CANADIAN THESES ON MICROFICHE

THÈSES CANADIENNES SUR MICROFICHE

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

Canada

NL 339 (r. 85/09)

A Nonparametric Approach to Flood Frequency Analysis

by

Charles W. Labatiuk

A thesis
presented to the University of Ottawa
in fulfillment of the
thesis requirement for the degree of
Master of Applied Sciences
in
Civil Engineering

OTTAWA, Ontario, 1985

The University of Ottawa requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

ABSTRACT

Parametric statistical techniques are generally used by hydrologists in flood frequency analysis. The assumption is made that the underlying distribution of the flood data is known, for example, it is a log-Pearson type III distribution. Such an assumption is not always justified, and can sometimes lead to considerable variability in the estimation of design floods. Nonparametric density estimation provides an alternative method of analysis which does not require any a priori distributional assumption.

Since being introduced by Rosenblatt (1956) and Parzen (1962), the kernel estimator has become one of the better known nonparametric methods. In a kernel estimate both the kernel function K(.) and the smoothing factor h represent empirical choices. Various methods to optimally compute h directly from the data have been proposed; including those of Scott, Tapia and Thompson (STT); Adamowski and Feluch (AF), and Wagner (W). Method STT is based on asymptotic expansion of the integrated mean square error criterion, method AF is based on a similar criterion at the distribution function, and method W is heuristic. These methods of obtaining h in the kernel estimator have been investigated for use in flood frequency analysis.

In this thesis design floods for the Northeast Margaree River at Margaree Valley in Canada and the Warta River at Poznan in Poland have been computed using nonparametric methods. The results are compared with the results of traditional parametric analysis. A Monte-Carlo simulation study was performed to permit a comparison of flood computation methods and various data-based algorithms for h. In the simulation a LP III parent was used and skews of 0.444 (Data 1) and 1.467 (Data 2) have been considered.

The principle finding is that design floods up to the 100 year return period can reasonably be estimated using the nonparametric methods W, STT and AF log. It is shown that nonparametric methods are especially suitable for multimodal data. With highly skewed data, a transformation would improve the estimates of floods.

ACKNOWLEDGEMENTS

The author expresses his sincere gratitude and respect to his thesis supervisor, Dr. K. Adamowski, for suggesting the topic and providing guidance during the study.

The author is grateful to Professor D. W. Scott of Rice University, Houston, Texas for providing the algorithm STT and to Professor R. E. Taylor of the University of Georgia for making the program of Mr. M. Liu available.

Appreciation is expressed to Dr. L. Birta, Dr. L. Devroye, Dr. W. Feluch and Mr. W. O. Thomas, Jr. for useful discussions. Thanks is also due Dr. B. Silverman, Dr. E. Shuster and Dr. S. Yakowitz for sending research papers.

Data for the Monte Carlo simulation and the results for parametric distributions was kindly provided by Mr. P. Pilon, Water Resources Branch, Environment Canada, Ottawa.

Financial assistance for this research was supplied by a grant from the National Research Council, Grant No. 7467.

Special thanks are due to my parents for their encouragement and moral support during the period of this research.

TABLE OF CONTENTS

ABSTRACT iv

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF FIGURES x

LIST OF TABLES xii

NOTATION xiv

Chapter Page

I. INTRODUCTION 1

1.1 Motivation 1

1.2 Mathematical Models in Hydrology 3

1.3 Flood Frequency Analysis 7

1.4 Objectives of Study 8

1.5 Thesis Overview 11

II. LITERATURE REVIEW OF PARAMETRIC METHODS 13

2.1 A Historical Perspective 13

2.2 Plotting Position 16

2.3 Theoretical Frequency Distributions 20

2.3.1 General 20

2.3.2 Lognormal Distributions 21

2.3.3 Extreme Value Distributions 24

2.3.4 Log-Pearson Type III Distribution 26

2.3.5 Wakeby Distribution 30

2.4 Fitting Techniques 31

2.5 Goodness-of-Fit Tests 35

2.6 Recent Research Trends 37

2.7 Limitations of Theoretical Frequency Distributions 40

III. LITERATURE REVIEW AND DEVELOPMENT OF NONPARAMETRIC TECHNIQUES 42

3.1 General 42

3.2 Graphical Nonparametric Density Estimation Methods 51

3.2.1 Histogram 51

3.2.2 Frequency Polygon 54

3.3 Kernel Estimator 56

3.4 Data-Based Algorithms 66

3.4.1 Method of Adamowski and Feluch 66

3.4.2 Method of Scott, Tapia and Thompson . . . 70

3.4.3 Method of Duin 72

3.4.4 Method of Flanagan 73

3.4.5 Mtheod of Wagner 74

3.5 Nonparametric Discretized Maximum Penalized Likelihood Estimator 76

IV. NUMERICAL ANALYSIS 78

4.1 Data Selection 78

4.2 Real Data 79

4.2.1 Graphical Analysis 79

4.2.2 Parametric Analysis 89

4.2.3 Nonparametric Analysis 91

4.3 Monte-Carlo Simulation Design 97

V. RESULTS AND DISCUSSION 101

5.1 Real Data 101

5.2 Synthetic Data 125

VI. CONCLUSIONS 139

6.1 Major Conclusions 139

6.2 Recommendations for Further Research 143

REFERENCES 145

Appendix Page

A. Parametric Flood Frequency Analysis Results 152

B. Computer Programs 173

LIST OF FIGURES

Figure		Page
1.1	Models in Hydrology	4
2.1	Standard deviation of empirical exceedance probability versus sample size for various plotting formulae	19
3.1	Ratio of histogram bin width to theoretical Gaussian bin width with same variance versus skewness	55
3.2	Relative Weighting of Three Kernels	60
3.3	Approximation of a density function by the sum of quadratic kernels	61
3.4	Effect of varying the smoothing factor on kernel density estimates	62
3.5	Density Function Plots for Test Data	69
4.1	Discharge-Rank-Time Plots, Northeast Margaree River at Margaree Valley	83
4.2	Discharge Histograms, Northeast Margaree River at Margaree Valley	84
4.3	Optimal Discharge Histogram and Frequency Polygon, Northeast Margaree River at Margaree Valley	85
4.4	Discharge-Rank-Time Plots, Polish River Warta at Poznan	86
4.5	Discharge Histograms, Polish River Warta at Poznan	87
4.6	Optimal Discharge Histogram and Frequency Polygon, Polish River Warta at Poznan	88
4.7	Flowchart for RECKER Method of Adamowski and Feluch	93
4.8	Flowchart for MVBKER Method of Wagner	94

4.9	Errors Computed in Simulation	100
5.1	Nonparametric Density Plots, Margaree River Untransformed Data	108
5.2	Nonparametric Density Plots, Margaree River Log Transformed Data	109
5.3 to 5.6	Margaree River Nonparametric Distributions	110-113
5.7	Nonparametric Density Plots, Warta River Untransformed Data	114
5.8	Nonparametric Density Plots, Warta River Log Transformed Data	115
5.9 to 5.12	Warta River Nonparametric Distributions	116-119
5.13	Comparison of Warta River Parametric and Nonparametric Distributions	124
A.1 to A.4	Northeast Margaree River at Margaree Valley Parametric Distributions	158-161
A.5 to A.8	Polish River Warta at Poznan Parametric Distributions	169-172

LIST OF TABLES

Table		Page
2.1	Plotting Position Formulae	18
3.1	Some advantages and disadvantages of various types of estimators	50
3.2	Typical kernels	58
4.1	Data Used for Numerical Analysis	78
4.2	Tests for Independence, Trend, Randomness and Homogeneity; Northeast Margaree River at Margaree Valley	80
4.3	Tests for Independence, Trend, Randomness and Homogeneity; Polish River Warta at Poznan	81
4.4	LP III Parameters Used in Generation	98
4.5	Statistics of the Generated Data	98
5.1	Flood Estimates for Northeast Margaree River at Margaree Valley	102
5.2	Flood Estimates for Polish River Warta at Poznan	102
5.3	Variability of Nonparametric versus Parametric Flood Estimates for Northeast Margaree River at Margaree Valley	104
5.4	Variability of Nonparametric versus Parametric Flood Estimates for Polish River Warta at Poznan	105
5.5	Nonparametric Density Estimate Statistics for Northeast Margaree River at Margaree Valley	123
5.6	Nonparametric Density Estimate Statistics for Polish River Warta at Poznan	123
5.7 to 5.10	Simulation Results	126-129

5.11 to 5.14	Smoothing Parameters in Simulation	133-134
5.15	Effect of Record Length on Smoothing Parameter	135
5.16	Computer Times in Simulation	136
5.17	Comparison of Parametric and Nonparametric Simulation Results	137
A.1	Data Listing; Northeast Margaree River at Margaree Valley	152
A.2 to A.5	Parametric Frequency Analysis; Northeast Margaree River at Margaree Valley	154-157
A.6	Data Listing; Polish River Warta at Poznan	162
A.7 to A.10	Parametric Frequency Analysis; Polish River Warta at Poznan	165-168

NOTATION

a = scale parameter

b = shape parameter

c = location parameter

C_v = coefficient of variation

$E(\ )$ = expected value of ( )

$f(x)$ = probability density function

$F(x)$ = probability distribution function

F_j = plotting position formula

g = coefficient of skewness

G = skew coefficient of logarithms

h = smoothing factor

h^* = optimal histogram bin width

h^{**} = optimal frequency polygon bin width

k = parameter, number of nearest neighbors

K = frequency factor

$K(.)$ = kernel

L = likelihood

n = sample size, number of degrees of freedom

N = number of observations

p = exceedance probability, number of parameters

q = quantile

Q = discharge (cubic meters per second)

Q_T = T-year discharge(cubic meters per second)

r = radius, weighting factor

R = risk of failure

s = sample standard deviation

S = standard deviation of logarithms

t = standard normal deviate

T = return period (years)

X = random variable

$\overline{X}$ = mean of logarithms

$(\hat{})$ = estimate of ( )

α = constant, level of significance

$\delta(.)$ = Dirac delta function

γ = coefficient of skewness

θ = parameter

μ = population mean

μ_r' = rth absolute moment

μ_r = rth central moment

σ = population standard deviation

σ^2 = variance

χ^2 = statistical chi-square variable

Abbreviations

AF = method of Adamowski and Feluch
AF log = method of Adamowski and Feluch on log transformed data
AFS = annual flood series
CDF = cumulative distribution function
CMS = cubic meters per second
CFA1 = Consolidated Frequency Analysis Program
C.K. = coefficient of kurtosis
C.S. = coefficient of skewness
C.V. = coefficient of variation
DHH = method of Duin
EV1,EV2,EV3 = extreme value types I, II & III distributions

FF	= formula method of Flanagan
GEV	= general extreme value distribution
IMSE	= integrated mean square error
IMSL	= International Mathematical and Statistical Libraries
LN3	= three-parameter lognormal distribution
LP III	= log-Pearson type III distribution
MLE	= maximum likelihood estimate
MSE	= mean square error
NDMPLE	= Nonparametric Discretized Maximum Penalized Likelihood Estimator
PDS	= partial duration series
RMSE	= root mean square error
STT	= method of Scott, Tapia and Thompson
STDV	= standard deviation
TCEV	= two-component extreme value distribution
W	= method of Wagner
Wak.	= Wakeby distribution
WRC	= U.S. Water Resources Council
WSC	= Water Survey of Canada

Chapter I

INTRODUCTION

1.1 MOTIVATION

Water resource management involves not only the beneficial uses of water resources, but also the prevention or minimization of the harmful effects of water excess or deficiency. The prediction of these extremes of hydrologic processes is one of the most basic tasks faced by the hydrologist.

Man has built many of his cities close to large rivers for reasons of water supply, wastewater disposal and navigation. Population pressures have caused further development, often on flood plains. A major flood on the Fraser River, British Columbia caused $20 million of damage to some 2,000 homes in 1948. Engineering works such as dikes, storage reservoirs and flood control dams can mitigate flood damage, but not avert it altogether. The Winnipeg Floodway is a good example of such measures being taken to reduce flood damage.

Knowledge of the magnitude and frequency of floods are crucial to the proper design of hydraulic structures and drainage facilities. Kite (1977) has cited a number of studies which indicate that inadequate spillway design is one

of the major causes of dam failure. Inadequate spillway design can usually be traced back to inadequate flood frequency analysis. The objective of flood frequency analysis has been stated as "the inference of the form and parameters of the statistical distribution which characterizes the population from the sample of recorded data", (Hall, 1984, p.86).

Drought analysis, although not pursued in this study, is similar to flood frequency analysis, but at the other end of the water spectrum.

Clearly, the prediction of extreme hydrologic events is of great practical importance. Parametric statistical techniques are normally used in single station flood frequency analysis. A basic problem with parametric analysis is that a theoretical frequency distribution for the population of events must be assumed, and there is no general agreement among hydrologists on the best one.

During approximately the last 25 years new developments in statistical theory have provided an alternative approach of nonparametric analysis. This approach involves estimating the density nonparametrically, and has the big advantage of bypassing a prior distribution choice altogether.

In this study, nonparametric techniques are investigated within the context of statistical hydrology, and for the specific problem of flood frequency analysis.

1.2 MATHEMATICAL MODELS IN HYDROLOGY

The term 'model' is so widely used in hydrology that its meaning has become vague. A model is an abstraction of reality which is used as a tool to derive answers about the behavior of hydrologic systems or processes. As nature is constantly original and inventive, a model can only approximate certain of its aspects. The fundamental questions which then arise are, which aspects of reality must be modeled and how good must these approximations be?

Some models which have been found useful in hydrology are shown in Fig. 1.1. This figure should not be considered rigid, since actual models tend to overlap categories, however it is illustrative of the relative characteristics of various models.

Optimization models concern themselves with finding the best set of conditions satisfying a set of objectives, given a set of constraints. Nonoptimization models can be classified as physical/deterministic or statistical. The former are perhaps more deeply rooted in hydrology than the latter, and can be further divided into causal and empirical models. Causal or conceptual models attempt to explain and

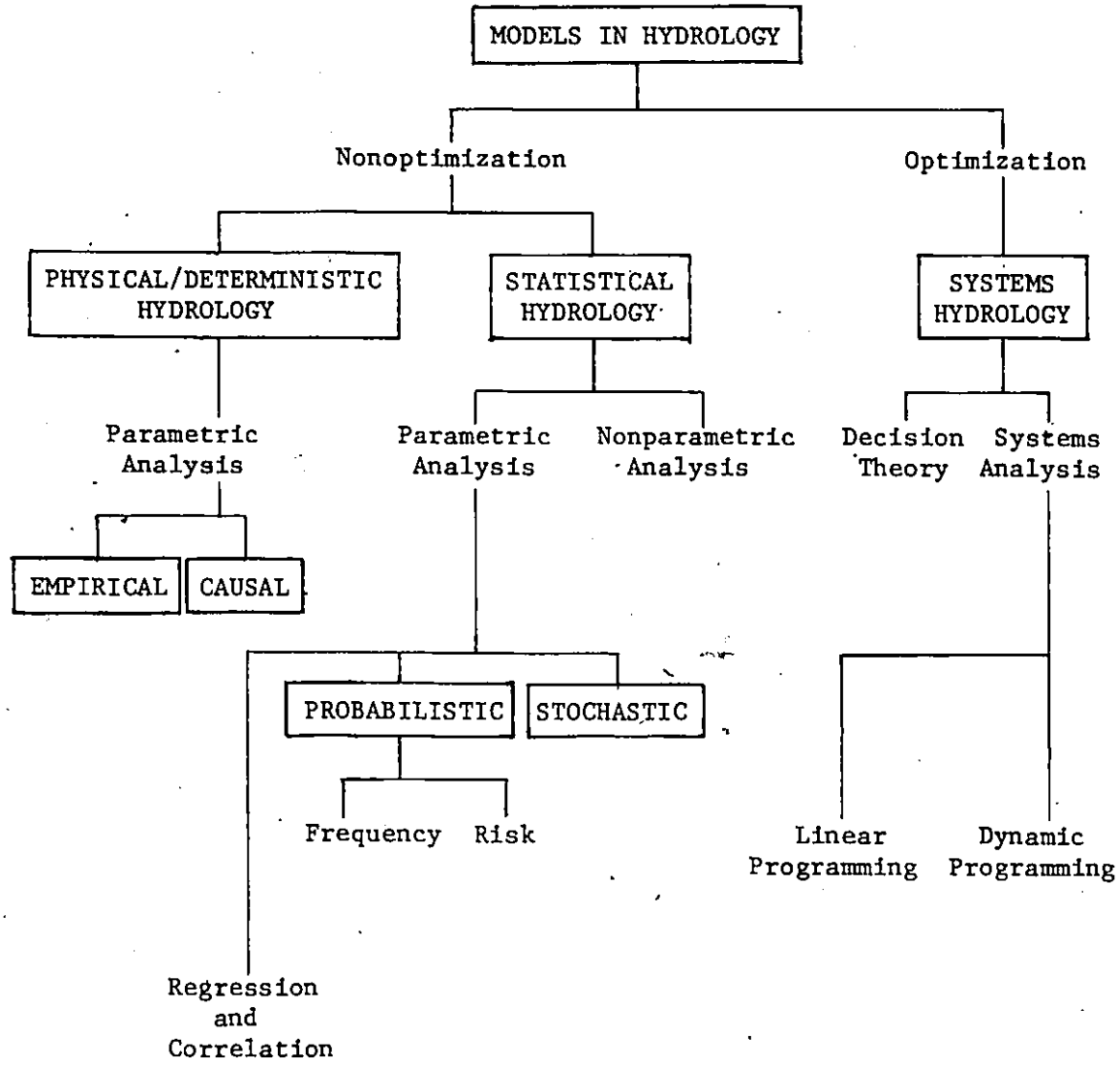


FIG. 1.1 Models in Hydrology
(After Fleming, 1975).

understand hydrologic processes based on the theoretical application of the principles of basic sciences. Empirical models offer a practical, immediate answer; however they are quite specific to the data for which they are developed. Statistical inference has been found necessary in the planning of water-resource systems, and provides an intermediate view in which statistical principles help bridge the gap when components in causal relationships are unknown or perhaps unknowable.

Statistical hydrologic models can be subdivided into parametric or nonparametric. In parametric analysis it is assumed that the unknown probability density takes a known functional form

$$f(x) = f(x;\theta) \tag{1.1}$$

where θ is a vector of p parameters which completely specify the density function. Parametric procedures are relatively efficient if the correct functional form for the sample data has been chosen. In nonparametric analysis the probability density function is directly estimated with no assumption about the functional form. This density estimation approach is more complex since the density cannot be given by an equation, but is computed numerically from the data. Computer use is essential with nonparametric procedures.

Nonparametric analysis should not be thought of as assumption free. However very little is assumed, and the

assumptions made are mild, such as working in a particular space. A series of increasingly stronger assumptions must be made in moving from nonparametric to parametric analysis.

Hydrologic processes are known to be continuous, time-varying and nonlinear. Hydrologists often assume time-invariance, or stationarity, and linearity in their models since it leads to a simpler explanation of the process, one with workable mathematics. The assumption of stationarity has been challenged by Sagan et al. (1979). They contend that "considering only global albedo changes, it is possible that human alteration of the environment has made a significant contribution to the observed global climatic shifts of both the last several decades and the last several millennia." Thus the longer the record, the more likely it is nonstationary. Homogeneity may also be an inappropriate assumption in many applications. High floods and low floods tend to be treated as coming from the same causes. In Canada, floods may be generated by either snowmelt or rainfall, and different types of storms can give rise to different flood characteristics. With nonparametric analysis the density is not assumed to be unimodal, or of any functional form, so nonhomogeneous data can readily be considered.

In statistical hydrology, a nonparametric class of models offers promise as they appear to be powerful while making few assumptions regarding underlying distributions.

1.3 FLOOD FREQUENCY ANALYSIS

To the hydrologist, flood frequency analysis has come to mean the science and art of fitting distributions to annual flood series (AFS) and the associated problems of estimation of parameters, goodness-of-fit testing, treatment of outliers, plotting position, etc. However, to the statistician the expression is a bit of a misnomer and might be better known as flood magnitude frequency analysis, as the usual type of problem is the estimation of flood magnitude corresponding to a specified return period, T.

If the qth quantile of a random variable X is a value x_q, then:

$$F(x_q) = q = \int_0^{x_q} f(x)dx = 1 - \frac{1}{T} \tag{1.2}$$

where F is the cumulative distribution function (CDF), f the density function, and T the return period. Thus it can be seen that the density function is the derivative of the cumulative distribution function, or simply the distribution function, i.e.:

$$f(x) = \frac{dF(x)}{dx} \geq 0 \tag{1.3}$$

Furthermore

$$\int_{-\infty}^{+\infty} f(x) = 1 \qquad (1.4)$$

which indicates that the area under the density function is 1.

The 100-year flood $x_{.99}$ is often the selected design value for flood control structures. It is sometimes convenient to work directly in terms of exceedance probability, p.

$$p = 1 - q = \int_{x_q}^{\infty} f(x)dx = \frac{1}{T} \qquad (1.5)$$

From eq. 1.5 it can be seen that when an AFS is analysed, the probability of the T-year flood being exceeded in any given water year is 1/T. A more pertinent value is the risk of failure, R, which is the probability of the T-year flood being exceeded during the design life of the structure. If the design life is taken as N years, then

$$R = 1 - (1 - (\tfrac{1}{T}))^N \qquad (1.6)$$

For example, the risk of failure in a 100-year design life using the 100-year flood is 0.63.

1.4 OBJECTIVES OF STUDY

Various authors have noted that the potential of nonparametric probability density estimation is not being fully realized (Tarter and Kronmal (1976), Scott and Factor

(1981)). One of the reasons for this is that many of the results have been of the large sample type, and little guidance exists as to their practical implementation. However, the use of nonparametric estimates for very large samples is a questionable procedure. Intuitively, large samples are more suited to a parametric approach, since with the amount of information available the distribution should be known. This line of reasoning supports the use of nonparametric techniques for small and intermediate sample sizes, such as the sample sizes available to hydrologists for flood frequency analysis.

Rosenblatt (1956) was one of the first researchers of nonparametric density estimation, and while the field has since burgeoned, the techniques are generally known and used only by statisticians. For over the past decade Yakowitz has studied nonparametric estimation methods and their suitability for time series modeling. Yakowitz has recommended wider use of nonparametric techniques and has stated that "flood prediction is a wide open field" (Yakowitz, 1981, p.91). This study is an attempt to address this issue.

The specific objectives of this thesis are:

1. To investigate the possibilities for the application of nonparametric methods to flood frequency analysis. Existing

methods are reviewed and evaluated in light of the new methods.

2. To present various nonparametric techniques including graphical estimators, kernel estimators and NDMPLE (Nonparametric Discretized Maximum Penalized Likelihood Estimator). The presentation is not intended to be a rigorous theoretical explanation. Instead the methods will be described from an engineering viewpoint and given with the goal of proper application in flood frequency analysis.

3. To explore the kernel estimator and some of its modifications in considerable detail. The investigation will include problems such as (i) selection of a kernel and (ii) the determination of the smoothing factor, especially by data-based algorithms. As the kernel estimator is the most popular and theoretically developed nonparametric estimator, it warrants special attention.

4. To apply the nonparametric techniques to two sets of annual flood series data, and compare the results with those of conventional parametric flood frequency analysis.

5. To perform a Monte-Carlo simulation study so that the relative merits and suitability of the various approaches can be compared numerically. The simulation study will also be designed to allow some preliminary conclusions on the

effects of sample length and skewness on the nonparametric flood estimates.

It should be noted that this thesis deals within the field of statistical hydrology, and as such its aim is more modest than that of deterministic hydrology in that the objective is to predict only the frequency or magnitude of future occurrences without attempting to specify the times of these occurrences. Information on soils, watersheds, etc. is not applicable, as the model is statistical, not physically-based. Regional frequency analysis can be very useful, especially with limited data, however this study is limited to single station frequency analysis. Basically, it is the intention of this thesis to use nonparametric statistical analysis as a new tool in flood frequency analysis.

1.5 THESIS OVERVIEW

Chapter I gives background material, the problem statement and a review of this thesis.

In Chapter II, beginning with a brief historical review, a survey is made of methods currently used in parametric statistical flood frequency analysis. Attention is focused on the limitations of these methods, which in turn provide the rationale for investigating nonparametric methods. Within this chapter general concepts and relevant terminology are introduced.

Chapter III develops various nonparametric methods which appear suitable for flood frequency analysis. Graphical nonparametric density estimation methods, such as the conventional histogram and lesser known frequency polygon are introduced. Next the kernel estimator is presented. Various data-based algorithms which automatically choose the smoothing factor required for the kernel estimator are emphasized. Finally NDMPLE, a computer routine available from IMSL (1980) for nonparametric density estimation, is discussed.

In Chapter IV, two sets of data are selected for analysis. The data is presented graphically and flow charts are given for computer programs which effect the nonparametric analysis. Chapter IV includes a description of the Monte-Carlo simulation study.

Chapter V compares the parametric and nonparametric analysis for the two data sets, and discusses the simulation study results.

Chapter VI summarizes the main conclusions of this thesis, and gives suggestions for further research.

Chapter II

LITERATURE REVIEW OF PARAMETRIC METHODS

2.1 A HISTORICAL PERSPECTIVE

According to the late V.T. Chow (1964), the various periods of development of hydrology since the turn of the century may be classified as:

(1)	Period of empiricism	1900-1930
(2)	Period of rationalization	1930-1950
(3)	Period of theorization	1950-to date

During the period of empiricism much emphasis was placed on probability plots and envelope curves. Probability paper was introduced by Hazen in 1919. The period of theorization has given rise to increasingly complex hydrologic models and the use of more sophisticated statistical techniques, with the result that the importance of probability plots has decreased. However, probability plots and plotting formula, or empirical distribution formula, remain useful for the graphical presentation of data and the detection of outliers. Few engineering decisions based on the analytical fitting of distributions to data are made without the use of a graphical display.

While the most recent period has been labelled as the period of theorization, it has also been marked by the advent of the computer with ever increasing raw data handling and computational capacity. Hall (1984) has observed that "the last 30 years have seen the subject of hydrology mature from being largely an engineering design aid into a recognized branch of the science of geophysics."

Data series which have been used in the modelling of floods include the annual flood series (AFS) and the partial duration series (PDS). The AFS consists of the sequence of maximum daily or instantaneous discharges recorded in successive water years. It is preferable to use instantaneous values in design, however due to the longer record lengths available maximum daily discharge values are often used in flood frequency analysis, and then transformed to instantaneous design values by the Fuller equation. The PDS consists of all independent flood peaks within the record above a specified threshold discharge. Thus the PDS is based on the truncation of the hydrograph to establish the stochastic nature of flood peaks, while the AFS is based on the empirical fit of distributions. The late Dr. Chow (1964) has shown that the probability distributions obtained by either method are essentially the same for return periods greater than 10 years. For frequent events, i.e. return

periods less than 10 years, the PDS should be used; otherwise the AFS is the best choice.

One concept which is now generally considered outdated is that of the Probable Maximum Flood. It is a totally deterministic method which relies on the identification of a relationship between rainfall and streamflow. To the layman it is implied that the maximum flood cannot exceed this limit, however as there is no associated probability level, the protection provided is illusionary. On the other hand, flood-frequency analysis accepts events of any magnitude as being possible although the occurrence probability decreases as the magnitude increases.

There are a large number of distributions available for the analysis of AFS. In 1967 at an International Hydrological Decade meeting in Leningrad it was said that over 100 distributions had been tried. Often used distributions include lognormal, Gumbel, and log-Pearson Type III (LP III). The normal distribution has been studied in more detail by statisticians than any other distribution, so its use, with a log-transformation to account for the skewness of flood data, was inevitable. In 1941 Gumbel initially postulated the use of the double exponential Type-I Extreme Value Distribution law for evaluating design floods. In 1977 the United States Water Resources Council issued Bulletin 17A, which recommended the use of the LP III

distribution as a base method for all U.S. Federal agencies. McCuen and Rawls (1979) have commented that the large number of procedures available creates the burdensome task for potential users of sorting and selecting a method for immediate use. They recommend guidelines be developed so that consistent information on data requirements, accuracy, reproducibility, and practicality of various procedures can be included with published results.

2.2 PLOTTING POSITION

The plotting position can be thought of as the return period or cumulative probability assigned to each observation in a sample. Plotting positions are used to get a general idea of how well the observations fit a theoretical frequency distribution on a plot. Plotting position is a function of the sample size and the rank of an event, and the plotting position of a particular observation will change as the data expands. Thus, for small samples the plotting position may be unrepresentative of the true probability.

Many plotting position formulae are of the following general form:

$$F_j = (j-\alpha)/(N+1-2\alpha) \qquad (2.1)$$

where F_j are ordered probability values $F_1 < F_2 < \ldots < F_N$; j is the j^{th} value of N ordered observations, and α is a constant depending on the distribution. For the normal distribution $\alpha = 0.375$ (the Blom formula), for the extreme value Type-I distribution $\alpha = 0.44$ (the Gringorten formula), and for the uniform distribution $\alpha = 0$ (the Weibull formula). Cunnane (1978) has noted that the often recommended Weibull formula can provide biased results, and he hypothesized $\alpha = 0.4$ as a compromise value. Cunnane's value is supported for the LP III distribution by Srikanthan and McMahon (1981), who note that α varies with g, the coefficient of skewness, and recommend that an optimum value of α for all values of g lies close to 0.4 based on a Monte-Carlo analysis. Zhang (1982) has derived a generalized plotting position formula based on order statistics, which is suitable for samples containing both extraordinarily large and small historical values. He notes that the Weibull formula is a special case of his generalized formula.

