values. It is suggested that the approach used in this study is superior because of the considerable sampling variability in R_n^{**} and K.

11.3 Annual Flows in Some Australian Rivers

The comparison of the theoretical and observed rescaled adjusted range functions appear to give an insight into model behaviour in the case of the St. Lawrence and Niger River models. Attention will now be focussed on annual flow series of some Australian rivers.

The series of annual flows in the Darling River at Wilcannia (1886-1971) was analysed to obtain the sample autocorrelation and partial autocorrelation functions which are shown as Figures 11.5 and 11.6. The series exhibits a lag-one autocorrelation coefficient

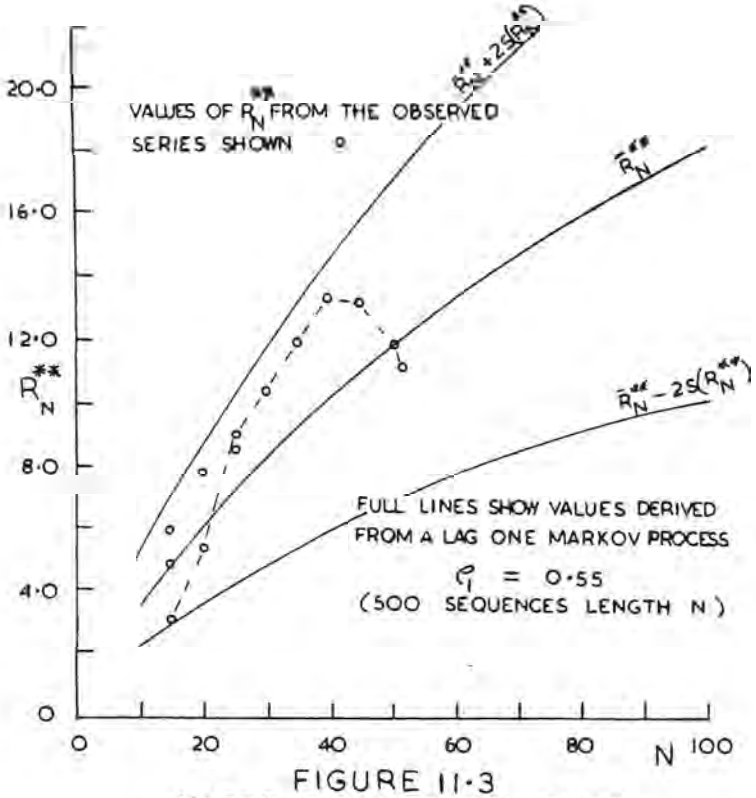

FIGURE 11·3
NIGER RIVER AT KOULICORO
ANNUAL FLOWS (1906-1957)

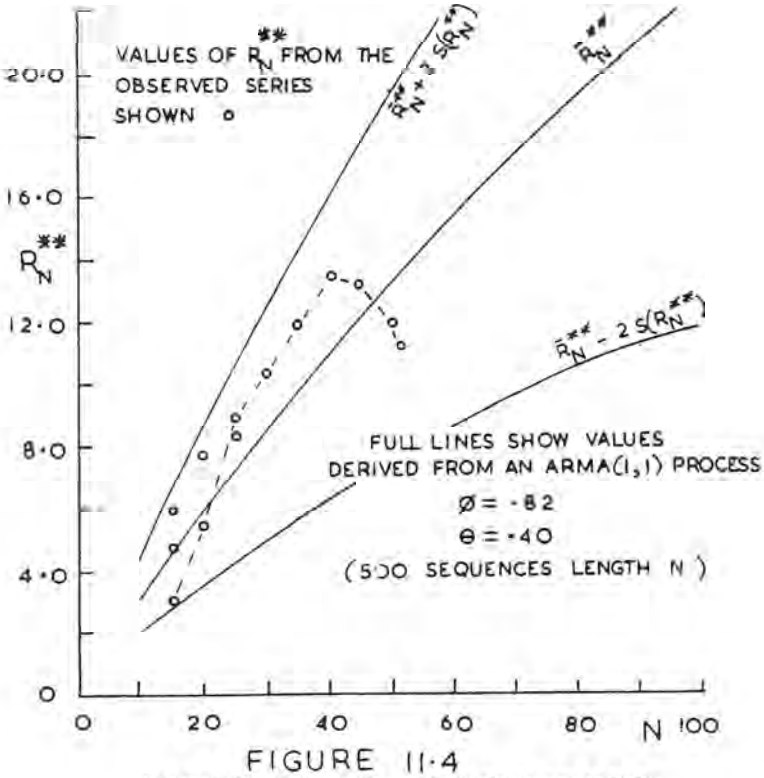

FIGURE 11·4
NIGER RIVER AT KOULICORO
ANNUAL FLOWS (1906 - 1957)

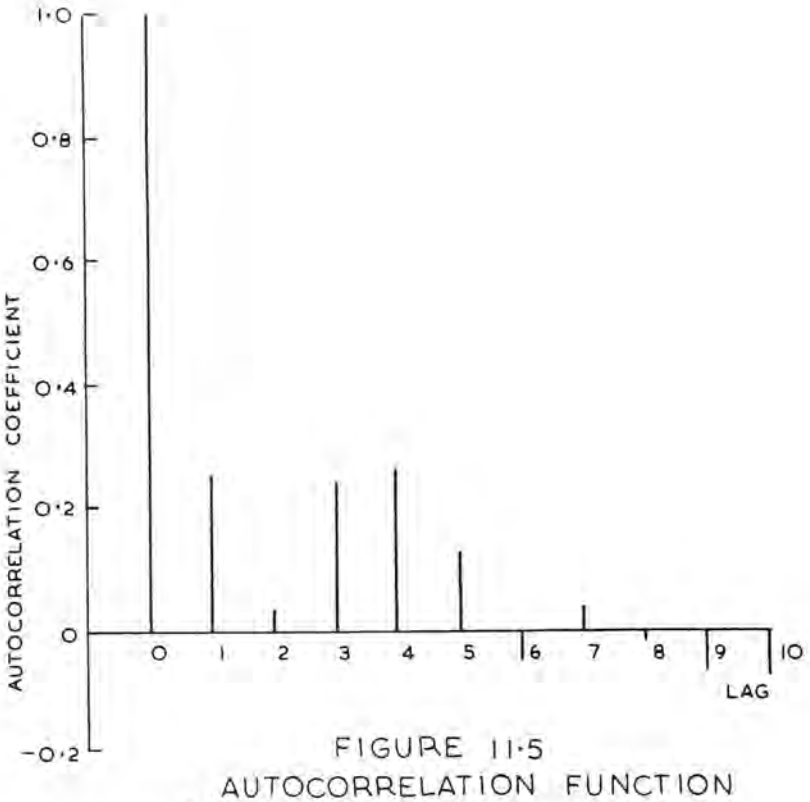

FIGURE 11·5
AUTOCORRELATION FUNCTION
DARLING RIVER AT WILCANNIA
ANNUAL FLOWS (1886-1971)

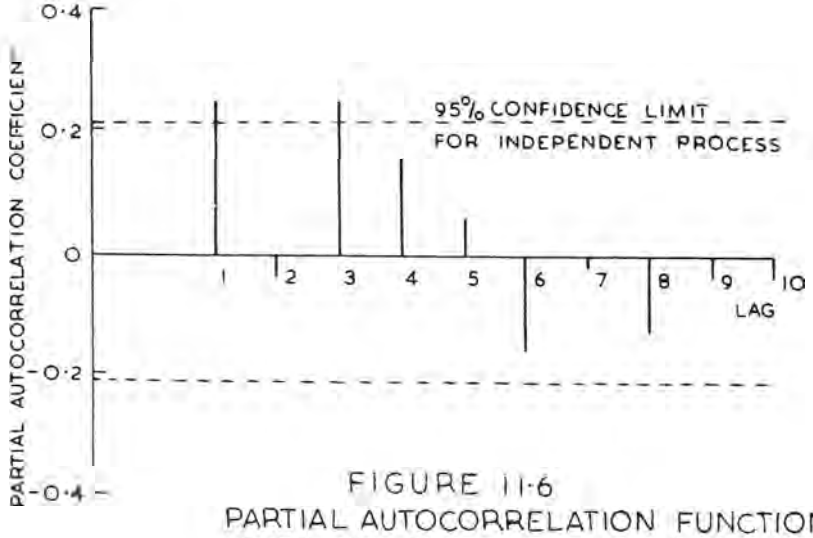

FIGURE 11·6
PARTIAL AUTOCORRELATION FUNCTION
DARLING RIVER AT WILCANNIA
ANNUAL FLOWS (1886-1971)

of 0.25 and hence a model of the form:

$$q_t = 0.25\, q_{t-1} + v_t \qquad \text{---(11.7)}$$

might be proposed. However the autocorrelation function shows similarly large values at lags three and four while the partial autocorrelation function has values outside the 95% confidence interval at lags one and three. The value at lag three is unlikely to result by chance if the 'underlying' process is assumed autoregressive and of order less than three. A three lag autoregressive process would therefore seem a better model than equation (11.7). Solving the Yule-Walker equations for a three lag process (see section 9.2 of this report) yields the following model form:

$$q_t = .26\, q_{t-1} - .09\, q_{t-2} + .26\, q_{t-3} + v_t \qquad \text{---(11.8)}$$

The observed and theoretical rescaled adjusted range functions for the two models are shown as Figures 11.7 and 11.8. Once again it appears that plots of this kind can discriminate between models as regards the appropriateness of their autocorrelation structure. The three lag model appears to give an improved fit to the observed rescaled adjusted range values.

It should be noted that, in this study, models have not been rigorously identified and estimated. The aim here is not to find the best model but to observe trends in the rescaled adjusted range function plots with increased model component identification effort. Accordingly model parameters have been determined by solution of the Yule-Walker equations (9.5) rather than by the more efficient maximum likelihood techniques.

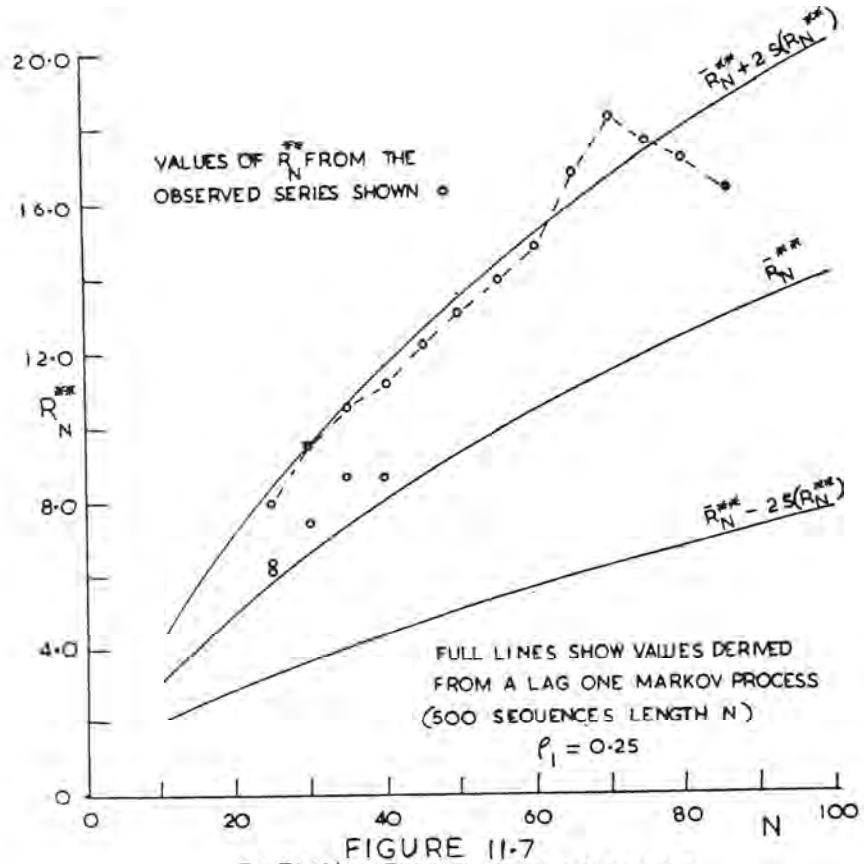

FIGURE 11·7
DARLING RIVER AT WILCANNIA
(ANNUAL FLOWS 1886-1971)

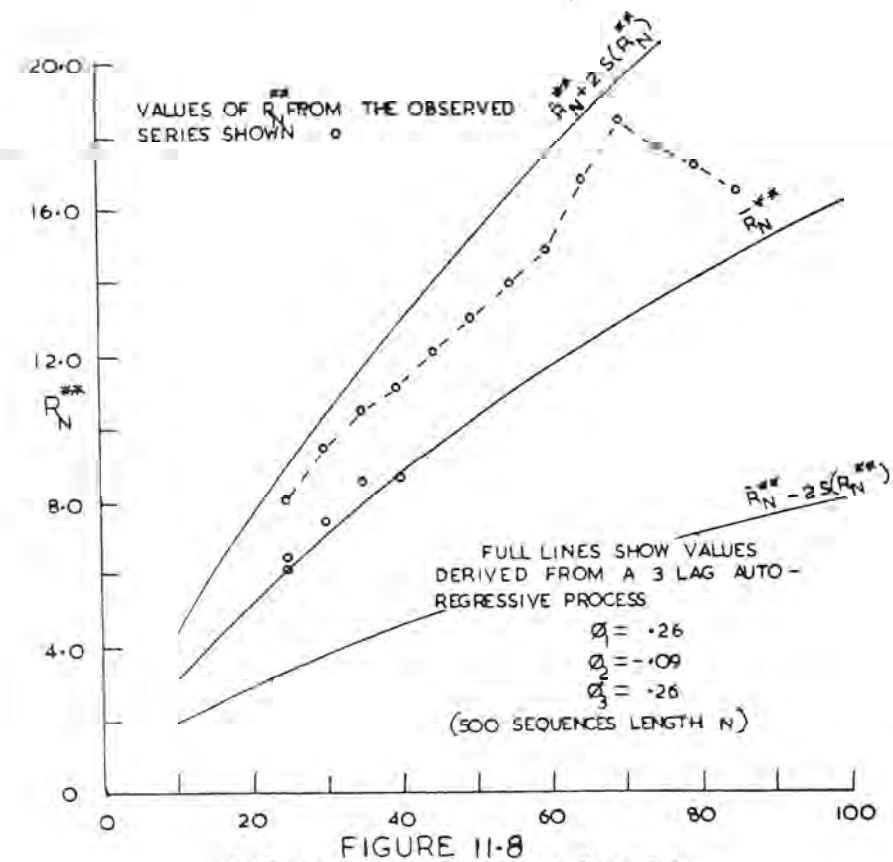

FIGURE 11·8
DARLING RIVER AT WILCANNIA
ANNUAL FLOWS (1886-1971)

Figures 11.9 to 11.13 show rescaled adjusted range function plots for annual flow series of the Snowy, Barron, Brisbane, Burdekin and Kiewa Rivers. In these cases the evidence available from the sample autocorrelation and partial autocorrelation functions points to either independent random processes or lag-one autoregressive (Markov) processes being appropriate models. This finding is reinforced in each of the cases by the rescaled adjusted range function plot. These simple models appear to preserve the rescaled adjusted range quite well.

