University of Manitoba

UNCERTAINTY IN FLOOD RISK ANALYSIS

by

LEONARD M.F.P. LYE

A THESIS SUBMITTED TO THE FACULTY OF
GRADUATE STUDIES IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Department of Civil Engineering
Winnipeg, Manitoba

November, 1987

ISBN 0-315-44117-8

UNCERTAINTY IN FLOOD RISK ANALYSIS

BY

LEONARD M.F.P. LYE

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

This thesis is concerned with four aspects of flood risk analysis. The use of Bayesian estimation theory is central in all four aspects.

The first aspect concerns the parameter estimation for probability distributions used for flood data. It will be shown that a Bayesian approach gives better estimates than the usually preferred method of maximum likelihood. Lindley's approximation technique greatly simplifies the computation of all Bayes estimates.

The second aspect concerns the estimation of the probability that a flood will be exceeded in a future period. Then the uncertainty in the parameters of the probability distribution must be taken into account. This is done by using the predictive distribution as distinct from the descriptive distribution.

Next the customary assumption of stochastic independence for annual flood peak series is waived. A calculation of the Hurst statistic for about fifty annual flood series from all over Canada indicates that long term serial correlation is present in many rivers. This is shown to increase the uncertainty of the sample statistics and leads to a substantial upward assessment of the flood risk. A simple but efficient technique of modelling series with a high Hurst statistic is described.

The fourth aspect of flood risk analysis is an attempt at reducing the uncertainty in the estimation of the probability of exceedence of extreme floods. Taking the Red River at Emerson as a case study, a physically-based stochastic flood simulation model is developed using soil moisture, snowfall, snowmelt, and rainfall as input. The predictive distribution of flood peaks obtained from this model shows less uncertainty than the predictive distribution based only on the recorded flood peaks. This is not necessarily the whole answer. Updating the predictive distribution with historic or regional information using the simulation model is still possible, but has not been attempted in this thesis.

The research described in the thesis shows that parameter uncertainty appears to be more important than the question of plotting positions, parameter estimation by one method or another, or the choice between 2-parameter and 3-parameter probability models.

TO MY LATE GRANDFATHER,

MR. JOHN JOSEPH LYE KON LEN

(1903 - 1987)

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Professor Caspar Booy, for his constant encouragement and guidance throughout the preparation of this thesis. In addition to providing financial support, he has taught me not only about hydrology and statistics, but through his own example, humility.

I am indebted to Professor S.K. Sinha whose helpful suggestions have contributed significantly towards this thesis. Special thanks to Professor B.D Macpherson and Professor Ian Goulter for their helpful comments and suggestions.

My gratitude to my friends Ansari Khan, Mike Buchko, Musfique Ahmed, Wendy Severson, Muriel Innes, Rob Boswick and many others for their help, friendship and laughter.

To my grandmother, my parents, brother and sisters, I would like to thank them for their constant moral support.

I would also like to mention my children, Jeannie and Sandy, whose love and trust have been an immense motivation.

My sincere thanks to my wife, Brenda, for her unfailing understanding and sharing of my hopes and ambitions.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 CONTEXT AND OBJECTIVES

Floods have always been a recurring menace in most inhabited parts of the world. While protective works have been designed from time immemorial, the risk of flooding can seldom be eliminated. An essential step in a flood protection study is therefore an assessment of flood frequencies aimed at determining the probability that peak flows of various magnitude will be exceeded.

The procedure traditionally involves:

(i) Obtaining a record of annual flood events (usual length is less than 50 years in Canada).

(ii) Choosing a probability distribution that seems to fit the data. This is usually done by arranging the observed data in order of magnitude, plotting them on different types of probability graph paper and, observing the shape suggested by the plotted points.

(iii) Estimating the parameters of the distribution, and

(iv) Making inferences about the occurrence of future flood events.

Within steps (ii) and (iii) there exist a number of sources of uncertainty. Three different types of uncertainty may be distinguished (Benjamin and Cornell, 1970):

(a) Stochastic or natural uncertainty of the probabilistic phenomenon itself, here, the annual flood peaks.

(b) Parameter uncertainty, caused by limited data and aggravated by serial correlation.

(c) Model uncertainty, associated with the uncertainty of the best model to describe the stochastic process.

How to deal with the three types of uncertainty depends on what one wants to analyse the flood data for, i.e. to understand or to predict. The purpose of the analysis may be simply to extend our knowledge of the flood potential of the river. This is usually the goal of a scientist-statistician who is after the probability distribution that best describes the stochastic variability, the first source of uncertainty.

The scientist is, of course, also interested in the confidence one may have in the probability distribution, that is, in the second type of uncertainty. But the two types of uncertainty must be left entirely separate. The first type is an objective issue, being a property of nature. The second type depends on the knowledge and information available. As such, it depends to a certain

extent on the investigators and is partly subjective; hence the classical statistician treats the two types of uncertainty quite differently.

The confidence one has in the chosen probability distribution to describe the stochastic variability of flood peaks depends on how well its parameters are estimated. The method of maximum likelihood is usually preferred by most scientists/hydrologists.

The first objective of this study is to show that better estimates of the parameters can be obtained by using the Bayesian approach.

Lindley's Bayesian approximation procedure (Lindley, 1980) which greatly simplifies computation is used to obtain all Bayes estimates. A 2-parameter distribution (Gumbel) and a 3-parameter distribution (3-parameter lognormal) fitted to some Canadian rivers are used as examples to demonstrate this technique.

The purpose of the analysis may also be to guide engineering decisions. The engineer, unlike the scientist, must go beyond a mere description of the randomness of nature and of the limits of our knowledge. He must make decisions based on predictions. He must answer the question: Is that dyke high enough to provide flood protection within the planning horizon of, say, the next 50 or 100 years? And then he or she must act on that prediction. It makes no difference whether a dyke fails because nature produced a very unusual event or because

the engineer underestimated the design flood. Engineers therefore cannot separate the first two types of uncertainty. They must be combined in the decision making process. The use of Bayesian probability theory permits one to combine stochastic and parameter uncertainty, provided that both can be quantified.[1] The result is the predictive distribution of the flood peaks (Zellner, 1971). Its frequency curve is steeper (degree depending on the length of record) than that of the descriptive probability distribution, which merely describes the

---

[1] A careful distinction must be made between two kinds of uncertainty. On the one hand, there is uncertainty that can be quantified objectively on the basis of available information. One can deal with it rationally. On the other hand, there is ignorance, leading to a different kind of uncertainty. This uncertainty cannot be quantified objectively or meaningfully. Stochastic uncertainty belongs to the first category. So does parameter uncertainty. There is no reason in the theory of mathematical statistics why the model parameters, such as the mean and the standard deviation, cannot be treated as random variables in an appropriate statistical model. And there is no physical reason why their variability cannot be quantified objectively, on the basis of observations, provided a reasonable analogy exists between the mathematical model and the physical reality.

Uncertainty regarding the appropriate type of probability distribution belongs to the second category. The guiding principle here is to choose the simplest model that is compatible with the information one has about the entire class of flood peak series to which the flood peak series we are interested in belongs and that is capable of adequately reproducing the significant features of that individual flood peak series.

The choice between a two-parameter and a three-parameter model can not be based on a balance of probabilities, at least not with the current state of relevant information. It is also not meaningful to say that there is a 50% probability that the flood peaks on Moose Jaw Creek follow a Gumbel distribution and a 50% probability that they are log-normally distributed, simply because one does not know which model to choose.

stochastic state of nature on the basis of available information.

Obtaining predictive distributions for continuous distributions requires rather sophisticated mathematics. However, one can make use of the property that the probability of exceedence of the predictive probability distribution is the Bayes estimator of the probability of exceedence of a flood discharge under the squared-error loss function (Sinha, 1985).

The second objective of this study is therefore to obtain Bayes estimates of the probability of exceedence.

Lindley's approximation is used to obtain the Bayes estimates and this procedure is demonstrated for some Canadian rivers fitted to two 2-parameter distributions, the Gumbel distribution and the 2-parameter lognormal distribution. The flood data are assumed to be serially independent.

From an engineering point of view, a nice feature of the predictive probability distribution is that it automatically provides a safety factor which is substantial when the record is short and the safety factor becomes small when the frequency curve is based on a large number of data.

Engineers, however, use the descriptive rather than the predictive probability distribution neglecting the uncertainty in the information. This seems strange, for in almost every other field of engineering the

uncertainty caused by our limited knowledge is not ignored but is the reason for the use of a safety factor. This state of affairs is probably partly caused by the fact that statistics courses for engineers at the Universities are mostly designed for and taught by scientists instead of engineers. Another reason is undoubtedly the confidence engineers have in the accuracy of the descriptive probability distribution based on the available information. That confidence seems justified at first glance since it can be shown that the effect of parameter uncertainty on flood risk is relatively small if one has a reasonable length of record, and if the flood data are serially independent which is the standard assumption in most analysis.

One must, however, seriously question the assumption that annual peak flows are serially independent for practically all rivers. It is true that for annual flood peak series the low lag serial correlation coefficients are usually small. As such any observed serial correlation seldom, if ever, passes the customary tests of statistical significance at the 10% or 5% level. But an analysis of about 50 flood peak series from all over Canada indicated that while short term serial correlation seems absent, significant long term serial correlation as measured by a high Hurst statistic (Hurst, 1951) is present in a large number of rivers analysed, and this serial correlation substantially increases the uncertainty of the parameters

of a flood probability distribution.

The third objective of this study is to demonstrate the effect of serial correlation on flood risk analysis.

This is demonstrated through the use of a discretised predictive probability distribution proposed by Russell (1982). This requires, however, a time series model that will reproduce both the short term and the long term serial correlation structure in addition to the marginal distribution properties. A number of models are available to do this; namely, the Fast Fractional Gaussian Noise model, the ARMA (1,1) model, the Broken-line model and the ARMA-Markov model. But these models are either too cumbersome to use because of their complexity or they require computer time far in excess of that required for autoregressive models.

The fourth objective of this study is, therefore to develop a simple and efficient time series model with the desired attributes.

Thus far, the underlying probability distribution of the annual flood peaks has to be determined 'a priori'. The model chosen may not be the most suitable model. The probability distribution derived from observed flood records may be reasonably well defined in the middle reach where many observations are located. However, in the upper tail on which flood protection decisions are often based, there is considerable uncertainty due to the lack of data in this region. One can reduce this uncertainty only

by obtaining additional information. Such additional information can be obtained by a study of the physical factors that determine the magnitude of flood peaks. An attempt is made to construct a simulation model in which additional information about the factors is combined with a knowledge of the physics of the runoff process. Sampling from this simulation model provides a probability distribution that can be expected to be more reliable than the distribution based on the record only. The Red River at Emerson is used as a case study. The simulation model also sheds some light on the possible causes of significant long term serial correlation in the annual spring peak flows on the Red River.

The final objective of this study is the development of such a physically-based simulation model.

1.2 OUTLINE OF THESIS

The study was performed by first outlining some basic principles of Bayesian analysis and to point out distinctive differences between the sampling theory (classical statistics) and Bayesian methods of inference. This is given in Chapter Two. In Chapter Three, the parameters and T-year floods of a 2-parameter and a 3-parameter probability distributions are obtained using the maximum likelihood method and Bayesian method and

their results compared. In Chapter Four, flood frequency analysis using the predictive probability distribution approach is considered. Chapter Five presents an analysis of the serial correlation structure of about 50 flood peak series from all over Canada. The development of a new time series model capable of reproducing the Hurst effect as well as short term serial correlation structure is presented in Chapter Six. Chapter Seven presents the study on the effect of serial correlation on flood risk analysis. The development and evaluation of the physically-based flood simulation model for the Red River is given in Chapter Eight. Chapter Nine presents conclusions and recommendations from the study.

CHAPTER 2

BAYESIAN ANALYSIS

2.1 GENERAL

Detailed discussions of the Bayesian approach to statistical inference are given in Jeffreys (1961), Box & Tiao (1973) and Zellner (1971).

This chapter will introduce only some of the basic principles and concepts of Bayesian analysis. Some important differences between the classical approach and the Bayesian approach to statistical inference are outlined and the relevance of the Bayesian approach to flood risk analysis discussed.

2.2 BAYES' THEOREM

This theorem is named after Reverend Thomas Bayes (1702 - 1761). It is derived, in fact, from a basic law of probability theory and is regarded by a growing number of statisticians and engineers as being fundamental to the revision of probability in the light of additional evidence.

Bayes' theorem, which follows from the definition of conditional probability, involves a prior (or a priori)

distribution based on theoretical considerations or on the investigator's own beliefs about the possible states of nature; the prior probabilities are not necessarily associated with repeatable experiments or the analogs thereof. This distribution describes all the relevant information prior to the receipt of a sample of data or additional information as appropriate. Given the prior distribution and the additional information, by combining these and using Bayes' theorem, the posterior (or a posteriori) distribution could be evaluated. The posterior distribution then embodies all the available information about the state of nature.

Bayes' theorem is derived as follows:

Let θ_1, θ_2, θ_3 . . . . θ_m, denote all possible states of nature which may refer to the state of weather, water level in the reservoir or any other variable or parameter which is subjected to uncertainty and let x represent a sample of data (additional information). The prior probabilities estimated before the receipt of the data can be expressed by $P'(\theta_i)$ and the conditional probabilities of the sample x subject to the states of nature θ, are denoted by $P(x/\theta_i)$. Also, let the posterior probabilities $P''(\theta_i/x)$ represent the probabilities of the states of θ_i of nature, given the sample x. If $P(\theta_i, x)$ denotes the joint probability of θ_i and x, by using conditional probabilities, it can be stated that:

$$P(\theta_i, x) = P'(\theta_i) \cdot P(x/\theta_i)$$

$$= C(x) \cdot P''(\theta_i/x)$$

where the normalizing constant C(x) is given by:

$$C(x) = \sum_{i=1}^{m} P'(\theta_i) \cdot P(x/\theta_i) \quad \ldots (2.1)$$

This leads to Bayes' theorem:

$$P''(\theta_i/x) = \frac{P'(\theta_i) \cdot P(x/\theta_i)}{\sum_{i=1}^{m} P'(\theta_i) \cdot P(x/\theta_i)} \quad \ldots (2.2)$$

Note that (2.2) should be changed to integral form for continuous states and probability density functions.

In the following section, parameter estimation of a continuous variable using Bayes' theorem is described.

2.3 BAYESIAN PARAMETER ESTIMATION

In Bayesian parameter estimation, Bayes' theorem is used to combine statistical information.

Let $\underline{x} = (x_1, x_2 \ldots x_n)$ be a random sample of observations with probability density function $f(x/\theta)$ which depends on the parameter θ but is otherwise completely known. The parameter θ may be vector valued or a real valued parameter. Let $g(\theta)$ be a prior probability distribution of θ obtained before observing x. The available information embodied in the observed data x

and the prior probability of θ can then be combined using Bayes' theorem. The posterior distribution of θ which is the probability density of θ conditional upon x is given by:

$$\pi(\theta/\underline{x}) = K \cdot g(\theta) \cdot \ell(\underline{x}/\theta)$$

where $\pi(\theta/x)$ is the posterior distribution of θ, $\ell(x/\theta)$ is called the likelihood of x, and K is the normalising constant given by:

$$K^{-1} = \int_{\Omega} \pi(\theta/\underline{x})d\theta = \int_{\Omega} g(\theta) \cdot \ell(\underline{x}/\theta)d\theta$$

where Ω is the parameter space of θ.

Hence,

$$\pi(\theta/\underline{x}) = \frac{g(\theta) \cdot \ell(\underline{x}/\theta)}{\int_{\Omega} g(\theta) \cdot \ell(\underline{x}/\theta)d\theta} \qquad \ldots(2.3)$$

The posterior density $\pi(\theta/\underline{x})$ then embodies all the information one has about θ. Thus in Bayesian inference, all inferences about θ are based on the posterior distribution of θ.

In practical applications, one may wish to characterize the posterior distribution in terms of a small number of measures such as measures of central tendency, dispersion, and skewness, with a measure of central tendency serving as a point estimate of θ.

In this thesis, only the mean and the variance of the posterior distribution of θ will be used as measures of central tendency and dispersion respectively in view of

their common usage in classical statistics and everyday life. In Bayesian terminology using the mean of the posterior distribution of θ as a point estimate, is tantamount to obtaining the Bayesian estimator of θ under a squared-error loss function.

For example, the Bayes estimator of θ, given the data x, is by definition the expectation of the posterior density of θ under a squared-error loss function. This is given by:

$$\theta^* = E(\theta/\underline{x}) = \int_{\Omega} \theta \, . \, \pi(\theta/\underline{x}) d\theta \qquad \ldots (2.4)$$

and the posterior variance of θ is given by:

$$Var(\theta/\underline{x}) = E(\theta^2/\underline{x}) - \theta^{*2} \qquad \ldots (2.5)$$

In (2.4) and (2.5), observational data and prior information are both used and combined in a systematic way to estimate the underlying parameter θ. In the next section, it will be shown how given available sample information, the probability density function of as yet unobserved observations can be obtained by the Bayesian approach.

2.4 PREDICTIVE PROBABILITY DENSITY FUNCTION

When a dyke fails, it makes no difference whether it failed because nature produced a very unusual flood or because the engineer underestimated the design flood discharge. As such, the natural variability or stochastic uncertainty of the flows as well as the uncertainty concerning the parameters of the flood probability distribution must be combined. This can be done by applying compound distribution theory in a Bayesian framework (Wood et al., 1975). This procedure results in what Benjamin and Cornell (1970) called the Bayesian distribution or the predictive density of a future observation (Zellner, 1971) of flood discharges x. This distribution is given by:

$$\tilde{f}(x) = \int_{\Omega} f(\underline{x}/\theta) \cdot \pi(\theta/\underline{x})\, d\theta \qquad \ldots(2.6)$$

where $f(\underline{x}/\theta)$ is the model distribution of the flood discharges, conditional upon the parameters θ; $\pi(\theta/\underline{x})$ is the posterior density function for θ; and $\tilde{f}(x)$ is the Bayesian predictive distribution of the flood discharges, now parameter free.

The predictive distribution $\tilde{f}(x)$ can be interpreted as an average of conditional predictive pdf's, $f(x/\theta)$, with the posterior probability distribution function (pdf) for θ, $\pi(\theta/\underline{x})$ serving as the weighting function. Updating the predictive distribution when new information

become available is achieved by updating the distributions of the uncertain parameters through Bayes' theorem and then updating the predictive distribution using (2.7). It is incorrect to try to update $\tilde{f}(x)$ directly.

Obtaining the probability density function of the predictive distribution for continuous probability distributions is quite complicated. However, one can more easily obtain the mean and variance of the predictive distribution. Take, for example, a model distribution $f(x/\theta)$ with 2 parameters (μ, σ) and, $-\infty < \mu < \infty$ and $\sigma > 0$. Assume that σ is known and fixed, and only μ is uncertain. From (2.6),

$$\tilde{f}(x) = \int_{-\infty}^{\infty} f(x/\mu, \sigma) \cdot \pi(\mu)\, d\mu \qquad \ldots (2.7)$$

The predictive mean of $x = \tilde{m}_x$

$$\tilde{m}_x = \int_{-\infty}^{\infty} x \cdot \tilde{f}(x)\, dx$$

$$= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f(x/\mu, \sigma) \cdot \pi(\mu)\, d\mu \right] dx \qquad \ldots (2.8)$$

By a change of the order of integration,

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x \cdot f(x/\mu, \sigma)\, dx \right] \pi(\mu)\, d\mu \qquad \ldots (2.9)$$

The inner brackets is the expectation of x, m_x.

Therefore,

$$\tilde{m}_x = \int_{-\infty}^{\infty} m_x \cdot \pi(\mu)\, d\mu \qquad \ldots (2.10)$$

Thus, the mean of the predictive distribution is just the average of the means of x for various values of μ.

The predictive variance of x is given by:

$$\tilde{\sigma}_x^2 = \tilde{E}(x^2) - \tilde{m}_x^2 \qquad \ldots (2.11)$$

where:

$$\tilde{E}(x^2) = \int_{-\infty}^{\infty} x^2 \left[\int_{-\infty}^{\infty} f(x/\mu, \sigma) \cdot \pi(\mu)\, d\mu \right] dx \qquad \ldots (2.12)$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x^2 \cdot f(x/\mu, \sigma)\, dx \right] \cdot \pi(\mu)\, d\mu \qquad \ldots (2.13)$$

$$= \int_{-\infty}^{\infty} (\sigma^2 + m_x^2) \cdot \pi(\mu)\, d\mu \qquad \ldots (2.14)$$

Therefore,

$$\tilde{\sigma}_x^2 = \int_{-\infty}^{\infty} \sigma^2 \cdot \pi(\mu)\, d\mu + \int_{-\infty}^{\infty} (\mu - \tilde{m}_x)^2\, \pi(\mu)\, d\mu \qquad ..(2.15)$$

If $\pi(\mu)$ is a symmetrical distribution with mean μ and variance σ_μ^2, then

$$\tilde{m}_x = \mu \qquad \ldots (2.16)$$

and $$\tilde{\sigma}_x = \sigma^2 + \sigma_\mu^2 \qquad \ldots (2.17)$$

Hence, the predictive distribution $\tilde{f}(x)$ has a larger variance than the model distribution $f(x/\theta)$, since the former incorporates both stochastic and parameter uncertainty.

When both μ and σ are uncertain, the predictive mean of x, $\tilde{m}_x$ is given by:

$$\tilde{m}_x = \int_0^{\infty} \int_{-\infty}^{\infty} \mu \cdot \pi(\mu, \sigma)\, d\mu\, d\sigma \qquad \ldots (2.18)$$

and the predictive variance of x, $\tilde{\sigma}_x^2$, is given by:

$$\tilde{\sigma}_x^2 = \int_0^\infty \int_{-\infty}^{\infty} \sigma^2 . \pi(\mu,\sigma)\, d\mu d\sigma + \int_0^\infty \int_{-\infty}^{\infty} (\mu - \tilde{m}_x)^2 . \pi(\mu,\sigma)\, d\mu d\sigma \quad \ldots (2.19)$$

2.5 PRIOR DISTRIBUTIONS AND LIKELIHOOD FUNCTIONS

The mathematical soundness of the Bayesian aproach has been widely accepted; the present controversy is with regard to application, in particular the choice of prior distribution $g(\theta)$ (Kottegoda, 1980).

Zellner (1971) distinguishes between two types of prior probabilities; those obtained from past samples of data are termed data based, and those obtained from personal or theoretical considerations are termed non-data based.

With data based prior distribution, no subjectivity is involved. On the other hand, if one has no prior information, the prior distribution $g(\theta)$ must be decided on a subjective basis or on theoretical grounds in which the theory of invariance of Jeffreys (1961) is widely used.

In a state of in-ignorance about the parameter(s) θ, Jeffreys (1961) suggested the following rules for the choice of the prior distribution $g(\theta)$ which according to Jeffreys, "... cover the commonest cases".

(i) If $\Omega_1 = (-\infty, \infty)$ choose θ_1 to be uniformly distributed, i.e. $g_1(\theta_1) = \text{constant}$.

(ii) If $\Omega_2 = (0, \infty)$, choose log θ_2 to be uniformly distributed, i.e. $g_2(\theta_2) \propto \frac{1}{\theta_2}$

where Ω_i is the range space of θ_i.

Rule (i) is invariant under any linear transformation and Rule (ii) is invariant under any power transformation of θ. In the literature, such priors are described as vague, diffuse, improper or non-informative.

If θ_1 and θ_2 are both unknown, we can assume a-priori that θ_1 and θ_2 are independent since any prior knowledge one may have about θ_1 is not likely to influence one's prior belief about θ_2 (Box and Tiao, 1973).

Thus,

$$g(\theta_1, \theta_2) = g_1(\theta_1) \cdot g_2(\theta_2/\theta_1)$$

$$\approx g_1(\theta_1) \cdot g_2(\theta_2)$$

$$\propto \frac{1}{\theta_2} \qquad \ldots (2.20)$$

Such 'vague' or 'diffuse' prior is quite useful because with a relatively modest sample size, the shape of the posterior distribution will be virtually identical to that of the sample likelihood function. In this case, any prior ideas about the parameters will be overshadowed by information obtained from the data. The role of the dominant likelihood in the analysis of scientific experiments is discussed in detail by Box and Tiao (1973).

The other element necessary in a Bayesian analysis is the likelihood function. This is the function through which the sample data $\underline{x}$ modify prior knowledge of θ. It can be regarded as the function that represents the information about θ contained in the sample data. This likelihood function is the same one used in some classical techniques of estimation and hypothesis testing, namely, maximum likelihood estimators and likelihood ratio tests. The likelihood function is defined as follows. Let $\underline{x} = (x_1, x_2, \ldots x_n)$ be a sample of independent observations, and let the density function with unknown parameter θ of each observation be $f(x_i/\theta)$. Since the trials are independent, the likelihood function is simply the product of the n density functions. That is,

$$\ell(\underline{x}/\theta) = \prod_{i=1}^{n} f(x_i/\theta) \qquad \ldots (2.21)$$

In the following section, the relevance of the Bayesian approach to flood risk analysis is discussed. Some distinctive differences between the classical and Bayesian approach to statistical inference are also outlined.

2.6 CLASSICAL VERSUS BAYESIAN

Martz and Waller (1982) have outlined in detail the distinctive differences between the classical and Bayesian approach to statistical inference. In this section, only a brief comparison is given with a discussion on the relevance of the Bayesian approach to flood risk analysis.

In the classical approach, inferences are based on the likelihood function in which the unknown parameter θ is assumed to be a fixed constant. In the Bayesian approach, however, θ is treated as a random variable having a probability distribution which represents a formalization of information about θ before observing a sample. In flood risk analysis such prior information can come about from regional hydrologic and geomorphic information (Wood et al., 1974; Vicens et al., 1975; Wood and Rodriquez-Iturbe, 1975) as well as sample data. It would be imprudent to neglect such additional information if it is available.

Another distinctive difference between the two methods of inference is the method of reasoning.

The classical method of inference is depicted in Figure 2.1. The process starts out by postulating a tentative sampling model. Inductive reasoning is then used in conjunction with the sample observations to produce inferences about the unknown parameters in the assumed model.

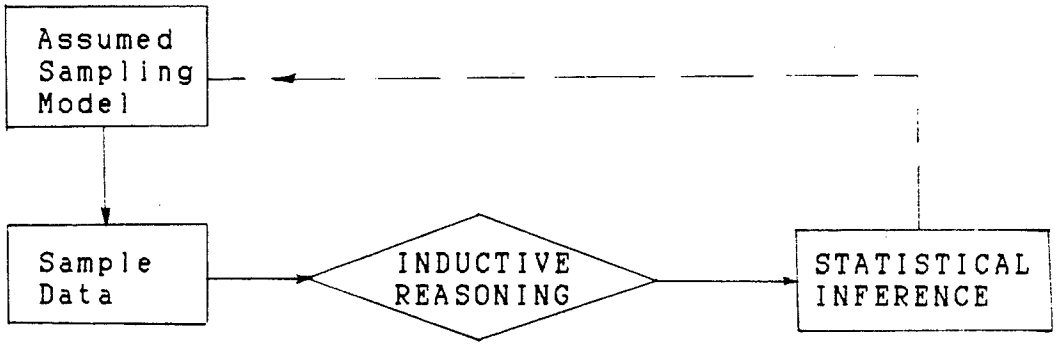


Figure 2.1 Inferences Based On Classical Theory (after Martz and Waller, 1982)

Figure 2.2 depicts the Bayesian method of inference. The process also starts with an assumed sampling model. A prior probability distribution is also postulated for those unknown parameters in the assumed sampling model. Bayes' theorem is then used to combine the sample data and the prior distribution. Deductive reasoning is then used in conjunction with the resulting posterior distribution to produce the desired inferences about the parameters of the assumed sampling model.

A further distinctive difference between the classical and Bayesian approach is that the Bayesian approach usually requires less sample data to arrive at the same quality of inferences than the classical approach. This is due again to the use of prior information. This is especially important in flood risk analysis due to the lack of data on floods in most parts of

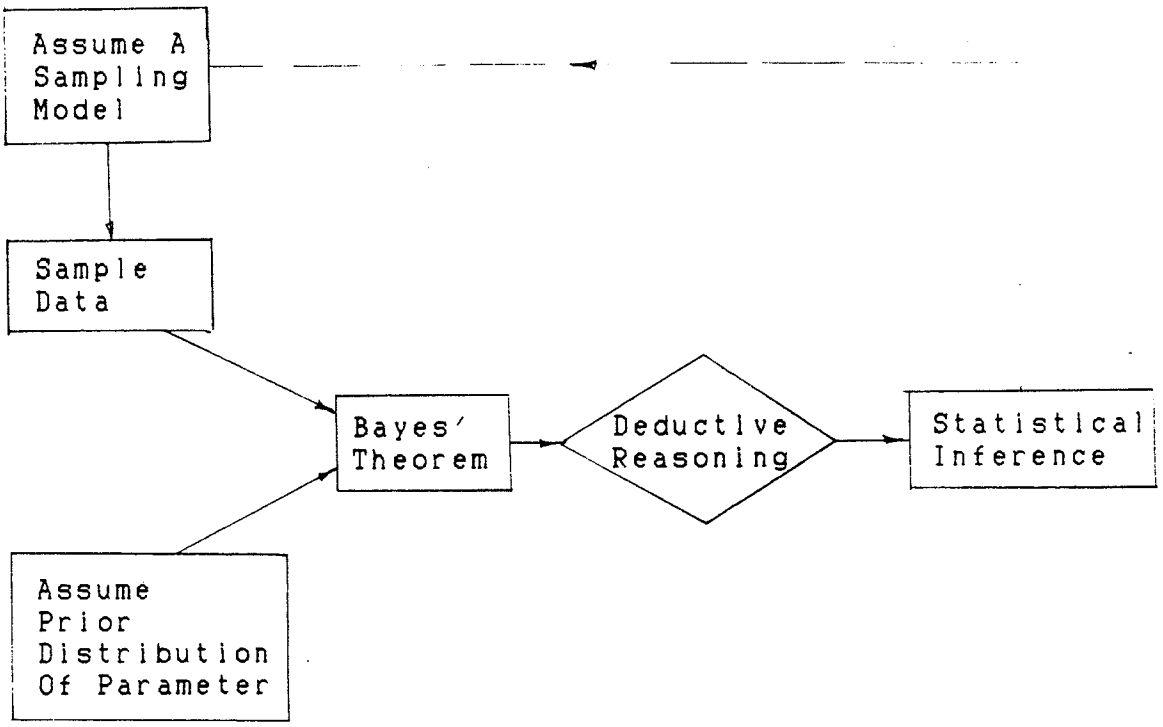


Figure 2.2 Bayesian Inference
(after Martz and Waller, 1982)

the world.

Another important aspect of the Bayesian approach is that Bayes' theorem provides a mathematical framework for processing new sample data as such data becomes sequentially available over time. The theorem provides a mechanism for continually updating available knowledge about θ as more sample data become available. How this is done is explained in Martz and Waller (1982), and Box and Tiao (1973). This basic fact concerning Bayes theorem is the basis of real-time processing of sample data by use of well-known mathematical devices such as the Kalman filter

(Kalman and Bucy, 1961), which have been used in real-time flood forecasting, satellite tracking, etc.

Other advantages of the Bayesian approach to flood risk analysis and other engineering designs are discussed in Davis et al., 1972 and Kottegoda, 1980. A summary of some characteristics of the classical theory and Bayesian methods of statistical inference is shown in Table 2.1.

2.7 SUMMARY

An introduction to the Bayesian approach and some comparison with classical theory was presented. The Bayesian approach has a number of attributes that are especially important to flood risk analysis. These are the explicit use of prior information, ability to continuously update our knowledge about the parameters, and the ability of the Bayesian approach to combine parameter uncertainty and stochastic uncertainty.

In the following chapter, the estimation of parameters and T-year flood by the Bayesian approach is considered.

TABLE 2.1

A SUMMARY OF CERTAIN CHARACTERISTICS OF CLASSICAL THEORY AND BAYESIAN METHODS OF STATISTICAL INFERENCE (after Martz & Waller, 1982)

Characteristic	Classical Theory	Bayesian
Parameter(s) of Interest	Unknown constant(s)	Random variable(s)
Prior Distribution	Does not exist	Exists and explicitly assumed
Sampling Model	Assumed	Assumed
Posterior Distribution	Does not exist	Explicitly derived
Method of Reasoning	Inductive	Deductive
Type of Interval Estimate	Confidence interval	Credible interval
Role of Past Experience	Not applicable	Applicable
Purpose of Sampling Experiment	Supply the data for making inferences	Confirm or deny expected performance as predicted from past experience
Quality of Inferences	More restrictive than Bayes' because of exclusive use of sample data	Depends on ability to quantitatively relate past experience to the sample data
Quantity of Sample Data	Bayes' approach usually requires less because it utilizes relevant past data	

CHAPTER 3

BAYESIAN ESTIMATES OF PARAMETERS AND T-YEAR FLOOD

3.1 GENERAL

In flood frequency analysis, the method preferred by hydrologists to estimate the 'true' parameters and T-year events of flood probability distribution is the method of likelihood. This is because the maximum likelihood estimates possess the properties of consistency and asymptotic efficiency (Kendall and Stuart, 1973). In addition, in most cases, it gives a smaller standard error of estimate of the T-year flood when compared to other methods. In this chapter, the Bayes estimates of the parameters and T-year flood for two commonly used flood probability distributions are obtained and the posterior variances of these estimates compared to the corresponding Maximum Likelihood Estimate's (MLE's). Lindley's Bayesian Approximation procedure (Lindley, 1980) is used to obtain the Bayes estimates, thus avoiding the need to evaluate unwieldy ratios of multiple integrals necessary in Bayesian analysis. A 'vague' prior distribution described in Section 2.5 is used to obtain all Bayes estimates. The probability distributions considered are the Gumbel and the 3-parameter lognormal distributions. The 2-parameter

lognormal is not considered as it is a special case of the 3-parameter lognormal distribution.

3.2 MAXIMUM LIKELIHOOD ESTIMATES

For a given probability density function $f(x/\underline{\theta})$, where $\underline{\theta} = (\theta_1, \theta_2, \ldots \theta_m)$ are parameters, the joint probability that a sample of n values $(x_1, x_2, \ldots x_n)$ came from that distribution is:

$$\ell(\underline{x}/\underline{\theta}) = f(x_1/\underline{\theta}) \cdot f(x_2/\underline{\theta}) \ldots f(x_n/\underline{\theta}) \quad ..(3.1)$$

where $\ell(\underline{x}/\underline{\theta})$ is the likelihood. The principle of maximum likelihood states that the values of θ_1, θ_2, etc. should be chosen to maximize $\ell(\underline{x}/\underline{\theta})$. This is obtained by partially differentiating $\ell(\underline{x}/\underline{\theta})$ with respect to each of the parameters and equating to zero. Usually for continuous distributions, it is easier to maximize the natural logarithm of the likelihood function L. The system of equations which when solved to achieve this maximization are the maximum likelihood estimators for the parameters of the distribution. This system is:

$$\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial \theta_2} = \frac{\partial L}{\partial \theta_3} \quad \text{etc} = 0 \qquad \ldots (3.2)$$

3.3 STANDARD ERROR OF ESTIMATES

A measure of the variability of an estimated value is the standard error of estimate. This is defined as:

$$S = \left[\frac{\sum_{i=1}^{n}(q_i - \hat{q}_i)^2}{n}\right]^{1/2} \quad \ldots (3.3)$$

where $\hat{q}_i$ is the computed estimate of recorded value q_i (Kite, 1977). The standard error measures the errors in the estimated parameters of the chosen population distribution that may be inaccurate due to the lack of data and/or sampling fluctuation.

The standard error of estimate by maximum likelihood is obtained as follows: Assume a particular distribution with parameters θ_1, θ_2 and θ_3 which have been estimated by the method of maximum likelihood. If Z is a function of $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$, each of which is subject to sampling error then the variance Z can be obtained from first order Taylor series expansion (Benjamin and Cornell, 1970). This is given by:

$$\begin{aligned} Var(Z) = &\left(\frac{\partial Z}{\partial \theta_1}\right)^2 Var(\hat{\theta}_1) + \left(\frac{\partial Z}{\partial \theta_2}\right)^2 Var(\hat{\theta}_2) + \left(\frac{\partial Z}{\partial \theta_3}\right)^2 Var(\hat{\theta}_3) \\ &+ 2\left(\frac{\partial Z}{\partial \theta_1}\right)\left(\frac{\partial Z}{\partial \theta_2}\right) Cov(\hat{\theta}_1, \hat{\theta}_2) + 2\left(\frac{\partial Z}{\partial \theta_1}\right)\left(\frac{\partial Z}{\partial \theta_3}\right) Cov(\hat{\theta}_1, \hat{\theta}_3) \\ &+ 2\left(\frac{\partial Z}{\partial \theta_2}\right)\left(\frac{\partial Z}{\partial \theta_3}\right) Cov(\hat{\theta}_2, \hat{\theta}_3) \quad \ldots (3.4) \end{aligned}$$

Equation (3.4) is the general variance of estimate equation and is applicable to a function of any number of variables. The partial derivatives may be obtained directly from the function of Z evaluated at the maximum likelihood estimates $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$. The required variances and covariances are obtained from the inverse of the symmetric matrix (Kendall and Stuart, 1973) given by:

$$[I] = \begin{bmatrix} -\frac{\partial^2 L}{\partial \theta_1^2} & -\frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} & -\frac{\partial^2 L}{\partial \theta_1 \partial \theta_3} \\ -\frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & -\frac{\partial^2 L}{\partial \theta_2^2} & -\frac{\partial^2 L}{\partial \theta_2 \partial \theta_3} \\ -\frac{\partial^2 L}{\partial \theta_3 \partial \theta_1} & -\frac{\partial^2 L}{\partial \theta_3 \partial \theta_2} & -\frac{\partial^2 L}{\partial \theta_3^2} \end{bmatrix} \quad \ldots (3.5)$$

where L is the log-likelihood function of the chosen probability distribution. That is:

$$\begin{bmatrix} \mathrm{Var}(\hat{\theta}_1) & \mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_3) \\ \mathrm{Cov}(\hat{\theta}_2, \hat{\theta}_1) & \mathrm{Var}(\hat{\theta}_2) & \mathrm{Cov}(\hat{\theta}_2, \hat{\theta}_3) \\ \mathrm{Cov}(\hat{\theta}_3, \hat{\theta}_1) & \mathrm{Cov}(\hat{\theta}_3, \hat{\theta}_2) & \mathrm{Var}(\hat{\theta}_3) \end{bmatrix} = [I]^{-1} \quad .(3.6)$$

The standard error of estimate of Z is then defined as the square root of (3.4).

In the following section, Lindley's Bayesian Approximation is discussed.

3.4 LINDLEY'S BAYESIAN APPROXIMATION PROCEDURE

One of the reasons why the Bayesian approach is not widely used in flood analysis is the mathematical complexity. Bayes estimators are often obtained as a ratio of multiple integrals which cannot be expressed in closed forms and numerical approximations are necessary. Here, Bayes estimates approximated by an asymptotic expansion of the ratio of two integrals due to Lindley (1980) are discussed.

The posterior expectation of an arbitrary function $u(\theta)$ with prior probability distribution $v(\theta)$ and log-likelihood function $L(\theta)$:

$$E\,[u(\theta)/\underline{x}] = \frac{\int_{\Omega} u(\theta) \cdot v(\theta) \cdot \exp[L(\theta)]\, d\theta}{\int_{\Omega} v(\theta) \cdot \exp[L(\theta)\, d\theta} \quad .. (3.7)$$

which is the Bayes estimator of $u(\theta)$ under the squared error loss function may be asymptotically estimated by:

$$E\,[u(\theta/\underline{x}] = \Big[\, u + \frac{1}{2}\sum_i\sum_j (u_{ij} + 2u_i\rho_j)\sigma_{ij} + \frac{1}{2}\sum_i\sum_j\sum_k\sum_\ell L_{ijk}\ \sigma_{ij}\ \sigma_{k\ell}\ u_\ell \,\Big]_{\hat\theta} + \text{terms of the order } 1/n^2 \text{ and smaller} \quad \ldots(3.8)$$

All functions being evaluated at the MLE of θ, and where $i, j, k, l = 1, 2, \ldots m$; $\theta = (\theta_1, \theta_2, \ldots \theta_m)$, $\hat\theta$ = MLE(θ), Ω is the range space of θ, $v(\theta)$ is the prior distribution of θ, $u \equiv u(\theta)$, $L \equiv L(\theta)$ is logarithmic likelihood function, $u_{ij} = \partial^2 u/\partial\theta_i\ \partial\theta_j$, $L_{ijk} = \partial^3 L/\partial\theta_i\partial\theta_j\partial\theta_k$, $\rho \equiv \rho(\theta) = \log v(\theta)$, $\rho_j = \partial\rho/\partial\theta_j$ and σ_{ij} = (i, j)th element of the variance-covariance matrix, (Lindley, 1980). See Appendix B for the derivation of Lindley's expansion.

Sinha (1985, 1986a, b, c), and Sinha and Sloan (1985) have used the linear Bayes estimator (3.8) for the Bayesian estimation of the reliability function of various distributions. Gren (1980) also states that (3.8) is a "very good and operational approximation for the ratio of the multi-dimensional integrals". Sinha (1987) has also shown that up to order of $1/n^2$, the linear Bayes estimator is more efficient than the MLE. Although the method requires that $\hat\theta$ be the unique MLE of θ, in most instances the local MLE produce acceptable estimates (Sinha and Sloan, 1988).

In the next section, Bayes estimates of the parameters and T-year flood for flood data fitted by the Gumbel distribution will be obtained and their posterior variances compared to the corresponding MLE's. This will be followed by the 3-parameter lognormal distribution.

3.5 GUMBEL DISTRIBUTION

The Gumbel distribution is widely used for frequency analysis of extremes in meteorology and hydrology. Lettenmaier and Burges (1982), Phien and Arbhabhirama (1980), and Jain and Singh (1987) gave several reasons for its popularity.

The Gumbel distribution, despite its extensive use, generally has no accepted method of estimating its parameters. In an extensive study, Jain and Singh (1987), compared seven methods of estimating the parameters. They found the method of maximum likelihood to be the best method based on various criteria. Studies by Phien et al. (1980) and Lettenmaier et al. (1982) also concluded that the method of maximum likelihood gave better estimates than the method of moments and other methods. These conclusions are based on the criteria of goodness of fit and estimation variability.

In this section, Bayes estimates of the parameters and of the T-year events will be obtained and the

posterior variances of these estimates compared to the corresponding MLE's.

3.5.1 Maximum Likelihood Estimates

The probability density function (pdf) of the Gumbel distribution is given by:

$$f(x/\mu, \sigma) = \frac{1}{\sigma} \exp \left\{ - \frac{x - \mu}{\sigma} - \exp \left[- \frac{x - \mu}{\sigma} \right] \right\} \quad ..(3.9)$$

$$\sigma > 0, \quad -\infty < x, \quad \mu < \infty$$

Given a random sample $\underline{x} = (x_1, x_2, \ldots x_n)$ from the pdf (3.9), the logarithmic likelihood is given by:

$$L = - n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} (x_i - \mu) - \sum_{i=1}^{n} \left[\exp - \left(\frac{x_i - \mu}{\sigma} \right) \right] \quad \ldots (3.10)$$

Taking partial derivatives with respect to σ and μ and equating to zero, one gets:

$$\frac{\partial L}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) - \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu) \cdot \exp - \left(\frac{x_i - \mu}{\sigma} \right) = 0 \quad \ldots (3.11)$$

$$\frac{\partial L}{\partial \mu} = \frac{n}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^{n} \left[\exp - \left(\frac{x_i - \mu}{\sigma} \right) \right] = 0 \quad \ldots (3.12)$$

From (3.12), one obtains:

$$n = \sum_{i=1}^{n} \exp\left[-\frac{1}{\sigma}(x_i - \mu) \right] \qquad \ldots (3.13)$$

$$\text{or} \quad \exp\left(\frac{\mu}{\sigma}\right) = \frac{n}{\sum_{i=1}^{n} \exp\left(\frac{-x_i}{\sigma}\right)} \qquad \ldots (3.14)$$

From which one obtains:

$$\hat{\mu} = \hat{\sigma} \log\left[n \Big/ \sum_{i=1}^{n} \exp\left(-\frac{x_i}{\hat{\sigma}}\right) \right] \qquad \ldots (3.15)$$

From (3.11), one has:

$$-n + \frac{n(\bar{x} - \mu)}{\sigma} - \frac{\exp(\mu/\sigma)}{\sigma} \sum_{i=1}^{n} (x_i - \mu) \exp\left(-\frac{x_i}{\sigma}\right) = 0 \qquad \ldots (3.16)$$

Substituting for $\exp(\mu/\sigma)$ from (3.14), one obtains a function of $\hat{\sigma}$ only.

$$F(\hat{\sigma}) = \sum_{i=1}^{n} x_i \exp\left(-\frac{x_i}{\hat{\sigma}}\right) - (\bar{x} - \hat{\sigma}) \sum_{i=1}^{n} \exp\left(-\frac{x_i}{\hat{\sigma}}\right) = 0 \qquad \ldots (3.17)$$

where $\bar{x}$ is the arithmetic average of x. Using the moment estimate of σ (Gumbel, 1958) as the starting value, (3.17) may be solved for $\hat{\sigma}$ iteratively using Newton-Raphson routine. Having obtained $\hat{\sigma}$, $\hat{\mu}$ may be estimated from (3.15).

3.5.2 T-Year Return Period Event

Using the parameters estimated by maximum likelihood, the T-year return period event or T-year event of the Gumbel distribution is given by:

$$x_T = \hat{\mu} + \hat{\sigma} Y_T \qquad \ldots (3.18)$$

where, $Y_T = - \log [-\log(1 - \frac{1}{T})]$ (Kimball, 1949); and T corresponds to a given return period in years of the extreme event.

S_T, the standard error of x_T, may be obtained from (3.4). This is given by:

$$S_T^2 = \left(\frac{\partial x_T}{\partial \sigma}\right)^2 \mathrm{Var}(\hat{\sigma}) + \left(\frac{\partial x_T}{\partial \mu}\right)^2 \mathrm{Var}(\hat{\mu}) + 2 \left(\frac{\partial x_T}{\partial \sigma}\right)\left(\frac{\partial x_T}{\partial \mu}\right) \mathrm{Cov}(\hat{\sigma}, \hat{\mu}) \qquad \ldots (3.19)$$

From (3.18), one obtains:

$$\frac{\partial x_T}{\partial \sigma} = Y_T \qquad \ldots (3.20)$$

$$\frac{\partial x_T}{\partial \mu} = 1 \qquad \ldots (3.21)$$

The variance-covariance matrix,

$$\sigma_{ij} = \begin{bmatrix} \mathrm{Var}(\hat{\sigma}) & \mathrm{Cov}(\hat{\sigma}, \hat{\mu}) \\ \mathrm{Cov}(\hat{\mu}, \hat{\sigma}) & \mathrm{Var}(\hat{\mu}) \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \qquad \ldots (3.22)$$

is estimated from:

$$[I]^{-1} = \begin{bmatrix} -\frac{\partial^2 L}{\partial \sigma^2} & -\frac{\partial^2 L}{\partial \sigma \, \partial \mu} \\ -\frac{\partial^2 L}{\partial \mu \, \partial \sigma} & \frac{\partial^2 L}{\partial \mu^2} \end{bmatrix}^{-1}_{\hat{\mu},\ \hat{\sigma}} \qquad \ldots (3.23)$$

where,

$$\frac{\partial^2 L}{\partial \sigma^2} = \frac{1}{\sigma^4} \left\{ n\sigma^2 - 2\sigma \sum_{i=1}^{n} (x_i - \mu) + 2\sigma \sum_{i=1}^{n} \left[(x_i - \mu) \exp - \left(\frac{x_i - \mu}{\sigma} \right) \right] \right\} \ldots (3.24)$$

$$\frac{\partial^2 L}{\partial \mu \, \partial \sigma} = -\frac{1}{\sigma^3} \sum_{i=1}^{n} \left[(x_i - \mu) \exp - \left(\frac{x_i - \mu}{\sigma} \right) \right] \qquad \ldots (3.25)$$

$$\frac{\partial^2 L}{\partial \mu^2} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \left[\exp - \left(\frac{x_i - \mu}{\sigma} \right) \right] = -\frac{n}{\sigma^2} \qquad \ldots (3.26)$$

3.5.3 Bayes Estimates

Using the principles suggested by Jeffreys (1961) and described in Section 2.5,

$g_1(\mu)$ = constant

$g_2(\sigma) \propto 1/\sigma$

Assuming μ and σ to be independent 'a priori', the joint prior distribution of μ and σ is given by:

$$\begin{aligned} v(\mu, \sigma) &= g_1(\mu) \cdot g_2(\sigma/\mu) \\ &\approx g_1(\mu) \cdot g_2(\sigma) \\ &\propto \frac{1}{\sigma} \end{aligned} \qquad \ldots (3.27)$$

Combining the prior with the likelihood function and using Bayes' theorem, the joint posterior distribution of (μ, σ) is obtained:

$$\Pi(\mu, \sigma/\underline{x}) = K \cdot \ell(\underline{x}/\mu, \sigma) \cdot v(\mu, \sigma) \qquad \ldots (3.28)$$

where K is the normalising constant.

Under the squared-error loss function, Bayes estimator of a function is its posterior expectation. Hence, the Bayes estimators of σ, μ and T-year event x_T are given by:

$$\begin{aligned} \sigma^* &= E(\sigma/\underline{x}) \\ &= \int_{-\infty}^{\infty} \int_{0}^{\infty} \sigma \cdot \Pi(\mu, \sigma/\underline{x}) \, d\sigma \, d\mu \\ &= \frac{\int_{-\infty}^{\infty} \int_{0}^{\infty} \sigma \cdot v(\mu, \sigma) \cdot \ell(\underline{x}/\mu, \sigma) \, d\sigma \, d\mu}{\int_{-\infty}^{\infty} \int_{0}^{\infty} v(\mu, \sigma) \cdot \ell(\underline{x}/\mu, \sigma) \, d\sigma \, d\mu} \end{aligned} \qquad \ldots (3.29)$$

$$x_T^* = \frac{\int_{-\infty}^{\infty} \int_{0}^{\infty} x_T \cdot v(\mu, \sigma) \cdot \ell(\underline{x}/\mu, \sigma) \, d\sigma \, d\mu}{\int_{-\infty}^{\infty} \int_{0}^{\infty} v(\mu, \sigma) \cdot \ell(\underline{x}/\mu, \sigma) \, d\sigma \, d\mu} \qquad \ldots (3.30)$$

and similarly for μ^*.

3.5.4 Bayesian Approximation

Lindley's asymptotic expansion of (3.7) for the 2-parameter case is given by:

$$
\begin{aligned}
E[u(\theta)] = u &+ \frac{1}{2}(u_{11}\sigma_{11} + u_{22}\sigma_{22}) + u_{12}\sigma_{12} \\
&+ u_1(\sigma_{11}\rho_1 + \sigma_{21}\rho_2) + u_2(\sigma_{12}\rho_1 + \sigma_{22}\rho_2) \\
&+ \frac{1}{2}\Big[L_{30}(u_1\sigma_{11}^2 + u_2\sigma_{11}\sigma_{12}) \\
&+ L_{21}\{3u_1\sigma_{11}\sigma_{12} + u_2(\sigma_{11}\sigma_{22} + 2\sigma_{12}^2)\} \\
&+ L_{12}\{u_1(\sigma_{11}\sigma_{22} + 2\sigma_{12}^2) + 3u_2\sigma_{12}\sigma_{22}\} \\
&+ L_{03}(u_1\sigma_{12}\sigma_{22} + u_2\sigma_{22}^2)\Big]_{\hat{\theta}} \qquad \ldots(3.31)
\end{aligned}
$$

where, $u \equiv u(\theta)$, $u_1 = \dfrac{\partial u}{\partial \sigma}$, $u_2 = \dfrac{\partial u}{\partial \mu}$, $u_{12} = \dfrac{\partial^2 u}{\partial \sigma\, \partial \mu}$,

$u_{11} = \dfrac{\partial^2 u}{\partial \sigma^2}$, $u_{22} = \dfrac{\partial^2 u}{\partial \mu^2}$; $\rho \equiv \log v(\theta)$, $\rho_1 = \dfrac{\partial \rho}{\partial \sigma}$,

$\rho_2 = \dfrac{\partial \rho}{\partial \mu}$; $L_{30} = \dfrac{\partial^3 L}{\partial \sigma^3}$, $L_{03} = \dfrac{\partial^3 L}{\partial \mu^3}$, $L_{12} = \dfrac{\partial^3 L}{\partial \sigma\, \partial \mu^2}$,

$L_{21} = \dfrac{\partial^3 L}{\partial \sigma^2 \partial \mu}$,

σ_{ij} = (i, j)th element of the variance-covariance matrix given by (3.22), and all constants are to be evaluated at the MLE of (σ, μ). See Appendix C for evaluation of L_{ij}'s.