Adamowski (1981) has recommended using $\alpha = 0.25$ for a distribution free plotting formula based on the mean square error criterion. Other recommended values of α include $\alpha = 0.3$ (Chegodajew formula which is used in U.S.S.R.) and $\alpha = 0.5$ (the Hazen formula).

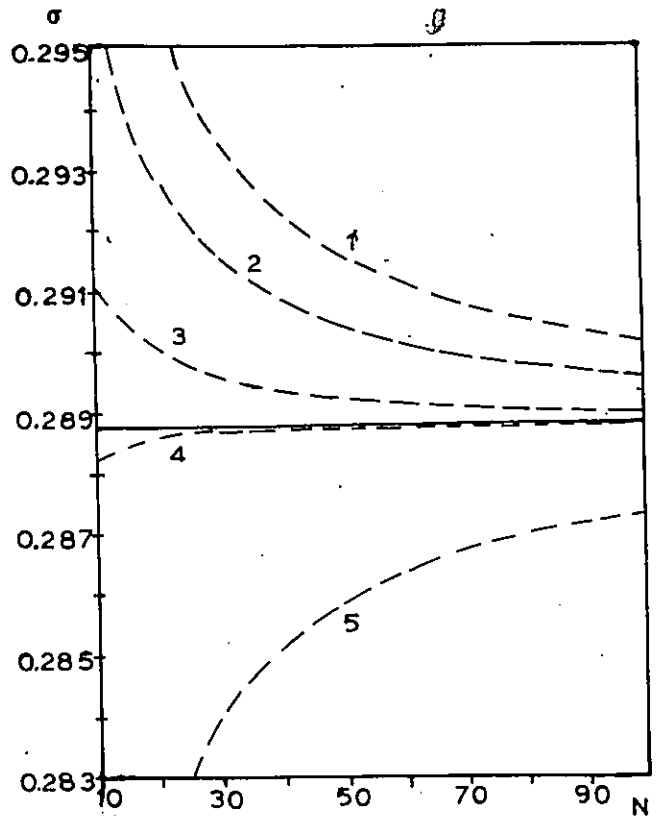
A comparison of the return periods assigned to the largest value, given a hypothetical AFS of 100 values, for various plotting formulae is made in Table 2.1.

TABLE 2.1

Plotting Position Formulae

Name	Formula F_j	α in Eq. 2.1	Values of F_j and T For j=1 and N=100 F_j	$T=1/F_j$
Blom	(j-0.375)/(N+0.25)	0.375	0.0062	160.4
Gringorten	(j-0.44)/(N+0.12)	0.44	0.0058	178.8
Weibull	j/(N+1)	0.00	0.0099	101.0
Cunnane	(j-0.4) /(N+0.20)	0.40	0.0060	167.0
Adamowski	(j-0.25)/(N+0.50)	0.25	0.0075	134.0
Chegodajew	(j-0.3)/(N+0.4)	0.30	0.0070	143.4
Hazen	(j-0.5)/ N	0.50	0.0050	200.0

With so many plotting formulae available, some direction has been provided by Cunnane (1978) in his review article. He suggests that plotting formulae should be used which "confer the best statistical qualities on:
(1) judging whether a given sample might reasonably have been drawn randomly from the type of distribution in question; and (2) estimating quantiles." Estimated quantiles should be unbiased and have minimum variance among



Number	Name	Formula, F_j
1	Hazen	(j-0.50)/N
2	Cunnane	(j-0.40)/(N+0.2)
3	Chegodajew	(j-0.30)/(N+0.4)
4	Adamowski	(j-0.25)/(N+0.5)
5	Weibull	j/(N+1.0)

FIG. 2.1 Standard deviation (σ) of empirical exceedance probability vs. sample size (N) for various plotting formulae (After Marnatnikov and Kudajbergenov, 1980).

graphical estimates. Various plotting formula are compared in Fig. 2.1.

2.3 THEORETICAL FREQUENCY DISTRIBUTIONS

2.3.1 General

Almost all current techniques of flood frequency analysis are based on a parametric approach. The AFS is assumed to follow a known distribution, and the parameters of the distribution are estimated from the sample data. In effect, the data is summarized in a few distribution parameters. Two basic sources of error exist. One is that the form of the distribution is not known and must be guessed. The effect of this may be considerable, as most of the sample data which provide some information occupy the central portion of the distribution, while the event magnitudes of interest are in the tails. Two different distributions may fit the same data equally well, but probabilities of extreme values could be vastly different. The second source of error is due to poor fitting of the distribution. The sample data which provide estimates of the parameters of the distribution are often relatively short in length and are in themselves subject to error.

Once a distribution has been selected and its parameters estimated, one method of performing the frequency

analysis is through the use of frequency factors. Chow (1964) proposed the general equation

$$x_T = \mu + K\sigma \qquad (2.2)$$

where x_T is the event magnitude at a given return period, T, μ and σ are the population mean and standard deviation estimated from sample moments, and K is a frequency factor which is a function of the return period and the distribution parameters. Frequency factors for distributions commonly used in flood frequency analysis are given in Kite (1977).

For parametric analysis the standard error, which is a measure of the difference between computed estimates and recorded events, can be derived for various methods of parameter estimation. However, the standard error of estimate accounts only for errors in parameter estimation, and does not include errors due to incorrect theoretical frequency distribution choice.

The lognormal, extreme value, LP III and Wakeby distributions are presented in the following subsections.

2.3.2 Lognormal Distributions

If the causative factors of a random variable X are multiplicative, that is if $X = X_1 X_2 \ldots . X_N$, then from the Central Limit Theorem, $Y = \ln X$ can be expected to be

normally distributed with mean μ_y and variance σ_y^2. The probability density function for the two-parameter lognormal distribution is:

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma_y^2}(\ln x - \mu_y)^2\} \qquad (2.3)$$

In effect the transformation $Y = \ln X$ is being used to normalize a skewed distribution, which then makes analysis easier as the theoretical properties of the normal distribution are well known.

The two-parameter lognormal distribution cannot have negative skewness, and its variates take on values in the range $[0,+\infty]$. The parameter μ_y determines the scale of the distribution while σ_y^2 determines the distribution's shape. The two-parameter lognormal distribution will reproduce a coefficient of skewness equal to:

$$g = C_v^3 + 3C_v \qquad (2.4)$$

where C_v is the coefficient of variation of the recorded events. For this distribution the theoretical coefficients of skew and kurtosis of the log-transformed data are 0.0 and 3.0 respectively, which allows for a quick and easy check on the validity of the fit.

Due to its simplicity, and because many hydrologic variables such as discharge and precipitation are measured

from zero and have no upper bound, the two-parameter lognormal distribution has been widely used in hydrology. However, Condie et al. (1981) have found with Canadian data that the log-transform can overdo the normalizing process, turning positively skewed basic data into data with a substantial negative skew coefficient.

A generalization of the two-parameter lognormal distribution is the shifted lognormal or three-parameter lognormal distribution obtained by letting $Z = \ln (X-c)$ be normally distributed. The probability density function for the three-parameter lognormal distribution is

$$f(x) = \frac{1}{(x-c)\sigma_z\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_z^2}\left[\ln(x-c) - \mu_z\right]^2\right] \qquad (2.5)$$

where μ_z and σ_z^2 are the mean and variance of Z. When the location parameter c is set to zero, eg. 2.5 reduces to the two-parameter lognormal distribution. The third parameter, c, provides the extra degree of freedom required to fit the distribution to the first three moments of the AFS. Thus the three-parameter lognormal distribution will reproduce the coefficient of skewness of the observed data. If the data is positively skewed, c will be a lower bound; while if the data is negatively skewed, c will be an upper bound. While there are fitting techniques available for the upper bounded case, if the distribution has a finite upper bound

close to the maximum observed flood in a data series, then its physical basis becomes questionable.

The three-parameter lognormal distribution is very flexible and in general is more successful as a transformation towards normality than the two-parameter lognormal. The theoretical values for the skew and kurtosis of the transformed data for the three-parameter lognormal distribution are 0.0 and 3.0 respectively.

2.3.3 Extreme Value Distributions

Fisher and Tippett found that the distribution of maximum or minimum events selected from N samples approached a limiting form as the sample size n increased. This limiting form is given by the functional equation (Kite, 1977)

$$f^n(x) = F(a_n x + b_n) \qquad (2.6)$$

where a_n and b_n are functions of n. There are three possible solutions to this functional equation, the unbounded EV1, the lower bounded EV2 and the upper bounded EV3 distributions.

The EV1 or Gumbel distribution results when the initial distributions within the samples are exponential. Since E. J. Gumbel's extensive studies, the Gumbel distribution, also known as the double exponential distribution, has been much favored in flood frequency analysis. The theoretical

values for the coefficients of skewness and kurtosis for the Gumbel distribution are 1.14 and 5.4 respectively.

The Gumbel distribution has been generalized to the general extreme value (GEV) distribution, the CDF of which may be written as (Hall, 1984)

$$F(x) = \exp\{-[1-k(x-c)/a]^{1/k}\} \qquad (2.7)$$

where a is the scale parameter and c the location parameter. As k tends to zero, the EV1 distribution is obtained, for which the CDF is

$$F(x_1) = \exp\{-\exp[-(x_1-c)/a]\} \qquad (2.8)$$

and the probability density function is

$$f(x_1) = a^{-1}\exp\{-[(x_1-c)/a]-\exp[(x_1-c)/a]\} \qquad (2.9)$$

When the value of k in eq. 2.7 is negative, the EV2 distribution is obtained. The EV2 distribution has a coefficient of skewness greater than 1.14, and is lower bounded at c + a/k. When the k-value in eq. 2.7 is positive, the EV3 distribution is obtained, which is upper bounded at c + a/k and has a coefficient of skewness less than 1.14. There are polynomial approximations available for k which only require knowledge of the coefficient of skewness.

The Weibull distribution is closely related to the EV3 distribution, being the negative of the EV3 variate. The Weibull distribution is used in hydrology as the distribution of low stream flows.

It should be noted that the GEV distribution is in real space, not log space. In the United Kingdom a preference has been expressed for use of the GEV distribution in flood frequency analysis (Hall, 1984).

2.3.4 The Log-Pearson Type III Distribution (LP III)

In an effort to promote uniform and consistent flood-flow frequency determination, the U.S. Water Resources Council (WRC) in June 1977 published Bulletin 17A, an extension of several earlier publications. In Bulletin 17A the Pearson Type III Distribution with log transformation of the flood data (LP III) was recommended as the basic distribution for defining the AFS. It was further recommended that the method of moments be used for determining the distribution parameters. Subsequently Bulletin 17B, which provided new relations for modifying the station skew coefficient based on a generalized skew study and revised procedures for treating outliers and computing confidence limits, has been issued by the U.S. Geological Survey's Interagency Advisory Committee on Water Data. The latest update of Bulletin 17B is March 1982.

In 1895 Karl Pearson proposed choosing the density f(x) which solves the differential equation (Kite, 1977)

$$\frac{d \log f(x)}{dx} = \frac{x - a}{b_0 + b_1 x + b_2 x^2} \qquad (2.10)$$

where the four constants are based on the first four sample moments, and $b_1 = -a$. The gamma and beta distributions are two members of the Pearsonian family of distributions. In particular, the three-parameter gamma or Pearson Type III distribution has found wide use in hydrology.

If the logarithms, $Y = \ln X$ of a random variable X follow the Pearson Type III distribution, then the variate itself conforms to the log-Pearson Type III distribution. The probability density function for the LP III distribution is

$$f(x) = \frac{\exp\{-(\ln x - c)/a\}}{|a| x \Gamma(b)} \{(\ln x - c)/a\}^{b-1} \qquad (2.11)$$

where a, b and c are the scale, shape and location parameters respectively, and $\Gamma(.)$ is the gamma function. When $a > 0$, the distribution has positive skewness and is unbounded; but when $a < 0$, i.e. negative skewness, an upper bound exists at exp c.

Bobee (1975) has shown that the overall geometrical shape of the distribution depends on the scale and shape

parameters, a and b as follows:

$(1/a) > 0$:	$0 < b \leq 1$	reverse J-shape
	$b > 1$	unimodal bell-shape
$(1/a) < 0$:	$-1 < (1/a) < 0$; $0 < b \leq 1$	U-shape
	$(1/a) < -1$; $0 < b \leq 1$	J-shape
	$(1/a) < -1$; $b > 1$	unimodal bell-shape

Not all of these shapes are plausible for use in flood frequency analysis, especially those where a is negative. Condie (1977) has noted that a generally is negative for the flood series of Canadian rivers.

As the CDF of eq. 2.11 cannot be expressed in closed form, by making the substitution $z = (\ln x - c)/a$ and using the Wilson-Hilferty approximation for chi-square one obtains

$$\ln x_t = c + a\{t/(3b^{1/6}) - 1/(9b^{2/3}) + b^{1/3}\}^3 \qquad (2.12)$$

where $\ln x_t$ is the logarithm of the T-year event and t is the standard normal deviate for the required return period.

The WRC recommended fitting the LP III distribution to an AFS by computing the base 10 logarithms of the discharge Q, at selected exceedance probability p, from the equation

$$\text{Log } Q = \overline{X} + KS \qquad (2.13)$$

where $\overline{X}$ and S are defined below and K is a frequency factor that is a function of the skew coefficient and selected

exceedance probability (see also eq. 2.2).

$$\overline{X} = \frac{\Sigma X}{N} \tag{2.14}$$

$$S = \left[\frac{\Sigma X^2 - (\Sigma X)^2/N}{N-1}\right]^{0.5} \tag{2.15}$$

$$G = \frac{N^2\Sigma X^3 - 3N\Sigma X\Sigma X^2 + 2(\Sigma X)^3}{N(N-1)(N-2)S^3} \tag{2.16}$$

where

X = logarithm of annual peak flow

N = number of items in sample

$\overline{X}$ = mean of logarithms

S = standard deviation of logarithms

G = skew coefficient of logarithms

Thus the LP III distribution will reproduce the measured coefficient of skewness of the observed data, as well as its mean and variance.

One consequence of the WRC endorsement of the LP III distribution is that it sparked considerable research on distributional analysis by hydrologists. While some hydrologists contend that this choice was ill-advised, it is evident from the literature that there is no particular frequency distribution which is clearly preferable to the others. In view of this, the imposition of a choice, such as LP III by the WRC is perhaps not such a bad idea. It should be noted that the Australian Institution for

Engineers also recommended the use of the LP III distribution for Australian Rivers in 1977, and they too have become involved in the LP III debate.

2.3.5 Wakeby Distribution

The fact that nature has been given more inherently unstable skews than most statistical distributions has led to the Wakeby distribution (Houghton 1978). The Wakeby has been described as a grand parent and is similar to a five-parameter member of the Tukey family of lambdas. While the Wakeby has five parameters, the higher order moments are not used to estimate these parameters. Some of the 'separation effect' presented by Matalas et al. (1975) can be explained by the Wakeby, which is a very flexible distribution with its left-hand side (small observations) divorced from its right-hand side (large observations).

The Wakeby is most conveniently defined as an inverse distribution function

$$x = -a(1-F)^{b} + c(1-F)^{-d} + e \qquad (2.17)$$

where F is the CDF, F(x). Equation 2.17 is formulated such that a, b, c and d define the shape of each of the tails and e is the location parameter. Kuczera (1982) has observed that the Wakeby is robust because the argument F is an

integral which is more stable than a point or functional value.

Due to the versatility of the Wakeby distribution, it can be used to generate synthetic flows in the pattern of any of the more traditional distributions. However there are configurations of the Wakeby that cannot be mimicked by any three-parameter distribution. For this reason the Wakeby has been called a grand parent.

2.4 FITTING TECHNIQUES

Four techniques of fitting distributions (i.e. parameter estimation) in increasing order of efficiency are: graphical, least squares, moments and maximum likelihood. Referring back to eq. 1.1, the objective of parameter estimation is to obtain $\hat{\theta}_i$, an estimate of parameter θ_i for the p parameters of the density. The parameters $\hat{\theta}_{1,..,p}$ must be inferred from the AFS. Certain terms should be defined. An unbiased estimator is one whose expected value equals the quantity being estimated, i.e. $E(\hat{\theta}) = \theta$. If an estimator, $\hat{\theta}$, converges to the true value θ as the number of observations increase, the estimator is said to be consistent. An estimator $\hat{\theta}$ of a parameter θ is said to be the most efficient estimator if it is unbiased and its variance is at least as small as that of any other unbiased estimator.

The graphical method of parameter estimation consists of visually fitting a function through the coordinate pairs. Use must be made of plotting position (refer to Section 2.2) and various types of graph paper with transformed coordinates are availabe to simplify the process. The basic disadvantage of this method is that it is very subjective.

Legendre in 1805 was the first to publish the theory of least squares and Snyder in his development of the unit hydrograph was the first to apply the technique in hydrology. Least squares estimation techniques minimize the sum of squared differences between values predicted from the theoretical function to be fitted and the observed values from an empirical distribution. Partial derivatives of the objective function are taken, yielding a number of equations equal to the number of parameters to be estimated.

The method of moments depends on relating either absolute or central moments to the parameters of the distribution. The general equation for the rth absolute moment is

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x)dx \qquad (2.18)$$

and for the rth central moment is

$$\mu_r = \int_{-\infty}^{\infty} (x-\mu_1')^r f(x)dx \qquad (2.19)$$

where μ_1' is the first absolute moment. Most distributions are adequately defined by the first two or three moments. The location parameter is analogous to the first absolute moment, the scale parameter to the square root of the second central moment (i.e., standard deviation), and if present the shape parameter depends on the third central moment. The third central moment rescaled by standard deviation so as to be dimensionless is called the coefficient of skewness.

Maximum likelihood estimation theory was developed by R. A. Fisher approximatly 1910. The principle of maximum likelihood states that the joint probability, L, of obtaining a sample of n values from a distribution with probability density function f(x;a,b,...) is proportional to the product

$$L = \prod_{i=1}^{n} f(x_i; a,b,...) \qquad (2.20)$$

L is the likelihood, and the parameters a,b,..., are chosen to maximize L. It is usually more convenient to maximize the log likelihood function, lnL. Maximization is obtained by taking the partial derivatives of lnL with respect to each parameter and equating to zero

$$(\partial \ln L)/(\partial a)=0; \quad (\partial \ln L)/(\partial b)=0; \ldots \qquad (2.21)$$

Parameters estimated by the method of maximum likelihood are generally considered as the most efficient obtainable.

Recently, a new fitting technique based on probability weighted moments has been introduced (Landwehr, 1979) which is especially useful for distributions, such as the Wakeby, that can only be expressed in inverse form. Bayesian analysis has been investigated by Vicens et al. (1975) as a method for combining regional information with single station records in order to reduce parameter uncertainties.

Monte-Carlo experiments by Nozdryn-Plotnicki and Watt (1979) investigated the small sample properties of three alternative fitting techniques for the LP III distribution. The techniques were the method of moments as recommended by the WRC, a maximum likelihood method in log space proposed by Condie (1977), and a third method developed by Bobee (1975) which preserved the moments of the original data rather than those of the logs of the flow. While maximum likelihood was best for large samples ($n > 100$), no one method was superior to the others for small samples. In an effort to circumvent the problem of skew coefficent determination with the method of moments, Rao (1980) has proposed the method of mixed moments. This method does not require the skew coefficient, but uses a combination of lower-order moments in both real and log space.

Due to the sampling error of the third moment for small sample sizes ($n < 100$), the WRC has recommended that the sample skew estimate be replaced by a linear combination of the sample skew and a regional skew value obtained from a skew map. Matalas et al. (1975) have shown that skewness estimates obtained from historical flood sequences were more variable than any of the commonly used frequency distributions, a phenomenon that they called the 'condition of separation'. In an extension of this work into log space Landwehr and Matalas (1978) have shown that the usefulness of regional skew maps in log space is questionable. Different skew contours in real space could give rise to identical contours in log space and vice versa.

An alternative fitting approach for the three-parameter lognormal distribution which uses the median instead of the coefficient of skewness has been proposed by Sangal and Biswas (1970). However Burges et al. (1975) in their investigation of skewness coefficient bias correction have shown that estimates using this method are less efficient than those based on skewness.

2.5 GOODNESS-OF-FIT TESTS

The choice between frequency distributions can be made by classical statistical tests such as the chi-square test or the Kolmogorov-Smirnov test.

The statistic

$$\chi_c^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i \qquad (2.22)$$

is distributed as chi-square with n = k-p-1 degrees of freedom where k is the number of class intervals, p the number of parameters, O_i the observed and E_i the expected (according to distribution being tested) number of observations in the ith class inverval. The hypothesis that the data are from the specified distribution is accepted at the α level of significance if

$$\chi_c^2 < \chi_{n;\alpha}^2 \qquad (2.23)$$

where the value $\chi^2_{n;\alpha}$ is available from tables.

One problem with the chi-square test is the loss of information due to grouping. The Kolmogorov-Smirnov test provides bounds within which every observation must lie if the sample is drawn from the specified distribution.

Passing a goodness-of-fit test is a necessary, but not a sufficient criteria for acceptance of a certain distribution. The difficulty is these conventional goodness-of-fit tests do not reflect on the quality of fit in the right-hand tail, an important consideration when obtaining design floods. Recently a test which weights specifically on the high or low flood data as desired has

been developed by Lin and Bothe (1984) of the Hydrology Branch, Alberta Environment.

2.6 RECENT RESEARCH TRENDS

Greis (1983) in a review article notes that major accomplishments in flood frequency analysis between 1979 and 1982 are in the areas of (1) distributional analysis (2) robust parameter estimation and (3) regionalized and geomorphologic approaches to the flood frequency problem.

In the area of distributional analysis, some hydrologists have come to feel that for flood frequency analysis, unimodal models are suspect. Most of the common parametric forms, including the ones discussed in Section 2.3, are unimodal. However, it has been noted that riverflow is the final link in a long cause-and-effect chain which probably makes it a more complex process than its causative factors (Klemes, 1980), and various more complex models, either bimodal or multimodal, have been proposed. Wood and Rodriquez-Iturbe (1975) have suggested a composite model consisting of a linear combination of individual distributions, each weighted by a parameter. K. P. Singh and M. Nakashima (1983) have concluded that "the mixed distribution method coupled with the detection and modification of outliers and inliers may be the best available method for flood frequency analysis at present."

Rossi (1984) has suggested that the two-component extreme value (TCEV) distribution can be assumed as a parent flood distribution. The TCEV distribution assumes individual floods arise from a mixture of two exponential components and its four parameters can be estimated by the maximum likelihood method.

Kuczera (1982) has argued that effort should be directed towards developing robust distribution functions that yield stable estimates of the recurrence probabilities. The word 'robust' has many connotations, but generally signifies insensitivity in the model given a moderate perturbation among the data. In the context of flood frequency models two necessary properties of a robust model are resistance over a wide parameter range and efficiency, i.e. smaller mean square error. Kuczera notes that density functions which contain exponential terms produce wide fluctuations in extreme events estimated from the tails given even small errors in the parameters. Based on sampling experiments, Kuczera (1982) concluded that the two-parameter lognormal maximum likelihood procedure was the most robust of the single station procedures considered.

Recently efforts have been made to use nonhydrologic information to increase effective record lengths. Costa (1978) suggests that geologic methods such as (1) pollen analysis and radiocarbon dating of organic flood deposits

(2) dating of submerged landforms and (3) tree ring analysis are applicable to flood frequency analysis. Benson (1950) documents a method of combining censored historical data and recent data. Condie and Lee (1982) treat the AFS together with any historic information as a censored sample from a three-parameter lognormal population, and derive the maximum likelihood solution.

Some nonparametric studies have appeared in the hydrologic literature, however they are quite different than the methods that will be investigated in this thesis. The approach taken by Lettenmaier and Burges (1978) is "based on a linear programming formulation for the maximum and minimum exceedance probabilities of an arbitrary distribution with a given set of moments taken as constraints." A cubic spline fit at pre-specified nodes was used to approximate the probability density function. The problem dealt with by Cooper and Clarke (1980) involved inferring the probabilistic characteristics of a streamflow variable from a correlated longer record, with the joint probability being distribution free. Unny et al. (1981) have considered the nonparametric approach in the context of pattern recognition as it pertains to time series synthesis. Yakowitz (1985) has presented a nonparametric Markov modeling approach for daily river flow forecasting.

A final research trend is that there is currently much interest in a greater linking between hydrology and climate, especially through the processing of remote sensing data.

2.7 LIMITATIONS OF THEORETICAL FREQUENCY DISTRIBUTIONS

The major problems and limitations of using theoretical frequency distributions in flood frequency analysis are:

1. A priori distributional selection. The nonparametric approach, as discussed in the next chapter, may well provide an alternative to assuming a particular theoretical frequency distribution for use in flood frequency analysis.

2. The estimation of the distribution's parameters. Yevjevich et al. (1984) have noted that the estimation of skewness, or of parameters related to skewness, is a weak link in the application of mathematical statistics to hydrology. A large area of the WRC log skew map has negative values which implies a finite upper bound. This bound is not physically meaningful, and in some cases is less than observed floods, or those inferred from high water marks.

3. The identification and treatment of outliers. Outliers can significantly affect distribution parameters, especially for small samples. If there is not a rationale for removing

outliers, then they should remain in the data set, otherwise incorrect estimates will result.

4. The assumption of homogeneity in flood data. The causative factors of floods vary and as well local changes in the watershed, such as increased urbanization, can alter flood data.

5. The handling of multimodal densities. Multimodal densities cannot be handled by the conventional parametric forms, however recently there has been increased research interest in the treatment of mixed distributions.

6. A final problem with conventional analysis is tail pathologies. Several different distributions may fit an AFS equally well, but yield very different flood magnitudes due to their different tail behaviors.

The nonparametric approach is not a magic panacea for these problems of conventional flood frequency analysis. However, in addition to not needing a distributional assumption, the nonparametric approach requires only one design parameter, and can readily estimate multimodal densities and handle nonhomogeneous data. Comments on the tail behavior of nonparametric estimates and the effect of outliers are premature at this stage.

CHAPTER III

LITERATURE REVIEW AND DEVELOPMENT OF NONPARAMETRIC TECHNIQUES

3.1 GENERAL

The objective of this section is to introduce and trace the historical development of nonparametric techniques, ending with a tabular comparison of various estimators.

Many different types of nonparametric density estimators are available, including the histogram, k-nearest neighbor estimator, maximum likelihood estimator, kernel estimator, orthogonal series estimator, spline estimator and penalized likelihood estimator. During the last twenty-five years a very extensive literature has developed in the area of nonparametric density estimation. Almost all of the literature is in the form of papers, and the following review articles provide a useful summary of the field to the uninitiated: Tarter and Kronmal (1976), Fryer (1977), Bean and Tsokos (1980), Wegman (1972a, 1972b). These articles are listed in order of increasing complexity. The first Wegman article gives an outline of methods to 1972 and their properties including a detailed discussion on types of consistency. In the latter paper, Wegman gives results of a Monte-Carlo study comparing the effectiveness of a trigonometric orthogonal series estimator, a kernel

estimator, and a maximum likelihood estimator. The review articles provide many references and a detailed bibliography up to 1977 is given in Wertz and Schneider (1979). Many of the over half dozen pattern recognition texts published in the early 1970's contained a chapter on nonparametric techniques, however the first English language book entirely devoted to the topic is Tapia and Thompson (1978). This excellent book considers the histogram, Pearson distributions, series estimators and kernel estimators; but concentrates on maximum penalized likelihood estimation methods.

In Chapter I it was mentioned that Rosenblatt (1956) was one of the first researchers of nonparametric density estimation. His research stemmed from an investigation into the properties of histograms. Histogram construction is such a routine process we may fail to realize that even when using identical class interval widths, different choices of their locations may result in very different looking histograms for the same data. This fallibility of the classical histogram prompted development of the shiftable histogram in which the intervals are shifted so that every sample point is in the center of an interval. The Rosenblatt estimator, also known as a running or floating histogram, is given by

$$\hat{f}(x) = \frac{\#\text{ observations in } (x-h, x+h)}{2Nh} \qquad (3.1)$$

where $X_1, X_2, \ldots, X_N$ is a random sample of observations from a continuous but unknown density F, and h is a constant known as the smoothing factor. Alternately,

$$\hat{f}(x) = \frac{F(x+h) - F(x-h)}{2h} \tag{3.2}$$

where

$$F(x) = \frac{\#\text{ observations} < x}{N} \tag{3.3}$$

If the sample points are likened to an energy source, the potential at any of these points attains a maximum value, but decreases rapidly at any point away from the sample point. Rosenblatt (1956) demonstrated his class of univariate estimators to be pointwise and integratedly consistent in quadratic mean as long as h is suitably chosen. The 'shiftable' histogram has been generalized by Parzen, and is now known as the kernel estimator.

Parzen (1962) imposed constraints on the kernel, which he called the weighting function, and investigated the asymptotic properties of seven suitable kernels. Parzen showed uniform consistency in probability and popularized the kernel estimator to such an extent that it is also called the Parzen estimator, or Parzen-window approach. Since Parzen's work, various authors have obtained stronger consistency results under differing conditions. Another has extended much of Parzen's work to the multivariate case.

The kernel estimator will be discussed in greater detail in Section 3.3 of this chapter.