A comparison of Figures 11.13 and 11.14 illustrate the sensitivity of the rescaled adjusted range to autocorrelation. In Figure11.14 the sample range function of the Kiewa River annual flow series is superimposed on the theoretical function for an independent random process. The fit is substantially poorer than for the model with the appropriate degree of autocorrelation.

The remaining Australian annual stream flow series analysed was the Macquarie River at Burrendong (1886-1964). The sample autocorrelation and partial autocorrelation functions are shown as Figures 11.15 and 11.16. The partial autocorrelation function shows no values outside the 95% confidence interval and hence there is no strong indication that the appropriate model would be other than an independent random process. However the rescaled adjusted range function plot (Figure 11.17) shows that an independent random process model does not appear to yield satisfactory rescaled adjusted range values. The disparity could be due to chance but the modeller might be advised to re-examine the data and should certainly be cautious in applying such a model.

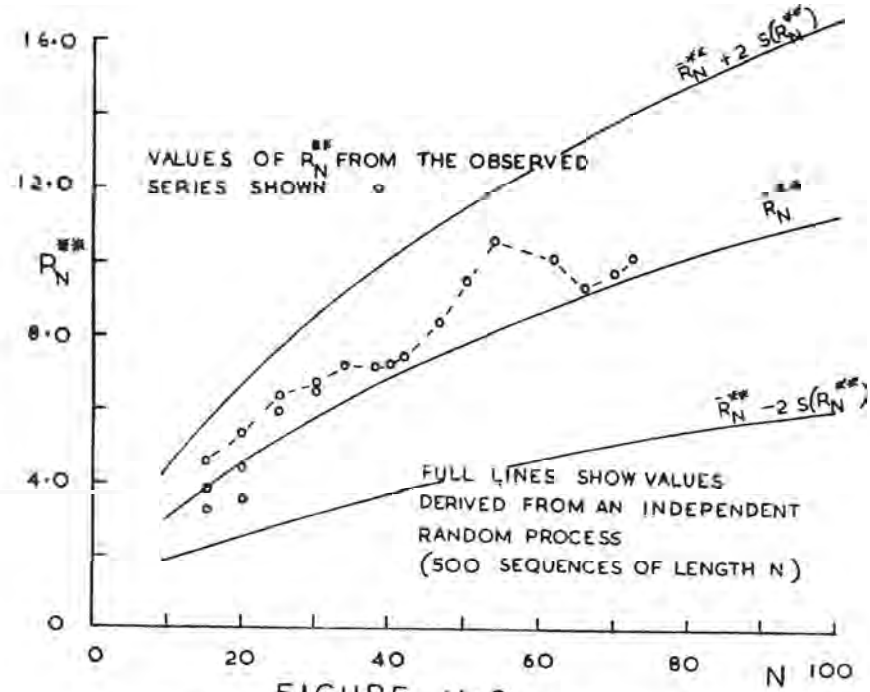

FIGURE 11·9
SNOWY RIVER AT JINDABYNE
ANNUAL FLOWS (1905-1977)
(CORRECTED FOR REGULATION)

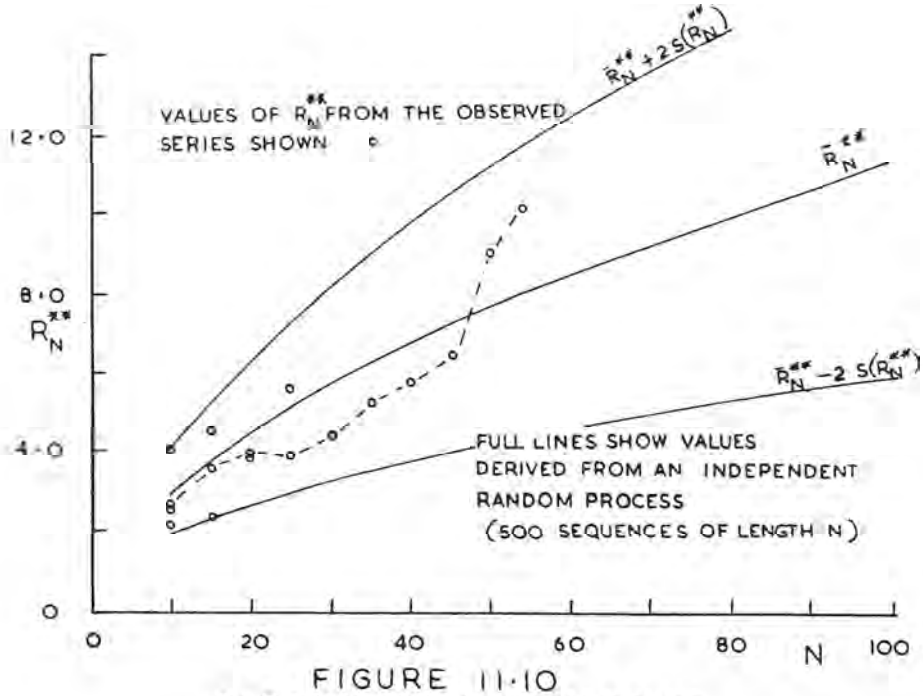

FIGURE 11·10
BARRON RIVER AT MAREEBA
ANNUAL FLOWS (1916-1969)

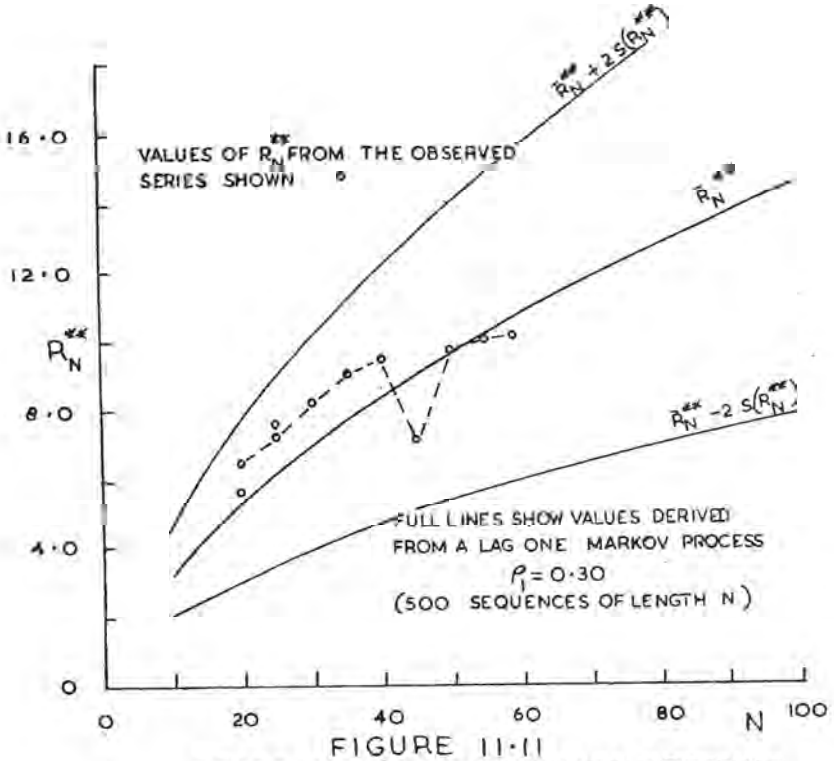

FIGURE 11·11
BRISBANE RIVER AT SAVAGES CROSSING
ANNUAL FLOWS (1910-1968)

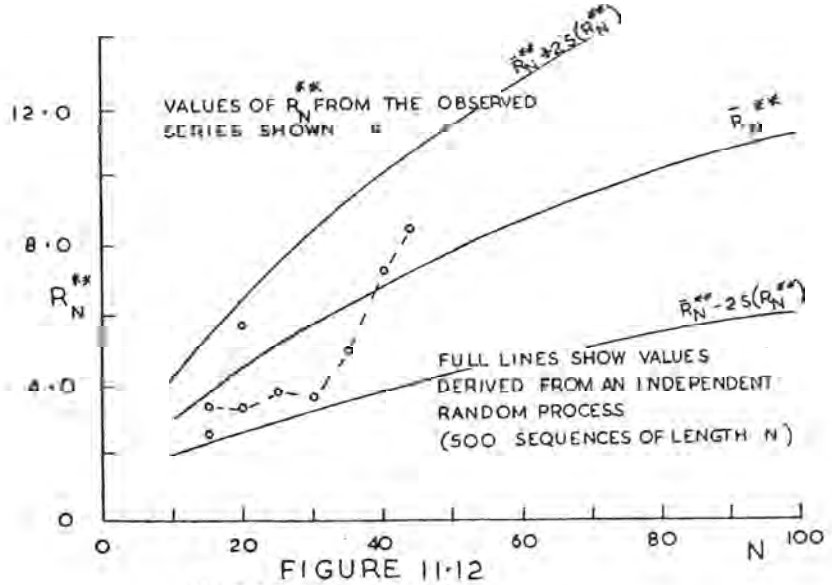

FIGURE 11·12
BURDEKIN RIVER
ANNUAL FLOWS (1920-1963)

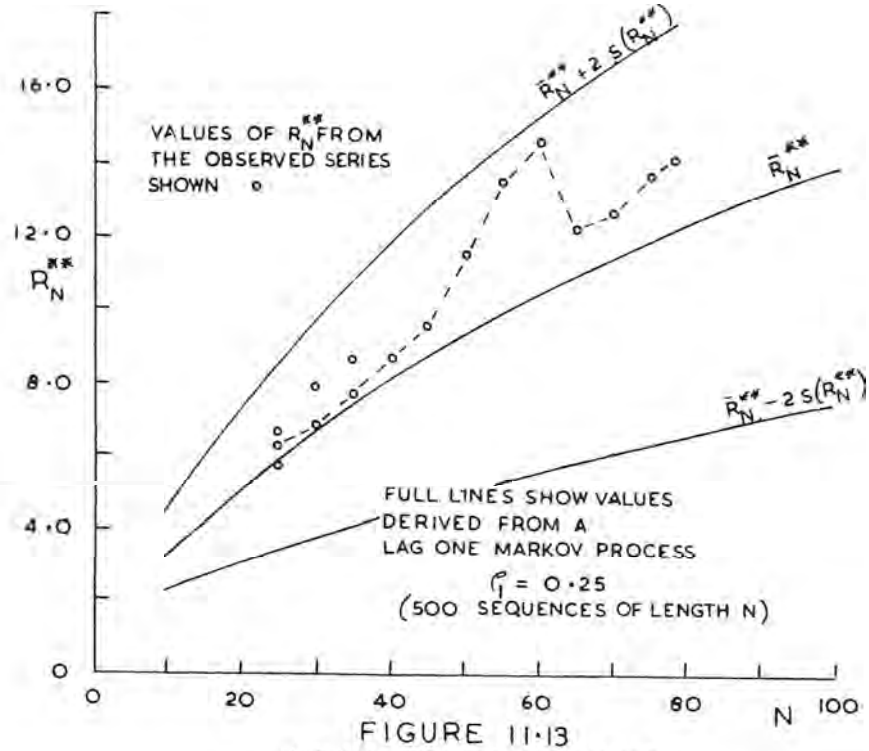

FIGURE 11·13
KIEWA RIVER AT KIEWA
ANNUAL FLOWS (1893-1970)

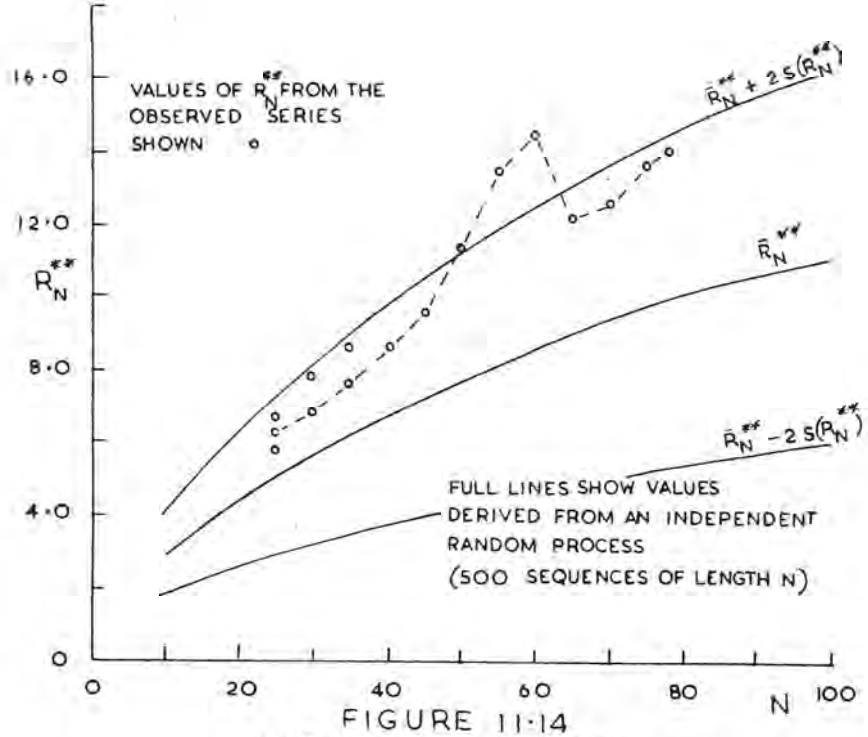

FIGURE 11·14
KIEWA RIVER AT KIEWA
ANNUAL FLOWS (1893-1970)