3.5.5 Numerical Example

The annual maximum flows of the Turtle River at Mine Center, Ontario, is used as an example. Figure 3.1 shows the fit of the Gumbel distribution to the observed data by maximum likelihood. The maximum likelihood estimators for this river are:

$$n = 58, \quad \hat{\sigma} = 45.810, \quad \hat{\mu} = 101.270$$

$$[\sigma_{ij}] = \begin{bmatrix} 22.5061 & 9.5444 \\ 9.5444 & 40.2302 \end{bmatrix}$$

$$\rho = \log(1/\sigma), \ \rho_1 = -0.02183, \ \rho_2 = 0$$

$$L_{30} = 0.0054401, \quad L_{03} = -0.0006033$$

$$L_{12} = 0.0014625, \quad L_{21} = -0.0014985$$

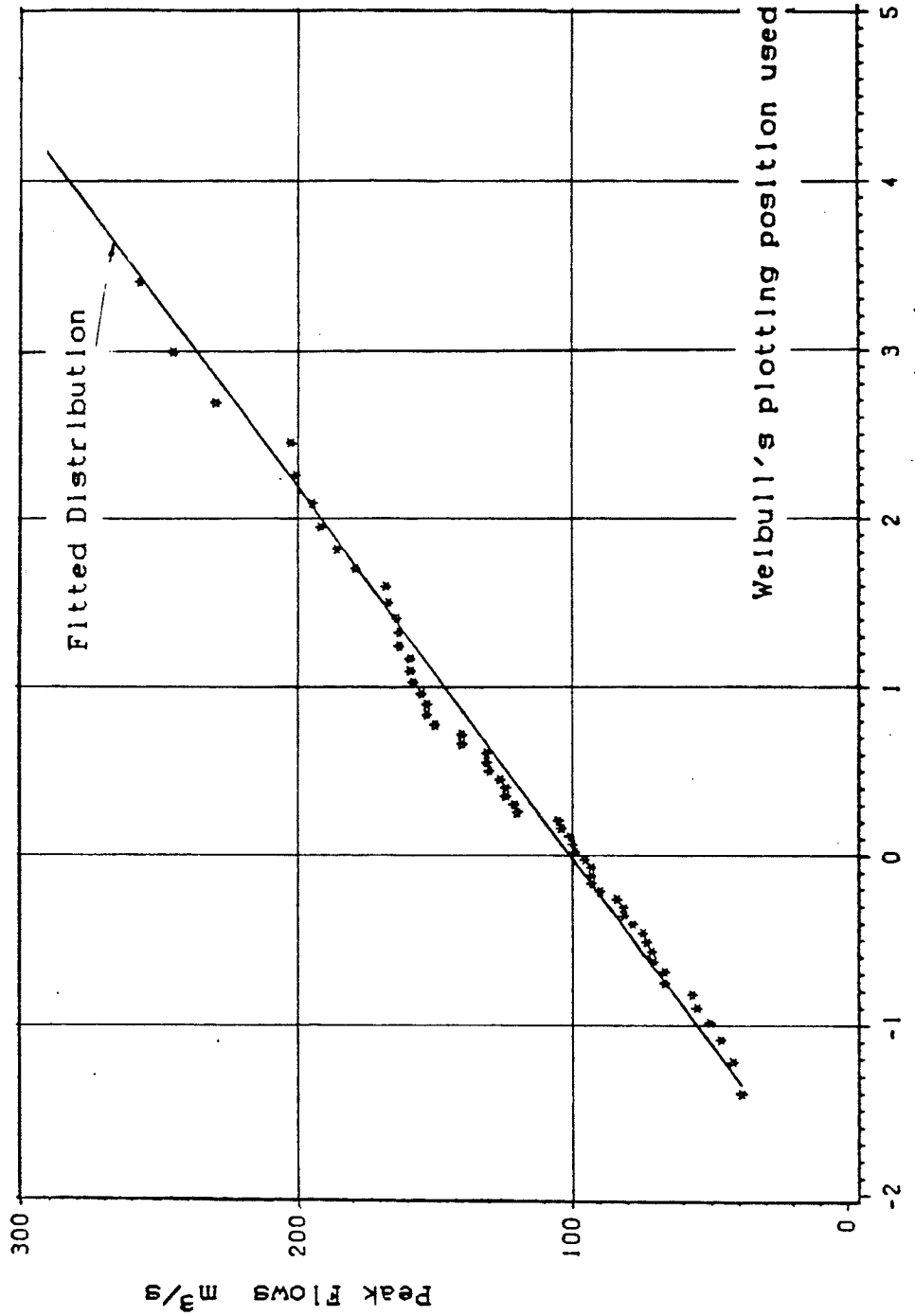


Figure 3.1 Gumbel Distribution - Turtle River at Mine Center

Let $u = \sigma$, $u_1 = 1$, $u_2 = 0$, $u_{ij} = 0$ for all of i, js.
Substituting into (3.31):

$$E[\sigma/\underline{x}] = \sigma^* = \hat{\sigma} + \sigma_{11}\rho_1 + \frac{1}{2}\left[L_{30}\sigma_{11}^2 + L_{21}(3\sigma_{11}\sigma_{12}) + L_{12}(\sigma_{11}\sigma_{12} + 2\sigma_{12})^2 + L_{03}\sigma_{12}\sigma_{22} \right]$$

$$= 46.894$$

Also let $u = \sigma^2$, $u_1 = 2\sigma$, $u_{11} = 2$, $u_2 = u_{12} = u_{22} = 0$.
Substituting into (3.31):

$$E[\sigma^2/\underline{x}] = \hat{\sigma}^2 + \sigma_{11} + 2\hat{\sigma}\sigma_{11}\rho_1 + L_{30}\hat{\sigma}\sigma_{11}^2 + 3L_{21}\hat{\sigma}\sigma_{11}\sigma_{12} + L_{12}\hat{\sigma}(\sigma_{11}\sigma_{12} + 2\sigma_{12}^2) + L_{03}\hat{\sigma}\sigma_{12}\sigma_{22}$$

$$= 2222.380$$

The posterior variance of σ is then given by:

$$Var(\sigma/\underline{x}) = \sigma_{11}^* = E(\sigma^2/\underline{x}) - [E(\sigma/\underline{x})]^2$$

$$= 21.3329$$

which is less than $\sigma_{11} = 22.5061$.
Similarly,

$$E(\mu/\underline{x}) = \mu^* = 101.182$$

and $Var(\mu/\underline{x}) = \sigma_{22}^* = 40.2230 < \sigma_{22} = 40.2302$

Table 3.1 summarizes the results.

TABLE 3.1

Bayes (*) and MLE (^) Estimates of (σ, μ)
For Turtle River At Mine Center (n=58)
(Gumbel Distribution)

Parameter	^	*	Var(^)	Posterior Variance
σ	45.810	46.894	22.5061	21.3329
μ	101.270	101.182	40.2302	40.2230

Bayes estimates of the T-year event is obtained as follows:

Let $u = x_T = \mu + \sigma Y_T$

$u_1 = Y_T$, $u_2 = 1$, $u_{11} = u_{12} = 0$;

where $Y_T = -\log [-\log (1 - \frac{1}{T})]$

For T = 100 (100 year return period flood), $Y_T = 4.60015$. Substituting into (3.31), one gets:

$$E(x_{T=100} / \underline{x}) = x_T^* = 316.90$$

Also let $u = x_T^2$,

$$u_1 = 2Y_T(\mu + \sigma Y_T), \quad u_2 = 2(\mu + \sigma Y_T);$$

$$u_{11} = 2Y_T^2; \quad u_{22} = 2; \quad u_{12} = 2Y_T$$

Substituting into (3.31):

$$E\,(x^2_{T=100} / \underline{x}) = 101{,}005.9381$$

Hence, the posterior variance of $x_{T=100}$ is given by:

$$Var(x_{T=100}/\underline{x}) = E\,(x^2_{T=100}/\underline{x}) - \left\{E\,(x_{T=100}/\underline{x})\right\}^2$$

$$= 580.3281$$

If one defines the posterior standard error of the T-year event as:

$$S_T(x_T / \underline{x}) = \sqrt{Var(x_T / \underline{x})}$$

The following is obtained:

$$S_T\,(x_{T=100} / \underline{x}) = 24.090 < \hat{S}_T = 24.583$$

Table 3.2 summarizes the results for return periods T = 2, 5, 10, 20, 50, 100 and 1000 years for the Turtle River at Mine Center (n = 58).

TABLE 3.2

Standard Error Of Estimate Of T-Year Events
Fitted By Gumbel Distribution
- Turtle River At Mine Center
(n = 58)

T-year	2	5	10	20	50	100	1000
$\hat{x}_T$	118.06	169.98	204.36	237.33	280.02	312.00	417.69
$\hat{S}_T$	7.089	10.931	14.041	17.189	21.386	24.583	35.297
x_T^*	118.37	171.52	206.71	240.46	284.18	316.90	425.09
S_T^*	7.082	10.822	73.843	16.902	20.981	24.090	34.513

(^) MLE, (*) Bayes estimates

The tables above shows that the posterior variances of (σ, μ) and posterior standard errors of the T-year events are less than the corresponding MLE's even with a 'vague' prior.

Lindley's procedure was repeated for seven other maximum annual flood events of rivers from all over Canada. The results are shown in Table 3.3. The results show that for all seven rivers, Bayes estimates of the parameters (σ, μ) and the T-year flood have smaller posterior standard error than that obtained by the method of maximum likelihood.

Table 3.3 Comparison of MLE and Bayes Estimates (Gumbel Distribution)

RIVER		PARAMETER		VARIANCE		STANDARD ERRORS OF X_T IN PARENTHESIS						
		σ	μ	σ	μ	T = 2	5	10	20	50	100	1000
RED RIVER AT REDWOOD BRIDGE MANITOBA, N=93	^	497.400	838.340	1696.11	2951.29	1,010.60 (60.779)	1,574.40 (94.206)	1,947.70 (121.25)	2,305.70 (148.62)	2,769.20 (185.09)	3,116.50 (212.85)	4,264.00 (305.90)
	H	504.833	827.775	1640.89	2950.97	1,012.80 (60.741)	1,585.00 (93.610)	1,963.80 (120.17)	2,327.20 (147.06)	2,797.60 (182.89)	3,150.10 (210.18)	4,314.80 (301.66)
TURTLE RIVER AT MINE CENTER ONTARIO, N=58	^	45.810	101.270	22.5061	40.2302	118.06 (7.088)	169.98 (10.931)	204.36 (14.041)	237.33 (17.189)	280.00 (21.386)	312.00 (24.583)	417.69 (35.297)
	H	46.894	101.182	21.3329	40.2230	118.37 (7.082)	171.52 (10.822)	206.71 (13.843)	240.46 (16.902)	284.16 (20.981)	316.90 (24.090)	425.09 (34.513)
SAUGEEN RIVER NEAR WALKERTON ONTARIO, N=70	^	98.417	241.390	83.4775	153.7240	277.46 (13.827)	389.01 (21.188)	462.86 (27.159)	533.71 (33.209)	625.41 (41.280)	694.12 (47.430)	921.18 (68.051)
	H	100.340	241.230	79.7817	153.6990	278.01 (13.816)	391.73 (21.012)	467.03 (26.837)	539.26 (32.741)	632.75 (40.622)	702.81 (46.628)	934.30 (66.775)
NORTH MAGNETAWAN RIVER, BURKS FALL ONTARIO, N=69	^	11.462	37.342	1.1430	2.1189	41.54 (1.624)	54.53 (2.485)	63.14 (3.184)	71.39 (3.892)	82.07 (4.836)	90.07 (5.556)	116.51 (7.968)
	H	11.687	37.323	1.0924	2.1185	41.61 (1.622)	54.85 (2.465)	63.62 (3.146)	72.04 (3.837)	82.92 (4.759)	91.08 (5.462)	118.05 (7.819)
PIGEON RIVER AT MIDDLE FALLS ONTARIO, N=41	^	45.297	111.050	20.5248	37.2982	127.65 (6.812)	178.99 (10.463)	212.99 (13.426)	245.59 (16.428)	287.80 (20.431)	319.42 (23.482)	423.93 (33.709)
	H	46.326	110.968	19.4649	37.2916	127.95 (6.806)	180.45 (10.361)	215.22 (13.239)	248.56 (16.156)	291.73 (20.050)	324.07 (23.017)	430.95 (32.969)
SKEENA RIVER AT USK B. COLUMBIA, N=37	^	1033.000	4571.700	17655.300	31984.700	4,950.10 (199.550)	6,120.96 (306.690)	6,896.23 (393.620)	7,639.72 (481.670)	8,602.19 (599.120)	9,325.42 (688.590)	11,706.81 (988.570)
	H	1071.390	4568.510	16180.400	31974.700	4,961.20 (199.250)	6,175.50 (301.820)	6,979.50 (384.710)	7,750.70 (468.730)	8,749.00 (580.880)	9,497.10 (666.370)	11,969.00 (953.190)
BULKLEY RIVER AT QUICK B. COLUMBIA, N=54	^	131.640	511.830	184.8720	358.8740	560.07 (21.098)	709.30 (32.033)	808.05 (40.892)	902.81 (49.875)	1,025.48 (61.866)	1,117.38 (71.006)	1,421.10 (101.670)
	H	134.858	511.551	174.5210	358.7970	560.98 (21.079)	713.83 (31.709)	815.03 (40.295)	912.10 (49.004)	1,037.80 (60.635)	1,131.90 (69.504)	1,443.00 (99.272)
HARRICANA RIVIERE A AMOS QUEBEC, N=52 N=53	^	37.329	171.000	14.6225	29.8424	184.68 (6.057)	226.99 (9.097)	255.00 (11.573)	281.87 (14.089)	316.65 (17.451)	342.72 (20.017)	428.84 (28.630)
	H	38.278	170.910	13.7208	29.8346	184.94 (6.052)	228.33 (8.999)	257.05 (11.390)	284.60 (13.821)	320.27 (17.073)	347.00 (19.554)	435.31 (27.889)

In the next section, the 3-parameter lognormal distribution is considered.

3.6 3-PARAMETER LOGNORMAL DISTRIBUTION

The 3-parameter lognormal distribution is another widely used probability distribution in fitting annual flood events. This distribution is a general skewed distribution and is exceptionally flexible and well suited for flood frequency analysis.

The pdf of the 3-parameter lognormal distribution is given by:

$$f(x/a,\ \mu,\ \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma(x - a)} \exp\left\{-\frac{1}{2\sigma^2}\left[\log(x_i - a) - \mu\right]^2\right\} \qquad \ldots (3.32)$$

$$-\infty < a < x_{(1)}; \quad -\infty < \mu < \infty; \quad \sigma^2 > 0$$

where $x_{(1)}$ is the smallest value of x, a is the threshold parameter, and (x - a) is the reduced variable. The parameters μ and σ^2 are the mean and variance of log(x-a) respectively. In other words,

$$\log(x - a) \sim N(\mu,\ \sigma^2)$$

The parameter, a, has to be estimated from the available sample in terms of the random variable x.

Many methods have been proposed to estimate the parameters of the 3-parameter distribution. Sangal and Biswas (1970), Burges et al. (1975) and Stedinger (1980) have discussed some of these methods. In view of the desirable asymptotic properties of the maximum likelihood estimators, the MLE is still preferred by most hydrologists. Sinha (1986a) and Kite (1977) have shown how to obtain maximum likelihood estimates of the parameters (a, μ, σ^2). The procedure by Kite (1977) is used here.

3.6.1 Maximum Likelihood Estimates

Given a random sample $\underline{x} = (x_1, x_2, \ldots x_n)$, the logarithm of the likelihood function of the pdf (3.32) is given by:

$$L = \text{constant} - n\log\sigma - \sum_{i=1}^{n} \log (x_i - a) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[\log(x_i - a) - \mu \right]^2 \quad \ldots (3.33)$$

Taking partial derivatives with respect to a, μ and σ^2 and equating to zero, one gets:

$$\frac{\partial L}{\partial a} = \sum_{i=1}^{n} \log(x_i - a)(x_i - a)^{-1} + (\sigma^2 - \mu) \sum_{i=1}^{n} (x_i - a)^{-1} = 0 \quad \ldots (3.34)$$

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[\log (x_i - a) - \mu \right] = 0 \qquad \ldots (3.35)$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma} \sum_{i=1}^{n} \left[\log (x_i - a) - \mu \right]^2 = 0 \quad \ldots (3.36)$$

From (3.35), one obtains:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \log (x_i - \hat{a}) \qquad \ldots (3.37)$$

and from (3.36):

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \log^2 (x_i - \hat{a}) - \left[\frac{1}{n} \sum_{i=1}^{n} \log (x_i - \hat{a}) \right]^2 \quad \ldots (3.38)$$

Substituting (3.37) and (3.38) into (3.34) a function in $\hat{a}$ only is obtained,

$$\begin{aligned} f(\hat{a}) = {} & n^2 \sum_{i=1}^{n} \log (x_i - \hat{a}) (x_i - \hat{a})^{-1} \\ & + \sum_{i=1}^{n} (x_i - \hat{a})^{-1} \left\{ n \sum_{i=1}^{n} \log^2 (x_i - \hat{a}) \right. \\ & \left. - \left[\sum_{i=1}^{n} \log (x_i - \hat{a}) \right]^2 + n \sum_{i=1}^{n} \log (x_i - \hat{a}) \right\} = 0 \end{aligned}$$

$$\ldots (3.39)$$

This equation can be solved by iteration using an appropriate starting value. Having obtained $\hat{a}$, one can then calculate for $\hat{\mu}$ and $\hat{\sigma}^2$ from (3.37) and (3.38). In this study, a value of $0.8x_{(1)}$ is used as the starting value. A Newton-Raphson procedure is then used to solve (3.39).

3.6.2 T-Year Return Period Event

The T-year event for the 3-parameter lognormal distribution is given by:

$$x_T = \hat{a} + \exp[\hat{\mu} + \hat{\sigma}t] \qquad \ldots (3.40)$$

where t is the standard normal deviate. Since T is not a variable, the standard error of x_T, S_T, may be obtained from (3.4). That is:

$$S_T^2 = \left(\frac{\partial x_T}{\partial a}\right)^2 \mathrm{Var}(\hat{a}) + \left(\frac{\partial x_T}{\partial \mu}\right)^2 \mathrm{Var}(\hat{\mu}) + \left(\frac{\partial x_T}{\partial \sigma^2}\right)^2 \mathrm{Var}(\hat{\sigma}^2)$$

$$+ 2\left(\frac{\partial x_T}{\partial a}\right)\left(\frac{\partial x_T}{\partial \mu}\right) \mathrm{Cov}(\hat{a}, \hat{\mu}) + 2\left(\frac{\partial x_T}{\partial a}\right)\left(\frac{\partial x_T}{\partial \sigma^2}\right) \mathrm{Cov}(\hat{a}, \hat{\sigma}^2)$$

$$+ 2\left(\frac{\partial x_T}{\partial \mu}\right)\left(\frac{\partial x_T}{\partial \sigma^2}\right) \mathrm{Cov}(\hat{\mu}, \hat{\sigma}^2) \qquad \ldots (3.41)$$

From (3.40):

$$\frac{\partial x_T}{\partial a} = 1 \qquad \ldots (3.42)$$

$$\frac{\partial x_T}{\partial \mu} = \exp[\hat{\mu} + \hat{\sigma}t] \qquad \ldots (3.43)$$

$$\frac{\partial x_T}{\partial \sigma^2} = \frac{t \cdot \exp[\hat{\mu} + \hat{\sigma}t]}{2\hat{\sigma}} \qquad \ldots (3.44)$$

Let $w = \exp[\hat{\mu} + \hat{\sigma}t]$, then,

$$S_T^2 = \mathrm{Var}(\hat{a}) + w^2\mathrm{Var}(\hat{\mu}) + \frac{w^2t^2}{4\hat{\sigma}^2}\mathrm{Var}(\hat{\sigma}^2) + \frac{tw}{\hat{\sigma}}\mathrm{Cov}(\hat{a}, \hat{\sigma}^2) + 2w\,\mathrm{Cov}(\hat{a}, \hat{\mu}) + \frac{tw^2}{\hat{\sigma}}\mathrm{Cov}(\hat{\sigma}^2, \hat{\mu}) \quad \ldots (3.45)$$

The variance-covariance matrix is the inverse of the symmetric matrix

$$[I] = \begin{bmatrix} -\frac{\partial^2 L}{\partial a^2} & -\frac{\partial^2 L}{\partial a\,\partial \mu} & -\frac{\partial^2 L}{\partial a\,\partial \sigma^2} \\ -\frac{\partial^2 L}{\partial \mu\,\partial a} & -\frac{\partial^2 L}{\partial \mu^2} & -\frac{\partial^2 L}{\partial \mu\,\partial \sigma^2} \\ -\frac{\partial^2 L}{\partial \sigma^2\,\partial a} & -\frac{\partial^2 L}{\partial \sigma^2\,\partial \mu} & -\frac{\partial^2 L}{\partial (\sigma^2)^2} \end{bmatrix} \quad \ldots (3.46)$$

That is,

$$[\sigma_{ij}] = \begin{bmatrix} \mathrm{Var}(\hat{a}) & \mathrm{Cov}(\hat{a}, \hat{\mu}) & \mathrm{Cov}(\hat{a}, \hat{\sigma}^2) \\ \mathrm{Cov}(\hat{\mu}, \hat{a}) & \mathrm{Var}(\hat{\mu}) & \mathrm{Cov}(\hat{\mu}, \hat{\sigma}^2) \\ \mathrm{Cov}(\hat{\sigma}^2, \hat{a}) & \mathrm{Cov}(\hat{\sigma}^2, \hat{\mu}) & \mathrm{Var}(\hat{\sigma}^2) \end{bmatrix} = [I]^{-1} \quad \ldots (3.47)$$

It can be shown (Appendix D) that:

$$\sigma_{11} = \frac{1}{2nD} \qquad \ldots (3.48)$$

$$\sigma_{12} = \frac{-\exp(\sigma^2/2 - \mu)}{2nD} \qquad \ldots (3.49)$$

$$\sigma_{13} = \frac{\sigma^2}{nD} \cdot \exp(\sigma^2/2 - \mu) \qquad \ldots (3.50)$$

$$\sigma_{22} = \frac{\sigma^2}{nD} \cdot \frac{\sigma^2 + 1}{2\sigma^2} \cdot \exp[\, 2(\sigma^2 - \mu)\,] - \exp(\sigma^2 - 2\mu) \qquad \ldots (3.51)$$

$$\sigma_{23} = \frac{-\sigma^2}{nD} \cdot \exp(\sigma^2 - 2\mu) \qquad \ldots (3.52)$$

$$\sigma_{33} = \frac{\sigma^2}{nD} \cdot (\sigma^2 + 1) \cdot \exp[\, 2(\sigma^2 - \mu)\,] - \exp(\sigma^2 - 2\mu) \qquad \ldots (3.53)$$

where D is the determinant of (3.46).

$$D = \frac{\sigma^2 + 1}{2\sigma^2} \cdot \exp\,[2(\sigma^2 - \mu)] - \frac{\exp(\sigma^2 - 2\mu)}{2\sigma^2} \cdot (2\sigma^2 + 1) \qquad \ldots (3.54)$$

3.6.3 Bayes Estimates

If one is 'in-ignorance' about the parameters (a, μ, σ^2), the 'vague' prior

$$g(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \quad \text{and}$$

$$h(a) = \text{constant}$$

would be appropriate (Jeffreys, 1961). It is reasonable to believe that a is distributed independently of σ^2 and μ since any prior knowledge about a is not likely to be much influenced by one's knowledge about the values of these parameters (Box and Tiao, 1973).

Thus, the joint prior distribution of (a, μ, σ^2) may be written as:

$$\begin{aligned} v(a, \mu, \sigma^2) &\propto g(\mu, \sigma^2)\, h(a/\mu, \sigma^2) \\ &\approx g(\mu, \sigma^2)\, h(a) \\ &\propto \frac{1}{\sigma^2} \end{aligned} \qquad \ldots (3.55)$$

Let $\underline{x} = (x_1, x_2, \ldots x_n)$ be a random sample of size n from the pdf given by (3.32). The likelihood function is:

$$\ell(\underline{x}/a, \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\ \sigma}\right)^n \prod_{i=1}^{n} (x_i - a) \cdot \exp\left\{-\frac{1}{2\sigma^2}[\log(x_i - a) - \mu]^2\right\}$$

Combining the likelihood with the prior distribution given by (3.55) and applying Bayes' theorem, the joint posterior distribution is obtained:

$$\Pi(a, \mu, \sigma^2) = K.\ell(\underline{x}/a, \mu, \sigma^2).v(a, \mu, \sigma^2) \quad \ldots (3.57)$$

where K is the normalizing constant given by:

$$K^{-1} = \int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{x_{(1)}} \ell(\underline{x}/a, \mu, \sigma^2)\, v(a, \mu, \sigma^2)\, da\, d\mu\, d\sigma^2 \quad ..(3.58)$$

Under the squared-error loss function, Bayes estimators of a function is the posterior expectation of that function. For example, the Bayes estimate of a is given by:

$$E(a/\underline{x}) = a^* = \frac{\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{x_{(1)}} a.v(a, \mu, \sigma^2)\ell(\underline{x}/a, \mu, \sigma^2)\, da d\mu d\sigma^2}{\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{x_{(1)}} \ell(\underline{x}/a, \mu, \sigma^2).v(a, \mu, \sigma^2)\, da d\mu d\sigma^2} \quad \ldots (3.59)$$

3.6.4 Bayesian Approximation

Lindley's expansion of (3.7) for the 3-parameter case is given by:

$$E[u(\theta)] = u + (u_1 a_1 - u_2 a_2 + u_3 a_3 + a_4 + a_5) + \frac{1}{2}\left[A(u_1\sigma_{11} + u_2\sigma_{12} + u_3\sigma_{12}) + B(u_1\sigma_{21} + u_2\sigma_{22} + u_3\sigma_{23}) + C(u_1\sigma_{31} + u_2\sigma_{32} + u_3\sigma_{33}) \right]_{\hat{\theta}} \quad \ldots (3.60)$$

which has to be evaluated at $\hat{\theta} = (\hat{a}, \hat{u}, \hat{o}^2)$.

Also,

$$a_1 = \rho_1\sigma_{11} + \rho_2\sigma_{12} + \rho_3\sigma_{13}$$

$$a_2 = \rho_1\sigma_{21} + \rho_2\sigma_{22} + \rho_3\sigma_{23}$$

$$a_3 = \rho_1\sigma_{31} + \rho_2\sigma_{32} + \rho_3\sigma_{33}$$

$$a_4 = u_{12}\sigma_{12} + u_{13}\sigma_{13} + u_{23}\sigma_{23}$$

$$a_5 = \frac{1}{2}(u_{11}\sigma_{11} + u_{22}\sigma_{22} + u_{33}\sigma_{33})$$

$$A = \sigma_{11}L_{111} + 2\sigma_{12}L_{121} + 2\sigma_{13}L_{131} + 2\sigma_{23}L_{231} + \sigma_{22}L_{221} + \sigma_{33}L_{331}$$

$$B = \sigma_{11}L_{112} + 2\sigma_{12}L_{122} + 2\sigma_{13}L_{133} + 2\sigma_{23}L_{232} + \sigma_{22}L_{222} + \sigma_{33}L_{332}$$

$$C = \sigma_{11}L_{113} + 2\sigma_{12}L_{123} + 2\sigma_{13}L_{133} + 2\sigma_{23}L_{233} + \sigma_{22}L_{223} + \sigma_{33}L_{333}$$

where,

$$u_{ij} = \frac{\partial^2 u}{\partial\theta_i\ \partial\theta_i}; \quad L_{ijk} = \frac{\partial^3 L}{\partial\theta_i\ \partial\theta_j\ \partial\theta_k};$$

$$\rho \equiv \rho(\theta) = \log v(\theta); \quad \rho_j = \frac{\partial\rho}{\partial\theta_j}; \quad \text{and}$$

σ_{ij} = (i, j)th element of the variance-covariance matrix (3.47). The subscripts 1, 2, 3 refer to (a, μ, σ^2) respectively. See Appendices D and E for the evaluation of the σ_{ij} matrix, L_{ijk}'s and u_{ij}'s.

3.6.5 Numerical Example

The annual extreme flows of the St. Marys River at Stillwater (n = 69) is used as a numerical example. Figure 3.2 shows the fit of the 3-parameter lognormal distribution to the observed data where the parameters are estimated by the maximum likelihood method. The maximum likelihood estimates are:

St. Marys River
at Stillwater

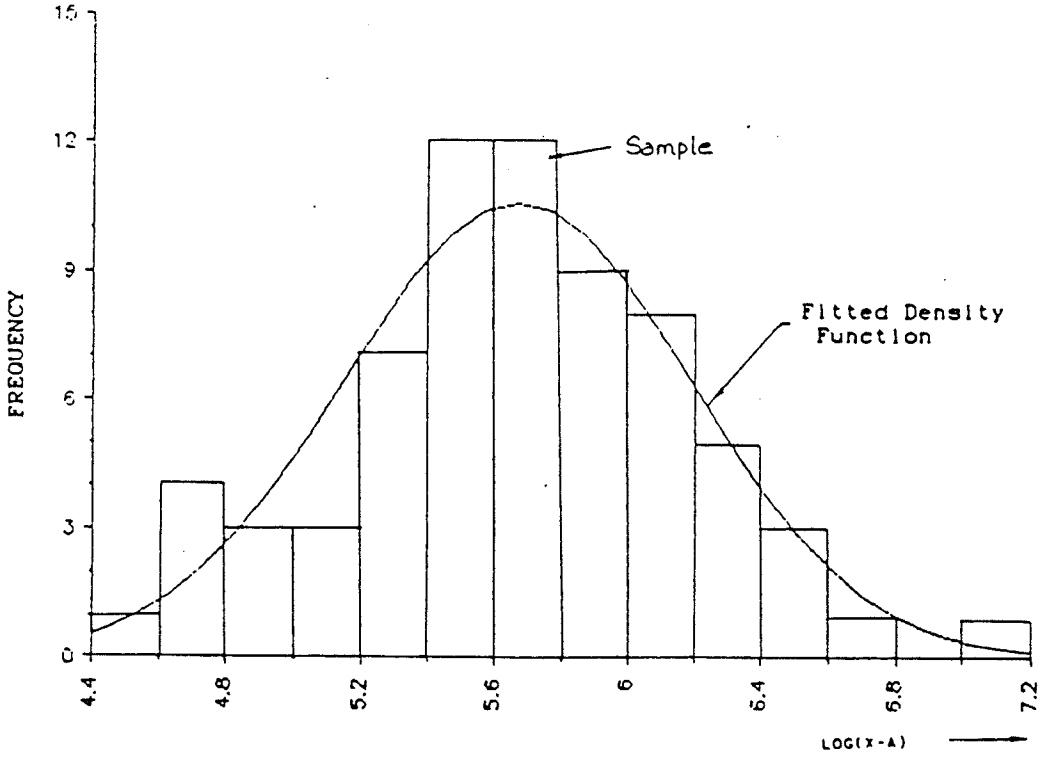


Figure 3.2 3-parameter Lognormal Distribution

$$\hat{a} = 116.13 \qquad \hat{\mu} = 5.6647 \qquad \hat{\sigma}^2 = 0.27026$$

$$\sigma_{ij} = \begin{bmatrix} 2007.811 & -7.96666 & 4.30617 \\ -7.96666 & 0.03553 & -0.01709 \\ 4.30617 & -0.01709 & 0.01135 \end{bmatrix}$$

$$L_{ijk} = \begin{bmatrix} -1.464 \times 10^{-4} & -5.2664 \times 10^{-3} & 3.0010 \times 10^{-2} \\ 0 & 0 & 9.4468 \times 10^{2} \\ -7.49642 & 0 & 6.9909 \times 10^{3} \end{bmatrix}$$

and $L_{123} = 3.74821$.

Also, $\rho_1 = \rho_2 = 0$, $\rho_3 = -3.70014$.

$$a_1 = -15.93343, \qquad a_2 = 6.32354 \times 10^{-2}$$

$$a_3 = -4.19966 \times 10^{-2}, \qquad A = -1.64713 \times 10^{-1}$$

$$B = -10.5824, \qquad C = 48.90154$$

Let $u = \hat{a}$, u_1, $u_2 = u_3 = 0$, $u_{ij} = 0$, $\forall_{ij}$

Substituting into (3.60),

$$E(a/\underline{x}) = a^* = 82.283$$

Similarly,

$$E(\mu/\underline{x}) = \mu^* = 5.77818$$

$$E(\sigma^2/\underline{x}) = \sigma^{2*} = 0.24157$$

Also let $u = a^2$, $u_1 = 2a$, $u_{11} = 2$; $u_2 = u_3 = u_{ij} = 0$ for all other i, js.

Substituting into (3.60),

$$E\,(a^2/\underline{x}) = 7{,}632.673$$

Therefore,

$$\sigma_{11}^{*} = \mathrm{Var}(a/\underline{x}) = E(a^2/\underline{x}) - a^{*2}$$

$$= 862.181 \quad \text{which is less than } \sigma_{11} = 2007.811$$

Similarly,

$$\sigma_{22}^{*} = 0.02266 < \hat{\sigma}_{22} = 0.03553$$

$$\sigma_{33}^{*} = 0.01052 < \hat{\sigma}_{33} = 0.01135$$

Table 3.4 summarizes the results.

TABLE 3.4

Bayes (*) And MLE (^) Estimates Of (a, μ, σ^2) For The St. Marys River At Stillwater

(n = 69)

Parameter	^	*	Var(^)	Posterior Var.
a	116.13	82.283	2007.811	862.181
μ	5.6647	5.7782	0.03553	0.02266
σ^2	0.2703	0.2416	0.01135	0.01052

Bayes estimates of the T-year event is obtained from (3.40). For T = 2 year return period, t = 0.

Let

$$u = x_T = 404.626$$

$$u_1 = 1, \quad u_{22} = 288.491, \quad u_{ij} = o \text{ for all other ijs.}$$

$$a_4 = 0, \quad a_5 = 5.12465$$

Substituting into (3.60),

$$E(x_T/\underline{x}) = 408.674$$

Also let $u = x_T^2 = 163,722.04$, $u_1 = 809.252$

$$u_2 = 233,462.18, \quad u_3 = u_{13} = 0, \quad u_{12} = 576.983$$

$$u_{11} = 2, \quad u_{22} = 399,916.82, \quad u_{23} = u_{33} = 0$$

$$a_4 = -4596.629, \quad a_5 = 9111.782$$

Substituting into (3.60),

$$E(x_T^2/\underline{x}) = 167,365.8606$$

Hence,

$$Var(x_T/\underline{x}) = 167,365.8606 - 408.674^2$$

$$= 351.6375$$

and standard-error, S_T is given by:

$$S_T^* = \sqrt{Var(x_T/\underline{x})} = 18.752 < \hat{S}_T = 19.184$$

Table 3.5 summarizes the result for return periods T = 2, 5, 10, 50 and 100 years for the St. Marys River at Stillwater.

TABLE 3.5

Standard Error Of Estimates Of T-Year Event
St. Marys River At Stillwater n = 69
3-Parameter Lognormal Distribution

T-year	2	5	10	20	50	1000
t	0.0	0.8416	1.2816	1.6449	2.0538	2.3264
$\hat{x}_T$	404.63	562.97	677.87	794.57	955.27	1083.0
$\hat{S}_T$	19.184	32.542	48.809	70.633	107.66	141.78
x_T^*	408.67	569.14	683.65	799.12	957.01	1082.0
S_T^*	18.752	31.952	48.458	70.486	107.64	141.77

(^) MLE, (*) Bayes estimates

From the tables above, the posterior variances of the parameters and the posterior standard-errors of the T-year event are less than the corresponding MLE's.

Bayes estimates using Lindley's procedure was repeated for 11 other maximum annual flows from rivers all over Canada. The results are summarized in Table 3.6.

Four of the 12 flood data analysed showed negative posterior variances for some of the parameters. The negative posterior variances, however, do not affect the estimate of the T-year flood and its standard error. The Lepreau River (n = 68) in the Atlantic Provinces is one example where the posterior variance of a is negative. It is possible that regularity conditions for the maximum likelihood estimation of the threshold parameter a are not met for this particular flood series. This may be a reason why a negative posterior variance was obtained from the calculation. Also $E(a^2/\underline{x})$ may have been underestimated and $[E(a/\underline{x})]^2$ overestimated. In addition, Lindley's procedure being a asymptotic expansion, some rounding off errors could have occurred.

3.7 SUMMARY

This chapter has shown that Bayes estimates have somewhat smaller posterior variances to their MLE counterparts, indicating that Bayes estimates of the flood events at various return periods are as reliable or more reliable than the MLE's.

Table 3.6 Comparison of MLE and Bayes Estimates (3 Parameter Lognormal Distribution)

RIVER		PARAMETER			VARIANCE			STANDARD ERRORS OF X_T IN PARENTHESIS					
		a	μ	σ^2	a	μ	σ^2	T = 2	5	10	20	50	100
BOW RIVER AT BANFF, ALBERTA N=76	^	-10.448	5.4194	.05956	6885.845	.1442	.00213	215.30 (6.799)	266.77 (9.041)	298.19 (11.959)	326.81 (15.888)	362.20 (22.292)	387.84 (27.876)
	*	-86.375	5.6898	.04757	1120.8667	.07108	.00198	216.69 (6.655)	268.50 (8.875)	299.44 (11.894)	327.27 (15.881)	361.33 (22.275)	385.80 (27.801)
ST. MARY'S RIVER AT STILLWATER ATLANTIC, N=69	^	116.13	5.6647	.27026	2007.811	.03553	.01135	404.63 (19.184)	562.97 (32.542)	677.81 (48.809)	794.57 (70.633)	955.27 (107.66)	1,083.00 (141.78)
	*	82.283	5.7783	.24161	861.889	.02262	.01053	408.67 (18.752)	569.14 (31.952)	683.65 (48.458)	799.12 (70.486)	957.01 (107.64)	1,082.00 (141.77)
LAHAVE RIVER AT WEST NORTHFIELD ATLANTIC, N=69	^	53.704	5.0421	.35734	384.655	.02813	.01542	208.50 (11.782)	309.70 (21.441)	386.72 (33.273)	467.50 (49.306)	582.08 (76.987)	675.59 (102.99)
	*	38.2264	5.1452	.32074	145.1001	.01749	.01408	211.16 (11.476)	313.71 (21.063)	390.46 (33.063)	470.25 (49.230)	582.66 (76.985)	673.97 (102.97)
ROSEWAY RIVER NEAR LOWER OHIO ATLANTIC, N=67	^	14.933	3.8801	.22317	75.5992	.04363	.00951	63.36 (2.977)	87.00 (4.840)	103.65 (7.109)	120.27 (10.139)	142.71 (15.231)	160.27 (19.873)
	*	8.3286	4.0074	.19766	31.981	.02742	.00886	63.98 (2.913)	87.94 (4.749)	104.54 (7.053)	120.96 (10.116)	142.99 (15.228)	160.15 (19.872)
LEPREAU RIVER AT LEPREAU, ATLANTIC N=68	^	28.641	3.7189	.68621	8.8377	.02042	.03331	69.86 (4.316)	111.41 (9.684)	147.81 (16.578)	189.65 (26.360)	254.56 (44.332)	311.80 (62.275)
	*	25.497	3.8161	.60735	-1.0413	.01097	.02709	71.14 (4.120)	113.09 (9.538)	148.96 (16.538)	189.63 (26.360)	252.01 (44.259)	306.55 (62.054)
SLOCAN RIVER NEAR CRESENT VALLEY B.C., N=60	^	-89.806	6.2817	.04956	59666.156	.22013	.00224	444.89 (16.544)	555.06 (21.597)	621.42 (28.279)	681.33 (37.312)	754.82 (52.023)	807.67 (64.805)
	*	-357.881	6.6804	.03508	-12198.099	.06115	.00203	448.85 (16.061)	559.76 (21.079)	624.56 (28.104)	682.06 (37.305)	751.52 (51.918)	800.88 (64.449)

Column 2: Estimates of the parameters (a,μ,σ^2)
Column 3: Variances of the parameters (a,μ,σ^2)
Column 4-9: Estimate of T-year flood with standard error in parenthesis

Table 3.6 Continue

RIVER		PARAMETER			VARIANCE			STANDARD ERRORS OF X_T IN PARENTHESIS					
		a	μ	σ^2	a	μ	σ^2	T = 2	5	10	20	50	100
RED RIVER AT REDWOOD BRIDGE MANITOBA, N=93	^	-12612	10.771	.18937	65173958.48	.03678	.00576	34,997.00 (2293.1)	56,054.00 (3605.9)	70,545.00 (5208.2)	84,788.00 (7344.1)	103,760.00 (10909.0)	118,410.00 (14131.0)
	M	-17835.8	10.870	.17294	37884180.40	.02686	.00549	35,392.00 (2258.8)	56,650.00 (3556.4)	71,101.00 (5178.5)	85,215.00 (7331.7)	103,920.00 (10907.0	118,320.00 (14131.0)
ASSINIBOINE RIVER AT HEADINGLEY MANITOBA, N=73	^	-2315.1	9.3214	.27823	2733720	.03273	.01107	8,859.50 (732.68)	15,104.00 (1251.4)	19,654.00 (1883.2)	24,295.00 (2731.6)	30,700.00 (4173.1)	35,805.00 (5504.1)
	M	-3532.84	9.4275	.2505	1250732.38	.02146	.01031	9,010.70 (716.90)	15,334.00 (1230.0)	19,872.00 (1870.5)	24,464.00 (2726.3)	30,763.00 (4172.7)	35,764.00 (5503.9)
SAUGEEN RIVER NEAR WALKERTON ONTARIO, N=70	^	-83.943	5.8948	.10076	10771.226	.09178	.00396	279.19 (14.784)	390.39 (20.953)	461.48 (28.681)	528.15 (38.995)	612.99 (55.916)	675.98 (70.856)
	M	-169.035	6.0912	.08513	3530.504	.05318	.00371	282.11 (14.493)	394.46 (20.554)	464.92 (28.475)	530.31 (38.935)	612.81 (55.916)	673.69 (70.819)
NORTH MAGNETAWAN RIVER NEAR BURKS FALL ONTARIO, N=69	^	-2.776	3.7979	.08934	189.0823	.10519	.00355	41.83 (1.724)	54.59 (2.404)	62.66 (3.262)	70.16 (4.410)	79.65 (6.289)	86.64 (7.942)
	M	-14.474	4.0154	.07424	52.226	.05792	.00332	42.18 (1.688)	55.07 (2.357)	63.04 (3.239)	70.39 (4.404)	79.58 (6.289)	86.32 (7.935)
ENGLISH RIVER NEAR SIOUX LOOKOUT ONTARIO, N=60	^	-14.599	5.6169	.20764	2993.72	.05217	.00984	260.43 (17.251)	388.98 (27.628)	478.58 (40.273)	567.37 (57.148)	686.57 (85.414)	779.30 (111.08)
	M	-58.533	5.7641	.18056	1063.525	.03049	.0091	264.17 (16.841)	394.65 (27.041)	483.90 (39.920)	571.50 (56.999)	688.22 (85.398)	778.55 (111.08)
MCEACHERN CREEK NEAR INTERNATIONAL BOUNDARY, SASK. N=53	^	-1.472	2.9196	1.3875	.38801	.0307	.10749	17.06 (3.063)	48.47 (9.423)	82.39 (18.645)	127.19 (33.107)	206.80 (62.859)	285.67 (95.769)
	M	-3.10902	3.0872	1.0396	-2.26986	.0026	-.0136	18.82 (2.510)	49.54 (9.362)	80.11 (18.504)	118.35 (31.903)	182.82 (58.102)	243.92 (86.189)

Column 2: Estimates of the parameters (a,μ,σ^2)
Column 3: Variances of the parameters (a,μ,σ^2)
Column 4-9: Estimate of T-year flood with standard error in parenthesis

Once the MLE have been obtained, obtaining Bayes estimates using Lindley's procedure is simple and can be worked out on a desk top calculator.

It has been found that for the 3-parameter lognormal distribution, negative posterior variances of the estimates of some parameters are obtained. This 'irregularity', however, does not affect the posterior estimates and standard error for the T-year events. The determination of Bayes estimates for (a, μ, σ^2) are independent of the determination of posterior expectation and standard errors of the T-year events. For the Gumbel distribution, the MLE of (μ, σ) are easily obtained. Here the problem of negative posterior variance does not arise.

In spite of the minor shortcoming of Lindley's method, this procedure is a useful technique in flood frequency analysis.

In the next chapter, the predictive distribution approach to flood analysis is considered.

CHAPTER 4

PREDICTIVE DISTRIBUTION

4.1 GENERAL

In the previous chapter, flood analysis was considered from a scientist/statistician's point of view. That is, the object of the analysis is the description of the stochastic variability of the observed floods.

Engineers, however, make use of the flood data to guide engineering decisions, for example, when determining the height of a dyke or the spillway capacity for a dam. Therefore, the engineer, unlike the scientist, must go beyond a mere description of the variability of nature. Since he is concerned with predictions, he is also concerned with the uncertainty in that description. Hence stochastic and parameter uncertainty must be combined in the decision making process. This can be done by using the predictive probability distribution as described in Section 2.4. The predictive distribution quantifies the risk of a future flood event on the basis of present information. By avoiding the relative frequency definition of probability and using the Bayesian concept of risk, the analysis sidesteps the conceptual problems associated with designations such as the thousand or ten thousand year flood. There is no conceptual problem with

a risk designation of 0.1, 0.01, or even 0.001 that a dyke will be overtopped within the next 50 or 100-year planning period. This risk is defined here as the probability based on all available information that a future flood discharge will be equalled or exceeded in any year during the period bounded by the planning horizon.

The predictive distribution approach to flood risk analysis has been advocated by a number of researchers (e.g. Wood et al., 1974; Vicens et al., 1975; Bodo and Unny, 1976; Stedinger, 1983; Bernier, 1967; Russell, 1982). Except for Russell (1982) who obtained the predictive distribution by discretization and computer calculation, the others obtained the predictive distribution either analytically which requires rather sophisticated mathematics or by numerical integration which is a problem when integrating to infinity (Bodo and Unny, 1976). Obtaining the predictive distribution for distributions other than the normal or lognormal distribution such as the popular Gumbel and Gamma distributions is very difficult (Stedinger, 1983).

In this chapter, the probability of exceedence of a future flood discharge or the risk that a future flood discharge is exceeded or equalled is obtained by using an important property of the predictive distribution. The property is that the probability of exceedence of the predictive distribution is the Bayes estimator of the probability of exceedence (under a squared-error loss

function) [Sinha, 1985]. The Bayes estimates of the probability of exceedence for the Gumbel distribution and 2-parameter lognormal distribution will be obtained using a 'vague' prior distribution and Lindley's Bayesian approximation procedure. The posterior variances of these estimates are then compared to their corresponding MLE's. In addition, Bayes estimates of the probability of exceedence for the 2-parameter lognormal distribution obtained by Lindley's method will be compared to those obtained analytically.

Russell's (1982) method of obtaining the predictive distribution will also be discussed in this chapter. Some modifications are made to his original scheme to improve the accuracy of the estimates and to deal with serially correlated data. It will be assumed that the random variable is adequately described by a normal distribution.

4.2 ESTIMATES OF THE PROBABILITY OF EXCEEDENCE

The risk that a flood discharge, q, will be exceeded or equalled in any year within the next 50 or 100 year planning horizon is by definition the probability of exceedence of q, namely P_q. That is:

$$P_q = P(X \geq q) \qquad \ldots (4.1)$$

$$= \int_q^{\infty} f(x/\theta)\, dx \qquad \ldots (4.2)$$

where $f(x/\theta)$ is the underlying probability density function of x with parameter(s) θ.

The maximum likelihood estimate of P_q is obtained by substituting the MLE of θ into (4.2).

$$\hat{P}_q = \int_q^{\infty} f(x/\hat{\theta})\, dx \qquad \ldots (4.3)$$

The asymptotic standard error of $\hat{P}_q$, $\hat{S}_q$ can be obtained from (3.4). For a two parameter case, this is given by:

$$\hat{S}_q^2 = \left(\frac{\partial P_q}{\partial \theta_1}\right)^2 Var(\hat{\theta}_1) + \left(\frac{\partial P_q}{\partial \theta_2}\right)^2 Var(\hat{\theta}_2) + 2\left(\frac{\partial P_q}{\partial \theta_1}\right)\left(\frac{\partial P_q}{\partial \theta_2}\right) Cov(\hat{\theta}_1, \hat{\theta}_2) \qquad \ldots (4.4)$$

where $Var(\hat{\theta}_1)$, $Var(\hat{\theta}_2)$, and $Cov(\hat{\theta}_1, \hat{\theta}_2)$ are given by (3.5) - (3.6). The standard-error of $\hat{P}_q$ is then the square-root of (4.4).

To obtain the probability of exceedence of the predictive distribution, $f(x/\theta)$ in (4.2) is replaced by $\tilde{f}(x)$ given in (2.7).

$$\tilde{P}_q = \int_q^{\infty} \tilde{f}(x)\, dx \qquad \ldots (4.5)$$

$$= \int_q^{\infty} \int_{\Omega} f(x/\theta)\, \pi(\theta)\, d\theta\, dx \qquad \ldots (4.6)$$

Under the squared-error loss function, Bayes estimate of the probability of exceedence is given by:

$$P_q^* = E[P(X \geq q/\theta)] = \int_\Omega P(X \geq q/\theta)\, \pi(\theta)\, d\theta \quad \ldots (4.7)$$

$$= \int_\Omega \int_q^\infty f(x/\theta)\, dx\, \pi(\theta)\, d\theta$$

$$= \int_q^\infty \int_\Omega f(x/\theta)\, \pi(\theta)\, d\theta\, dx \quad \ldots (4.8)$$

$$= \tilde{P}_q$$

Therefore, Bayes estimate of the probability of exceedence is equal to the probability of exceedence of the predictive distribution (Sinha, 1985). This property is useful because it is easier to obtain Bayes estimate using Lindley's expansion than to obtain the probability of exceedence from the predictive distribution which requires rather difficult mathematics.

The posterior standard-error of P_q^*, S_q^* is obtained as follows:

$$Var(P_q/\underline{x}) = E(P_q^2/\underline{x}) - P_q^{*2} \quad \ldots (4.9)$$

and define, $S_q^* = \sqrt{Var(P_q/\underline{x})}$ $\ldots (4.10)$

In the following Sections 4.3 and 4.4, maximum likelihood estimates and Bayes estimates of the probability of exceedence will be obtained for the Gumbel and the lognormal distributions.

4.3 GUMBEL DISTRIBUTION

The Gumbel distribution has been described in Section 3.5.

4.3.1 Maximum Likelihood Estimate Of P_q

The probability of exceedence, P_q, for the Gumbel distribution (3.9) is given by:

$$P_q = 1 - \exp\left\{-\exp\left[-\frac{q - \mu}{\sigma}\right]\right\} \qquad \ldots(4.11)$$

The maximum likelihood estimates of P_q, $\hat{P}_q$, is obtained by substituting $(\hat{\mu}, \hat{\sigma})$ obtained in (3.15) and (3.17) into (4.4) and,

$$\frac{\partial P_q}{\partial \sigma} = \frac{q - \hat{\mu}}{\hat{\sigma}^2} \cdot w_1 . w_2 = u_1 \qquad \ldots (4.12)$$

$$\frac{\partial P_q}{\partial \mu} = \frac{1}{\hat{\sigma}} \cdot w_1 w_2 = u_2 \qquad \ldots (4.13)$$

where, $w_1 = \exp\left\{-\exp\left[-\frac{q - \hat{\mu}}{\hat{\sigma}}\right]\right.$; $w_2 = \exp\left[-\frac{q - \hat{\mu}}{\hat{\sigma}}\right]$

This leads to:

$$\hat{S}_q = u_1^2\sigma_{11} + u_2^2\sigma_{22} + 2u_1u_2\sigma_{12} \quad \ldots (4.14)$$

where, $\sigma_{11} = \text{Var}(\hat{\sigma})$; $\sigma_{22} = \text{Var}(\hat{\mu})$; and $\sigma_{12} = \text{Cov}(\hat{\sigma}, \hat{\mu})$ and are given by (3.22) - (3.23).

4.3.2 Bayes Estimate Of P_q

The 'vague' prior given by (3.27) and Lindley's expansion (3.31) will be used to obtain Bayes estimate of the probability of exceedence. To obtain P_q^*, let $u(\theta) = P_q$ in (3.31). And,

$$u_1 = \frac{\partial P_q}{\partial \sigma}, \quad u_2 = \frac{\partial P_q}{\partial \mu}, \quad u_{12} = \frac{\partial^2 P_q}{\partial \sigma\, \partial \mu}, \quad u_{11} = \frac{\partial^2 P_q}{\partial \sigma^2},$$

$$u_{22} = \frac{\partial^2 P_q}{\partial \mu^2}.$$

All other constants are as defined in Section 3.5.3 and deviations of u_{ij}'s are given in Appendix C.

4.3.3 Numerical Example

The annual maximum flows of the Turtle River at Mine Center (n = 58) are again used as an example. The maximum likelihood estimates for this river are:

$$\hat{\sigma} = 45.810, \qquad \hat{\mu} = 101.270$$

$$[\sigma_{ij}] = \begin{bmatrix} 22.5061 & 9.5444 \\ 9.5444 & 40.2302 \end{bmatrix}$$

$o_1 = -0.02183$, $L_{30} = 0.0054401$, $L_{03} = -0.0006033$, $L_{12} = 0.0014625$, $L_{21} = -0.0014985$.

Let $q = 300\ m^3/s$. The maximum likelihood estimate of P_q is given by:

$$\hat{P}_q = 1 - \exp\left\{-\exp\left[-\left(\frac{\hat{q} - \hat{\mu}}{\hat{\sigma}}\right)\right]\right\} = 0.01298$$

The asymptotic standard-error, $\hat{S}_q$, is obtained as follows: From (4.12) and (4.13),

$$\frac{\partial P_q}{\partial \sigma} = u_1 = 0.0012208; \qquad \frac{\partial P_q}{\partial \mu} = u_2 = 0.0002814$$

Substituting into (4.14),

$$\hat{S}_q = 0.006579$$

Bayes estimate of the probability of exceedence is obtained as follows:

Let $u = \hat{P}_q = 0.01298$

$$u_1 = \frac{\partial P_q}{\partial \sigma} = 0.0012208, \quad u_2 = \frac{\partial P_q}{\partial \mu} = 0.0002814,$$

$$u_{12} = \frac{\partial^2 P}{\partial \sigma \, \partial \mu} = \frac{q - \mu}{\sigma^3} \cdot w_2 \left[-\sigma u_2 + w_1 - \frac{w_1 \sigma}{(q - \mu)} \right]$$

$$= 0.00002016$$

$$u_{11} = \frac{\partial^2 P_q}{\partial \sigma^2} = - \frac{(q - \mu)}{\sigma^4} \cdot w_2 \left[\sigma^2 u_1 - w_1 (q - \mu) + 2\sigma w_1 \right.$$

$$= 0.000060798$$

$$u_{22} = \frac{\partial^2 P_q}{\partial \mu^2} = - \frac{1}{\sigma} \cdot w_2 . u_2 = -0.00000008023$$

where $w_1 = \exp\left[-\exp - \left(\frac{q - \mu}{\sigma} \right) \right]$

$$w_2 = \exp\left[- \left(\frac{q - \mu}{\sigma} \right) \right]$$

Substituting into (3.31), we get:

$$E[P(X \geq 300/\underline{x}] = P_q^* = 0.01515$$

Also let $v = u^2 = P_q^{\,2} = 0.0001684$, and

$$v_1 = 2u \cdot u_1 = P_q^2 = 0.00003168,$$

$$v_2 = 2u \cdot u_2 = 0.000007303,$$

$$v_{12} = 2u.u_{12} + 2u_1u_2 = 0.0000012101,$$

$$v_{11} = 2u.u_{11} + 2u_1^2 = 0.000004558,$$

$$v_{22} = 2u.u_{22} + 2u_2^2 = 0.0000001563.$$

Substituting into (3.31) , one gets:

$$E[P_q^2/\underline{x}] = 0.00026805$$

The posterior variance of P_q is then given by:

$$Var(P_q/\underline{x}) = E[P_q^2/\underline{x}] - \{E[P_q/\underline{x}]\}^2$$

$$= 0.00003856$$

Define the posterior standard-error of P_q as

$$S_q(P_q/\underline{x}) = S_q^* = \sqrt{Var(P_q/\underline{x})}$$

$$= 0.006210$$

which is less than $\hat{S}_q = 0.006579$.

Table 4.1 summarizes the results for q = 200 to 500 in steps of 50 m^3/s, for the Turtle River at Mine Center, n = 58.

TABLE 4.1

PROBABILITY OF EXCEEDENCE AND STANDARD-ERROR
TURTLE RIVER AT MINE CENTER, n = 58

q	200	250	300	350	400	450	500
$\hat{P}_q$	0.10941	0.03816	0.01298	0.00435	0.00147	0.00049	0.00017
$\hat{S}_q$	0.03071	0.01505	0.00658	0.00271	0.00107	0.00042	0.00016
P_q^*	0.11468	0.04204	0.01515	0.00545	0.00196	0.00071	0.00026
S_q^*	0.03026	0.01454	0.00621	0.00249	0.00096	0.00036	0.00013

$\hat{P}_q$ = MLE of P_q

$\hat{S}_q$ = Standard error of $\hat{P}_q$

P_q^* = Bayes estimate of P_q

S_q^* = Posterior standard error of P_q^*

q = Flood discharge in m^3/s.

The table above shows that the posterior standard-errors are smaller than the corresponding standardf errors of the MLE's by a minute amount. However, Bayes estimates of the probabillty of exceedence are higher than the corresponding MLE's. This is also to be expected since Bayes estimates incorporate both stochastic as well as parameter uncertainty. See also Figure 4.1.