Although much of the statistical literature on nonparametric estimation deals with properties such as convergence and asymptotic behavior, uniqueness and optimality, strong or weak and global or pointwise consistency; the engineer is inclined to accept such results on faith and an in-depth discussion of these properties is beyond the scope of this thesis. However it is useful to stress an important difference between the histogram and the kernel estimator at this point. While the histogram is computationally efficient, it is quite inefficient in terms of integrated mean squared error (IMSE). The IMSE is widely accepted as a global measure of performance of an estimator $\hat{f}$ of a density f and is defined by

$$\text{IMSE} = E\int[\hat{f}(x) - f(x)]^2 \, dx \tag{3.4}$$

Given a random sample $x_1, \ldots, x_n$, the histogram IMSE is asymptotically minimized proportional to $N^{-2/3}$ when the histogram class interval width is optimally chosen (Scott, 1979). On the other hand, the kernel estimator is statistically efficient with IMSE rate convergence in the order of $N^{-4/5}$. The advantage of $N^{-4/5}$ convergence is reminiscent of the advantage of central differences in numerical analysis. Yakowitz (1985) has pointed out that

even if the true density function was known a priori, then on the basis of the sample size N the IMSE of a parametric estimate is usually N^{-1}.

In 1951 Fix and Hodges wrote a paper on nonparametric discrimination which influenced the development of the k-nearest neighbor density estimator of Loftsgaarden and Quesenberry (1965). This estimator is given by

$$\hat{f}(x) = [(k(n)-1)/n][1/Ar_{k(n)},x] \qquad (3.5)$$

where $x \in R^d$, $Ar_{k(n)}$, x is the volume of the d-dimensional hypersphere about point x with radius $r_{k(n)}$, which is the distance to the k(n)th nearest neighbor. Thus the number of observations is fixed and the volume is random. In the univariate case the 'sphere' reduces to a line segment and the 'volume' becomes the length of the line. The counterpart of choosing h in the kernel estimator is choosing the number of nearest neighbors in the k-nearest neighbor estimator. While the estimator of Loftsgaarden and Quesenberry does not require the choice of a kernel, it is essentially the uniform kernel with the smoothing factor chosen by a different criteria. The k-nearest neighbor estimator was quite popular during the 1970's, but the interest in it has since decreased. As the k-nearest neighbor estimator is not a probability density function, it is of limited use in flood frequency analysis.

Various types of orthogonal series estimators have been considered, including those based on trigonometric, Laguerre and Hermite functions. Cencov (1962) was the first to suggest that Fourier series be used for the estimation of probability density functions. The general formula for an orthogonal series estimate is

$$f(x) = \sum_{k \in I} a_k g_k(x) \qquad (3.6)$$

with

$$\hat{a}_k = \left(\frac{1}{N}\right) \sum_{j=1}^{N} g_k(X_j) r(X_j) \qquad (3.7)$$

where g_k is a family of orthogonal functions with respect to some weighting function r. This estimator is not always nonnegative. However negative estimates have been interpreted as evidence there is insufficient data to provide estimates in those regions. One approach to smoothing the series estimator is that of Cencov (1962), who has proposed a stopping rule for the number of terms in the series based on the truncation of high frequency terms. Tarter and Kronmal (1976) recommend an additional inclusion rule and have considered series estimators for the cumulative distribution function as well as the density function. A second approach to smoothing the series estimator has been to apply a low-pass window to the series expansion. Orthogonal series estimators have certain operational advantages, including the fact that during

computation all samples need not be stored and the final result is easily stored. However, one disadvantage is that the IMSE depends substantially on the choice of the series, which is not made by the data, but is left up to the user.

The spline estimator is somewhat related to the orthogonal series estimator. The 'histospline' approach of Boneva, Kendall and Stefanov (1971) begins with a histogram and ends up solving a cubic spline in order to obtain a smooth estimate of an unknown probability density function. A problem with the estimator is that it can assume negative values. Schoenberg (1972) has proposed an alternative called the 'spliegram' which provides a smoother estimate. The procedure again begins with the histogram but uses parabolic arcs in a way which preserves the nonnegativity. Wahba, in a series of papers beginning in 1970, has made an important contribution to the formulation of spline smoothing in its modern form.

As noted by Wegman (1972a), in general maximum likelihood estimates (MLE) for density functions do not exist; however, if the estimate is selected from a suitably restricted class of densities, then a maximum likelihood estimate for that class may exist. Robertson has considered the case of a MLE for a unimodal density with known mode, while Wegman has investigated the case where the mode is unknown. In his maximum likelihood approach, Wegman has

obtained a modified histogram estimator where the interval widths are inversely proportional to the density of sample points in the interval. Tapia and Thompson (1978, p.45) have proved that the histogram estimator is actually a maximum likelihood estimator. Wegman (1972b) has noted that MLE's are very data sensitive, have fast computation times and few arbitrary choices, however have very poor IMSE properties.

The penalized likelihood estimator will be discussed in Section 3.5 of this chapter.

An important concept common to all the types of estimators discussed here is that they all require some sort of factor which controls the degree of smoothing in the estimate. Density estimates can be thought of to summarize the data. Estimates with different degrees of smoothing will emphasize different features. Thus the selection of the smoothing factor is critical. Even the choice between methods sometimes is less serious than the choice of the smoothing factor.

Fryer (1977) has tabulated some of the advantages and disadvantages of the kernel estimator, along with some other estimators, as shown in Table 3.1.

TABLE 3.1

Some advantages and disadvantages of the various types of estimators (After Fryer (1977))

Orthogonal Series	Kernel	Histogram MLE
For:		
Optimal Stopping Rule exists	Can be density	Few arbitrary choices
Easy to compute	Can have infinite support	Can have infinite support
Good IMSE	Good IMSE and rate of convergence	Easy to compute Error rate seems higher than $O(N^{-4/5})$
Against:		
Choice of series critical	Kernel to be chosen although the choice is not critical	Very poor IMSE
Not usually a density	Critical choice of h	
Often finite support	Large amount of computer time required	Interval length or number of observations per interval to be chosen

For reasons summarized in Table 3.1, the kernel estimator is emphasized in this thesis. The relevant theory is presented in Sections 3.3 and 3.4 of this chapter. However, as graphical representation is very important in engineering, especially when used to inform and convince clients, graphical nonparametric methods are discussed in Section 3.2. Primarily since the computer routine NDMPLE is

readily available from IMSL, the Nonparametric Discretized Maximum Penalized Likelihood Estimator is detailed in Section 3.5.

3.2 GRAPHICAL NONPARAMETRIC DENSITY ESTIMATION METHODS

3.2.1 Histogram

The histogram is probably the oldest probability density estimator, and it is the classic nonparametric statistical tool for the graphical display of data. Examples of near histograms may be found in the Old Testament, i.e. the censuses of able-bodied fighting men per tribe by Moses (Numbers, Chs. 1 and 26). In 1661 John Graunt, a London haberdasher, represented birth and death data in a manner which was perhaps the first formal histogram. Statistician Karl Pearson gave the histogram its name in 1895.

The process of using histograms is so simple we may not think of the decisions which must be made prior to histogram construction. These decisions include (1) the number of intervals to use, (2) the size of these intervals, and (3) their location. Various rules of thumb exist for the number of intervals, some completely arbitrary like between 5 and 15 intervals, others based on the sample size n, i.e. the equation $1 + 2.2 \log n$ has been suggested. Interval size is normally determined by dividing the sample range by the

desired number of intervals. However this is complicated by the fact that equal-probability histograms with intervals of varying size can also be constructed. The location of the intervals is chosen so that observations do not fall on interval boundaries. Finally, the last step in histogram construction is determining the frequency of observations in each interval, and graphing the results.

In the more precise language of mathematics, given the random sample $X_1, X_2, \ldots, X_n \in (a,b)$, consider an equally spaced mesh t_i where $a = t_1 < t_2 < \ldots < t_{m+1} = b$. Let q_i be the number of sample points falling in the ith interval. Clearly, $\sum_{i=1}^{m} q_i = N$. The histogram may be defined as

$$\hat{f}(x) = \frac{q_i}{N(t_{i+1} - t_i)} \quad , \quad \text{for } x \in (t_i, t_{i+1}) \tag{3.8}$$

If F is the cumulative distribution function as defined by eq. 3.3, the histogram may also be written as

$$\hat{f}(x) = \frac{F(t_{i+1}) - F(t_i)}{t_{i+1} - t_i} \text{ for } x \in (t_i, t_{i+1}), \; i=1,2,\ldots,m \tag{3.9}$$

It is instructive to compare this equation with that for the shiftable histogram, eq. 3.2.

The considerable latitude in histogram construction might lead the investigator to force the histogram to 'look right', thereby biasing the presentation of the data. In addition to being completely ad hoc, problems with the histogram include the fact that it is difficult to update

and as the sample size goes to infinity, the number of characterizing parameters increases without limit. Furthermore, histograms should not be used to compare the distributions of two populations, as the process is so arbitrary that even histograms constructed from samples from the same population can appear very different.

In an effort to decrease the subjectivity involved in histogram construction, Scott (1979) has shown that choosing a histogram interval size or bin width of

$$h^* = [6/\int_{-\infty}^{\infty} f'(x)^2 dx]^{1/3} N^{-1/3} \qquad (3.10)$$

is the optimal choice based on an IMSE criterion. It is assumed that f is a continuous probability density function with two continuous and bounded derivatives. For any density and any positive constant c, the IMSE using a nonoptimal bin width ch* is larger than the minimum IMSE by the factor $(c^3 + 2)/(3c)$. Thus using a bin width 50% too small implies an IMSE 42% too large. For Gaussian data, eq. 3.10 becomes

$$h^* = 3.49sN^{-1/3} \qquad (3.11)$$

where s is the sample estimate of the standard deviation. Scott (1979) considered the lognormal density as a model of skewed data, and noted that the presence of significant skewness requires a reduction in the bin width from that

given by eq. 3.11. As flood data is invariably skewed, the correction factor is reproduced in Fig. 3.1. The exact form of the equation graphed in Fig. 3.1 is quite complicated, however it is well-approximated over the range shown by

$$\alpha(\gamma) = (1 - 0.0060\gamma + 0.27\gamma^2 - 0.0069\gamma^3)^{-1} \qquad (3.12)$$

where γ is the coefficient of skewness and α is the correction factor.

3.2.2 Frequency Polygon

The frequency polygon is constructed from an equal interval histogram by connecting the midpoints of the bars with straight lines. Its advantage over the histogram is its extra visual smoothness, which makes it more useful for comparing sample estimates from differing populations. However some statisticians, including Fisher, have belittled the frequency polygon, saying that any advantage is illusionary.

In an insightful but as yet unpublished paper, Scott (1984) has shown that while the frequency polygon is computationally equivalent to the histogram, it is statistically similar to the kernel estimator. Scott proves that an optimally constructed frequency polygon achieves an

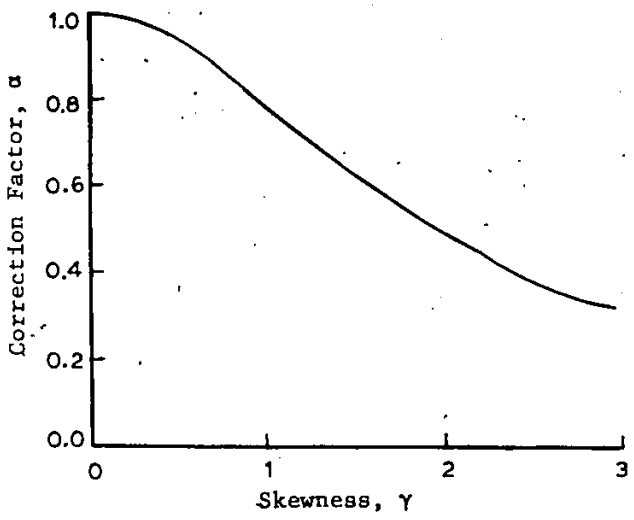


FIG. 3.1 Ratio of histogram bin width to theoretical Gaussian bin width with same variance (α) versus skewness (After Scott, 1979).

IMSE rate convergence of $N^{-4/5}$. Choosing the bin width to be

$$h^{**} = 2[15/(49 \int_{-\infty}^{\infty} f''(x)^2 dx)]^{1/5} N^{-1/5} \qquad (3.13)$$

asymptotically minimizes the frequency polygon IMSE. For any density and any positive constant c, the IMSE using a nonoptimal bin width ch** is larger than the minimum IMSE by the factor $(c^5 + 4)/(5c)$. Thus using a bin width 50% too small implies an IMSE 61% too large. Two observations can be made. The first is that frequency polygons (also kernel estimators) are more sensitive to nonoptimal choices of smoothing factors than histograms. The second is that the optimal frequency polygon will in general be constructed from a nonoptimal histogram. For Gaussian data, eq. 3.13 becomes

$$h^{**} = 2.15sN^{-1/5} \qquad (3.14)$$

where s is the sample standard deviation. Again, for skewed data a correction factor from Fig. 3.1 can be applied.

3.3 KERNEL ESTIMATOR

The kernel estimator, already defined by eqns. 3.1 and 3.2, can be written in more general form as

$$f(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x-X_i}{h}\right) \qquad (3.15)$$

where $X_1, X_2, \ldots, X_N$ is a random sample of observations, f(.) the unknown density, K(.) the kernel, and h a design parameter commonly called the smoothing factor or window width. The kernel must satisfy various conditions of unbiasedness and consistency which are given in Wegman (1972a) as:

$$\int_{-\infty}^{\infty} K(y)dy = 1 \tag{3.16}$$

$$\int_{-\infty}^{\infty} |K(y)|^2 dy < \infty \tag{3.17}$$

$$\sup_{-\infty<y<\infty} |K(y)| < \infty \tag{3.18}$$

$$\lim_{y\to\infty} |yK(y)| = 0 \tag{3.19}$$

$$K(-y)=K(y) \geq 0 \tag{3.20}$$

Conditions on h are:

$$\lim_{N\to\infty} h = 0 , \quad \lim_{N\to\infty} Nh = \infty , \quad \text{and} \quad h > 0 \tag{3.21}$$

The choice of h is critical and in practise not easy. Too large an h implies large bias, an oversmoothed estimate, and consequent loss of information. Too small an h implies large variance and too rough an estimate. An analogy is that in the former case we see too much of the forest; while in the latter case we see too many trees in the forest. As h approaches zero, K(.) approaches a Dirac delta function centered at X_i, and f(.) approaches a

superposition of delta functions centered at the samples, where the Dirac delta function $\delta(.)$ is defined as

$$\int_{-\infty}^{\infty} \delta(y)dy = 1$$
$$\delta(y) = 0 \text{ if } y \neq 0 \qquad (3.22)$$

Many investigators, including Anderson (1969) and Schuster (1982) have determined that the choice of kernel seems to matter little. Three useful kernels, including their interval of support and $\alpha(K_i)$ (see eq. 3.26) are given in Table 3.2

TABLE 3.2

Typical kernels (After Tapia and Thompson (1978, p. 60))

Kernel, K_i	Interval of Support	$\alpha(K_i)$
$K_1(y) = 1/2$	$\|y\| \leq 1$	1.3510
$K_2(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$	$\|y\| < \infty$	0.7764
$K_3(y) = \frac{3}{4}(1-y^2)$	$\|y\| \leq 1$	1.7188

The rectangular kernel, K_1, is computationally efficient due to finite support and use of absolute value; however estimates tend to be rough. The Gaussian kernel, K_2, is computationally inefficient due to infinite support and use of exponentiation; however it is still quite popular. Silverman (1982) has noted that with the Gaussian kernel, computing the kernel estimate directly from eq. 3.15 is

highly inefficient. He has proposed a more efficient method in which the estimate is obtained as a multiple of the product of the Fourier transform of the data and the kernel. The quadratic kernel of the general form given by K_3 has been shown by Epanechnikov (1969) to be optimal in terms of IMSE rate of convergence for univariate kernel estimation regardless of the sample size or true underlying density. The precise form of the kernel derived by Epanechnikov is

$$K_4(y) = \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3y^2}{20\sqrt{5}} & \text{for } |y| \leq \sqrt{5} \\ 0 & \text{for } |y| > \sqrt{5} \end{cases} \tag{3.23}$$

Even earlier, Bartlett (1963) proposed the following asymptotically optimal kernel

$$K_5(y) = \begin{cases} \frac{9}{8h}(1-5y^2/3h^2) & \text{for } |y| \leq h \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

Greblicki and Krzyzak (1979) have also recommended the parabolic kernel. It should be noted that many symmetric kernels, including the Gaussian kernel are nearly optimal. The Epanechnikov kernel only gives a 4% improvement in the IMSE lower bound over the Gaussian kernel (Tapia and Thompson, p. 76).

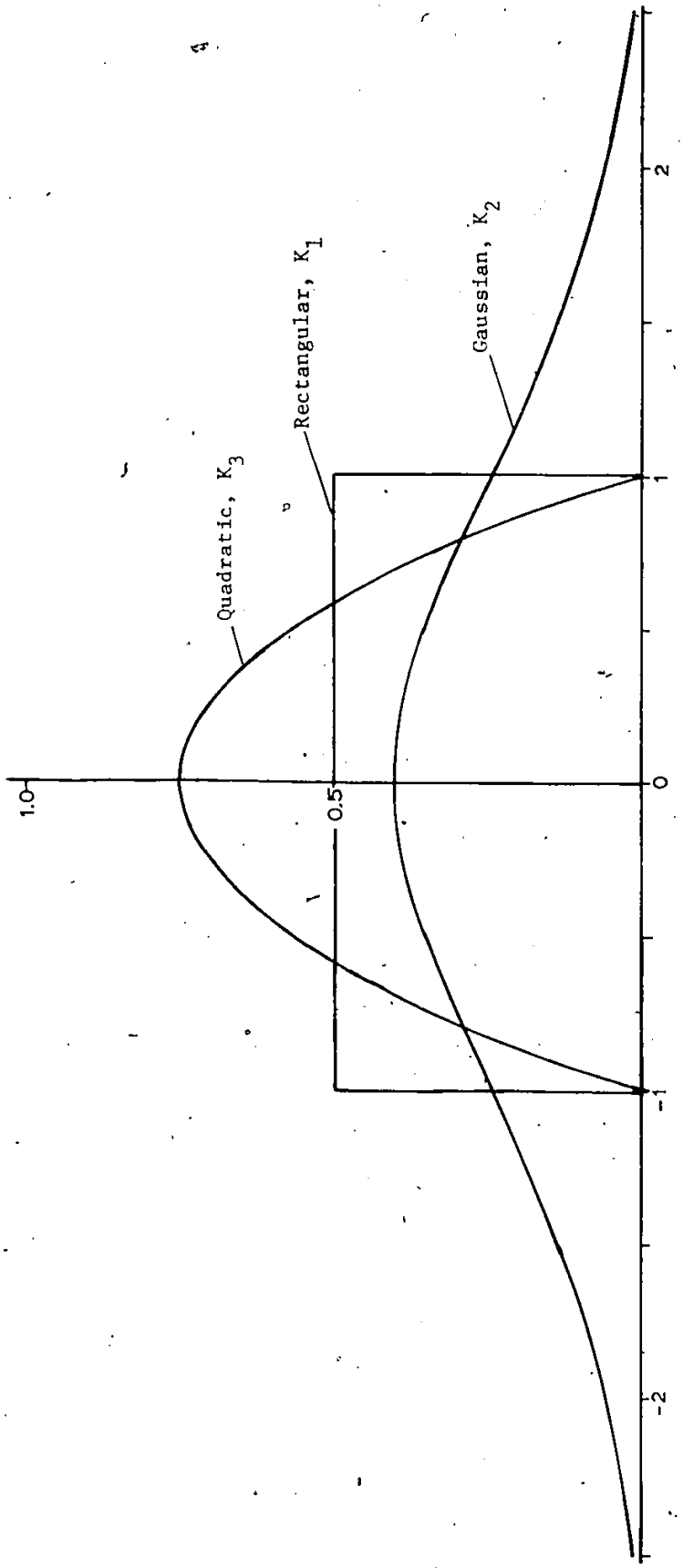
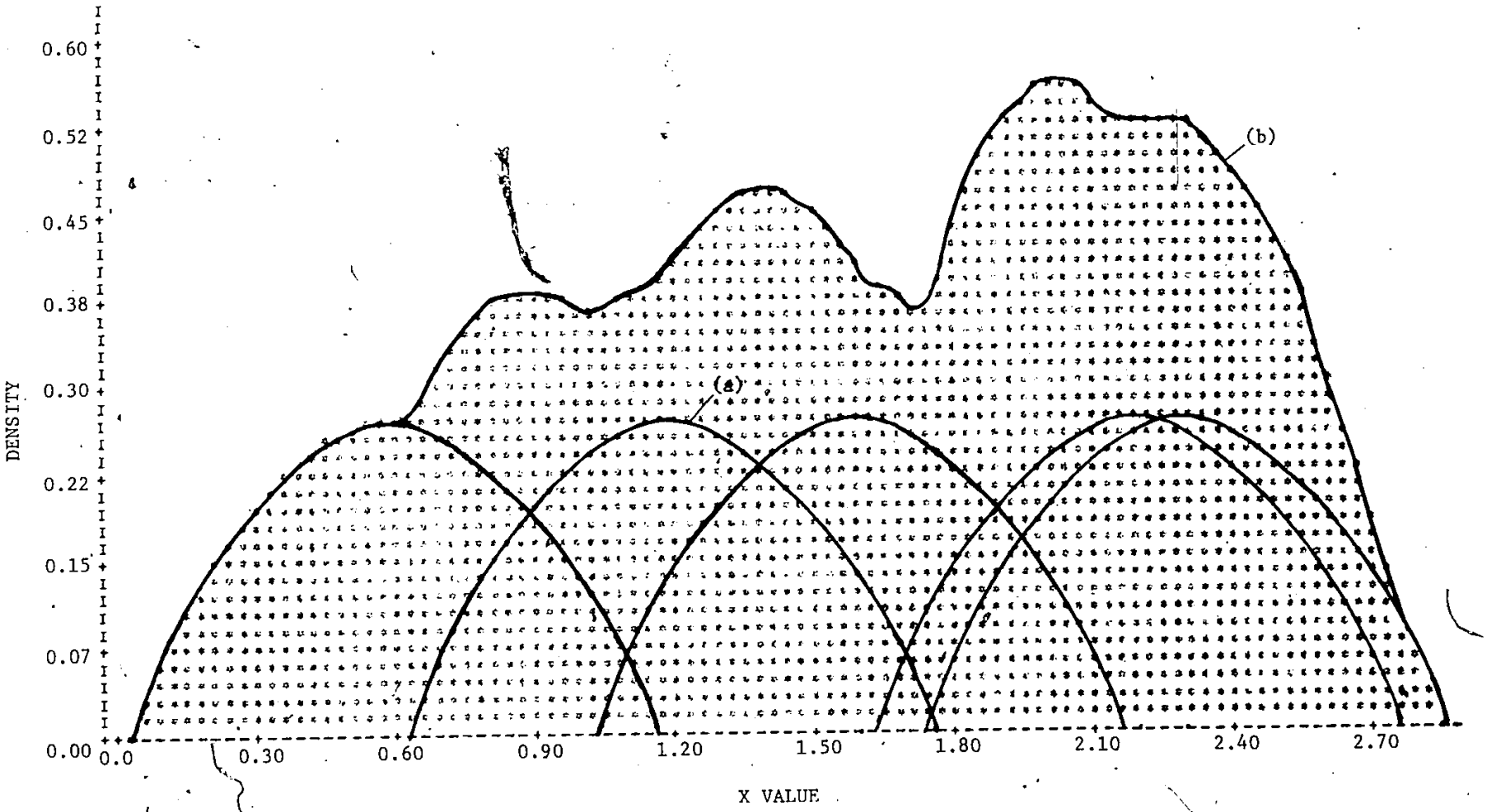


FIG. 3.2 Relative Weighting of Three Kernels



Smoothing factor, h = 0.562256 h by method of DHH Kernel = Quadratic, K_3 Data used is {0.6,1.2,1.6,2.2,2.3}

FIG. 3.3 Approximation of a density function (b) by the sum of quadratic kernels (a).

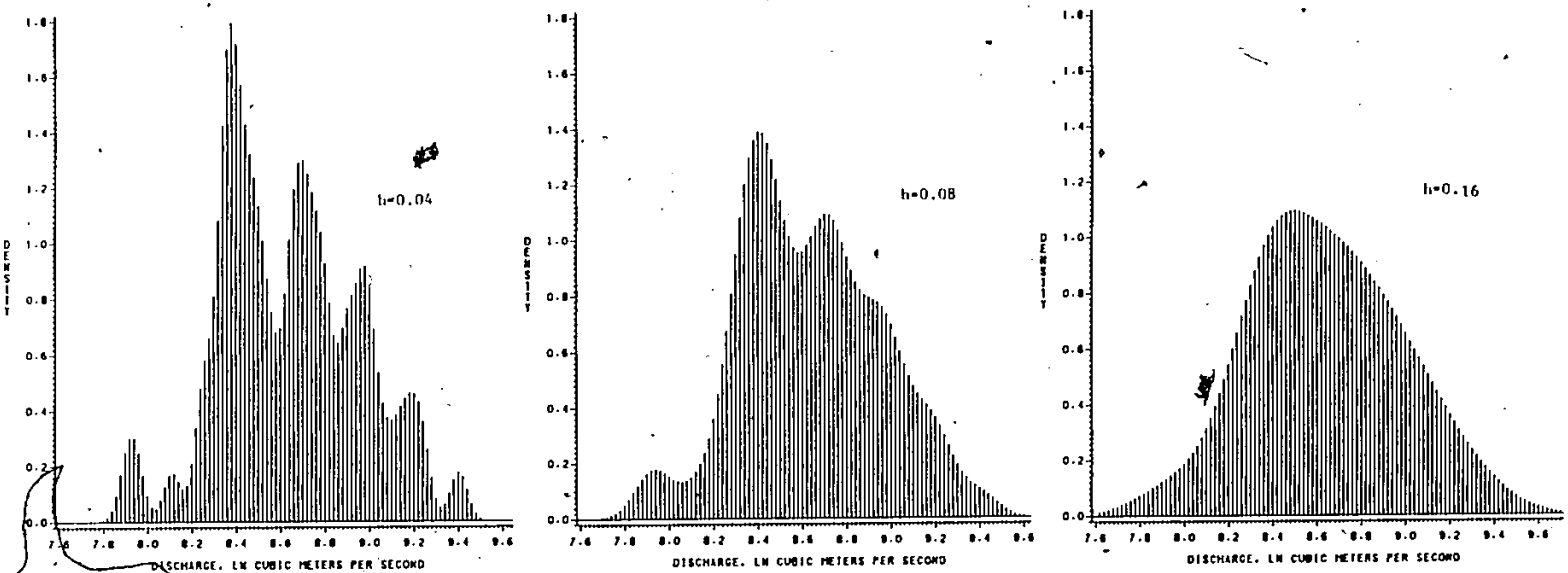


FIG. 3.4 Effect of varying the smoothing factor (h) on kernel density estimates (Data used is WSC 01FB001,1917-1973).

The rectangular, Gaussian and quadratic kernels, are all used in this thesis. The shapes of these three commonly used kernels are shown in Fig. 3.2. In this figure the difference between kernels with finite support (K_1, K_3) and infinite support (K_2) is readily apparent.

In a kernel estimate the density is constructed by summing N kernels centered on the N samples. An illustration of the kernel method approximation of a density function by the sum of quadratic kernels is shown in Fig. 3.3. The data used in Fig. 3.3 consists of the following five sample points, {0.6, 1.2, 1.6, 2.2, 2.3}. This test data will be used to illustrate further points in this chapter. How the smoothing factor h controls the degree of smoothing in a kernel estimate is demonstrated in Fig. 3.4.

Both the kernel function and the smoothing factor represent empirical choices. Since all error measures depend on the unknown density, they generally cannot be used in deriving analytical expressions for selecting the smoothing factor. However, a global optimal h can be obtained by minimizing the IMSE and substituting $\hat{f}(x)$ for f. Scott et al. (1981) present this asymptotically optimal choice for the smoothing factor h as follows:

$$h = \alpha(K)\ \beta(f)\ N^{-1/5} \qquad (3.25)$$

where

$$\alpha(K) = [\int K(y)^2 dy]^{1/5} [\int K(y) y^2 dy]^{-2/5} \qquad (3.26)$$

and

$$\beta(f) = [\int f''(x)^2 dx]^{-1/5} \qquad (3.27)$$

Thus it can be seen that the smoothing factor h depends on the true density f, the kernel $K(.)$, and the sample size N. When one has obtained a good estimate of the smoothing factor for a particular kernel, say h_1 for K_1, a good estimate h_2 for any other kernel K_2 may be obtained by scaling as follows:

$$h_2 = \frac{\alpha(K_2)}{\alpha(K_1)} h_1 \qquad (3.28)$$

Some values for $\alpha(K_i)$ are listed in Table 3.2.

As h is an empirical choice, some statisticians have advocated the ad hoc procedure where plots of $\hat{f}$ are examined, starting with h values which are too large and then decreasing h until density estimates which are too rough are obtained. The general idea of this procedure can be ascertained from Fig. 3.4. Tapia and Thompson (1978, p.67) have noted that "the point where further attempts at resolution by decreasing h drives one into a highly noisy estimate is generally fairly sharp, and readily observable." This method is very useful in data exploration; however it

requires an interactive environment, is time consuming and subjective, and is best done by a skilled statistician. All of these factors limit the applicability of ad hoc procedures in flood frequency analysis.

An interesting modification is to examine plots of f" rather than of f. The second derivative is much more sensitive to small changes in h then the density function itself. This procedure has been developed by Silverman (1978) and results in pictures he calls "test graphs." Silverman has shown that if the random fluctuations of f" are of maximum size $\pm k \sup f''$, then the h corresponding to this test graph should be used to construct the density. The constant k depends on the kernel, and the factor $\sup f''$ is estimated from the test graph. Unfortunately the test graph procedure is still subjective. Silverman (1981) has also published an article describing a statistical test which uses kernel density estimtes to investigate the number of modes in a population.