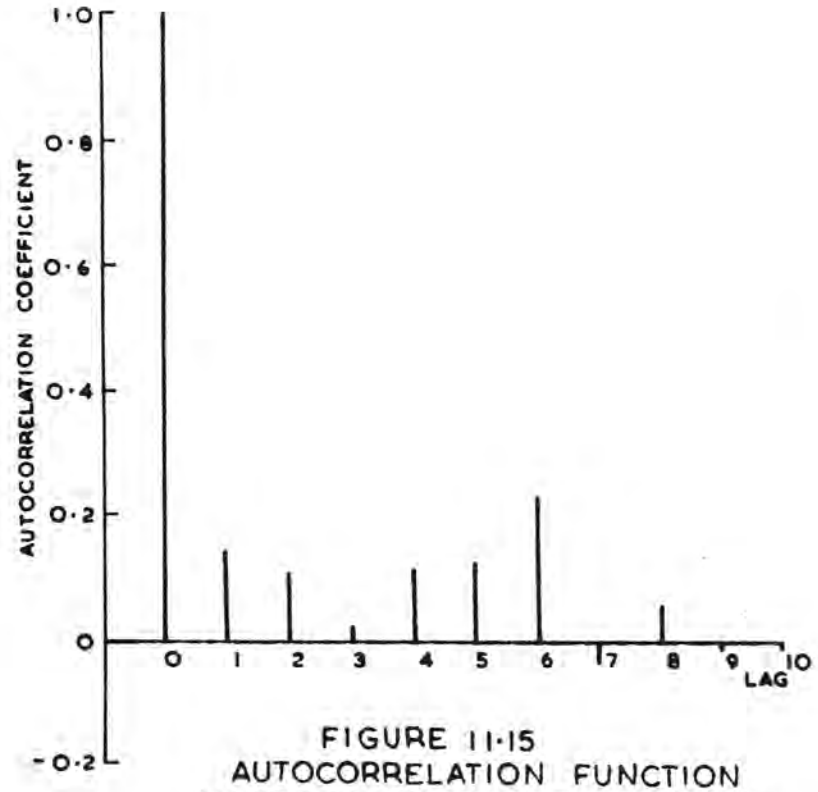

FIGURE 11·15
AUTOCORRELATION FUNCTION
MACQUARIE RIVER AT BURRENDONG
ANNUAL FLOWS (1886-1964)

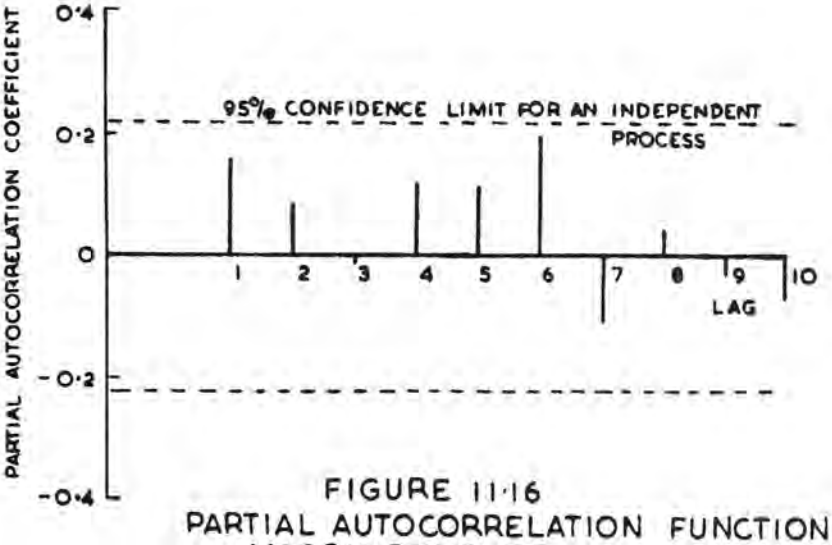

FIGURE 11·16
PARTIAL AUTOCORRELATION FUNCTION
MACQUARIE RIVER AT BURRENDONG
ANNUAL FLOWS (1886-1964)

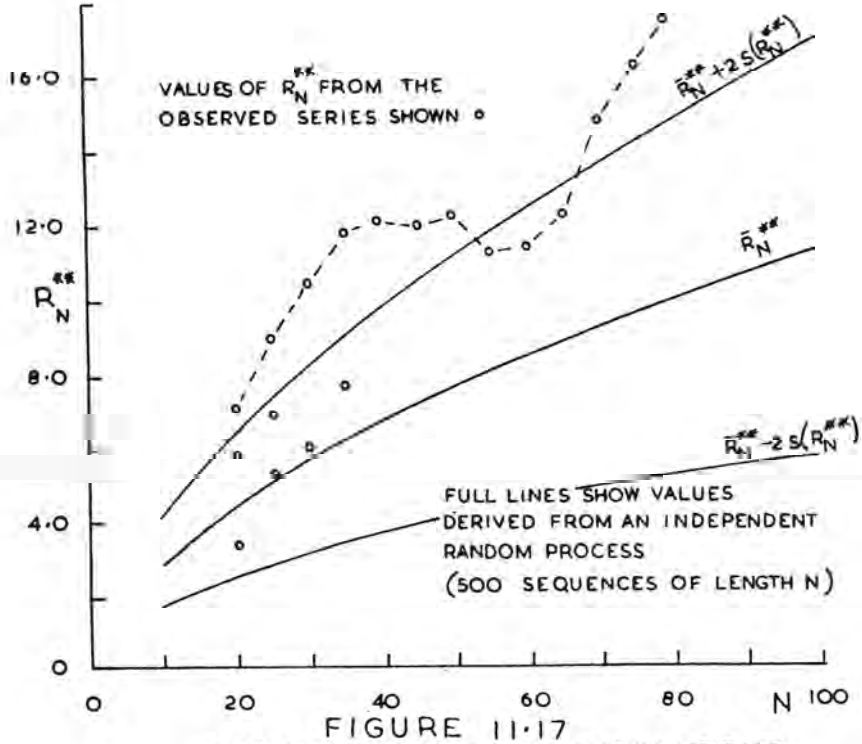

FIGURE 11·17
MACQUARIE RIVER AT BURRENDONG
ANNUAL FLOWS (1886-1964)

11.4 Monthly Flows in Some Australian Rivers

Monthly flows are of great practical interest to the hydrologist and water resource system designer as the period of one month allows for realistic simulation and analysis in many cases. Monthly flows of course have the complication of periodicity and often high autocorrelation. One commonly used approach to the modelling of monthly flow series is to transform the observed periodic series into one which is approximately stationary by subtracting monthly mean values and dividing by monthly standard deviations. The resulting 'standardised' series can then be modelled with the type of models already discussed in this report. The standardised series is stationary with respect to mean and standard deviation. Periodicities may remain however in higher moments such as skewness and also in the autocorrelations from one monthly value to the next. It is of great interest therefore to know whether the observed series can be adequately modelled by viewing it as a single 'standardised' entity. If so, autocorrelation structure may be typified by say, a single overall value of lag-one autocorrelation rather than different values between different pairs of calendar months.

Wright (1975) analysed monthly flows from 12 Australian streams including the Kiewa River at Kiewa and the Macquarie River at Burrendong. He removed periodicities from the series and found that the partial autocorrelation functions of the resulting series indicated that autoregressive models of order greater than one were indicated in nine of the cases. Models of various lags were fitted to each of the series and sequences of flows generated. Wright found that the marginal statistics of the observed series such as the means, standard deviations and skewness were reasonably well preserved by the models irrespective of the number of lags included

in them. He concluded that there is little justification for using autoregressive structures of order higher than one unless the preservation of the serial correlogram (autocorrelation function) is important.

The importance of preserving the series autocorrelation is clearly seen in Figures 11.18 and 11.19. These figures show the sample rescaled adjusted range function of the standardised series of monthly flows in the Kiewa River at Kiewa (1893-1970) together with the range values obtained from the appropriate one lag and three lag autoregressive models. The three lag model clearly appears superior as regards preserving the rescaled adjusted range. A similar effect is noticed in Figures 11.20 and 11.21 for the Macquarie River at Purrendong (1886-1964) standardised monthly series. Examination of the partial autocorrelation function in this case indicates that a 12 lag model is appropriate. This multilag model also appears to be more successful than the lag-one model in preserving the rescaled adjusted range.

In the case of the standardised monthly flow series for the Snowy River at Jindabyne (1905-1977) the sample autocorrelation and partial autocorrelation functions indicate that a lag-one autoregressive model should be appropriate. Figure 11.22 shows that in this case such a model appears to adequately preserve the rescaled adjusted range.

11.5 Annual Rainfalls at Some Australian Localities

Figures 11.23 to 11.26 show rescaled adjusted range function plots for annual rainfall series at Adelaide, Alice Springs, Windsor (N.S.W.) and Balranald. The sample autocorrelation and partial autocorrelation functions indicate that an independent

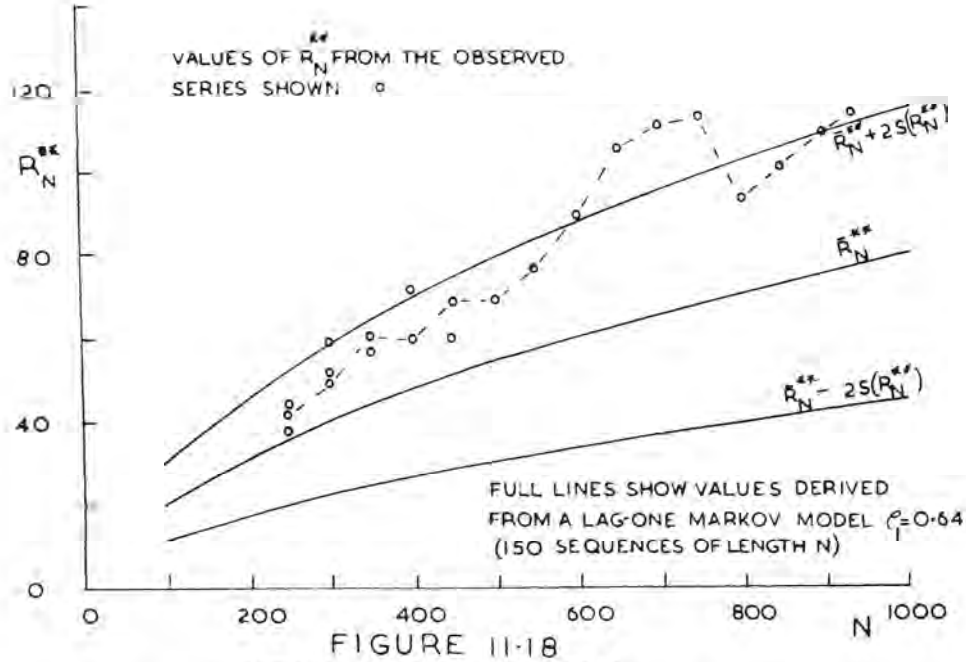

FIGURE 11·18
KIEWA RIVER AT KIEWA
STANDARDISED MONTHLY FLOWS (1893-1970)

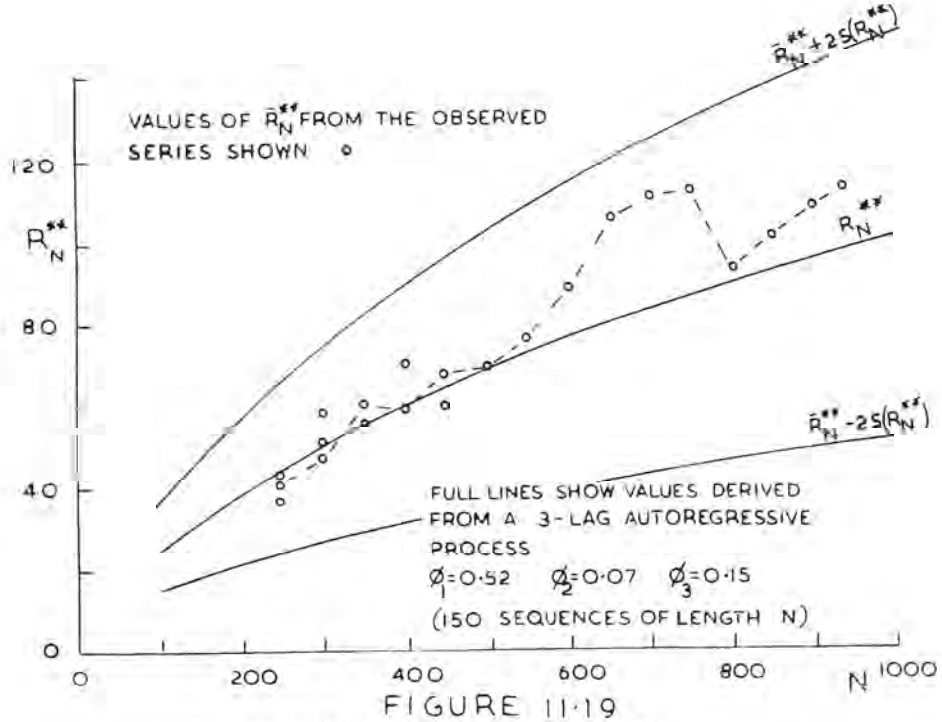

FIGURE 11·19
KIEWA RIVER AT KIEWA
STANDARDISED MONTHLY FLOWS (1893-1970)

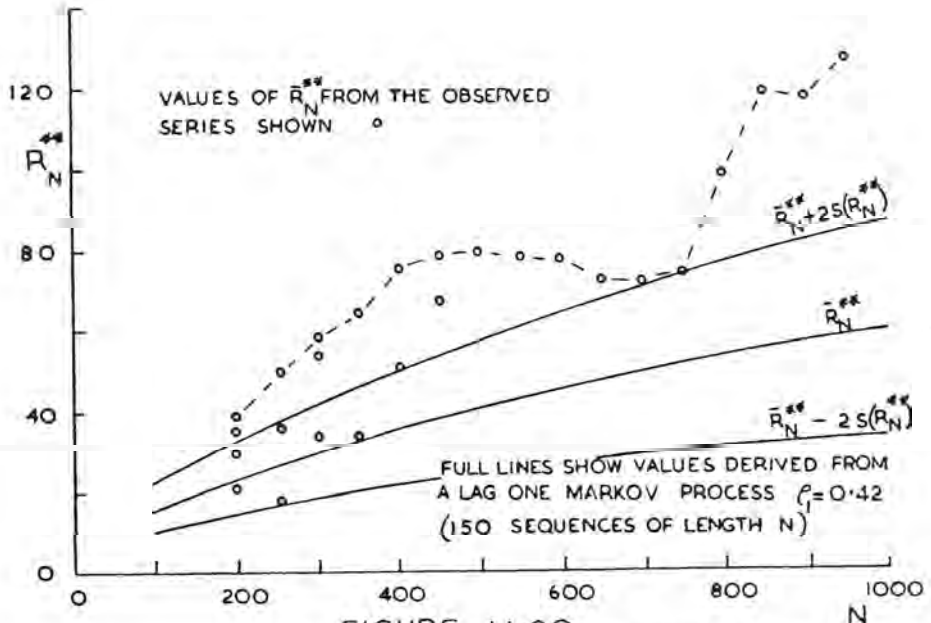

FIGURE 11·20
MACQUARIE RIVER AT BURRENDONG
STANDARDISED MONTHLY FLOWS (1886-1964)