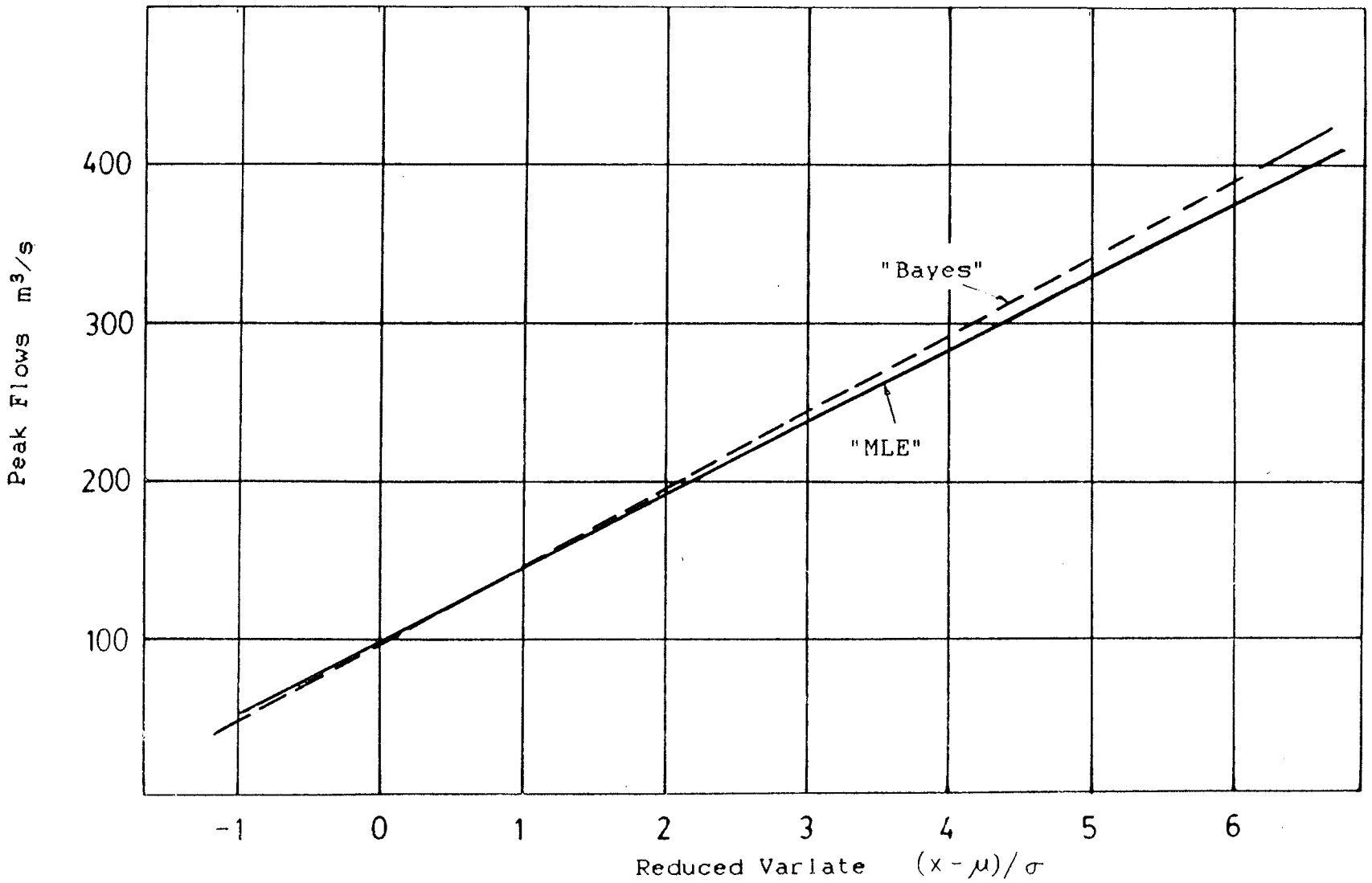


Figure 4.1 Bayes Probability of Exceedence (Gumbel Distribution) - Turtle River at Mine Center

The analysis was repeated for seven other maximum annual flows events or rivers all over Canada. The results are shown in Table 4.2.

In the next section, the 2-parameter lognormal distribution will be considered.

4.4 2-PARAMETER LOGNORMAL DISTRIBUTION

The 2-parameter lognormal or simply, the lognormal distribution has long been a favourite among hydrologists for describing the distribution of floods and other phenomena. Beard (1974) also concluded that among those distributions considered for modelling flood flow distributions across the United States, none were found to be superior to the lognormal distribution. Hence, it is reasonable to assume that flood flows in much of North America can be modelled by the simple and physically reasonable lognormal distribution when one lacks evidence to the contrary for the site in question (Stedinger, 1983). In addition, it is very easy to generate serially correlated lognormal variates needed for Monte Carlo simulations.

The probability density function of the lognormal distribution is given by:

Table 4.2 Bayes (*) and MLE (^) Estimates of the Probability of Exceedence of a Flood Discharge q. (q in m^3/s)

Red River at Redwood Bridge, n=93

q	1500	2000	2500	3000	4000	4500	5000
$\hat{P}q$	0.22830	0.09048	0.03411	0.01262	0.00170	0.00062	0.00023
$\hat{S}q$	0.03583	0.02171	0.01104	0.00513	0.00097	0.00041	0.00017
Pq*	0.23087	0.09374	0.03644	0.01397	0.00204	0.00079	0.00030
Sq*	0.03574	0.02146	0.01079	0.00494	0.00091	0.00037	0.00015

Saugeen River near Walkerton, n=70

q	400	500	600	700	800	900	1000
$\hat{P}q$	0.18091	0.06970	0.02581	0.00942	0.00342	0.00124	0.00045
$\hat{S}q$	0.03661	0.02069	0.01010	0.00457	0.00197	0.00083	0.00034
Pq*	0.18490	0.07369	0.02844	0.01088	0.00416	0.00159	0.00061
Sq*	0.03639	0.02030	0.00975	0.00433	0.00184	0.00075	0.00030

North Magnetawan River near Burk's Falls, n=69

q	50	60	70	80	90	100	110
$\hat{P}q$	0.28209	0.12934	0.05624	0.02390	0.01006	0.00422	0.00176
$\hat{S}q$	0.04456	0.03076	0.01797	0.00958	0.00485	0.00237	0.00113
Pq*	0.28455	0.13370	0.05999	0.02644	0.01159	0.00507	0.00223
Sq*	0.04449	0.03045	0.01758	0.00924	0.00460	0.00221	0.00104

Pigeon River at Middle Falls, n=61

q	150	200	250	300	350	400	450
$\hat{P}q$	0.34506	0.13094	0.04547	0.01531	0.00510	0.00169	0.00056
$\hat{S}q$	0.05023	0.03303	0.01652	0.00725	0.00297	0.00117	0.00045
Pq*	0.34652	0.13598	0.04946	0.01759	0.00623	0.00221	0.00079
Sq*	0.05021	0.03264	0.01603	0.00688	0.00275	0.00105	0.00039

Skeena River at Usk, n=37

q	6000	7000	8000	9000	10000	11000	12000
$\hat{P}q$	0.22190	0.09089	0.03555	0.01365	0.00521	0.00198	0.00075
$\hat{S}q$	0.05553	0.03402	0.01775	0.00851	0.00389	0.00172	0.00075
Pq*	0.22850	0.09897	0.04141	0.01715	0.00709	0.00293	0.00122
Sq*	0.05514	0.03305	0.01675	0.00776	0.00340	0.00143	0.00059

Bulkley River at Quick, n=54

q	700	800	900	1000	1100	1200	1400
$\hat{P}q$	0.21293	0.10597	0.05105	0.02422	0.01140	0.00535	0.00117
$\hat{S}q$	0.04473	0.03054	0.01874	0.01079	0.00597	0.00328	0.00091
Pq*	0.21742	0.11140	0.05556	0.02740	0.01346	0.00661	0.00160
Sq*	0.04450	0.03005	0.01819	0.01031	0.00560	0.00296	0.00078

Harricana Riviere á Amos, n = 52

q	200	250	300	350	400	450	500
$\hat{P}q$	0.36860	0.11349	0.03107	0.00823	0.00216	0.00057	0.00015
$\hat{S}q$	0.05465	0.03180	0.01297	0.00456	0.00149	0.00047	0.00014
Pq*	0.36955	0.11913	0.03476	0.00993	0.00284	0.00081	0.00023
Sq*	0.05445	0.03130	0.01243	0.00423	0.00133	0.00040	0.00011

$$f(x/\mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma\,x} \exp\left[-(\log x - \mu)^2/2\sigma^2\right] \quad \ldots (4.15)$$

$$-\infty < \mu, \quad \log x < \infty, \quad \sigma > 0$$

where x is the value of the random variable, μ and σ^2 are the mean and variance of log x respectively. That is,

$$\log X \sim N(\mu, \sigma^2)$$

4.4.1 Maximum Likelihood Estimates

Given a random sample $\underline{x} = (x_1, x_2, \ldots x_n)$, the logarithm of the likelihood function of the pdf (4.15) is given by:

$$L = \log \ell = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log o^2 - \sum_{i=1}^{n}\log x_i - \sum_{i=1}^{n}\frac{(\log x_i - \mu)^2}{2\sigma^2} \quad \ldots (4.16)$$

Taking partial derivatives of L with respect to μ and σ and equating to zero, the well-known maximum likelihood estimators of μ and σ are obtained. This is given by:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\log x_i \quad \ldots (4.17)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\log x_i - \hat{\mu})^2}{n} \qquad \ldots (4.18)$$

The variance-covariance matrix of (μ, σ) is the inverse of the symmetric matrix

$$[I] = \begin{bmatrix} -\frac{\partial^2 L}{\partial \mu^2} & -\frac{\partial^2 L}{\partial \mu \, \partial \sigma} \\ -\frac{\partial^2 L}{\partial \sigma \, \partial \mu} & -\frac{\partial^2 L}{\partial \sigma^2} \end{bmatrix} \qquad \ldots (4.19)$$

That is,

$$\sigma_{ij} = [I]^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{bmatrix} \qquad \ldots (4.20)$$

The maximum likelihood estimator of the probability of exceedence $\hat{P}_q$ is given by:

$$\hat{P}_q = 1 - \int_{q}^{\infty} \frac{1}{\sqrt{2\pi}\,\hat{\sigma}\,x} \exp\left[-(\log x - \hat{\mu})^2 / 2\,\hat{\sigma}^2\right] dx \qquad \ldots (4.21)$$

$$= 1 - \Phi\left[\frac{\log q - \hat{\mu}}{\hat{\sigma}}\right] \qquad \ldots (4.22)$$

where $\Phi(z)$ is the standard normal cumulative distribution function evaluated at $z = (\log q - \hat{\mu})/\hat{\sigma}$. For the sake of computer calculation, an algorithm (26.2.17) from Abramowitz and Stegun (1970) is used to calculate (4.22).

From (4.4) and (4.20), the asymptotic standard-error of $\hat{P}_q$ is obtained, and where,

$$\frac{\partial P_q}{\partial \mu} = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} \exp\left[-(\log q - \hat{\mu})^2 / 2\hat{\sigma}^2 \right] = u_1 \quad \ldots (4.23)$$

$$\frac{\partial P_q}{\partial \sigma} = \frac{\log q - \hat{\mu}}{\hat{\sigma}} \cdot u_1 = u_2 \quad \ldots (4.24)$$

This leads to:

$$\hat{S}_q^2 = \frac{u_1^2\,\hat{\sigma}^2}{n} + \frac{u_2^2\,\hat{\sigma}^2}{2n} \quad \ldots (4.25)$$

4.4.2 BAYES ESTIMATES

The principles of Jeffreys (1961) described in Section 2.5 will be used to obtain the joint prior distribution of (μ, σ). This leads to:

$$v(\mu, \sigma) \propto 1/\hat{\sigma} \quad \ldots (4.26)$$

The joint posterior distribution of (μ, σ) is obtained by combining the prior with the likelihood function of (4.15) and applying Bayes' theorem as in (3.28),

Under the squared-error loss function, Bayes estimator of the probability of exceedence, P_q, is given by (4.8). For the lognormal distribution, the analytical solution to (4.8) is given by:

$$\tilde{P}_q = 1 - t\left[\sqrt{\frac{n}{n+1}} \cdot \frac{\log q - \hat{\mu}}{s};\ v\right] \qquad \ldots (4.27)$$

where $v = n - 1$, $s^2 = \dfrac{\sum_{i=1}^{n} (\log x_i - \hat{\mu})^2}{v}$, and

$t(z;\ v)$ is the cumulative distribution function of a t random variable with v degrees of freedom evaluated at z, (Martz and Waller, 1982; Stedinger, 1983).

In this section, a Bayes estimator of P_q will also be obtained by using Lindley's Bayesian approximation procedure and compared to (4.27), and the MLE $\hat{P}_q$ given by (4.22).

Lindley's expansion (3.31) for the 2-parameter lognormal distribution is given by:

$$E[u(\theta)/\underline{x}] = u + \frac{1}{2}(u_{11}\sigma_{11} + u_{22}\sigma_{22}) + \rho_1 u_1 \sigma_{11} + \rho_2 u_2 \sigma_{22}$$

$$+ \frac{1}{2}(L_{03} u_2 \sigma_{22}^2 + L_{21} u_2 \sigma_{11} \sigma_{22}) \qquad \ldots (4.28)$$

From (4.20) $\sigma_{12} = \sigma_{21} = 0$ and $L_{12} = L_{30} = 0$ (see Appendix F). Also $u(\theta) = P_q \equiv u$ in (4.21) and

$$u_1 = \frac{\partial u}{\partial \mu}, \quad u_2 = \frac{\partial u}{\partial \sigma}, \quad u_{11} = \frac{\partial^2 u}{\partial \mu^2}, \quad u_{22} = \frac{\partial^2 u}{\partial \sigma^2},$$

$$L_{03} = \frac{\partial^3 L}{\partial \sigma^3} = \frac{10n}{\sigma^2}, \quad L_{21} = \frac{\partial^3 L}{\partial \mu^2} = \frac{2n}{\sigma^3},$$

$$\sigma_{11} = \frac{\sigma^2}{n}, \quad \sigma_{22} = \frac{\sigma^2}{2n} \quad \text{from (4.20)}$$

$$\rho_1 = \frac{\partial \rho}{\partial \mu} = 0, \quad \rho_2 = \frac{\partial \rho}{\partial \sigma} = -\frac{1}{\sigma}$$

all evaluated at $(\hat{\mu}, \hat{\sigma})$. See Appendix F for the derivatives of u_{ij}'s and L_{ij}'s. Substituting in (4.28) one obtains P_q^*. Similarly, the posterior standard-error of P_q^* can be obtained letting $u(\theta) = P_q^2$, then,

$$S_q^* = \sqrt{E(P_q^2/\underline{x}) - P_q^{*2}} \qquad \ldots (4.29)$$

4.4.3 Numerical Example

The annual maximum flows of the Sturgeon River at Fort Saskatchewan (n = 50) is used as an example. The flood data for this river were found to be well fitted by the lognormal distribution. The maximum likelihood estimates for this river are:

$\hat{\mu} = 3.06423, \quad \hat{\sigma} = 0.739286, \quad \rho_2 = -1.352656$

$\hat{\sigma}_{11} = 0.010931, \quad \hat{\sigma}_{22} = 0.0029871$

$L_{21} = 247.492387, \quad L_{03} = 1237.461933$

Let q = 250 m^3/s, from (4.22) $\hat{P}_q = 0.00044408 = u$

$u_1 = 0.00215354, \quad u_2 = 0.00715791, \quad u_{11} = 0.0096822$

$u_{22} = 0.08760007$

Substituting into (4.25)

$\hat{S}_q = 0.00045138$

Bayes estimate of P_q^* by Lindley's expansion is obtained by substituting the evaluated constants into (4.28). This leads to:

$E(P_q/\underline{x}) = P_q^* = 0.00066735$

Also let $v = u(\theta) = P_q^2 = 0.000000197208$

$v_1 = 0.000001912694, \quad v_2 = 0.0000063574,$

$v_{11} = 0.0000178747, \quad v_{22} = 0.000180274$

Substituting into (4.28),

$E(P_q^2/\underline{x}) = 0.00000059925$

$Var(P_q/\underline{x}) = E(P_q^2/\underline{x}) - P_q^{*2}$

$$= 0.00000015389$$

Hence $S_q^* = 0.00039229 < \hat{S}_q = 0.00045138$.

P_q from the predictive distribution (4.27) is given by:

$$\tilde{P}_q = 1 - t[3.291;\ 49]$$

$$= 0.00090$$

Table 4.3 summarizes the results for q = 50 to 250 in steps of 50 m^3/s for the Sturgeon River at Fort Saskatchewan (n = 50). Figure 4.2 shows the results graphically.

TABLE 4.3

PROBABILITY OF EXCEEDENCE AND STANDARD-ERROR
STURGEON RIVER AT FORT SASKATCHEWAN
(n = 50)

q	50	100	150	200	250
$\hat{P}_q$	0.12574	0.018564	0.0042341	0.00012558	0.000444
$\hat{S}_q$	0.034082	0.009509	0.002999	0.0010969	0.0004514
P_q^*	0.12944	0.020825	0.0052497	0.0017192	0.000668
S_q^*	0.03388	0.009232	0.002822	0.000942	0.0003922
$\tilde{P}_q$	0.12900	0.02152	0.00576	0.00210	0.00090

MLE(^), Bayes(*), Predictive(~)

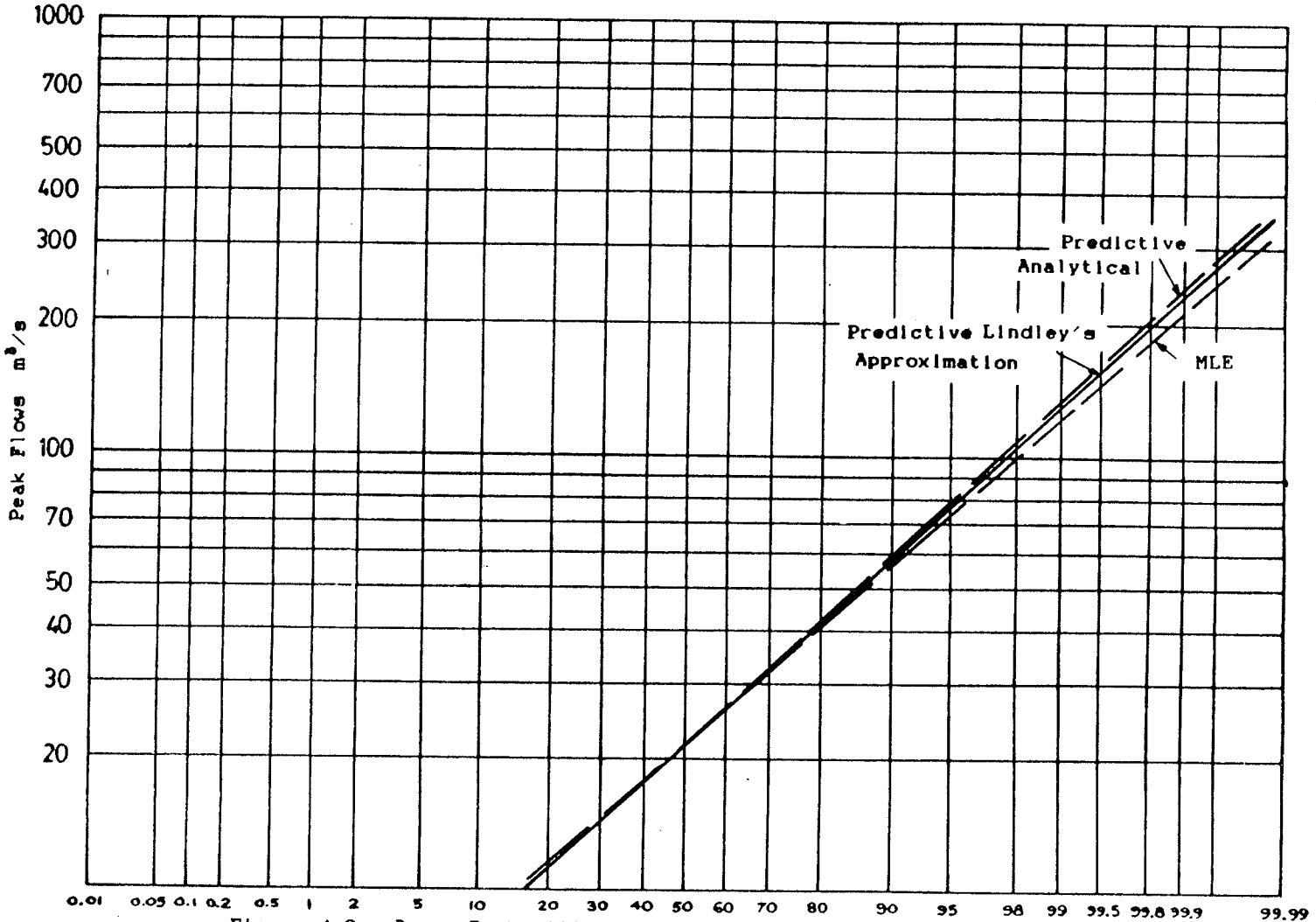


Figure 4.2 Bayes Probability of Exceedence (Lognormal Distribution) - Sturgeon River Near Fort Saskatchewan

The table above shows that the posterior standard-errors are smaller than the corresponding standard errors for the MLE's, and Bayes estimates of the probability of exceedence are higher, as expected. Furthermore, Bayes estimates of the probability of exceedence obtained by Lindley's expansion is a fairly good approximation to those estimates obtained analytically. Further comparison between using Bayesian approximation/Predictive distribution are given in Sinha (1986a).

The analysis was repeated for seven other maximum annual flow events of rivers from all over Canada that fit the lognormal distribution. The results are shown in Table 4.4. Only the MLE and Bayes estimates of P_q are compared.

In the next section, the Bayes estimate of the probability of exceedence by a discrete approach is discussed.

4.5 PREDICTIVE DISTRIBUTION BY DISCRETE APPROXIMATION

In this Section, the probability of exceedence of a future value of X is obtained by a discrete approach.

As noted in Section 4.1, Bayes' theorem for continuous probability models may often be difficult to apply because of potential problems in evaluating the

TABLE 4.4

BAYES (*) AND MLE (^) ESTIMATES OF THE PROBABILITY OF EXCEEDENCE OF A FLOOD DISCHARGE q (lognormal distribution) (q in m^3/s)

BOW RIVER AT BANFF, n = 76

q	250	300	400	500	550	600
$\hat{P}q$	0.27793	0.09764	0.00770	0.00049	0.00012	0.00003
$\hat{S}q$	0.03869	0.02019	0.00266	0.00024	0.00007	0.000019
Pq^*	0.27917	0.09821	0.00809	0.00055	0.00014	0.00004
Sq^*	0.03867	0.02013	0.00263	0.00023	0.00006	0.000018

RED RIVER AT EMERSON, n = 70

q	550	1000	2000	3000	3500	4000
$\hat{P}q$	0.48197	0.20325	0.04092	0.01153	0.00667	0.00402
$\hat{S}q$	0.04765	0.03701	0.01438	0.00570	0.00372	0.00248
Pq^*	0.48213	0.20569	0.04326	0.01288	0.00767	0.00477
Sq^*	0.04765	0.03693	0.01419	0.00554	0.00358	0.00237

ROSEWAY RIVER AT LOWER OHIO, n = 67

q	800	1000	1200	1600	2000	2200
$\hat{P}q$	0.27489	0.11058	0.04142	0.00553	0.00077	0.00030
$\hat{S}q$	0.04122	0.02413	0.01182	0.00230	0.00042	0.00018
Pq^*	0.27629	0.11232	0.04269	0.00593	0.00087	0.00035
Sq^*	0.04120	0.02407	0.01175	0.00226	0.00040	0.00017

ENGLISH RIVER AT UMFREVILLE, n = 63

q	150	200	300	500	700	800
$\hat{P}q$	0.42988	0.23172	0.06465	0.00611	0.00080	0.00032
$\hat{S}q$	0.04959	0.03977	0.01817	0.00295	0.00053	0.00024
Pq^*	0.43039	0.23346	0.06641	0.00671	0.00095	0.00040
Sq^*	0.04958	0.03973	0.01808	0.00289	0.00050	0.00022

SLOCAN RIVER NEAR CRESENT VALLEY, n = 60

q	450	500	600	700	800	1000
$\hat{P}q$	0.47409	0.32425	0.12878	0.04419	0.01393	0.00123
$\hat{S}q$	0.05140	0.04659	0.02776	0.01268	0.00497	0.00061
Pq^*	0.47429	0.32555	0.13073	0.04557	0.01465	0.00137
Sq^*	0.05140	0.04658	0.02768	0.01260	0.00492	0.00059

SOUTH THOMSON RIVER AT CHASE, n = 48

q	1000	1200	1500	1600	1800	2000
$\hat{P}q$	0.45038	0.17267	0.02583	0.01268	0.00285	0.00060
$\hat{S}q$	0.05715	0.03729	0.00907	0.00501	0.00137	0.00034
Pq^*	0.45087	0.17513	0.02710	0.01350	0.00313	0.00069
Sq^*	0.05715	0.03721	0.00896	0.00494	0.00134	0.00033

BABINE RIVER AT BABINE, n = 40

q	150	200	250	300	350	400
$\hat{P}q$	0.25265	0.05798	0.01147	0.00220	0.00043	0.00009
$\hat{S}q$	0.05109	0.01945	0.00533	0.00130	0.00030	0.00007
Pq^*	0.25517	0.06037	0.01248	0.00252	0.00052	0.00011
Sq^*	0.05103	0.01930	0.00523	0.00125	0.00029	0.00006

integral in the denominator of the equation. This difficulty does not arise if the prior distribution is a conjugate prior distribution. A conjugate prior distribution, say $g(\theta)$, for a given sampling distribution, say $f(x/\theta)$, is such that the posterior distribution $\Pi(\theta/x)$ and the prior $g(\theta)$ are members of the same family of distributions. A sampling distribution and its conjugate when combined using Bayes' theorem can be integrated without difficulty. Conjugate distributions are given in most standard texts such as Raiffa and Schlaiffer (1961), Box and Tiao (1973) and Ang and Tang (1975). However, with the possible exception of the normal distribution, these are too restrictive in form to be useful to a practising hydrologists (Russell, 1982).

In this study, the probabilistic model adopted takes either the annual flows themselves, their logarithms, or some other transformation of the flows to be normally distributed with parameters (μ, σ). The uncertainty in these parameters must be expressed in a joint probability distribution for μ and σ. This joint distribution is evidently continuous but for the sake of computer calculations, Russell (1982) adopted discrete approximations for the joint distribution of μ and σ. This essentially amounts to approximating the joint distribution by a discrete mass function and applying Bayes' theorem for discrete probability models. In addition to making computer analysis possible,

discretization has the advantage of avoiding mathematical complexities that add little to the understanding of the process. Arguments and derivations can thus be kept simple and straight forward (Booy and Morgan, 1985).

In general terms, then, the joint distribution of X when both the mean μ and standard deviation σ are uncertain is defined by:

$$X \sim N(\mu, \sigma) \qquad \ldots (4.30)$$

$$P(\mu = \mu_i, \sigma = \sigma_j) = P_{ij} \qquad \ldots (4.31)$$

for i = 1 to I, and j = 1 to J, and

$$\sum_{j}^{J} \sum_{i}^{I} P_{ij} = 1.0 \qquad \ldots (4.32)$$

In words: X is normally distributed with a mean μ and standard deviation σ. The range of values these parameters can have is represented by I.J parameter conditions or combinations (μ_i, σ_j), the probability of each condition being P_{ij}. Since these conditions are mutually exclusive and cumulatively exhaustive, their probability must add up to unity.

Equation (4.31) defines an array of conditional distributions which can represent either the joint prior or posterior distribution of (μ, σ).

Under the squared-error loss function, Bayes estimate of the probability of exceedence is given by:

$$P_q^* = E[P(X \geq q)/\underline{x}] \qquad \ldots (4.33)$$

$$= \sum_{j}^{J} \sum_{i}^{I} (P(X \geq q)/\mu_i, \sigma_j) \cdot P_{ij} \qquad \ldots (4.34)$$

$$= \tilde{P}_q \qquad \ldots (4.35)$$

where $\tilde{P}_q$ is the probability of exceedence of the predictive distribution of a future value of X. This relationship was proved earlier for the continuous case (see Section 4.2). Hence, (4.34) represents the probability that X is equal to or greater than a given value q, and the probability of this event is governed by a set of mutually exclusive and collectively exhaustive conditions (μ_i, σ_j). Further, the probability that the event occurs together with the condition is equal to the conditional probability of the event given that condition. Summing over all possible conditions results in the average probability of the event $(X \geq q)$ weighted by each probability condition.

4.5.1 Probability Distribution of μ and σ

In this study it will be assumed that the distributions of the parameters are entirely data-based and that there is very little prior information concerning the parameters (μ, σ) relative to the information contained in the sample. The only subjective element is the choice of the underlying distribution of the basic

random variable X, which is assumed to be normally distributed.

With a moderately sized sample ($n \geqslant 30$) and relatively little prior information, the posterior distribution of the parameters will depend almost solely on the sample information, as summarised by the likelihood function. This produces a posterior distribution quite similar to the distribution obtained by classical inference. The important difference, however, is in interpretation. The Bayesian statistician treats the parameters as random variables and is willing to make probability statements concerning the parameters, whereas, the classical statistician considers the parameters as fixed. The Bayesian approach thus allows the uncertainty in the parameters to be quantified and using Bayes' theorem permits updating the parameter distribution as additional information becomes available.

When little prior knowledge is assumed about μ relative to the information which would be supplied from the data, and a sample of n observations from a normal distribution with known variance σ^2 are given, then the posterior distribution of μ is also normally distributed (Box and Tiao, 1973). If the n observations are stochastically independent, the posterior distribution of μ is given by:

$$\Pi(\mu/\sigma, \underline{x}) \sim N(\bar{x}, \sigma^2/n) \qquad \ldots (4.36)$$

That is, the posterior distribution $\Pi(\mu/\sigma, \underline{x})$ is normally distributed with mean $\bar{x}$ and variance σ^2/n.

Similarly, given a sample $\underline{x}$ of n observations from a normal distribution $N(\mu, \sigma^2)$, with u known and with little prior information about σ relative to that supplied by the data, the posterior distribution of σ is approximately "chi-squared" (χ^2). However, for ($n \geqslant 30$), the χ^2 distribution is approximately normal (Benjamin and Cornell, 1970).

Hence in this study, both the mean and standard deviation are assumed to be approximately normally distributed. Consequently, only the mean of the means m_μ and the standard deviation of the means σ_μ and, the mean of the standard deviation m_σ, and the standard deviation of the standard deviation σ_σ are required to define the distributions of (μ, σ) respectively.

The mean of the means m_μ corresponds to the sample mean $\bar{x}$, and the mean of the standard deviation m_σ corresponds to the sample standard deviation s. The standard deviation of the mean σ_μ, and standard deviation of the standard deviation σ_σ are more difficult to determine. They depend very much on the serial correlation structure of the random variable. Estimates can be obtained by Monte Carlo techniques and the procedure is discussed in Chapter Seven.

4.5.2 Discretization And Assignment Of Probabilities

Having defined the distributions of the mean and standard deviation, the next step is to discretize and assign probabilities to the discrete values of the means and standard deviations. This allows the joint distributions to be defined.

The discrete approximation is accomplished by dividing the set of possible values of the parameter of interest into a number of intervals and determining the probability of each interval. This probability is then assumed to be a probability-mass located at the mid-point of the interval.

Eleven discrete values of each of the two parameters are chosen to represent the normal distribution. These values are as shown in Figure 4.3 for the distribution of the mean, with the initial probabilities assigned as the corresponding areas.

Thus for both the parameters, the eleven ordered values are 0.0113, 0.0280, 0.0660, 0.1210, 0.1750, 0.1974, 0.1750, 0.1210, 0.0660, 0.0280, and 0.0113 respectively. Since the mean and standard deviation of the normal distribution are independent, the joint probabilities are simply the product of the two individual probabilities, and in the eleven discrete value case, the corresponding joint probability distribution may be represented by a 11 x 11 matrix. The normalised probability matrix is shown in

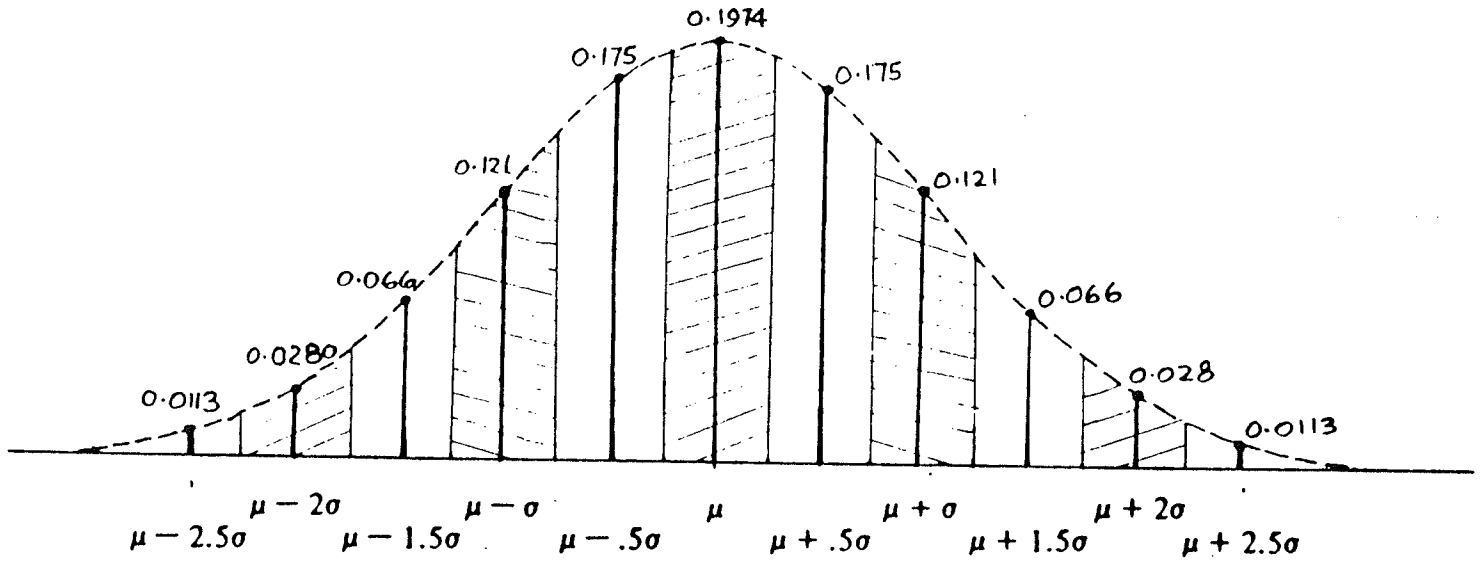


Figure 4.3 Discrete Representation of Normal Distribution

Figure 4.4.

Once the joint probabilities P_{ij} are established, the probability of exceedence of a future flood discharge q, P_q^* can be obtained from (4.34). Also if additional information becomes available which may be in the form of newly recorded flows, historical flows or regional information, then P_q^* can be updated using the new information. This updating is achieved in an indirect way by changing the likelihood of the parameter conditions, P_{ij}, in the light of the new information. This procedure explained in detail in Booy and Morgan (1985) is as follows:

Let the P_{ij} be the prior probabilities of the condition (μ_i, σ_j) for all values of i and j. This array is called the prior probability distribution of the parameters, since it was established prior to the availability of the new information. Let an event A be observed, and for this event all conditional probabilities $P(A/(\mu_i, \sigma_j))$ have been determined. $P(A/(\mu_i, \sigma_j))$ is the likelihood of A given the corresponding conditions (μ_i, σ_j). The observation of the event A changes the probability assessment of the condition (μ_i, σ_j) in that one is in a better position to revise one's degree of belief about (μ, σ). The revised probabilities of the conditions (μ_i, σ_j) after the observation of A are obtained by applying Bayes' theorem as follows:

Figure 4.4

Joint Probability of the Mean and Standard Deviation

PROBABILITY MATRIX

MEANS: / STD. DEV:	$m_\mu - 2.5\sigma_\mu$	$m_\mu - 2.0\sigma_\mu$	$m_\mu - 1.5\sigma_\mu$	$m_\mu - 1.0\sigma_\mu$	$m_\mu - 0.5\sigma_\mu$	m_μ	$m_\mu + 0.5\sigma_\mu$	$m_\mu + 1.0\sigma_\mu$	$m_\mu + 1.5\sigma_\mu$	$m_\mu + 2.0\sigma_\mu$	$m_\mu + 2.5\sigma_\mu$
$\mu_\sigma - 2.5\sigma_\sigma$	0.000128	0.000316	0.000746	0.001367	0.001978	0.002231	0.001978	0.001367	0.000746	0.000316	0.000218
$\mu_\sigma - 2.0\sigma_\sigma$	0.000316	0.000784	0.001848	0.003388	0.004900	0.005527	0.004900	0.003388	0.001848	0.000784	0.000316
$\mu_\sigma - 1.5\sigma_\sigma$	0.000746	0.001848	0.004356	0.007986	0.011550	0.013028	0.011550	0.007986	0.004356	0.001848	0.000746
$\mu_\sigma - 1.0\sigma_\sigma$	0.001367	0.003388	0.007986	0.014641	0.021175	0.023885	0.021175	0.014641	0.007986	0.003388	0.001367
$\mu_\sigma - 0.5\sigma_\sigma$	0.001978	0.004900	0.011550	0.021175	0.030625	0.034545	0.030625	0.021175	0.011550	0.004900	0.001978
μ_σ	0.002231	0.005527	0.013028	0.023885	0.034545	0.038967	0.034545	0.023885	0.013028	0.005527	0.002231
$\mu_\sigma + 0.5\sigma_\sigma$	0.001978	0.004900	0.011550	0.021175	0.030625	0.034545	0.030625	0.021175	0.011550	0.004900	0.001978
$\mu_\sigma + 1.0\sigma_\sigma$	0.001367	0.003388	0.007986	0.014641	0.021175	0.023885	0.021175	0.014641	0.007986	0.003388	0.001367
$\mu_\sigma + 1.5\sigma_\sigma$	0.000746	0.001848	0.004356	0.007986	0.011550	0.013028	0.011550	0.007986	0.004356	0.001848	0.000746
$\mu_\sigma + 2.0\sigma_\sigma$	0.000316	0.000784	0.001848	0.003388	0.004900	0.005527	0.004900	0.003388	0.001848	0.000784	0.000316
$\mu_\sigma + 2.5\sigma_\sigma$	0.000128	0.000316	0.000746	0.001367	0.001978	0.002231	0.001978	0.001367	0.000746	0.000316	0.000128

$$P((\mu_i, \sigma_j)/A) = C \cdot P(A/(\mu_i, \sigma_j)) \cdot P_{ij} \quad \ldots (4.37)$$

for $i = 1$ to I and $j = 1$ to J.

where $P((\mu_i, \sigma_j)/A)$ is the posterior or updated probability distribution of the parameters (μ_i, σ_j) conditioned on the observation of A; each value of $P((\mu_i, \sigma_j)/A)$ being calculated as the product of the likelihood of A, $P(A/(\mu_i, \sigma_j))$ given the corresponding condition (μ_i, σ_j) and the prior probability of that condition, P_{ij}. The normalizing factor C ensures that the sum of the posterior probabilities for all conditions add up to unity.

The updated probability of exceedence of a future flood discharge q, P_q^* can now be obtained from the updated parameter array by means of the total probability theorem in the form of (4.34).

If more than one event is available for updating, the procedure can be repeated for each event, using the posterior parameter distribution after the first updating as the prior for the second event, etc.

4.6 SUMMARY

In this chapter, the predictive distribution approach to flood risk analysis was described. The predictive distribution combines the stochastic uncertainty that is inherent in the probabilistic phenomenon and the parameter uncertainty of the model it defines. It was shown that the Bayes estimate of the probability of exceedence (under a squared-error loss function) is the probability of exceedence of the predictive distribution. Using this property, Bayes estimates of the probability of exceedence were calculated using Lindley's expansion which greatly simplifies the computation. It was shown that Bayes estimates give a smaller standard error of estimate than the method of maximum likelihood.

A discrete approximation approach to calculating Bayes estimate of the probability of exceedence was also described. This approach is entirely data-based. It can be used for serially correlated or serially independent data. Also the distributions can be updated with different types of information.

In the following chapter, the serial correlation structure of annual peak flows of Canadian rivers are analysed.

CHAPTER 5

SERIAL CORRELATION STRUCTURE OF ANNUAL PEAK FLOWS OF CANADIAN RIVERS

5.1 GENERAL

In previous chapters, the customary assumption that annual peak flow series are independent was used. For the vast majority of rivers this assumption appears to be supported by statistical evidence.

The point that is overlooked, however, is that most statistical tests of serial independence are designed to show up only short term serial correlation. They are insensitive to the long term serial correlation structure of the data which is generally far more important (Wallis and Matalas, 1971).

To demonstrate this issue, the serial correlation structure of the recorded or reconstructed natural annual peak flow series of 49 rivers from all over Canada was analysed. The years of record range from about 40 to 90 years. The average length of record is 60 years. The rivers are listed in Table 5.1.

The results show that while short term serial correlation is practically absent for most of the peak flow series, significant long term serial correlation is present for a large number of peak flow series tested.

Table 5.1 Hurst's K and lag-one serial correlation of natural annual peak flows of some Canadian Rivers.

No.	River	Province	Year of record	R/S	K	ρ1
1	ELBOW RIVER AT BRAGG CREEK	ALBERTA	1935 - 1984	9.46	.70	.00
2	RED DEER RIVER AT RED DEER	ALBERTA	1935 - 1980	7.68	.65	.30*
3	ELBOW RIVER ABOVE GLENMORE DAM	ALBERTA	1934 - 1977	7.71	.66	.03
4	BOW RIVER AT BANFF	ALBERTA	1909 - 1984	10.99	.66	-.08
5	ST. JOHN RIVER AT POKIOK	ATLANTIC	1919 - 1967	7.74	.64	.05
6	ST. MARY RIVER AT STILLWATER	ATLANTIC	1916 - 1984	13.84	.74	.04
7	ROSEWAY RIVER AT LOWER OHIO	ATLANTIC	1918 - 1984	13.46	.74	.08
8	NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY	ATLANTIC	1917 - 1984	16.94	.80	.36*
9	LAHAVE RIVER AT WEST NORTHFIELD	ATLANTIC	1916 - 1984	13.00	.72	.00
10	ST. JOHN RIVER AT FORT KENT	ATLANTIC	1927 - 1984	12.31	.75	.22*
11	LEPREAU RIVER AT LEPREAU	ATLANTIC	1917 - 1984	9.48	.64	-.03
12	ST. JOHN RIVER AT GRANDFALLS	ATLANTIC	1931 - 1984	9.91	.70	.13
13	SOUTHWEST MARGAREE RIVER NEAR UPPER MARGAREE	ATLANTIC	1919 - 1984	17.71	.82	.21*
14	BABINE RIVER AT BABINE	B. COLUMBIA	1945 - 1984	7.00	.65	.07
15	SOUTH THOMSON RIVER NEAR CHASE	B. COLUMBIA	1911 - 1958	9.12	.70	.17
16	SLOCAN RIVER NEAR CRESENT VALLEY	B. COLUMBIA	1925 - 1984	12.06	.73	.14
17	SKEENA RIVER AT USK	B. COLUMBIA	1948 - 1984	6.23	.63	-.28
18	MOYIE RIVER AT EASTPORT	B. COLUMBIA	1930 - 1984	15.74	.83	.08
19	KETTLE RIVER NEAR LAURIER	B. COLUMBIA	1930 - 1984	11.66	.74	.15
20	KETTLE RIVER NEAR FERRY	B. COLUMBIA	1926 - 1984	11.62	.74	.13
21	BULKLEY RIVER AT QUICK	B. COLUMBIA	1931 - 1984	8.10	.64	.21
22	BOUNDARY CREEK NEAR PORTHILL	B. COLUMBIA	1928 - 1984	15.64	.82	.36*
23	FRASER RIVER AT HOPE	B. COLUMBIA	1912 - 1951	9.45	.75	.21
24	ATHABASCA RIVER AT ATHABASCA	B. COLUMBIA	1942 - 1984	5.70	.57	-.20
25	RED RIVER AT EMERSON	MANITOBA	1915 - 1985	14.44	.75	.17
26	ROSEAU RIVER NEAR CARIBOU	MANITOBA	1920 - 1984	9.80	.66	.20
27	STURGEON RIVER NEAR FORT SASKATCHEWAN	MANITOBA	1935 - 1984	7.47	.63	-.14
28	ROSEAU RIVER NEAR DOMINION CITY	MANITOBA	1940 - 1984	5.80	.57	.03
29	RED RIVER AT REDWOOD BRIDGE (NO BREAKS)	MANITOBA	1893 - 1985	18.65	.76	.11
30	ASSINIBOINE RIVER AT BRANDON	MANITOBA	1902 - 1985	10.80	.63	.17
31	ASSINIBOINE RIVER AT HEADINGLEY	MANITOBA	1913 - 1985	11.00	.67	-.04
32	PEMBINA RIVER AT NECHE (INTERNATIONAL)	MANITOBA	1904 - 1984	13.83	.72	.05
33	TURTLE RIVER NEAR MINE CENTRE	ONTARIO	1921 - 1978	10.85	.71	.08
34	SAUGEEN RIVER NEAR WALKERTON	ONTARIO	1915 - 1984	10.37	.66	.11
35	SAUGEEN RIVER NEAR PORT ELGIN	ONTARIO	1915 - 1984	12.04	.70	.05
36	PIGEON RIVER AT MIDDLE FALLS	ONTARIO	1924 - 1984	10.63	.69	.01
37	NORTH MAGNETAWAN RIVER NEAR BURK'S FALLS	ONTARIO	1916 - 1984	6.85	.54	-.04
38	NAMAKAN RIVER AT OUTLET OF LAC LA CROIX	ONTARIO	1923 - 1984	12.90	.75	.22*
39	MISSINAIBI RIVER AT MATTICE	ONTARIO	1920 - 1984	13.26	.74	.17
40	ENGLISH RIVER NEAR SIOUX LOOKOUT	ONTARIO	1922 - 1981	13.51	.77	.12
41	ENGLISH RIVER AT UNFREVILLE	ONTARIO	1922 - 1984	14.01	.77	.05
42	BLACK RIVER NEAR WASHAGO	ONTARIO	1916 - 1984	13.57	.74	.13
43	BEAURIVAGE RIVIERE A SAINTE-ETIENNE	QUEBEC	1926 - 1984	12.45	.75	.16
44	HARRICANA RIVIERE A AMOS	QUEBEC	1933 - 1984	7.37	.61	-.15
45	RICHELIEU RIVIERE AUX RAPIDES FRYERS	QUEBEC	1938 - 1984	8.45	.68	.14
46	POPLAR RIVER AT INTERNATIONAL BOUNDARY	SASKATCHEWAN	1933 - 1984	5.46	.52	-.23
47	MCEACHERN CREEK AT INTERNATIONAL BOUNDARY	SASKATCHEWAN	1924 - 1976	7.44	.61	.01
48	HORSE CREEK AT INTERNATIONAL BOUNDARY	SASKATCHEWAN	1919 - 1961	8.11	.68	.07
49	SOUTH SASKATCHEWAN RIVER AT SASKATOON	SASKATCHEWAN	1912 - 1968	8.38	.63	.13
			MEAN:		.70	.08

5.2 STATISTICAL TESTS OF INDEPENDENCE

Each of the flood peak series was subjected to the following parametric and non-parametric tests of independence.

a) Anderson test for first order serial correlation coefficient.

b) Median crossing test.

c) Rank difference test.

d) Spearman rank order correlation test.

e) Wald-Wolfowitz test.

f) Cumulative periodogram test.

g) Turning point test.

h) Rescaled-range or Hurst coefficient.

Tests (a) to (g) are the commonly used tests of independence.

These tests are well documented in most statistical texts on time series analysis such as Kendall and Stuart (1973).

For the 49 rivers tested, tests (a) - (g) gave similar results, and confirms the general opinion that on the whole, any short term serial correlation is small. An average value of 0.08 for the lag-one serial correlation ceofficient was obtained. Six of the 49 series showed a significant lag-one serial correlation at the 10% level using Anderson's test (Anderson, 1942).

The situation, however, is quite different for the long term serial correlation as measured by the Hurst coefficient.

The long term serial correlation structure of geophysical time series was first investigated by Hurst (1951). The strength of this type of correlation is measured by means of the Hurst coefficient, h. A more detailed discussion of this parameter is given in Appendix G. It is sufficient to mention here that the Hurst coefficient is estimated by Hurst's K value in which K is theoretically 0.5 for series of independent data and it increases when there is a greater degree of persistence and that it cannot exceed 1.0. A comparison between long term serially correlated series and serially independent series is presented in the next section.

5.3 LONG TERM SERIAL CORRELATION

For the purpose of comparison between long term serially correlated series and serially independent series, Figure 5.1 is shown first. It shows a theoretical time series with no serial correlation. The graph shows, nevertheless, periods of relatively high and periods of relatively low values of the variable. This chance grouping may mask, or indeed destroy, any grouping that might be caused by true long term serial correlation. It

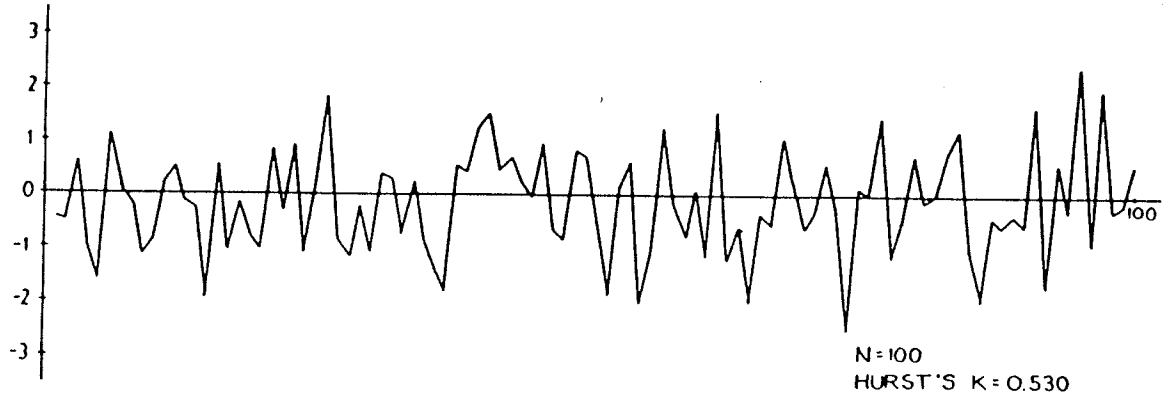


Figure 5.1 Sample function of independent data (N=100)

is therefore not sufficient to examine a time series visually to assess its serial correlation structure; a proper statistical analysis is essential. Yet, even a visual inspection can be useful in some respects as can be seen from the next series of figures.

Figures 5.2a - 5.2e illustrate a few of the observed annual peak flow series that were studied. The series are all approximately log-normally distributed. To make them comparable, the standardized log-transform is shown for each series as well as the flow in cubic metres per second. These series demonstrate to a greater or lesser degree a characteristic grouping of high and low peak flows in periods of irregular length. The degree of grouping is measured by the Hurst's K-value, which is indicated for each series. The question now arises: Is this grouping statistically significant, or is it due to chance?

To answer this question, the Hurst's K-values were calculated for all 49 peak flow series. They are shown on Table 5.1. To analyse them, a histogram of the 49 observed K-values was plotted and approximated by a normal probability density function. This is shown in Figure 5.3. The calculations were then repeated for an equal number of independent series of the same length that were obtained by theoretical simulation. The probability density function of K obtained from the generated independent series is also plotted on Figure 5.3.