Woodroofe (1970) was one of the first to tackle the problem of objective determination of the smoothing factor h. He considered a two step procedure; the first step estimating the kernel density in the usual way, then using f to estimate h. Woodroofe's procedure is theoretically based on minimizing the mean square error at a point, but it is slow to converge and computationally impractical.

In recent years there has been considerable research interest in algorithms that automatically choose a smoothing factor suitable for a given set of data. One such procedure, a new method of cross-validation called least squares cross-validation has been proposd by Rudemo (1982) and Bowman (1982). Hall (1983) has investigated the procedure, showing it to be asymptotically optimal in addition to being consistent. Various other data-based algorithms which experimentally compute the h value are presented in detail in the following section.

3.4 DATA-BASED ALGORITHMS

3.4.1 Method of Adamowski and Feluch (AF)

The method described in this subsection was first used with the Cauchy kernel for its simplicity of implementation and physical justification (i.e., Cauchy kernel is unbounded). However, the density estimate obtained was much too sensitive to the data points and in fact was a Dirac solution. De Montricher, Tapia and Thompson (1975) note that continuous functions, differentiable functions, infinitely differentiable functions, and polynomials can all approximate Dirac delta functions.

Further experimentation, including split-record testing, was done with a number of other kernels, including the triangular and rectangular ones, and it was concluded

that the rectangular kernel was preferable among those tested. This kernel, also known as the uniform or boxcar kernel, is

$$K(y) = \begin{cases} 1/2 & |y| < 1 \\ 0 & |y| > 1 \end{cases} \tag{3.29}$$

Substituting eq. 3.15 into eq. 1.5 gives

$$p(x) = \frac{1}{Nh} \sum_{i=1}^{N} \int_{x}^{\infty} K(\frac{x-X_i}{h})dx = \frac{1}{Nh} \sum_{i=1}^{N} C_i(x) \tag{3.30}$$

$$\text{where } C_i(x) = \int_{x}^{\infty} K(\frac{x-X_i}{h})\, dx \tag{3.31}$$

For the rectangular kernel,

$$C_i(x) = \begin{cases} h & \text{for} \quad x \leq X_i - h \\ \frac{1}{2}(X_i+h-x) & \text{for} \quad X_i - h < x < X_i + h \\ 0 & \text{for} \quad x \geq X_i + h \end{cases} \tag{3.32}$$

The smoothing factor h can be determined by minimizing the following mean square error expression

$$\sum_{j=1}^{N} [p_j - p(x_j)]^2 \xrightarrow{\min_h} \tag{3.33}$$

where p_j can be obtained empirically using plotting formula. Differentiating eq. 3.33 and equating to zero gives

$$\sum_{j=1}^{N} [p_j - p(x_j)] \frac{dp(x_j)}{dh} = 0 \tag{3.34}$$

where

$$\frac{dp(x_j)}{dh} = \frac{1}{Nh} \sum_{i=1}^{N} [\frac{dC_i(x_j)}{dh} - \frac{1}{h} C_i(x_j)] \tag{3.35}$$

and, when using the rectangular kernel, one obtains the following

$$\frac{dC_i(x)}{dh} = \begin{cases} 1 & \text{for} \quad x < X_i - h \\ 1/2 & \text{for} \quad X_i - h < x < X_i + h \\ 0 & \text{for} \quad x > X_i + h \end{cases} \qquad (3.36)$$

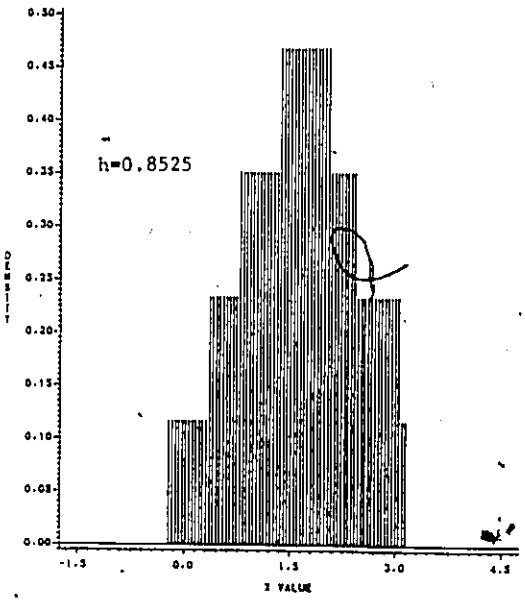
The probability of exceedance of an ordered set of observations to be used in eq. 3.33 can be computed by using a distribution-free plotting formula (Adamowski, 1981)

$$P_j = \frac{j-0.25}{N+0.50} \qquad (3.37)$$

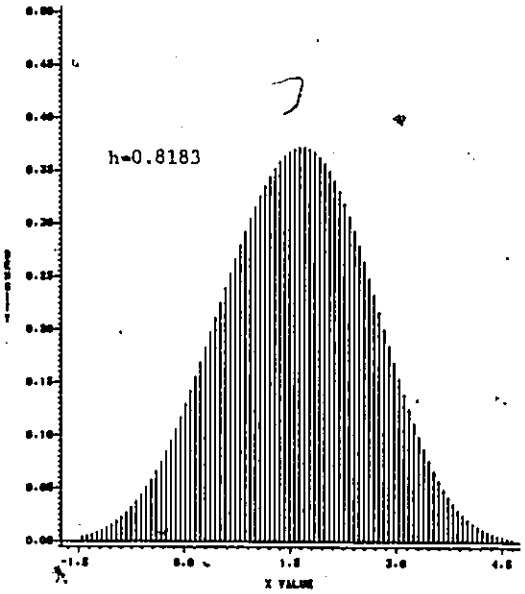
The Adamowski formula is used due to its good performance for N>25 as shown in Fig. 2.1.

The value of the smoothing factor h can be obtained by solving eq. 3.34. The method of successive approximations is used. Then from eq. 3.30 for any assumed value of x, a probability can be determined. Similarly, for any assumed probability p (or return period T = 1/p) the corresponding value of x_p can be calculated.

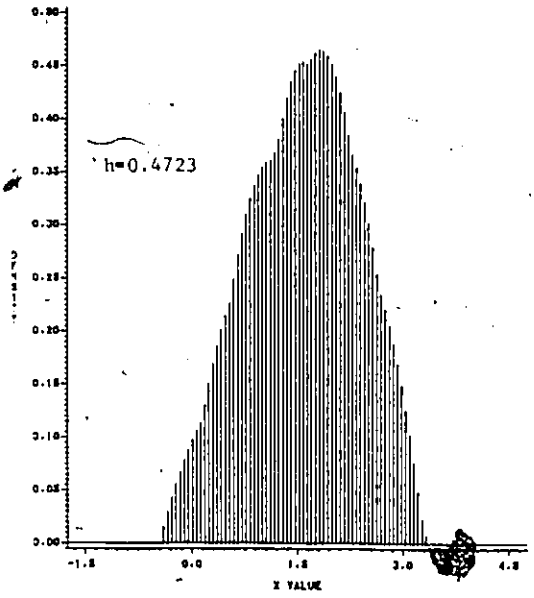
As the method of Adamowski and Feluch is quite complicated computationally, a computer is required to analyse the test data sample {0.6, 1.2, 1.6, 2.2, 2.3}. The program, which will be discussed in the next chapter, returns a smoothing factor h = 0.8525. A plot of the resulting density is shown in Fig. 3.5(a). It is apparent how the density is constructed by summing rectangular 'blocks'. It is also apparent that the density is very



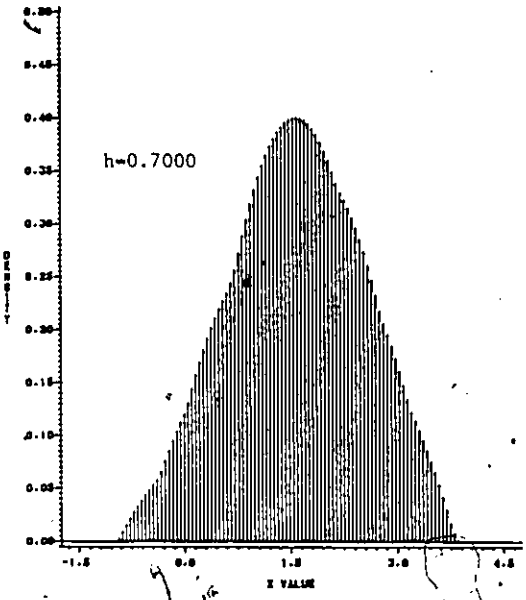
a) h^o by Adamowski/Feluch Procedure



b) h by Method of Scott, Tapia & Thompson



c) h by Flanagan's Formula



d) h by Method of Wagner

FIG. 3.5 Density Function Plots for Test Data {0.6,1.2,1.6,2.2,2.3}

rough. However it should be noted that a density is not really required in order to obtain flood estimates, as the method of Adamowski and Feluch works directly in terms of exceedance probability. This is in contrast to the methods discussed in the following subsections, whose basis is to first estimate the density, and then integrate numerically to obtain the probability distribution.

3.4.2 Method of Scott, Tapia and Thompson (STT)

Scott, Tapia and Thompson (1977) developed an algorithm based on functional iteration to calculate a quasi-optimal kernel smoothing factor h for the Gaussian kernel given a random sample from an unknown density. Given an initial choice h(0), they form the sequence

$$h(i+1) = \alpha(K)\ \beta(f_{h(i)})N^{-1/5} \tag{3.38}$$

where i denotes the iteration number,

$$\beta(f_h)^{-5} = \frac{3}{8\sqrt{\pi}N^2h^9} \sum_{j=1}^{N} \sum_{k=1}^{N} [h^4 - (x_j - x_k)^2h^2 + \frac{1}{12}(x_j-x_k)^4] \exp[-(x_j-x_k)^2/4h^2] \tag{3.39}$$

and $\alpha(K)$ has been given in eq. 3.26. They choose the largest nonnegative solution of eq. 3.38 and have shown that choosing h(0) to be the sample range guarantees convergence

to this solution. If the largest solution is h=0, they call the resulting Dirac delta estimate degenerate.

Eq. 3.39 is almost zero for large h, however as h approaches zero, eq. 3.39 increases at least as fast as an exponential. The quasi-optimal smoothing factor lies between these two extremes. In the algorithm Newton's method is employed and usually about 5 iterations are required to have $|h(i+1) - h(i)| < 10^{-5}$. However, a problem with the method of STT is that in practical applications it fails 5 to 10% of the time.

Scott et al. (1981) in a sensitivity analysis of three data-based algorithms demonstrated that the h values obtained by the method of STT were quite insensitive to the presence of a single outlier, and nearly independent of its value. The method of Duin (refer to next subsection) and Wahba's series estimator method were more sensitive.

The STT program required three iterations to determine the smoothing factor for the test data sample as follows:

Iteration	h
0	1.00000
1	0.80776
2	0.81828
3	0.81830

The density function plot is shown in Fig. 3.5(b). While the plot is smooth, being constructed from only five sample points it reveals as much information about the kernel

(i.e. Gaussian kernel was used) as about the true underlying density.

3.4.3 Method of Duin (DHH)

The method of Duin (1976), independently arrived at by Hermans and Habbema (1976), can be called a leave-out-one-at-a-time nonparametric maximum likelihood method. To avoid the degeneracy problem, the contribution of the sample itself is omitted in the estimation of the density at that point:

$$\max_{h>0} L(h) = \prod_{j=1}^{N} f_j(x_j) \qquad (3.40)$$

where

$$f_j(x_j) = \frac{1}{Nh} \sum_{\substack{i=1 \\ i \neq j}}^{N} K\left(\frac{x_j - x_i}{h}\right) \qquad (3.41)$$

In the terminology of Wahba eq. 3.40 is a cross-validation likelihood function, however the method was not shown to be cross-validatory until some years after being introduced (Hall, 1983). Schuster (1982) has noted that this method can produce inconsistent density estimates for long-tailed distributions such as the double exponential but works quite well for short-tailed distributions. Bowman (1982) showed that this method of cross-validation can be derived by minimizing the Kullback-Leibler loss function.

Duin (1976) has reported that the optimization of eq. 3.40 was "always executed by finding the zero crossings of its first derivative, using a special adapted version of the regula falsi." A 'canned' single parameter optimization routine was not available , but it was first attempted to find the maximum working directy with eq. 3.40 and repeatingly fitting quadratics. As this approach often did not converge, a new program was written working with the first derivative instead. In this thesis the method of Duin is used in conjunction with the quadratic kernel, K_4.

For the test data sample the program gives a value for the smoothing factor of h=0.5623. A plot of the resulting density function is shown in Fig. 3.3. The fine structure of the density is visible in this plot.

3.4.4 Formula Method of Flanagan (FF)

In his M.S. thesis entitled "Computer Programs for Kernel Estimates of a Probability Density Function", completed in 1980 at the University of South Carolina under the supervision of Dr. R. Taylor, P. Flanagan gave the following formula for the optimal bandwidth when using the quadratic kernel (K_4):

$$h = \frac{\sum_{i=2}^{N} \sum_{j=1}^{i-1} (X_i - X_j)}{\sqrt{5}\, N\, (N-10/3)} \qquad (3.42)$$

This formula is based on the theoretical development of Tapia and Thompson (1978) for a global optimal h, presented here as eqns. 3.25, 3.26 and 3.27.

For the test data sample {0.6, 1.2, 1.6, 2.2, 2.3} the numerator of eq. 3.42 works out to:

$$\begin{aligned} & 0.6 \\ + \quad & 1.0 + 0.4 \\ + \quad & 1.6 + 1.0 + 0.6 \\ + \quad & 1.7 + 1.1 + 0.7 + 0.1 \\ = \quad & 8.8 \end{aligned}$$

and

$$h = \frac{8.8}{\sqrt{5(5)(5-10/3)}} = 0.47226$$

The density function plot is shown in Fig. 3.5(c). Again, as with the method of DHH, the fine structure of the density is shown.

3.4.5 Method of Wagner (W)

In an effort to allow the data to play a role in the degree of smoothing about observation points, Wagner (1975) proposed the modified variable bandwidth kernel density estimator. Liu (1983) has written a computer program for this estimator.

The modified variable bandwidth estimator can be considered a combination of the kernel estimator and the nearst neighbor estimator. The nearest neighbor estimator

does not have a fixed bandwidth, but one which varies with the data.

In the modified variable bandwidth kernel the smoothing factor h is some symmetric function of the observations X_1, X_2, ..., X_n. Wagner (1975) suggests letting $k(n) = [n^a]$, where $0 < a < 1$, and D_{jn} be the distance from X_j to its k(n) nearest neighbour from X_1, X_2, ..., X_n (with X_j omitted) where $1 < j < n$. Then h may be selected at random from D_{1n}, D_{2n}, ..., D_{nn}. The quantities $\sum_{j=1}^{n} \frac{D_{jn}}{n}$, max $(D_{1n}, \ldots, D_{nn})$ and min $(D_{1n}, \ldots, D_{nn})$ will also work for h in the estimator.

In Liu's program $h = \sum_{j=1}^{n} \frac{D_{jn}}{n}$ is used to avoid problems caused by data duplicates (which often occur in flood data) in the random sample. In this thesis the method of Wagner is used in conjunction with the quadratic kernel, K_4.

For the test data sample {0.6, 1.2, 1.6, 2.2, 2.3}, choosing a=0.5, $k(n) = [n^a] = [5^{0.5}] = 2$. Then D1 = 1.0, D2 = 0.6, D3 = 0.6, D4 = 0.6, D5 = 0.7; and $h = \sum_{j=1}^{n} \frac{D_{jn}}{n}$ works out to h = (1.0+0.6+0.6+0.6+0.7)/5 = 0.70. A plot of the density function is shown in Fig. 3.5(d). The quadratic kernel K_4 has an interval of support of $\sqrt{5}$, so the density goes to zero at $0.6 - \sqrt{5}(.7) = -1.0$ and $2.3 + \sqrt{5}(.7) = 3.9$ as can be seen in the plot.

3.5 NONPARAMETRIC DISCRETIZED MAXIMUM PENALIZED LIKELIHOOD ESTIMATOR (NDMPLE)

Good and Gaskins (1971) suggested a means of choosing the smoothing parameter in a penalized maximum likelihood method. The basic concept was a trade off between goodness-of-fit to the data, as measured by the sum of squared residuals; and smoothness, as measured by the integrated squared second derivative. De Montricher, Tapia and Thompson (1975) showed that the problem introduced by Good and Gaskins was well defined regarding both existence and uniqueness of solutions. Good and Gaskins (1980) continued their work, finding the use of Fourier series in the computation of their estimates far preferable to the use of Hermite series.

For a given sample $X_1, \ldots, X_N$ the Φ - penalized likelihood of f is defined by

$$\hat{L}(f) = \prod_{i=1}^{N} f(x_i)\exp(-\Phi(f)) \tag{3.43}$$

where f is any probability density function and $\Phi(f)$ a penalty functional such as

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} \left| \frac{d^2 f(t)}{dt^2} \right| dt \tag{3.44}$$

Without the penalty functional there would be a degeneracy problem. The second derivative is used in the penalty term rather than the first derivative as it results in less local 'roughness'. Note that practically $\ln L(f)$ is maximized, and

eq. 3.43 separates into a log likelihood term and a log penalty term.

To implement this estimator a nonlinear constrained optimization problem must be solved. First the estimator must be discretized, with outlying data points being truncated. A modification of Newton's method can be used which makes use of the fact that the Hessian of the Lagrangian is a diagonally dominant tridiagonal matrix. The operation count per iteration is then $0(N^2)$ instead of $0(N^3)$ as is usually the case with Newton's method. Scott (1976) presents details of the method and NDMPLE, a computer program for this estimator, is part of the IMSL (1980) library.

Chapter IV

NUMERICAL ANALYSIS

4.1 DATA SELECTION

Two sets of data were chosen for numerical analysis, namely the Northeast Margaree River at Margaree Valley (Station WSC 01FB001) in Canada from 1917-1983 (N=67) and the Warta River at Poznan in Poland from 1826-1965 (N=140). The AFS consisted of maximum daily flows, with statistical characteristics shown in Table 4.1. The data itself can be found in Appendix A.

TABLE 4.1

Data Used for Numerical Analysis

River	Mean	STDV	C.V.	C.S.	C.K.
Margaree					
X Series	171.827	57.367	0.334	1.073	4.500
LNX Series	5.096	0.319	0.063	0.166	3.135
Warta					
X Series	436.457	319.690	0.732	1.770	6.046
LNX Series	5.863	0.646	0.110	0.272	2.759

NOTE: STDV = standard deviation; C.V., C.S., C.K. = coefficients of variation, skew and kurtosis respectively.

Both stations are for natural, unregulated watersheds. The location of Station WSC 01FB001 is LAT 46 22 10 N, LONG 60 58 36 W and its drainage area is 368 square kilometers.

The Margaree River for the period 1917-1973 has been analyzed using four different distributions (Gumbel, lognormal, three-parameter lognormal, LP III) by Condie et al. (1981). The Warta River was selected due to its long record length.

4.2 REAL DATA

4.2.1 Graphical Analysis

The two data sets were analyzed graphically and screened using tests available in the Consolidated Frequency Analysis Program documented by Pilon et al. (1985). The assumption that on an annual basis, floods are independently and identically distributed (iid.) can be checked by discharge-time and rank-time plots. For the Margaree River station these plots are given in Fig. 4.1a and b. There is no evidence of time-dependence or trend in the plots, and it seems unlikely that significant watershed changes have occurred during the period of record. Fig. 4.1c, a discharge-rank plot, is useful for identifying high or low outliers, which would appear as near-vertical lines at either end. There is no evidence of outliers in Fig. 4.1c.

The independence and randomness of the data can be verified by statistical hypothesis testing, as in Table 4.2

TABLE 4.2

Test for Independence, Trend, Randomness and Homogeneity
Northeast Margaree River at Margaree Valley (1917-1983)

Test	Critical Value at 5% Level	Test Result
Spearman Test for Independence: Spearman Rank Order Serial Coeff = 0.112 D.F. = 64 Corresponds to Students T = 0.898 Conclusion: There is no significant serial dependence.	1.670	Not Significant
Spearman Test for Trend: Spearman Rank Order Correlation Coeff = -0.018 D.F. = 65 Corresponds to Students T = -0.148 Conclusion: The Data do not display significant trend.	-1.998	Not Significant
Run Test for General Randomness: Runs Above and Below Median = 29 Runs Above Median = 33 Runs Below Median = 34 For this test, Z = 1.353 Conclusion: The sample is significantly random.	1.960	Not Significant
Mann-Whitney Homogeneity Test: Split by Time Span, Size of Subsample 1 = 33 Size of Subsample 2 = 34 For this test, Z = -0.684 Conclusion: The samples appear to be from the same population.	-1.645	Not Significant

TABLE 4.3

Test for Independence, Trend, Randomness and Homogeneity
Polish River Warta at Poznan (1826-1965)

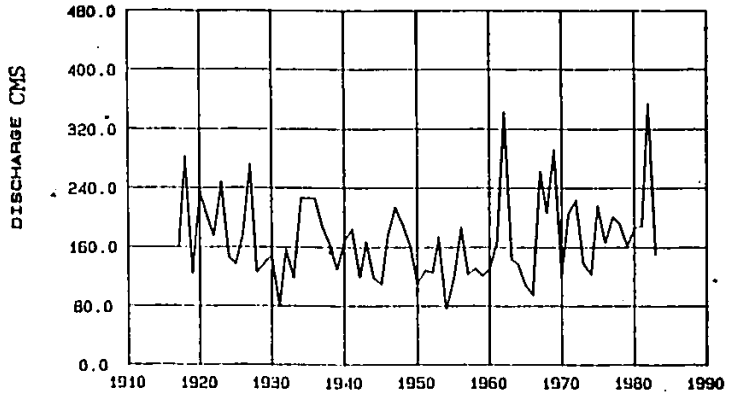
Test	Critical Value at 5% Level	Test Result
Spearman Test for Independence: Spearman Rank Order Serial Coeff = 0.161 D.F. = 137 Corresponds to Students T = 1.911 Conclusion: At the 5% level of significance, the correlation is significantly different from zero, but it is not so at the 1% level. That is, the dependence is significant, but not highly so.	1.645	Significant
Spearman Test for Trend: Spearman Rank Order Correlation Coeff = 0.123 D.F. = 138 Corresponds to Student T = 1.452 Conclusion: The data do not display significant trend.	1.960	Not Significant
Run Test for General Randomness: Runs Above and below median = 62 Runs Above Median = 70 Runs Below Median = 70 For the test, Z = 1.527 Conclusion: The sample is significantly random.	1.960	Not Significant
Mann-Whitney Homogeneity Test: Split by Time Span, Size of Subsample 1 = 68 Size of Subsample 2 = 72 For this test, Z = -1.632 Conclusion: The samples appear to be from the same population.	-1.645	Not Significant

for the Margaree River data. To test for homogeneity, the Mann-Whitney test can be used (Kite, 1977, p. 11).

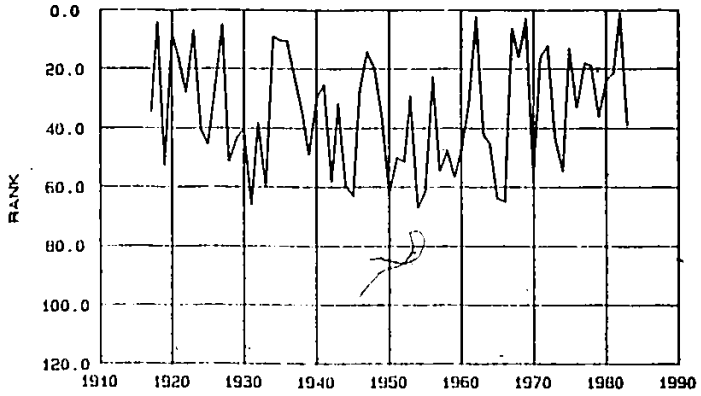
Discharge histograms for Margaree River are given in Fig. 4.2. The histogram in Fig. 4.2b is constructed using seven intervals. The histogram in Fig. 4.2a, by presenting the data in terms of % maximum discharge, avoids requiring any choice as to how many intervals to use. Fig. 4.3 gives the optimal histogram and frequency polygon for the same data. The interval sizes have been calculated from eqns. 3.11 and 3.14, using the skewness correction factor from Fig. 3.1. Note the similarity between Fig. 4.2a and Fig. 4.3a. While the frequency polygon gives the smoothest result, all the plots in Figs. 4.2 and 4.3 are similar and all adequately describe the data, perhaps indicating that the data is 'well-behaved'.

Fig. 4.4 gives the discharge-rank-time plots for the Warta River. The data appears random and there is no evidence of outliers. Statistical tests for the Warta River data are presented in Table 4.3.

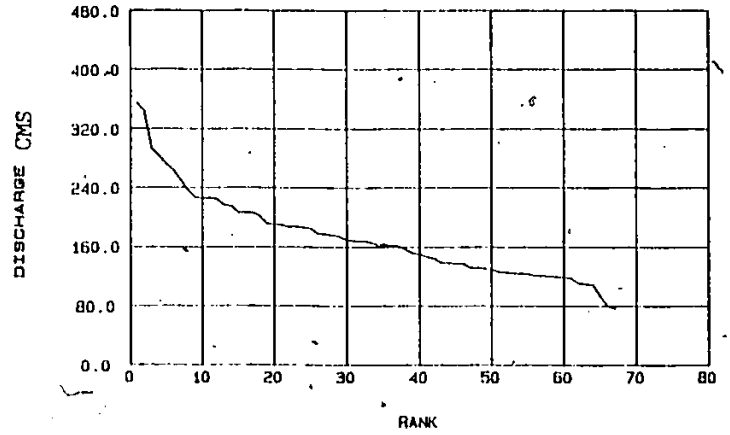
Discharge histograms for Warta River are given in Fig. 4.5. The optimal histogram and frequency polygon are given in Fig. 4.6. The % maximum discharge histogram in Fig. 4.5a and the frequency polygon in Fig. 4.6b are fairly similar in shape. Both have three modes and show 14 observations to be



a) Discharge versus Time

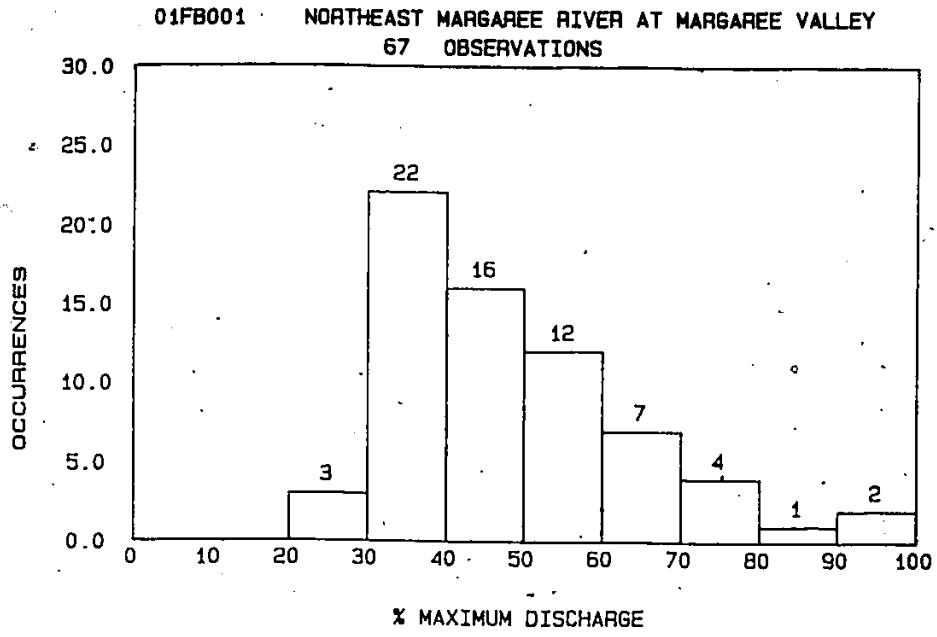


b) Rank (Descending Order) versus Time

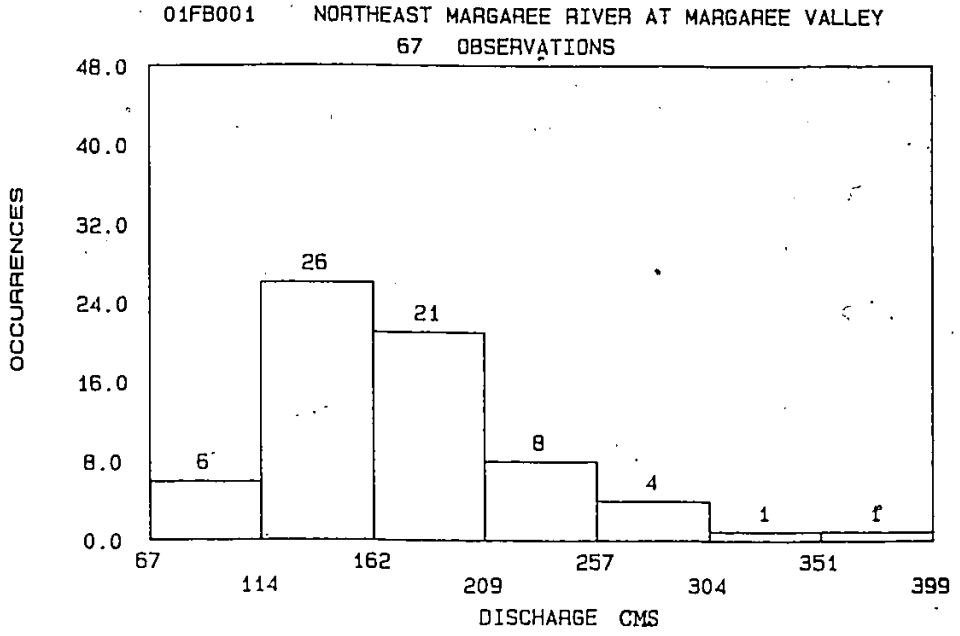


c) Discharge versus Rank

FIG. 4.1 Northeast Margaree River at Margaree Valley (1917-1983) Discharge-Rank-Time Plots