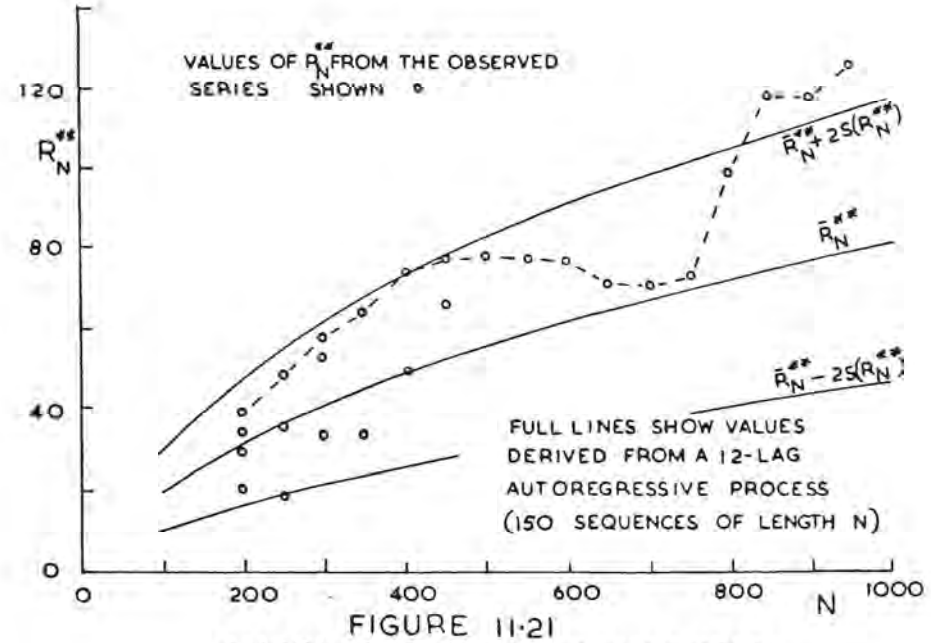

FIGURE 11·21
MACQUARIE RIVER AT BURRENDONG
STANDARDISED MONTHLY FLOWS (1886-1964)

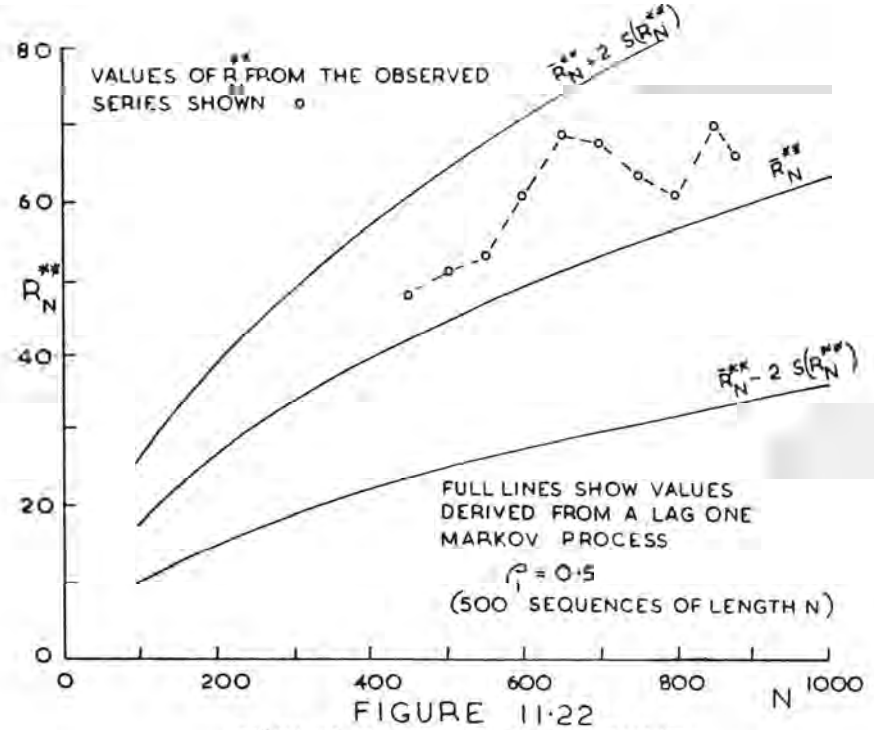

FIGURE 11·22
SNOWY RIVER AT JINDABYNE
STANDARDISED MONTHLY FLOWS
(1905-1977)

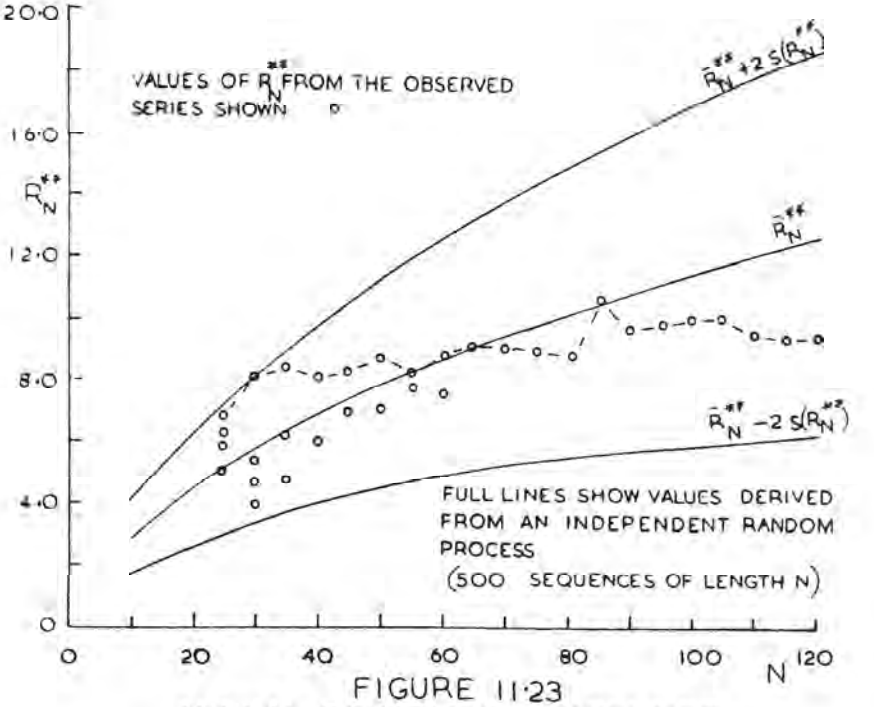

FIGURE 11·23
ANNUAL RAINFALLS AT ADELAIDE
(1839 - 1960)

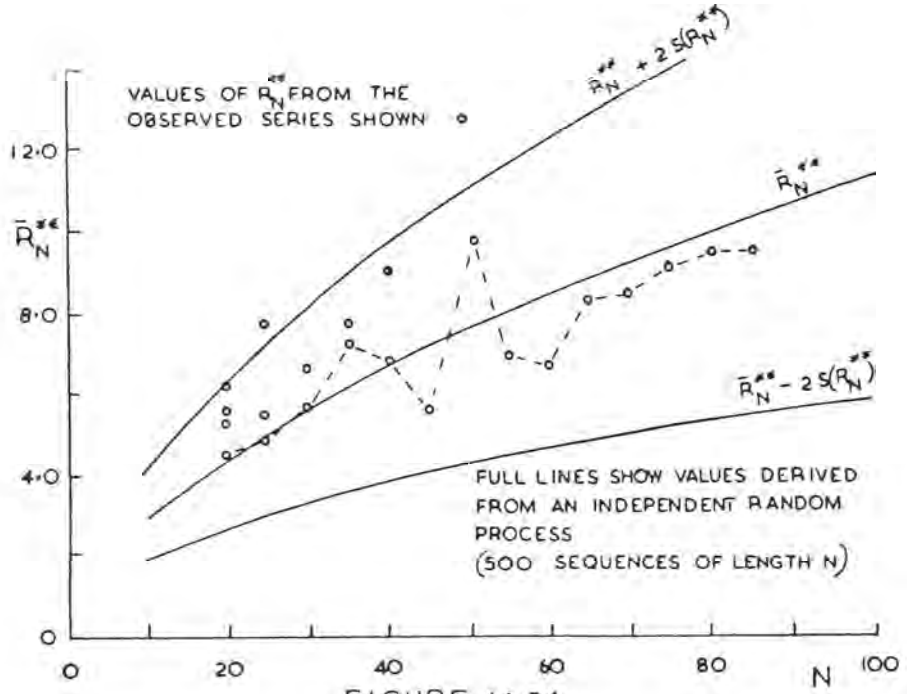

FIGURE 11·24
ANNUAL RAINFALLS AT ALICE SPRINGS
(1874 - 1960)

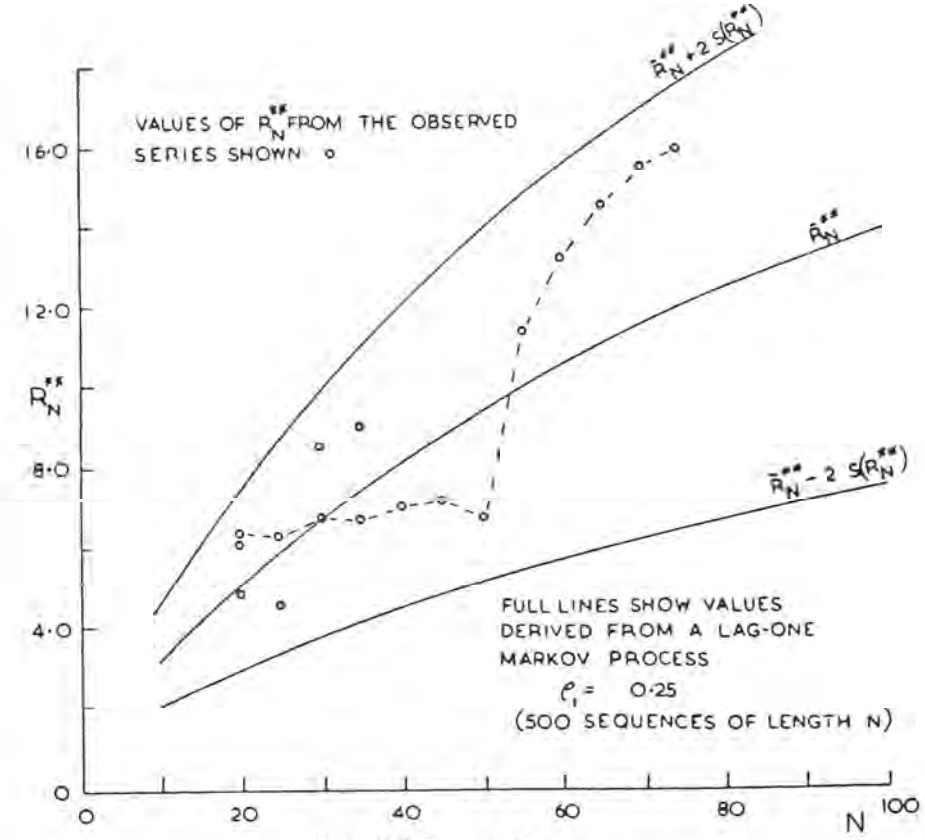

FIGURE 11·25
ANNUAL RAINFALLS AT WINDSOR - NSW
(1898-1971)

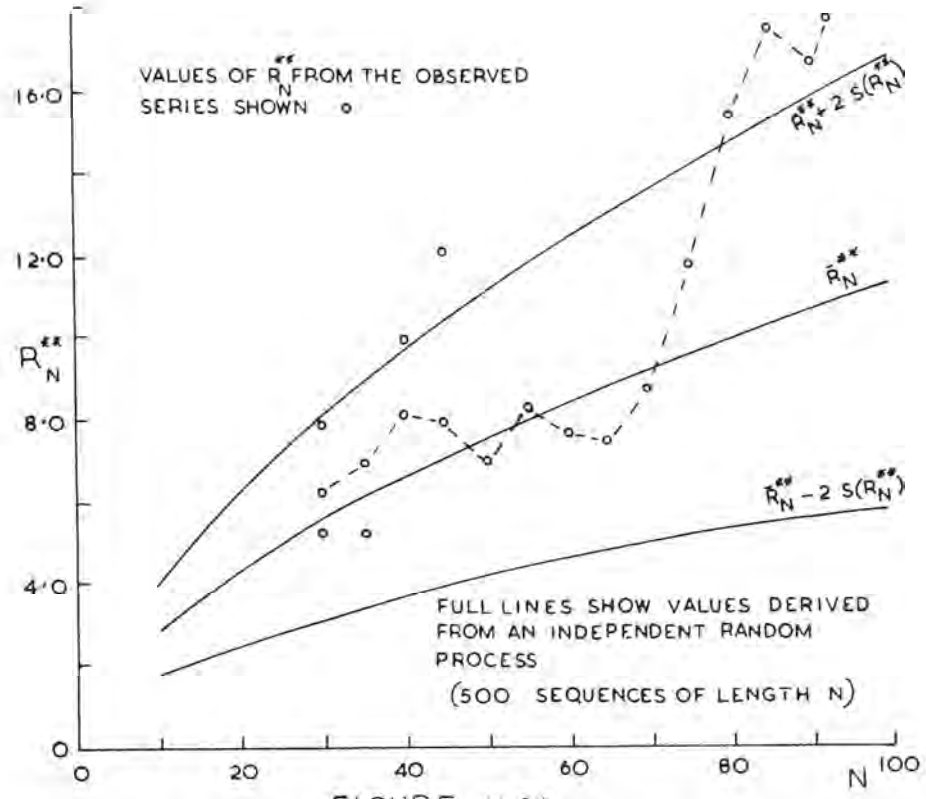

FIGURE 11·26
ANNUAL RAINFALLS AT BALRANALD
(1879 - 1970)

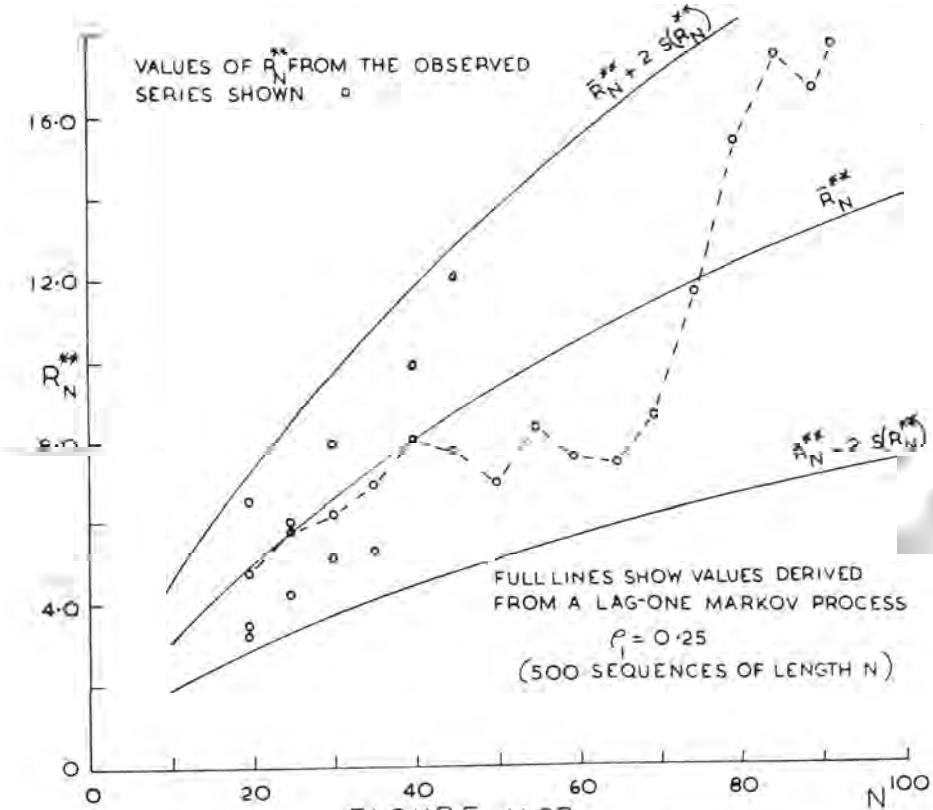

FIGURE 11·27
ANNUAL RAINFALLS AT BALRANALD
(1879 - 1970)

random process would be an appropriate model for the Adelaide (1839-1963) and Alice Springs (1874-1960) series. Figures 11.23 and 11.24 indicate that in both cases the rescaled adjusted range appears to be rather well preserved by such a model. The autocorrelation and partial autocorrelation functions of the Windsor (1898-1971) series indicate a lag-one autoregressive model with an autocorrelation coefficient of 0.25. Figure 11.25 shows that the indicated model appears to preserve the rescaled adjusted range in this case also.