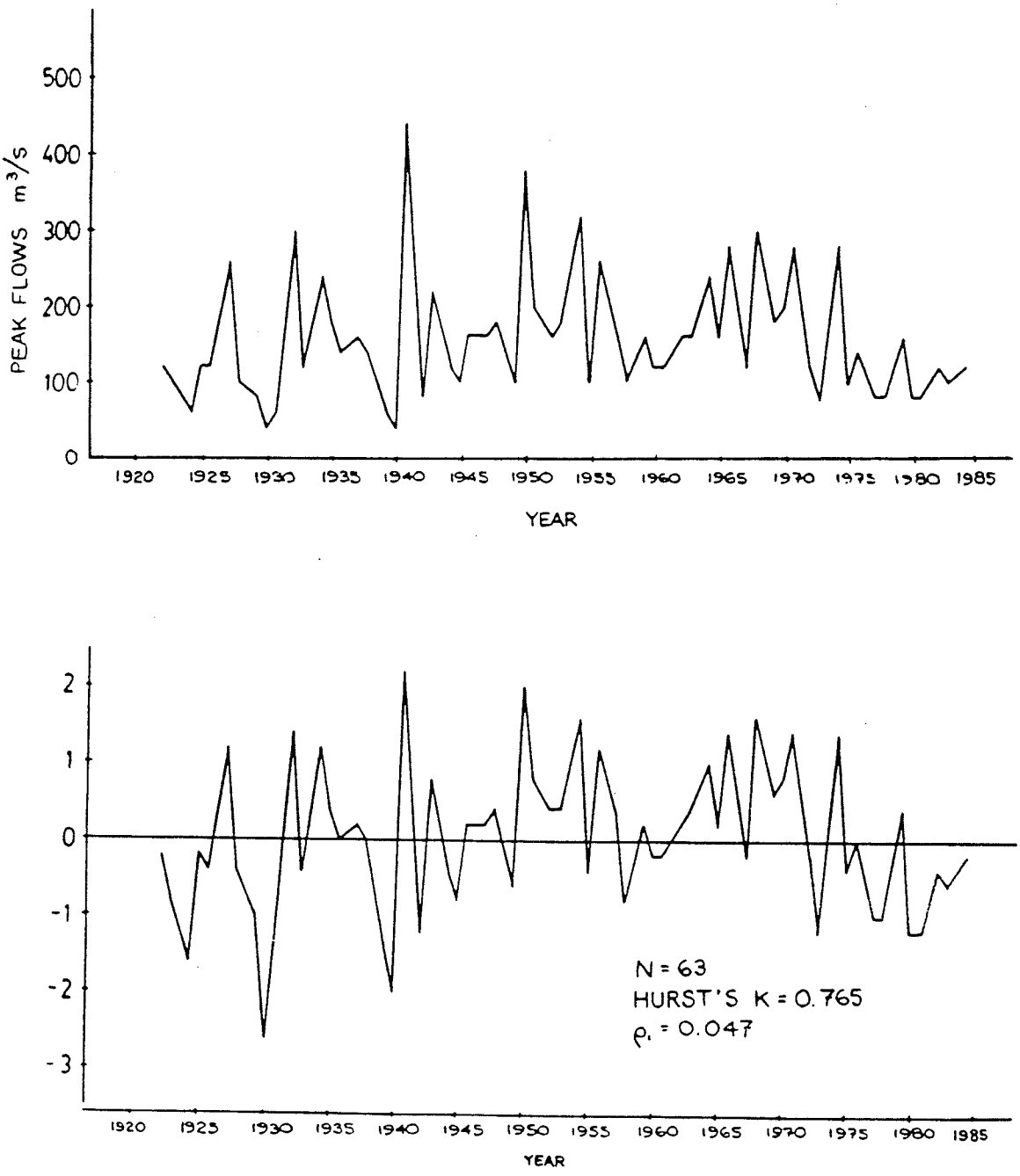


Figure 5.2a Peak Flow Time Series - English River At Umfreville

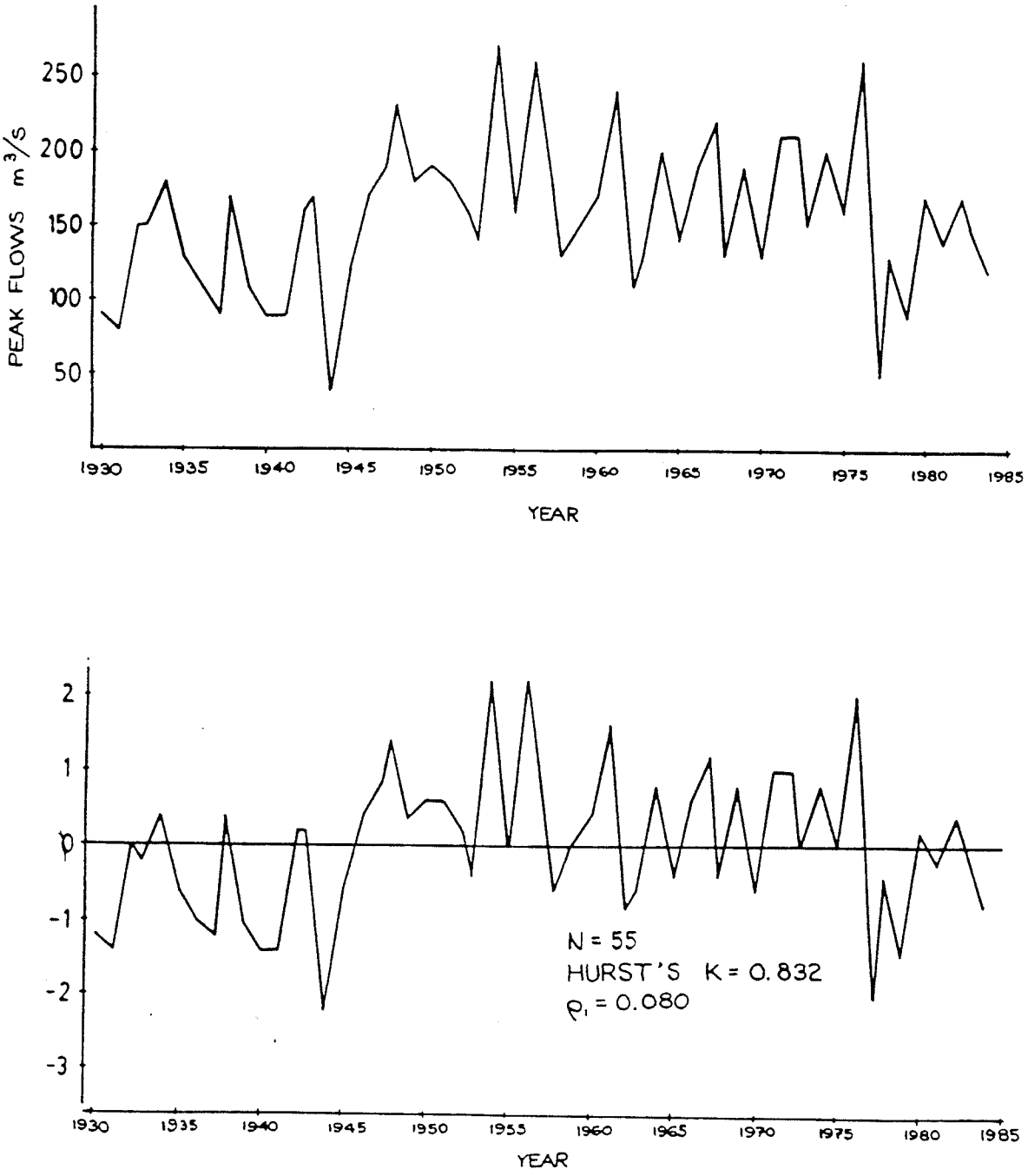


Figure 5.2b Peak Flow Time Series - Moyle River at Eastport

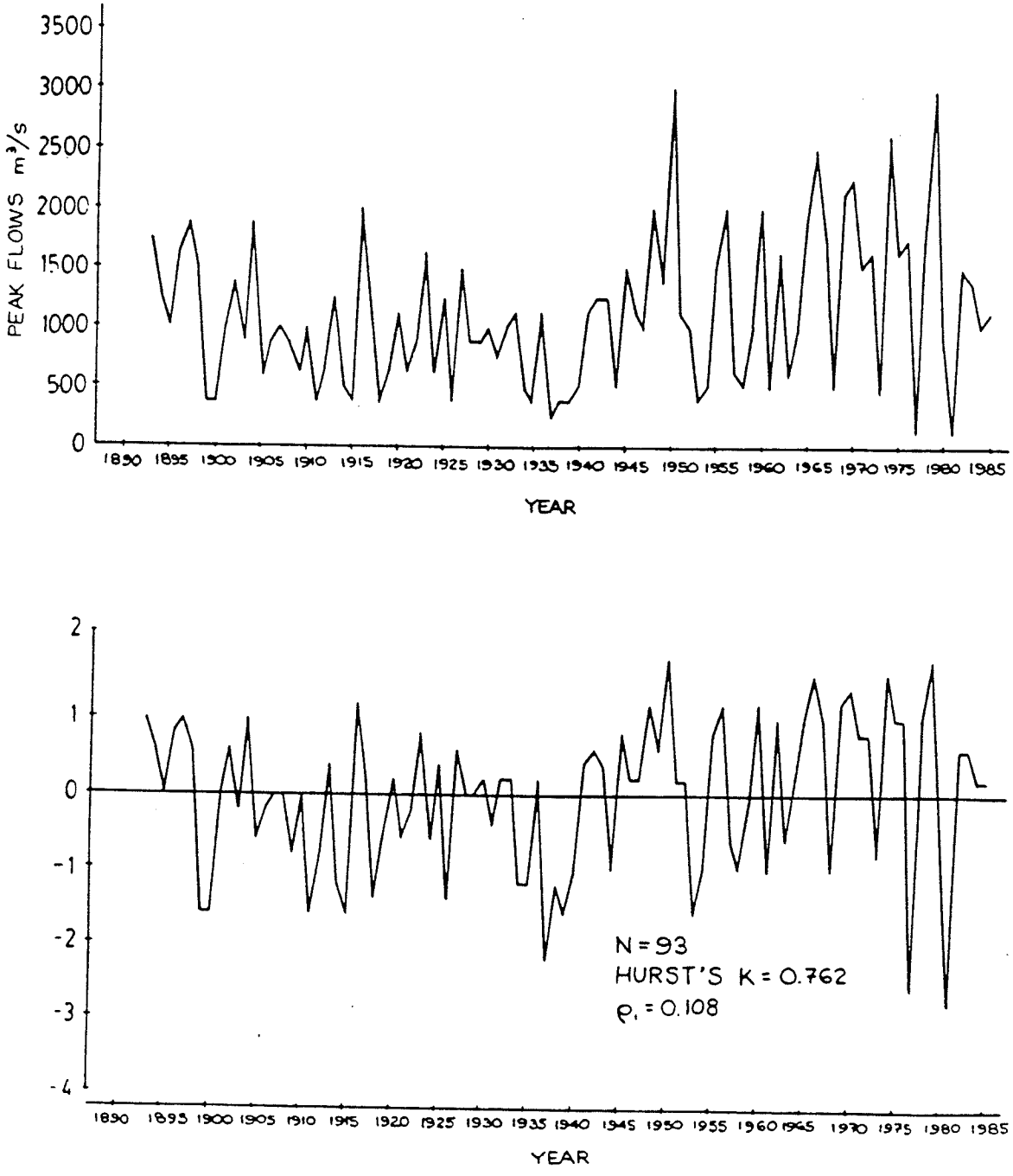


Figure 5.2c Peak Flow Time Series - Red River at Redwood Bridge

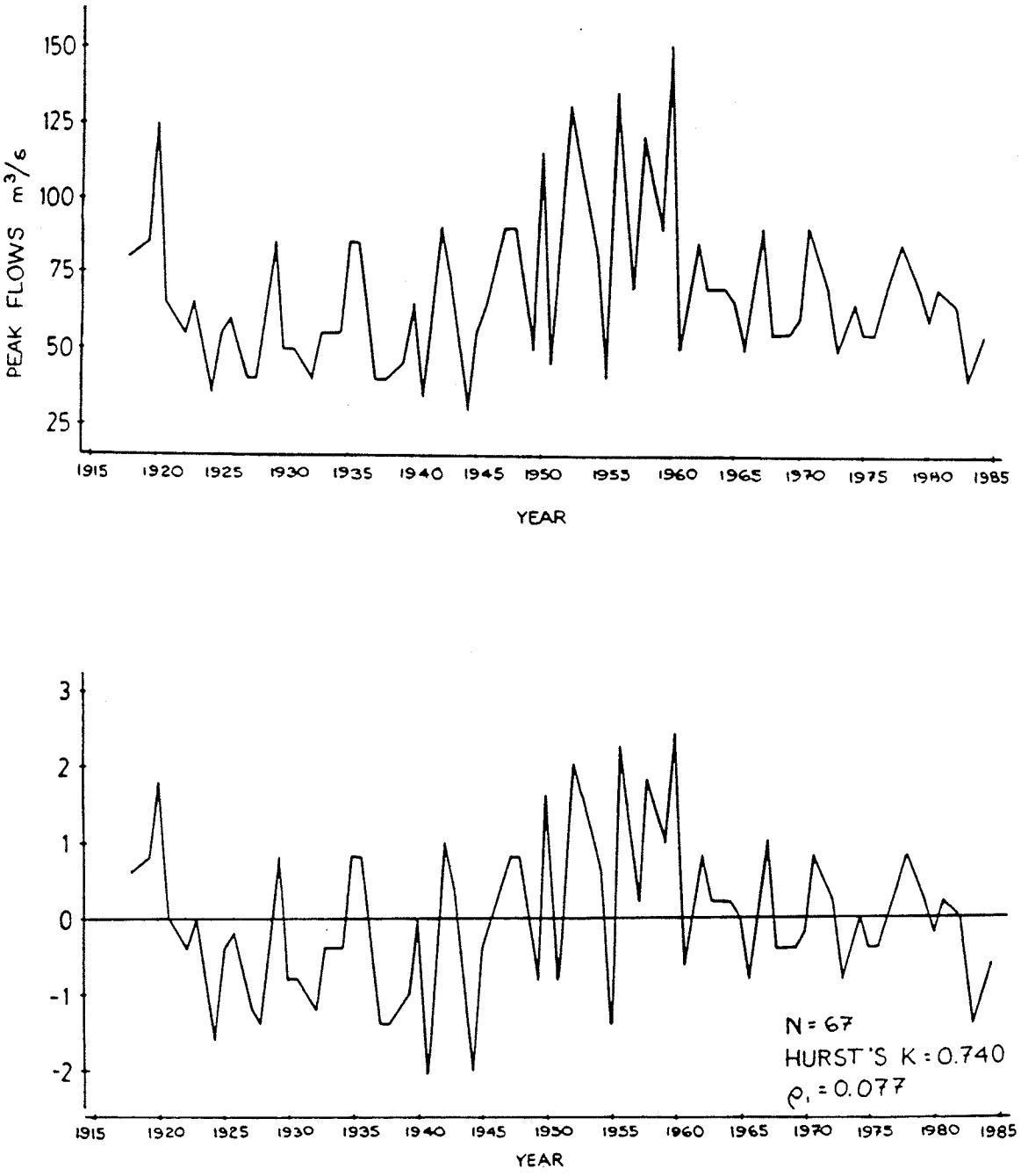


Figure 5.2d Peak Flow Time Series - Roseway River at Lower Ohio

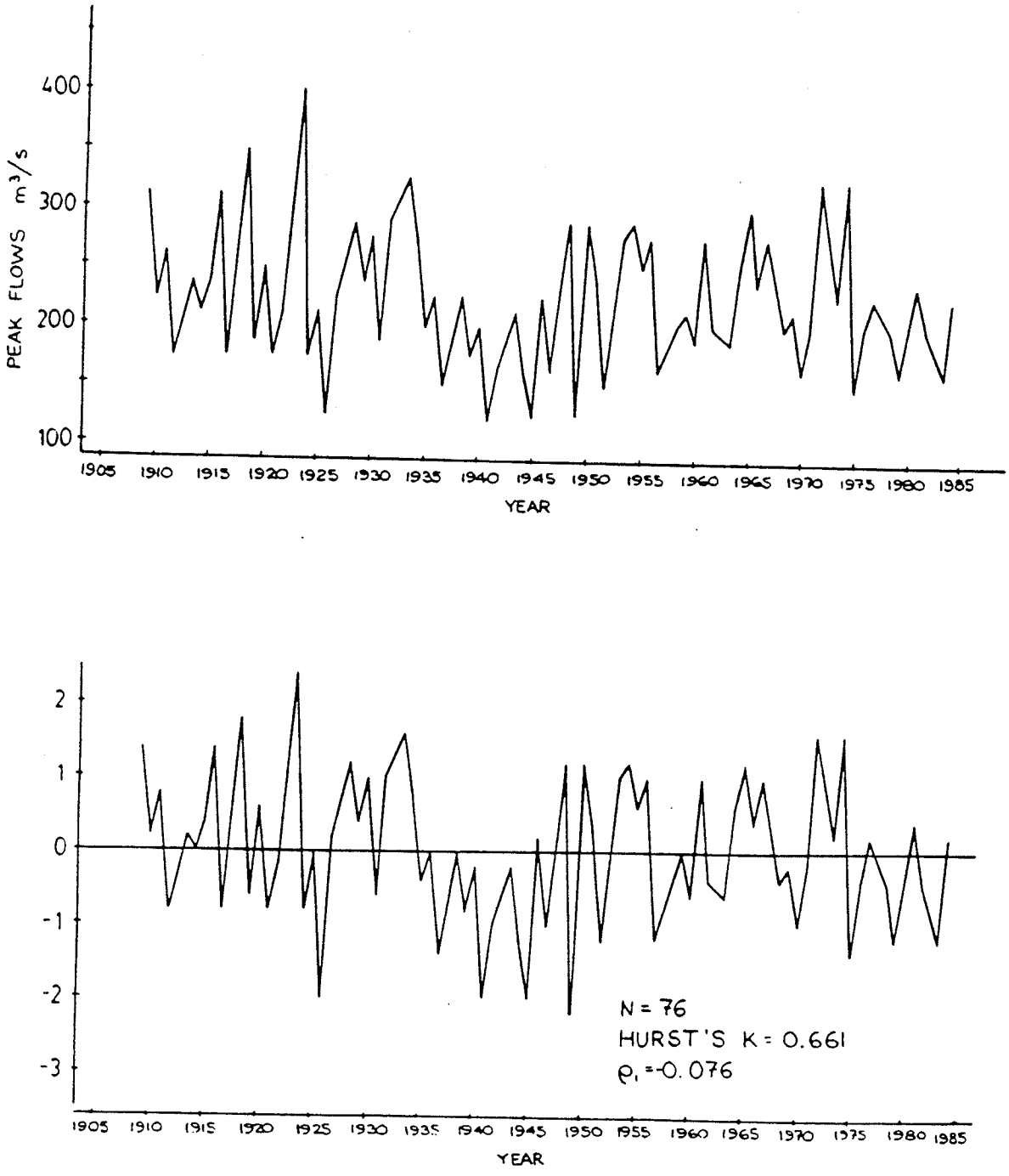


Figure 5.2e Peak Flow Time Series - Bow River at Banff

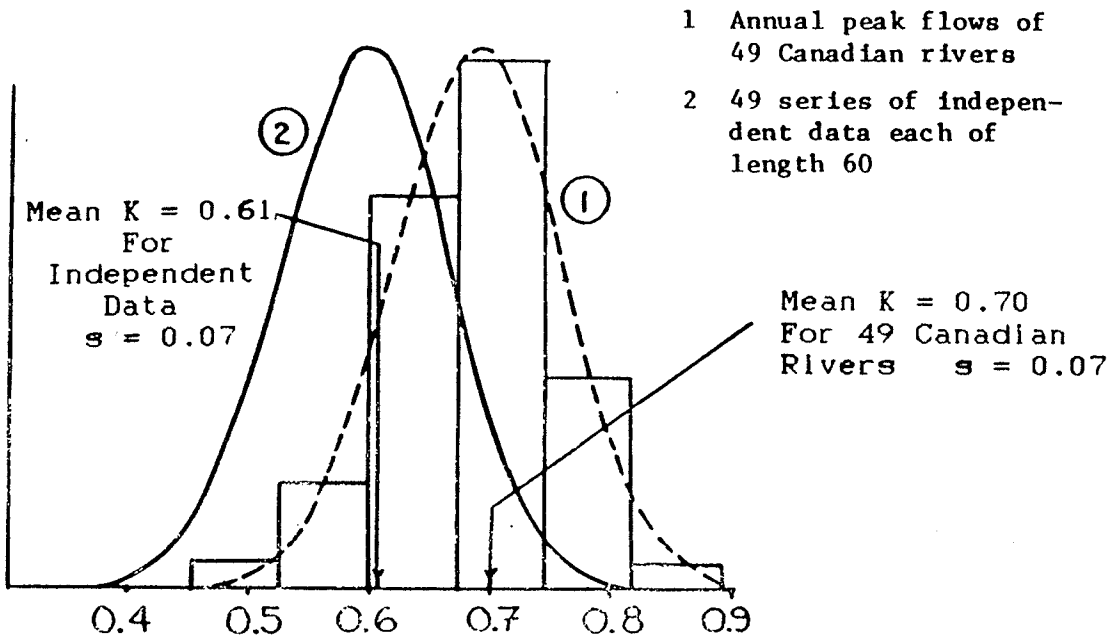


Figure 5.3 Distribution of Hurst's K

A comparison of the two probability density functions shows that the standard deviation of the K-values, which is partly caused by chance groupings, is the same and equal to 0.07. But the mean value of K is significantly greater for the observed peak flow series than for the independent data.

The probability that K = 0.70 is equalled or exceeded, based on the alternate assumption of independence, is about 10%. For the 49 Canadian rivers tested, about half the rivers showed a Hurst's K of 0.70 or greater. In fact, about 20% of the rivers showed a Hurst's K to be as large or larger than 0.75. The probability that K = 0.75 is equalled or exceeded, based on the assumption of independence, is only about 2%.

Hence, it is clear that long term serial correlation is as much a reality for the peak flow series of the Canadian rivers as it is for the many geophysical series Hurst found.

Unfortunately, in the extensive discussions of the Hurst phenomenon, much attention was focussed on the presumed "infinite memory" supposedly implied by a K-value that is significantly and persistently larger than 0.5. The grouping of high and low values does not reflect some mysterious memory of what happened in a distant past. It is simply a form of variability which one encounters in natural time series, and which is of a different nature than the variability one finds in controlled repeatable

experiments. The grouping reflects a greater variability of the same mean and standard deviation compared to a sequence of serially independent data.

Much has been written to explain the Hurst phenomenon or to explain it away. However, what remains is Hurst's monumental achievement which exposed the long-term correlation structure of a vast variety of geophysical time series and his legacy of a useful parameter that serves as an objective measure of long term serial correlation. Denying or ignoring the existence of the Hurst phenomenon as a characteristic of natural variability reflects pre-conceived notions that variability does not stem from observation but from the fact that statistical models do not fit the actual situation.

5.4 SUMMARY

In this chapter, the serial correlation structure of annual flood series from all over Canada was analysed. It was found that significant long term serial correlation as measured by the Hurst coefficient is present in a large number of rivers.

In the next chapter, the modelling of long term serial correlation is discussed. This will be followed by

the implications of long term serial correlation on flood risk analysis.

CHAPTER 6

TIME SERIES MODELS OF PEAK FLOWS

6.1 GENERAL

In the previous chapter, it was shown that annual peak flows while having negligible short term serial correlation often possess a significant long term serial correlation structure. Long term serial correlation is known to substantially increase the uncertainty in parameter estimation from the flood data. The effect of this uncertainty on flood risk will be demonstrated in Chapter 7 through the use of the discretized predictive distribution approach as described in Section 4.5. Before this can be carried out, however, one would require the probability distribution of these parameters. For a normally distributed time series with members that are independent or that follow a simple correlation structure, the distribution of the parameters (μ, σ) can be obtained analytically. For flood series with a complicated serial correlation structure, an analytical approach is out of the question. One must resort to Monte Carlo techniques.

To obtain the distribution of the parameters one requires a theoretical time series model that will reproduce the correct correlation structure of the peak flow series. A number of such time series models which can model series characterized by both a small first order

serial correlation coefficient and a high Hurst coefficient are currently available. These models are: the Fast Fractional Gaussian Noise Model (FFGN); the Broken-Line Model (BL); the ARMA (1,1) Model; and the ARMA-MARKOV (AM) Model. It will be shown, however, that these models are either too difficult to use by practising engineers or that they require excessive computer time. An added difficulty is that small sample biases in these models are not adequately documented. An efficient model that is simple to use is therefore needed.

In this chapter, the procedure for modelling a hydrological time series, including the correction for bias of parameter estimation, is first presented. This is followed by a review of the currently available models that can reproduce the correct short and long term serial correlation. Finally, a simple and relatively efficient time series model that is capable of reproducing short and long term serial correlation as well as the relevant marginal distribution properties, i.e., the mean and variance, is presented here.

6.2 MODELLING HYDROLOGIC TIME SERIES

To model a hydrological time series one must estimate a set of statistics, θ_i, i = 1, 2, . . . m, which form the model parameters. In the context of this study, these are the mean, variance, first order serial correlation coefficient and the Hurst coefficient. Each parameter must be estimated from the observed data. The model is then used to transform K sets of random numbers into K synthetic sequences, which provide equally probable random examples of the manner in which the time series may evolve in the future. The validity of the model is generally assessed by comparing the statistics of the synthetic data sequences, $\tilde{\theta}_i$, with the historical statistics θ_i. One must distinguish between statistical resemblance in the long run (for sequences whose length approaches infinity) and in the short run (for sequences that are about the same length as the planning horizon). According to Matalas (1977), statistical resemblance in the long run is achieved if the $\tilde{\theta}_i$ approach the θ_i as the length of the sequences increases; statistical resemblance in the short run is obtained when the θ_i^*, the averages of the $\tilde{\theta}_i$ over the K synthetic sequences, approach the θ_i as K increases.

Matalas and Wallis (1976) and Matalas (1977) have noted that, although short run resemblance implies long run resemblance, the converse does not necessarily hold

since the $\tilde{\theta}_i$ may be biased estimates of the θ_i. To obtain short run resemblance, the estimated values from the historic series must be corrected for bias before being used in the generation model. Of the various statistics that are employed as model parameters, only the mean is statistically unbiased; the others are not. The correction for bias in all other parameters depends upon:

(i) the length of synthetic sequences to be generated;

(ii) the time series model employed; and

(iii) the distribution function used to generate the random inputs to the model.

In the presence of serial correlation, expressions for bias corrections are difficult to obtain analytically, and must often be evaluated by Monte Carlo methods. Unfortunately, the problem of bias has attracted relatively little attention. A complete treatment of the problem of bias is beyond the scope of this study. The discussion presented in the next section illustrates how the bias problem was addressed.

6.3 BIAS CORRECTION IN PARAMETER ESTIMATION

Although the performance of a time series model should not be judged solely on its ability to preserve the parameters explicitly built into its structure, such preservation is a necessary criterion of the formal correctness of the model (Klemes, 1972). Nevertheless, as noted in the previous section, short run resemblance may not be achieved due to the bias in the estimation of the parameters. It will be shown in this section that the bias in the estimation of the parameters may be substantial. Some results given, which relate to the standard deviation, first order serial correlation coefficient and Hurst coefficient will serve to illustrate the order of magnitude of the necessary corrections for bias.

One of the few studies of bias in estimates of the standard deviation was presented by Wallis and Matalas (1972). They derived bias corrections for sample sizes of 100 for a lag one Markov model [AR(1)] using Monte Carlo methods. The results are shown in Table 6.1. They apply only to AR(1) models with normally distributed random terms. The use of the correction factors is illustrated by an example.

TABLE 6.1

CORRECTION FACTORS, α, FOR THE STANDARD DEVIATION OF AN AR(1) PROCESS WITH A SAMPLE SIZE OF 100 (after Wallis and Matalas, 1972)

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
α	1.0	0.99	0.99	0.99	0.98	0.98	0.97	0.95	0.94	0.90

The expected values of the first serial correlation coefficient, $E[\rho(1)]$, given ρ for an AR(1) model are also obtained by Wallis and Matalas (1972), for sample sizes of 100. These results are shown in Table 6.2.

TABLE 6.2

EXPECTED VALUES OF THE FIRST SERIAL CORRELATION $E[\rho(1)]$, FOR AN AR(1) PROCESS WITH A GIVEN VALUE OF ρ AND A SAMPLE SIZE OF 100 (after Wallis and Matalas, 1972)

ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$E[\rho(1)]$	0.09	0.18	0.28	0.38	0.47	0.57	0.67	0.76	0.86

The use of Tables 6.1 and 6.2 may be illustrated by means of the following example quoted by Matalas and Wallis (1976). Given that realizations are to be generated for an AR(1) model with $\hat{s}$ = 10 units and ρ = 0.8, attention is paid firstly to the latter. The value of the first serial correlation coefficient to be used in the model in order to ensure that $E[\rho(1)]$ = 0.8 may be found from Table 6.2 to be 0.84. Using this value for σ, Table 6.1 yields a correction factor for the standard deviation of $\propto$ = 0.92. The unbiased estimate of the standard deviation is therefore equal to 10/0.92 = 10.87. Synthetic flows would then be generated by using in the generation process the values of 10.87 and 0.84 for the standard deviation and first serial correlation coefficient respectively.

Various algorithms for estimating the first serial correlation coefficient were investigated by Wallis and O'Connell (1972). They found that for an AR(1) model with normally distributed random term, the algorithm suggested by Jenkins and Watts (1968) and Box and Jenkins (1970) gives a simple and satisfactory estimate. The general algorithm is given by:

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2} \quad \ldots (6.1)$$

in which the most commonly used measure of short term serial correlation is the first serial correlation coefficient $\hat{\rho}(1)$.

However, the estimate $\hat{\rho}(1)$, obtained through (6.1) is biased downwards, and Wallis and O'Connell (1972) suggested that the computed value of o should be replaced by:

$$\rho(1) = \left[\hat{\rho}(1) + \frac{1}{n} \right] / \left[1 - \frac{4}{n} \right] \quad \ldots(6.2)$$

where n is the sample size, and $\hat{\rho}(1)$ is estimated from (6.1). This bias correction, however, is not satisfactory for other processes such as the ARMA(1,1) or FFGN. The bias in $\rho(1)$ for these processes is a function of other parameters in the model as well as the sequence length to be generated.

Before the correction of bias in the estimation of the Hurst coefficient is discussed, a clear distinction must be drawn between the estimation procedure suggested by Hurst, (1951, 1956) and that proposed by Mandelbrot and Wallis (1969) and extended by Wallis and Matalas (1970). The Hurst estimator, K, and the Wallis and Matalas estimator, H, are discussed in Appendix G. It is sufficient to mention here that both H and K are biased estimators of the Hurst coefficient. The estimator K shows greater bias but smaller variance than H. However, both H and K are unbiased around 0.7, and that the bias

for both H and K decreases with increasing sample size n but at a very slow rate (Wallis and Matalas, 1971). The magnitude of bias for both H and K, however, would depend on the time series model used. A discussion of the degree of bias in Hurst's K for a modified fast fractional gaussian noise process for a sample length of 70 years is given in a later section. Hence, to preserve a given value of the Hurst coefficient, appropriate adjustments have to be made to the input parameters of the generation model.

The brief discussion on bias correction in parameter estimation given in this Section has shown that both the variability and serial correlation of generated series may be seriously underestimated if the input parameters of the generation model are not properly corrected for bias.

In the following sections, a brief review of the currently available models that are able to preserve simultaneously a high Hurst coefficient and a low first serial correlation coefficient which is typical of peak flow series, are presented. The models are presented in chronological order of their development.

6.4 FAST FRACTIONAL GAUSSIAN NOISE MODEL (FFGN)

In this section, the Fractional Gaussian Noise (FGN) process for modelling long term serial correlation is first introduced. The difference between the FGN process and short term serial correlation processes is demonstrated in terms of correlation functions and sample functions. Finally, the most efficient approximation to FGN, the Fast Fractional Guassian Noise (FFGN) generator, is presented together with the procedure to modify FFGN to fit the low lag serial correlation coefficient.

A mathematical explanation of the Hurst phenomenon, as described in Appendix G, was provided by Mandelbrot (1965). He reproduced the Hurst phenomenon by using a self-similar stochastic process called fractional gaussian noises. The properties of such self similar processes were published by Mandelbrot and Van Ness (1968) where the terminology of Fractional Brownian Motion (FBM) and Fractional Gaussian Noises (FGN) was introduced to the hydrological community. Readers are referred to the papers by Mandelbrot et al. (1968), Mandelbrot and Wallis (1969a, b, c, d, e) and, Lawrence and Kottegoda (1977) for further details.

FGN employs an autocorrelation function which is independent of any observed correlogram but will automatically reproduce the desired Hurst coefficient.

The covariance of two terms s time units apart in a series of a discrete-time FGN standardized variables is given by:

$$C(s;h) = \frac{1}{2}\left[\,|s+1|^{2h} - 2|s|^{2h} + |s-1|^{2h}\right] \quad \ldots (6.3)$$

The comparison of the correlogram for FGN and the lag-one Markov process [AR(1)] on the basis of equal values of $\rho(1)$, the first serial correlation coefficient, is shown in Figure 6.1.

By definition, fractional noises are continuous parameter processes with "infinite memories"; that is, there is a small but non-negligible statistical relationship between members indefinitely far apart in the series. For computer simulation, it must therefore be approximated. Mandelbrot and Wallis (1969a) proposed two discrete approximations, Type I and Type II, which consist of weighted moving averages of independent Gaussian variables for the generation of flows.

Perhaps the best way to appreciate the difference between white Gaussian noise and fractional noise is to compare sample functions. Figure 6.2 shows a 1000 point realization of white noise and Figure 6.3 is a 1000 point realization of approximated fractional noise with a Hurst coefficient, h = 0.9.

It can be seen that Figure 6.3 exhibits considerably more low frequency behaviour or long term serial correlation than does Figure 6.2.

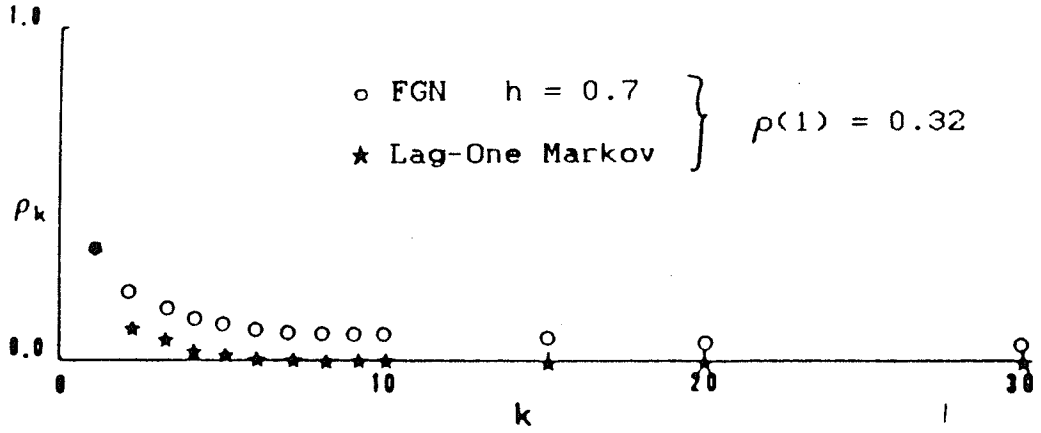


Figure 6.1 Comparison of FGN and Lag-one Markov process on the Basis of Equal values of $\rho(1)$

(after O'Connell, 1977)

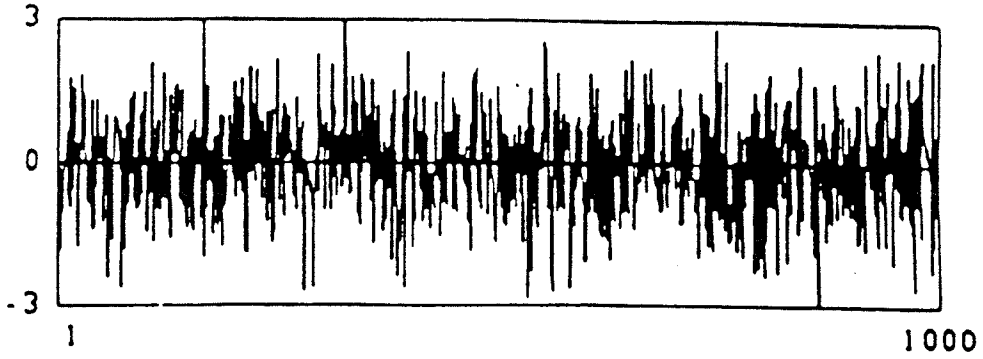


Figure 6.2 Plot of 1000 points of Standardized Gaussian White Noise

(after O'Connell, 1977)

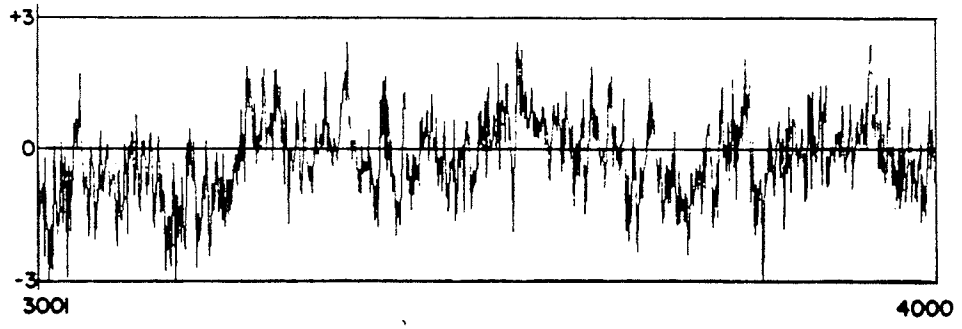


Figure 6.3 Plot of 1000 points of Standardized Type 1 Fractional Gaussian Noise with h = 0.9

Table 6.3 shows a further comparison between AR(1) process, FGN and ARMA (1,1) correlation function with same lag one correlation coefficient.

TABLE 6.3

COMPARISON OF SERIAL CORRELATION COEFFICIENTS OF AR(1), FGN, AND ARMA (1,1) PROCESSES (after Loucks et al., 1981)

Lag	AR(1) ($\rho(1)$ = 0.376)	FGN (h = 0.73)	ARMA (1,1) (ϕ = 0.92, θ = 0.706)
1	0.376	0.376	0.376
4	0.0202	0.160	0.293
16	1.6×10^{-7}	0.075	0.108
32	2.4×10^{-14}	0.052	0.028
64	6.0×10^{-28}	0.036	0.002
125	6.8×10^{-54}	0.025	1.2×10^{-5}
500	2.0×10^{-213}	0.012	3.2×10^{-19}

Both the Type I and Type II approximations of discrete fractional noise developed by Mandelbrot and Wallis (1969a) have several drawbacks. Type I approximations proved to be expensive computationally, while the Type II approximations was notably deficient in high frequencies. Although some improvements in computer

time was possible by using the filtered Type II approximations proposed by Matalas and Wallis (1971), the reduction is not significant. Mandelbrot (1971) proposed a new approximation, Fast Fractional Gaussian Noise (FFGN), which is a more efficient and flexible approximation to FGN. In addition, it uses the Hurst coefficient as an explicit parameter. The reduction in computer time is such that about 2000 FFGN variates can be generated in the time it takes to general a single Type II variate (Kottegoda, 1980).

In order to derive FFGN, the low frequency properties of the covariance function given by (6.3) is reproduced. For large lags, the covariance function is given by:

$$C(s;h) \approx h(2h - 1)\, s^{2h-2} \qquad \ldots (6.4)$$

The construction of the standardized FFGN variates, $X_f(t;\ h)$, requires two additive components. The first component, $X_L(t;\ h)$, concerns low frequency effects and is formed by weighting NT independent Markov-Gauss [AR(1)] processes with Equation (6.4) as its large lag covariance structure. The second component, $X_H(t)$, is a single Markov-Gauss process with zero mean. The second component is added to correct the high frequency error from $X_L(t;h)$. The FFGN variates are obtained by summing the low frequency term and high frequency term:

$$X_f(t;h) = X_L(t;h) + X_H(t) \qquad \ldots (6.5)$$

The low frequency component as defined by Mandelbrot (1971) takes the form:

$$X_L(t;h) = \sum_{m=1}^{NT} W_m \, X(t;r_m) \qquad \ldots (6.6)$$

where $X(t; r_m)$ is the mth Markov-Gauss process with correlation parameter r_m and theoretical weight W_m. W_m is given by:

$$W_m = \left\{ \frac{h(2h - 1)(B^{1-h} - B^{h-1}) \, B^{-2(1-h)m}}{\Gamma(3 - 2h)} \right\}^{1/2} \qquad \ldots (6.7)$$

where B is a parameter, and Γ is a gamma function. It follows that the covariance function of $X_L(t; h)$ for lag s is:

$$C_L(s;h) = \sum_{m=1}^{NT} W_m^2 \, r_m^s \qquad \ldots (6.8)$$

and the variance of the process is given by:

$$C_L(0;h) = \sum_{m=1}^{NT} W_m^2 \qquad \ldots (6.9)$$

The Markov-Gauss process $X(t; r_m)$ is a AR(1) model with zero mean, unit variance and takes the form:

$$X(t;r_m) = r_m \, X(t-1;r_m) + (1 - r_m^2)^{1/2} \, \varepsilon_t \qquad \ldots (6.10)$$

and $r_m = \exp(-B^{-m})$ $\qquad \ldots (6.11)$

with ε_t a white noise term.

Mandelbrot (1971) used a complicated method for determining the values of B and the number of Markov-Gauss processes NT. He found values 2 through 4 to be convenient for B. NT is obtained from:

$$NT = \left\| \frac{\log(QT)}{\log(B)} \right\| \qquad \ldots (6.11)$$

where Q is a quality factor and the vertical lines means the smallest integer above the value enclosed. The parameter B and quality factor Q together determine the quality of approximation.

Others, notably Chi et al. (1973) and Kottegoda (1974) recommend values of about 15 to 20 for NT. In this study, it was found that with B = 4 and NT = 15 produces good results for a wide range of h.

As a result of neglecting the high-frequency and some of the very low frequency effects in deriving the expression for the low frequency component, $X_L(t; h)$, the variance of the latter will be less than unity. To make up this deficiency in high frequency, Mandelbrot (1971) suggested that a simple Markov process can be added to the low frequency variance. Chi et al. (1973) used the following procedure:

The high frequency variance is given by:

$$\sigma_H^2 = 1 - C_L(0, h) \qquad \ldots (6.13)$$

$$= 1 - \sum_{m=1}^{NT} W_m^2 \qquad \ldots (6.14)$$

The high frequency first serial correlation coefficient, ρ_H, is therefore:

$$\rho_H = \frac{\rho(1) - C_L(1;\ h)}{\sigma_H^2} \qquad \ldots (6.15)$$

where $\rho(1)$ is obtained from (6.3) with s = 1. That is:

$$\rho(1) = 2^{2h-1} - 1 \qquad \ldots (6.16)$$

Similarly, $C_L(1;\ h)$ is obtained from (6.8) with s = 1 giving:

$$C_L(1;\ h) = \sum_{m=1}^{NT} W_m^2 \exp(-B^{-m}) \qquad \ldots (6.17)$$

The steps involved in generating normally distributed FFGN variates are briefly as follows (Srikanthan, 1979).

Step 1. Obtain the values of the mean ($\bar{x}$), standard deviation (s), lag-one serial correlation coefficient [$\hat{\rho}(1)$] and the Hurst coefficient (h) from the historical sequence. Specify the values of B, NT and the required length T to be generated.

Step 2. Compute the weighting coefficients W_m and the autocorrelations of the low frequency Markov processes r_m.

Step 3. Compute the variance σ_H^2 and lag-one serial correlation coefficient of the high frequency term ρ_H.

Step 4. NT independent random numbers, $G_m(t)$, are assumed equal to the NT Markov processes in the low frequency term to start the data generation procedure. Also, set the high frequency Markov process equal to another random number, $G(t)$.

Step 5. Compute all the Markov terms in the low frequency expression and obtain the weighted sum as in (6.6) to give $X_L(t; h)$.

Step 6. Obtain the high frequency Markov term from $X_H(t) = \rho_H X_H(t-1) + (1 - \rho_H^2)^{1/2} \cdot G(t)$

Step 7. Finally the FFGN variate is obtained from $X_f(t;h) = X_L(t; h) + \sigma_H X_H(t)$

Repeat steps 4 to 7 until the required length T is generated.

Figure 6.4 shows the flowchart for the generation of normally distributed FFGN variates with a specified length T.

The construction of FFGN described above will reproduce correctly the covariance structure given by Equation (6.3). However, for annual peak flow series and other natural time series, the first order serial correlation coefficient is usually close to zero instead of that given by (6.16). In order to match the observed first serial correlation, FFGN must be modified.

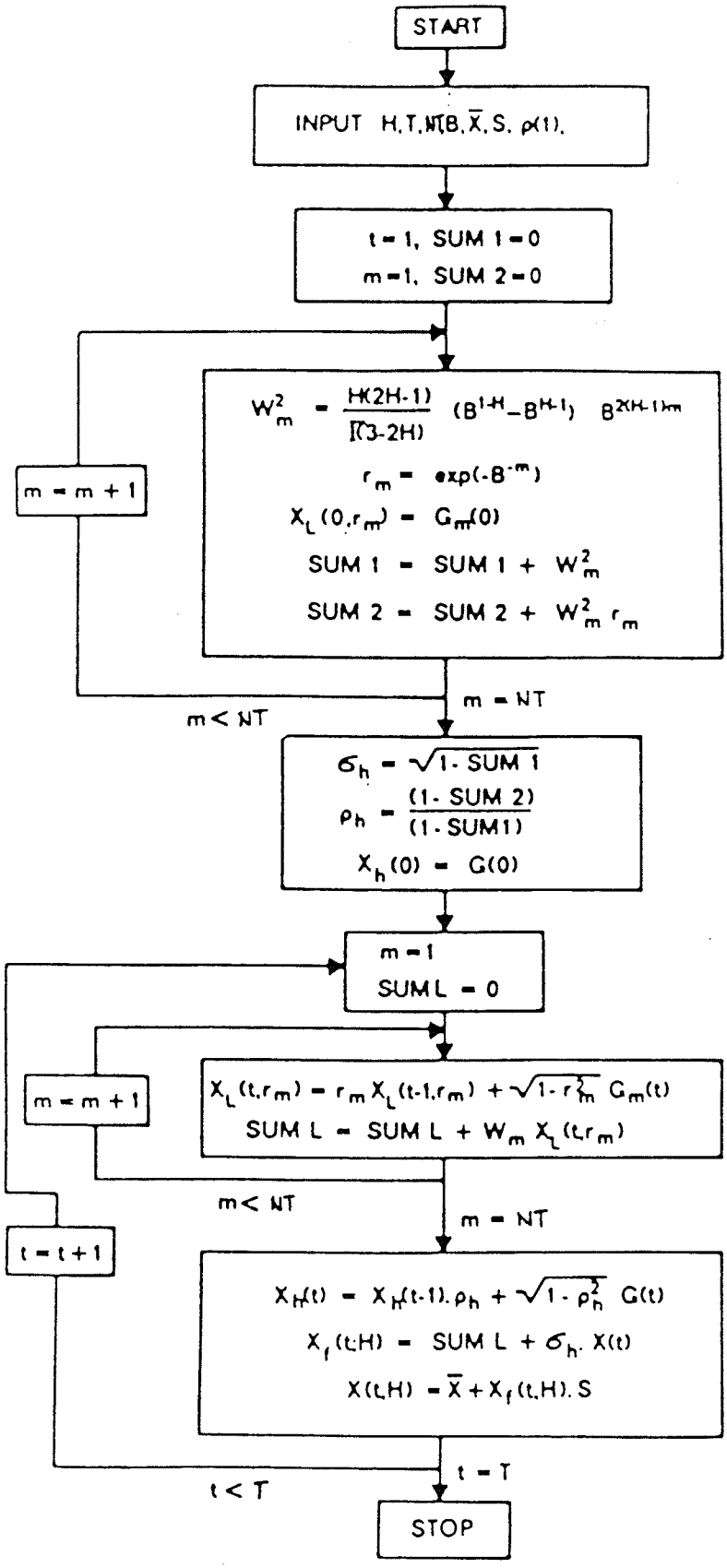


Figure 6.4 Flow Chart for the Generation of Fast Fractional Noise Variates

(after Srikanthan and McMahon, 1978)

The method proposed by Srikanthan and McMahon (1978) to modify the FFGN is to choose o_H given in (6.15) such that o(1) is equal to the observed first serial correlation coefficient. That is:

$$\rho(1) = \hat{\rho}(1) = \rho_H(1 - \sum_{m=1}^{NT} W_m^2) + \sum_{m=1}^{NT} W_m^2 \exp(-B^{-m}) \quad ..(6.18)$$

Using this method would invariably mean that $\rho_H < 0$ in order that a low enough $\rho(1)$ be obtained. Even though Mandelbrot (1971) did not impose the condition of non-negativeness for ρ_H, Chi et al. (1973) and Burges and Lettenmaier (1975) insisted that $\rho_H > 0$. No strong justification can be found for this restriction. Since the correlation function of a Markov process dies out very quickly with lag for small $\rho(1)$, the correlation function of the modified FFGN depends only on the low frequency approximation for large lags. However, the use of a negative ρ_H introduces a transient zone at the beginning in which the estimated Hurst coefficient, Hurst's K, from the generated series may be lower than the desired value. But, it will tend to the desired value as the sequence length increases. Srikanthan et al. found that with $\rho_H > -0.2$ there is no noticeable effects on the values of Hurst's K for model input h < 0.8, and a B value of 3 and NT of 10.

To produce generated sequences from the modified FFGN process that "on average" reproduce statistics equal to the historical values, the parameters used in the model

must be corrected for bias. Analytical expressions for bias correction for the modified FFGN process may be possible. However, bias correction derived from the Monte Carlo method is sufficient for most practical purposes.

Only the case when the modified FFGN sequences are normally distributed with zero mean and unit variance is investigated. The range of ρ_H used was -0.25 to 0.20 and the model Hurst coefficient, h, range from 0.55 to 0.90. In addition, B = 4 and NT = 15. The sequence lengths used were 50, 70 and 100. For each combination of ρ_H and h, and a given sequence length, 500 independent samples were generated. From these 500 samples, the expected values of Hurst's K, mean and variance can be determined. Since the expected values of Hurst's K are not significantly affected by ρ_H, their values are averaged for the various values of h. The relationship for the mean expected values of Hurst's K can then be plotted against h. Figure 6.5 shows the result for the bias in Hurst's K for a sequence length of 70.

Similarly, the relationship of the expected values of the first serial correlation coefficient $\rho(1)$ can be plotted for the various values of h at different levels of ρ_H. Figure 6.6 shows the result for a sequence length of 70.

Finally, the expected values of the standard deviation can also be plotted against h. Figure 6.7 shows the result for sequence lengths of 50, 70 and 100. The

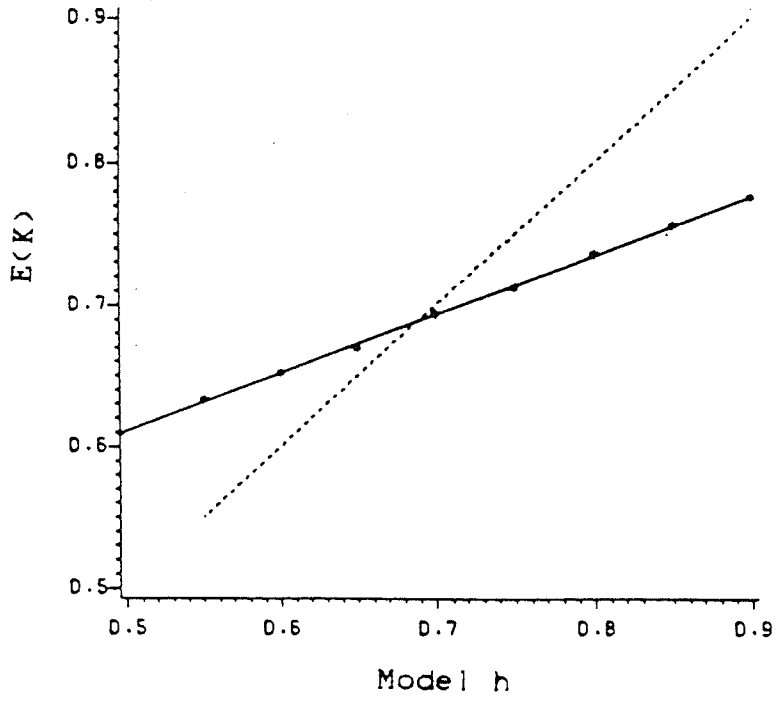


Figure 6.5 Bias in Hurst's K for modFFGN process (N=70)

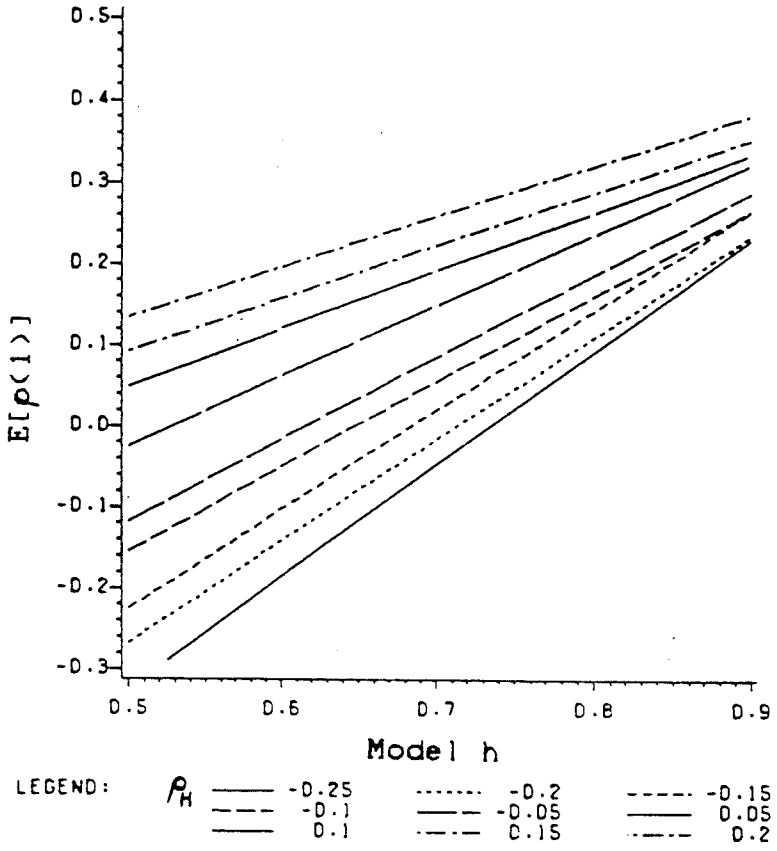


Figure 6.6 Bias in $\rho(1)$ at Various Values of h for modFFGN Process (N = 70)

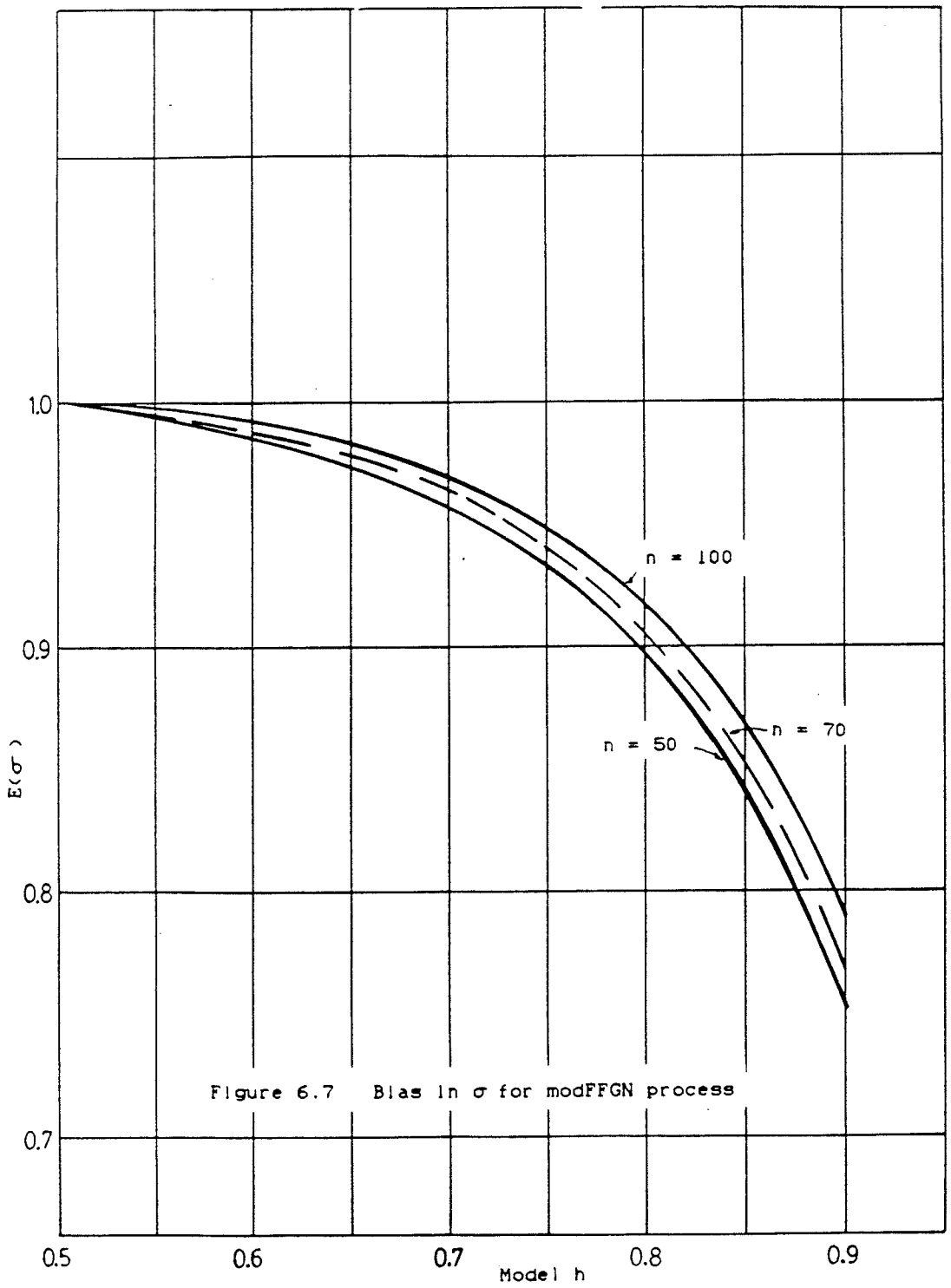


Figure 6.7 Bias in σ for modFFGN process

standard deviation does not seem to be affected by ρ_H.

Once the relationships between the expected values and the input parameters are obtained, they can be used to select the appropriate ρ_H, h, and bias correction for the variance to obtain the desired $\rho(1)$ and Hurst's K. For example, a standardized series of length 70 is observed. The estimated first order serial correlation is 0.1 and Hurst's K is 0.72. Using Figure 6.5, with E(K) = 0.72, h = 0.765. With h = 0.765 and r_1 = 0.1, ρ_H = -0.15 from Figure 6.6. Finally, from Figure 6.7 with h = 0.765, the model variance should be increased to 1/0.958. Hence with model parameters h = 0.765, ρ_H = -0.15 and standard deviation = 1.044, the generated sequences will give E[ρ(1)] = 0.1, E(K) = 0.72 and unit variance. Of course, it is not possible to obtain exactly the historical statistics since they are random variables, but the difference will be small. The possibility also exists that E[ρ(1)] and E(K) may be incompatible with or lie outside the simulation results. For example, if E[ρ(1)] = 0.0 and E(K) = 0.80, then it is practically impossible to obtain ρ_H and h to fit. However, for most practical ranges of E[ρ(1)] and E(K), there is little problem.