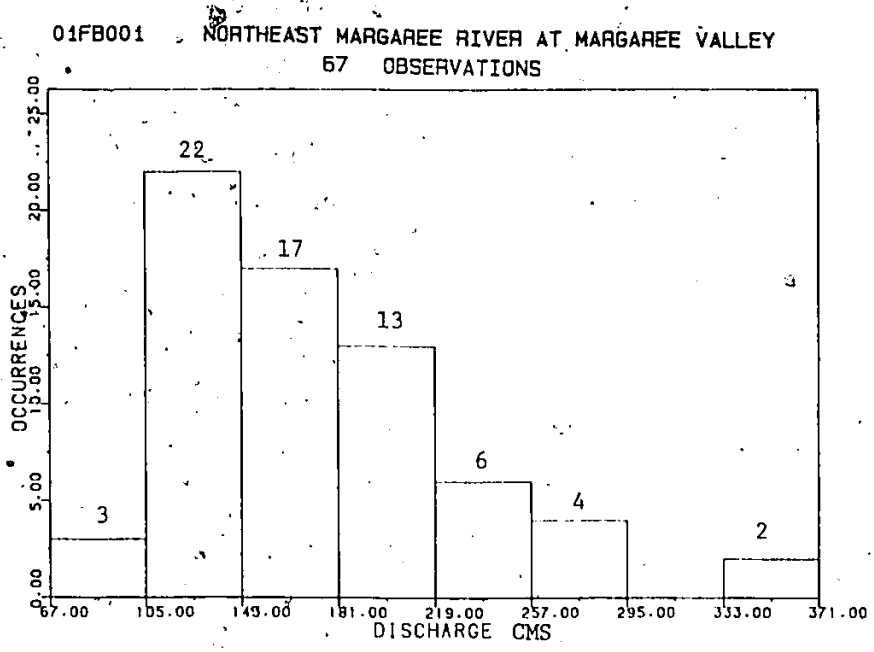


a) HISTOGRAM FOR DISCHARGE BY % MAXIMUM DISCHARGE: 1917 TO 1983

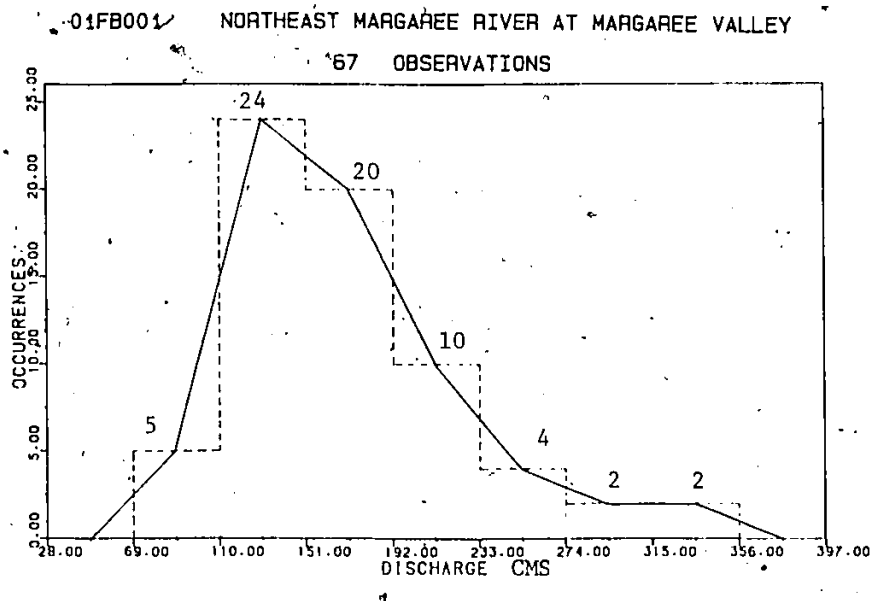


b) HISTOGRAM FOR DISCHARGE: 1917 TO 1983

FIG. 4.2 Northeast Margaree River at Margaree Valley (1917-1983) Discharge Histograms

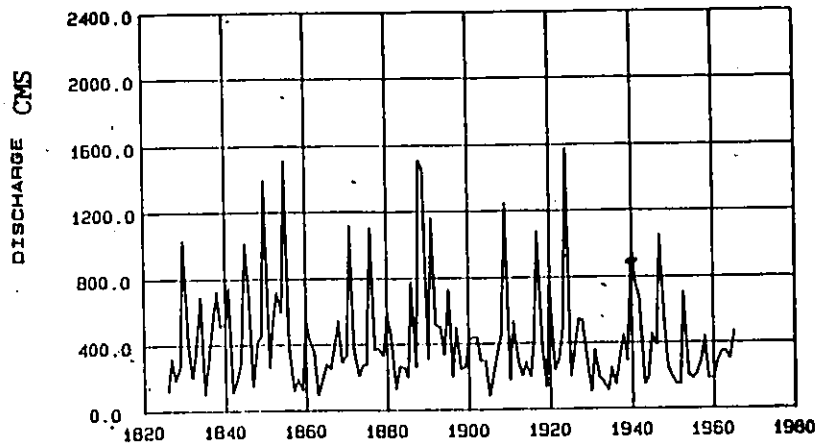


a) Optimal Histogram for Discharge: 1917 to 1983

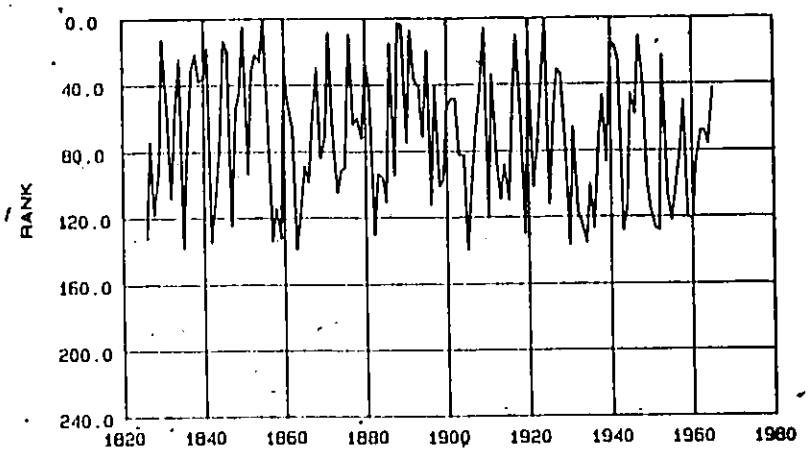


b) Optimal Frequency Polygon for Discharge: 1917 to 1983

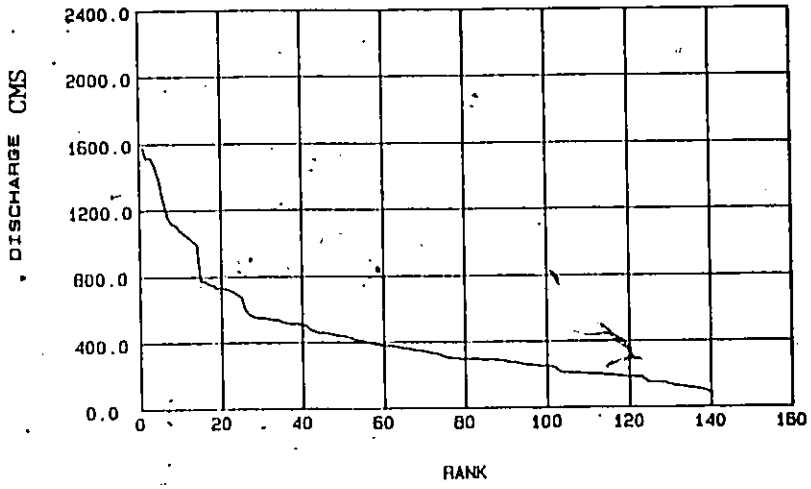
FIG. 4.3 Northeast Margaree River at Margaree Valley (1917-1983) Optimal Discharge Histogram and Frequency Polygon



a) Discharge versus Time

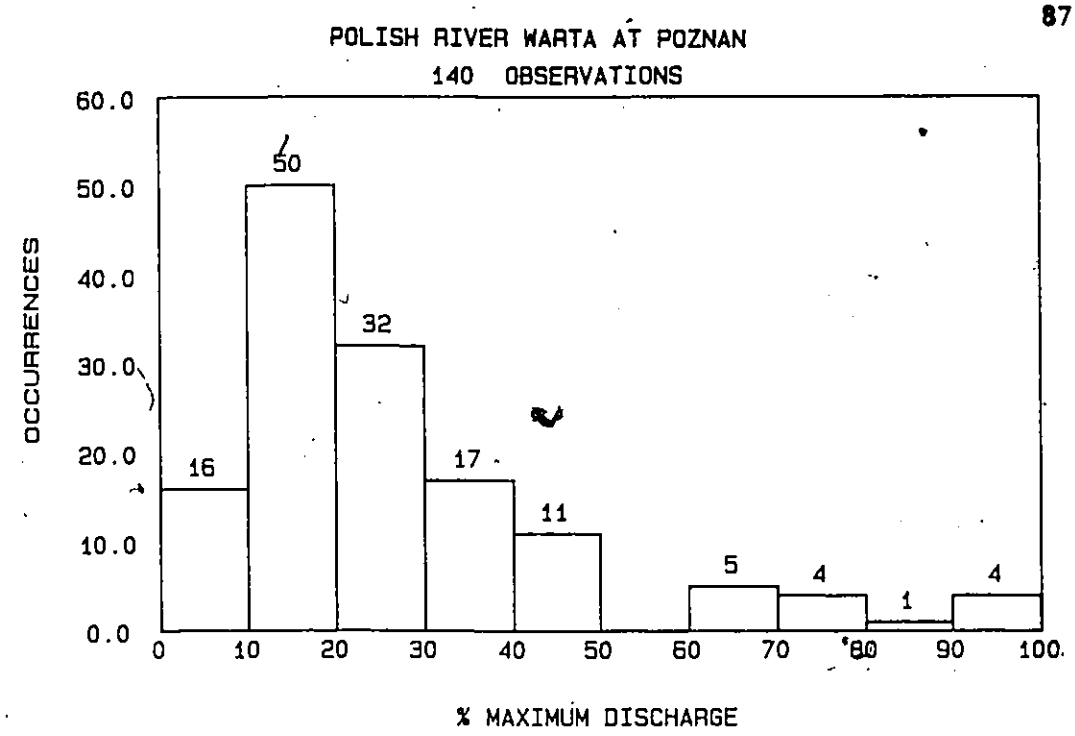


b) Rank (Descending Order) versus Time

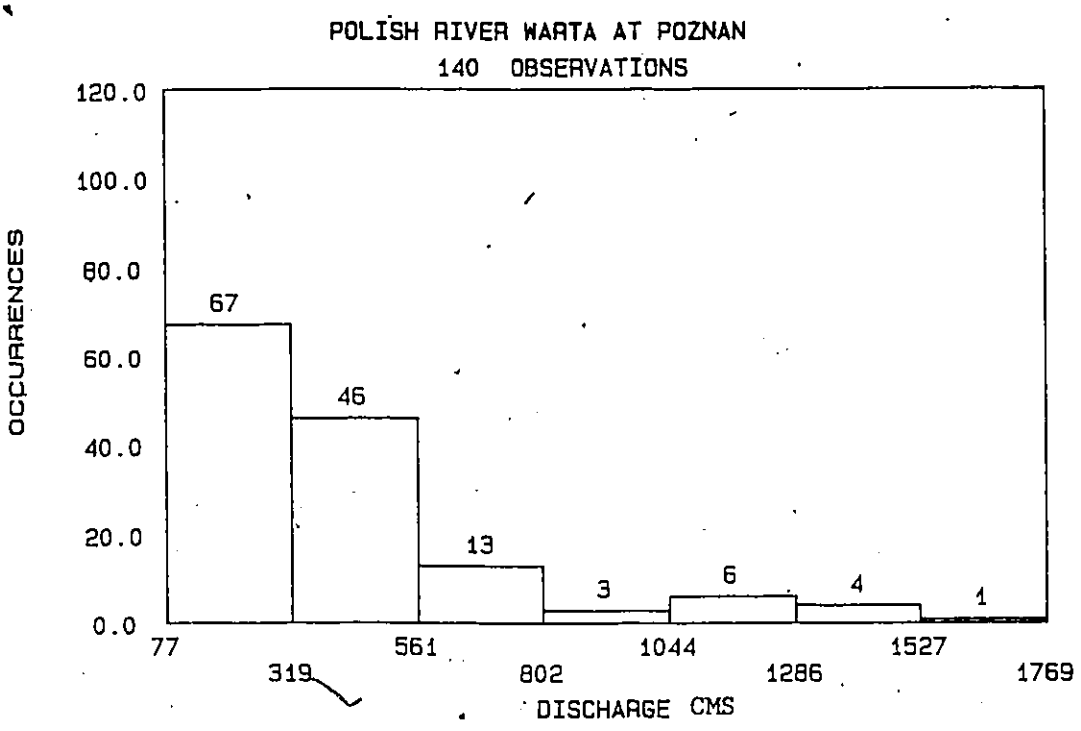


c) Discharge versus Rank

FIG. 4.4 Polish River Warta at Poznan (1826-1965). Discharge-Rank-Time Plots

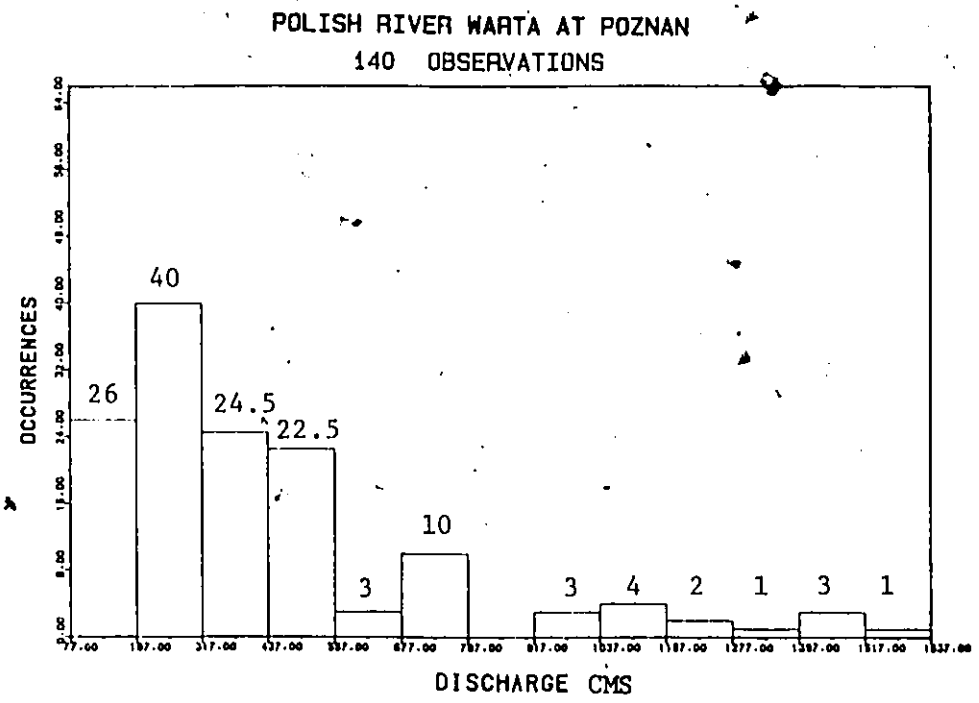


a) HISTOGRAM FOR DISCHARGE BY % MAXIMUM DISCHARGE: 1826 TO 1965

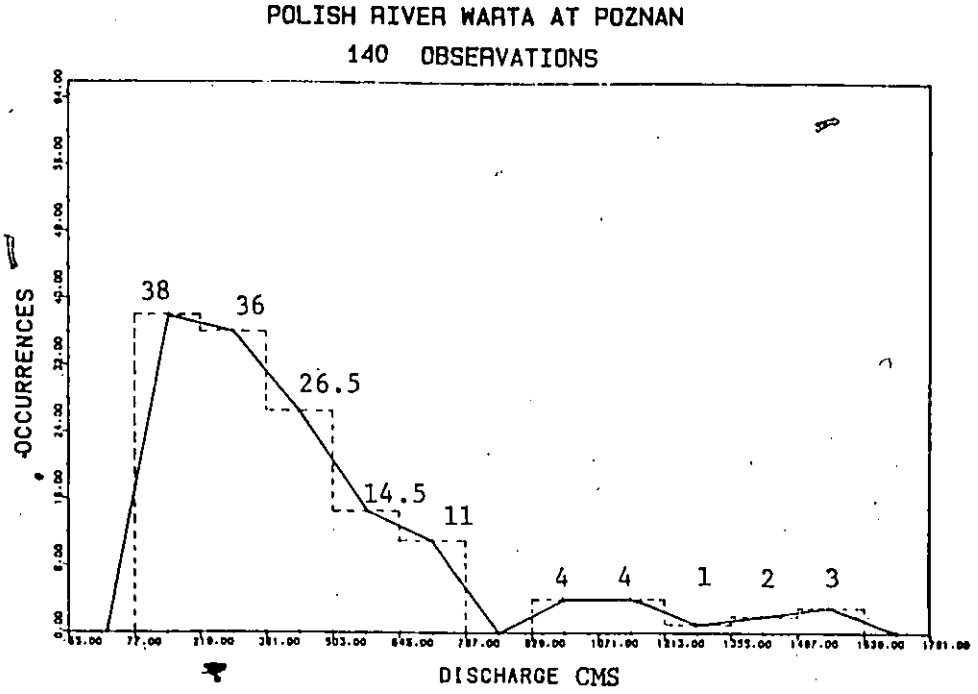


b) HISTOGRAM FOR DISCHARGE: 1826 TO 1965

FIG. 4.5 Polish River Warta at Poznan (1826-1965) Discharge Histograms



a) Optimal Histogram for Discharge: 1826 to 1965



b) Optimal Frequency Polygon for Discharge: 1826 to 1965

FIG. 4.6 Polish River Warta at Poznan (1826-1965) Optimal Discharge Histogram and Frequency Polygon

quite a bit higher than the rest. The optimal frequency polygon is constructed using 11 intervals of width 142 while the optimal histogram is constructed using 13 intervals of width 120. The optimal histogram in Fig. 4.6a has an additional mode. The histogram in Fig. 4.5b is constructed using an arbitrarily selected seven intervals. It does not adequately describe the data, suffering from loss of information. The optimal histogram has about twice as many intervals as the histogram in Fig. 4.5b.

These histograms and frequency polygon provide a preliminary look at the data and can be compared with the kernel densities in the next chapter.

4.2.2 Parametric Analysis

Flood frequency analysis of the two data sets were performed for the parametric distributions discussed in Section 2.3 using the Consolidated Frequency Analysis Program (CFA1) documented by Pilon et al. (1985). The complete results including frequency curves can be found in Appendix A.

The CFA1 program is a more powerful and easier to use update of the Flood Damage Reduction Program documented by Condie et al. (1981). CFA1 does single station analysis for the three-parameter lognormal, the generalized extreme value, the log-Pearson Type III and Wakeby distributions.

CFA1 is capable of handling non-standard cases involving combinations of historic information, low outliers or zeros. In the program parameters are estimated by maximum likelihood except in cases where they are unobtainable and then moment estimates are used.

The frequency analysis for the Northeast Margaree River at Margaree Valley and Warta River at Poznan stations were standard. No historic information was available, and there were no low outliers or zero flows. All solutions were obtained via maximum likelihood. Both stations have a positive coefficient of skewness, so in the three-parameter lognormal (LN3) distribution the location parameter c is a lower bound. For the LN3 distribution the sample estimates of skewness and kurtosis for the transformed data can be compared with their theoretical equivalents of 0 and 3. These are 0.002 and 3.210 for the Margaree River and -0.067 and 2.870 for the Warta River. For both stations with the GEV distribution, k is less than zero, which indicates a lower bounded EV2 distribution. For both stations with the LPIII distribution, a and b are greater than zero, indicating an unbounded unimodal bell-shaped distribution. For Margaree River the Wakeby distribution is upper bounded at 1605. As the flood estimate for the 500 year return period is only 441, this is not necessarily a problem.

Now comes the question of deciding which distribution gives the best representation of the flood frequency regime. The appearance of the frequency plots, Figs. A.1 - A.8 can aid in the decision.

It is apparent that for the Northeast Margaree River at Margaree Valley all the distributions fit well and give similar estimates of flood magnitude values. As the histograms indicate that the Margaree River station is 'good', this could be anticipated. Condie et al. (1981) recommended using the LN3 distribution for this station based on its very successful normalizing transformation.

For the Polish River Warta at Poznan the choice is more difficult. There is considerable variability between flood estimates using the GEV and Wakeby distributions. Either the LN3 or LPIII distributions could be recommended based on their frequency curves. However neither of these distributions fit the higher observations very well. These problems illustrate why nonparametric methods are being considered in this thesis.

4.2.3 Nonparametric Analysis

Computer programs for the kernel estimator with the smoothing factor automatically selected by methods presented in Section 3.4 can be found in Appendix B. The programs have been designated RECKER, NORKER, QUAKER and MVBKER;

acronyms for rectangular kernel, normal kernel, quadratic kernel and modified variable bandwidth kernel. These programs select h using the data-based algorithms of Adamowski and Feluch, Scott, Tapia and Thompson, Flanagan, and Wagner respectively.

The formula for the kernel estimator (eq. 3.15) should not be used directly for its calculation. This would be inefficient since if the density is to be estimated at m x-coordinates given a sample of N observations, the kernel would have to be evaluated mN times. If the kernel has finite support, both the data points X_i and x-coordinates can be ordered, and the interval of support specified so as to achieve greater efficiency.

Flowcharts for RECKER and MVBKER are given in Figs. 4.7 and 4.8. The MVBKER flowchart is also valid for NORKER and QUAKER, the only differences being that h is selected by a different method and used with a different kernel. Detailed comments describing the input expected, important variables and subprograms can be found in the programs themselves. NORKER requires specifying an initial choice for the smoothing factor, h(0). If the user has no prior knowledge, the sample range can be used. RECKER requires specifying EPS, an accuracy limit to stop the method of bisection. For untransformed data EPS=0.01 was used and for log transformed data EPS=0.0005 was used. These values allow the smoothing

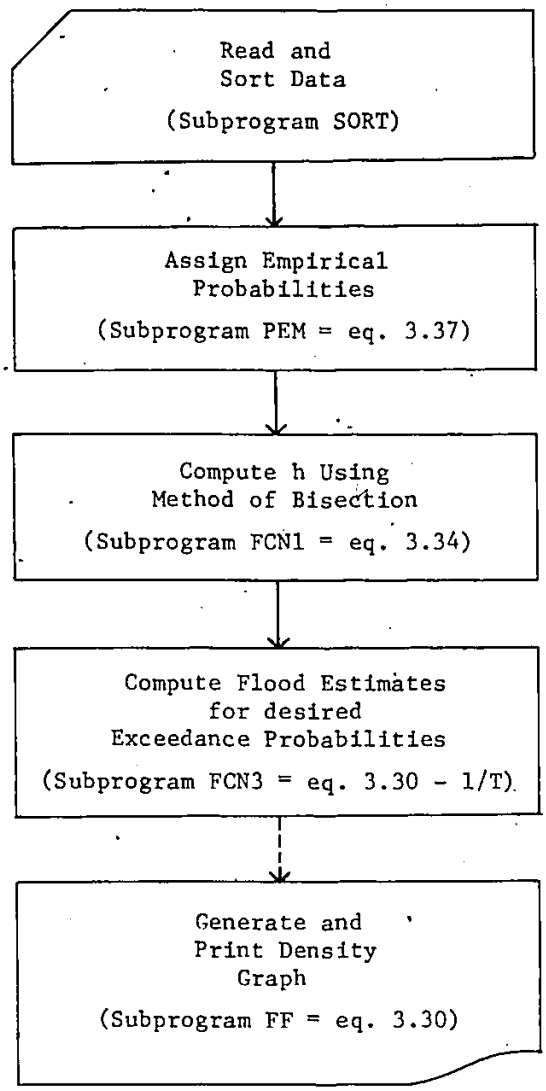


FIG. 4.7 Flowchart for RECKER
Method of Adamowski and Feluch

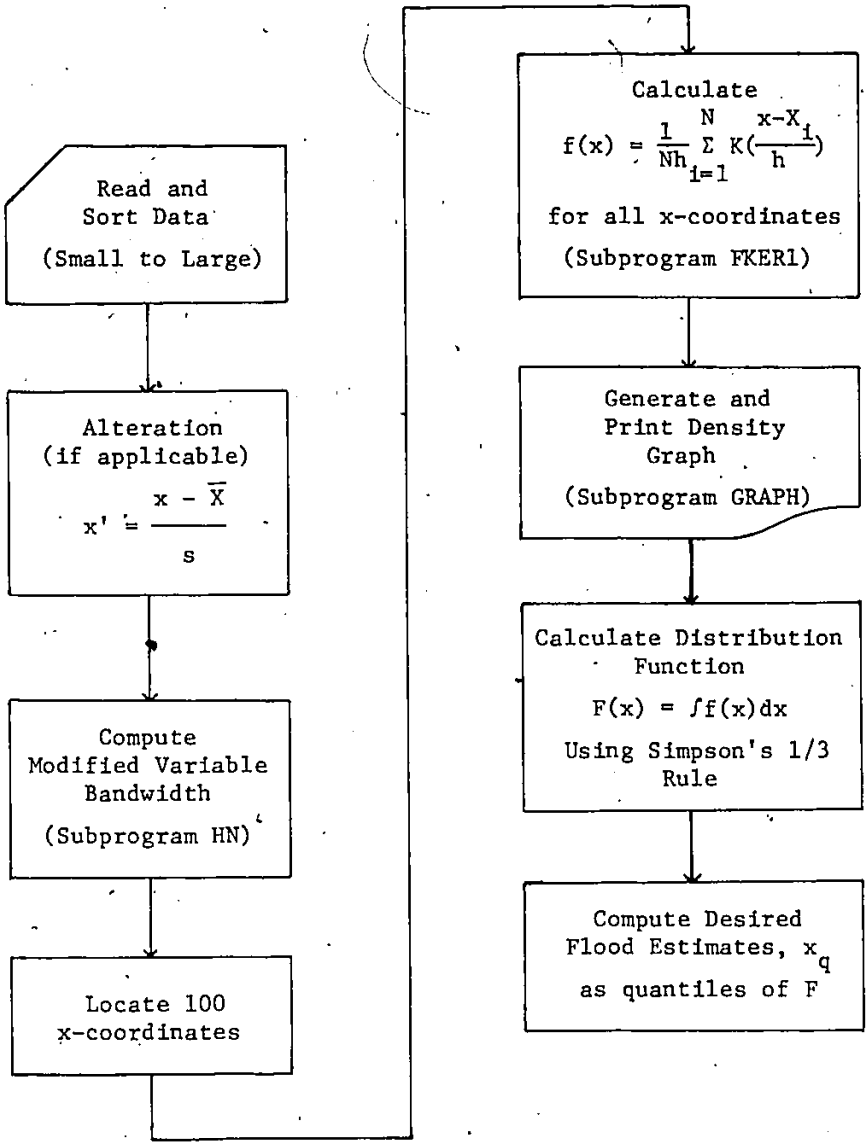


FIG. 4.8 Flowchart for MVBKER
Method of Wagner (After Liu, 1983)

factor to be calculated to at least three significant digits. All the programs are in single precision except NORKER which is in double precision as required for the STT algorithm. NORKER calls IMSL (1980) routine NDKER to evaluate the density function at the desired x-coordinates.

In evaluating the densities a minimum of 100 x-coordinates were used. To enable consistent comparison of densities for Margaree River or Warta River for the various methods, the same x-coordinates (spacings) were used. It should be emphasized that calculation of the density function is not essential in RECKER. In the other programs the density function is essential as it is integrated numerically to obtain the distribution function. In all cases numerical integration was done using Simpson's 1/3 Rule.

Experimentation was done with two additional nonparametric methods, the method of Duin and NDMPLE. It was determined that these methods do not appear workable for use in flood frequency analysis.

The method of Duin is described in Section 3.4.3. Programs were written both to maximize the likelihood function and to find the zero crossings of its first derivative. While the programs worked well for the test data sample (N=5), they failed to converge to an h value 50%

of the time with untransformed or log transformed Margaree River (N=67) and Warta River (N=140) data. Devroye et al. (1985, p. 153) states that the method of Duin works in most, but not all, situations. Hall (1984, p. 94) notes that computational problems including programming difficulty, time-consuming operation and unobtainable solutions for certain data sets can arise in maximum likelihood based methods.

The theory behind the routine NDMPLE is described in Section 3.5. Although this routine is widely used now that it is part of the IMSL library, it requires the user to specify the penalty weighting factor α, the interval of support for the estimate (B1, B2) and the number of mesh nodes MMAX. Unlike the histogram which becomes too rough with too many intervals, NDMPLE approaches a smooth limit as MMAX approaches infinity.