In the case of the Balranald (1879-1970) series a random independent process model is indicated. The rescaled adjusted range function plot (Figure 11.26) however shows a lack of fit which, while being possibly due to chance, may cause the modeller to doubt the adequacy of the model. Figure 11.27 shows the performance of a lag-one autoregressive model with an autocorrelation coefficient of 0.25. In this case the fit between the theoretical and observed rescaled adjusted range values is quite good. Given that the preservation of the storage characteristics of the series is important, and in the light of uncertainties surrounding model identification on the basis of sample autocorrelation and partial autocorrelation functions, it is suggested that the modeller might consider selecting the auto-regressive model as being more suitable.

Potter (1976) examined six annual rainfall series ranging in timespan from 100 to 155 years for locations along the east coast of the United States. In contrast to the Australian series discussed above, Potter found the U.S. series to be quite highly autocorrelated with lag-one coefficients ranging from 0.22 to 0.59. He also calculated the Hurst coefficient K for the series and found values ranging from 0.73 and 0.88. From this and other analysis Potter concluded that the series exhibited nonstationarity of the mean which

was a reflection of shifts in climate. He claimed that the series could not therefore be modelled by Markov (lag-one autoregressive) models. This conclusion is not well supported by the results for the Australian series examined in this study. Independent and lag-one autoregressive models appear quite satisfactory for at least the Adelaide, Alice Springs and Windsor series.

As discussed in detail earlier in this report, considerable caution must be used in interpreting one-point Hurst coefficient (K) values. For example the K value in the case of the Windsor series (Figure 11.25) as determined from the single rescaled adjusted range value at the end of the series is 0.76 and is obviously subject to large sampling error.

11.6 Tree Rings and Mud Varves

In section 8.3 of this report, plots of log R_n^{**} versus log n were examined for two very long annual series believed to have climatic or hydrologic significance. The series were the North Finland Pine Tree-Ring index (Siren 1961) of 780 years timespan and the Lake Saki Mud Varve series (Shostakovitsch 1934) of 4,180 years. The tree ring index is believed to correlate with mean summer temperatures and the mud varve thicknesses with annual lake inflows.

These two series are further examined by means of rescaled adjusted range function plots drawn to natural scales in Figures 11.28 and 11.29.

The sample autocorrelation and partial autocorrelation functions of the tree ring data indicate that a four lag auto-regressive model might be appropriate. The rescaled adjusted range function plot (Figure 11.28) indicates that such a model produces rescaled adjusted range values which appear rather high. This is an

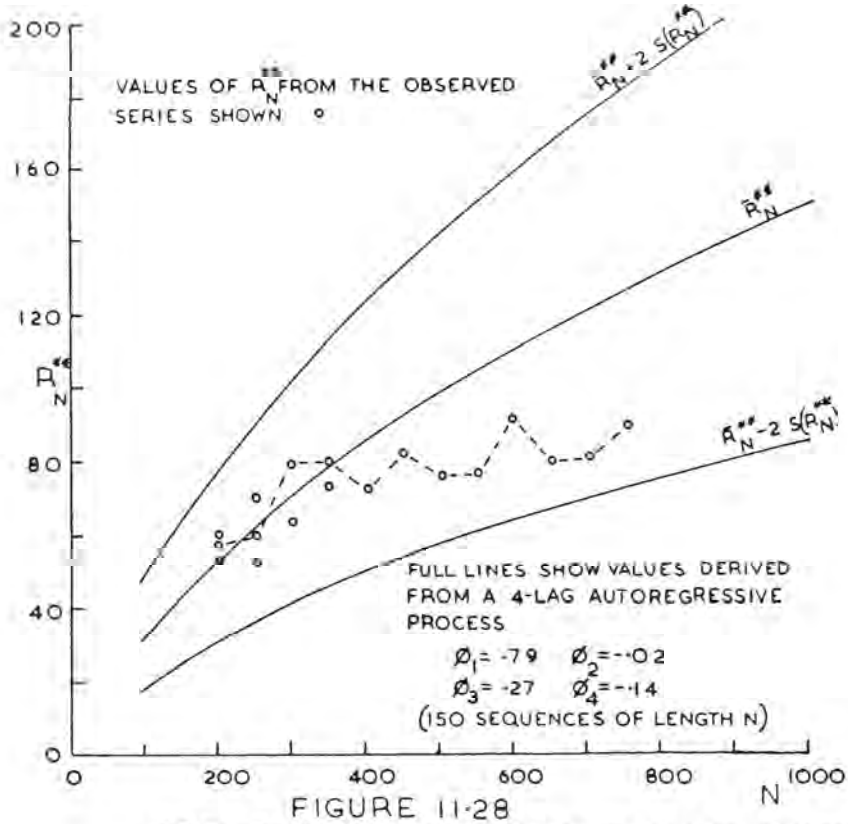

FIGURE 11·28
TREE RING INDEX - NORTH FINLAND PINES
(770 YEARS)

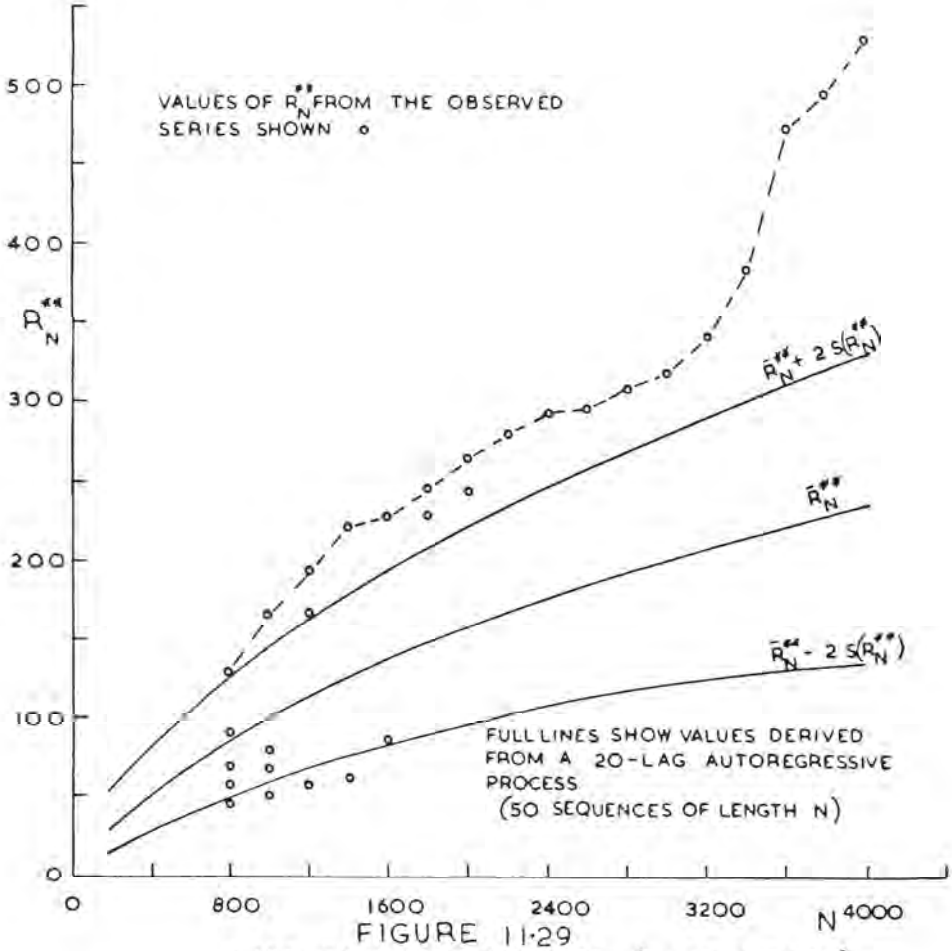

FIGURE 11·29
LAKE SAKI MUD VARVES (4180 YEARS)

interesting result as 'short memory' autoregressive models have often been cited in the literature as being incapable of reproducing the high rescaled adjusted range values observed in very long geophysical series.

The rescaled adjusted range function plot for the Lake Saki data (Figure 11.29) shows a quite different result. The sample autocorrelation function for this series shows a relatively small lag-one coefficient of 0.24 and a lag-two coefficient of 0.17. Beyond this value the autocorrelation function attenuates extremely slowly. The partial autocorrelation function exhibits apparently significant values out to about lag 20 and hence a 20 lag autoregressive model was fitted by solution of the Yule-Walker equations (9.5). Figure 11.29 shows clearly that such a model gives values of the rescaled adjusted range which are too low for values of the subseries length n greater than about 1600. The shape of the autocorrelation and partial autocorrelation functions in this case give some indication that a mixed autoregressive moving average (ARMA) process might be appropriate. An ARMA model with one autoregressive and one moving average term was fitted but the rescaled adjusted range values produced were considerably smaller than for the 20 lag autoregressive model. ARMA models of higher order than ARMA (1,1) were not examined but such a study would be of interest.

11.7 The Rescaled Adjusted Range Function

The preceding analysis of various data series and models shows quite clearly the value of the rescaled adjusted range function as another weapon in the time-series modeller's armoury. The term 'rescaled adjusted range function' has been used deliberately because of the analogy with the autocorrelation function. There is also an

analogy between the comparison of the theoretical and sample autocorrelation functions and the theoretical and sample rescaled adjusted range functions. The former are functions in the domain of lag and the latter in the domain of series length.

The analyses carried out in this chapter could perhaps have been performed using the autocorrelation function. An inferred model structure could have been used to generate many sequences of the same length as the observed series. For each lag value the mean and standard deviation or if necessary, the full empirical distribution of the autocorrelation coefficient, could be determined and a theoretical autocorrelation function plot prepared in the same way as for the rescaled adjusted range. The sample autocorrelation function could then be superimposed on the theoretical function to see whether the model appeared to adequately preserve the observed autocorrelation. Such a procedure might be regarded as assisting in model identification as well as diagnostic checking, two tasks kept separate in the Box and Jenkins approach to modelling (see section 9.6 of this report).

The theoretical rescaled adjusted range and autocorrelation functions have been seen to be related. The expected value of the autocorrelation coefficient as well as presumably its sampling distribution depend only on the process autocorrelation structure. The expected value of the rescaled adjusted range appears also to have the same attribute whilst its sampling variance has been seen to be largely unaffected by changes in the marginal distribution of the process as seen in Chapter 10.

The close relationship between the two functions raises the question as to what additional insight is gained by the use of the rescaled adjusted range function. Indeed Hipel and McLeod (1978a) show that in a sample of 23 geophysical time series the construction

of autoregressive and mixed autoregressive-moving average models involving close attention to the preservation of the sample autocorrelation, led to the general preservation of the rescaled adjusted range. It can be argued that good modelling practice, where 'good' implies the paying of careful attention to autocorrelation without explicit consideration of the rescaled adjusted range, will in fact lead to models which perform satisfactorally as far as the rescaled adjusted range is concerned.

The value of the use of the rescaled adjusted range function in model building lies in the direct hydrological significance of the rescaled adjusted range statistic. Autocorrelation coefficient values measure the degree of association between series values at different lags. The rescaled adjusted range statistic however measures the consequence of the various degrees of association at all lags as they affect the amount of storage required to deliver a yield equal to the series mean. The modeller is able therefore to work directly with a statistic which is closely related to what is in many cases the aim of the modelling effort, that of generating realistic synthetic sequences for storage analysis. Even if storage analysis is not the aim, it is difficult to imagine a situation in time-series modelling where the preservation of the sample autocorrelation would not be desirable. The rescaled adjusted range function assists by providing an alternative integrated view of the sample autocorrelation.

It should be noted that in the case of two of the series analysed (Figures 11.17 and 11.26), the autocorrelation structures indicated by the autocorrelation and partial autocorrelation functions led to models giving apparently unsatisfactorily low values of the rescaled adjusted range. It is possible that more rigorous model identification and parameter estimation procedures would have

corrected this. At the very least the modeller is warned of apparent deficiencies in his model.

In the situation where a particular time series is to be modelled it would be desirable to be able to show by statistical test that a particular model preserved the rescaled adjusted range. However, the main interest in model performance as regards the rescaled adjusted range is at higher values of the sub-series length where only one independent value is available and statistical testing is not possible. The autocorrelation function also suffers this drawback as only one independent value of the autocorrelation coefficient is available at each lag for the particular sample series length. Statistical inference can be used however in both cases to reject, but not accept, the null hypothesis that the sample value is drawn from the theoretical population. Rescaled adjusted range function values falling more than, say, two standard deviations away from the mean theoretical value would be unlikely to occur by chance given the truth of the null hypothesis. This would indicate at the 95% significance level, given the approximate normality of the distribution of the rescaled adjusted range, that the particular model does not preserve the statistic.

This chapter has presented a rather exploratory examination of the usefulness of the rescaled adjusted range function in time series modelling. Model fitting has been of necessity approximate but hopefully adequate for the purpose of illustrating the worth of the approach. The remaining chapter will conclude this report with some general discussion on the rescaled adjusted range and the 'Hurst Phenomenon'.

CHAPTER 12: CONCLUDING REMARKS

12.1 Introduction

This report has presented a study which might be regarded as a general exploration of the usefulness of the rescaled adjusted range statistic in the field of stochastic hydrology. An inseparable part of the study has been an examination of the so-called 'Hurst Phenomenon'.