In the next section, the Broken Line (BL) Model is discussed.

6.5 BROKEN-LINE MODEL (BL)

The extensive computing resources required to implement the FFGN processes have prompted a search for alternative processes that reproduce the Hurst phenomenon. One such model is the Broken Line Model (BL) proposed by Mejia et al. (1972, 1974).

As introduced by Mejia et al. (1972), the BL model consists of a sum of series of simple BL processes, each of which results from linear interpolation between equally-spaced normally and independently distributed (NID) variables with random displacement of the starting point of the series in order to make it stationary. The simple broken line process is illustrated in Figure 6.8.

Algebraicly, the simple BL process is given by:

$$\xi(t) = \xi(t' - ca) = \sum_{j=0}^{\infty} \left[\eta_j + \frac{\eta_{j+1} - \eta_j}{a} (t' - ja)\right] I(t') \qquad \ldots (6.19)$$

where η = independent and identically distributed random variables with zero mean and variance σ^2,

c = a random variable uniformly distributed over the interval (0, 1),

a = time distance between the η_j, also referred to as the memory parameter, and

$I(t') = 1 \qquad ja \leqslant t' \leqslant (j + 1)a$

$= 0 \qquad$ otherwise

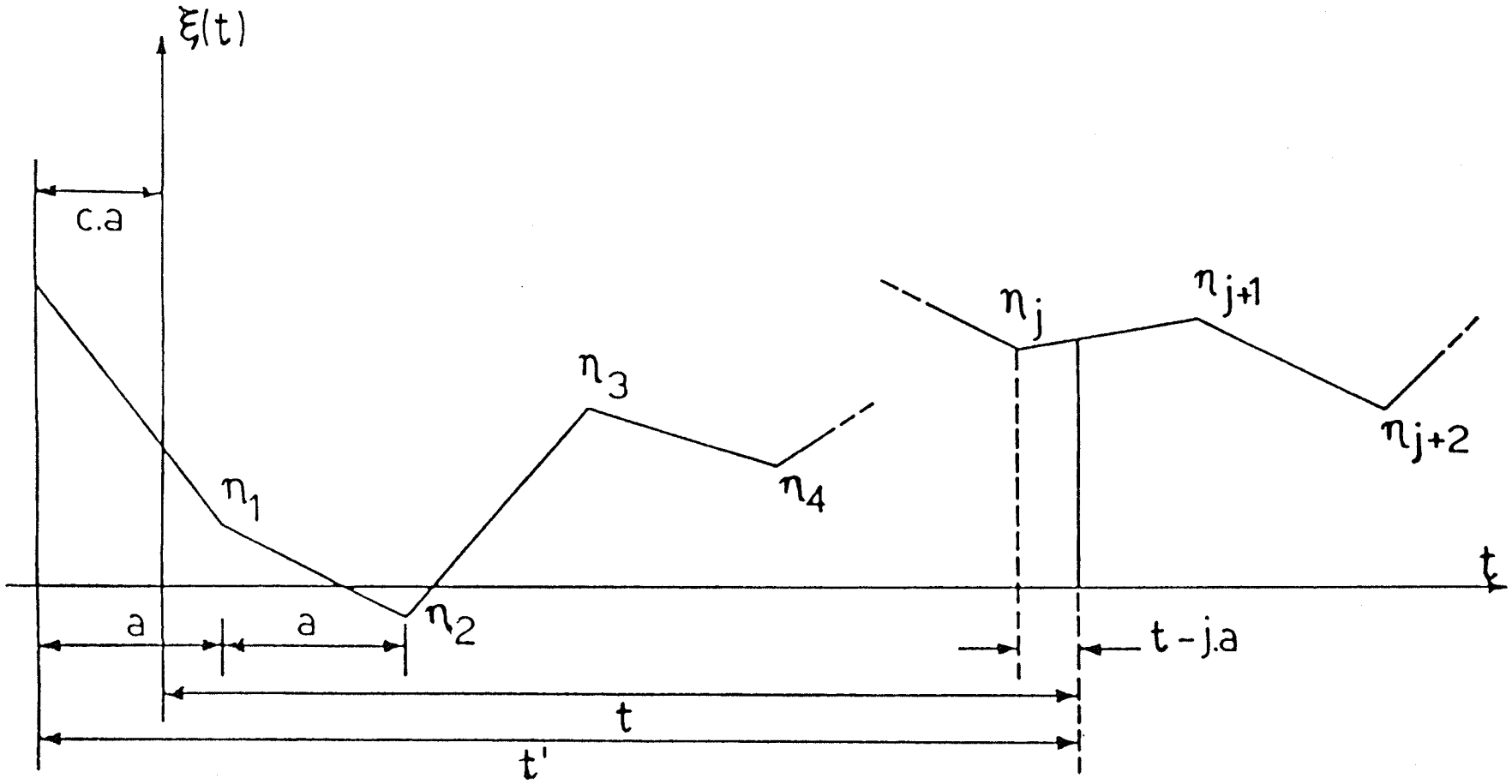


Figure 6.8 Schematic Representation of the Simple Broken Line Process

The mean of the process is zero and the variance is $(2/3)\sigma^2$. The autocorrelation function of the process is:

$$\rho_s = \begin{cases} 1 - 0.75(s/a)^2[2 - 2/a] & 0 \leqslant s \leqslant a \\ 0.25\,[2 - s/a]^3 & a \leqslant s \leqslant 2a \\ 0 & 2a \leqslant s \end{cases} \quad ..(6.20)$$

The general BL process, $X_b(t)$, which is capable of reproducing a particular value of the Hurst coefficient, h, is obtained from a weighted summation of NT independent, standardized simple BL processess. This approach is directly comparable to that of FFGN applied by Mandelbrot (1971), but with BL used in place of AR(1) processes. Therefore, as emphasised by Mandelbrot (1972), the BL process is more an alternative to FFGN than to the FGN process itself.

As with FFGN, a low frequency component, $X_L(t; h)$, is employed to approximate the large lag covariance structure of discrete time FGN, and a high frequency component, $X_H(t)$, is added to ensure that the BL model has unit variance. The low frequency term is defined as:

$$X_L(t; h) = \sum_{m=1}^{NT} V_m \, \xi_m(t) \quad \ldots (6.21)$$

where $\xi_m(t)$ is a simple BL process with parameters a_m, c_m, zero mean and unit variance,

$$V_m = \left\{ \frac{a_1 b^{2n-2} (B^{h-1} - B^{1-h})}{2(h-1)} \, B^{2(h-1)m} \right\}^{1/2} \quad \ldots (6.22)$$

$$a_m = a_1 B^m$$

$$b = \frac{h(2h-1)(2h-2)(2h-3)(2h-4)(2h-5)}{6(2^{3-2h}-1)} \quad \ldots (6.23)$$

and a_1, which should be greater than unity, is obtained from the first serial correlation coefficient:

$$\rho(1) = \frac{a_1^{2h-2}\, b(B^{h-1} - B^{1-h})}{2(h-1)} \sum_{m=0}^{NT} \left\{ 1 - \frac{3}{4a_1^2 B^{2m}} \left(2 - \frac{1}{a_1 B^m}\right) \right\}$$

$$\cdot\ B^{2(h-1)m} \quad \ldots (6.24)$$

The number of simple broken line processes (NT) required to develop the BL process is obtained from:

$$NT = \text{smallest integer above } [\log(QT)/\log B] \quad \ldots (6.25)$$

where T is the sequence length, Q and B are 5 and 4 respectively (Mejia et al., 1972).

The high frequency term is then defined as an independent process with variance:

$$\sigma_H^2 = 1 - \sum_{m=0}^{NT} V_m^2 \quad \ldots (6.26)$$

Mejia et al. (1972) suggest that the high frequency term should be a BL process with $a = a_1$. However, Srikanthan and McMahon (1978) advocated assuming a value for a_1 (> 1) and then evaluating the right hand side of (6.24) to give ρ_L (say). The high frequency term may then

be taken as an AR(1) model with variance given by (6.26) and first serial correlation coefficient given by:

$$\rho_H = (\rho_{(1)} - \rho_L) / \sigma_H^2 \qquad \ldots (6.27)$$

Srikanthan and McMahon (1978) also suggest that a value of $a_1 = 2$ will give satisfactory results.

The computer time per generated realization with BL model was found to be roughly a quarter of that required for the FFGN model.

Although the BL model is fully capable of modelling both long and short term serial correlation simultaneously, its greatest deterrent to the hydrologist is the estimation of its parameters which require iterative procedure (Lawrence and Kottegoda, 1977). In addition, small sample bias of the parameters of the BL model are largely unknown.

6.6 ARMA (1,1) MODEL

In this section, the first-order autoregressive-first-order-moving average [ARMA(1,1)] process proposed by O'Connell (1971, 1974) as an approximation to FGN is presented.

The defining equation for the ARMA (1,1) process is given by:

$$X_t = \phi X_{t-1} + a_t - \theta a_{t-1} \qquad \ldots (6.28)$$

where ϕ and θ are the parameters of the model, X_t is the process value at time t, and a_t is an independent random variate at time t.

The autocorrelation function of the process is given by:

$$\rho(1) = \frac{(\phi - \theta)(1 - \phi\theta)}{1 - 2\phi\theta + \theta^2} \qquad \ldots (6.29)$$

and $$\rho_k = \phi^{|k|-1} \rho(1) \qquad |k| > 1 \qquad \ldots (6.30)$$

The autocorrelation function of the ARMA (1,1) process exhibits an exponential decay from the first autocorrelation coefficient onwards, with the rate of decay being controlled by the autoregressive parameter, ϕ.

O'Connell (1974) has shown that with $0.8 < \phi < 0.99$ and $0.5 < \theta < 0.95$, the ARMA (1,1) was in certain instances able to model long term serial correlation. More specifically, for large values of ϕ, a value of θ can be found which preserves a predetermined value of the first serial correlation coefficient yet provides a slowly decaying autocorrelation function. Therefore, although the Hurst coefficient, h, is asymptotically equal to 0.5 for

the ARMA (1,1) process, careful choice of ø and θ will produce values of h substantially greater than 0.5.

Fitting of an ARMA (1,1) model to a sequence of observed flows is complicated by small sample bias in the estimates of the variance, first serial correlation coefficient and Hurst coefficient. However, O'Connell (1974) has shown that the small sample estimate of the variance, s^2, may be written as:

$$E(s^2) = \sigma^2 \cdot f(n, \rho(1), ø) \qquad \ldots (6.31)$$

with, $$f[n, \rho(1), ø] = 1 - \frac{2\rho(1)}{n(n-1)} \left\{ \frac{n(1-ø) - (1-ø^n)}{(1-ø)^2} \right\} \qquad \ldots (6.32)$$

where σ^2 and $\rho(1)$ are the population variance and first serial correlation coefficient, and n is the sample size.

O'Connell (1974) used Monte Carlo methods to derive the small sample expectations of $\rho(1)$ and Hurst's K, for sample sizes of 25, 50 and 100, and selected values of ø and θ. The results for ø = 0.92 are shown in Table 6.4.

When modelling flows, the ARMA (1,1) model may be written as:

$$X_t = \bar{x} + ø[X_{t-1} - \bar{x}] + s[a_t - \theta a_{t-1}] \qquad \ldots (6.33)$$

where $\bar{x}$ and s are the mean and standard deviation of the process and a_t are normally and independently distributed

TABLE 6.4

SMALL SAMPLE ESTIMATES OF THE HURST COEFFICIENT, E[K], AND FIRST SERIAL CORRELATION COEFFICIENT, E[ρ(1)] FOR Ø = 0.92 AND VARIOUS VALUES OF θ IN AN ARMA (1,1) MODEL (after O'Connell, 1974)

		E[K]			E[ρ(1)]		
θ	ρ(1) \ n	25	50	100	25	50	100
0.88	0.049	0.657	0.664	0.654	0.001	0.012	0.028
0.84	0.114	0.687	0.680	0.686	0.005	0.046	0.079
0.80	0.189	0.686	0.705	0.709	0.075	0.093	0.123
0.76	0.269	0.699	0.735	0.745	0.082	0.160	0.208
0.72	0.349	0.725	0.751	0.756	0.116	0.199	0.240
0.68	0.426	0.740	0.782	0.783	0.169	0.269	0.332
0.64	0.496	0.772	0.783	0.800	0.218	0.309	0.390
0.60	0.560	0.773	0.796	0.803	0.273	0.364	0.437
0.56	0.616	0.774	0.820	0.825	0.285	0.432	0.516
0.52	0.665	0.794	0.825	0.828	0.335	0.467	0.532

variates with zero mean and variance given by:

$$\sigma_a^2 = (1 - \phi^2) / (1 + \theta^2 - 2\phi\theta) \qquad \ldots (6.34)$$

The fitting procedure for this model is given in O'Connell (1974, 1977). The steps are as follows:

i) Compute $\bar{x}$, s, Hurst's K and the first serial correlation, $\hat{\rho}(1)$, from the n-year historical series;

ii) Using the tables of O'Connell (1974), find values of ϕ and θ for which:

$$E[K] = K; \quad E[\rho(1)] = \hat{\rho}(1)$$

iii) Obtain the unbiased estimate of the variance, S_{ub}^2, from (6.31). This is given by:

$$S_{ub}^2 = s^2 / f[n, \rho(1), \phi]$$

iv) Substitute the values of $\bar{x}$, S_{ub}, ϕ and θ into (6.33) to obtain the required generating mechanism.

The ARMA (1,1) model has two principle advantages. Firstly, it has only two parameters; secondly, it requires substantially less computing time than the FFGN and BL process.

However, the major drawback of the ARMA (1,1) model is that the Hurst coefficient is not an explicit parameter

of the model as it is for the FFGN and BL models. No equivalence has been found between the ARMA (1,1) parameters and h (Lettenmaier and Burges, 1977a).

6.7 ARMA-MARKOV (AM) MODEL

While the ARMA (1,1) model is computationally efficient, the inability to preserve a given Hurst coefficient as an explicit model parameter may be a major drawback in operational applications (Lettenmaier and Burges, 1977a). On the other hand, the FFGN model, while it models h as an explicit parameter, is relatively expensive to run, especially when the number of Markov-Gauss terms is large. To preserve the economy of the ARMA (1,1) and also use the Hurst coefficient as an explicit parameter, Lettenmaier and Burges (1977a) modified Mandelbrot's (1971) approach to deriving FFGN. Instead of fitting a series of independent processes to 'build' a desired autocorrelation function theoretically, a similar effect was achieved by making the approximation to the FGN autocorrelation function on a geometric basis. Lettenmaier and Burges (1977a) proposed a mixed model called the ARMA-Markov which uses the Hurst coefficient as an explicit model parameter to achieve this fit.

For zero mean and unit variance process, the ARMA-Markov model consists of five parameters: the Markov

and ARMA variance fractions C_1 and C_2, respectively; the Markov and ARMA lag one autocorrelation coefficients ρ_M and ρ_{AM}, respectively; and the autoregressive parameter ϕ of the ARMA model. The moving average parameter, θ, of the ARMA process is uniquely defined by ϕ and the lag one correlation condition $|\theta| < 1$ is imposed.

The generating equation for the zero mean and unit variance ARMA-Markov process is given by:

$$X_t = \rho_M X_{t-1}^{(M)} + \varepsilon_t^{(M)} + \phi X_{t-1}^{(AM)} - \theta \varepsilon_{t-1}^{(AM)} + \varepsilon_t^{(AM)} \quad \ldots (6.35)$$

where $\varepsilon_t^{(M)}$ and $\varepsilon_t^{(AM)}$ are independent processes having variance $C_1(1 - \rho_M^2)$ and $C_2[(1 - \phi^2)/(1 + \theta^2 - 2\phi\theta)]$, respectively.

The autocorrelation function of this process is fitted to the theoretical autocorrelation function of FGN given by (6.3) at three specified lags, K_1, K_2 and K_3. The lag one autocorrelation coefficient may be arbitrarily specified. The parameters of the model are obtained by solving the following system of equations:

$$\begin{aligned}
C_1 + C_2 &= 1 \\
C_1\rho_M + C_2\rho_{AM} &= \rho(1) \\
C_1\rho_M^{K_1} + C_2\rho_{AM}^{K_1-1} &= C(K_1; h) \\
C_1\rho_M^{K_2} + C_2\rho_{AM}^{K_2-1} &= C(K_2; h) \\
C_1\rho_M^{K_3} + C_2\rho_{AM}^{K_3-1} &= C(K_3; h)
\end{aligned} \quad \ldots (6.36)$$

where C_1, C_2, ρ_M, ρ_{AM} and ϕ are all constrained to lie between 0 and 1, $\rho(1)$ is the desired first serial correlation coefficient and $C(K; h)$ is the autocorrelation function of FGN given by (6.3). Lettenmaier and Burges (1977a) suggest to take K_1, K_2 and K_3 to be approximately n/8, n/2 and n, where n is the length of the sequence to be generated. The system of equations given by (6.36) may be solved by using Newton's method to give the model parameters for a given $\rho(1)$ and h. The second ARMA parameter, θ, is obtained from:

$$\rho_{AM} = \frac{(1 - \phi\theta)(\phi - \theta)}{1 + \theta^2 - 2\phi\theta} \qquad \ldots (6.37)$$

The value of θ which satisfies the invertibility condition $|\theta| < 1$ is taken.

The ARMA-Markov model has also been extended to generate skewed variates by Srikanthan (1979).

According to Srikanthan (1979), the ARMA-Markov model, while it is able to give similar values for the Hurst coefficient as FFGN and BL, has the following disadvantages:

i) It does not preserve the mean as well as FFGN or BL models.

ii) It considerably underestimates the first serial correlation coefficient. Both FFGN and BL give better results, and

iii) It is not possible to obtain parameter estimates for series with negative first serial correlation.

In addition to the disadvantages listed by Srikanthan (1979), the small sample biases of the ARMA-Markov model parameters are unknown and the parameters of the models are difficult to obtain.

However, the ARMA-Markov do have two desired attributes. Firstly, it requires less computer time than the FFGN or BL models, and secondly, it uses both h and o(1) explicitly to derive its parameters.

In the next section, a simple and relatively efficient model, developed along the lines of the ARMA-Markov model is described.

6.8 MIXED-NOISE MODEL (MN)

In the previous Section, it was shown that while the ARMA-Markov model is computationally efficient, its major drawback is in the estimation of the model's parameters.

This problem prompted the development of the Mixed-Noise model.

In the development of the MN model, the Hurst coefficient, h, and first order serial correlation $\rho(1)$ are used explicitly to estimate the model's parameters. These are very easily obtained.

In principle, the MN model is obtained as the sum of three independent AR(1) process each with a suitable weight so as to reproduce approximately the autocorrelation function characterized by a given first order correlation coefficient and a long term correlation structure corresponding to fractional noise with a given Hurst coefficient. The autocorrelation function of a mixed-noise process is:

$$\rho_{MN}(K) = a^2\rho_H{}^K + b^2\rho_M{}^K + c^2\rho_L{}^K \qquad \ldots (6.38)$$

where, a^2, b^2, c^2, are the variance fraction (or weights) which sum to unity, ρ_H, ρ_M, ρ_L, are the autocorrelation coefficients of the three independent AR(1) processes. The first AR(1) process models the high frequency effects, the second AR(1) process models the intermediate or medium frequency effects, and the third AR(1) process models the low frequency effects of the time series. Hence, essentially the technique is to fit the autocorrelation function of FGN with the given three weighted auto-correlation function of the AR(1) process.

The MN model has six parameters. The three variance fractions (a^2, b^2 and c^2) and the autoregressive parameters (ρ_H, ρ_M, ρ_L).

The generating equation for a zero mean, unit variance MN process is given by:

$$X_t = a(\rho_H X_{t-1}^{(H)} + \varepsilon_t^{(H)}) + b(\rho_M X_{t-1}^{(M)} + \varepsilon_t^{(M)}) + c(\rho_L X_{t-1}^{(L)} + \varepsilon_t^{(L)}) \quad \ldots (6.39)$$

where, $\varepsilon_t^{(H)}$, $\varepsilon_t^{(M)}$ and $\varepsilon_t^{(L)}$ are normal independent processes having variance $a^2(1 - \rho_H^2)$, $b^2(1 - \rho_M^2)$ and $c^2(1 - \rho_L^2)$, respectively.

The autocorrelation function of this process is fitted to the theoretical autocorrelation function of FGN given by (6.3) at four specified lags, K_1, K_2, K_3, and K_4. The lag one autocorrelation coefficient, like the ARMA-Markov model, may be arbitrarily specified. To obtain the parameters of the model requires the solution of the following system of equations:

$$a^2 + b^2 + c^2 = 1 \quad \ldots (6.40a)$$

$$a^2\rho_H + b^2\rho_M + c^2\rho_L = \rho(1) \quad \ldots (6.40b)$$

$$a^2\rho_H^{K_1} + b^2\rho_M^{K_1} + c^2\rho_L^{K_1} = c(K_1; h) \quad \ldots (6.40c)$$

$$a^2\rho_H^{K_2} + b^2\rho_M^{K_2} + c^2\rho_L^{K_2} = c(K_2; h) \quad \ldots (6.40d)$$

$$a^2\rho_H^{K_3} + b^2\rho_M^{K_3} + c^2\rho_L^{K_3} = c(K_3; h) \quad \ldots (6.40e)$$

$$a^2\rho_H^{K_4} + b^2\rho_M^{K_4} + c^2\rho_L^{K_4} = c(K_4; h) \quad \ldots (6.40f)$$

where, a^2, b^2, c^2, ρ_M and ρ_L are constrained to lie between 0 and 1, $\rho(1)$ is the desired first order serial correlation coefficient, and $c(K; h)$ is the autocorrelation function of FGN given by (6.3). ρ_H within limits is allowed to be negative if necessary.

In this study, it was found convenient to take K_1 = 4, K_2 = 15, K_3 = 54, and K_4 = 200. On a logarithmic scale, these chosen lags are equally spaced. The value of K_4 = 200 is chosen to reflect the planning period of most major flood protection schemes. Also, the chosen spacings facilitate estimation of the model parameters.

Since the autocorrelation function of an AR(1) process 'dies off' rapidly with increasing lags, the system of equations (6.40) can be evaluated sequentially starting with the low frequency end. From (6.40e) and 6.40f), and assuming ρ_H^K and ρ_M^K to be negligible at lags K_3 and K_4,

$$\frac{C(K_3; h)}{C(K_4; h)} = \frac{c^2 \rho_L^{K_3}}{c^2 \rho_L^{K_4}}$$

$$= \rho_L^{K_3 - K_4} \qquad \ldots (6.41)$$

For a given h value, K_3 and K_4, the left hand side of (6.41) is defined. Therefore ρ_L can be calculated. Substituting ρ_L into (6.40f) and ignoring the high and medium frequency terms, c^2 is obtained.

From (6.40d), and assuming ρ_H^K to be negligible at lag K_2.

$$c(K_2 ; h) = b^2\rho_M^{K_2} + c^2 \rho_L^{K_2} \quad \ldots (6.42)$$

From which one gets:

$$b^2 = \frac{C(K_2; h) - c^2\rho_L^{K_2}}{\rho_M^{K_2}} \quad \ldots (6.43)$$

Substituting into (6.40c), one gets ρ_M.

Then from (6.43), one obtains b^2.

From (6.40a), $a^2 = 1 - b^2 - c^2$,

and from (6.40b), $\rho_H = [\rho(1) - b^2\rho_M - c^2\rho_L)/a^2]$ ...(6.44)

Perhaps the best way to illustrate the above computational procedure is by an example. Let the desired Hurst coefficient h and first serial correlation coefficient $\rho(1)$ be 0.70 and 0.2 respectively.

From (6.41),

$$\frac{C(54; 0.70)}{C(200; 0.70)} = \frac{0.0256}{0.0117} = \rho_L^{-146}$$

Therefore, $\rho_L = 0.99465$

Substituting into (6.40f), assuming the high and medium frequency term to be negligible,

$$c^2 = \frac{C(200; 0.7)}{\rho_L} = 0.03419$$

From (6.43),

$$b^2 = \frac{0.02355}{\rho_M^{15}}$$

Substituting into (6.40c), one gets:

$$C(4;\ 0.7) = 0.02355\ \rho_M^{-11} + c^2\rho_L^{4}$$

Therefore, $\rho_M = 0.88612$.

From (6.43), $b^2 = 0.14442$.

From (6.40a), $a^2 = 1 - 0.14442 - 0.03419$

$= 0.82139$

and from (6.40b), $\rho_H = (0.2 - b^2\rho_M - c^2\rho_L)/a^2$

$= 0.04629$.

Figure 6.9 shows the autocorrelation function of the fitted mixed-noise process.

Once the parameters (a^2, b^2, c^2, ρ_H, ρ_M, ρ_L) are determined, one can proceed with generating synthetic sequences.

Figure 6.10 shows on double-log graph paper the autocorrelation function of the MN and theoretical FGN process for h = 0.55(0.05)0.95. The lag one correlation coefficients of the FGN and MN models are kept identical in this case.

Figure 6.11 shows graphically the relationships among the variance fractions (a^2, b^2, c^2) and the correlation coefficient (ρ_H, ρ_M, ρ_L) at various values of h to

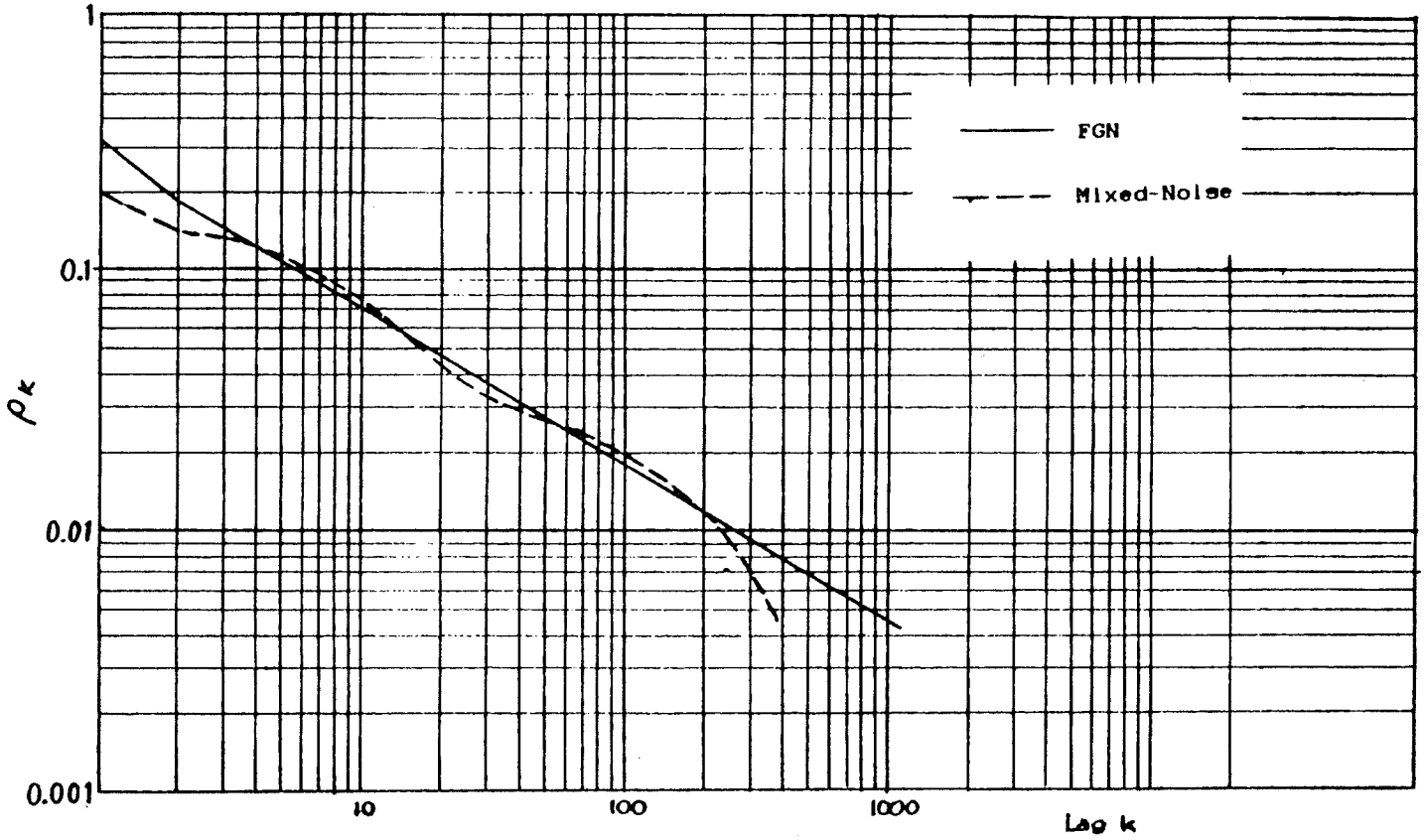


Figure 6.9 Autocorrelation Function of Mixed-Noise Process with h = 0.70 and $\rho(1) = 0.20$

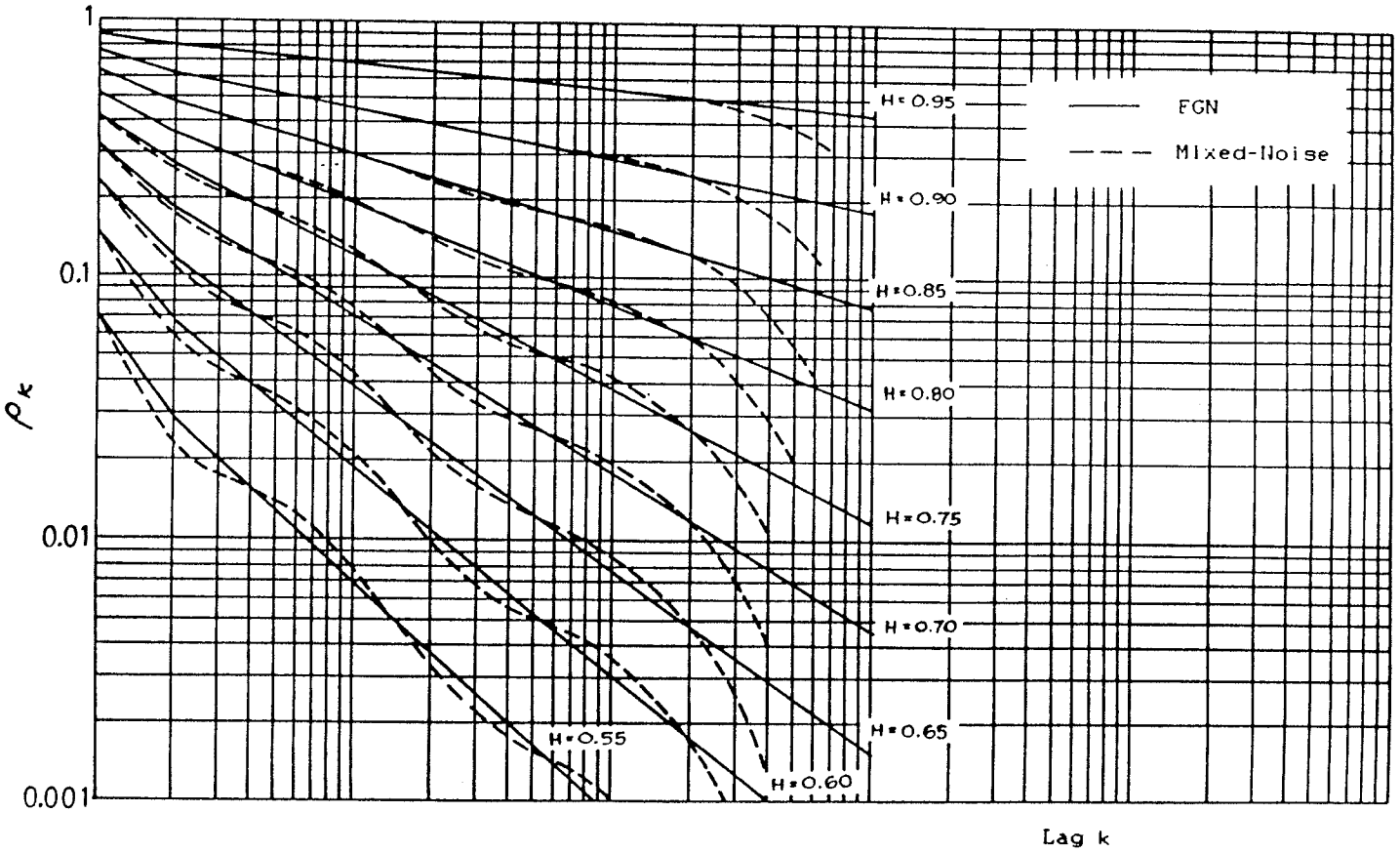


Figure 6.10 Autocorrelation Function of FGN and Mixed-Noise Process on Equal Values of $\rho(1)$

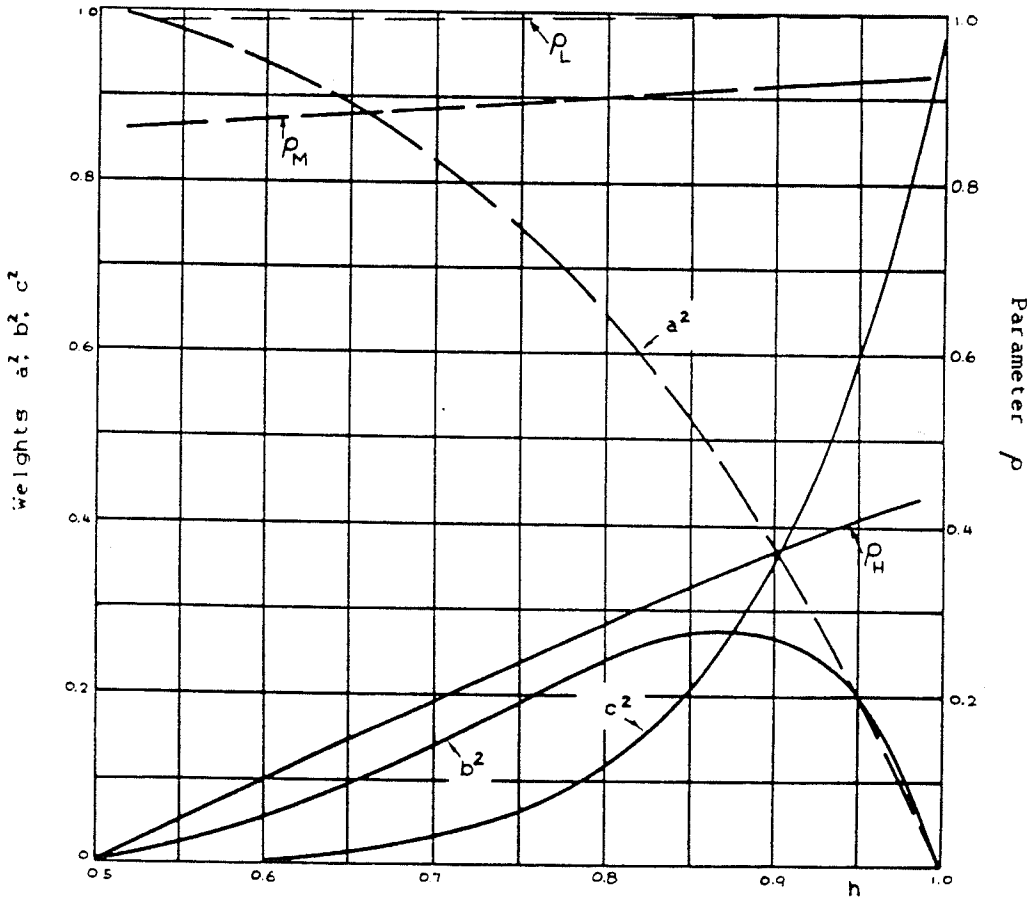


Figure 6.11 Parameter Relationship of Mixed-Noise Process

preserve the lag one correlation coefficient of FGN given by (6.16).

Determining the parameters of the MN model so that it correctly reproduces a sequence of observed flows for a given period length is complicated by small sample bias in the estimates of the Hurst coefficient, first serial correlation coefficient and variance.

Proper bias corrections for these parameters can however be obtained by the Monte Carlo method. Figure 6.12 shows the expected values of the standard deviation σ for sample sizes of 50, 70 and 100 at various values of h. It can be seen that the bias in the standard deviation is quite substantial at large values of h. As expected, the bias decreases with increasing sample size. Examination of the data reveals that the bias in σ did not seem to be affected significantly by the high frequency component of the model.

To obtain small sample expectation of $\rho(1)$ and Hurst's K, ρ_H in (6.44) is assumed to take on values of -0.3 to +0.3 for a given value of h. This range of ρ_H will cover most practical situations. The resulting expectation of the first serial coefficient, $E[\rho(1)]$, and Hurst's K, E(K), can be determined. Figures 6.13 - 6.15 show the small sample expectation of $\rho(1)$ and Hurst's K for sample sizes of 50, 70 and 100. Each point on the graphs is based on 500 replications.

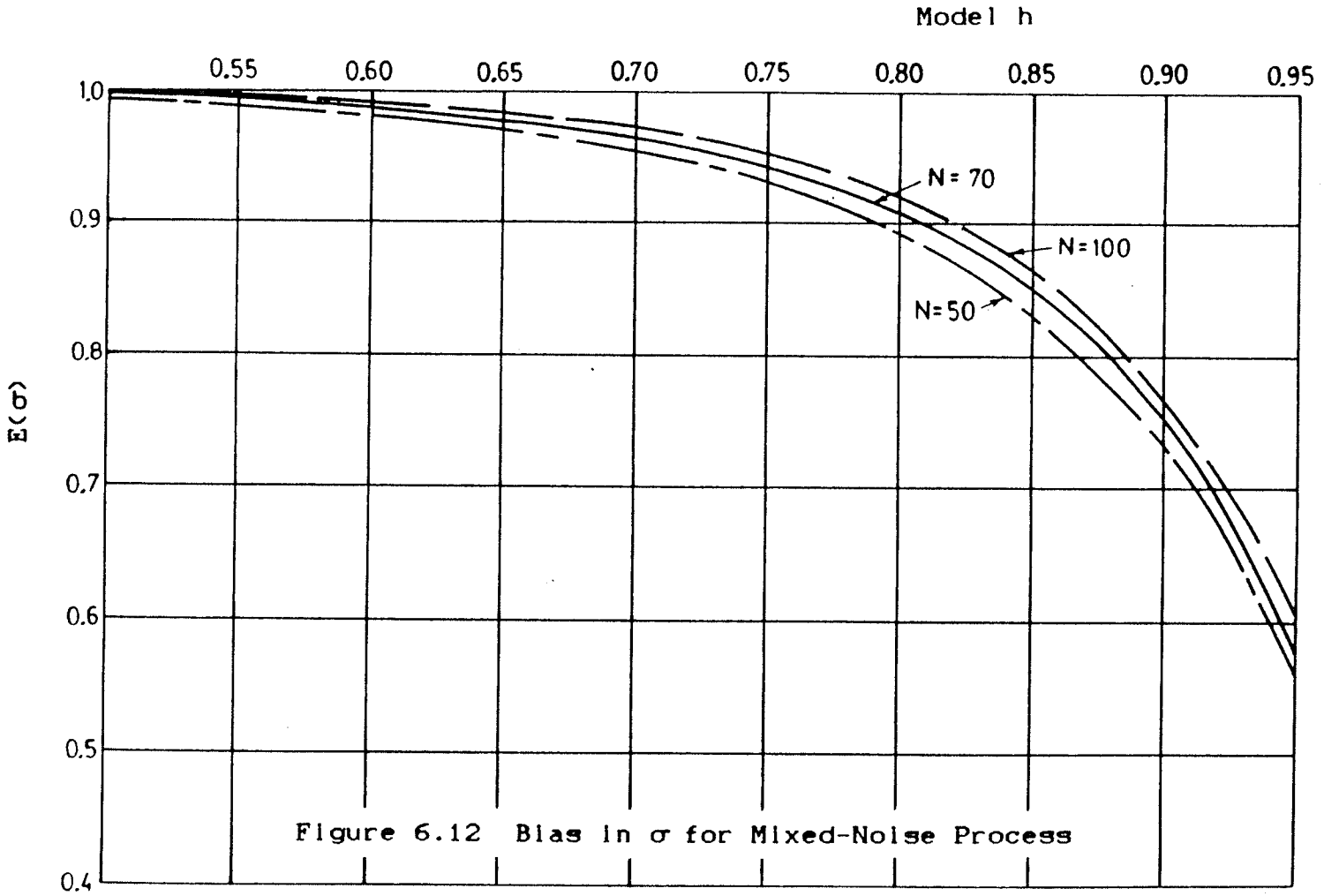


Figure 6.12 Bias in σ for Mixed-Noise Process

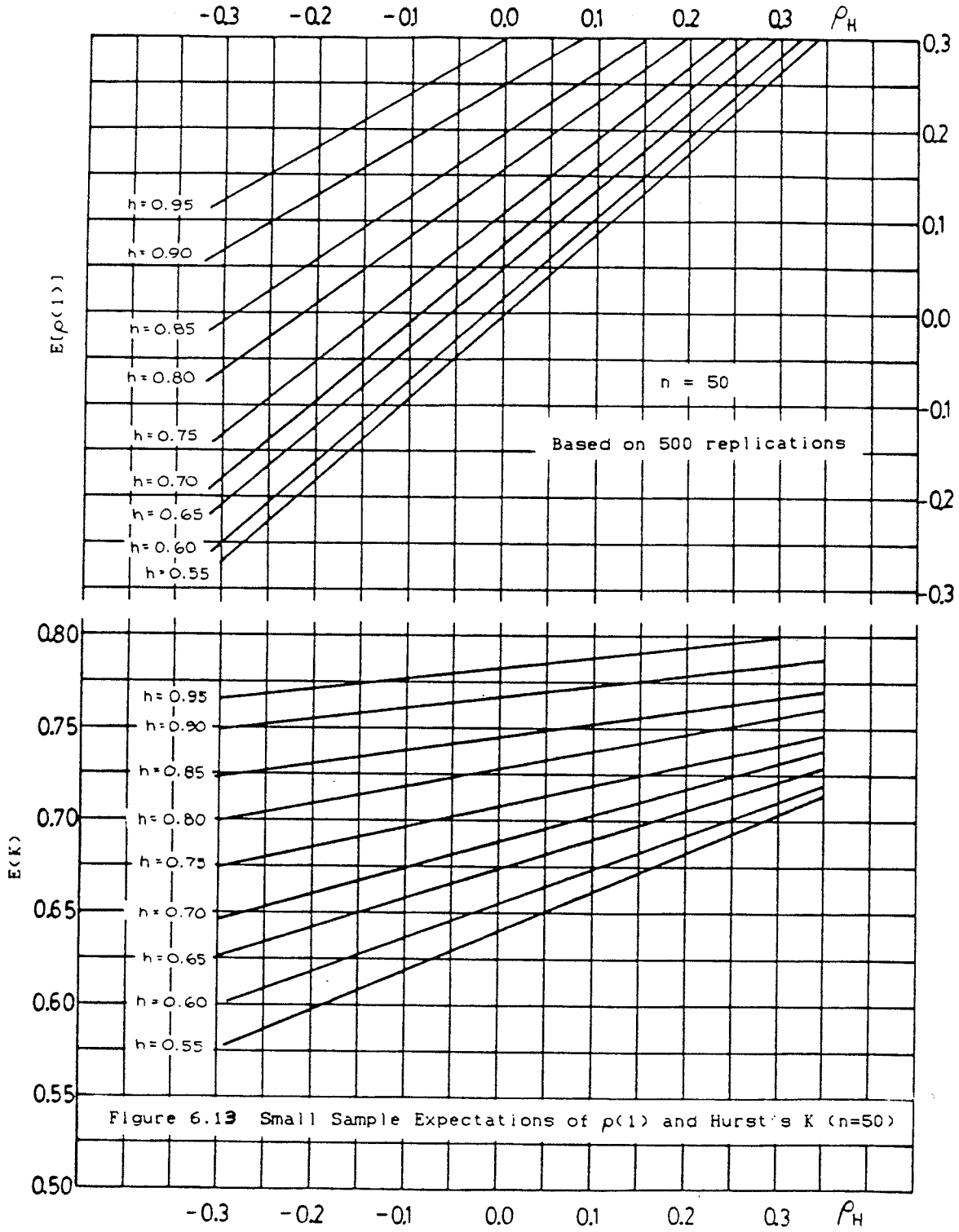


Figure 6.13 Small Sample Expectations of $\rho(1)$ and Hurst's K (n=50)

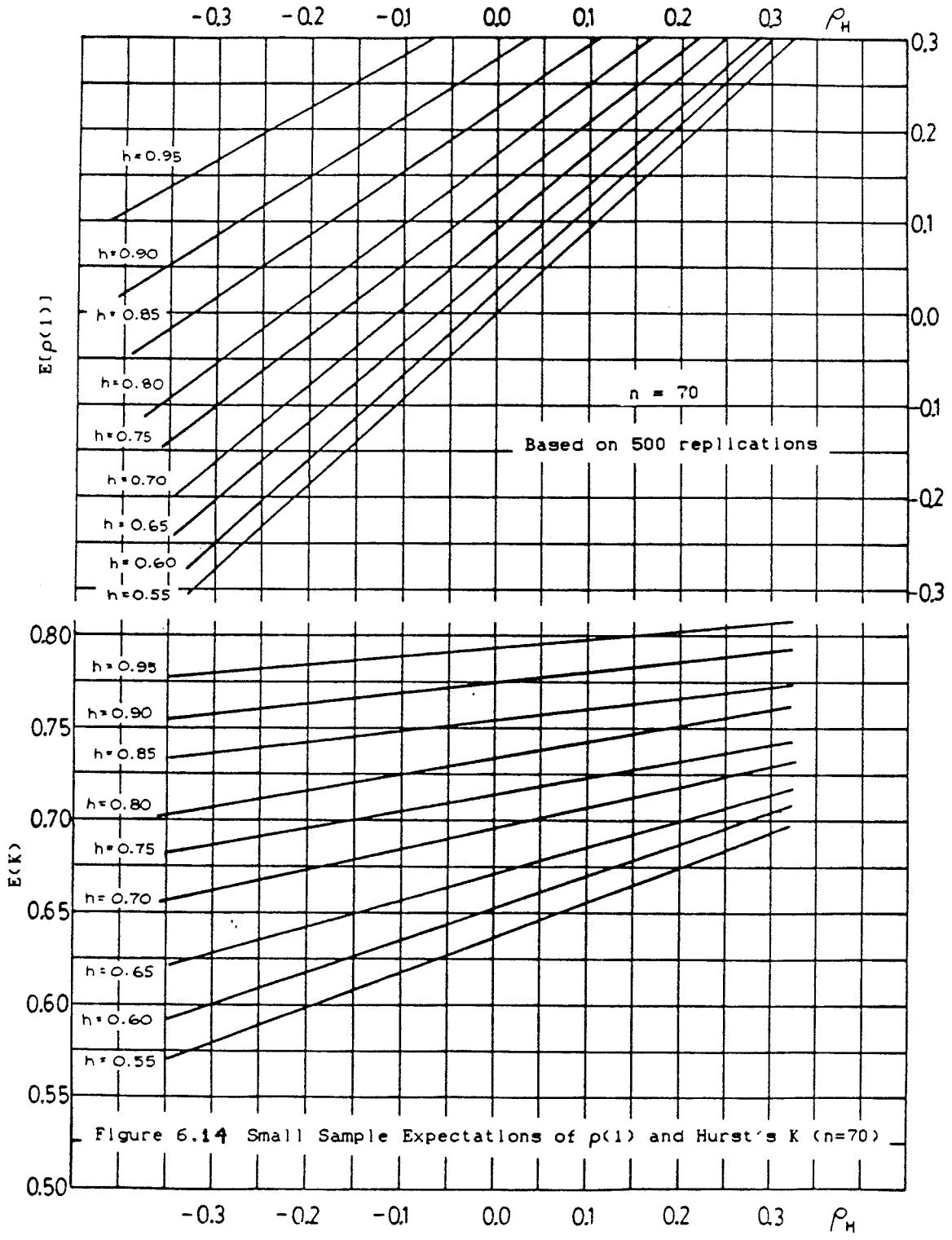


Figure 6.14 Small Sample Expectations of ρ(1) and Hurst's K (n=70)

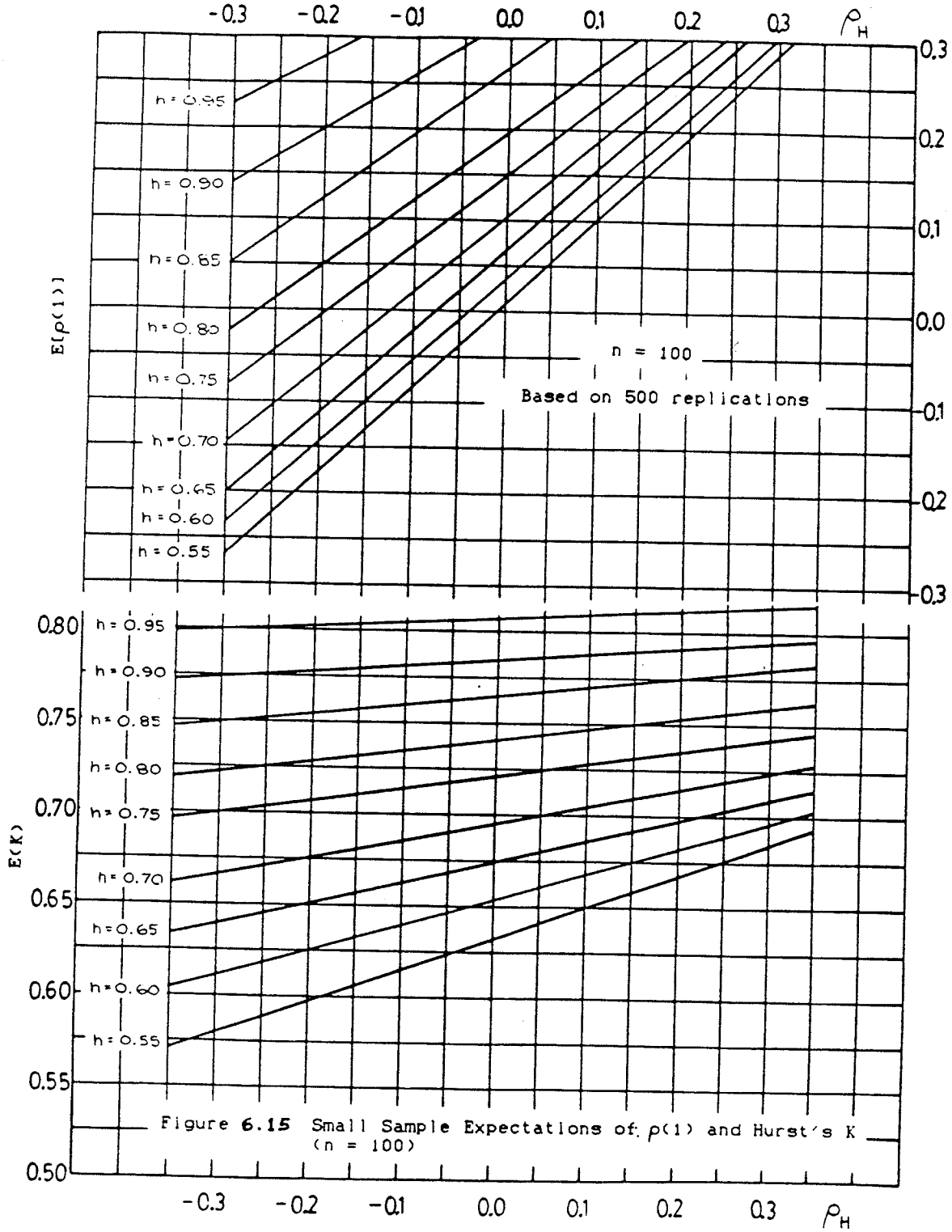


Figure 6.15 Small Sample Expectations of $\rho(1)$ and Hurst's K (n = 100)

The fitting procedure for this model is as follows:

i) Compute mean, $\bar{x}$; standard deviation, s; Hurst's K and first serial correlation coefficient, $\hat{\rho}(1)$, from the n-year historical series.

ii) Using Figures 6.13 to 6.15, find values of ρ_H and h for which:

$$E(K) = K; \quad E[\rho(1)] = \hat{\rho}(1)$$

iii) Obtain unbiased estimate of the standard deviation, S_{ub}, from Figure 6.12. This is given by:

$$S_{ub} = s / E(\sigma)$$

iv) Use S_{ub}, ρ_H, h and $\bar{x}$ in the generating equation.

To illustrate the procedure, let the standard deviation, Hurst's K and first serial correlation, estimated from a sample of size 100 be 3.0, 0.725 and 0.15 respectively.

Using Figure 6.15, with h = 0.76 and ρ_H = -0.01, we get E(K) = 0.725 and E[ρ(1)] = 0.15. From Figure 6.12, with h = 0.76, E(σ) = 0.95. Therefore, the unbiased standard deviation, S_{ub} = 3.0/0.95 = 3.16. Hence, parameters of h = 0.76, ρ_H = -0.01 and S_{ub} = 3.6, must be used in the model to generate synthetic sequences which on average produce the required sample statistics.

Skewed MN sequences may be generated by using a logarithmic transformation so that the logarithms of the flows are assumed to be normally distributed. Another way to generate skewed MN variates is by modifying the random numbers used in the generation process (Lettenmaier and Burges, 1977b). This procedure is described in Appendix H.

Although the procedure for generating skewed deviates, described in Appendix H, is fully operational, it is not used in this thesis. The simple logarithmic transformation to normality is used instead.

The mixed-noise process has several advantages. Firstly, it uses both h and $\rho(1)$ explicitly to derive its parameters. Secondly, the parameters can be estimated easily. Thirdly, it has a simple structure, and, finally it is relatively efficient when compared to the computational time of the FFGN model. It remains, however, to determine the optimum values of the lags K_1, K_2, K_3, and K_4 where the MN correlation function is forced to match FGN correlation function. In addition, a comprehensive comparison with the ARMA-Markov model and other contending models remains to be carried out.

Only the MN model will be used in subsequent studies carried out in this thesis.

6.9 SUMMARY

In this chapter, models capable of simultaneously reproducing sequences with a high Hurst coefficient and low first order correlation coefficient were reviewed. It was shown that most currently available model are either too difficult to use by practising engineers or requires excessive computer time. In addition, small sample biases for these models are largely unknown.

A new model called the Mixed-Noise model was described and was shown to have several advantages over those currently available.

In the next chapter, the effect of serial correlation on flood risk analysis will be presented.

CHAPTER 7

SERIAL CORRELATION AND FLOOD RISK ANALYSIS

7.1 GENERAL

An examination of time series with a high Hurst's coefficient shows that a sample obtained from a short record may be far less representative than a similar sample obtained from an independent series. This is due to the greater variability of the sample statistics, primarily the mean, which produces a higher variability of the flood peaks. However, when speaking of the variability of the flood peaks, it is important to define the time frame one is considering. There is no point in making pronouncements about the variability of hydrologic events on a geological time scale. It is not unreasonable, however, to assume that within the planning horizon the variability pattern of hydrologic events can be assessed by studying the pattern observed during the period of record. Both periods are in general somewhere between 50 and 100 years in length. While climatic fluctuations do occur over this period, there is no evidence of climatic change that would invalidate the basic assumption of stationarity, although in the rather weak sense that the observed pattern of variability characterized by the Hurst coefficient will persist. The engineering question then is: "What floods can one expect during the planning period

on the basis of our hydro-meteorologic experience with peak flow variability during the period of record?