Over 100 NDMPLE runs were done using different combinations of α, B1, B2, and MMAX for the Margaree River flood data. Attempts were made to choose MMAX based on the number of observations in the sample. Plotting formula were used to help select B1, B2. However the resulting flood estimates didn't follow any pattern, and seemed to depend heavily upon the discretization. The most promising results were when the mean and standard deviation of the estimate matched those calculated from the sample.

Another problem is the recommendation that when using NDMPLE, extreme observations be truncated. This may be fine for large samples of items other than floods. However hydrologists cannot simply truncate extreme observations, especially high ones. It was finally decided that too many variables must be subjectively selected in NDMPLE to make rational use of it in flood frequency analysis.

4.3 MONTE-CARLO SIMULATION DESIGN

Flood magnitude values for the parametric and nonparametric analysis methods can be compared with each other. However, as the true flood magnitudes are not known when using real data it is not possible to determine which estimates are the correct ones. For this reason a simulation study was designed.

The Wakeby distribution is well suited for Monte Carlo studies because it can generate flows in the pattern of any of the more traditional distributions usually employed in hydrology. However in this study synthetic data from an LPIII parent was used. The data was available in already generated form from Environment Canada, and had been used by them to assess the various parametric distributions in their new Consolidated Frequency Analysis Program (CFA1).

The technique used to generate the LPIII data is described in Nozdryn-Plotnicki and Watt (1979). As the CDF

of the gamma distribution does not exist in explicit form, use was made of the Wilson-Hilferty transformation, eq. 2.12. The random variates were generated by the power residue method on a CDC computer using IBM routine RANF. To minimize the effect of the starting value, five samples were generated and discarded with the sixth being kept for use.

The values of the LPIII parameters used in the generation are given in Table 4.4, and represent typical values for Canadian rivers. Each sample had 50,000 values, and statistics for the samples are presented in Table 4.5.

TABLE 4.4

LPIII Parameters Used in Generation

	a	b	c
Data 1	-.06	25	7
Data 2	-.06944	64	10

TABLE 4.5

Statistics of the Generated Data

	Mean	STDV	C.S
Data 1			
Untransformed	255.546	73.829	.444
LP3 Transform	5.500	.300	-.402
Data 2			
Untransformed	299.803	167.979	1.467
LP3 Transform	5.552	.557	-.249

In calculating population flood values from eq. 2.12, the substitution t=-t must be used with negative skewness.

The Monte Carlo experiments were carried out for one underlying distribution F(x), four data-based algorithms for selecting h in the kernel method on both untransformed and log transformed data, two values of γ , two values of N, and four QT events. In summary, the experimental sample space was:

F(x)	LPIII;
h	AF, STT, FF, W on untransformed data AF, STT, FF, W on log transformed data;
γ	0.444 (Data 1), 1.467 (Data 2);
N	50, 100;
QT	Q10, Q50, Q100, Q200.

The errors computed were the bias and root mean square error (RMSE) for each QT event where

$$\begin{aligned} MSE &= (Bias)^2 + Variance \\ RMSE &= \sqrt{MSE} \end{aligned} \tag{4.1}$$

In the mean square error (MSE) the bias term describes the systematic portion of the error and variance the random portion. These errors are presented graphically in Fig. 4.9. The RMSE of a QT event can be thought of as the standard deviation about its true population mean. Ideally, it is desired to have estimates of QT events which are both precise (small variability) and accurate (unbiased).

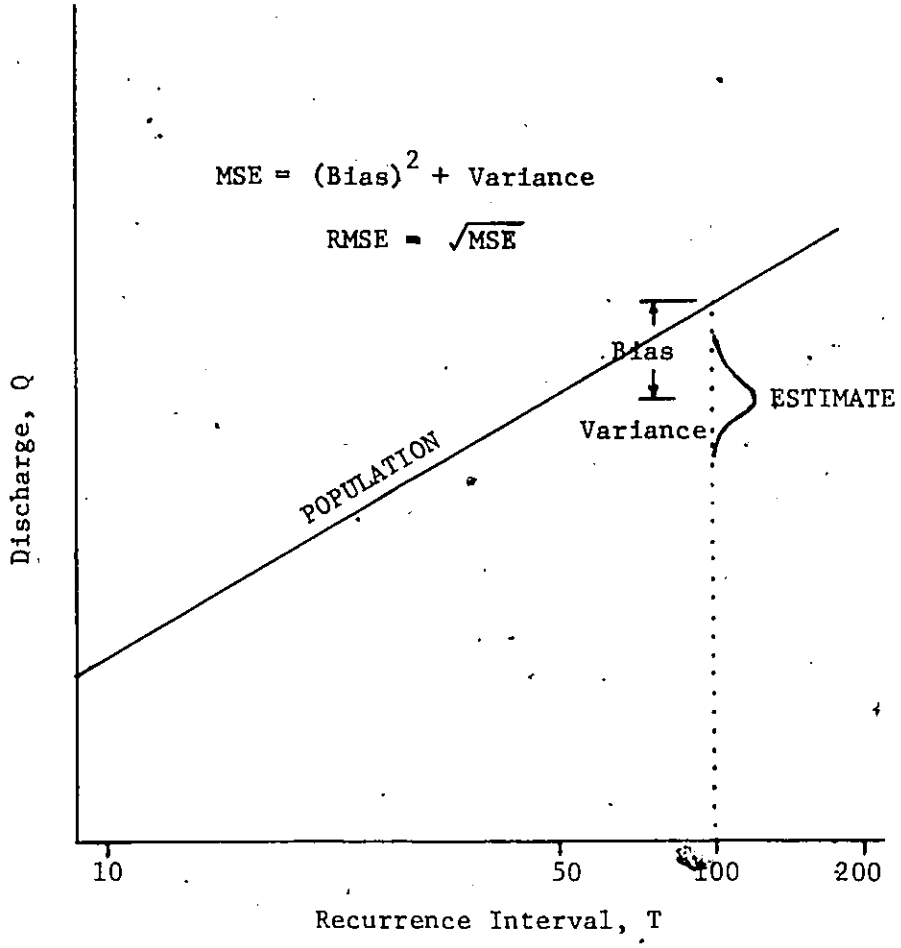


FIG. 4.9 Errors Computed in Simulation

Chapter V

RESULTS AND DISCUSSION

5.1 REAL DATA

Flood frequency analysis results for the two sets of real data are summarized in Tables 5.1 and 5.2. Flood magnitude values are given for the 10, 50, 100 and 200 year return periods for both the nonparametric analysis methods considered and four commonly used parametric distributions. Nonparametric analysis was done on both X, the original data and on LNX, the log transformed data. Natural logs were used. The kernel estimator requires only one parameter, the smoothing factor h, which is given in Tables 5.1 and 5.2 for the various data-based algorithms of Section 3.4.

Due to the sparsity of data in the tails with the sample sizes considered, it was not justified to determine flood magnitude values beyond the 200 year return period. From the theory and examples in Chapter III, it is clear that the kernel method cannot be used to estimate a density in an area where there are no observations. Using the Cunnane plotting formula, the largest observation in the Northeast Margaree River at Margaree Valley data set is assigned a return period of 112.0 years. Similarly, for the Warta River at Poznan station, the largest observation is assigned a return period of 233.7 years. Details of these

TABLE 5.1

Flood Estimates for Northeast Margaree River at Margaree Valley (N = 67, Record from 1917-1983)

	Nonparametric Analysis								Parametric Analysis			
	Untransformed Data				Log Transformed Data				LN3	GEV	LP3	
	AF	STT	FF	W	AF	STT	FF	W	ML	ML	ML	Wak.
Event:												
Q10	247	248	249	250	247	250	251	252	247	246	246	249
Q50	345	344	343	341	340	339	339	339	323	323	322	332
Q100	353	354	358	359	358	362	367	370	355	357	355	366
Q200	359	361	367	370	370	378	387	392	387	391	389	399
Parameter:												
h	12.65	9.51	14.49	17.20	0.0832	0.0797	0.0837	0.0935				

TABLE 5.2

Flood Estimates for Polish River Warta at Poznan (N = 140, Record from 1826-1965)

	Nonparametric Analysis								Parametric Analysis			
	Untransformed Data				Log Transformed Data				LN3	GEV	LP3	
	AF	STT	FF	W	AF	STT	FF	W	ML	ML	ML	Wak.
Event:												
Q10	807	905	901	907	861	913	899	894	834	819	822	836
Q50	1489	1485	1485	1486	1468	1575	1505	1490	1490	1620	1510	1420
Q100	1531	1537	1560	1557	1566	1802	1698	1649	1830	2130	1890	1700
Q200	1560	1570	1611	1606	1617	2013	1855	1771	2220	2780	2340	2000
Parameter:												
h	31.64	38.35	72.65	65.42	0.1042	0.2493	0.1667	0.1200				

calculations can be found in columns 5, 6 and 7 of Tables A.1 and A.6.

It has been observed (Fryer, 1977) that the kernel method performs best for symmetric f(.). However, as flood data is not symmetrical but invariably skewed, analysis on log transformed data has been done in addition to original data analysis. That the log transformation is a transformation towards near-symmetry can be demonstrated by comparing the C.V.'s for the data sets in Table 4.1 or alternatively the coefficients of variation of the inter-observation distances:

Inter-observation Distance	C.V., Original Data	C.V., Log Transform
Margaree River	1.62	1.35
Warta River	2.33	1.47

Of course other transformations are possible including those discussed by Singh et al. (1983). However an in-depth consideration of transformations is beyond the scope of this thesis.

It should be noted that with flood data in hydrology, even the effect of data alteration by standardization is not completely known. While there are theoretical reasons with the kernel estimate to "uniformize" the data on a closed [0,1] interval (Devroye et al., 1985, p. 121), the choice of

a suitable transform when one is mainly interested in the upper tail is not easy. Devroye et al. (1985, pgs. 80, 245) note that the isosceles triangular density is the easiest to estimate by the kernel method, and this forms the basis of their transformed kernel estimate.

In parametric analysis, the choice of assumed distribution matters only when different choices lead to different designs. It can be seen from the QT estimates in Table 5.1 that for the Margaree River, any method of analysis would result in a similar design. An indication of the relative variability of the nonparametric versus parametric analysis methods can be given by the ratios of the span between the highest and lowest estimate values. This is given in Tables 5.3 and 5.4 for the two sets of

TABLE 5.3

Variability of Nonparametric versus Parametric Flood Estimates for Margaree River (N=67)

	Estimate Range			Ratio	
Parametric	Nonparametric Untransformed	Nonparametric Log Transform	(2)/(1)	(3)/(1)	
(1)	(2)	(3)	(4)	(5)	
Event:					
Q10	3	3	5	1.00	1.67
Q50	10	4	1	0.40	0.10
Q100	11	6	12	0.55	1.09
Q200	12	11	22	0.92	1.83

TABLE 5.4

Variability of Nonparametric versus Parametric Flood Estimates for Polish River Warta At Poznan (N=140)

	Estimate Range		Ratios	
Parametric	Nonparametric Untransformed	Nonparametric Log Transform	(2)/(1)	(3)/(1)
	(1)	(2)	(3)	(4)	(5)
Event:					
Q10	17	100	52	5.88	3.06
Q50	200	4	107	0.02	0.54
Q100	430	29	236	0.07	0.55
Q200	780	51	396	0.07	0.51

data. For the Margaree River, the variability of the nonparametric and parametric methods are comparable. For the Warta River, except for the Q10 event, the variability of the nonparametric estimates is approximately 50 to 60% that of the parametric estimates. Note that in Tables 5.3 and 5.4, excepting the Q10 events, the estimate spans increase for the higher events as would be expected.

The nonparametric densities and distributions for the two sets of data are presented in Figs. 5.1 through 5.12.

Fig 5.1 shows the nonparametric densities, for Margaree River for the untransformed data. Fig. 5.1a is quite rough. There are a large number of peaks (14) in the density and bumps occur at outlying observations. However, Fig. 5.1d is very smooth. The number of peaks in a density is indicative

of its smoothness. These are given in Tables 5.5 and 5.6 for the two data sets. In general, the data-based algorithms ranked from roughest to smoothest estimate are: AF, STT, FF, W. A final observation which can be made from Fig. 5.1 is that the shapes of the estimated density functions are quite similar for the methods of AF and STT, although the methods are very different. The densities estimated by the methods of FF and W are also quite similar. These similarities are also evident with the other data set, as seen in Fig. 5.7.

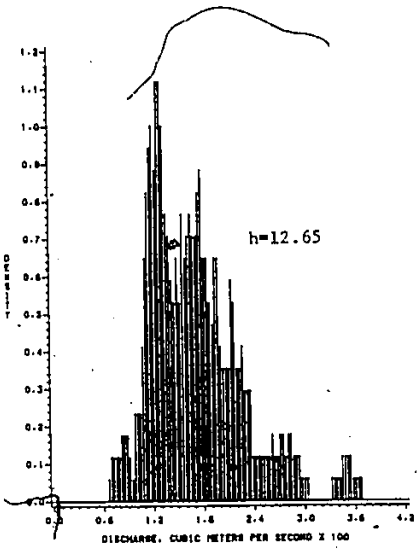
Fig. 5.2 shows the nonparametric densities for Margaree River for the log transformed data. Note that all of the densities are distinctly bimodal, and the problem of bumps at outlying observations has been resolved. Scott et al. (1981) have indicated that nonparametric techniques are particularly well suited for multimodal data. As stated in Chapter I, flood data can be multimodal as the causative factors of floods vary. Spring floods are usually caused by rapid snowmelt, heavy rainfall or both combined while fall floods may be caused by heavy rains accompanying thunderstorms or hurricanes. It was postulated that the right-hand mode in Fig. 5.2d might be due to the effect of fall floods. Means were computed from the nonparametric density without the right-hand mode and for the right-hand

mode in isolation and compared with seasonal means computed from the data as shown below:

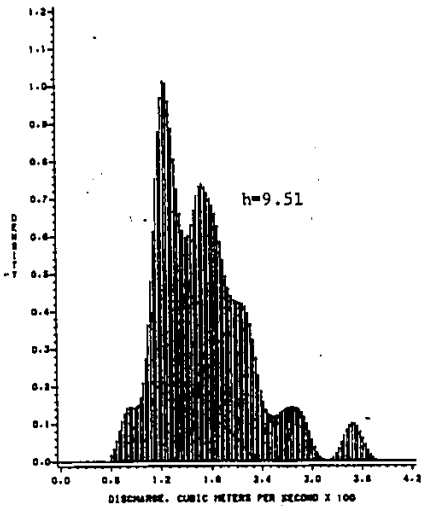
Means from Nonparametric Density	Means from Data
Right Hand Mode Only: 199.7 CMS	Fall Floods (Sept.- Dec.): Mean of 14 is 185.5 CMS
Rest of Density: 167.3 CMS	Rest of Year (Jan.- Aug.): Mean of 53 is 168.2 CMS
Overall: 172.6 CMS	Overall: Mean of 67 is 171.8 CMS

While the comparison is interesting, it must be remembered that it is for log transformed data. The log transform is a nonlinear transformation, and none of the Margaree River untransformed data densities appear bimodal. However other stations such as WSC 05QA001 English River Near Sioux Lookout exhibit multimodality with untransformed data. A final note is that in Fig. 5.2d the fine structure of the density is visible. One criterion which has been suggested for choosing the smoothing factor is when rapid fluctuations just start to appear in the density.

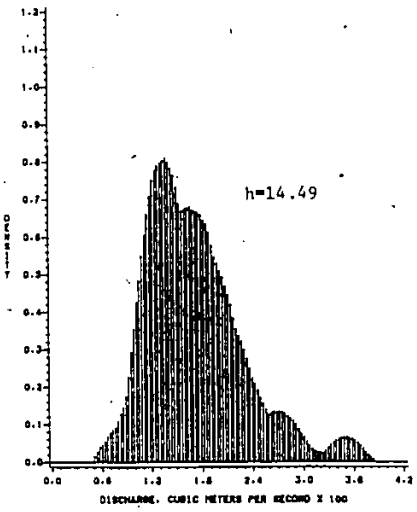
In an application such as flood frequency analysis, the distribution function plots are equally if not more important than the density function plots. The distributions corresponding to the densities in Figs. 5.1 and 5.2 are given in Figs. 5.3-5.6. The observations are represented using the empirical distribution formula given by eq. 3.37. In Fig. 5.3, while the distribution functions 'fit' the observations, the sudden 'zig' towards the second largest data point is not desirable due to potential overfitting. Klemes (1980) has



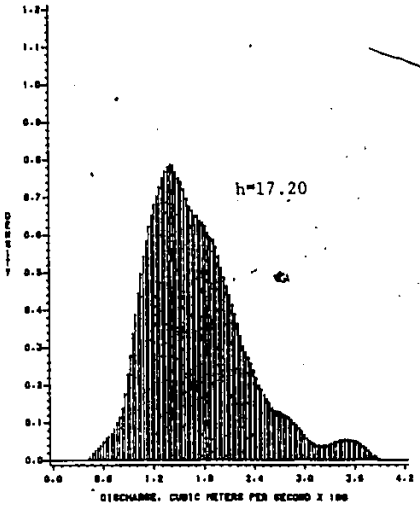
a) h by Adamowski/Feluch Procedure



b) h by Method of Scott, Tapia & Thompson

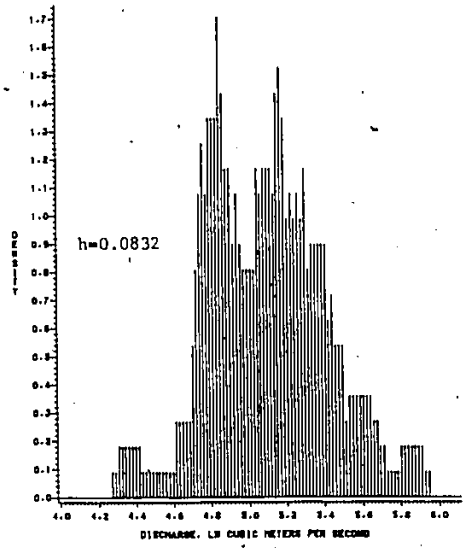


c) h by Flanagan's Formula

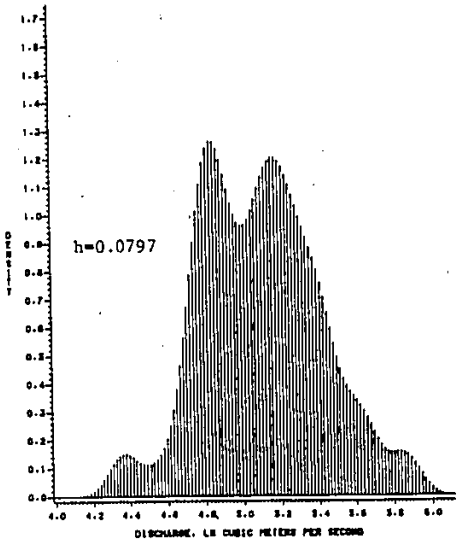


d) h by Method of Wagner

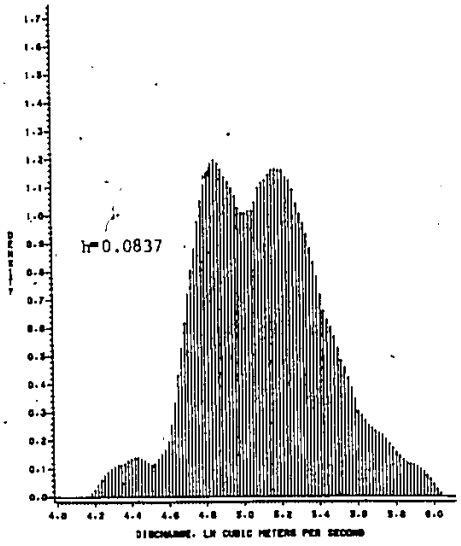
FIG. 5.1 Margaree River Density Plots (N=67, Untransformed Data)



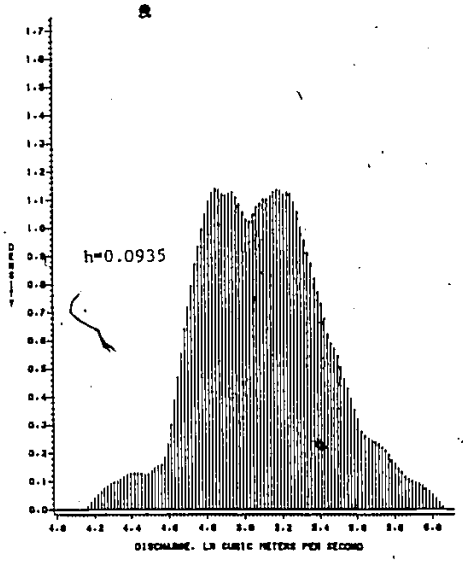
a) h by Adamowski/Feluch Procedure



b) h by Method of Scott, Tapia & Thompson

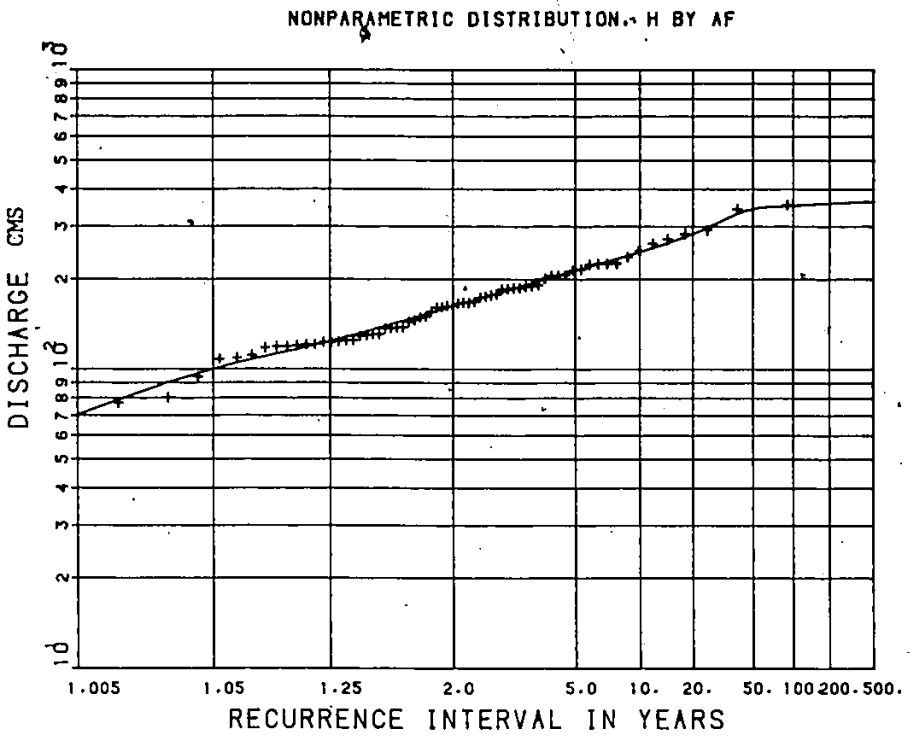


c) h by Flanagan's Formula



d) h by Method of Wagner

FIG. 5.2 Margaree River Density Plots (N=67, Log Transformed Data)



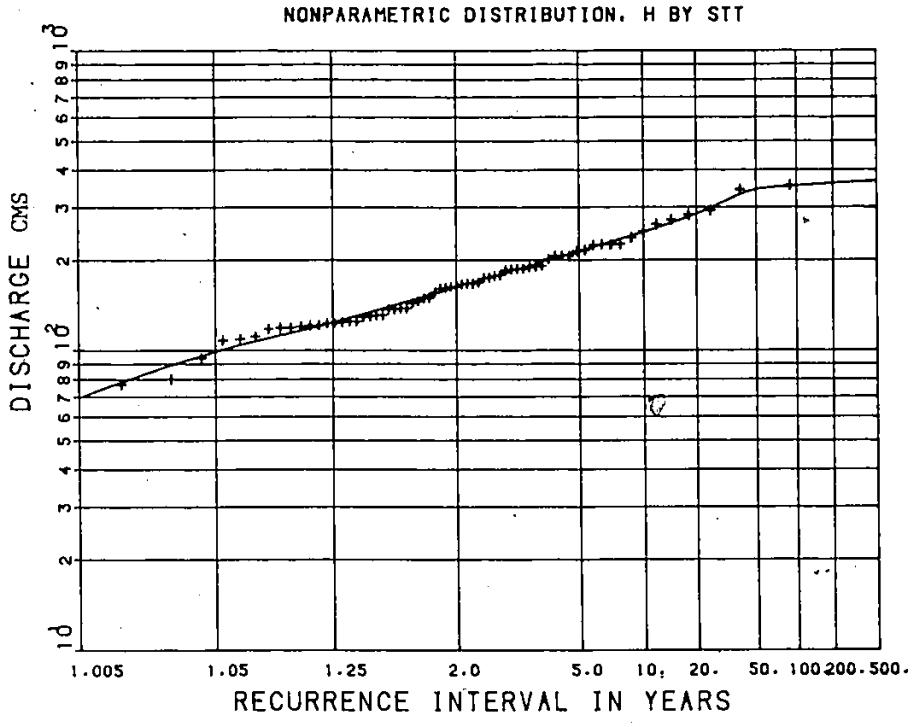
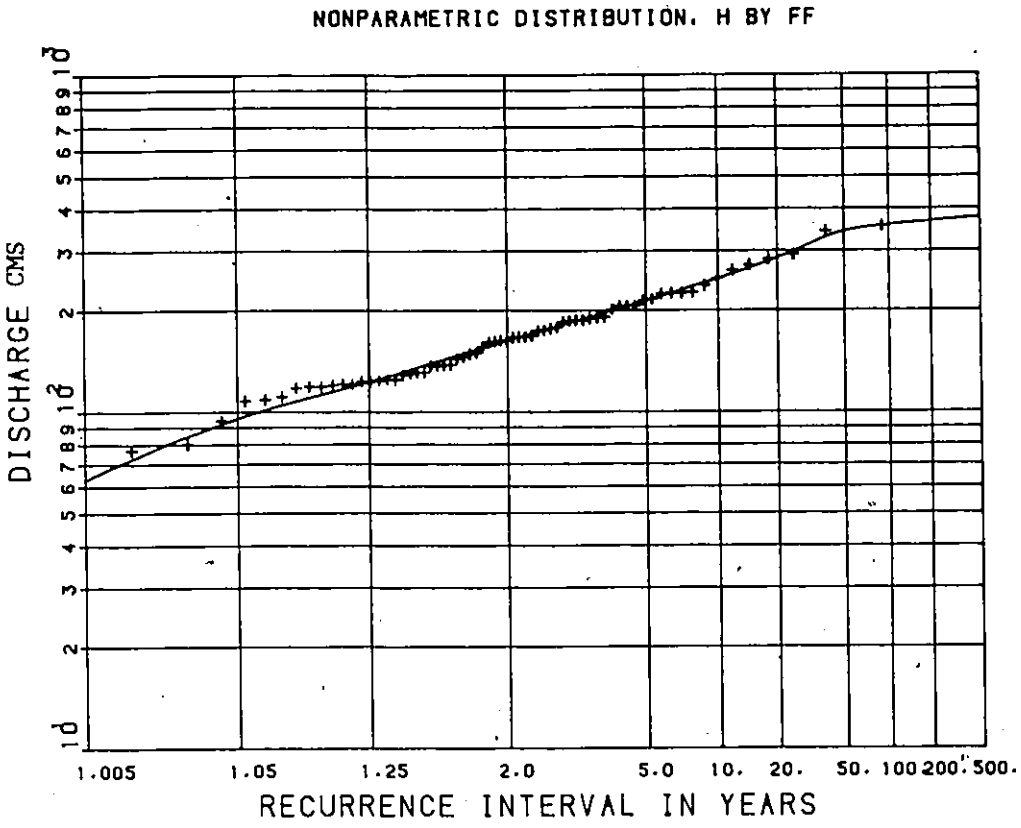


FIG. 5.3 Margaree River Distributions (Untransformed Data)