The main thrust of the study has been towards the comparison of observed and theoretical rescaled adjusted range values presented in Chapter 11. It is in this series of figures that the real value of the statistic R_n^{**} in synthetic hydrology can be seen.

This chapter will present some concluding remarks about the three main areas of interest in the study;

(i) Properties of the rescaled adjusted range which provide a basis for the comparisons of observed and theoretical rescaled adjusted range functions presented in Chapter 11.

(ii) The results of the comparisons presented in Chapter 11.

(iii) The 'Hurst Phenomenon'.

12.2 Properties of the Rescaled Adjusted Range

A large part of this report has been concerned with properties of the R_n^{**} statistic in theoretical processes and the structure of the processes themselves. This has been necessary to establish the validity of the comparisons made in Chapter 11 and the conclusions drawn from them.

The process models examined were of the general class of short-memory autoregressive and moving average models of which the simple lag-one Markov model is a member. These models are in

common use in synthetic hydrology. It has been repeatedly claimed in the literature that they are not capable of generating synthetic sequences which have R_n^{**} values similar to those observed in real data series.

It was shown in the report that for the processes examined the rescaled adjusted range is approximately normally distributed for medium to large values of series length n. This feature enables the construction of an approximate 95% confidence region around the expected values by drawing confidence lines at plus or minus two standard deviations distant from the expected values.

It was also shown for the processes examined that the expected values and variance of R_n^{**} are unaffected by the process mean or variance. The expected values of R_n^{**} are also unaffected, and the variance of R_n^{**} only slightly affected, by changes in process skewness. These properties of R_n^{**} mean that the rescaled adjusted range function of a process depends only on the autocorrelation structure of the process. The confidence region around the function is as well quite insensitive to factors other than the autocorrelation structure of the process. It is not necessary therefore to determine model parameters relating to process mean, variance and skewness when examining the rescaled adjusted range function of the process.

It was noted that the sample rescaled adjusted range function obtained from the observed series can be regarded as an alternative view of the underlying autocorrelation structure to that provided by the sample autocorrelation function.

12.3 Comparison of Observed and Theoretical Rescaled Adjusted Range Functions

The figures in Chapter 11 presenting comparisons of observed

and theoretical R_n^{**} functions show the sample function from the observed series superimposed on the approximate theoretical expected value function and its surrounding confidence region. The comparison does not provide a statistical test designed to prove that the proposed model preserves the rescaled adjusted range of the observed series, but it does give an indication of the model's performance in this regard.

Twenty different data series were examined in Chapter 11. They consisted of annual and standardised monthly river flows, annual rainfalls, North Finland tree ring indices and the Lake Saki mud varves. The autocorrelation and partial autocorrelation functions of each series were used to identify an appropriate short-memory process model.

Reasonable agreement was found between the observed and process R_n^{**} functions for sixteen of the twenty cases. In these cases the sample R_n^{**} values all lie within plus or minus two standard deviations of the estimated expected value of R_n^{**} for the process. In nine of the cases the sample R_n^{**} values lie largely within plus or minus one standard deviation.

In the remaining four cases it was found that sample R_n^{**} values fall outside the approximate 95% confidence region indicated by lines at plus or minus two standard deviations from the process expected value. The series in question are the annual flows and standardised monthly flows in the Macquarie River at Burrendong (Figures 11.17 and 11.21), the annual rainfalls at Balranald (Figure 11.26) and the series of Lake Saki mud varves (Figure 11.29). It was pointed out in the report that, for the first three of these series at least, the assumption of a moderate increase in the degree of autocorrelation above that indicated by the sample autocorrelation

and partial autocorrelation functions would lead to reasonable agreement between the sample and process R_n^{**} functions. The uncertainties surrounding the identification of the true underlying autocorrelation from the sample autocorrelation and partial autocorrelation functions were pointed out. The suggestion was made that in such cases, and where the storage character of generated synthetic sequences is of importance, the autocorrelation structure of the assumed model should perhaps be adjusted to give reasonable agreement between the sample and process R_n^{**} functions.

In the case of the Lake Saki mud varve series (Figure 11.29) the assumed 20-lag autoregressive model is clearly quite incapable of generating sequences having values of R_n^{**} large enough to match those of the observed series. The inadequacy of the model is apparent at values of series length n above about 1600. The distinguishing feature of the Lake Saki series is its very long length compared with the other series examined. The reasons for the inadequacy of the assumed model are certainly an interesting area of further investigations.

From this study it appears that short-memory autoregressive and moving average models are quite adequate in most cases for generating sequences which have realistic values of the rescaled adjusted range. This is particularly so for the range of series lengths likely to be of interest in hydrological design. In any case the comparison of sample and process R_n^{**} functions provides a method for judging the adequacy of the model in this regard.

It was also seen in the report that a valuable feature of the comparison of R_n^{**} functions is the ability to discriminate between various models proposed for a given series. This is clearly seen in Figures 11.18 and 11.19 for example, in which the advantage in using

the appropriate three-lag autoregressive model for the series of standardised monthly flows in the Kiewa River at Kiewa, rather than the single lag model, is quite apparent. The single lag model would generate sequences having unrealistically low values of R_n^{**} which could have serious implications for hydrological design.

12.4 The Relevance of the Rescaled Adjusted Range to Reservoir Storage Design

This study has been very much concerned with the problem of choosing models which generate synthetic sequences having realistic values of the rescaled adjusted range. The question which now arises is that of just how important is this aspect of model performance, particularly in the area of reservoir storage design.

The role of a stochastic model in reservoir storage design is to generate many realistic synthetic sequences, all equally likely to occur in the future. The length of each sequence might be set equal to some assumed 'economic' life of the project. The operation of the reservoir system can be simulated using each of the synthetic sequences and a relationship obtained between risk of failure to supply and storage size for some assumed release rule. This relationship can then be used in an economic study involving benefit and cost functions to determine the economic optimum storage size.

Now assume that the above-mentioned stochastic model preserves the rescaled adjusted range in the observed series. The many synthetic sequences generated by the model can be analysed and the value of R_n^{**} determined from each at n equal to the sequence length. Each value of R_n^{**} represents the minimum reservoir size required to deliver a constant supply equal to the mean inflow for the sequence. The mean of all the reservoir sizes determined from the

sequences will approximate $E(R_n^{**})$, the expected value of R_n^{**} for the stochastic process underlying the observed series. It is apparent therefore that the model preserves the 'storage character' of the observed series. However a knowledge of the value of $E(R_n^{**})$ gives no information about the probability that a reservoir of such size will fail to deliver a constant supply equal to the mean inflow in the case of a single sequence sampled from those generated by the model. It is this latter type of information that is required for economic decision making.

The value of preserving the rescaled adjusted range lies however in the need for <u>realistic</u> synthetic sequences in the design procedure. Fiering (1967) stated this point quite clearly in terms of the adjusted, but not rescaled, range R_n^*. "It is appropriate to repeat that no special nobility is ascribed to $E(R_n^*)$ as a design decision; rather any generating model recommended for design purposes should, as a matter of consistency, be capable of reproducing the essence of observed storage behaviour ....."

12.5 <u>The Fiering (1967) Approach to Synthesis of Streamflow Data</u>

Fiering (1967) discussed synthetic data generation in the context of the reservoir storage design problem. As the statement quoted in the previous section reveals, he saw the preservation of the adjusted range as a necessary attribute of a stochastic model if it is to produce realistic data to be used in economic decision making.

The technique presented in this study, of comparing R_n^{**} functions to establish whether or not a particular model appears to preserve the rescaled adjusted range, is to some extent a refinement of the approach used by Fiering.

Fiering saw a need to consider multi-lag autoregressive

models in some cases to preserve the adjusted range. He established a set of criteria unrelated to the adjusted range statistic for deciding the number of autoregressive terms to be used in the model. The maximum number of terms considered was twenty. Additional autoregressive terms were added until there was no improvement in the degree of multiple correlation between the actual terms in the data series and the preceding terms at the lags included in the model. Numerical instability in the calculated autoregressive parameters was another criteria for ceasing to add additional terms.

Fiering used the approach described above to fit models to fifteen data series. For each of the models identified he obtained by computer simulation the expected value of the adjusted range, $E(R_n^*)$, at the value of n equal to the length of the observed series. He prepared a plot of the log $E(R_n^*)/s$ versus log n values in the same manner as Hurst (1951). He found that the regression equation fitted by Chow (1951) to Hurst's data also was a close fit to his own. Fiering took this result to be an indication that models fitted using his technique would in general reproduce the 'Hurst Phenomenon' or, expressed differently, would preserve the adjusted range.

The difference between Fiering's approach and that presented in this study is that in the latter the ability of the individual model to preserve the rescaled adjusted range is explicitly examined.

Fiering's (1967) results have been incorrectly used by other authors to argue that short-memory models in general require very many autoregressive terms to preserve the rescaled adjusted range for even short series lengths. O'Connel (1977) makes the following statement. "In applying multi-lag autoregressive models Fiering (1967) found that he required a 20-lag model to ensure Hurst's law .... with $h \geq 0.5$ held for $n \leq 60$. Computational, not

statistical, grounds prevented an extension of this approach". A similar statement is also made by Mandelbrot and Wallis (1968). In fact Fiering may well have chosen other criteria leading to models with fewer autoregressive terms and still have found that they preserved the 'Hurst Phenomenon'. The results of this study show that simple short-memory models are in general capable of preserving the rescaled adjusted range over the series lengths of interest in hydrological design.

12.6 The Hurst Phenomenon

It would appear from the discussion in this report that much of the behaviour of the rescaled adjusted range in real data series can be simulated by simple short-memory stochastic models. It has been strongly argued in the report that such models are generally satisfactory for hydrological design purposes. It is not surprising however that simple stochastic models may not be able to reproduce all the complexities of real data series. Stochastic models are, of course, mere mathematical abstractions and their structures bear no relationship to the real physical processes they attempt to mimic. The important question is that of whether the synthetic data produced by the model is realistic enough to be useful for the purpose at hand.

Some authors have argued that there are important differences between the behaviour of the rescaled adjusted range in long series of real data and that in long synthetic series produced by short-memory process models. In particular it is claimed that the slope of the log R_n^{**} versus log n plot constructed for real series does not exhibit the convergence to an asymptotic slope of 0.5 that is exhibited by series produced by short-memory models. It has been illustrated in this study (Sections 5.3 and 8.3) that with short-memory processes the

convergence to 0.5 occurs very slowly and is slower with increasing process autocorrelation. Real data series can be thought of as one realisation of an underlying stochastic process. It was shown also (Section 8.1) that for single series realisations of short-memory processes, the sampling variability of R_n^{**} led to large fluctuations in the slope of the log R_n^{**} versus log n plot, particularly at larger n values. This was seen to make estimation of the true underlying slope quite difficult. In this study log R_n^{**} versus log n plots of five different series were examined. Two of the five plots (Figure 8.3) gave the visual impression of convergence of the slope to something of the order of 0.5. The other three plots (Figures 8.4 and 8.5), which included the results of the analysis of the longest series available, the Lake Saki mud varves (n = 4180), did not indicate convergence of the slope to 0.5. All plots showed large fluctuations in slopes at the higher n values.

From the results of this study it would appear that it can not be stated without doubt and as a general truth that short-memory processes fail to preserve the behaviour of the slope of the log R_n^{**} versus log n plot in long series of real data. However, should this be the case, the question remains as to the significance to hydrological design of such a failure.

One physically plausible reason why short-memory process models may not fully reproduce the behaviour of the rescaled adjusted range in long geophysical series data series is non-stationarity. It is possible that climatic changes occur more or less randomly bringing about abrupt shifts in the mean level of geophysical processes. Models of non-stationary stochastic processes have been proposed by Hurst (1957), Klemes (1974), Potter (1975) and Boes and Salas (1978). These models are capable of producing synthetic data series in which

convergence of the log R_n^{**} versus log n slope to a value of 0.5 occurs extremely slowly. These models have been proposed as alternatives to complex stationary process models such as fractional gaussian noise (Mandelbrot and Wallis, 1968).

In regard to the relative merits of non-stationary and stationary stochastic models, O'Connel (1977) makes the following statement. ".... non-stationarity is a rather intractable assumption if the ultimate aim is to generate synthetic flows; provided stationary stochastic processes which can reproduce the Hurst Phenomenon are available, these would appear to be more desirable for application in the planning of water resource systems, provided strong physical grounds do not inhibit their use". It has been shown in this report that simple stationary short-memory stochastic processes do, in many cases, "reproduce the Hurst Phenomenon" over the range of values of series length encountered in available hydrological records.

Concern with the slope of the log R_n^{**} versus log n plot at very large series lengths may be appropriate where it is desired to synthesise a very long record which is to preserve features consistent with the possible non-stationary nature of long geophysical series. This is not the usual aim of data synthesis for hydrological design. The usual aim is to produce many equally-likely sequences of some fairly short length equal perhaps to the assumed economic life of the project. The use of a stationary short-memory model in this setting, implies that the short-term past and the short term future may be assumed to be manifestations of a stochastic process which can be regarded as stationary at least over such a period of time. Attempts to use non-stationary models or complex models such as fractional gaussian noise in this setting imply that the uncertainties associated

with, say, climatic change in the future are to be gathered into the many short term futures used in the design process. The choice between the two approaches is a problematical one for the designer.

It is appropriate to conclude this discussion with the words of Klemes (1974), "There is no doubt about the importance of the Hurst Phenomenon. It seems, however, that its import is not in reducing the uncertainties of storage reservoir design but rather in helping us to understand them, to realise the complexity and unpredictability of hydrologic processes and the limits of our knowledge".

REFERENCES

AKAIKE, H. (1974), A new look at the statistical model identification. IEEE Trans. Automat. Contr., AC-19 (6), 716-723

ANIS, A.A. and LLOYD, E.H. (1976), The Expected Value of the Adjusted Rescaled Hurst Range of Independent Normal Summands. Biometrika, 63, 1, pp. 111-6

ANIS, A.A. and LLOYD, E.H. (1977), On the Distribution of the Hurst Range of Independent Normal Summands. International Institute for Applied Systems Research, RR-77-16

BARNES, F.B. (1954), Storage Required for a City Water Supply. J. Inst. Engrs. Aust., 26 (9)

BOES, D.C. and SALAS, J.D. (1978), Nonstationarity of the Mean and the Hurst Phenomenon. Water Resour. Res., 14 (1), 135-143

BOX, G.E.P., and JENKINS, G.M. (1970), Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

BRITTAN, M.R. (1961), Probability Analysis to the Development of a Synthetic Hydrology for the Colorado River, Part IV, Past and Probable Future Variations in Streamflow in the Upper Colorado River. Univ. of Colorado.