It will be shown in this chapter that, even with average Hurst's K values and long term time period, the additional uncertainty in the assessment of the floods which one may expect within the planning horizon is significant. Neglecting the parameter uncertainty, as is current practice, can lead to serious underestimation of the potential flood risk.

In this chapter, the effect of short and long term serial correlation on the variability of sample statistics is first discussed. This will be followed by a demonstration of the effect of serial correlation and sample length on the flood risk for a hypothetical peak flow series. Finally, the flood risk for peak flow series for some Canadian rivers will be analysed taking into account their proper serial correlation structure. This will be compared to the flood risk based on the customary assumption of serial independence. All peak flow series are assumed to be lognormally distributed, and the discrete predictive distribution approach described in Section 4.5 will be used throughout the analysis.

7.2 EFFECT OF SHORT TERM SERIAL CORRELATION ON SAMPLE STATISTICS

This section will show that when the observations of a random variable X are serially correlated, the variance of the sample mean and sample variance are greater than that for independent observations. In addition, the bias in the estimation of the variance of X is also greater.

Let x_1, x_2, x_t, x_n, be the observed values of a stationary stochastic process. The sample mean

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} x_t \qquad \ldots (7.1)$$

is an unbiased estimate of the mean of the process, μ_x.

$$E[\bar{X}] = \mu_x \qquad \ldots (7.2)$$

However, correlation among the x_t's, so that $\rho_x(K) \neq 0$ for $K \neq 0$, affects the variance of $\bar{X}$ as follows (Loucks et al., 1981):

$$Var(\bar{X}) = E[\bar{X} - \mu_x{}^2] = \frac{1}{n^2} E\left\{ \sum_{t=1}^{n} \sum_{k=1}^{n} (x_t - \mu_x)(x_k - \mu_x) \right\}$$

$$= \frac{\sigma_x^2}{n} \left\{ 1 + 2 \sum_{K=1}^{n-1} \left(1 - \frac{k}{n}\right) \rho_x(k) \right\} \qquad \ldots (7.3)$$

The variance of $\bar{X}$, equal to $\sigma_x{}^2/n$ for independent observations, is increased by the factor within the braces. For $\rho_x(K) > 0$, as is often the case, this factor is a non-decreasing function of n, so that the variance of

$\bar{X}$ is increased by a factor whose magnitude does not decrease with increasing sample size (Loucks, et al., 1981).

For a Markov lag-one or AR(1) stochastic process, where

$$\rho_X(k) = \rho^k$$

The variance of $\bar{X}$ is given by:

$$\text{Var}(\bar{X}) = \frac{\sigma_x^2}{n}\left(1 + \frac{2\rho}{n}\,\frac{[n(1-\rho) - (1-\rho)]^n}{(1-\rho)^2}\right) \quad \ldots (7.4)$$

Table 7.1 illustrates the effect of correlation among x_t's on the standard error of their mean.

TABLE 7.1

STANDARD ERROR OF $\bar{X}$ WHEN $\sigma_x = 0.25$
AND $\rho(k) = \rho^k$ (after Loucks et al., 1981)

	Correlations of Observations		
n	$\rho = 0$	0.3	0.6
25	0.05	0.067	0.096
50	0.035	0.048	0.069
100	0.025	0.034	0.050

The properties of the estimate of the variance of X,

$$v_x^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 \qquad \ldots (7.5)$$

are also affected by correlation among the X_t's. The expected value of v_x^2 becomes:

$$E[v_x^2] = \sigma_x^2\left(1 - \frac{1}{n} - \frac{2}{n}\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_x(k)\right) \qquad \ldots (7.6)$$

If the x_t's are assumed to be generated by a Markov lag one process, then the expected value of v_x^2 is given by:

$$E[v_x^2] = \sigma_x^2\left\{1 - \frac{1}{n} - \frac{2}{n^2}\left[\frac{n\rho(1-\rho) - \rho(1-\rho^n)}{(1-\rho)^2}\right]\right\} \ldots (7.7)$$

where $\rho = \rho(1)$. If $\rho(k) = 0$ for all $|k| > 0$, then the above equation reduces to $E[v_x^2] = S_x^2$ in which case S_x^2 is an unbiased estimator of σ_x^2 for an independent process. For $\rho > 0$, the term in braces is positive and less than unity, whereby, S_x^2 tends to under estimate σ_x^2. To obtain an unbiased estimate of σ_x^2, $E[v_x^2]$ must be divided by the terms in braces where ρ is replaced by its sample estimate. Fortunately, the bias in v_x^2 decreases with n and is generally unimportant when compared to its variance (Loucks et al., 1981). Table 7.2 shows the approximate bias in v_x^2 for a Markov lag one process corresponding to Table 7.1.

TABLE 7.2

BIAS IN v_x^2 FOR MARKOV LAG ONE PROCESS

$E[v_x^2] / \sigma_x^2$

n	$\rho = 0$	0.3	0.6
25	0.96	0.9277	0.8520
50	0.98	0.9633	0.9230
100	0.99	0.9816	0.9608

Correlation among the observations also affects the variance of v_x^2. Assuming that X has a normal distribution, the variance of v_x^2 for large n is approximately (Kendall and Stuart, 1966) given by:

$$\mathrm{Var}(v_x^2) \approx 2 \frac{\sigma_x^4}{n} \left(1 + 2 \sum_{K=1}^{\infty} \rho_x^2(k) \right) \qquad \ldots (7.8)$$

where for $\rho(k) = \rho^k$, becomes

$$\mathrm{Var}(v_x^2) \approx 2 \frac{\sigma_x^4}{n} \left(\frac{1 + \rho^2}{1 - \rho^2} \right) \qquad \ldots (7.9)$$

Like the variance of $\bar{X}$, the variance of v_x^2 is increased by a factor whose magnitude does not decrease with n. This is illustrated by Table 7.3, which gives the

coefficient of variation of v_x^2 as a function of n and ρ when the observations have a normal distribution and $\rho(k) = \rho^k$.

TABLE 7.3

COEFFICIENT OF VARIATION OF v_x^2 WHEN OBSERVATIONS HAVE A NORMAL DISTRIBUTION AND $\rho(k) = \rho^k$
(after Loucks et al., 1981)

n	$\rho = 0$	0.3	0.6
25	0.28	0.31	0.41
50	0.20	0.22	0.29
100	0.14	0.15	0.21

7.3 EFFECT OF LONG TERM SERIAL CORRELATION ON SAMPLE STATISTICS

7.3.1 Fractional Noise Process

It was shown in the previous section that when the observations $(x_1, x_2, \ldots, x_n)$ follows a typical short term correlated process, the variance of the sample mean and sample variance of the random variable X are greater than that for independent observations. In addition, the bias in the estimation of the variance of X is also

greater. It will be shown in this section that when the series of observations exhibits long term serial correlation exemplified by the fractional gaussian noise process described in Chapter 6, the variance of the sample statistics are greater than that for either short term correlated or independent processes.

Mandelbrot and Wallis (1969c) showed that the sample mean or FGN is normally distributed with a mean equal to the mean of the basic random variable X. That is:

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} x_t \qquad \ldots (7.10)$$

The variance of $\bar{X}$ however, is given by:

$$\mathrm{Var}(\bar{X}) = \sigma_x^{\,2}\, n^{2h-2} \qquad \ldots (7.11)$$

where $\sigma_x^{\,2}$ is the true variance of X and h is the Hurst coefficient (see Appendix G). This equation reduces to the well known $\sigma_x^{\,2}/n$ with h = 0.5 (independent data). Table 7.4 shows the effect of long term serial correlation on the standard error of X for various values of n and h.

It can be seen that with a typical h value of 0.7, and n of 50, the uncertainty in the sample mean is almost twice as large when the data are independent.

TABLE 7.4

STANDARD ERROR OF $\bar{X}$ FOR FGN PROCESS
WHEN $\sigma_X = 0.25$

	h = 0.5	0.6	0.7	0.8	0.9
n = 25	0.5	0.069	0.095	0.131	0.181
50	0.035	0.052	0.077	0.114	0.169
100	0.025	0.040	0.063	0.100	0.158

The expected value of v_x^2 for an FGN process is given by:

$$E[v_x^2] = \sigma_x^2 (1 - n^{2h-2}) \quad \ldots (7.12)$$

where σ_x and h were defined above. (7.12) reduces to the familiar Bessel bias correction for small samples when the observations are independent, that is h = 0.5.

Table 7.5 shows the bias in v_x^2 for FGN for various values of n and h. It can be seen that the bias in the estimate of the sample variance for an FGN process is greater than for independent data and that for the Markov lag one process for typical values of h and ρ.

TABLE 7.5

BIAS IN v_x^2 FOR FGN

	h = 0.5	0.6	0.7	0.8	0.9
n = 25	0.96	0.9239	0.8550	0.7241	0.4747
50	0.98	0.0563	0.9044	0.7909	0.5427
100	0.99	0.9749	0.9369	0.8415	0.6019

It is known that the variance of v_x^2 is affected by serial correlation. However, no analytical expression could be found in the literature for the variance of v_x^2 for the FGN process nor could the appropriate type of distribution of the sample variance be determined analytically.

For mixed processes in which both short and long term serial correlation are present as in peak flow series, the probability distribution of the sample statistics cannot be easily determined analytically. The approximate probability distribution, however, can be obtained using the Monte Carlo method.

7.3.2 Probability Distribution Of Sample Statistics By Monte Carlo Method

Using Monte Carlo methods to obtain the probability distribution of sample statistics requires a time series model that produces "on average" samples with the same statistics as the observed sample. The mixed-noise model described in Chapter 6, will be used here. The sample statistics, after correcting for small sample bias, are used as inputs to the model to generate sample time series of a specified length. The sample length to be generated is usually the same length as the sample observations. A large number (say 500) of these samples is then generated, and, for each sample, the mean, standard deviation or any other statistic can be calculated. From the 500 data available for each statistic, an approximate probability distribution can be determined. For example, assume the distribution of the mean is of interest. From the 500 samples each of length n, 500 estimates of the mean are available to construct an approximate probability distribution for the mean. The mean of the mean and the standard deviation of the mean which measures the uncertainty in estimating the mean, can then be obtained. This uncertainty is dependent on the sample length and the strength and nature of the serial correlation structure.

7.4 EFFECT OF SERIAL CORRELATION AND SAMPLE LENGTH ON FLOOD RISK ANALYSIS

In this section, the effect of serial correlation and sample length will be demonstrated for a hypothetical lognormally distributed peak flow series.

The effect of sample length for a serially independent series is shown in Figure 7.1. It can be seen that even for moderately sized samples, parameter uncertainty is very small for serially independent data.

However, if the peak flow series is characterized by a mixed-noise process with an average Hurst's K value of 0.70 and a lag-one serial correlation coefficient of 0.2, there is some difference between the observed frequency curve (independent data with no parameter uncertainty) and the predictive distributions (risk curves) for records that are about 25 years long. This is shown in Figure 7.2.

With a K value of 0.75, the difference between the observed frequency curve and the risk curves is quite substantial, even for records of 50 and 100 years. This is shown on Figure 7.3.

A further comparison between the risk values (probability of exceedence) obtained for serially correlated peak flow data and serially independent data for sample length of 100, is shown in Table 7.6. The peak

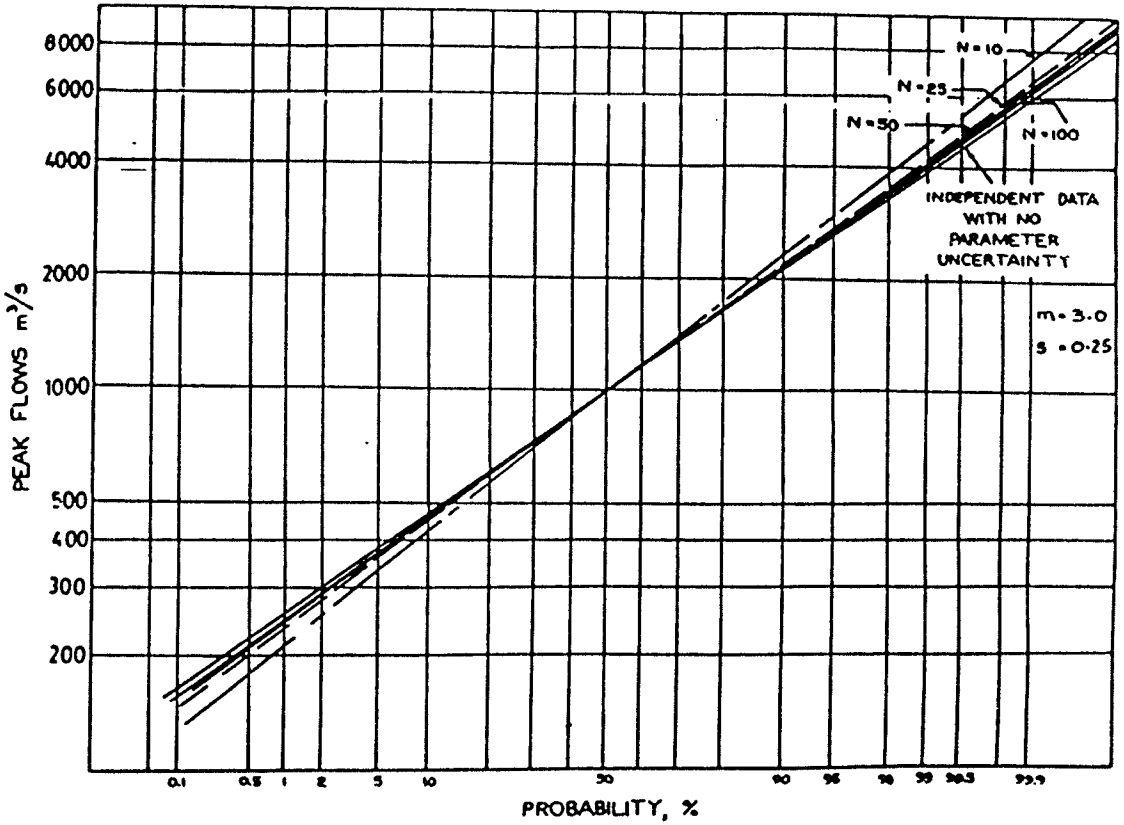


Figure 7.1 Risk Curves for a Log-Normally Distributed Serially Independent Variable of Length n = 10, 25, 50 and 100 Years

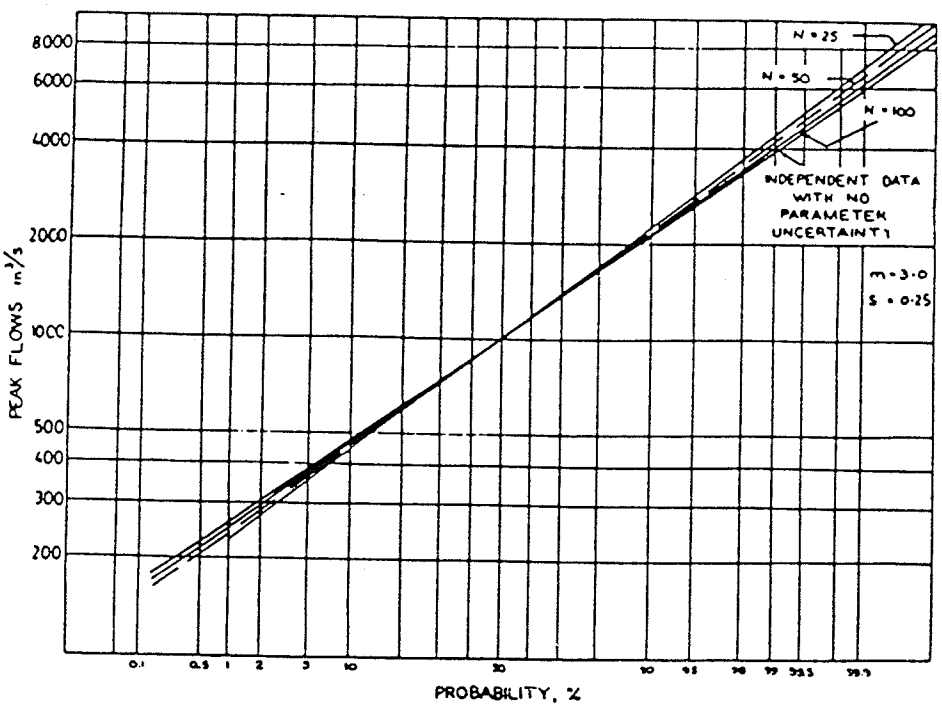


Figure 7.2 Risk Curves for a Log-Normally Distributed Mixed-Noise Process (K = 0.70, ρ(1) = 0.20) of Length n = 25, 50 and 100 Years

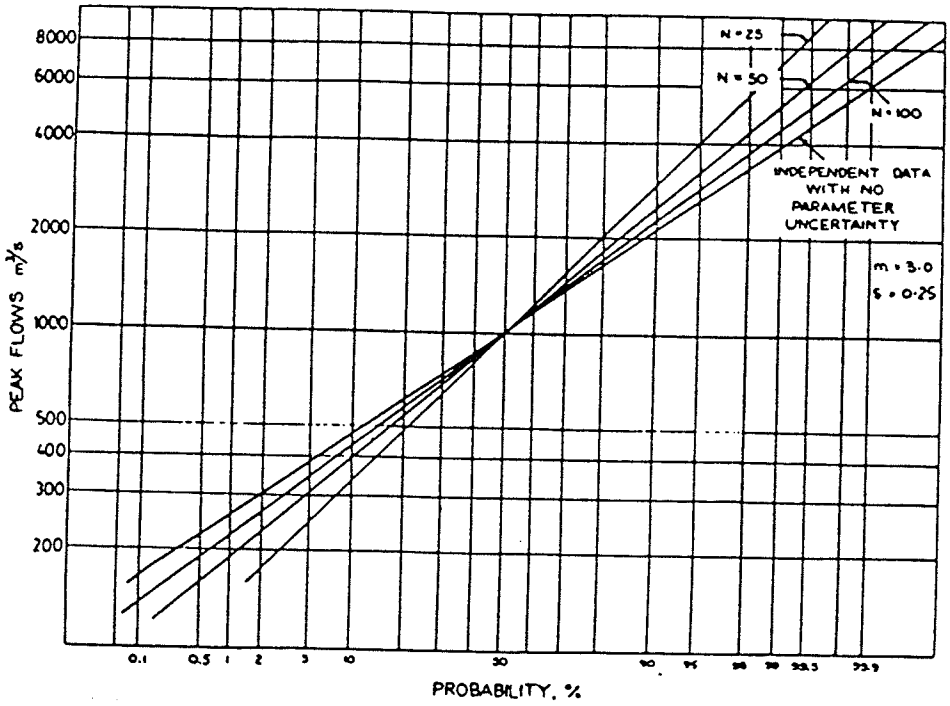


Figure 7.3 Risk Curves for a Log-Normally Distributed Mixed-Noise Process (K = 0.75, ρ(1) = 0.20) of Length n = 25, 50 and 100 Years

TABLE 7.6

EFFECT OF SERIAL CORRELATION ON FLOOD RISK (N = 100)

Prob(X $\geq$ q), $M_{\log_{10}x} = 3.0$, $S_{\log_{10}x} = 0.25$

(1)	(2)	(3)	(4)	(5)	(6)
q m^3/s	$\rho_{(1)}=0.2$ K=0.75	$\rho_{(1)}=0.2$ K=0.70	$\rho_{(1)}=0.2$	Independent data with parameter uncertainty	Independent data without parameter uncertainty
1000	0.5	0.5	0.5	0.5	0.5
1200	0.38675	0.37782	0.37607	0.37578	0.37573
2000	0.13761	0.11916	0.11577	0.11515	0.11427
3000	0.04255	0.03149	0.02966	0.02929	0.02816
4000	0.01524	0.00978	0.00896	0.00877	0.00802
5000	0.00616	0.00348	0.00311	0.00302	0.00259
6000	0.00274	0.00138	0.00121	0.00117	0.00093
7000	0.00132	0.00060	0.00052	0.00049	0.00036
8000	0.00068	0.00028	0.00024	0.00022	0.00015

Column (1): Future flow q m^3/s.

Column (2): Mixed-Noise process with $\rho(1) = 0.2$ and K = 0.75.

Column (3): Mixed-Noise process with $\rho(1) = 0.2$ and K = 0.70.

Column (4): Markov process $\rho(1) = 0.2$.

Column (5): Independent data with parameter uncertainty.

Column (6): Independent data without parameter uncertainty.

flows are assumed to be lognormally distributed. Table 7.7 shows the situation when the sample length is 50.

It can be seen that even for a sample length of 100, the probability that a future flow of 6000 m^3/s (say) is equalled or exceeded in any year during the period bounded by the planning horizon is seriously affected by serial correlation. For that flow, the risk is increased by a factor of 3 if the peak flows are derived from a mixed-noise process with K = 0.75 instead of the customary assumption of serial independence with no parameter uncertainty. For a sample length of 50, the risk is increased by a factor of 6.

These results illustrate three points:

i) If there is sufficient reason to believe that in the time frame bounded by the planning horizon the annual peak flows are serially independent or show at most short term serial correlation, then the predictive distribution seems a simple and attractive way of producing a safety factor to account for the parameter uncertainty in short records.

ii) If there is evidence of long term serial correlation leading to a K value in excess of 0.70, then the parameter uncertainty is substantial even when long records are available. The nature and the effect of the serial correlation should then be made the subject of a special study.

TABLE 7.7

EFFECT OF SERIAL CORRELATION ON FLOOD RISK (N = 50)

Prob(X $\geq$ q), $M_{\log_{10}x} = 3.0$, $S_{\log_{10}x} = 0.25$

(1)	(2)	(3)	(4)	(5)	(6)
q m^3/s	$\rho_{(1)}=0.2$ K=0.75	$\rho_{(1)}=0.2$ K=0.70	$\rho_{(1)}=0.2$	Independent data with parameter uncertainty	Independent data without parameter uncertainty
1000	0.5	0.5	0.5	0.5	0.5
1200	0.39511	0.37826	0.37653	0.37594	0.37573
2000	0.15687	0.12091	0.11755	0.11617	0.11427
3000	0.05624	0.03320	0.03133	0.03040	0.02816
4000	0.02321	0.01088	0.00100	0.00950	0.00802
5000	0.01071	0.00412	0.00370	0.00344	0.00259
6000	0.00539	0.00174	0.00154	0.00140	0.00093
7000	0.00291	0.00081	0.00070	0.00062	0.00036
8000	0.00166	0.00040	0.00034	0.00030	0.00015

Column (1): Future flow q m^3/s.

Column (2): Mixed-Noise process with $\rho(1) = 0.2$ and K = 0.75.

Column (3): Mixed-Noise process with $\rho(1) = 0.2$ and K = 0.70.

Column (4): Markov process $\rho(1) = 0.2$.

Column (5): Independent data with parameter uncertainty.

Column (6): Independent data without parameter uncertainty.

iii) Parameter uncertainty appears to be a more important issue than the questions of plotting positions, parameter estimation by method of moments or maximum likelihood, or the choice between the standard 2-parameter and 3-parameter models.

In the next section, flood risk analysis for some Canadian peak flow series of Canadian rivers are presented.

7.5 FLOOD RISK ANALYSIS FOR SOME CANADIAN RIVERS

In this section, the flood risk for some Canadian peak flow series will be analysed taking into account their proper correlation structure. This will be compared to the flood risk obtained based on the customary assumption of serial independence.

Five peak flow series are used as examples. Each peak flow series is approximately lognormally distributed and the statistics of each river is shown in Table 7.8. The five rivers are: the Bow River at Banff, the Red River at Emerson, the Roseway River at Lower Ohio, the South Thomson River near Chase, and Slocan River near Cresent Valley.

The probability distribution of the mean and standard deviation is then obtained for each peak flow series using the Monte Carlo method described in Section 7.3.2.

TABLE 7.8

SUMMARY STATISTICS OF PEAK FLOW SERIES

River	N	$\hat{\mu}$	$\hat{\sigma}$	K	$\rho(1)$	S_μ	S_σ
Bow River at Banff	76	5.37	0.256	0.66	-0.08	0.0563	0.0205
Red River at Emerson	70	6.28	0.762	0.75	0.17	0.4541	0.0824
Roseway River at Lower Ohio	67	4.17	0.356	0.74	0.08	0.2141	0.0401
Slocan River near Cresent Valley	60	6.09	0.270	0.73	0.14	0.1301	0.0291
South Thomson River near Chase	48	6.88	0.223	0.70	0.17	0.0606	0.0240

Note: N = record length (years)

$\hat{\mu}$ = sample mean ($\log_e$)

$\hat{\sigma}$ = sample standard deviation ($\log_e$)

K = Hurst's coefficient K

$\rho(1)$ = lag-one serial correlation coefficient

S_μ = standard deviation of mean $\hat{\mu}$

S_σ = standard deviation of standard deviation $\hat{\sigma}$.

Since the probability distributions for both the mean and standard deviation are approximately normally distributed, only the standard deviation of the mean and standard deviation of the standard deviation are necessary to characterize the uncertainty in the estimation of the mean and standard deviation. These are also summarized in Table 7.8.

The resulting predictive distributions (risk curves) for each of the rivers are shown in Figures 7.4 - 7.8. Also plotted on each of the figures are the frequency curves obtained by fitting a lognormal distribution to the observed data. One can see that the difference between the predictive distribution and the descriptive distribution can be substantial for rivers with large Hurst's K values, e.g. Red River and Roseway River.

7.6 SUMMARY

The object of probability analysis is to quantify the variability of the peak flows one may expect in the period bounded by the planning horizon. The only information available for pure statistical analysis is the peak flow records. A high Hurst coefficient means that there are low frequency components in the serial correlation structure of the observed time series. Although the information about the nature of these low

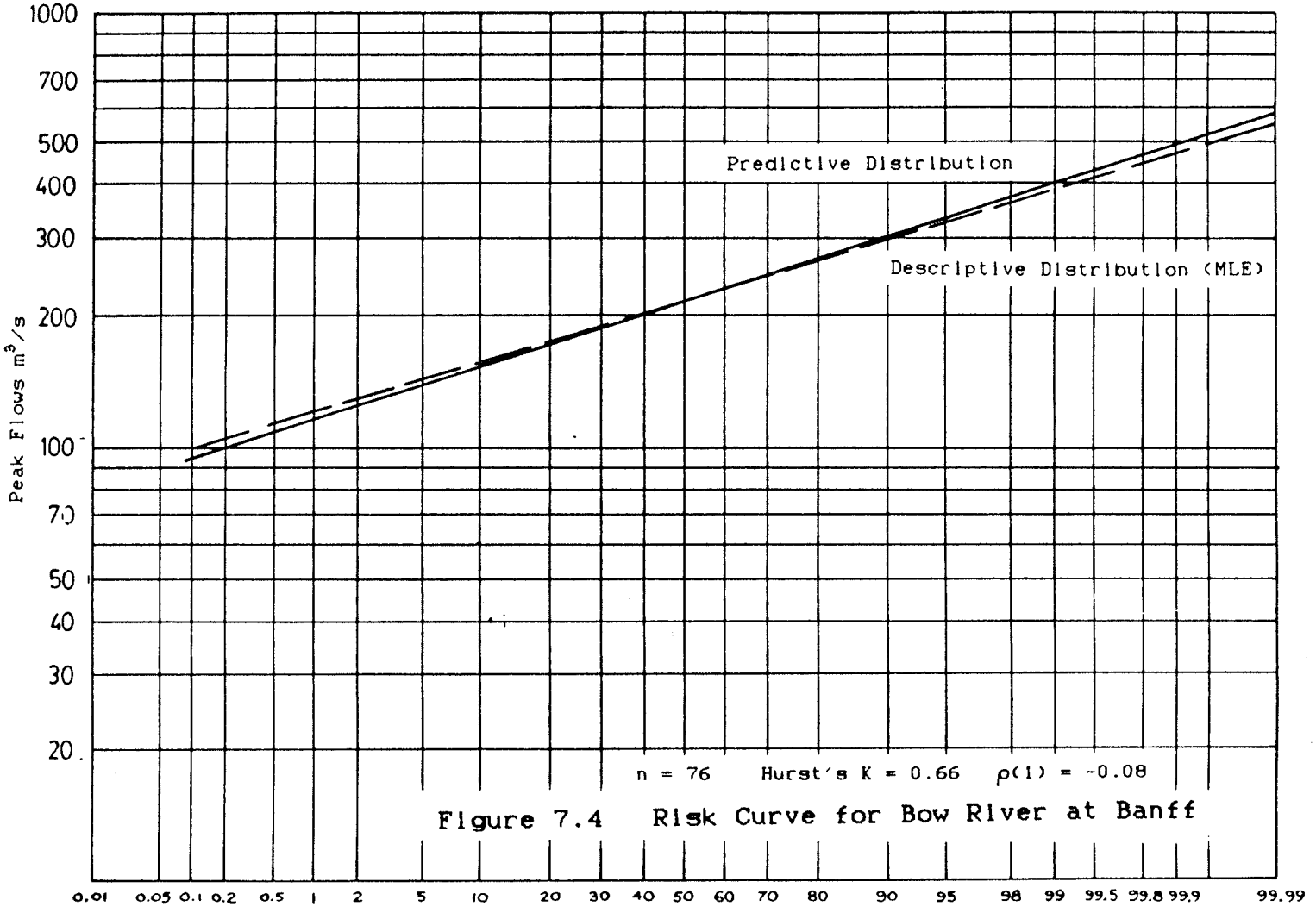


Figure 7.4 Risk Curve for Bow River at Banff

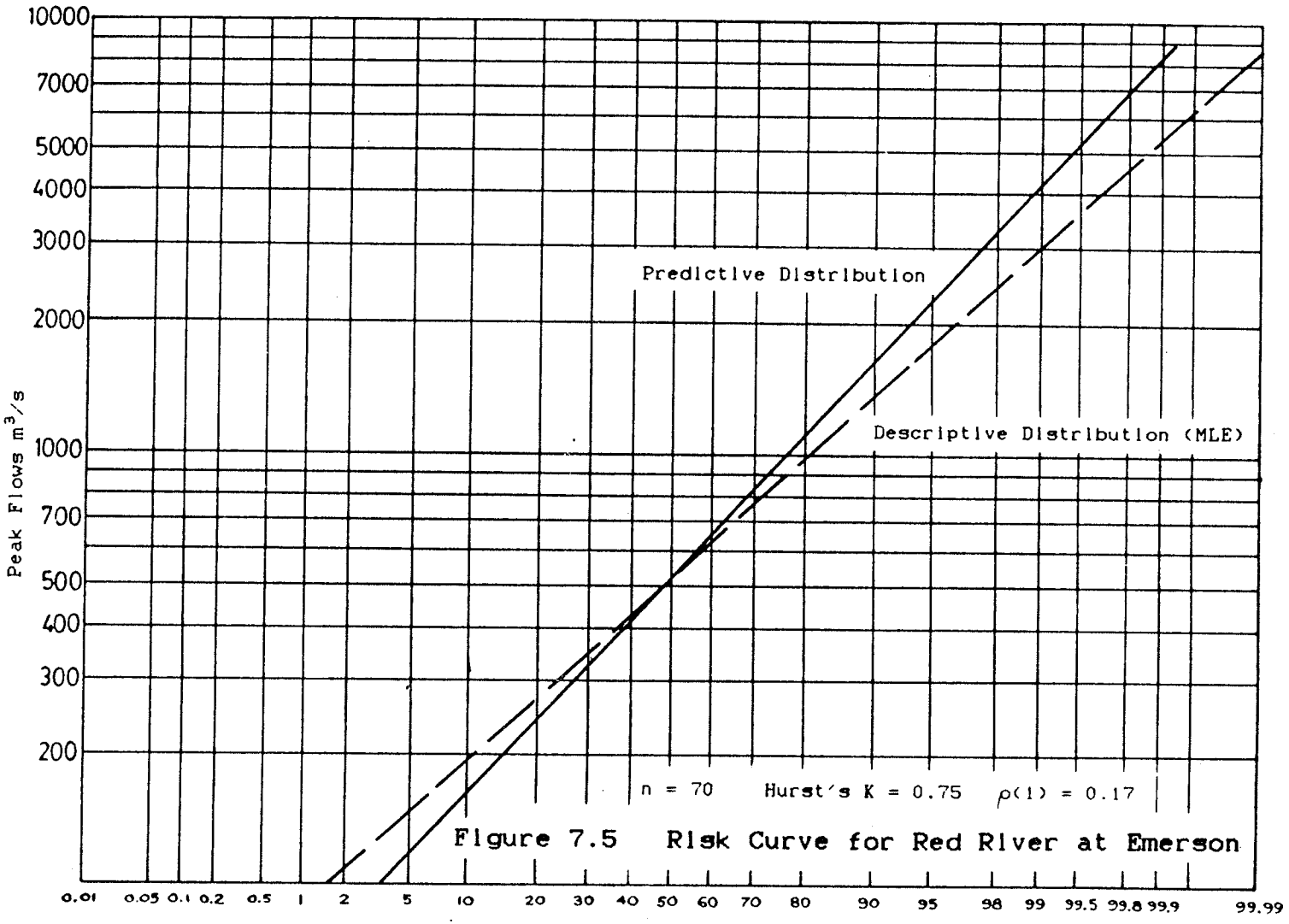


Figure 7.5 Risk Curve for Red River at Emerson

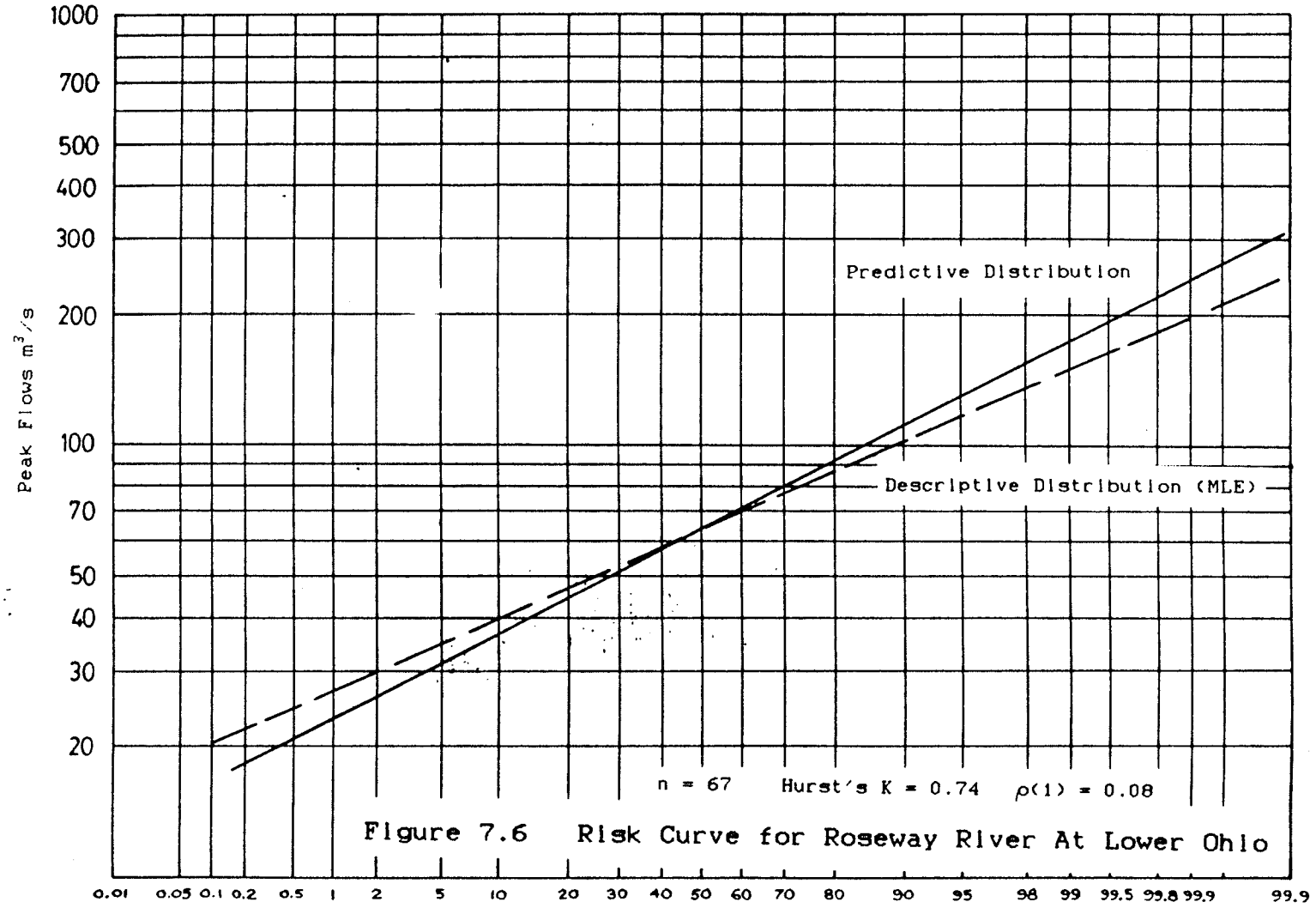


Figure 7.6 Risk Curve for Roseway River At Lower Ohio

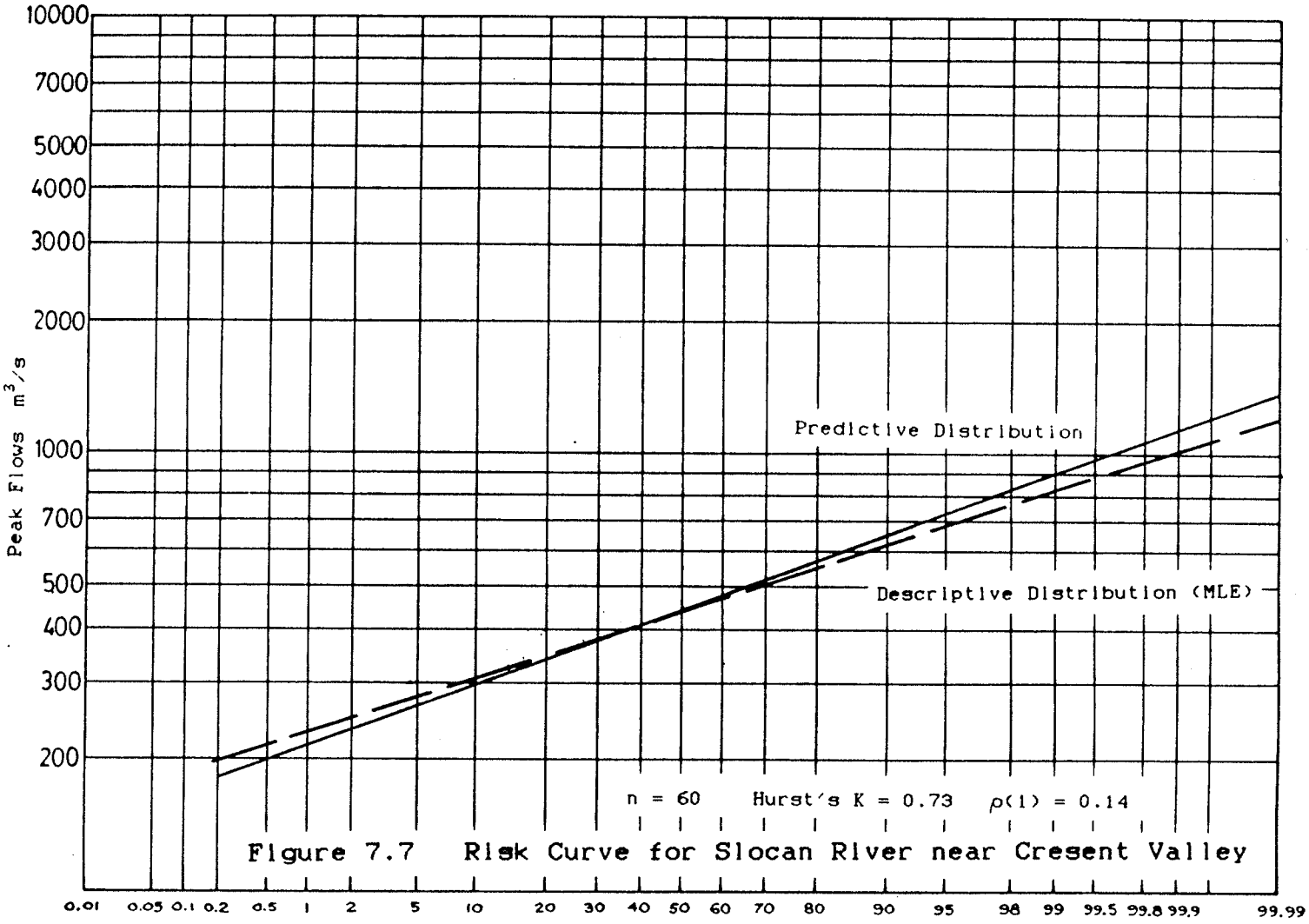


Figure 7.7 Risk Curve for Slocan River near Cresent Valley

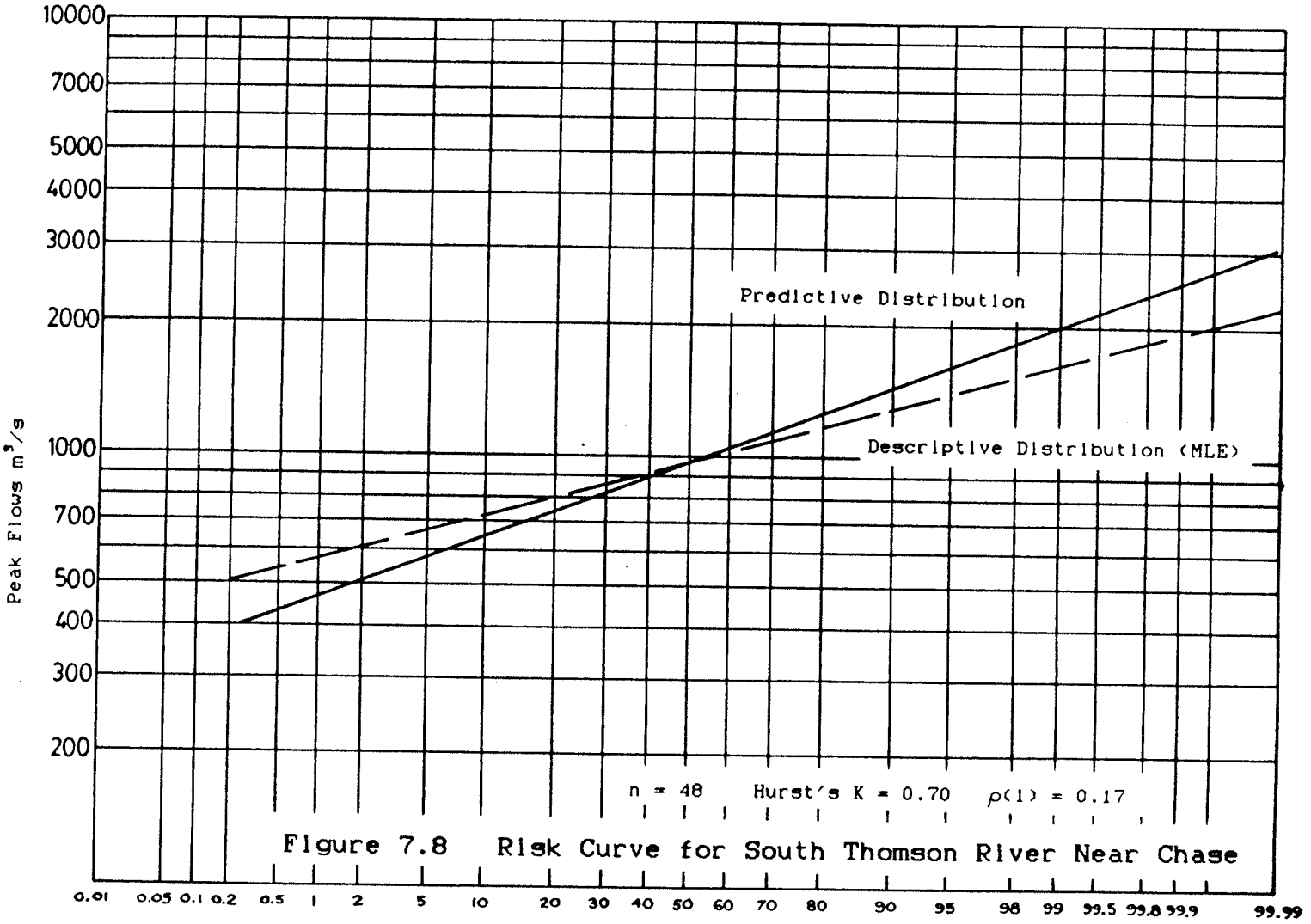


Figure 7.8 Risk Curve for South Thomson River Near Chase

frequency components is little, the information should not be discarded and denied by our probability analysis, as it may drastically affect the degree to which the peak flow sample provided by the record is representative of the population. This chapter has shown that taking into account the parameter uncertainty which is aggravated by serial correlation, significantly increases the risk associated with future peak flows.

CHAPTER 8

STOCHASTIC PEAK FLOW SIMULATION MODEL

8.1 GENERAL

In the previous chapters, the probability distribution of which the recorded annual peak flow series is considered to be a sample was determined 'a priori'. The choice of the probability distribution is usually based on considerations of parsimony of parameters and goodness of fit of observed data. The chosen probability distribution may be quite well supported in the middle reach where many observations are available; in the upper tail, upon which flood protection decisions are based, there is much uncertainty due to the lack of data. Moreover, as demonstrated in the previous chapter, the observed serial correlation structure of the peak flows may add substantially to this uncertainty.

One can reduce the uncertainty in the risk analysis of extreme events such as peak flows only by obtaining additional information. In principle, such additional information about the type of probability distribution and the magnitude of its parameters can be obtained by a statistical analysis of the physical factors, which as jointly distributed random variables determine the severity of the extreme event. This possibility is investigated in this chapter.

This investigation requires the development of a physically-based stochastic simulation model that takes as input the main hydro-meteorologic factors that determine the magnitude of the extreme event and that produce as output synthetic extreme event series statistically similar to the observed series.

The advantage in developing such a stochastic simulation model is that it allows for a full range of possible interactions of the input variables in that the output of the model reflects not only combinations that have occurred in the past, but also combinations that did not occur but that are equally probable, thereby providing more information. In addition, the simulation model provides the probability distribution function of the output variable.

The simulation model should, therefore, provide a better basis for a probabilistic assessment of the extreme event than a conventional extrapolation of the assumed probability distribution.

A disadvantage of the stochastic simulation model is that considerable simplification of physical and statistical relationships are necessary which introduces error in the model. Whether the gain in accuracy is positive can only be assessed by building and analysing the models.

This chapter describes an attempt to develop such a physically-based stochastic simulation model for the

generation of annual spring peak flows on the Red River at Emerson, Manitoba.

8.2 MODELLING METHODOLOGY

This section outlines the assumptions and the necessary steps involved in the construction and subsequent use of the stochastic peak flow simulation model for risk assessment.

Stochastic modelling of a process is, in principle, quite simple provided one can avoid or eliminate cross-correlation. Given the distribution functions of the input variables, values of the input variables are generated using values of random variables having the specified distribution functions and used to compute a single value of the output variable. This process is repeated until a sufficient number of values of the output variables have been generated to define the distribution function of the output variable.

The construction of the stochastic simulation model involves the following steps:

i) Identify the "contributing" or input variables.

ii) Determine the probability distribution functions of the input variables.

iii) Analyse and model the serial correlation structure of the input variables.

iv) Check the cross-correlations of the contributing variables.

v) Identify the physical-statistical relationship between the input variables and the output variable.

vi) Generate values of the output variable from generated values of the input variables and the relationship identified in (v).

vii) Check for statistical resemblance between the generated and the observed spring peak flow series making adjustments if necessary.

viii) Obtain the probability distribution function of the output variable.

ix) Obtain the predictive distribution function of the sping peak flows by sampling from the predictive distributions of the input variables.

x) Compare the predictive distribution function obtained from (ix) to that derived only from the observed data.

The contributing variables in step (i) are the hydro-meteorological input data. They may be used as observed or in the form of functions in which the effect of more than one variable is aggregated. In their effect, the contributing variables must simulate the main physical factors that determine the magnitude of the annual peak flows.

In step (ii), the probability distribution of each of the contributing variables is determined. It is expected that some of the contributing variables may be skewed.

A time series of observations corresponds to each contributing variable. In step (iii) the serial correlation structure of the time series is analysed. The analysis would involve the calculation of the serial correlation coefficients and the Hurst's coefficient in addition to the marginal distribution parameters such as the mean and standard deviation. In this step the appropriate time series model required to reproduce the observed serial correlation structure of each of the contributing variables would be identified and tested for adequacy.

In step (iv), the cross-correlations among the contributing variables must be checked. If significant cross-correlations between any two contributing variables are found, then it may be necessary to express one contributing variable in terms of the other to eliminate multicollinearity problems.

In step (v), the statistical relationship between the contributing variables and the annual spring peak flows is determined. The relationship must reflect the physical runoff process. That is, the proper interaction among the contributing variables and the manner in which these contributing variables affect the spring peak flows

must be correctly represented in the statistical model. It is expected that the statistical model would be nonlinear and the parameters of the model have to be determined from the observed data using appropriate statistical techniques.

With the time series model for each contributing variable indentified in step (iii), and the statistical model relating the spring peak flows to the contributing variables determined in step (v), generation of spring peak flow series can proceed. This is carried out in step (vi). A time series of each contributing variable is generated and combined using the statistical model to give one series of the spring peak flows. This process is repeated until a large number of independent series of the spring peak flows, each of length equal to the observed sample, are generated. Alternatively, one could generate only one very long series. But that would make the incorporation of the parameter uncertainty of the distribution function of the contributing variables due to short sample length impossible.

In step (vii) the statistical resemblance between the generated spring peak flow series and the observed series is tested. Statistical resemblance is achieved when the mean, standard deviation, serial correlation coefficients and the Hurst coefficient that are calculated from the observed series are equal to the expected values of these statistics from the large number of independently

generated series. Adjustments to the parameters of the statistical model may be necessary to ensure that statistical resemblance is achieved between the spring peak flows generated from the simulation model and those in the observed series.

The probability distribution function of the spring peak flows from generated data is determined in step (viii). A simple way of obtaining this distribution function is to calculate the relative frequency that a given flow is exceeded or equalled based on a large number of data. The number of data to be generated should be large enough to ensure that the distribution function obtained is well defined. A set of flows can be defined, and the calculated relative frequency or probability that each flow is equalled or exceeded can be plotted on probability graph paper to give the approximate distribution function of the spring peak flows. The probability distribution function obtained would be descriptive distribution function since the parameter uncertainty of the inputs distribution function has not been taken into account yet.

In step (ix), the predictive distribution function of the spring peak flows is determined. This is obtained by sampling from the predictive distribution function of each of the contributing variables. However, it is assumed that the parameters of the statistical model that relate spring peak flows to the contributing variables

are known and fixed. Uncertainties related to the determination of these parameters will not be incorporated in the model since its effect is considered to be very small compared to other variabilities.

The last step of the process is the comparison of the predictive distribution function of the spring peak flows obtained from the stochastic simulation model to the predictive distribution function of the observed data based on a assumed underlying probability distribution function. It can, therefore, be expected that the uncertainty of the predictive distribution function obtained from the simulation model to be less than that derived only from the observed data. This will indicate that a more reliable estimate of the flood risk can be made by using the predictive distribution obtained from the simulation model.

Using the methodology and assumptions set out in this section, a specific stochastic simulation model for the annual spring peak flows on the Red River at Emerson will be constructed. This is described in the following sections.

8.3 IDENTIFICATION AND MODELLING OF CONTRIBUTING VARIABLES

Spring peak flows on the Red River are caused by a combination of four primary hydro-meteorological variables (RRBI, 1953):

i) The basin storage conditions prior to the runoff.

ii) The snow pack that has accumulated over the winter.

iii) The rate at which the snow pack is melted.

iv) The spring precipitation coinciding with the melting snow.

Other factors such as vapor pressure, evaporation, depth of frost penetration, wind, etc., may also affect the magnitude of the spring peak flows. The effect of these factors on spring peak flows however are quite minor and are highly correlated to the four main factors. As such they can be absorbed in the other terms and in the random error term of the simulation model. The description and modelling of each of the four contributing variables is discussed next.

8.3.1 Basin Storage Condition

The first of the four contributing variables is the basin storage condition which is primarily in the form of soil moisture, prior to the runoff. Basin storage sets the stage for the flood event at hand; a high basin storage condition reduces the infiltration rate which increases the surface runoff component leading to a higher peak flow.

Soil moisture is seldom measured directly. Hydrologists, therefore, must resort to indicators. Two commonly used indicators are the baseflow rate in a river prior to the runoff and the antecedent precipitation index (API). Both have been found useful in multi-regressive type prediction models. Neither indicator, however, is suitable for probabilitistic analysis using simulation since both have a complicated interdependency with the probabilistic inputs that are required in a simulation study. For this reason, the "accumulated basin storage" (ABS), a physically-based parameter that could be used in a probabilistic flood analysis was developed (Booy and Lye, 1985, 1986). In this study, the ABS primarily quantifies the average soil moisture conditions for the Red River drainage basin upstream of Emerson, Manitoba.

The ABS has been compared to the API and the baseflow rate for the purpose of real time flood prediction (Booy and Lye, 1985). The results for the

three methods were rather similar. However, from the stochastic simulation point of view, the physically-based ABS is superior.

The ABS can be written as:

$$ABS_t = ABS_{t-1} + WP_{t-1} + SP_t - C \cdot U_t - R_t \quad \ldots (8.1)$$

where ABS is the accumulated basin storage, WP is the winter precipitation, SP is the summer precipitation, C the coefficient of evapotranspiration, U the potential evapotranpiration, and R is the annual runoff at Emerson. In this equation all terms are expressed in millimetres as spatial averages over the drainage basin. With regard to notation, the water year with index t refers to the period from April 1st of the year t to March 31st of year t+1. When attached to process parameter such as summer precipitation, the index t refers to the water year t in which the process takes place. When attached to a condition parameter, that is, the ABS, the index t refers to the end of the water year t. A distinction between summer and winter precipitation must be made. Summer precipitation is all that precipitation that falls between April 1st and October 31st of a given year. The precipitation in the rest of the water year is designated as winter precipitation. For simplicity it is assumed that none of the winter precipitation of year t contributes to the ABS until it melts in water year t+1.

The terms WP, SP and R are obtained from the historical records. Seventy years of record are available from 1915 to 1984. These are given in Table 8.1.

The initial value of the ABS and the evapotranspiration term C.U are estimated by a fitting procedure described in detail by Booy and Lye (1986). Figure 8.1 shows the final plot of the ABS on April 1st for the period of record.

The ABS as a factor in the correlation structure of the annual spring peak flows on the Red River at Emerson was also investigated by Booy and Lye (1986). They found that the auto-correlation of the ABS affected the low frequency grouping in the peak flows thereby producing a high Hurst coefficient, as well as, the low lags serial correlation coefficients. Therefore, the soil moisture fluctuations as measured by the ABS explain in part the observed serial correlation structure of the spring peak flows at Emerson.