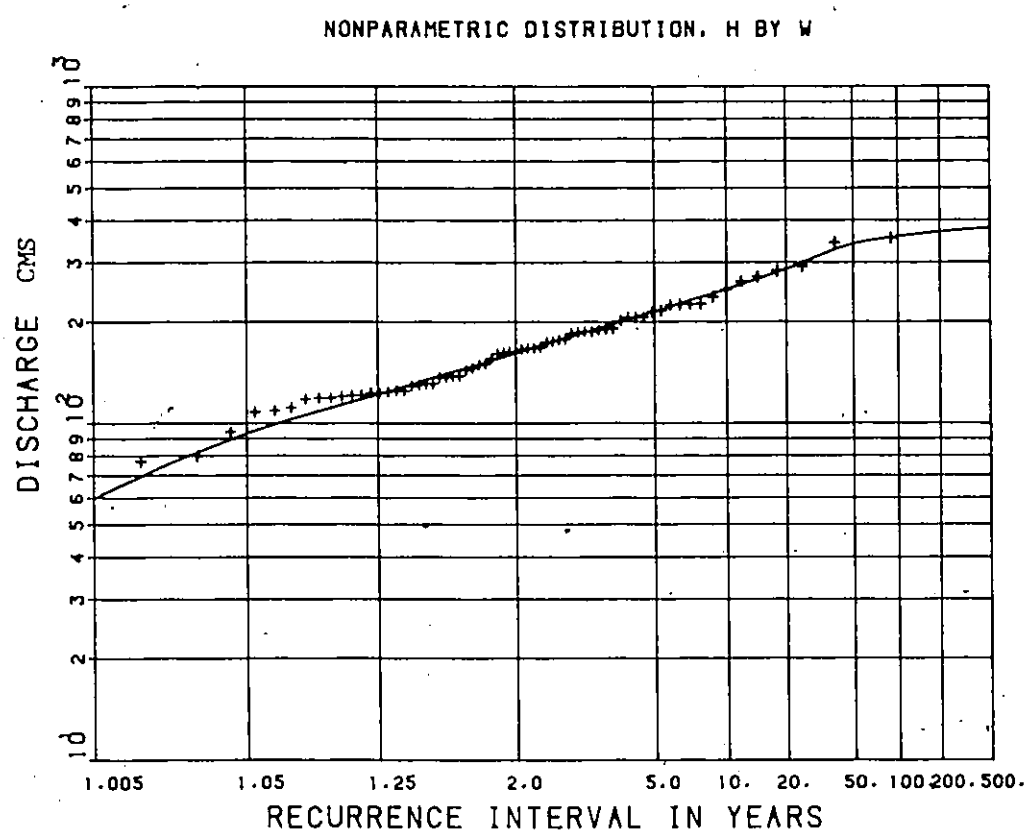
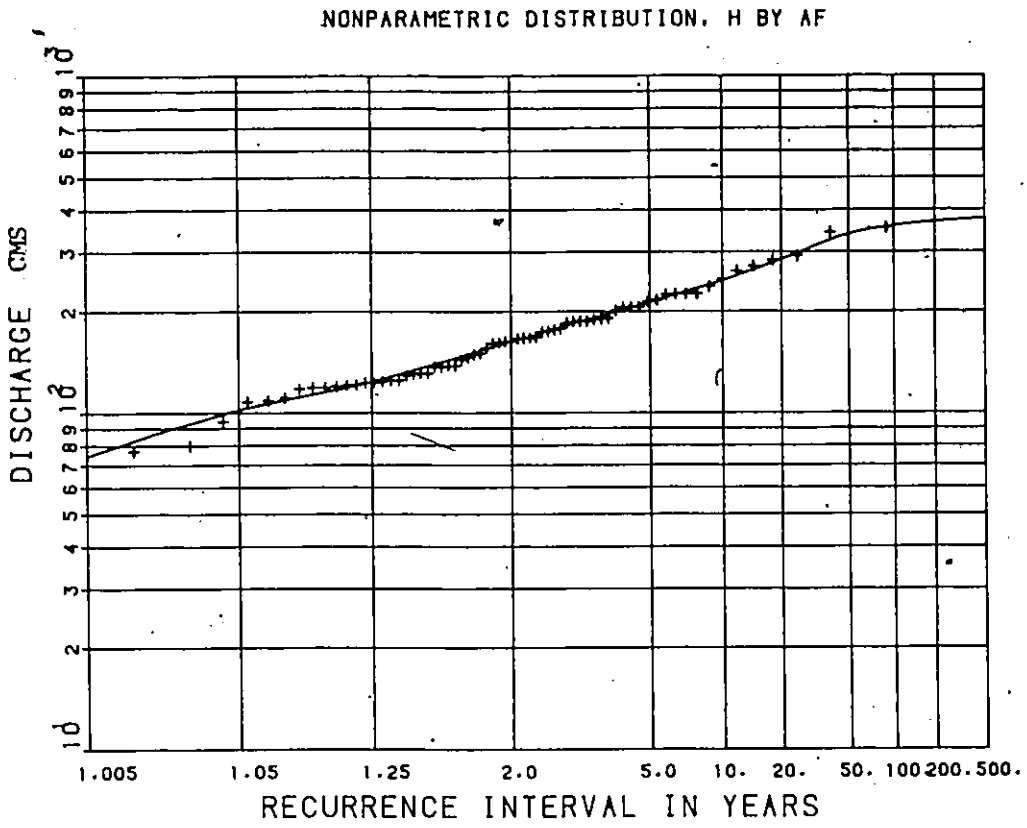


FIG. 5.4 Margaree River Distributions Continued (Untransformed Data)



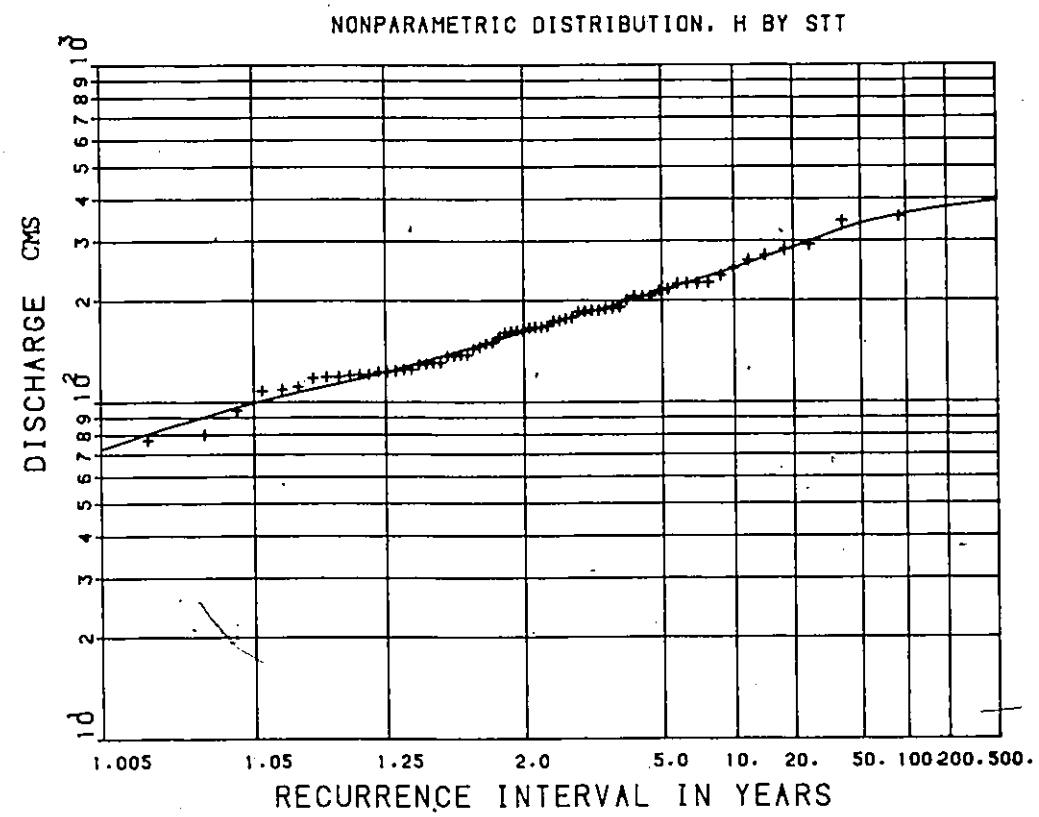
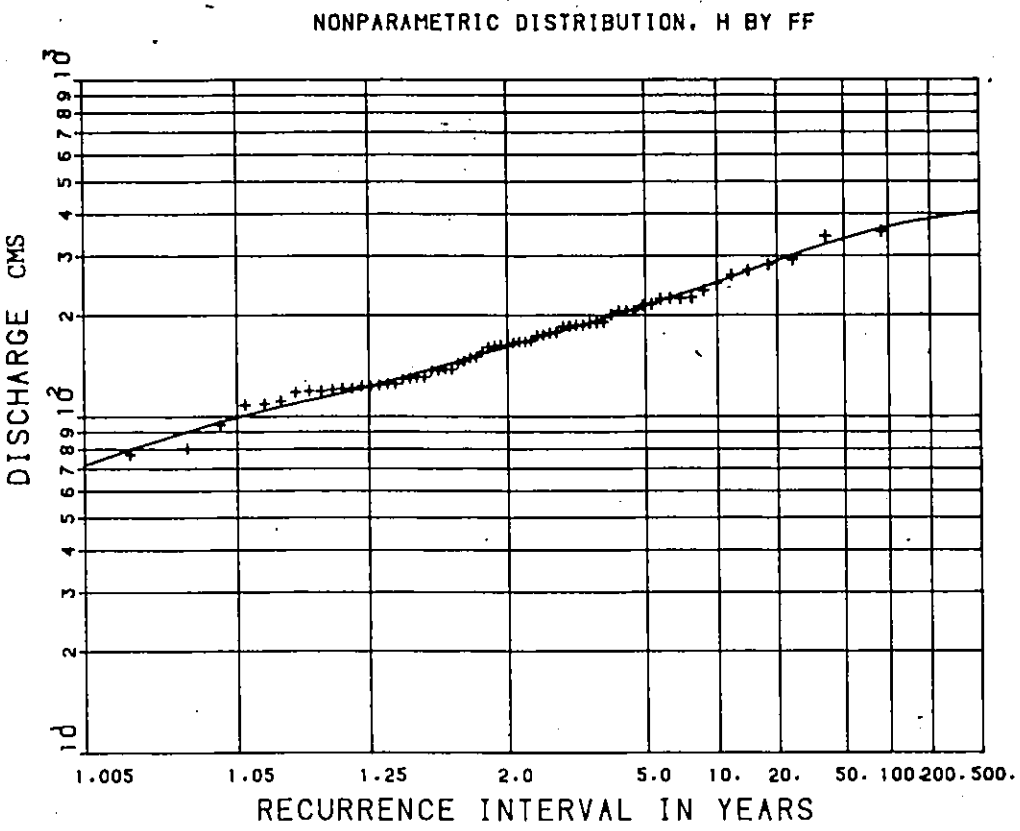


FIG. 5.5 Margaree River Distributions (Log Transformed Data)



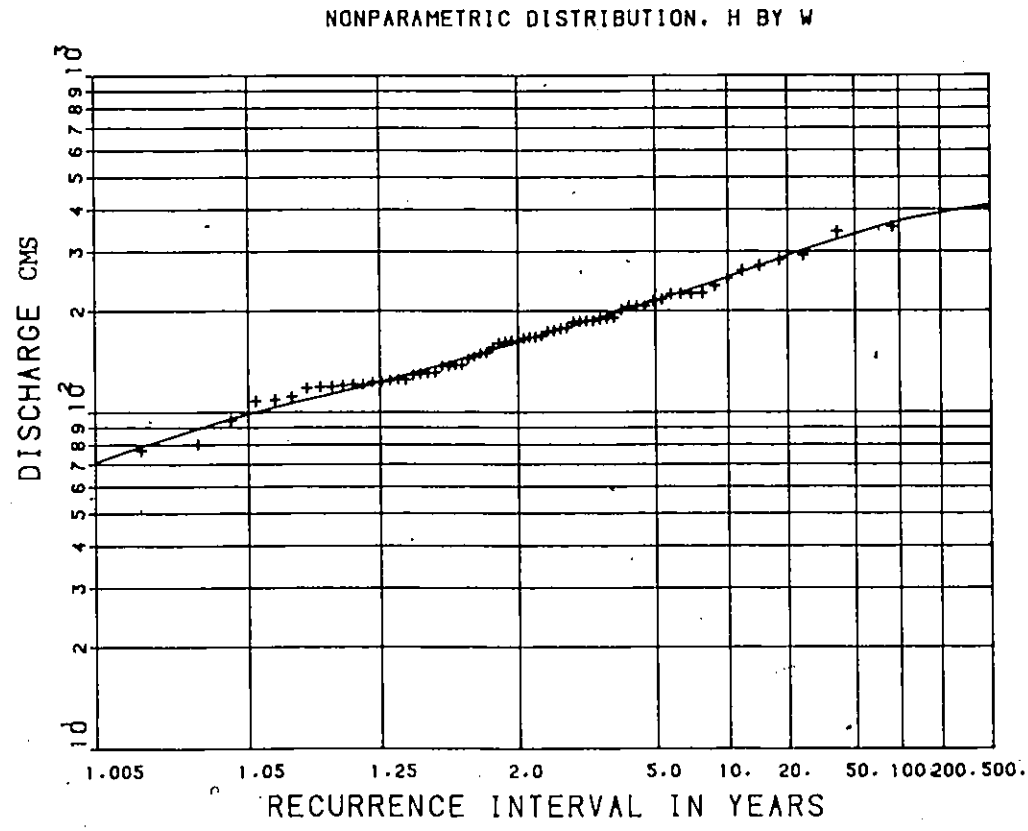
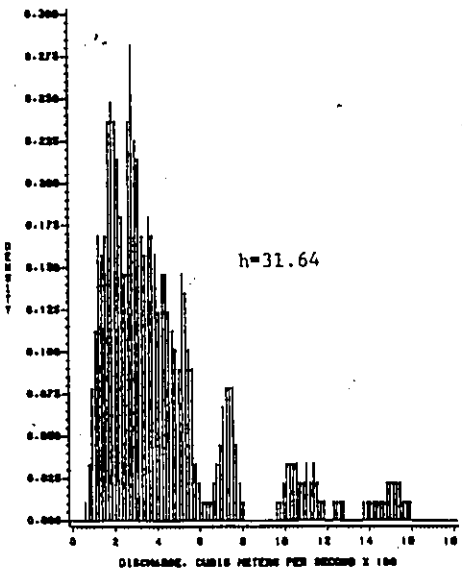
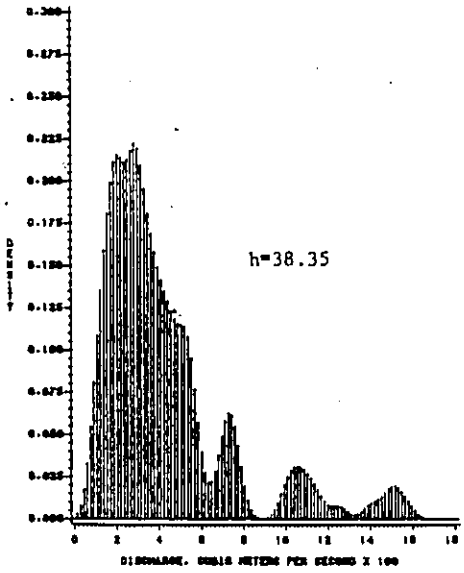


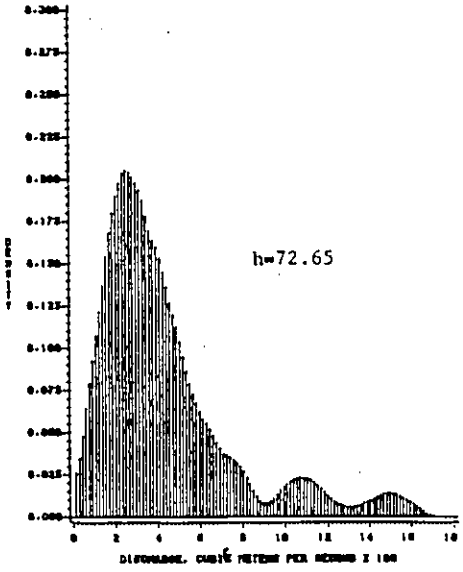
FIG. 5.6 Margaree River Distributions Continued (Log Transformed Data)



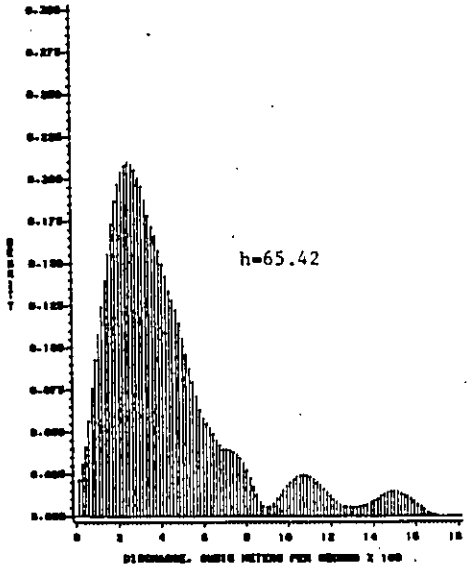
a) h by Adamowski/Feluch Procedure



b) h by Method of Scott, Tapia & Thompson

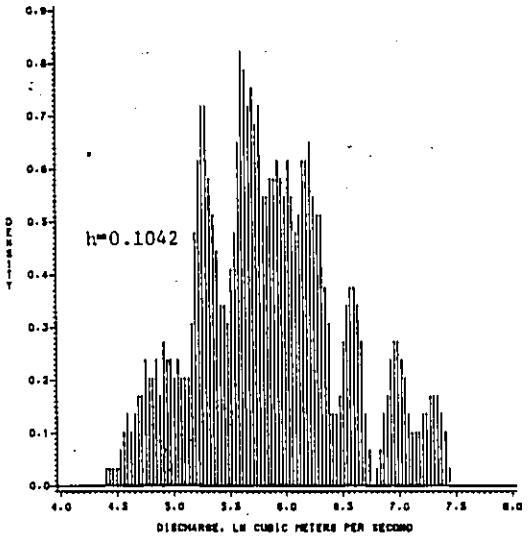


c) h by Flanagan's Formula

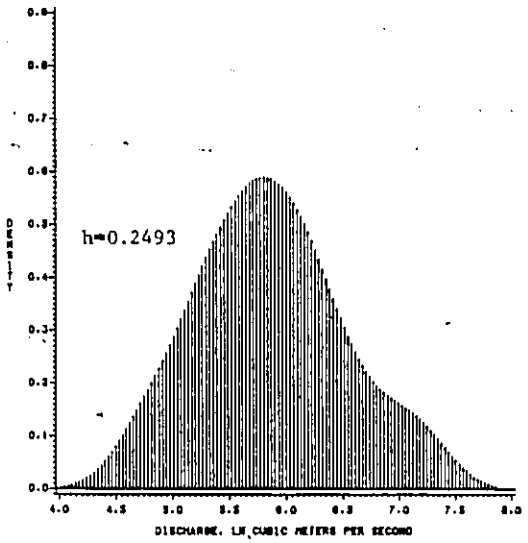


d) h by Method of Wagner

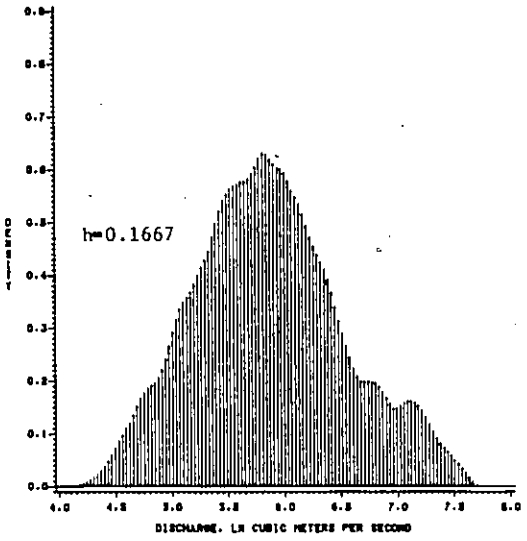
FIG. 5.7 Warta River Density Plots (N=140, Untransformed Data)



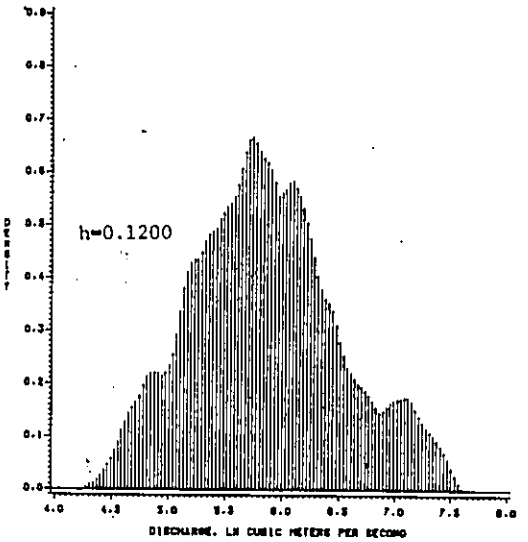
a) h by Adamowski/Feluch Procedure



b) h by Method of Scott, Tapia & Thompson

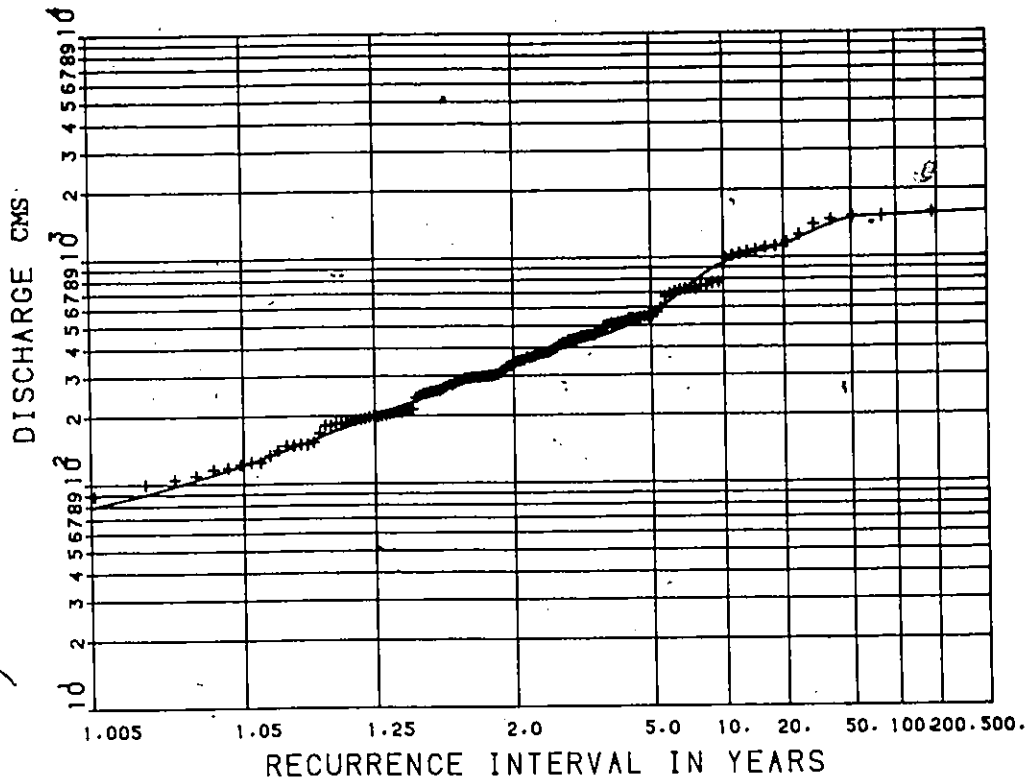


c) h by Flanagan's Formula



d) h by Method of Wagner

FIG. 5.8 Warta River Density Plots (N=140, Log Transformed Data)



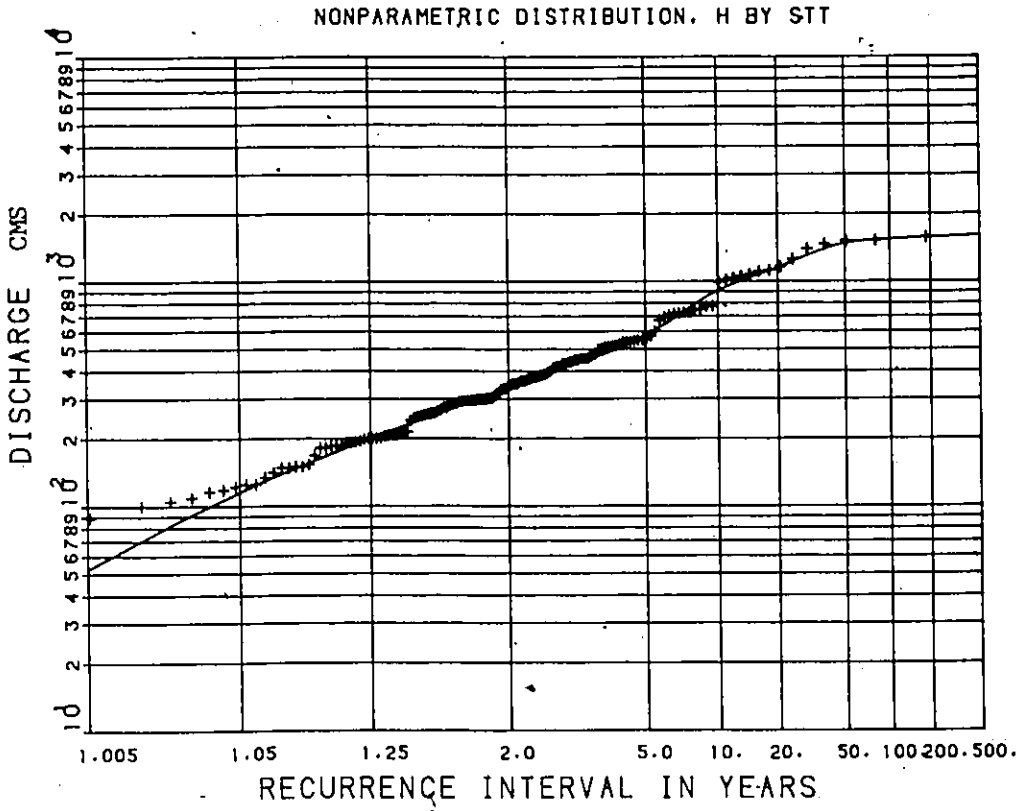
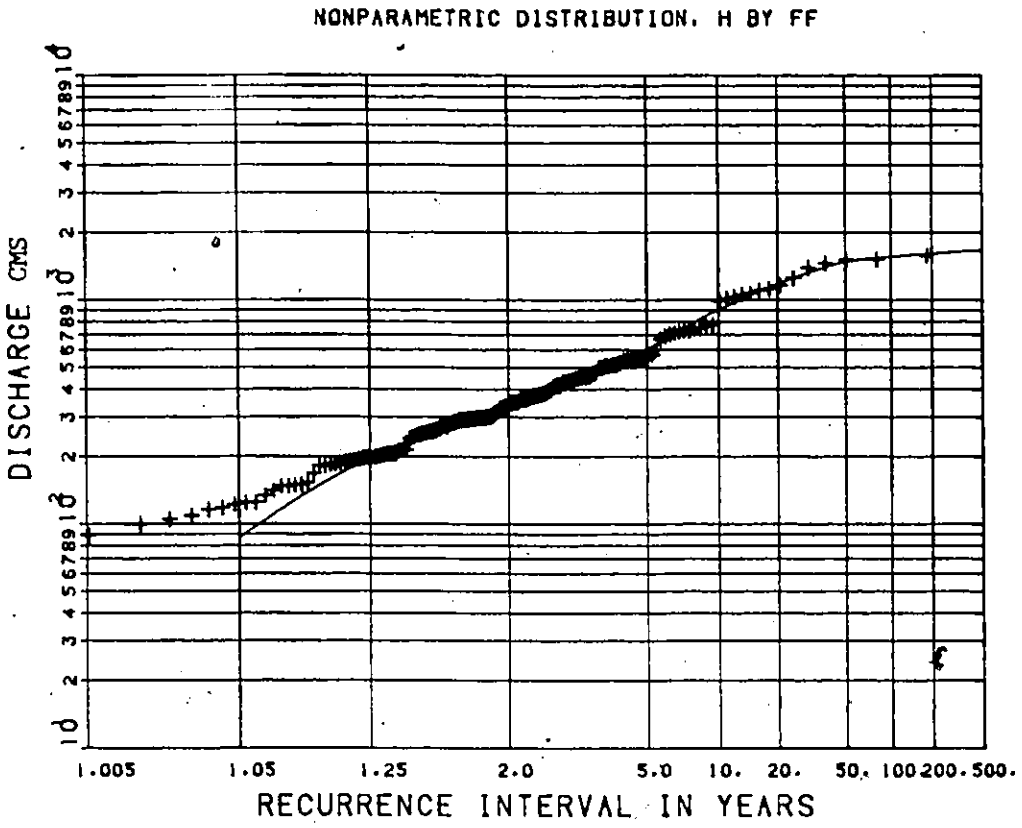


FIG. 5.9 Warta River Distributions (Untransformed Data)