CARLSON, R.F., MACCORMICK, A.J.A. and WATTS, D.G. (1970), Application of Linear Random Models to Four Annual Streamflow Series, Water Resour. Res., 6 (4).

CHOW, V.T. (1951), Discussion of 'Long-term Storage in Reservoirs' by H.G. Hurst. Trans. Amer. Soc. Civil Eng., 116, 770-808.

DORAN, D.G. (1975), An Improvement to the Probabilistic Discrete State Modelling of Reservoir Behaviour. Univ. of New South Wales, Water Res. Lab. Report, 141.

FELLER, W. (1951), The Asymptotic Distribution of the Range of Sums of Independent Random Variables. Ann. Math. Statist., 22 427-432.

FIERING, M.B. (1967), Streamflow Synthesis. Macmillan, London.

GOMIDE, F.L.S. (1975), Range and Deficit Analysis using Markov Chains. Hydrol, Pap. 79, Colorado State Univ., Fort Collins.

HAZEN, A. (1914), Storage to be Provided in Impounding Reservoirs for Municipal Water Systems. Trans. ASCE Vol. 77, 1539.

HIPEL, K.W., McLEOD, A.I. and LENNOX, W.C. (1977), Advances in Box-Jenkins Modelling - 1. Model Construction. Water Resour. Res. 13 (3).

HIPEL, K.W. and McLEOD, A.I. (1978a), Preservation of the Rescaled Adjusted Range. 2, Similation Studies using Box-Jenkins Models. Water Resour. Res., 14 (3), pp. 509-518.

HURST, H.E. (1951), Long-term storage in reservoirs. Trans. Amer. Soc. Civil Eng., 116, 770-808.

HURST, H.E. (1957), A Suggested Statistical Model of some Time Series which occur in Nature. Nature, 180, 494.

KIRBY, W. (1972), Computer-Oriented Wilson-Hilferty Transformation that Preserves the First Three Moments and the Lower Bound of the Pearson Type 3 Distribution. Water Resour. Res., 8 (5), pp. 1251-1254.

KLEMES, V. (1974), The Hurst Phenomenon: A Puzzle?, Water Resour. Res., 10 (4) 675-688.

LAMB, H.H. (1977), Climate Present, Past and Future, Vol. 2. Climatic History and the Future. Methuen, London.

LANGBEIN, W.B. (1956), Discussion of 'Methods of Using Long-Term Storage in Reservoirs', by H.E. Hurst. Proc. Instn. Civ. Engrs., 1, pp. 565-568.

McLEOD, A.I. (1975), Derivation of the Theoretical Autocovariance Function of Autoregressive-Moving Average Time Series. J. Roy. Statist. Soc., Ser. C, 24 (2), 255-256.

McLEOD, A.I., HIPEL, K.W. and LENNOX, W.C. (1977), Advances in Box-Jenkins Modelling. 2, Application. Water Resour. Res., 13 (3), pp. 577-586.

McLEOD, A.I. and HIPEL, K.W. (1978a), Preservation of the Rescaled Adjusted Range.1, A Reassessment of the Hurst Phenomenon, Water Resour. Res., 14 (3), pp. 491-507.

McMAHON, T.A. and MILLER, A.J. (1971), Application of the Thomas and Fiering Model to Skewed Hydrologic Data. Water Resour. Res., 7 (5), pp. 1338-1340.

MANDELBROT, B.B. and WALLIS, J.R. (1968), Noah, Joseph and Operational Hydrology. Water Resour. Res., 4 (5), pp. 909-918.

MANDELBROT, B.B. and WALLIS, J.R. (1969b), Computer Experiments with Fractional Gaussian Noises Part 2, Rescaled Ranges and Spectra. Water Resour. Res., 5 (1).

MANDELBROT, B.B. and WALLIS, J.R. (1969d), Some Long Run Properties of Geophysical Records. Water Resour. Res., 5 (2), pp. 321-340.

MANDELBROT, B.B. and WALLIS, J.R. (1969e), Robustness of the Rescaled Range R/S in the Measurement of Noncyclic Long Run Statistical Dependence. Water Resour. Res., 5 (5), pp. 967-988.

MATALAS, N.C. (1967), Mathematical Assessment of Synthetic Hydrology. Water Resour. Res. 3 (4).

MATALAS, N.C. and HUZZEN, C.S. (1967), A Property of the Range of Partial Sums. International Hydrology Symposium, Fort Collins.

MORAN, P.A.P. (1954), A Probability Theory of Dams and Storage Systems. Aust. J. Appl. Sci. 5, pp. 116-124.

NELSON, C.R. (1973), Applied Time Series Analysis for Managerial Forecasting. Holden-Day, San Francisco.

O'CONNEL, P.E. (1977), ARIMA Models in Synthetic Hydrology, in 'Mathematical Models for Surface Water Hydrology'. Edited by T.A. CIRIANI, U. MAIONE and J.R. WALLIS, Wiley Interscience.

POTTER, K.W. (1975), Comment on 'The Hurst Phenomenon: A Puzzle?' by V. Klemes. Water Resour. Res., 11 (2), 373-374.

POTTER, K.W. (1976), Evidence for Nonstationarity as a Physical Explanation of the Hurst Phenomenon. Water Resour. Res., 12 (5), 1047-1052.

RIPPL, W. (1883), Capacity of Storage Reservoirs for Water Supply. Minutes of Proc. I.C.E., 71, pp. 270-278.

SCHOSTAKOVITSCH, W.B. (1934), Bodenablagerungen der Seen und Periodische Schwankungen der Naturerscheinungen. Memoirs of the Hydrologic Institute, Leningrad, 23.

SEN, Z. (1975), Small Sample Properties of Stationary Processes and the Hurst Phenomenon in Hydrology. Ph.D. thesis, 284 pp., Imperial Coll., London.

SEN, Z. (1977a), The Small Sample Estimation of h. Water Resour. Res., 13 (6), pp. 971-974.

SIDDIQUI, M.M. (1976), The Asymptotic Distribution of the Range and other Functions of Partial Sums of Stationary Processes. Water Resour. Res., 12 (6), pp. 1271-1276.

SIREN, G. (1961), Skogsgranstallen som Indikator fer Klimatfluktuationerna i norra Fennoskandien under Historisk tid. Helsingfors, Communicationes Instituti Forestalis Fenniae, 54 (2).

SRIKANTHAN, R. and McMAHON, T.A. (1977), Stochastic Modelling of Streamflows. A Symposium on Developments in Simulation Theory. Statistical Society of Aust. NSW Branch, and Aust. Society for Operations Research, Sydney Chapter.

SUDLER , C.E. (1927), Storage Required for Regulation of Streamflow. Trans. ASCE, Vol. 9.

TAQQU, M. (1970), Note on Evaluation of R/S for Fractional Noises and Geophysical Records. Water Resour. Res. 6 (11), pp. 349-350.

WALLIS, J.R. and MATALAS, N.C. (1970), Small Sample Properties of H and K - Estimators of the Hurst Coefficient h. Water Resour. Res., 6 (6), pp. 1583-1594.

WILSON, E.B. and HILFERTY, M.M. (1931), Distribution of Chi-Square. Proc. Nat. Acad. Science, 17, pp. 684-688.

WRIGHT, G.L. (1975), Multilag Markov Models for Eastern Australian Streams. Institution of Engineers Australia, National Committee on Hydrology, Hydrology Symposium.

YEVJEVICH, V. (1963), Fluctuation of Wet and Dry Years, 1, Research Data Assembly and Mathematical Models. Hydrol. Pap. 1, Colorado State Univ., Fort Collins.

YEVJEVICH, V. (1972), Stochastic Processes in Hydrology. Water Resources Publications, Fort Collins, Colorado.

APPENDIX COMPUTER PROGRAMS DEVELOPED FOR THIS STUDY

General Information

Computer Environment: Control Data CYBER installation at the University of New South Wales. KRONOS operating system. BATCH mode.

Language: FORTRAN IV

Program GENRATE

Program Description: Generates and writes to a file a nominated number of values from a specified lag-one Markov or ARMA (1,1) process.

Program Listing: See page 141.

Further Comments: Sub-routine MHNRAND returns normally distributed pseudo-random numbers of zero mean and unit variance and was a pre-existing routine used within the School of Civil Engineering, University of New South Wales. This routine contains non-standard FORTRAN IV and it can be replaced by an equivalent local library routine.

Program RANGE

Program Description: Reads the record to be analysed from a file. Determines the rescaled adjusted range and the Hurst coefficient K for the available non-overlapping sub-series at the nominated sub-series lengths. Determines mean values of the rescaled adjusted range and K at each nominated sub-series length.

Program Listing: See page 142

Program DSRGE

Program Description: Generates a nominated number of independent series of nominated lengths using a defined ARMA (1,1) process model or autoregressive process model of up to 20 terms. The process random component may be distributed like GAMMA with a given skewness. Values of the rescaled adjusted range and the Hurst coefficient K are determined for each series. The mean values, standard deviations and skewness of R_n^{**} and K at the nominated lengths n are then determined.

Program Listing: See pages 143-146.