The ABS equation in the form of (8.1) cannot, however, be used for simulation purposes because the runoff term R and the potential evapotranspiration term U are correlated with other terms. Booy and Lye (1986) found that a workable simulation model for the ABS series can be obtained by absorbing the runoff and the evapotranspiration terms into other terms in the equation. Using this procedure, the simulation model equation for the ABS series can be written as:

Table 8.1 Hydro-meteorological Data - Red River at Emerson

OBS	YR	PF	WP	SP	ABS	MI	TP	MP	CP
1	1915	285	69	508	59	5.50	44	0.4	3.2
2	1916	1310	132	469	99	11.75	55	23.6	29.2
3	1917	733	85	518	229	4.67	43	0.0	1.2
4	1918	141	50	289	124	10.42	34	0.0	0.0
5	1919	312	97	407	103	5.17	43	1.8	37.8
6	1920	756	124	488	187	2.67	45	0.0	0.0
7	1921	362	85	385	177	12.00	46	13.6	2.4
8	1922	535	107	469	223	7.67	45	7.2	12.6
9	1923	736	148	376	193	1.75	56	22.4	9.2
10	1924	179	53	399	239	2.25	52	12.0	24.8
11	1925	428	59	477	298	2.67	33	0.0	0.4
12	1926	227	56	467	331	10.25	32	0.0	0.0
13	1927	580	91	403	303	4.08	45	15.8	23.0
14	1928	476	76	454	348	5.00	37	0.0	0.0
15	1929	544	85	453	379	1.08	32	0.0	0.0
16	1930	589	89	303	277	0.83	41	1.2	0.0
17	1931	225	99	365	227	2.50	41	0.0	2.4
18	1932	535	78	413	234	3.58	46	14.4	7.6
19	1933	311	97	388	205	7.25	40	0.0	23.8
20	1934	136	86	299	90	2.75	44	8.2	8.4
21	1935	155	112	324	9	2.17	33	0.0	2.4
22	1936	510	109	422	59	6.75	52	1.2	0.4
23	1937	165	87	250	19	6.00	68	88.0	32.0
24	1938	213	72	485	96	8.25	27	0.0	0.0
25	1939	190	81	369	26	3.67	41	0.0	6.6
26	1940	413	46	349	3	0.33	52	15.2	0.6
27	1941	787	89	401	0	5.00	47	32.6	27.4
28	1942	790	82	619	191	1.50	41	0.0	14.0
29	1943	835	109	450	229	13.83	51	0.0	0.0
30	1944	348	50	451	259	3.50	50	1.8	4.2
31	1945	833	131	566	348	18.83	35	0.0	6.6
32	1946	682	91	384	369	12.00	36	0.0	10.6
33	1947	804	66	435	387	3.08	59	62.4	1.4
34	1948	1470	117	467	395	10.08	58	36.6	26.6
35	1949	827	104	380	352	5.08	46	0.0	0.0
36	1950	2670	122	481	409	3.50	75	92.4	26.0
37	1951	753	86	440	406	7.75	46	0.0	5.2
38	1952	685	85	389	376	8.75	55	3.4	2.8
39	1953	255	74	342	266	2.25	36	0.0	4.0
40	1954	326	75	509	337	4.33	48	3.2	3.2
41	1955	680	67	404	332	5.42	41	0.0	4.0
42	1956	957	102	430	298	3.08	57	24.6	1.6
43	1957	351	68	455	347	10.50	60	22.0	2.0
44	1958	174	52	515	367	8.17	37	0.0	3.0
45	1959	445	80	340	251	8.42	41	4.0	0.0
46	1960	864	58	477	311	2.75	48	23.6	7.4
47	1961	122	54	398	245	5.67	31	0.0	0.0
48	1962	946	97	395	188	2.17	56	10.0	1.4
49	1963	391	50	528	268	7.17	44	38.8	4.8
50	1964	481	62	385	164	9.17	52	21.0	33.0
51	1965	1310	64	511	219	7.92	56	73.8	2.6
52	1966	1890	158	528	334	7.25	44	0.0	2.8
53	1967	951	88	417	362	8.25	40	0.0	12.0
54	1968	244	81	348	278	3.92	34	0.0	0.0
55	1969	1550	107	548	400	15.92	57	28.6	1.2
56	1970	1120	89	406	382	3.33	61	39.4	6.4
57	1971	753	99	447	378	7.08	47	2.2	0.0
58	1972	869	85	503	458	3.17	54	32.8	7.0
59	1973	416	63	386	403	1.33	27	0.0	0.0
60	1974	1230	85	444	400	11.42	60	44.6	43.8
61	1975	1210	105	440	402	11.25	69	70.4	10.0
62	1976	932	92	490	456	3.75	38	0.0	3.6
63	1977	130	86	249	268	4.75	43	3.6	0.4
64	1978	1430	109	501	344	7.42	49	14.4	1.6
65	1979	2630	124	403	316	12.67	61	51.6	29.0
66	1980	614	78	410	314	4.58	40	0.0	0.0
67	1981	107	64	386	262	2.33	32	0.0	0.0
68	1982	966	104	492	319	0.08	48	11.0	3.8
69	1983	732	94	478	393	1.83	40	0.0	8.6
70	1984	855	88	435	416	3.42	40	0.0	0.0

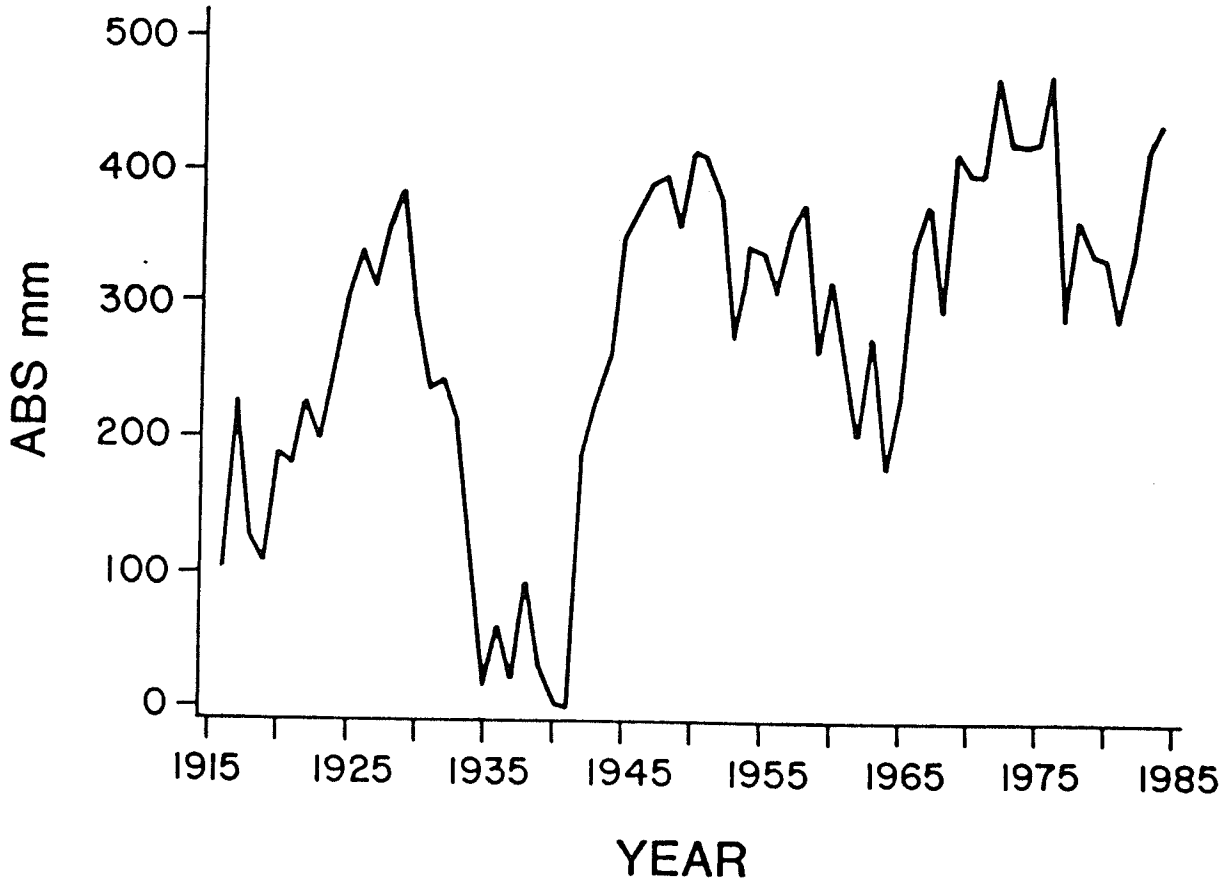


Figure 8.1 Accumulated Basin Storage Time Series

$$ABS_t = c_1 ABS_{t-1} + c_2 WP_{t-1} + c_3 SP_t + e_t \quad \ldots (8.2)$$

where $c_1 = 0.87$, $c_2 = 0.76$, $c_3 = 0.83$, and e_t is approximately normally distributed with a mean of -390 mm and a standard deviation of 14.0 mm. The terms in (8.2) are statistically independent of each other, and $R^2 = 0.97$.

To generate values of ABS, one needs to generate values of SP, WP and the error terms e. This can be done as long as the distribution function and the serial correlation structure of these terms are known. It was found that the summer precipitation, SP, and the random error term, e, are approximately normally distributed. The winter precipitation, WP, however requires a square-root transformation to produce an approximately normally distributed data set. The Hurst coefficient and the first serial correlation together with the mean and standard deviation of the terms in (8.2) are shown in Table 8.2. The appropriate parameter values for time series model for WP, SP and e are also shown in Table 8.2.

The fitting procedure and the determination of the input parameters and bias correction for the Mixed-Noise model has been described in detail in Section 6.8. Summer precipitation (SP) series, for example, can then be generated as:

$$SP_t = 428 + 73.8 * zs_t \quad \ldots (8.3)$$

TABLE 8.2

SUMMARY STATISTICS OF ABS, SP, WP AND e

Variable	Mean	s.d.	Hurst's K	$\rho(1)$	Model
ABS	267	120.6	0.88	0.81	Eqn (8.2)
SP	428	73.8	0.70	0.0	MN
WP	87	23.7	0.67	0.0	-
WP	9.26	1.26	0.68	0.02	MN
e	-390	14.0	0.63	0.10	IND

Note: s.d. = standard deviation

$\rho(1)$ = first serial correlation coefficient

MN = Mixed-noise model described in Section 6.8

IND = White noise model (independent series)

where zs_t is a standardized normally distributed Mixed-Noise process that will on average reproduce a Hurst's K of 0.7 and a $\rho(1)$ of 0.0.

The ABS series generated using the simulation model given by (8.2) was checked for statistical resemblance with the observed ABS series. A large number of ABS series, each 70 years long, were generated using the model and generated sequences of SP, WP and e. This analysis showed that, on average, the ABS simulation model was able to reproduce the observed mean, standard deviation, the first 15 serial correlation coefficients, and the Hurst coefficient reasonably well in that the expected values of these statistics are approximately equal to the observed sample values.

8.3.2 Snow Accumulation

The second contributing variable to the magnitude of the annual spring peak flows at Emerson is the amount of snow or winter precipitation that has accumulated during the winter. This variable was briefly described in the previous Section on the ABS. For simplicity, the winter precipitation, WP, is assumed to start on November 1st and end on March 31st of the water year t. This is of course strictly not correct because in some years snow may come earlier while in other years it may continue to snow

after March 31st. But deviation from the assumptions will not have much effect on the analysis.

The total amount of winter precipitation in the basin is perhaps the most important variable that contributes to the magnitude of the spring peak flows on the Red River. High winter precipitation invariably means that more water is available for the runoff. The severity of the flood however would still depend on the other contributing variables. But, if there is no snow (highly unlikely in the Red River basin), then severe spring floods are extremely unlikely even if a high ABS and spring rains would occur.

As discussed in the previous section, the winter precipitation series can be generated by using a Mixed-Noise model. The generating equation is:

$$WP_t^{1/2} = 9.26 + 1.26 * zw_t \qquad \ldots(8.4)$$

where zw_t is a standardized MN process that on average will reproduce a Hurst's K = 0.68 and $\rho(1) = 0.02$. The WP_t series is then obtained by squaring $WP_t^{1/2}$. A large number of $WP_t^{1/2}$ series each of 70 years long were generated using (8.4) and the square transformation to test the adequacy of the generating scheme. It was found that on average the observed mean, standard deviation, first order serial coefficient, and the Hurst's coefficient K of the WP series were reasonably well

reproduced in that the expected values of these statistics are approximately equal to the observed sample values.

8.3.3 Melt-Rate Of Snow

The third variable contributing to the magnitude of spring peak flows is the rate at which the snow pack melts in the spring. The melt-rate is the most difficult parameter to develop. It depends on many meteorological factors for which very little data are available for proper time series analysis. The only available data for which long records are available is temperature. The use of temperature as an index of melt-rate was used with some success by Johnson and Archer (1972) and by the Manitoba Water Resources Branch (Warkentin, personal communications) for short term flood forecasting. However, the method that was used to estimate the melt-rate is highly subjective.

An objective albeit simplified index based on temperature is used in this study. It may be described as follows.

The melt index (MI) used is the average degree-days above $32^{o}F$ for a six days period prior to a nine day lag period from the day of the observed flood peak. This combination of a six days average and a nine day lag was obtained by trial and error until the best correlation is

obtained with the magnitude of the peak flows. An R^2 value of 0.07 was obtained which is comparable to the index used by the Manitoba Water Resources Branch. This low R^2 value, although statistically significant at the 5% level, indicates that the rate of snow melt contributes little to the magnitude of spring peak flows on the Red River. It should be noted that only the temperature data from the Grand Forks station in North Dakota was used. This station is in the middle of the basin upstream of Emerson.

The derived melt-index, MI, is also given in Table 8.1. It was found that a square root transformation of MI will give approximately normally distributed transformed variates. The summary statistics for the melt index are listed below.

	Mean	s.d.	Hurst's K	$\rho(1)$	Model
MI	5.92	3.94	0.70	0.10	-
MI	2.29	0.82	0.71	0.12	MN

The melt-index can be generated from:

$$MI_t^{1/2} = 2.29 + 0.82 * zm_t \qquad \ldots (8.5)$$

where zm_t is a standardized normally distributed process that on average will give Hurst's K = 0.7 and $\rho(1) = 0.1$. MI is obtained by squaring $MI^{1/2}$.

The adequacy of the melt-index generation model was tested using the same procedure for the WP. It was found on average that the generation model reproduces the observed sample statistics and serial correlation reasonably well in that the expected values of these statistics are approximately equal to the observed sample values.

8.3.4 Spring Precipitation

The final variable contributing to the magnitude of the spring peak flows at Emerson is the precipitation that occurs between March 31st and the day of the peak flow. A distinction must be made between the precipitation that may fall as snow or rain, and that falls only as rain. Snow may continue to fall after March 31st with a late spring. Therefore, the precipitation that falls between March 31st and the day of the flood peak at Emerson must be divided into two parts, CP and MP. The variable MP is defined as the amount of precipitation that falls from April 1st to the date nine days before the date of the peak flow at Emerson. Hence, MP may be in the form of snow or rain. The variable CP is defined as the amount of

liquid precipitation that falls during the nine day snowmelt period before the date of the peak flow at Emerson. The "nine days" is the lag period as defined in the previous section on melt-index.

The total amount of precipitation CP and MP were found to be highly correlated with the date of the peak flow at Emerson as one would expect; the later the peak the larger the amount of CP and MP. Therefore, the "time to peak" (TP) must be treated as a random variable and must be generated as part of the simulation process.

The variable "time to peak" TP is defined as the number of days from March 1st. For example, TP = 32 if the day of the peak is on April 1st, and TP = 63 if the day of the peak is on May 2nd.

The square root of TP was found to be approximately normally distributed. The summary statistics of TP and its appropriate time series model are as follows:

	Mean	s.d.	Hurst's K	$\rho(1)$	Model
TP	46	10.12	0.68	0.0	-
$TP^{1/2}$	6.47	0.742	0.67	0.0	MN

The generating equation for TP is given by:

$$TP_t^{1/2} = 6.8 + 0.74 * zt_t \qquad \ldots (8.6)$$

where zt_t is a standardized normally distributed Mixed-Noise variate. The generated values of $TP_t^{1/2}$ are squared and the integer part is taken as TP.

After a minor adjustment to the transformed mean, from 6.47 to 6.80 the generating equation was found to be able to reproduce on average the observed mean, standard deviation, Hurst's K and the first serial correlation reasonably well.

From the record, there are six years in which TP was less than 32. That is, the peak flow occurred before April 1st. For these cases, CP and MP are both zero. Also by definition when TP is less than 41, MP is zero. There are 21 such cases in the observed data set. A simulation model for MP is obtained by regressing the remaining 49 values of MP to TP. A cube-root transformation on the values of MP used in the regression was necessary to obtain normally distributed random error terms. The simulation model is given by:

$$MP_t^{1/3} = 0.131 * TP_t + em_t \qquad TP_t > 40 \qquad \ldots (8.7a)$$

$$MP_t = 0 \qquad \text{otherwise} \qquad \ldots (8.7b)$$

where em, the error term, is normally distributed with mean = -4.68 and standard deviation = 0.96.

The generating scheme (8.7) was found to be able to reproduce the statistics of the transformed variate well.

Cubing the generated values of $MP^{1/3}$ however did not preserve the historical statistics of MP. However, this is not a problem in the construction of the final compound simulation model of spring peak flows as will be described in Section 8.4.

Similarly, a simulation model for CP was obtained by regression of the remaining 64 CP values to TP. A cube-root transformation on CP was also found necessary to give normally distributed random error terms in the regression model. The simulation model for CP is given by:

$$CP_t^{1/3} = 0.041 * TP_t + ec_t \qquad TP > 32 \qquad \ldots (8.8a)$$

$$CP_t = 0 \qquad \text{otherwise} \qquad \ldots (8.8b)$$

where ec, the error term, is normally distributed with mean = -0.381 and standard deviation = 0.90.

The generating scheme (8.8) was also found to be able to reproduce the observed statistics of the transformed variates well. As with the generation of MP, cubing $CP^{1/3}$ to obtain CP did not preserve the historical statistics of CP. Therefore only $CP^{1/3}$ without the inverse transformation is generated.

The summary statistics for $CP^{1/3}$ and $MP^{1/3}$ are as follows:

	Mean	s.d.	Hurst's K	ρ(1)
$CP^{1/3}$	1.442	1.056	0.55	0.0 *
$MP^{1/3}$	1.438	1.493	0.67	0.0 *

* Calculated ρ(1) were negative and has been set to zero here for physical reasons.

8.4 CONSTRUCTION OF SPRING PEAK FLOW SIMULATION MODEL

In this section, the cross-correlations among the contributing variables are checked, and the physical-statistical relationship between the annual spring peak flows at Emerson and the contributing variables defined.

Having identified the variables contributing to the magnitude of annual spring flows at Emerson; namely, the soil moisture condition, the winter precipitation, the rate of snowmelt and the spring precipitation, and procedures for simulating each of them, their cross-correlations must be checked before they can be used in the peak flow simulation model. One would expect that the melt-index, MI, to be correlated to the "time to peak", TP. Since the later the peak occurs, the greater are the chances of having rapid rise in temperature. In this study, however, no appreciable correlation between two variables were found. The reason for this is probably that the melt-index used did not reflect the actual melt

conditions in the basin. This is a major drawback of temperature based indices.

The variables CP and MP are correlated to each other as expected since they are both related to TP.

All other variables were found to be mutually independent. This means that ABS, WP, TP, and MI can each be generated as independent inputs into the peak flow simulation model.

A physically-based relationship between spring peak flows and the primary contributing variables is given by:

$$PF_t = k_1 WP_t^{a} (ABS_t - b)\, MI_t^{c} + k_2 MP_t^{d} (ABS_t - b)\, MI_t^{c} + k_3 CP_t^{e} (ABS_t - b) + E_t \quad \ldots (8.9)$$

where k_1, k_2, k_3 are scale parameters; a, c, d, e, are shape parameters; b is a threshold parameter related to effective soil moisture; and E is the random error term.

The explanation for (8.9) is as follows:
The three precipitation terms namely, WP, MP and CP are the factors that contribute to the magnitude of the spring peak flow additively. The contribution of the winter precipitation (WP), however, is influenced by the soil moisture condition (ABS), and the rate of melt (MI) of the snow pack. The contribution of MP (which may be part snow) is similarly influenced by the soil moisture condition (ABS), and the melt rate (MI). The contribution of CP is influenced only by the soil moisture condition.

The melt index is not included because it is assumed that CP consists only of rain and no snow. The last additive term in (8.9) is the error term which takes into account other factors that contribute to the spring peak flows but were not modelled explicitly.

To obtain the constants of the model given by (8.9), a non-linear least squares regression technique is used to fit the observed spring peak flow data (PF) to the observed data of the variables on the right hand side of (8.9). The following results were obtained.

$k_1 = 0.000081$ $k_2 = 0.01418$ $k_3 = 0.00140$ $a = 2.1092$

$b = -130$ $c = 0.0294$ $d = 1.0886$ $e = 1.5464$

and E_t is normally distributed with a mean of 93.5 and a standard deviation of 279.1.

Residual analysis was then used to check the adequacy of the regression model.

The plot of the predicted peak flow values against the residuals of the fitted model indicated that the variance of the residuals are not constant. That is, there is a problem of heteroscedasticity. A plot of each of the independent variables against the residuals showed that the problem of heteroscedasticity is caused by the winter precipitation (WP) variable; larger residual variance being associated with larger winter precipitation.

A simple way to deal with heteroscedasticity is by the method of "deflation" (Johnson, 1972) whereby the terms in the equation are divided by the variable causing heteroscedasicity. It is assumed that:

$$\text{Var}\,(E_t) = WP^2\,\sigma^2 \qquad \ldots (8.10)$$

Dividing (8.9) throughout by WP results in an equation with homoscedastic error term. Performing this division gives a modification of (8.9):

$$\frac{PF}{WP} = k_1 WP_t^{a-1}(ABS_t - b)\,MI_t^{c} + k_2 MP_t^{d}(ABS_t - b)\,MI_t^{e}\,\frac{1}{WP}$$

$$+ k_3 CP_t^{e}(ABS_t - b)\,\frac{1}{WP} + E_t^{*} \qquad \ldots (8.11)$$

where, $E_t^{*} = E_t/WP_t$, and the model constants now become; $k_1 = 0.0000265$, $k_2 = 0.00789$, $k_3 = 0.0992$, $a = 2.309$, $b = -106$, $c = 0.0282$, $d = 1.236$, $e = 0.2524$, and E^{*} is normally distributed with a mean $\mu_E = 2.0$ and a standard deviation $\sigma_E = 2.9$.

Equation (8.11) is then used to generate PF/WP data after which values of PF are obtained by multiplying through by WP.

The percentage of variance of the spring peak flow at Emerson explained by the simulation model (8.11) is about 74%.

The simulation model can be simplified further by dropping the melt-index term (MI) since the contribution

of this variable is very small and the real effect of melting condition is already taken into account by the "time to peak", TP.

The refitted model is given by:

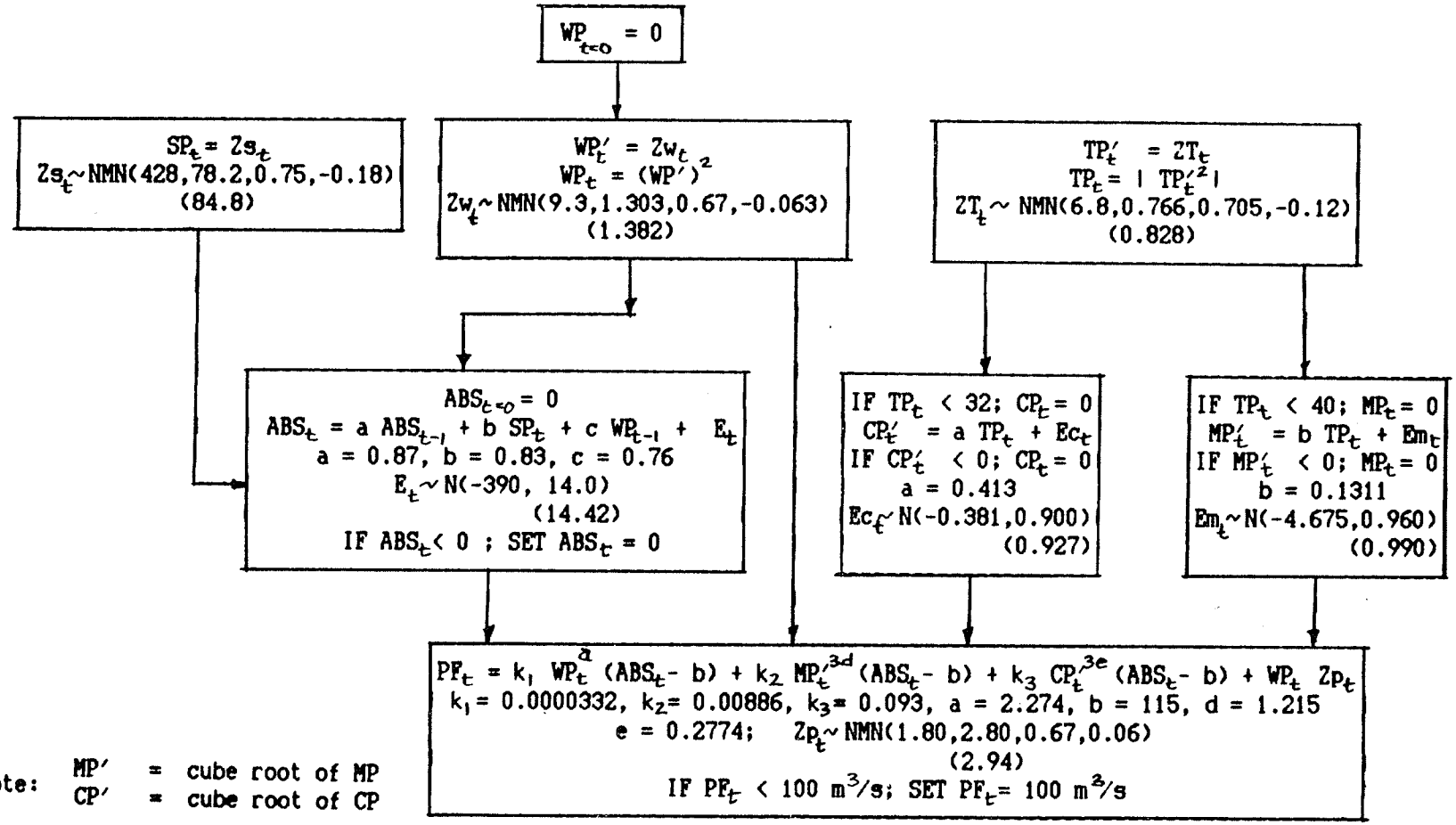
$$\frac{PF_t}{WP_t} = k_1' WP_t^{a'-1}(ABS_t - b') + k_2' MP_t^{d'}(ABS_t - b')\frac{1}{WP_t}$$

$$+ k_3' CP_t^{e'}(ABS_t - b')\frac{1}{WP_t} + E_t' \qquad \ldots (8.12)$$

where $k_1' = 0.0000332$, $k_2' = 0.00886$, $k_3' = 0.0926$, $a' = 2.274$, $b' = 115$, $d' = 1.215$, $e' = 0.2774$, E_t' is normally distributed with mean $\mu_{E'} = 2.0$ and standard deviation $\sigma_{E'} = 2.90$. The Hurst's coefficient K and first serial correlation coefficient $\rho(1)$ of E_t' is 0.68 and 0.1 respectively.

The difference in the percentage of variance of PF explained by this new model and (8.11) is negligible. Equation (8.12) is therefore preferred since it has fewer parameters.

The generation of PF can now proceed. The steps in the generation process are shown in Figure 8.2. In the figure, the coefficients d and e are multiplied by 3 because the cube-root of CP and MP are used as inputs instead of the original variable as explained in Section 8.3.4.



Note: MP' = cube root of MP
CP' = cube root of CP

NMN (μ, σ, h, ρ_H) = normally distributed mixed-noise process with mean μ, unbias standard deviation σ, model Hurst coefficient h, and high frequency serial coefficient term ρ_H.

N (μ, σ) = normally distributed independent process with mean μ and unbias standard deviation σ.

(σ_ρ) = predictive standard deviation.

Figure 8.2 Flow chart for stochastic generation of spring peak flows

To test the adequacy of the peak flow simulation model (8.12), a large number of series each 70 year length was generated.

The observed Hurst's K and first serial correlation were reasonably well reproduced by the model in that the values of these statistics are approximately equal to the observed sample values. However, minor bias corrections to the parameters μ_E' and σ_E' were required to reproduce on average the observed mean and standard deviation of the peak flow series. In addition, the minimum generated peak flow value is set at 100 m^3/s which is close to the observed minimum peak flow of 107 m^3/s. This is to avoid accidental generation of negative flows.

The observed spring peak flow statistics and those obtained on average by the simulation model is shown in Table 8.3.

TABLE 8.3

OBSERVED AND GENERATED SPRING PEAK FLOW STATISTICS

	Mean	s.d.	skew	Hurst's K	$\rho(1)$
PF_{obs}	692	518	1.8	0.745	0.20
PF_{gen}	692	514	2.1	0.740	0.16

Note: Unit of peak flow is m^3/s.
PF_{obs} = observed peak flow series.
PF_{gen} = generated peak flow series based on 20,000 replications of 70 years.

8.5 DESCRIPTIVE DISTRIBUTION FUNCTION OF SPRING PEAK FLOWS USING THE SUMULATION MODEL

To derive the complete probability distribution function of the spring peak flows at Emerson using the simulation model (8.12), a large number of peak flow data (PF) must be generated. It was found that 20 000 sequences of PF each of length equal to 70 years, giving a total of 1,400,000 peak flow data, gave stable probabilities based on relative frequencies.

The probability that a certain flow rate is equalled or exceeded in any one year, $P(X \geqslant Q)$, is estimated by:

$$P(X \geqslant Q) = \frac{\text{No. of PF} \geqslant Q}{1,400,000} \qquad \ldots (8.13)$$

Table 8.4 summarises the results for various values of Q, and Figure 8.3 shows the plot of the probability distribution function on probability graph paper. The observed flood data fitted by a 2-parameter lognormal distribution and is also shown in Figure 8.3.

The probability distribution obtained from the peak flow simulation model (8.12) is a descriptive rather than a predictive distribution since the parameter uncertainty of the distribution functions of the contributing variables have not been taken into account. Only the stochastic or inherent uncertainty of the contributing variables were considered.

TABLE 8.4

DESCRIPTIVE DISTRIBUTION FROM SIMULATION MODEL

(Total number of data = 1,400,000)

Q (m^3/s)	# $\geqslant$ Q	Probability
150	1,323,840	0.94560
200	1,262,996	0.90214
300	1,102,640	0.78760
500	758,688	0.54192
1000	263,396	0.18814
2000	44,296	0.03164
2670*	18,396	0.01314
3000	12,684	0.00906
3880**	5,558	0.00397
4500	3,444	0.00246
5153***	2,170	0.00155
6000	1,320	0.00094
7000	742	0.00053
8000	448	0.00032

* 1950 flood
** 1852 flood
*** 1826 flood (highest historical flood observed)

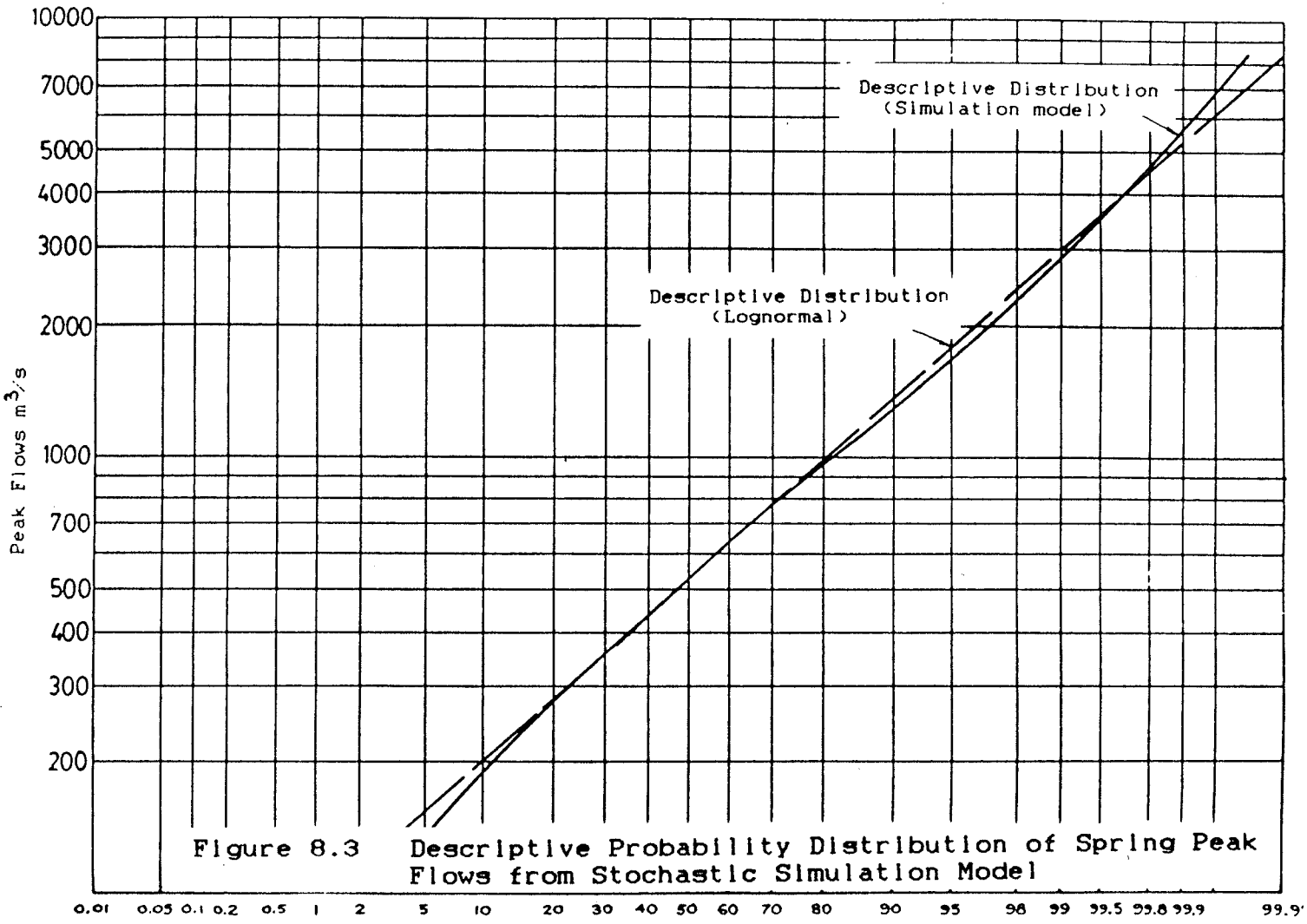


Figure 8.3 Descriptive Probability Distribution of Spring Peak Flows from Stochastic Simulation Model

One can see that the distribution function obtained from the simulation model is almost indistinguishable from that obtained by fitting a lognormal distribution to the observed peak flow data up to a risk level (probability of exceedence) of 1%. At the upper tails of the distributions however, the distribution obtained by simulation is steeper than that of the assumed lognormal distribution. This is an important consideration in view of the negative skew in the log transformed recorded series which has led previous investigators to curve the line downward.

Table 8.5 compares the probability of exceedences obtained by simulation and the lognormal distribution of some of the historical spring floods at Emerson.

However, when making probability assessments about future flood occurrences, the uncertainty in the parameters must be included in the distribution functions of each contributing variable. This is considered in the next section.

TABLE 8.5

PROBABILITY OF EXCEEDENCE OF HISTORICAL SPRING PEAKS ON THE RED RIVER AT EMERSON

Year	Q (m^3/s)	$Prob_{Sim}$	$Prob_{LN2}$
1826	5153	0.0016 (625)	0.0013 (770)
1852	3880	0.0040 (250)	0.0035 (286)
1950	2670	0.0130 (77)	0.0170 (59)

$Prob_{Sim}$: Probability obtained from descriptive simulation model

$Prob_{LN2}$: Probability obtained from lognormal model

Note: Approximate return periods in parenthesis.

8.6 PREDICTIVE DISTRIBUTION FUNCTION OF SPRING PEAK FLOWS USING SIMULATION MODEL

The predictive distribution was discussed in Chapter 4. It combines both stochastic and parameter uncertainty into the analysis.

To include parameter uncertainty into time series generation, it is necessary to sample from the predictive distribution of the random variable (Vicens et al., 1975). For normally distributed variables this essentially amounts to sampling from a distribution function with a larger variance. The mean remains the same. The generated peak flow sequences obtained this way would then include parameter uncertainty caused by serial correlation and sample length.

The predictive distribution functions of the contributing variables are obtained using the discrete predictive distribution approach described in Section 4.5 after doing the necessary transformation of the variables to normally distributed variates. The parameters used to generate the predictive distribution function are shown in Figure 8.2 in parenthesis.

As with the descriptive distribution, 20 000 sequences each 70 years long were generated to define the predictive distribution of the spring peak flows. The results are shown in Table 8.6.

TABLE 8.6

PREDICTIVE DISTRIBUTION FUNCTION FROM SIMULATION MODEL
(Total number of data = 1,400,000)

Q (m^3/s)	# $\geq$ Q	Prob $\geq$ Q
150	1,313,424	0.93816
200	1,251,194	0.89371
300	1,094,100	0.78150
500	770,280	0.55020
1000	296,254	0.21161
2000	60,802	0.04343
2670*	27,468	0.01962
3000	19,684	0.01406
3880**	9,548	0.00682
4500	6,174	0.00441
5153***	4,228	0.00302
6000	2,674	0.00191
7000	1,624	0.00116
8000	1,050	0.00075

* 1950 flood
** 1852 flood
*** 1826 flood

Figure 8.4 shows the results of Table 8.6 on probability graph paper. The predictive and descriptive distribution of the observed spring peak flows assuming a lognormal distribution are also shown in Figure 8.4.

The predictive distribution function that is obtained by simulation is, as expected, steeper than the descriptive distribution which does not take parameter uncertainty into account. But, the predictive distribution function based only on the observed data is considerably steeper (larger predictive variance) than that obtained from the simulation model. This indicates that the additional information obtained from the contributing variables to the spring peak flows have reduced the variance of the predictive distribution. That is, the uncertainty in the risk assessment is reduced by the contributing variables in the simulation model. Furthermore, since no 'a priori' probability distribution is necessary for the spring peak flows in the simulation model, the uncertainty associated with choosing the appropriate probability distribution to describe the observed data is also reduced.

Therefore, the predictive distribution obtained using the stochastic simulation model should provide a better basis for probability assessment of future floods on the Red River at Emerson.

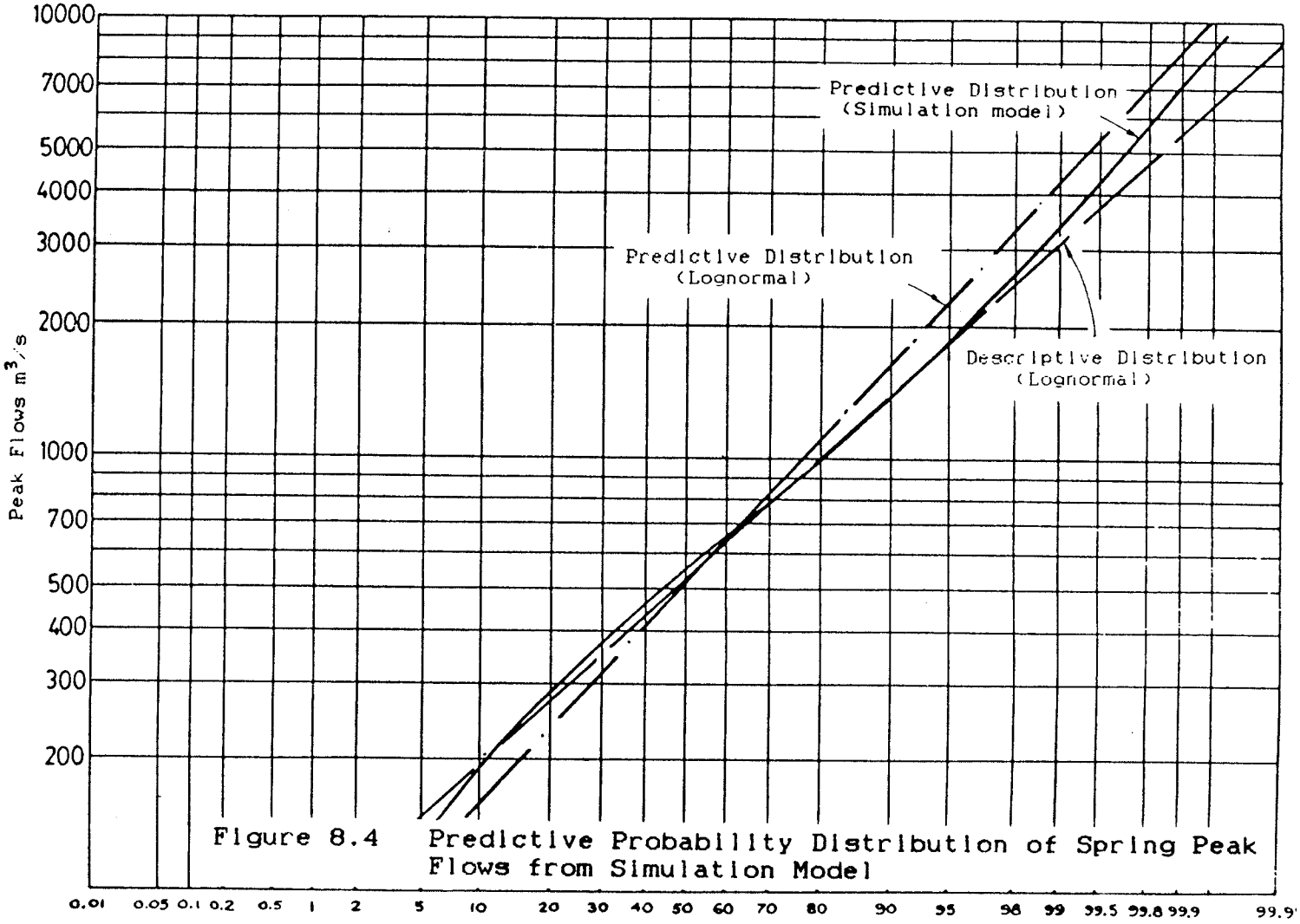


Figure 8.4 Predictive Probability Distribution of Spring Peak Flows from Simulation Model

8.7 SUMMARY

The construction of a physically-based stochastic simulation model for the spring peak flows on the Red River at Emerson was attempted in this chapter.

In developing the simulation model, the observed Hurst's K, first serial correlation coefficient and the parameters of the model from the non-linear least squares fit were assumed to be constant. These are also uncertain parameters. However, it does not seem to be unreasonable to assume at this stage that they are constant for the next 70 year period. Only the uncertainty related to the parameters of the distribution function of the contributing variables were taken into consideration in the simulation model. This uncertainty however included uncertainty caused by short sample length and serial correlation.

The predictive distribution obtained from the stochastic simulation model showed less uncertainty than that based only on the observed data.

CHAPTER 9

CONCLUSIONS AND RECOMMENDATIONS

9.1 CONCLUSIONS

The primary conclusions to be drawn from this thesis are:

1. If there is sufficient evidence to show that the observed series of annual peak flows are serially independent, and the only issue of interest is to describe the probabilistic phenomenon using the observed data, then a Bayesian approach of estimating the parameters of the flood probability distribution provides slightly more reliable estimates than the usually preferred method of maximum likelihood.

   The computation of Bayes estimates with a 'vague' prior distribution for the parameters, has been shown to be much simplified using Lindley's Bayesian approximation procedure.

2. For making decisions concerning future floods using the observed flood data, the parameter uncertainty of the flood probability distribution and the stochastic uncertainty of the flood data it defines must be combined. This can be achieved using the predictive distribution approach. If there are

sufficient reasons to believe that in the future time frame bounded by the planning horizon, the annual flood data are serially independent, either Lindley's method or Russell's discrete approach can be used to obtain the predictive distribution. In this case, parameter uncertainty is primarily due to short sample length.

3 If there is evidence of long term serial correlation in the observed flood data leading to a high Hurst's coefficient, then the parameter uncertainty is substantial even when long flood records are available.

The predictive distribution for serially correlated data can be obtained by using Russell's discrete predictive distribution approach. The uncertainty of the estimated parameters is obtained using a time series model that is capable of reproducing the observed serial correlation structure of the peak flows.

4. Annual peak flows series of Canadian rivers are found to exhibit significant long term serial correlation similar to the many geophysical time series examined by Hurst himself.

5. A new technique for modelling series with a high Hurst coefficient and a low first order serial correlation coefficient which is characteristic of

flood peak series was developed. The resulting model is called a Mixed-Noise model. It has been shown that the Mixed-Noise model is relatively efficient and that estimation of the model's parameters is extremely easy in comparison to the estimation of parameters for other models currently available. In addition, graphs for determining the small sample bias corrections of the model parameters were developed for ease of use.

6. A stochastic simulation model for the spring peak flows on the Red River at Emerson was described. The stochastic inputs to the simulation model are the primary contributing variables that determine the magnitude of the spring peak flows. These are: the soil moisture condition in the basin prior to runoff and measured by the ABS; the amount of winter precipitation that has accumulated over the winter; the melt-rate of the snow; and the precipitation that occurs during the melting period. Each contributing variable is generated as a stochastic time series, and the variables are related to the spring peak flows by a physically-based statistical relationship.

The predictive distribution obtained from the simulation model indicates that the additional information from the contributing variables has increased the reliability of risk assessment of

future spring peak flows when compared to the use of the predictive distribution based only on observed data. In addition, the simulation model also provides the distribution function of the peak flows thereby obviating the need to assume an 'a priori' flood distribution.

7. Long term soil moisture fluctuations as measured by the accumulated basin storage was found to explain in part the observed serial correlation structure of the spring peak flows at Emerson.

8. Parameter uncertainty due to short sample length and aggravated by serial correlation appears to be a more important issue than the questions of probability plotting positions, parameter estimation by the method of moments, maximum likelihood, etc., or the choice between a 2-parameter and 3-parameter distribution which have been the subjects of much research by engineers and statisticians alike.

9.2 RECOMMENDATIONS FOR FURTHER STUDY

There are a number of issues in which further study is desirable:

1. While Lindley's approximation procedure is an attractive method for engineers to obtain Bayes estimators, the technique at present is limited to non data-based prior distributions only. Techniques for inclusion of data-based priors would greatly enhance the attractiveness of Lindley's approach.

2. Investigation into the theoretical properties of the Mixed-Noise model and a comprehensive comparison with other available models are desirable. This includes developing simple procedures and appropriate time series models for practising engineers to select and use.

3. Since parameter uncertainty caused by long term serial correlation is quite substantial leading to a upward assessment of flood risk, physical reasons for the long term behaviour should be investigated for each river basin where this phenomenon is observed. The accumulated basin storage appears to be a good parameter by which to examine the flows.

4. Further advancement of flood risk assessment can probably only come from a better understanding of the physical processes that determine the magnitude

of floods itself. A synthesis of physical and statistical hydrology is indispensible for such a study. In addition methodology for combining additional information on floods via Bayesian techniques or other methods would be desirable.

REFERENCES

Abramowitz, M. and Stegun, I.A. (1970). Handbook of Mathematical Functions. Dover Publications Inc., New York. 1046 p.

Anderson, R.L. (1942). Distribution of the Serial Correlation Coefficient. Annals of Mathematical Statistics, 13, 1-13.

Ang, A.H-S. and Tang, Wilson. (1975). Probability Concepts in Engineering Planning and Design, Vol. 1. John Wiley and Sons, New York, 488 p.

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. Philos. Transactions Royal Society, London, 53, 370-418. Reprinted in Biometrika 45, 296-315.

Beard, L.R. (1974). Flood Flow Frequency Techniques. Technical Report CRWR 119, Center for Research in Water Resources. University of Texas, Austin.

Benjamin, J.R. and Cornell, C.A. (1970). Probability, Statistics and Decision for Civil Engineers. McGraw Hill, New York. 684 p.

Bernier, J. (1967). Les Methods Bayesiennes en Hydrologie Statistique. Proceedings: 1st International Hydrology Symposium, Vol. 1, Fort Collins, Colarado, 459-470.

Bodo, B. and Unny, T.E. (1976). Model Uncertainty in Flood Frequency Analysis and Frequency-based Design. Water Resources Research, 12(6), 1109-1117.

Booy, C. and Lye, L.M. (1985). The Use of Accumulated Basin Storage in Flood Peak Analysis. Proceedings 7th Canadian Hydrotechnical Conference, Saskatoon. 317-335.

Booy, C. and Lye, L.M. (1986). Accumulated Basin Storage as a Factor in the Correlation Structure of Annual Peak Flows on the Red River. Canadian Journal of Civil Engineering, 13(3), 365-374.

Booy, C. and Lye, L.M. (1987). Uncertainty in Flood Risk Analysis. Proceedings 8th Hydrotechnical Conference, Montreal, 401-418.

Booy, C. and Morgan, D.R. (1985). The Effect of Clustering of Flood Peaks on a Flood Risk Analysis for the Red River. Canadian Journal of Civil Engineering, 12(1), 150-165.

Box, G.E.P. and Jenkins, G.M. (1970). Time Series Analysis: Forecasting and Control. Holden Day, San Francisco, 553 p.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison Wesley, Massachusetts, 588 p.

Burges, S.J., Lettenmaier, D.P. (1975). Operational Comparison of Stochastic Streamflow Generation Procedures. Technical Report 45, Harris Hydraulics Lab., Dept. of Civil Engineering, University of Washington, Seattle, 112 p.

Burges, S.J., Lettenmaier, D.P. and Bates, C.L. (1975). Properties of the Three Parameter Lognormal Probability Distribution. Water Resources Research, Vol. 11, 229-235.

Chi, M., Neal, E. and Young, G.K. (1973). Practical Applications of Fractional Brownian Motion and Noise to Synthetic Hydrology. Water Resources Research, 9(6), 1523-1533.

Davis, D.R., Kisiel, C.S., and Duckstein, L. (1972). Bayesian Decision Theory Applied to Design in Hydrology. Water Resources Research. 8(1), 33-41.

Feller, W. (1951). The Asymptotic Distribution of the Range of Sums of Independent Random Variables. Annals of Mathematical Statistics, 22, 427-432.

Fiering, M.B. (1967). Streamflow Synthesis. MacMillan Ltd., London. 139 p.

Gren, J. (1980). Disscussant on D.V. Lindley's (1980) paper on Approximate Bayesian Methods, Trabajos Estadistica, 31, p. 241.

Gumbel, E.J. (1958). Statistics of Extremes. Columbia University Press, New York, New York.

Gumbel, E.J. (1967). Discussion - Some Aspects of Time Series Analysis in Hydrological Studies by N.C. Matalas. Symposium on Statistical Methods in Hydrology, Montreal.

Haan, C.T., Johnson, H.P. and Brakensiek, D.L. (1982). Hydrologic Modelling of Small Watersheds. ASAE Monograph No. 5, American Society of Agricultural Engineers, Michigan.

Hall, W.A., Askew, A.J. and Yeh, W.W-G. (1969). Use of the Critical Period in Reservoir Analysis. Water Resources Research, 5(6), 1205-1215.

Hipel, K.W. (1975). Contemporary Box-Jenkins Modelling in Hydrology. PhD. Thesis, University of Waterloo.

Hurst, H.E. (1951). Long Term Storage Capacity of Reservoirs. Trans. Am. Soc. Civ. Engrs., 116, 770-808.

Hurst, H.E. (1956). Methods of Using Long-Term Storage in Reservoirs. Proc. Inst. Civil Engineers, 1, 519-543.

Hurst, H.E. (1957). A Suggested Statistical Model of Some Time Series Which Occur in Nature. Nature, 180, p. 494.

Jain, D. and Singh, V.P. (1987). Estimating Parameters of EV1 Distribution for Flood Frequency Analysis. Water Resources Bulletin, 23(1), 59-71.

Jeffreys, Sir H. (1961). Theory of Probability. Oxford Clarendon Press.

Jenkins, G.M. and Watts, D.G. (1968). Spectral Analysis and Its Applications. Holden Day, San Francisco, Calif., 525 p.

Johnson, J. (1972). Econometric Methods, McGraw Hill, New York. 2nd Edition.

Johnson, P. and Archer, D.R. (1972). Current Research in British Snowmelt River Flooding. Hydrological Science Bulletin, 17(4), 443-451.

Kalman, R.E. and Bucy, R.S. (1961). New Results in Linear Filtering and Prediction Theory. A.S.M.E.J. Basic Engineering. 83D, 85-108.

Kendall, M.G. and Stuart, A. (1973). The Advanced Theory of Statistics, Vol. 2, Third Edition, Hafner Publishing, New York.

Kimball, B.F. (1949). An Approximation of the Sampling Variances of An Estimated Maximum Value of Given Frequency Based on Fit of Doubly Exponential Distribution of Maximum Values. Annals of Mathematical Statistics, 20, 110-113.

Kite, G.W. (1977). Frequency and Risk Analysis. Water Resources Publication, Littleton, Colorado, 224 p.

Klemes, V. (1972). Comments on - Adequacy of Markovian Models with Cyclic Components for Stochastic Streamflow Simulation by I. Rodriquez-Iturbe, D.R. Dawdy and L.E. Garcia, Water Resources Research, Vol. 8, 1613-1615.

Klemes, V. (1974). The Hurst Phenomenon - A Puzzle? Water Resources Research, 10(4), 675-688.

Kottegoda, N.T. (1974). Effect of Skewness in 3 Stochastic Pentad River Flow Models on Crossing Properties of Synthesised Data. Water Resources Research, Vol. 10, 446-456.

Kottegoda, N.T. (1980). Stochastic Water Resources Technology. MacMillan, London. 384 p.

Lawrence, A.J. and Kottegoda, N.T. (1977). Stochastic Modelling of River Flow Time Series. Journal Royal Statistical Society, Series A, 140, Part 1, 1-47.

Lettenmaier, D.P. and Burges, S.J. (1977a). Operational Assessment of Hydrologic Models of Long Term Persistence. Water Resources Research, 13(1), 113-124.

Lettenmaier, D.P. and Burges, S.J. (1977b). An Operational Approach to Preserving Skew in Hydrologic Models of Long Term Persistence. Water Resources Research, 13(2), 281-290.

Lettenmaier, D.P. and Burges, S.J. (1982). Gumbel's Extreme Value 1 Distribution: A New Look. Journal of the Hydraulics Division, Proc. American Society of Civil Engineers, 108(HY4), 502-514.

Lindley, D.V. (1980). Approximate Bayesian Methods. Trabajos Estadistica, 31, 223-237.

Loucks, D.P., Stedinger, J.R. and Haith, D.A. (1981). Water Resources System Planning and Analysis. Prentice Hall, Englewood Cliffs, N.J. 559 p.

Mandelbrot, B.B. (1965). Une Classe de Processus Stochastiques Homothetiques à Soi; Application à la Loi Climatologique de H.E. Hurst., Comptes Rendus de L'Academie des Sciences de Paris, 260, 3274-7.

Mandelbrot, B.B. (1971). A Fast Fractional Gaussian Noise Generator. Water Resources Research, 7(3), 543-553.

Mandelbrot, B.B. (1972). Broken Line Process Derived As An Approximation to Fractional Noise. Water Resources Research, 8(5), 1354-1356.

Mandelbrot, B.B. and Van Ness, J.W. (1968). Fractional Brownian Motions, Fractional Noises and Applications. S.I.A.M. Review, 10(4), 422-437.

Mandelbrot, B.B. and Wallis, J.R. (1968). Noah, Joseph and Operational Hydrology. Water Resources Research, 4(5), 909-918.

Mandelbrot, B.B. and Wallis, J.R. (1969a). Computer Experiments with Fractional Gaussian Noises. Part 1 - Averages and Variances. Water Resources Research, 5(1), 228-241.