Further Comments: For comments on sub-routine MHNRAND see previous comments on program GENRATE.

```
      PROGRAM GENRATE(OUTPUT,TAPE2=OUTPUT,TAPE3)
      COMMON INIT,V(100)
      DIMENSION AV(1000)
      NI=876
      INIT=1234567
      SMEAN=0.0
      STAND=1.0
      PHI=.92
      THETA=.70
      RONE=.28
      IGEN=1
      DO 100 I=1,100
      CALL MHNRAND(RN)
      V(I)=RN
  100 CONTINUE
      CALL DGDRAND(RN)
      RVNEW=RN
      AVOLD=0.
      DO 200 I=1,100
      RVOLD=RVNEW
      CALL DGDRAND(RN)
      RVNEW=RN
      IF(IGEN.EQ.1) WASTE=SMEAN+PHI*(AVOLD-SMEAN)+STAND*(RVNEW-THETA
     1*RVOLD)
      IF(IGEN.EQ.2) WASTE=SMEAN+RONE*(AVOLD-SMEAN)+RVNEW*STAND*SQRT
     1(1-RONE*RONE)
      AVOLD=WASTE
  200 CONTINUE
      DO 300 I=1,NI
      RVOLD=RVNEW
      CALL DGDRAND(RN)
      RVNEW=RN
      IF(IGEN.EQ.1) AV(I)=SMEAN+PHI*(AVOLD-SMEAN)+STAND*(RVNEW-THETA
     1*RVOLD)
      IF(IGEN.EQ.2) AV(I)=SMEAN+RONE*(AVOLD-SMEAN)+RVNEW*STAND*SQRT
     1(1-RONE*RONE)
      AVOLD=AV(I)
  300 CONTINUE
      WRITE (3,5) (AV(I),I=1,NI)
    5 FORMAT(6X,12F6.2,2X)
      STOP
      END
      SUBROUTINE MHNRAND (RANDNOR)
C
C     THIS ROUTINE GENERATES NORMALLY DISTRIBUTED PSEUDO-RANDOM NUMBERS
C     OF ZERO MEAN AND UNIT VARIANCE BY A MULTIPLICATIVE CONGRUENTIAL
C     PROCEDURE FOLLOWED BY A REVERSE BOX-MULLER TRANSFORMATION.
C
C     INITIALISING NUMBER SHOULD BE AN ODD INTEGER
C
      COMMON INIT,V(100)
      NINT = SHIFT(INIT,10)
      INT = INIT+INIT+INIT+NINT
      INT = INT .AND. 000037777777777777B
      REALNO = FLOAT(INT)*2.0**(-47)
      NINT = SHIFT(INT,10)
      INIT = INT+INT+INT+NINT
      INIT = INIT .AND. 000037777777777777B
      REALNP = FLOAT(INIT)*2.0**(-47)
      RANDNOR = SIN(REALNO*6.2831853071796)*SQRT(-2.0*ALOG(REALNP))
      RETURN
      END
      SUBROUTINE DGDRAND(R)
      COMMON INIT,V(100)
      IR = 100*RANF(AK) + 1
      R = V(IR)
      CALL MHNRAND(RR)
      V(IR) = RR
      RETURN
      END
```

```
      PROGRAM RANGE(OUTPUT,TAPE2=OUTPUT,TAPE3,TAPE4)
      DIMENSION QTS(4500),IN(100),X(4500) ,IFMT(3)
C
      READ (3,11) TITLE
      READ (3,10) NI
      READ (3,10) NT
      READ (3,10) (IN(I),I=1,NT)
      READ(3,10) ILIST
      READ (3,11) (IFMT(I),I=1,3)
      READ (4,IFMT) (QTS(I),I=1,NI)
C
      WRITE (2,19) TITLE
      WRITE (2,20) NI
      WRITE (2,21)
      WRITE (2,23) (IN(I),I=1,NT)
      WRITE (2,22)
C
      DO 900 I=1,NT
      RSUM=0.
      NA=IN(I)
      KN=NI/NA
      KR=0
      DO 800 K=1,KN
      DO 700 J=1,NA
      KR=KR+1
      X(J)=QTS(KR)
  700 CONTINUE
      CALL HART(X,NA,RS)
      IF(KN.GT.1.AND.ILIST.EQ.1) WRITE(2,35)RS
      RSUM=RSUM+RS
  800 CONTINUE
      RSUM=RSUM/KN
      WRITE (2,30) NA,RSUM,KN
  900 CONTINUE
C
   10 FORMAT (16I5)
   11 FORMAT (3A2)
   19 FORMAT(10X,'FILE  ',3A10,/)
   20 FORMAT(10X,'NUMBER OF  ITEMS IN SERIES',I5)
   21 FORMAT(10X,'DETERMINATION OF RESCALED ADJUSTED RANGE FOR N =')
   22 FORMAT(//,11X,'N',11X,' R/S  ',8X,'MEAN OF K VALUES',/,37X,'K',/)
   23 FORMAT(54X,6I5)
   35 FORMAT(20X,F10.2)
      STOP
      END
      SUBROUTINE HART(X,NA,RS)
      DIMENSION X(1)
      CALL MDEV(X,NA,XM,XD)
      SUMDEV=0.
      SURPLS=0.
      DEFICT=0.
      DO 900 I=1,NA
      SUMDEV=SUMDEV-XM+X(I)
      IF(SUMDEV.GT.SURPLS) SURPLS=SUMDEV
      IF(SUMDEV.LT.DEFICT) DEFICT=SUMDEV
      RGE=SURPLS-DEFICT
      RS=RGE/XD
  900 CONTINUE
      RETURN
      END

      SUBROUTINE MDEV (X,NA,XM,XD)
      DIMENSION X(1)
      SUMM=0.
      SUMD=0.
      DO 100 I=1,NA
      SUMM=SUMM+X(I)
  100 CONTINUE
      XM=SUMM/FLOAT(NA)
      DO 200 I=1,NA
      SUMD=SUMD+(X(I)-XM)*(X(I)-XM)
  200 CONTINUE
      SUMD=SUMD/FLOAT(NA-1)
      XD=SQRT(SUMD)
```

```
      PROGRAM DSRGE9(OUTPUT,TAPE2=OUTPUT,TAPE3,TAPE4,TAPE1)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      COMMON WHA,WHB,WHG
       DIMENSION VAR(4100),LSEQJ(20),RFILE(4100),PHIJ(20),VARP(20)
      DIMENSION RSUM(20),RSUM2(20),RSUM3(20),RLSUM(20),RLSUM2(20)
     1,RLSUM3(20)

      SAMPLING STRUCTURE

      READ(1,52) NTEST,NTFILE,LSEQT,NSEQ
      READ (1,52) (LSEQJ(I),I=1,NTEST)

      MODEL STRUCTURE

      IGEN=1,2,3 --ARMA(1,1),AR(1,0),AR(K,0)
      STAND IS ESTIMATED SERIES STD. DEV.
      SDRAN IS MODEL DEPENDENT SCALING FACTOR FOR STD.DEV. OF RANDOM
      VARIATE E.G SQRT(1-PHIJ(1)**2) FOR ARMA(1,0)
      NPHI IS NUMBER OF AUTOREGRESSIVE TERMS
      READ(1,52) IGEN
      READ(1,52) NPHI
      READ(1,54) (PHIJ(I),I=1,NPHI)
      READ(1,54) SMEAN,STAND,SDRAN
      READ(1,54)THETA
      READ(1,54) GAMA
      READ(1,57) WHA,WHB,WHG
      INIT=1234567
      IF(NTFILE.NE.0) WRITE(2,30) LSEQJ(NTFILE)
      WRITE(2,45) IGEN,NPHI,SMEAN,STAND,THETA,SDRAN,(PHIJ(I),I=1,NPHI)
      WRITE(2,46) GAMA
      WRITE(2,58) WHA,WHB,WHG
      WRITE (2,47)
      WRITE(2,10)

      INITIAL LOADING OF RANDOM VARIATE VECTOR V(I)

      DO 100 I=1,100
      CALL MHNRAND(RN)
      V(I)=RN
  100 CONTINUE

      INITIAL LOADING OF VARP(I)=SMEAN
      DO 65 I=1,20
      VARP(I)=SMEAN
   65 CONTINUE
      CALL GENRATE(PHIJ,200,VAR,VARP)
      WRITE(4,40) (VAR(LL),LL=1,200)
      DO 67 I=1,20
      K=I+150
      VARP(I)=VAR(K)
   67 CONTINUE
      KK=0
      M=0
      DO 50 L=1,NTEST
      RSUM(L)=0.
      RSUM2(L)=0.
      RSUM3(L)=0.
      RLSUM(L)=0.
      RLSUM2(L)=0.
      RLSUM3(L)=0.
   50 CONTINUE

      MAIN LOOP

      DO 900 I=1,NSEQ
      CALL SEED(INIT)

  GENRATE SINGLE SEQUENCE LENGTH LSEQT

      CALL GENRATE (PHIJ,LSEQT,VAR,VARP)

      ACUMULATE STATISTICS FOR EACH SUBSERIES-LENGTH

      DO 600 L=1,NTEST
      XSUM=0.
      XSUM2=0.
      LSEQ=LSEQJ(L)
      DO 300 J=1,LSEQ
      X=VAR(J)
      XSUM=XSUM+X
      XSUM2=XSUM2+X*X
  300 CONTINUE
      SN=FLOAT(LSEQ)
      XMEAN=XSUM/SN
      XSTD=SDEVR(SN,XSUM2,XMEAN)
```

```
      SUMDEV=0.
      SURPLS=0.
      DEFICT=0.
      DO 400 J=1,LSEQ
      X=VAR(J)
      SUMDEV=SUMDEV-XMEAN+X
      IF(SUMDEV.GT.SURPLS) SURPLS=SUMDEV
      IF(SUMDEV.LT.DEFICT) DEFICT=SUMDEV
      R=SURPLS-DEFICT
  400 CONTINUE
      R=R/XSTD
      IF(NTFILE.NE.L) GO TO 450
      M=M+1
      IF(NTFILE.EQ.L) RFILE(M)=R
  450 RSUM(L)=RSUM(L)+R
      RSUM2(L)=RSUM2(L)+R*R
      RSUM3(L)=RSUM3(L)+R*R*R
      RL=ALOG(R)/ALOG(SN/2.)
      RLSUM(L)=RLSUM(L)+RL
      RLSUM2(L)=RLSUM2(L)+RL*RL
      RLSUM3(L)=RLSUM3(L)+RL*RL*RL
  600 CONTINUE
  900 CONTINUE


      SN=FLOAT(NSEQ)
      DO 700 L=1,NTEST
      LSEQ=LSEQJ(L)
      RMEAN=RSUM(L)/SN
      RSTD=SDEVR(SN,RSUM2(L),RMEAN)
      RSKEW=SKEW(SN,RSUM3(L),RSUM2(L),RMEAN)
      RLMEAN=RLSUM(L)/SN
      RLSTD=SDEVR(SN,RLSUM2(L),RLMEAN)
      RLSKEW=SKEW(SN,RLSUM3(L),RLSUM2(L),RLMEAN)
      WRITE(2,20)LSEQ,RMEAN,RSTD,RSKEW,RLMEAN,RLSTD,RLSKEW,NSEQ
   20 FORMAT (I6,6X,3F12.4,3F12.5,I7)
   30 FORMAT(/,' R/S VALUES FOR N=',I5,' WRITTEN TO TAPE3(10F8.3)',//)
   40 FORMAT(10F8.3)
   45 FORMAT(3X,'IGEN = 1,2,3  -- ARMA(1,1),AR(1,0),AR(NPHI,0)',//,
     13X,'IGEN = ',I2,7X,'NPHI = ',I2,//,3X,'SERIES MEAN',4X,'SERIES ST.
     2DEV.',3X,'THETA',4X,'ST.DEV.OF RANDOM VARIATE',//,3X,F11.3,5X,
     3F11.3,5X,F5.3,14X,F6.3,//,3X,'PHI(NLAG)',/,8F10.3,//)
   46 FORMAT (/,3X,'SKEWNESS OF RANDOM VARIATE = ',F5.3)
   47 FORMAT(/,'  SAMPLE OF 200 GENERATED VARIABLES WRITTEN TO TAPE4')
   52 FORMAT(16I5)
   54 FORMAT(16F5.2)
   57 FORMAT(3F10.3)
   58 FORMAT(3X,'MODIFIED WILSON-HILFERTY TRANSFORMATION',
     1'PARAMETERS (KIRBY-W.RES.RES 8(5)-1972)---A,B,G,',3F10.5)
   10 FORMAT (//,5X,'N',9X,'MEAN R/S',2X,'STD.DEV R/S',4X,'SKEW R/S',
     13X,'MEAN  K  ',2X,'STD.DEV  K  ',3X,'SKEW  K  ','  SAMPLE SIZE')
  700 CONTINUE
      IF(NTFILE.NE.0) WRITE(3,40) (RFILE(I),I=1,NSEQ)
      STOP
      END
      FUNCTION SDEVR(SN,SUMX2,XMEAN)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      COMMON WHA,WHB,WHG
      SDEVR=SQRT((SUMX2-SN*XMEAN*XMEAN)/(SN-1.0))
      RETURN
      END
      FUNCTION SKEW(SN,SUMX3,SUMX2,XMEAN)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      COMMON WHA,WHB,WHG
      SKEW=SUMX3+3.0*SN*(XMEAN**3.0)-3.0*XMEAN*SUMX2-SN*(XMEAN**3.0)
      SKEW=SKEW*SN/((SN-1.0)*(SN-2.0))
      SKEW=SKEW/((SDEVR(SN,SUMX2,XMEAN))**3.0)
      RETURN
      END
      SUBROUTINE GENRATE(PHIJ,NI,AV,AVP)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      COMMON WHA,WHB,WHG
      DIMENSION AV(4100) ,PHIJ(20),AVP(20),TEMP(20)
      IF(IGEN.EQ.3) GO TO 400

      DGRAND(AN) RETURNS SINGLE RANDOM (0,1) VARIATE
      CALL DGRAND(RN)
      RVNEW=RN
      AVOLD=AVP(1)
      DO 200 I=1,50
      RVOLD=RVNEW
      CALL DGRAND(RN)
      RVNEW=RN
```

```
      IF(IGEN.EQ.1) WASTE=SMEAN+PHIJ(1)*(AVOLD-SMEAN)+STAND*
     1SDRAN*(RVNEW-THETA*RVOLD)
      IF(IGEN.EQ.2) WASTE=SMEAN+PHIJ(1)*(AVOLD-SMEAN)+RVNEW*STAND
      AVOLD=WASTE
  200 CONTINUE
      DO 300 I=1,NI
      RVOLD=RVNEW
      CALL DGDRAND(RN)
      RVNEW=RN
      IF(IGEN.EQ.1) AV(I)=SMEAN+PHIJ(1)*(AVOLD-SMEAN)+STAND*(RVNEW-THETA
     1*RVOLD)*SDRAN
      IF(IGEN.EQ.2) AV(I)=SMEAN+PHIJ(1)*(AVOLD-SMEAN)+RVNEW*STAND*SDRAN
      AVOLD=AV(I)
  300 CONTINUE
      GO TO 900
  400 CONTINUE
      NT=NI+50
      DO 700 I=1,NT
      CALL DGDRAND(RN)
      RVNEW=RN
      AVAL=SMEAN+STAND*SDRAN*RVNEW
      DO 550 JJ=1,NPHI
      AVAL=AVAL+(AVP(JJ)-SMEAN)*PHIJ(JJ)
  550 CONTINUE
      DO 600 JJ=1,NPHI
      TEMP(JJ)=AVP(JJ)
  600 CONTINUE
      DO 650 JJ=2,NPHI
      AVP(JJ)=TEMP(JJ-1)
  650 CONTINUE
      AVP(1)=AVAL
      IF(I.GT.50) AV(I-50)=AVAL
  700 CONTINUE
  900 CONTINUE
      RETURN
      END
      SUBROUTINE MHNRAND (RANDNOR)

      THIS ROUTINE GENERATES NORMALLY DISTRIBUTED PSEUDO-RANDOM NUMBERS
      OF ZERO MEAN AND UNIT VARIANCE BY A MULTIPLICATIVE CONGRUENTIAL
      PROCEDURE FOLLOWED BY A REVERSE BOX-MULLER TRANSFORMATION.

      INITIALISING NUMBER SHOULD BE AN ODD INTEGER

      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      COMMON WHA,WHB,WHG
      NINT = SHIFT(INIT,10)
      INT = INIT+INIT+INIT+NINT
      INT = INT .AND. 000037777777777777B
      REALNO = FLOAT(INT)*2.0**(-47)
      NINT = SHIFT(INT,10)
      INIT = INT+INT+INT+NINT
      INIT = INIT .AND. 000037777777777777B
      REALNP = FLOAT(INIT)*2.0**(-47)
      RANDNOR = SIN(REALNO*6.283185307796)*SQRT(-2.0*ALOG(REALNP))

      WILSON-HILFERTY TRANSFORMATION(KIRBY'S MODIFICATION)

      IF(GAMA.EQ.0.) GO TO 100
      H=(WHB-(2.0/GAMA)/WHA)**.3333333
      HT=1.0-(WHG/6.0)*(WHG/6.0)+(WHG/6.0)*RANDNOR
      IF(HT.GT.H) H=HT
      RANDNOR=WHA*(H**3.0-WHB)
  100 CONTINUE
      RETURN
      END
      SUBROUTINE DGDRAND(R)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      COMMON WHA,WHB,WHG
      IR = 100*RANF(AK) + 1
      R = V(IR)
      CALL MHNRAND(RR)
      V(IR) = RR
      RETURN
      END
      SUBROUTINE SEED(ITS)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      COMMON WHA,WHB,WHG
      S=SECOND(T)*1000.
      ITS=INT(S)
      ITS=ITS+3333333
      IF(MOD(ITS,2).EQ.0) ITS=ITS+1
      RETURN
      END
```

```
      CALL DGDRAND(RN)
      RVNEW=RN
      AVAL=SMEAN+STAND*SDRAN*RVNEW
      DO 550 JJ=1,NPHI
      AVAL=AVAL+(AVP(JJ)-SMEAN)*PHIJ(JJ)
  550 CONTINUE
      DO 600 JJ=1,NPHI
      TEMP(JJ)=AVP(JJ)
  600 CONTINUE
      DO 650 JJ=2,NPHI
      AVP(JJ)=TEMP(JJ-1)
  650 CONTINUE
      AVP(1)=AVAL
      IF(I.GT.50) AV(I-50)=AVAL
  700 CONTINUE
  900 CONTINUE
      RETURN
      END
      SUBROUTINE MHNRAND (RANDNOR)
C
C     THIS ROUTINE GENERATES NORMALLY DISTRIBUTED PSEUDO-RANDOM NUMBERS
C     OF ZERO MEAN AND UNIT VARIANCE BY A MULTIPLICATIVE CONGRUENTIAL
C     PROCEDURE FOLLOWED BY A REVERSE BOX-MULLER TRANSFORMATION.
C
C     INITIALISING NUMBER SHOULD BE AN ODD INTEGER
C
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      NINT = SHIFT(INIT,10)
      INT = INIT+INIT+INIT+NINT
      INT = INT .AND. 00003777777777777777B
      REALNO = FLOAT(INT)*2.0**(-47)
      NINT = SHIFT(INT,10)
      INIT = INT+INT+INT+NINT
      INIT = INIT .AND. 00003777777777777777B
      REALNP = FLOAT(INIT)*2.0**(-47)
      RANDNOR = SIN(REALNO*6.2831853071796)*SQRT(-2.0*ALOG(REALNP))
C
C     WILSON-HILFERTY TRANSFORMATION
C
      IF(GAMA.EQ.0.) GO TO 100
      RANDNOR=(2./GAMA)*((1.+GAMA*RANDNOR/6.-GAMA*GAMA/36.)**3.-1.)
  100 CONTINUE
      RETURN
      END
      SUBROUTINE DGDRAND(R)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      IR = 100*RANF(AK) + 1
      R = V(IR)
      CALL MHNRAND(RR)
      V(IR) = RR
      RETURN
      END
      SUBROUTINE SEED(ITS)
      COMMON INIT,V(100),SMEAN,STAND,THETA,NPHI,SDRAN,GAMA,IGEN
      S=SECOND(T)*1000.
      ITS=INT(S)
      ITS=ITS+3333333
      IF(MOD(ITS,2).EQ.0) ITS=ITS+1
      RETURN
      END
```