Mandelbrot, B.B. and Wallis, J.R. (1969b). Computer Experiments with Fractional Gaussian Noises. Part 2 - Rescaled Ranges and Spectra. Water Resources Research, 5(1), 242-259.

Mandelbrot, B.B. and Wallis, J.R. (1969c). Computer Experiments with Fractional Gaussian Noises. Part 3 - Mathematical Appendix. Water Resources Research, 5(1), 260-267.

Mandelbrot, B.B. and Wallis, J.R. (1969d). Some Long-Run Properties of Geophysical Records. Water Resources Research, 5(2), 321-340.

Mandelbrot, B.B. and Wallis, J.R. (1969e). Robustness of the Rescaled Range R/S in the Measurement of Non-cyclic Long-Run Statistical Dependence. Water Resources Research, 5(5), 967-988.

Martz, H.F. and Waller, R.A. (1982). Bayesian Reliability Analysis. John Wiley and Sons, New York, 745 p.

Matalas, N.C. and Huzzen, C.S. (1967). A Property of the Range of Partial Sums. Proc. Int. Hydrology Symposium, Fort Collins, Colorado State University, Vol. 1, 252-257.

Matalas, N.C. and Wallis, J.R. (1971). Statistical Properties of Multivariate Fractional Noise Processes. Water Resources Research, 7(6), 1460-1468.

Matalas, N.C. and Wallis, J.R. (1976). Generation of Synthetic Flow Sequences, in Biswas, A.K. (ed), Systems Approach to Water Management. McGraw Hill, 54-79.

Matalas, N.C. (1977). Generation of Multivariate Synthetic Flows. In: Mathematical Models for Surface Water Hydrology, Ciriani, T.A., Manione, U. and Wallis, J.R. (eds.). John Wiley and Sons. pp. 27-38.

Mejia, J.M., Dawdy, D.R. and Nordin, C.F. (1974). Streamflows Simulation 3. The Broken Line Process and Operational Hydrology. Water Resources Research, 10(2), 242-245.

Mejia, J.M., Rodriquez-Itrube, I. and Dawdy, D.R. (1972). Streamflow Simulation 2. The Broken Line Process as a Potential Model for Hydrologic Simulation. Water Resources Research, 8(4), 931-941.

O'Connell, P.E. (1971). A Simple Stochastic Modelling of Hurst's Law. Proc. Int. Symposium on Mathematical Models in Hydrology. Int. Association Hydrological Sciences, Warsaw, 169-187.

O'Connell, P.E. (1974). Stochastic Modelling of Long Term Persistence in Streamflow Sequences. PhD. Thesis, Imperial College, University of London, 284p.

Phien, H.N. and Arbhabhirama, A. (1980). A Comparison of Statistical Tests on the EV1 Distribution. Water Resources Bulletin, 16(5), 856-861.

Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory, Harvard University Press, Boston, Mass.

Red River Basin Investigation. (1953). Report on Investigations into Measures for the Reduction of the Flood Hazard in the Greater Winnipeg Area.

Russell, S.O. (1982). Flood Probability Estimation. Journal of the Hydraulics Division, Am. Soc. of Civil Engineers, 108(HY1), 63-73.

Salas, J.D. and Smith, R.A. (1981). Physical Basis of Stochastic Models of Annual Flows. Water Resources Research, 17(2), 428-430.

Sangal, B.P. and Biswas, A.K. (1970). The 3-parameter Lognormal Distribution and Its Application in Hydrology. Water Resources Research, 6(2), 505-515.

Sinha, S.K. (1985). Bayesian Estimation of the Reliability Function of Normal Life Testing Distribution. IEEE Transactions on Reliability. Vol. R-34, 4, 360-362.

Sinha, S.K. (1986a). Reliability and Life Testing. Wiley Eastern Ltd./Halsted Press. 252 p.

Sinha, S.K. (1986b). Bayesian Estimation of the Reliability Function of the Inverse Gaussian Distribution. Statistics and Probability Letters, 4(6), 319-323.

Sinha, S.K. (1986c). Bayesian Estimation of the Reliability Function and Hazard Rate of a Weibull Failure Time Distribution. To appear in Trabajor Estadistica.

Sinha, S.K. (1987). A Note on the Efficiency of Bayesian Approximation. Dept. of Statistics, University of Manitoba, Tech. Report No. 185.

Sinha, S.K. and Sloan, J. (1985). Bayesian Estimation of the Parameters and Reliability Function of a 3-parameter Weibull Life Distribution. To appear in IEEE - Transactions on Reliability, 1988.

Srikanthan, R. and McMahon, T.A. (1978). Generation of Skewed Flows Using a Fast Fractional Gaussian Noise Generator. Water Resources Research, 14(4), 665-671.

Srikanthan, R. and McMahon, T.A. (1978). Comparison of Fast Fractional Gaussian Noise and Broken Line Models for Generating Annual Flows. Journal of Hydrology, 81-92.

Srikanthan, R. (1979). Stochastic Generation of Annual and Monthly Flow Volumes. PhD Thesis, Monash University, Australia.

Stedinger, J.R. (1980). Fitting Lognormal Distributions to Hydrologic Data, Water Resources Research, 16(3), 481-490.

Stedinger, J.R. (1983). Design Events with Specified Flood Risk. Water Resources Research, 19(2), 511-522.

Vicens, G.J., Rodriquez-Iturbe, I. and Schaake, J.C. Jr. (1975). A Bayesian Framework for the Use of Regional Information in Hydrology. Water Resources Research, 11(3), 405-414.

Vicens, G.J., Rodriquez-Iturbe, I. and Schaake, J.C. Jr. (1975). A Bayesian Framework for the Use of Regional Information in Hydrology. Water Resources Research, 11(3), 405-414.

Wallis, J.R. and Matalas, N.C. (1970). Small Sample Properties of H and K - Estimators of the Hurst Coefficient h. Water Resources Research, 6(6), 1583-1594.

Wallis, J.R. and Matalas, N.C. (1971). Correlogram Analysis Revisited. Water Resources Research, 7(6), 1448-1459.

Wallis, J.R. and Matalas, N.C. (1972). Sensitivity of Reservoir Design to Generating Mechanism of Inflows. Water Resources Research, 8(3), 634-641.

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

Wallis, J.R. and O'Connell, P.E. (1973). Firm Reservoir Yield: How Reliable are Historic Hydrological Records. Bull. Int. Ass. Hydrol. Science, 18(3), 347-365.

Warkentin, A. (1985). Personal communications.

Wood, E.F., Rodriquez-Iturbe, I. and Schaake, J.C., Jr. (1974). The Methodology of Bayesian Inference and Decision Making Applied to Extreme Hydrologic Events, Rep. 178, Ralph M. Parsons Lab. for Water Resources and Hydrodym., Department of Civil Engineering, Mass. Inst. of Technology, Cambridge, 296 pp

Wood, E.F. and Rodriquez-Iturbe, I. (1975a). Bayesian Inference and Decision Making for Extreme Hydrologic Events. Water Resources Research, 11(4), 533-542.

Wood, E.F. and Rodriquez-Iturbe, I. (1975b). A Bayesian Approach to Analysing Uncertainty Among Flow Frequency Models. Water Resources Research, 11(6), 839-843.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. John Wiley, New York.

APPENDIX A

Predictive Moments of a 2 parameter probability distribution $f(x|\mu,\sigma)$ when both μ and σ are uncertain, and

$$-\infty<\mu,\ x<\infty\ ;\ \sigma>0$$

$$\tilde{f}(\chi) = \int_0^{\infty}\int_{-\infty}^{\infty} f(x|\mu,\sigma)\ \Pi(\mu,\sigma)\ d\mu d\sigma$$

Predictive mean of $x = \tilde{m}_x$

$$\tilde{m}_x = \int_{-\infty}^{\infty} x\ \tilde{f}(x)\ dx$$

$$= \int_{-\infty}^{\infty} x \left[\int_0^{\infty}\int_{-\infty}^{\infty} f(x|\mu,\sigma)\ \Pi\ (\mu,\sigma)\ d\mu d\sigma \right] dx$$

$$= \int_0^{\infty}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x\ f(x|\mu,\sigma)dx\right]\ \Pi\ (\mu,\sigma)d\mu d\sigma$$

$$= \int_0^{\infty}\int_{-\infty}^{\infty} m_{x|\mu,\sigma}\ \Pi(\mu,\sigma)\ d\mu,d\sigma$$

$$= \int_0^{\infty}\int_{-\infty}^{\infty} \mu\ \Pi\ (\mu,\sigma)\ d\mu d\sigma$$

Discrete case:

$$\tilde{m}_x = \sum_j^m \sum_j^m \mu_i\ P(\mu_i,\sigma_j)$$

Predictive variance of $x = \tilde{\sigma}_x^2$

$$= \tilde{E}(x^2) - \tilde{m}_x^2$$

$$\tilde{E}(x^2) = \int_{-\infty}^{\infty} x^2 \left[\int_0^{\infty} \int_{-\infty}^{\infty} f(x|\mu,\sigma)\ \Pi\ (\mu,\sigma)\ d\mu d\sigma \right] dx$$

$$= \int_0^{\infty} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x^2\ f(x|\mu,\sigma) dx \right] \Pi\ (\mu,\sigma)\ d\mu d\sigma$$

$$= \int_0^{\infty} \int_{-\infty}^{\infty} (\sigma^2 + \mu^2)\ \Pi\ (\mu,\sigma)\ d\mu d\sigma$$

$$= \int_0^{\infty} \int_{-\infty}^{\infty} \sigma^2\ \Pi\ (\mu,\sigma)\ d\mu d\sigma + \int_0^{\infty} \int_{-\infty}^{\infty} \mu^2\ \Pi\ (\mu,\sigma)\ d\mu d\sigma$$

$$\tilde{\sigma}_x^2 = \int_0^{\infty} \int_{-\infty}^{\infty} \sigma^2\ \Pi\ (\mu,\sigma)\ d\mu d\sigma + \int_0^{\infty} \int_{-\infty}^{\infty} \mu^2\ \Pi\ (\mu,\sigma)\ d\mu d\sigma - \tilde{m}_x^2$$

$$\tilde{\sigma}_x^2 = \int_0^{\infty} \int_{-\infty}^{\infty} \sigma^2\ \Pi\ (\mu,\sigma)\ d\mu d\sigma + \int_0^{\infty} \int_{-\infty}^{\infty} (\mu - \tilde{m}_x)^2\ \Pi\ (\mu,\sigma)\ d\mu d\sigma$$

Discrete case:

$$\tilde{\sigma}_x^2 = \sum_j^m \sum_i^n \sigma_j^2\ P\ (\mu_i,\sigma_j) + \sum_j^m \sum_i^n (\mu_i - \tilde{m}_x)^2\ P\ (\mu_i,\sigma_j)$$

APPENDIX B

Bayesian Approximation
(adapted from Sinha, 1986a)

Lindley (1980) discussed the approximate evaluation of the ratio of integrals of the form

$$I = \frac{\int_{\Omega} w(\theta) e^{L(\theta)} d\theta}{\int_{\Omega} v(\theta) e^{L(\theta)} d\theta}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ is the parameter of interest, $L(\theta)$ is the logarithm of the likelihood function, $w(\theta)$ and $v(\theta)$ are arbitrary functions of θ and Ω represents the range space of the parameter θ.

Let

$$L_{ijk} = \frac{\partial^3 L}{\partial\theta_i \partial\theta_j \partial\theta_k}, \qquad w_{ijk} = \frac{\partial^3 w}{\partial\theta_i \partial\theta_j \partial\theta_k},$$

$$W_i(\theta) = \frac{w_i(\theta)}{w(\theta)}, \qquad w(\hat{\theta}) \neq 0.$$

At the MLE $\hat{\theta}$, $L_i = \frac{\partial L}{\partial\theta_i} = 0$ at $\hat{\theta}$.

Expanding $L(\theta)$ and $w(\theta)$ by Taylor's series about $\hat{\theta}$, we have

$$L(\theta) = L(\hat{\theta}) + \sum_i (\theta_i - \hat{\theta}_i) L_i + \frac{1}{2} \sum_i \sum_j (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) L_{ij} +$$

$$+ \frac{1}{6} \sum_i \sum_j \sum_k (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)(\theta_k - \hat{\theta}_k) L_{ikj} +$$

$$+ \frac{1}{24} \sum_i \sum_j \sum_k \sum_\ell (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)(\theta_k - \hat{\theta}_k)(\theta_\ell - \hat{\theta}_\ell) L_{ijk\ell} + \ldots$$

$$w(\theta) = w(\hat{\theta}) + \sum_i (\theta_i - \hat{\theta}_i) w_i + \frac{1}{2} \sum_i \sum_j (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) w_{ij} +$$

$$+ \frac{1}{6} \sum_i \sum_j \sum_k (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)(\theta_k - \hat{\theta}_k) w_{ijk} +$$

$$+ \frac{1}{24} \sum_i \sum_j \sum_k \sum_\ell (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)(\theta_\ell-\hat{\theta}_\ell)w_{ijk\ell} + \ldots$$

$$\int w(\theta)e^{L(\theta)}d\theta = w(\hat{\theta})\,e^{L(\hat{\theta})} \int e^{\frac{1}{2}\sum_i \sum_j (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)} \left[1 + \sum_i(\theta_i-\hat{\theta}_i)W_i + \right.$$

$$+ \frac{1}{2} \sum_i \sum_j (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)W_{ij} + \frac{1}{6} \sum_i \sum_j \sum_k$$

$$\left. (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)W_{ijk} + \ldots \right] \cdot$$

$$\left[1 + \frac{1}{6} \sum_i \sum_j \sum_k (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)L_{ijk} \right.$$

$$\left. + \frac{1}{24} \sum_i \sum_j \sum_k \sum_\ell (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)(\theta_\ell-\hat{\theta}_\ell)L_{ijk\ell} + \ldots \right] d\theta$$

$$= w(\hat{\theta})e^{L(\hat{\theta})} \int e^{\frac{1}{2}\sum_i \sum_j (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)} \left[1 + \sum_i(\theta_i-\hat{\theta}_i)W_i \right.$$

$$+ \frac{1}{2} \sum_i \sum_j (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)W_{ij} + \frac{1}{6} \sum_i \sum_j \sum_k (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)L_{ijk}$$

$$+ \frac{1}{6} \left\{ \sum_i(\theta_i-\hat{\theta}_i)W_i \right\} \sum_i \sum_j \sum_k (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)L_{ijk}$$

$$\left. + \ldots \right] d\theta. \qquad \text{(B.2)}$$

Using $e^{\frac{1}{2}\sum_i \sum_j (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)L_{ij}}$ proportional to the density of multivariate normal distribution, the integration involves the moments of this distribution where the precision matrix has elements - L_{ij}^{-1}.

Let σ_{ij} = (i,j) th element of the inverse of the matrix $[-L_{ij}]$.

For multivariate normal distribution

$$E\left\{(\theta_i-\hat{\theta}_i)\right\} = 0, \quad E\left\{(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)\right\} = \sigma_{ij}.$$

$$E\left\{(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)\right\} = 0. \quad \text{Also}$$

$$E\left\{(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)(\theta_\ell-\hat{\theta}_\ell)\right\} = \sigma_{ij}\sigma_{k\ell} + \sigma_{ik}\sigma_{j\ell} + \sigma_{jk}\sigma_{i\ell}.$$

Substituting in (B.2)

$$\int W(\theta)e^{L(\theta)}d\theta = w(\hat{\theta})e^{L(\hat{\theta})}|\Sigma|^{\frac{1}{2}}(2\Pi)^{\frac{m}{2}}\left[1 + \frac{1}{2}\sum_i\sum_j W_{ij}\sigma_{ij} + \right.$$

$$\left. + \frac{1}{6}\sum_i\sum_j\sum_k\sum_\ell W(\sigma_{ij}\sigma_{k\ell} + \sigma_{ik}\sigma_{j\ell} + \sigma_{jk}\sigma_{i\ell})L_{ijk} + \cdots\right]$$

$$= w(\hat{\theta})e^{L(\hat{\theta})}|\Sigma|^{\frac{1}{2}}(2\Pi)^{\frac{m}{2}}\left[1 + \frac{1}{2}\sum_i\sum_j W_{ij}\sigma_{ij}\right.$$

$$\left. + \frac{1}{2}\sum_i\sum_j\sum_k\sum_\ell W_\ell\sigma_{ij}\sigma_{k\ell}L_{ijk} + \cdots\right] \qquad \text{(B.3)}$$

Since L_{ijk} is unchanged by permutation of its suffixes. Similarly

$$\int v(\theta)e^{L(\theta)}d\theta = v(\hat{\theta})e^{L(\hat{\theta})}|\Sigma|^{\frac{1}{2}}(2\Pi)^{\frac{m}{2}}\left[1 + \frac{1}{2}\sum_i\sum_j V_{ij}\sigma_{ij}\right.$$

$$\left. + \frac{1}{2}\sum_i\sum_j\sum_k\sum_\ell V_\ell\sigma_{ij}\sigma_{k\ell}L_{ijk} + \cdots\right] \qquad \text{(B.4)}$$

where $V_i(\theta) = \dfrac{v_i(\theta)}{v(\hat{\theta})}$, $v(\hat{\theta}) \neq 0$.

From (B.3) and (B.4) we have

$$I = \frac{u(\hat{\theta})}{v(\hat{\theta})}\left[1 + \frac{1}{2}\sum_i\sum_j W_{ij}\sigma_{ij} + \frac{1}{2}\sum_i\sum_j\sum_k\sum_\ell W_\ell\sigma_{ij}\sigma_{k\ell}L_{ijk} + \ldots\right]\left[1 + \frac{1}{2}\sum_i\sum_j V_{ij}\sigma_{ij} + \frac{1}{2}\sum_i\sum_j\sum_k\sum_\ell V_\ell\sigma_{ij}\sigma_{k\ell}L_{ijk} + \ldots\right]^{-1}$$

$$= \frac{u(\hat{\theta})}{v(\hat{\theta})}\left[1 + \frac{1}{2}\sum_i\sum_j (W_{ij}-V_{ij})\sigma_{ij} + \frac{1}{2}\sum_i\sum_j\sum_k\sum_\ell (W_\ell-V_\ell)\sigma_{ij}\sigma_{k\ell}L_{ijk} + \ldots\right] \qquad (B.5)$$

Let $w(\theta) = v(\theta)u(\theta)$. For simplicity, θ is dropped,

$$w = uv$$

$$w_i = u_i v + uv_i$$

$$w_{ij} = u_{ij}v + u_i v_j + u_j v_i + uv_{ij}$$

$$\frac{w_{ij}}{w} = \frac{u_{ij}}{u} + \frac{u_i v_j + u_j v_i}{uv} + \frac{v_{ij}}{v} .$$

Let $\rho(\theta) = \log v(\theta)$

$$v_i = v\rho_i , \quad v_j = v\rho_j$$

$$W_{ij} - V_{ij} = \frac{u_{ij}}{u} + \frac{u_i\rho_j + u_j\rho_i}{u}$$

$$W_\ell - V_\ell = \frac{u_\ell}{u} .$$

Hence from (B.5) given the data $\underline{x} = (x_1, x_2, \ldots, x_n)$, one obtains

$$E\left\{u(\theta)|\underline{x}\right\} = u\left[1 + \frac{1}{2u}\sum_i\sum_j (u_{ij} + u_i\rho_j + u_j\rho_i)\sigma_{ij}\right.$$

$$\left. + \frac{1}{2u}\sum_i\sum_j\sum_k\sum_\ell L_{ijk}u_\ell\sigma_{ij}\sigma_{k\ell} + \ldots\right]$$

$$= u + \frac{1}{2}\sum_i\sum_j (u_{ij} + 2u_i\rho_j)\sigma_{ij} +$$

$$+ \frac{1}{2}\sum_i\sum_j\sum_k\sum_\ell L_{ijk}u_i\sigma_{ij}\sigma_{k\ell} + \text{terms of the order } \frac{1}{n^2}$$

and smaller ...

all functions to be evaluated at the MLE $\hat{\theta}$.

APPENDIX C

Bayesian Approximation Constants of the Gumbel Distribution

$$L_{30} = -\frac{2n}{\sigma^3} + \frac{6n}{\sigma^4}(\bar{x}-\mu) + \frac{6}{\sigma^5}\sum_{i=1}^{n}[(x_i-\mu)^2\exp-(\frac{x_i-\mu}{\sigma})]$$

$$-\frac{1}{\sigma^6}\sum_{i=1}^{n}[x_i-\mu)^3\exp-(\frac{x_i-\mu}{\sigma})] - \frac{6}{\sigma^4}\sum_{i=1}^{n}[(x_i-\mu)\exp-(\frac{x_i-\mu}{\sigma})]$$

$$L_{03} = \frac{1}{\sigma^3}\sum_{i=1}^{n}\exp-(\frac{x_i-\mu}{\sigma}) = -\frac{n}{\sigma^3}$$

$$L_{21} = \frac{4}{\sigma^4}\sum_{i=1}^{n}[(x_i-\mu)\exp-(\frac{x_i-\mu}{\sigma})] - \frac{1}{\sigma^5}\sum_{i=1}^{n}[(x_i-\mu)^2\exp-(\frac{x_i-\mu}{\sigma})]$$

$$L_{12} = \frac{2n}{\sigma^3} - \frac{1}{\sigma^4}\sum_{i=1}^{n}[(x_i-\mu)\exp-(\frac{x_i-\mu}{\sigma})]$$

$$u = P_q = 1 - \exp\left\{-\exp-(\frac{q-\mu}{\sigma})\right\}$$

$$u_1 = \frac{\partial u}{\partial \sigma} = (\frac{q-\mu}{\sigma^2})\,\omega_1 \cdot \omega_2$$

$$u_2 = \frac{\partial u}{\partial \sigma} = \frac{1}{\sigma}\cdot\omega_1\cdot\omega_2$$

$$u_{12} = \frac{\partial^2 u}{\partial\sigma\partial\mu} = (\frac{q-\mu}{\sigma^3})\cdot\omega_2[-\sigma u_2+\omega_1-\frac{\omega_1\sigma}{(q-\mu)}]$$

$$u_{11} = \frac{\partial^2 u}{\partial\sigma^2} = -(\frac{q-\mu}{\sigma^4})\cdot\omega_2[\sigma^2 u_1-\omega_1(q-\mu)+2\sigma\omega_1]$$

$$u_{22} = \frac{\partial^2 u}{\partial \mu^2} = -\frac{1}{\sigma} \cdot \omega_2 \cdot u_2$$

$$v = u^2 = P_q^2$$

$$v_1 = 2u \cdot u_1$$

$$v_2 = 2u \cdot u_2$$

$$v_{11} = 2u \cdot u_{11} + 2u_1^2$$

$$v_{12} = 2uu_{12} + 2u_1 u_2$$

$$v_{22} = 2uu_{22} + 2u_2^2$$

where:

$$\omega_1 = \exp[-\exp-(\frac{q-\mu}{\sigma})]$$

$$\omega_2 = \exp-(\frac{q-\mu}{\sigma})$$

Variance-Covariance Matrix of the 3 Parameter Lognormal Distribution
[after Kite (1977)]

$$L = -\frac{n}{2}\log 2\Pi - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\log(x_i-a) - \sum_{i=1}^{n}\frac{[\log(x_i-a)-\mu]^2}{2\sigma^2}$$

$$\frac{\partial^2 L}{\partial a^2} = \frac{1}{\sigma^2}\left[(\sigma^2 - \mu - 1)\sum_{i=1}^{n}(x_i-a)^{-2} + \sum_{i=1}^{n}\log(x_i-a)(x_i-a)^{-2}\right]$$

$$\frac{\partial^2 L}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 L}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} - \sum_{i=1}^{n}\frac{[\log(x_i-a)-\mu]^2}{\sigma^6} = \frac{-n}{2\sigma^4}$$

$$\frac{\partial^2 L}{\partial a\partial\mu} = \frac{-1}{\sigma^2}\sum_{i=1}^{n}(x_i-a)^{-1}$$

$$\frac{\partial^2 L}{\partial a\partial\sigma^2} = -\frac{1}{\sigma^4}\left[\sum_{i=1}^{n}[\log(x_i-a)-\mu](x_i-a)^{-1}\right]$$

$$\frac{\partial^2 L}{\partial\mu\partial\sigma^2} = -\sum_{i=1}^{n}\frac{[\log(x_i-a)-\mu]}{\sigma^4} = 0$$

$$[\sigma_{ij}] = \begin{bmatrix} -\frac{\partial^2 L}{\partial a^2} & -\frac{\partial^2 L}{\partial a \partial \mu} & -\frac{\partial^2 L}{\partial a \partial \sigma^2} \\ & -\frac{\partial^2 L}{\partial \mu^2} & -\frac{\partial^2 L}{\partial \mu \partial \sigma^2} \\ & & -\frac{\partial^2 L}{\partial (\sigma^2)^2} \end{bmatrix}^{-1}$$

$$= \frac{n}{\sigma^2} \begin{bmatrix} (1+\sigma^2)e^{2\sigma^2-2\mu} & e^{\frac{\sigma^2}{2}-\mu} & -e^{\frac{\sigma^2}{2}-\mu} \\ & 1 & 0 \\ & & \frac{1}{2\sigma^2} \end{bmatrix}^{-1}$$

$$[\sigma_{ij}] = \begin{bmatrix} \frac{1}{2nD} & -\frac{e^{\frac{\sigma^2}{2}-\mu}}{2nD} & \frac{\sigma^2 e^{\frac{\sigma^2}{2}-\mu}}{nD} \\ & \frac{\sigma^2}{nD}\left[\frac{(\sigma^2+1)}{2\sigma^2}e^{2(\sigma^2-\mu)} - e^{\sigma^2-2\mu}\right] & -\frac{\sigma^2}{nD}e^{\sigma^2-2\mu} \\ & & \frac{\sigma^2}{nD}\left[(\sigma^2+1)e^{2(\sigma^2-\mu)} - e^{\sigma^2-2\mu}\right] \end{bmatrix}$$

where $D = \frac{1+\sigma^2}{2\sigma^2} e^{2(\sigma^2-\mu)} - \frac{e^{\sigma^2-2\mu}}{2\sigma^2}(1+2\sigma^2)$

APPENDIX E

Bayesian Approximation constants of the 3 Parameter lognormal distribution

$$u = a + e^{\mu+\sigma t}$$

$$u_1 = \frac{\partial u}{\partial a} = 1$$

$$u_2 = \frac{\partial u}{\partial \mu} = e^{\mu+\sigma t}$$

$$u_3 = \frac{\partial u}{\partial \sigma^2} = \frac{t\, e^{\mu+\sigma t}}{2\sigma}$$

$$u_{12} = \frac{\partial^2 u}{\partial a \partial \mu} = u_{13} = \frac{\partial^2 u}{\partial a \partial \sigma^2} = 0$$

$$u_{22} = \frac{\partial^2 u}{\partial \mu^2} = e^{\mu+\sigma t}$$

$$u_{23} = \frac{\partial^2 u}{\partial \mu \partial \sigma^2} = u_3$$

$$u_{33} = \frac{\partial^2 u}{\partial (\sigma^2)^2} = \frac{t^2}{4\sigma^2} e^{\mu+\sigma t} - \frac{t}{4\sigma^3} e^{\mu+\sigma t}$$

Let $p = e^{\mu+\sigma t}$

$$u_2 = u_{22} = p \quad ; \quad u_{33} = \frac{tp}{4\sigma^3} (\sigma t - 1)$$

$$u_3 = u_{23} = \frac{tp}{2\sigma}$$

$$u = (a+e^{\mu+\sigma t})^2 = a^2 + 2ae^{\mu+\sigma t} + e^{2(\mu+\sigma t)}$$

$$u_1 = \frac{\partial u}{\partial a} = 2a + 2e^{\mu+\sigma t} = 2(a+p) \; ; \; u_{11} = 2$$

$$u_2 = \frac{\partial u}{\partial \mu} = 2ae^{\mu+\sigma t} + 2e^{2(\mu+\sigma t)} = 2p(a+p)$$

$$u_{22} = \frac{\partial u}{\partial \mu^2} = 2ae^{\mu+\sigma t} + 4e^{2(\mu+\sigma t)} = 2p(2+2p)$$

$$u_{23} = \frac{\partial^2 u}{\partial \mu \partial \sigma^2} = \frac{tae^{\mu+\sigma t}}{\sigma} + \frac{2te^{2(\mu+\sigma t)}}{\sigma} = \frac{tp}{\sigma}(a+2p)$$

$$u_{12} = \frac{\partial^2 u}{\partial a \partial \mu} = 2e^{\mu+\sigma t} = 2p$$

$$u_{13} = \frac{\partial^2 u}{\partial a \partial \sigma^2} = \frac{2te^{\mu+\sigma t}}{2\sigma} = \frac{tp}{\sigma}$$

$$u_3 = \frac{\partial u}{\partial \sigma^2} = \frac{ate^{\mu+\sigma t}}{\sigma} + \frac{te^{2(\mu+\sigma t}}{\sigma} = \frac{tp}{\sigma}(a+p)$$

$$u_{33} = \frac{\partial^2 u}{\partial (\sigma^2)^2} = \frac{at^2e^{\mu+\sigma t}}{2\sigma^2} - \frac{ate^{\mu+\sigma t}}{2\sigma^3} + \frac{t^2e^{2(\mu+\sigma t)}}{\sigma^2}$$

$$- \frac{te^2}{2\sigma^3}{}^{2(\mu+\sigma t)}$$

$$= \frac{tp}{2\sigma^3}(\sigma at - a + 2tp\sigma - p)$$

$$L_{111} = \frac{\partial^3 L}{\partial a^3} = \frac{1}{\sigma^2}\left\{(2\sigma^2 - 2\mu - 3)\sum_{i=1}^{n}(x_i-a)^{-3} + 2\sum_{i=1}^{n}\log(x_i-a)(x_i-a)^{-3}\right\}$$

$$L_{112} = \frac{\partial^3 L}{\partial a^2 \partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-a)^{-2}$$

$$L_{113} = \frac{\partial^3 L}{\partial a^2 \partial \sigma^2} = \frac{1}{\sigma^4}\left[\sum_{i=1}^{n}(x_i-a)^{-2} - \sum_{i=1}^{n}\log(x_i-a)(x_i-a)^{-2} + \mu\sum_{i=1}^{n}(x_i-a)^{-2}\right]$$

$$L_{122} = \frac{\partial^3 L}{\partial a \partial \mu^2} = 0$$

$$L_{133} = \frac{\partial^3 L}{\partial a \partial (\sigma^2)^2} = \frac{2}{\sigma^6}\left[\sum_{i=1}^{n}\log(x_i-a)(x_i-a)^{-1} - \mu\sum_{i=1}^{n}(x_i-a)^{-1}\right]$$

$$L_{222} = \frac{\partial^3 L}{\partial \mu^3} = 0$$

$$L_{223} = \frac{\partial^3 L}{\partial \mu^2 \partial \sigma^2} = \frac{n}{\sigma^4}$$

$$L_{333} = \frac{\partial^3 L}{\partial (\sigma^2)^3} = -\frac{n}{\sigma^6} + 3\sum_{i=1}^{n}\frac{[\log(x_i-a) - \mu]^2}{\sigma^8} = \frac{2n}{\sigma^6}$$

$$L_{123} = \frac{\partial^3 L}{\partial a \partial \mu \partial \sigma^2} = \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-a)^{-1}$$

$$[L_{ijk}] = \begin{bmatrix} L_{111} & L_{112} & L_{113} \\ L_{221} & L_{222} & L_{223} \\ L_{331} & L_{332} & L_{333} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{n}{\sigma^2} e^{9\sigma^2/2 - 3\mu}(4\sigma^2+3) & -\frac{n}{\sigma^2} e^{2(\sigma^2-\mu)} & \frac{n}{\sigma^4} e^{2(\sigma^2-\mu)}(2\sigma^2+1) \\ 0 & 0 & \frac{n}{\sigma^4} \\ -\frac{2n}{\sigma^4} e^{\sigma^2/2 - \mu} & 0 & \frac{n}{\sigma^6} \end{bmatrix}$$

AND $L_{123} = \frac{ne}{\sigma^4}^{\sigma^2/2 - \mu}$

APPENDIX F

Bayesian Approximation Constants of the 2 Parameter Lognormal Distribution

$$f(x \mid \mu,\sigma) = \frac{1}{\sqrt{2\Pi}\ \sigma\ x}\ e^{-\frac{1}{2\sigma^2}(\log x-\mu)^2}$$

$$x>0,\ -\infty<\mu<\infty\ ,\quad \sigma>0$$

$$P_q = \int_q^{\infty} \frac{1}{\sqrt{2\Pi}\ x\ \sigma} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right] dx$$

Let $u = P_q$

$$u_1 = \frac{\partial u}{\partial \mu} = \frac{1}{\sqrt{2\Pi}\ \sigma} \exp\left[-\frac{1}{2\sigma^2}(\log q - \mu)^2\right]$$

$$u_2 = \frac{\partial u}{\partial \sigma} = \frac{\log q-\mu}{\sigma} \cdot u_1$$

$$u_{11} = \frac{\partial^2 u}{\partial \mu^2} = \frac{\log q-\mu}{\sigma^2} \cdot u_1$$

$$u_{12} = \frac{\partial^2 u}{\partial \mu \partial \sigma} = \left[\frac{(\log q-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right] \cdot u_1$$

$$u_{22} = \frac{\partial^2 u}{\partial \sigma^2} = [(\log q-\mu)^2 - 2\sigma^2] \cdot \frac{u_2}{\sigma^3}$$

$$L = -\frac{n}{2}\log 2\Pi - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\log x_i - \sum_{i=1}^{n}(\log x-\mu)^2/2\sigma^2$$

$$L_{12} = \frac{\partial^3 L}{\partial\mu\partial\sigma^2} = 0$$

$$L_{21} = \frac{\partial^3 L}{\partial\mu^2\partial\sigma} = \frac{2n}{\sigma^3}$$

$$L_{30} = \frac{\partial^3 L}{\partial\mu^3} = 0$$

$$L_{03} = \frac{\partial^3 L}{\partial\sigma^3} = \frac{10n}{\sigma^3}$$

APPENDIX G

HURST PHENOMENON

The use of the Hurst coefficient as a measure of long term serial correlation is introduced in this appendix. The estimation of the Hurst coefficient is discussed, and the test of significance of the estimated Hurst coefficient is presented.

In his study of the long term storage capacity of reservoirs, H.E. Hurst (1951, 1956) employed a statistic called the "range of cumulative departures from the sample mean", which equals the required storage volume of a reservoir which for a given inflow sequence can release in every year the mean inflow.

Let $x_1, x_2, \ldots x_n$ be a sequence of annual inflows into a reservoir over n years. Let the mean flow in the n year period be denoted by:

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

The accumulated departure of the flows from the mean flow after y years is:

$$S_y = \sum_{i=1}^{y} (x_i - \bar{x}_n)$$

In the last period, $S_n = 0$. The range of the cumulative departures from the mean is:

$$R_n = \max_y (S_y) - \min_y (S_y) = S_M - S_m$$

where S_M and S_m are the largest and smallest values in the set (S_y). Figure G1 illustrates, through the mass curve, the relationships between S_y and R_n.

Hurst studied how the average value of R_n changes as a function of n and found that the expected value of R_n divided by the standard deviation S_n of the n annual inflows is proportional to n raised to some power h.

$$E\left(\frac{R_n}{S_n}\right) \sim n^h \qquad \ldots \text{G1}$$

The exponent h which varies between 0 and 1 is called the Hurst statistic. The ratio R_n/S_n is called the rescaled range.

In addition to river discharges, Hurst investigated a host of other natural geophysical time series ranging from tree rings to clay varves. All in all, 75 different phenomenon were used. The total number of series was close to 900 and they vary in length from 40 to 2000 years.

Equation G1 implies that the relationship between log E(R/s) and log n is linear with slope h. See Figure G2. To determine h, Hurst defined:

$$\frac{R_n}{S_n} = \left(\frac{n}{2}\right)^K$$

or

$$K = \log\left(\frac{R_n}{S_n}\right) / \log\left(\frac{n}{2}\right) \qquad \ldots \text{G2}$$

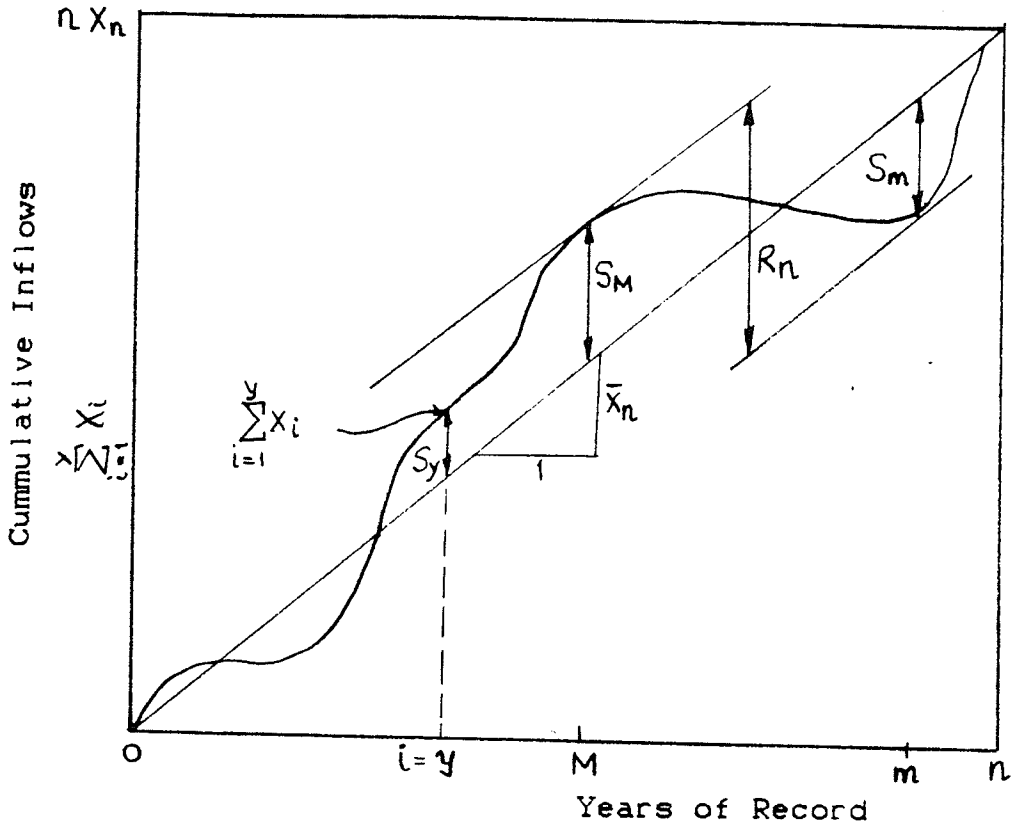


Figure G1 Mass Curve Representation of Range of Cumulative Inflows

(after Loucks et al., 1981)

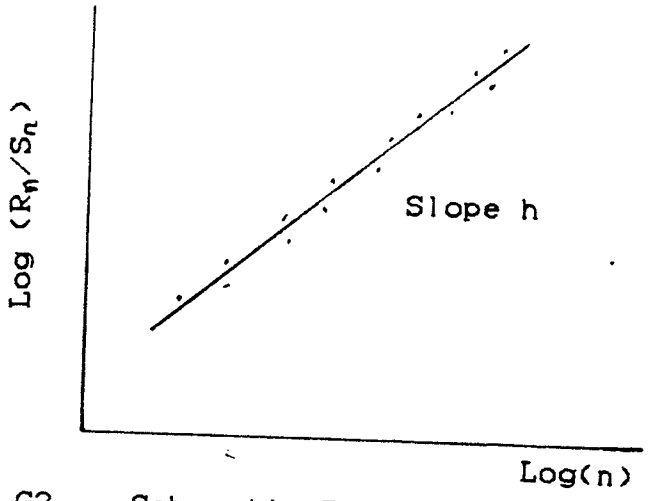


Figure G2 Schematic Plot of Hurst's Law

(after O'Connell, 1977)

where K represents an estimate of h for each of the 900 time series he investigated.[1] Over all phenomena Hurst's K was found to have an average value of 0.73 with a standard deviation of 0.08. Asymptotically, for independent normal random variables, Hurst (1956) and Feller (1951) showed that:

---

[1] Hurst used equation G2 so that E(R/S) = 1 for n = 2 in his attempts to find the best possible fit to the data used, but this restriction may cause bias. In view of its simplicity, however, many researchers still prefer it over the more general expression equation G1. Mandelbrot and Wallis (1969) proposed a graphical procedure for estimating h, through a so called pox diagram. To construct the pox diagram, the series is divided into a number of sub-series of length n' on double logarithmic paper. An estimate of h denoted by H is obtsined as the slope of a least squares line fitted to the logarithms of the mean values of $R_{n'}/S_{n'}$.

Mandelbrot and Wallis contend that actual pox diagrams have a straight trend line of slope H that fails to pass through the point of abscissa log 2 and ordinate 0. As such, Hurst's K is thus a very poor estimate of slope h. It tends to be too low when h > 0.72 and too high when h < 0.72. As a result, the trend line and Hurst's method may both suggest identical values for h, but Hurst's method will greatly underrate the variability of h around its typical value.

It must be pointed out that the use of H as an estimator of h would invariably introduce a subjective element into the computation. As such, different investigators may have a different estimate. Usually these differences are quite small. The sampling properties of both K and H for small samples have been extensively studied by Wallis and Matalas (1971). They found that K and H are both biased estimators of h. K is more biased but less variable and H is less biased but more variable. However, both K and H are unbiased around 0.7. They also found that the bias for both K and H decreases with increasing n but at a very slow rate. Only Hurst's K is used in this study in view of its simplicity.

Hipel and McLeod (1975) in an effort to standardize the estimation of h have proposed a maximum likelihood method. Although it is statistically appealing, the procedure is rather involved and it has not been widely used.

$$E(R_n) = (\pi/2)^{1/2} \sigma n^{1/2} \quad \ldots G3$$

In other words, $h \rightarrow 0.5$ as n becomes large. But Hurst also made the far reaching discovery that for many of the natural series he investigated, the slope h remains much steeper than 0.5 even for large values of n. The failure of natural series to accord with theory is termed the "Hurst phenomenon". This so called phenomenon generated considerable interest among hydrologists and mathematicians alike since it indicates a puzzling long term "memory" or "persistence" in the random process that generated the series. In the literature, a number of possible explanations for the Hurst phenomenon have been offered. So far none of these have been conclusive or statisfactory. Four of the most popular explanations are:

i) skewness

ii) transience

iii) non-stationarity

iv) autocorrelation

Hypothesis (i) may be discarded on the basis of studies by Matalas and Huzzen (1967) and Mandelbrot and Wallis (1969), who showed that the behaviour of the rescaled range for a number of stochastic processes is unaffected by the marginal distribution of the process. However, some erroneous explanations of the Hurst phenomenon have been based on the unscaled range R_n which is affected by the marginal distribution unlike the rescaled range R_n/S_n.

With regard to hypothesis (ii), the implication is that, if sufficiently long time series were available in nature, h would tend to a value of 0.5 corresponding to asymptotic independence. Rejection or acceptance of this hypothesis must await the availability of longer geophysical records. At this stage hypothesis (iii) and (iv) must be considered as either of these could be advanced as an explanation of the Hurst phenomenon.

In an attempt to account for his own findings, Hurst (1957) used a crude form of non-stationarity in the mean of simulated series, resulting in some very approximate agreements with the behaviour of the rescaled range observed in nature. Klemes (1974) used a similar but more sophisticated technique to arrive at the same conclusion that non-stationarity may be an explanation of the Hurst phenomenon. However, it the ultimate aim is to generate synthetic series, non-stationarity is a rather intractable assumption. It would be more desirable to use stationary stochastic processes which can reproduce the Hurst phenomenon for application in the planning of water resources system, unless strong physical grounds indicate the contrary (Lawrence and Kottegoda, 1977).

Early attempts at explaining the Hurst phenomenon through hypothesis (iv), addressed the issue of autocorrelation, and employed models of the autoregressive type, with the lag-one autoregressive or Markov process. However, these attempts were largely unsuccessful. In the

case of the lag-one Markov process, the autocorrelation function ρ_k is given by:

$$\rho_k = \rho(1)^k \qquad \ldots G4$$

A typical value of $\rho(1)$ for annual streamflow is 0.3 (for peakflows it is $\approx$ 0.0); hence ρ_k approaches zero very quickly and the memory of such a process is extremely short. These processes which include the more complex ARMA process are usually termed short memory models. The behaviour of the rescaled range for these short memory models is characterized by a short initial transient, where for small values of n, $h > 0.5$, followed by a break to the classical $h = 0.5$ law. Such behaviour is not in conformance with natural time series where no convergence to the $h = 0.5$ law has been observed. Fiering (1967) applied a multi-lag autoregressive model in an effort to reproduce the Hurst phenomenon. He found that he required a 20-lag model with $h > 0.5$ held for $n \leq 60$. Computational not statistical grounds prevented an extension of this approach.

Modellers were at a loss for some time to develop a model that reproduced the Hurst phenomenon until Mandelbrot and Van Ness (1965) developed a procedure that produces flows with a specified value of h. Their model is based fractional Brownian motion which can be used to obtain fractional Gaussian noise. This procedure is briefly described in Chapter 6.

HURST COEFFICIENT AS A MEASURE OF SERIAL CORRELATION

Coming back to the use of equation G1, the following can be deduced.

a) The greater the variability of X_i, the greater the value of R_n.

b) The longer the period, the more severe the wettest and driest period encountered will be, even if the data are completely independent. Hence as $n \rightarrow \infty$, the storage required would also tend to be infinte.

c) More importantly R_n/S_n also depends on the degree of clustering of wet and dry years caused by the serial correlation structure. Serial correlation whether short or long term, tends to increase the length and severity of dry and wet periods.

d) In terms of h, the higher the value of h the more persistent the series. Hence, a high Hurst statistic is synonymous with long term serial correlation.

VARIABILITY OF THE MEAN AND STANDARD DEVIATION

In addition to the expectation of the rescaled range, Hurst (1951) also compared the variability of means for 50 years and 100 years with what would be expected if the observations were independent, and not affected by serial correlation. It appears in his finding that the means are much more variable than they would be in the case of serially independent observations. The standard deviation of a 50-year mean being 2.5 times as great as it would be for serially independent observations. Similarly, the standard deviation of a 100-year mean is 3.2 times as great as in serially independent observations. This implies that, over a long period, the maximum value is likely to be higher and the minimum value lower than would be predicted from the application of the ordinary theory of probability to the records from a short period. Similar results were obtained for the standard deviation of the standard deviations. For the 50-year periods, the standard deviation of the standard deviations is 3.2 times higher than if the observations were serially independent, and for the 100-year periods, it is 4.6 times as great. Hence it appears that the standard deviation for these natural phenomena is more variable than one would expect in the case of independent observations.

TESTING THE SIGNIFICANCE OF K AND H

Like any other statistic, the computed value of K or H must be tested for significance.

SIGNIFICANCE TEST FOR K

It may be recalled from Equation G2 that Hurst's K is obtained from the rescaled range R/S for a given n. The distribution of R/S is known to be highly skewed when n is small (Mandelbrot and Wallis, 1969). Therefore, to use R/S as a test of statistical dependence, it is necessary to know not just $E(R/S)$ and $E(R/S)^2$ but the whole distribution. A closed form solution to the distribution of R/S for small n has not been obtained yet. However, it is feasible to obtain usable approximations by Monte Carlo simulation for any desired process.

The paper by Wallis and O'Connell (1973), gives in graphical form, the distribution of R/S for both normal independent processes and lag-one Markov processes for various values of n. The results were based upon 30,000 replications. Hence, one can use these figures to ascertain probability tests for R/S or K as a function of n for an independent process or a Markov process, and observed values of R/S or K from unknown distribution can then be compared with these levels. In essence the

hypothesis H_o being tested is whether or not a given value of a statistic, R/S or K, could reasonably be expected to have occurred if the generating mechanism were, say, a normal independent process. If only the probability of exceeding a certain value of R/S is considered then we have a one tailed test for the existence of persistence. Wallis et al., (1973) also considered the possibility of using the R/S statistic to distinguish short term persistence from long term persistence with a discussion of its power.

SIGNIFICANCE TEST FOR H

As pointed out earlier there is no protocol for the calculation of H. Generalised graphs like those for testing R/S or K cannot therefore be developed. However, one can still test the significance of H by the Monte Carlo procedure. If the null hypothesis is that of independence, then we can generate a large number, say 1000 replications, of independent data with the same length as the sample and compute H in exactly the same way as was used to calculate the sample H. The distribution of the 1000 Hs can then be plotted on probability paper. Depending on the level of significance, it can read off from the graph whether or not to accept of reject the null hypothesis.

APPENDIX H

SKEWED MIXED-NOISE VARIATES

The necessary skewness in the MN variates may be obtained in different ways as described below.

The MN process (6.39) can be written as:

$$X_t = aX_t^{(H)} + bX_t^{(M)} + cX_t^{(L)} \qquad \ldots (H.1)$$

Cubing both sides and taking expectations,

$$E(X_t^3) = a^3E(X_t^{(H)3}) + b^3E(X_t^{(M)3}) + c^3E(X_t^{(L)3}) \qquad \ldots (H.2)$$

Since $X_t^{(H)}$, $X_t^{(M)}$ and $X_t^{(H)}$ are independent of each other and have zero mean, the expected values of the cross-product terms are all zero. Also, $X_t^{(H)}$, $X_t^{(M)}$ and $X_t^{(L)}$ are AR(1) processes given by:

$$X_t = \rho X_{t-1} + (1 - \rho^2)^{1/2} \varepsilon_t \qquad \ldots (H.3)$$

Cubing both sides and taking expectations,

$$E(X_t^3) = \rho^3 E(X_{t-1}^3) + (1 - \rho^2)^{3/2} E(\varepsilon_t^3) \qquad \ldots (H.4)$$

That is, $$\gamma_x = \frac{(1 - \rho^2)^{3/2}}{1 - \rho^3} \gamma_\varepsilon \qquad \ldots (H.5)$$

where γ_x and γ_ε are the coefficient of skewness of the random variables, X, and random deviates, ε.

Substituting into (H.2), one gets:

$$\gamma_x = \frac{a^3(1 - \rho_H^2)^{3/2}}{1 - \rho_H^3}\,\gamma_{\varepsilon,H} + \frac{b^3(1 - \rho_M^2)^{3/2}}{1 - \rho_M^3}\,\gamma_{\varepsilon,M}$$

$$+ \frac{c^3(1 - \rho_L^2)^{3/2}}{1 - \rho_L^3}\,\gamma_{\varepsilon,L} \qquad \ldots (H.6)$$

i) Modify only the high frequency term. In this case $\gamma_{\varepsilon,M}$ and $\gamma_{\varepsilon,L} = 0$, and the required skewness of the random numbers used in the high frequency component is given by:

$$\gamma_{\varepsilon,H} = \frac{1 - \rho_H^3}{a^3(1 - \rho_H^2)^{3/2}} \cdot \gamma_x \qquad \ldots (H.7)$$

The Wilson-Hilferty transform can then be used to obtain the required skewed random variate. This transform is given by:

$$\eta_t = \frac{2}{\gamma_{\varepsilon,H}}\left[1 + \frac{\gamma_{\varepsilon,H} \cdot \varepsilon_t}{6} + \frac{\gamma_{\varepsilon,H}^2}{36}\right]^3 - \frac{2}{\gamma_{\varepsilon,H}} \qquad \ldots (H.8)$$

where, η_t is approximately gamma distributed with mean of zero, unit variance and skew $\gamma_{\varepsilon,H}$;

$\gamma_{\varepsilon,H}$ is the skewness of the random deviates required; and

ε_t is a normally distributed random deviate with zero mean and unit variance.

ii) Modify only the medium frequency term. In this case, $\gamma_{\varepsilon,H}$ and $\gamma_{\varepsilon,L} = 0$, and the required skewed deviates are obtained from (H.7) - (H.8) with replaced by $\gamma_{\varepsilon,M}$.

iii) Modify only the low frequency term. In this case, $\gamma_{\varepsilon,H}$ and $\gamma_{\varepsilon,M} = 0$, and the skewed random deviates are obtained from (H.7) - (H.8) with $\gamma_{\varepsilon,H}$ replaced by $\gamma_{\varepsilon,L}$.

Skewed random deviates can also be obtained by assuming the same skewness for each component. That is, $\gamma_{\varepsilon} = \gamma_{\varepsilon,H} = \gamma_{\varepsilon,M} = \gamma_{\varepsilon,L}$. From (H.6),

$$\gamma_{\varepsilon} = \gamma_{x} \left[\frac{a^3(1 - \rho_H^2)^{3/2}}{1 - \rho_H^3} + \frac{b^3(1 - \rho_M^2)^{3/2}}{1 - \rho_M^3} + \frac{c^3(1 - \rho_L^2)^{3/2}}{1 - \rho_L^3} \right]^{-1} \quad \ldots \text{(H.9)}$$

and the required skewed random variate is obtained from (H.8) by replacing $\gamma_{\varepsilon,H}$ with γ_{ε}.

VITA

Leonard Melvin Fung Piau Lye was born on November 2, 1956 in Kota Kinabalu, Malaysia. His undergraduate education was undertaken at the Bolton Institute of Technology, England. He was awarded a Sabah Electricity Board scholarship for this purpose. He graduated in 1981 with a Bachelor of Science degree in Civil Engineering with First Class Honours. For his outstanding undergraduate work, he was awarded the Institution of Civil Engineers Prize. As part of his undergraduate studies, he worked as an assistant engineer for Malaysia International Consultants for a year. Upon graduation, he worked for Sabah Electricity Board in the Hydro Section as an engineer. While there he was seconded to Nippon Koei Consultants of Tokyo, Japan, to supervise the construction of electrical sub-station works which was part of the first major hydroelectric project in Sabah. In September 1983, he began graduate study in the Department of Civil Engineering at the University of Manitoba. He was a teaching and research assistant, and he was awarded a University of Manitoba Graduate Fellowship for 1985-86 and 1986-87.

He is an associate member of the Canadian Society of Civil Engineers and a Registered Professional Engineer in the Province of Manitoba.

His publications are:

Booy, C. and L.M. Lye (1985). The Use of Accumulated Basin Storage in Flood Peak Analysis. Proceedings 7th Canadian Hydrotechnical Conference, Saskatoon. 317-335.

Booy, C. and L.M. Lye. (1986). Accumulated Basin Storage as a Factor in the Correlation Structure of Annual Peak Flows on the Red River. Canadian Journal of Civil Engineering. 13(3), 365-374.

Booy, C. and L.M. Lye. (1986). Discussion of "Correlation of Annual Peak Flows for Pennsylvania Streams", by D.J. Wall and Mary Englot. Water Resources Bulletin (to appear).

Booy, C. and L.M. Lye. (1987). Parameter Uncertainty: A Fundamental Issue in a Probabilistic Approach to Flood Control. Proceedings of the Workshop on Statistical Modelling and Flood Regionalization, McGill University, 1986.

Booy, C. and L.M. Lye. (1987). Uncertainty in Flood Risk Analysis. Proceedings Eight Canadian Hydrotechnical Conference, Montreal. 401-418.

L.M. Lye, S.K. Sinha and C. Booy. (1987). Bayesian Estimation of the T-year Events for Flood Data Fitted by a Three-parameter Lognormal Distribution. Paper submitted to Civil Engineering Systems.

L.M. Lye, S.K. Sinha and C. Booy. (1987). A Bayesian Analysis of Flood Frequency Data Fitted by the Extreme Value Type 1 Distribution. Paper submitted to Journal of the Royal Statistical Society, London.

Tamburi, A. and L.M. Lye. (1985). Discharge Measurements at Non-Ideal Sites. Proceedings 7th Canadian Hydrotechnical Conference, Saskatoon. 169-189.