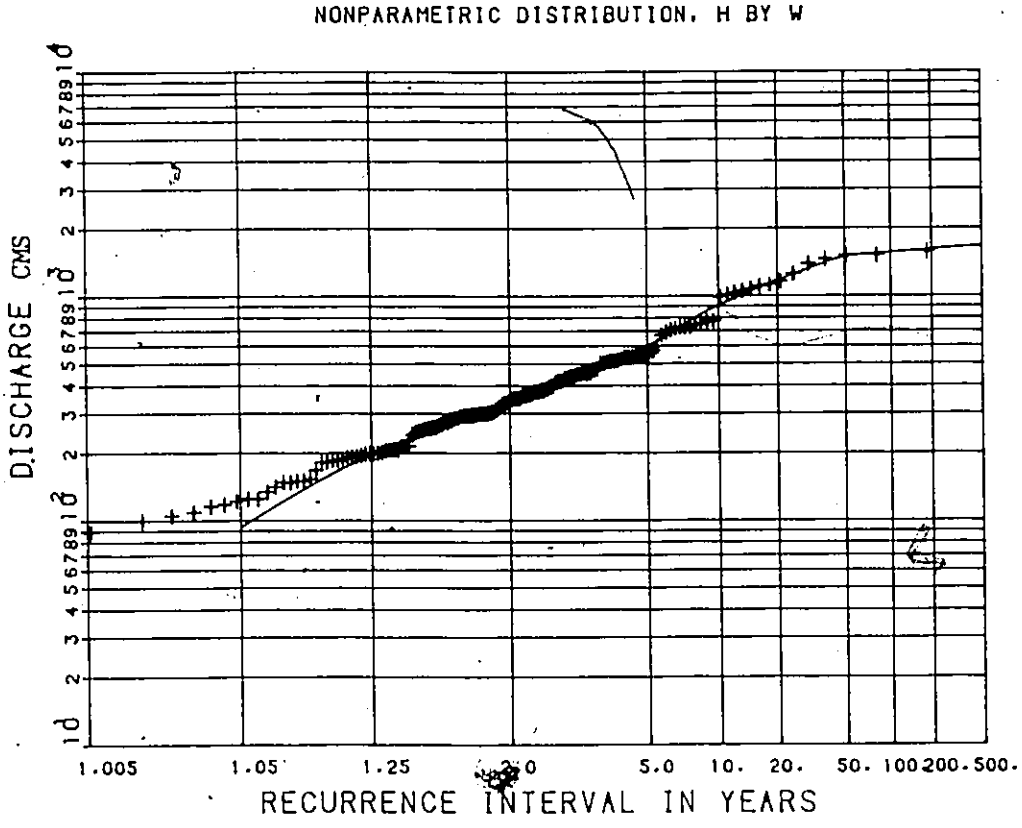
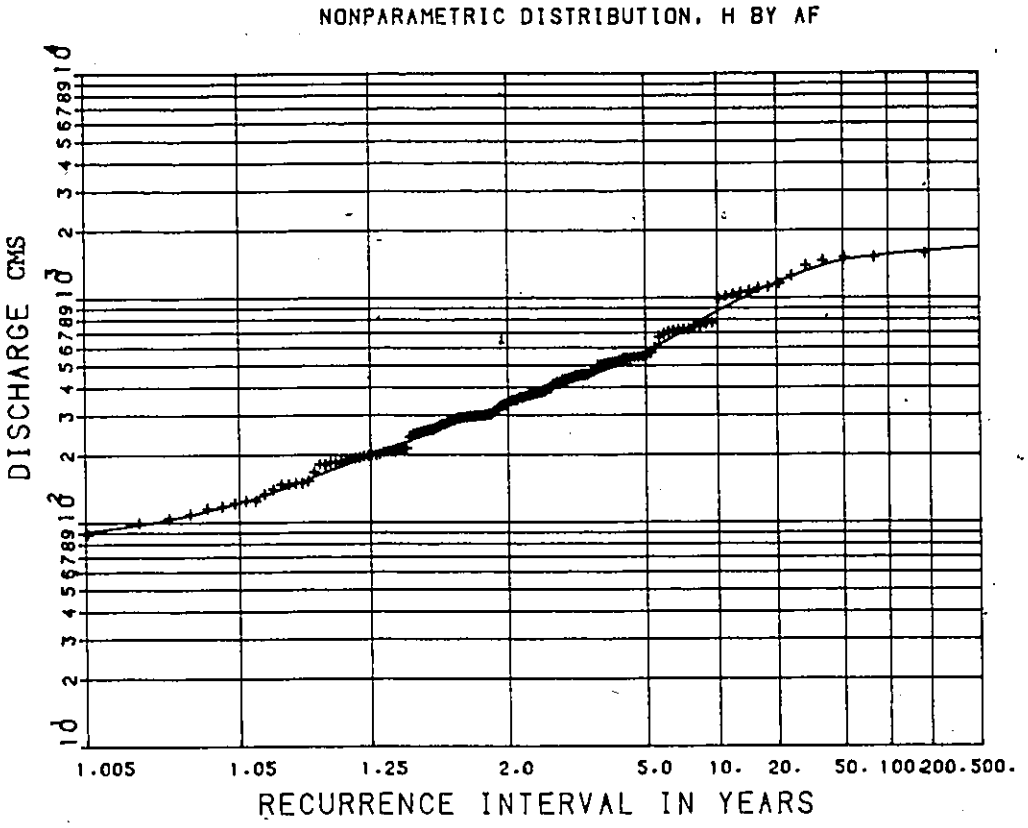


FIG. 5.10 Warta River Distributions Continued (Untransformed Data)



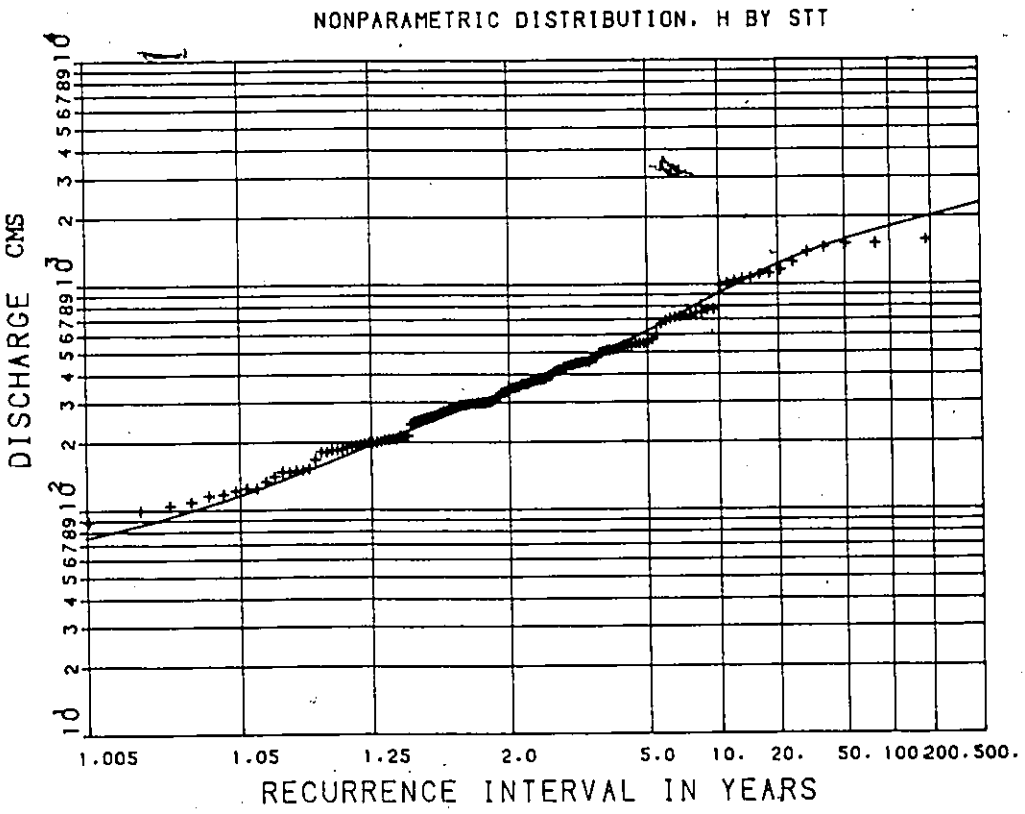
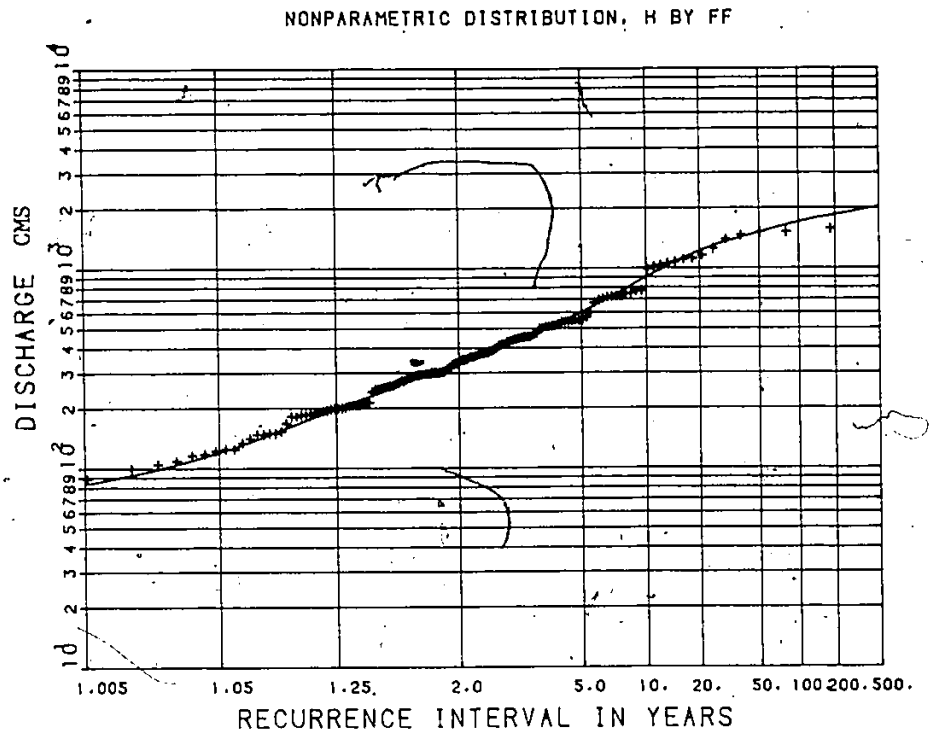


FIG. 5.11 Warta River Distributions (Log Transformed Data)



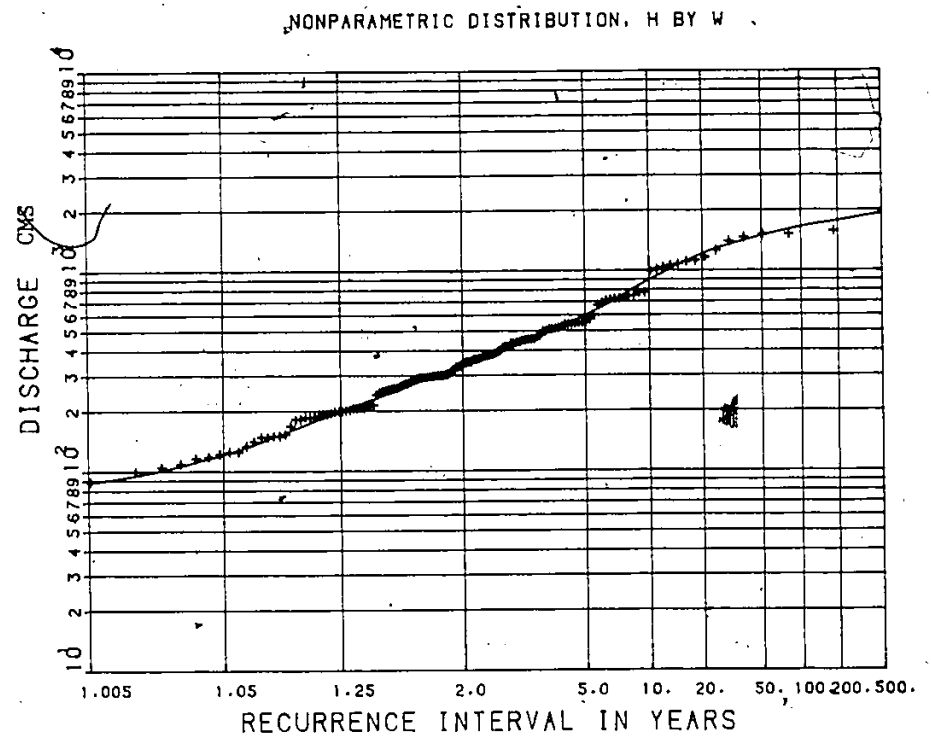


FIG. 5.12 Warta River Distributions Continued (Log Transformed Data)

warned against overfitting, i.e. mistaking noise for information. Note the effect of the log transform in Figs. 5.5 and 5.6. Here the distribution functions are no longer almost horizontal after the last observation as they are in Figs. 5.3 and 5.4.

Fig. 5.7 shows the nonparametric densities for Warta River for the untransformed data. With this set of data the right-hand tail is longer and more complex, as could be anticipated from the histograms in Chapter IV. In Fig. 5.7c and d there is a problem in that the left-hand side of the densities become slightly negative. As negative floods have no physical basis and also to allow uniform scales with the other plots on the page, these densities have been truncated at zero. The effect of this anomaly can be seen on the left-hand side of the distribution functions in Fig. 5.10. In any case however, estimates of floods with return periods less than 10 years should be computed using partial duration series, not annual flood series.

In Fig. 5.9 and 5.10 for the higher observations the distribution function matches the data points as though by interpolation. Even though the true plotting position for the observations is not known, it can be seen that the kernel estimator allows the data to "speak for itself".

Fig. 5.8 shows the nonparametric densities for Warta River for the log transformed data. Contrast Fig. 5.8a with Fig. 5.8b. In Fig. 5.8a the density is too rough while in Fig. 5.8b it is oversmoothed. Of the four density plots on the page, Fig. 5.8a has the smallest smoothing factor h and Fig. 5.8b the largest. These density estimators should be compared with respect to their distribution functions as well as the density function. In Fig. 5.11 for h by STT it is seen that there is a negative bias in the left-hand tail of the distribution function and a positive bias in the right-hand one. In Fig. 5.11 for h by AF the distribution function 'fits' the observations extremely well. This is hardly surprising as the same plotting formula is used in the formulation of h by AF as is used in representing the observations in the figure. More 'reasonable' densities for the Warta River are given in Fig. 5.8c and d, with their corresponding distributions being given in Fig. 5.12. As the true densities are unknown, this assessment is subjective; however the fine structure of the densities is apparent, and the appearance of the distributions pleasing.

To characterize the nonparametric densities, the following statistics were parametrically computed using moments:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} f_i X_i \quad \text{where} \quad n = \sum_{i=1}^{k} f_i \tag{5.1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i X_i^2 - n\bar{x}^2}{n - 1}} \qquad (5.2)$$

Results for the two data sets are given in Tables 5.5 and 5.6. They can be compared with the equivalent statistics calculated from the data, in Table 4.1. The estimate and data values agree very closely for the Margaree River, but less so for the Warta River. The difficulty with the Warta River might be anticipated from its higher coefficient of skewness. The estimate standard deviation decreases for all the methods except the method of STT with the log transformation. The method with the closest match between estimate and data for both the mean and standard deviation is AF log.

Figure 5.13 compares typical parametric and nonparametric distribution plots for the Warta River. It can be seen that out of the total sample of 140, only 14 observations are in the portion of the distributions to the right of the 10 year recurrence interval. In the parametric model, the parameters are largely determined by the central portion of the distribution, where most of the observations lie. However in the nonparametric model, more weight is given to the right-hand tail, ie. the 14 observations to the right of the 10 year recurrence interval. The nonparametric model seems more suitable for predictions in this region than the parametric one.

TABLE 5.5

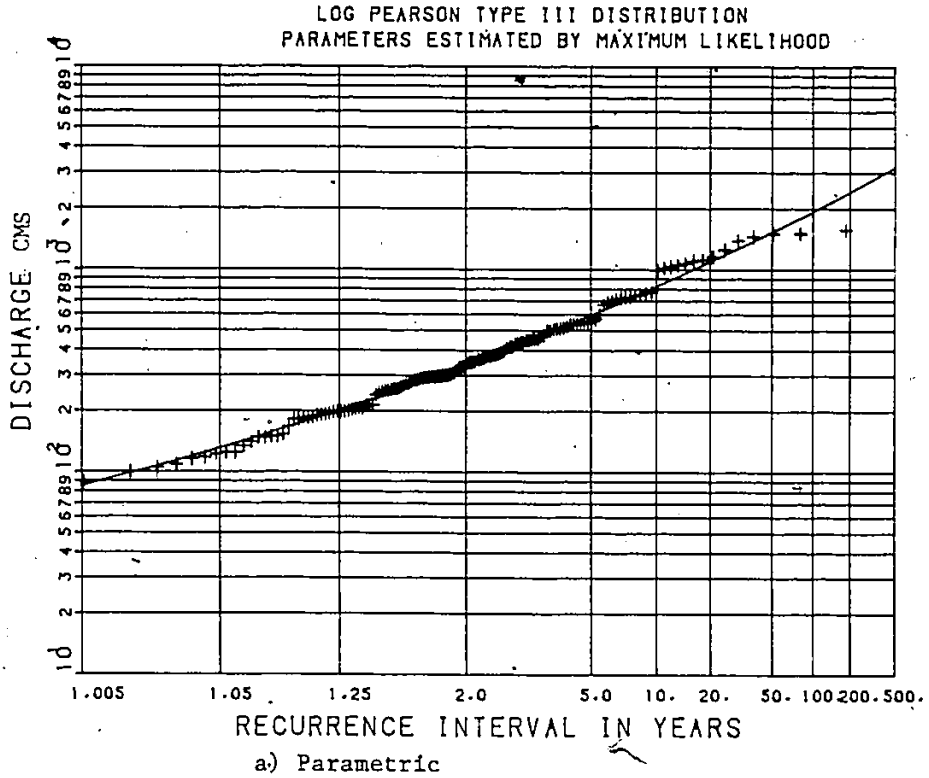
Nonparametric Density Estimate Statistics for Northeast Margaree River at Margaree Valley
(N=67, Record from 1917-1983)

Method	Number of Peaks	Estimate Mean X	Estimate Standard Deviation S
Untransformed:			
AF	14	172.19	58.39
STT	5	171.87	58.59
FF	4	171.84	59.72
W	2	171.84	60.46
Log Transform:			
AF	14	171.95	57.82
STT	4	172.36	59.41
FF	3	172.43	59.68
W	5	172.58	60.21

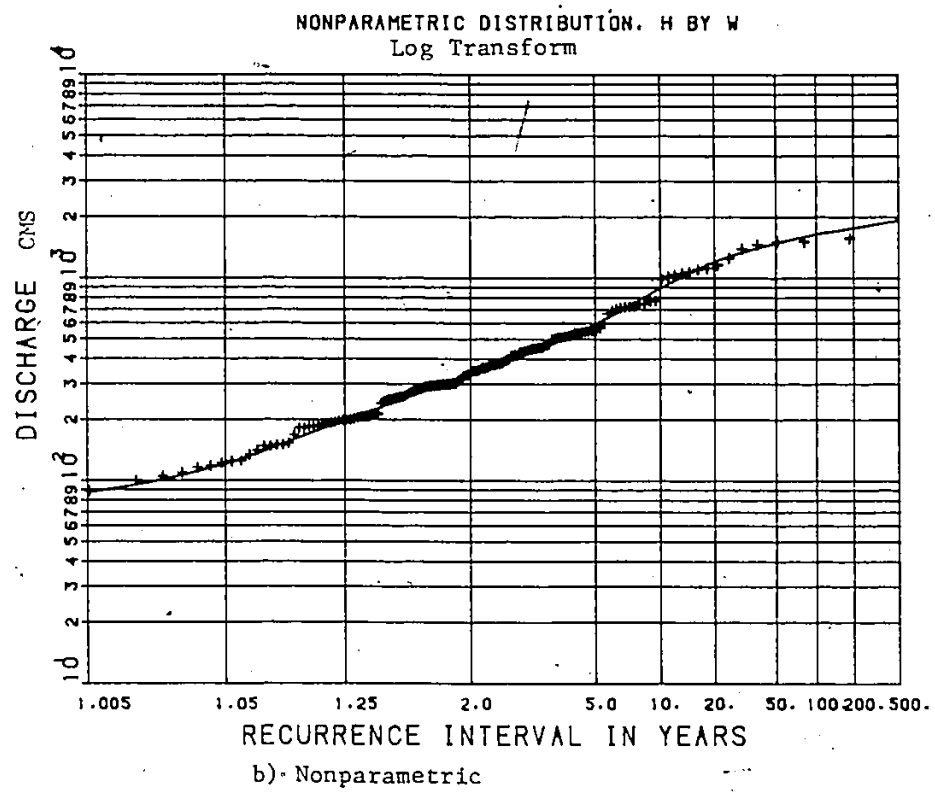
TABLE 5.6

Nonparametric Density Estimate Statistics for Polish River Warta at Poznan.
(N=140, Record from 1826-1965)

Method	Number of Peaks	Estimate Mean X	Estimate Standard Deviation S
Untransformed:			
AF	12	436.39	348.02
SST	6	436.57	349.62
FF	3	436.48	356.39
W	3	436.45	354.39
Log Transform:			
AF	15	436.59	325.77
STT	1	449.23	359.61
FF	3	442.56	341.10
W	4	439.62	332.79



a) Parametric



b) Nonparametric

FIG. 5.13 Warta River Parametric and Nonparametric Distributions

5.2 SYNTHETIC DATA

The complete simulation results for the experimental sample space and errors described in Section 4.3 are given in Tables 5.7 through 5.10. The true population values are now known and are shown. As with the real data, estimates were computed for the Q10, Q50, Q100 and Q200 events.

In the simulation results, some basic differences can be seen between the method of AF and the other three methods. These differences can be attributed to the fact that AF is really formulated based on a distribution function criteria, while the other methods are based on probability density function estimation.

First consider the bias results, untransformed data. For Data 1 the results are very unbiased, generally less than 2%. For Data 2 with sample size 100 this is also true, except for the Q100 event estimated by the methods of STT, FF, W. For Data 2 with sample size 50 the bias is less than 2% for only half the events and methods. The exceptions this time are the Q50 event estimated by the methods of STT, FF, W and most of the cases in which the bias is negative. Note that Data 2, sample size 50 is the most difficult to estimate as it has the largest coefficient of skewness with the least amount of data. With the kernel estimate, the increasing negative bias is evidence of the increasing sparsity of data

TABLE 5.7

Simulation Results: Data 1, Sample Size = 100, Replications = 500, STT 36 Failed, STT Log 34 Failed

			Estimate MEAN				BIAS (%)				Estimate STDV				RMSE			
	Event	Pop.	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W
Untransformed Data	Q10	354	354	359	357	356	0.1	1.4	0.8	0.7	14.7	13.9	14.0	14.2	14.7	14.7	14.3	14.4
	Q50	424	423	431	430	430	-0.3	1.5	1.4	1.3	24.1	22.3	24.4	25.1	24.1	23.2	25.1	25.7
	Q100	450	449	458	460	461	-0.3	1.7	2.3	2.5	30.9	28.8	34.2	35.6	30.9	29.7	35.7	37.3
	Q200	474	469	478	479	478	-1.1	0.8	1.1	1.0	39.1	35.4	39.2	39.9	39.4	35.6	39.5	40.2
Log Transformation	Q10	354	354	362	359	358	0.1	2.4	1.4	1.1	14.4	14.6	13.8	13.8	14.4	16.8	14.7	14.4
	Q50	424	425	443	437	436	0.2	4.4	3.1	2.8	23.9	24.4	22.9	23.2	23.9	30.6	26.4	26.1
	Q100	450	451	473	469	469	0.1	5.1	4.2	4.2	30.6	29.6	29.6	30.6	30.6	37.4	35.1	35.9
	Q200	474	472	500	498	498	-0.4	5.5	5.2	5.1	38.1	35.8	37.8	38.7	38.1	44.3	45.0	45.7

TABLE 5.8

*Simulation Results: Data 1, Sample Size = 50, Replications = 1000, STT 81 Failed, STT Log 63 Failed

			Estimate MEAN				BIAS (%)				Estimate STDV				RMSE			
	Event	Pop.	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W
Untransformed Data	Q10	354	355	361	357	358	0.2	2.0	0.8	1.1	19.9	19.1	19.3	19.0	19.9	20.4	19.5	19.4
	Q50	424	421	435	432	432	-0.7	2.4	1.8	1.7	32.5	30.7	35.8	34.0	32.6	32.3	36.5	34.7
	Q100	450	446	458	452	454	-0.8	1.7	0.4	0.8	41.5	37.0	41.2	40.7	41.7	37.8	41.2	40.8
	Q200	474	455	473	465	468	-3.9	-0.2	-2.0	-1.2	42.8	39.7	43.5	43.8	46.7	39.7	44.5	44.1
Log Transformation	Q10	354	355	366	359	362	0.4	3.4	1.4	2.1	19.5	20.7	19.0	18.8	19.5	24.0	19.6	20.2
	Q50	424	425	452	440	443	0.2	6.6	3.6	4.3	32.2	34.5	32.3	31.1	32.2	44.4	35.7	36.1
	Q100	450	450	484	469	473	0.1	7.5	4.2	5.1	40.7	41.9	39.9	39.2	40.7	53.8	44.2	45.5
	Q200	474	465	511	493	499	-1.9	7.8	3.9	5.4	44.6	49.0	45.5	45.8	45.4	61.4	49.1	52.4

TABLE 5.9

Simulation Results: Data 2, Sample Size = 100, Replications = 500, STT 50 Failed, STT Log 34 Failed

			Estimate MEAN				BIAS (%)				Estimate STDV				RMSE			
	Event	Pop.	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W
Untransformed Data	Q10	519	517	522	525	524	-0.2	0.6	1.1	1.1	44	41	42	43	44	41	43	43
	Q50	751	742	758	773	774	-1.2	1.0	3.0	3.1	93	95	103	103	93	95	105	105
	Q100	850	860	885	917	920	1.2	4.1	7.8	8.1	134	165	188	188	134	169	199	200
	Q200	950	942	950	967	967	-0.8	-0.0	1.7	1.8	194	192	199	200	194	192	200	201
Log Transformation	Q10	519	519	542	532	529	0.2	4.5	2.5	2.0	43	44	42	42	43	50	44	43
	Q50	751	754	813	792	787	0.4	8.2	5.5	4.8	93	95	93	95	93	113	102	101
	Q100	850	862	934	919	916	1.4	9.8	8.1	7.8	129	129	142	151	129	153	158	164
	Q200	950	954	1051	1040	1033	0.5	10.6	9.4	8.7	188	179	201	207	188	205	220	223

TABLE 5.10

Simulation Results: Data 2, Sample Size = 50, Replications = 1000, STT 110 Failed, STT Log 82 Failed

			Estimate MEAN				BIAS (%)				Estimate STDV				RMSE			
	Event	Pop.	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W	AF	STT	FF	W
Untransformed Data	Q10	519	519	526	524	527	0.1	1.4	1.1	1.7	60	57	59	58	60	58	59	59
	Q50	751	734	779	800	792	-2.3	3.8	6.5	5.5	121	157	176	165	122	159	182	170
	Q100	850	841	846	852	855	-1.1	-0.5	0.2	0.5	188	185	189	188	188	185	189	188
	Q200	950	859	876	882	891	-9.6	-7.8	-7.1	-6.2	190	188	195	197	210	202	207	205
Log Transformation	Q10	519	522	552	532	539	0.7	6.5	2.6	3.8	59	62	58	58	59	71	59	61
	Q50	751	757	843	794	806	0.8	12.2	5.8	7.3	126	137	135	133	126	165	142	143
	Q100	850	858	966	900	920	0.9	13.5	5.9	8.2	184	187	187	187	184	219	194	199
	Q200	950	914	1075	986	1019	-3.8	13.1	3.8	7.2	206	230	224	232	209	262	227	242

in the tail. A negative bias in a QT event estimate is usually more critical than a positive one, since the underdesign of flood control structures have more severe consequences than the overdesign.

Still considering the untransformed data bias results, note that for the density methods of STT, FF, W the QT event with the largest positive bias is QN where N is the sample size. This holds for both data sets, and for both N=50 and N=100. On the other hand, for the distribution criteria based method of AF, the bias is generally small but negative, with its magnitude simply increasing for the higher events.

Now consider the bias results, log transformed data. The methods of STT, FF, W all have a larger bias than with the untransformed data, and for the events considered, a positive one. However if higher events were calculated, say Q500, a negative bias could be expected. The effect of the log transform with these methods is to shift the QT event with the largest positive bias higher, to say about Q2N. With the method of AF, the log transformation results in QT event estimates which are less biased. This can be ascribed to the fact that the log transform decreases the skew of the data, and the method of AF makes use of an empirical distribution formula to assign probabilities based on rank information.

Consider now the standard deviation results. As would be expected, in all cases the standard deviation increases with increasing QT event.

The last columns in Tables 5.7 through 5.10 give the RMSE. For the untransformed data, the method of STT performed best in terms of RMSE. For the log transformed data, the method of AF performed best. Except for this method, the RMSE increased for all the data-based algorithms with the log transformation. Of all the methods, STT has the least variance, and AF log the smallest bias. With all the estimators, but especially with AF and AF log, the RMSE is mainly due to variance, as the estimates are generally very unbiased. With the higher value of skewness and the smaller sample size, the results are more biased.

Finally, consider the change in RMSE when the sample size doubles from N=50 to N=100. With more information available, one would expect the RMSE to decrease. This is almost always true, the ratio $RMSE_{100}/RMSE_{50}$ usually being in the range 0.70 to 0.90. However there are two exceptions with untransformed Data 2 for the Q100 event estimate by the methods of FF and W. The reason for this is the large bias for these methods as shown in Table 5.9 compared with the corresponding values in Table 5.10. The cause of the large bias can in turn be traced to the abnormally large smoothing parameters for these methods with untransformed Data 2 as

shown in Tables 5.13 and 5.14. In Table 5.13 for the method of FF the mean h is 40.53, the largest of any method. Similarly, in Table 5.14 for the method of W the mean h is 53.05, again the largest. This problem of oversmoothing for these methods does not exist with log transformed Data 2. Note that both FF and W are density estimation methods.

In general there is a very close correlation between bias in Tables 5.7 to 5.10 and smoothing parameter h in Tables 5.11 to 5.14. Small h corresponds to a small bias, and large h to a large bias. Some exceptions occur for the Q200 event estimate when the sample size is 50. Even though the different data-based algorithms make use of different kernels, there is still a close correlation, which confirms theory presented in Chapter III.

It is interesting to examine the coefficient of variation, C.V. of the smoothing parameter in Tables 5.11 through 5.14. The log transform does not have much effect. Decreasing the sample size from 100 to 50 increases the C.V. of h by a factor of 1.30 to 1.45, except for the method of STT, where the factor is 1.05. However the C.V. for the method of STT is much larger than the C.V. for the other methods to begin with. Perhaps this large variance in the smoothing parameter with STT explains why QT event estimates by this method have the least variance. In these tables, note that the method of STT failed to converge to a h value

TABLE 5.11

Smoothing Parameters in Simulation:
Data 1, Sample Size = 100, Replications = 500

Method	Degenerate Cases	% Degenerate Cases	Smoothing Parameter, h		
			Mean	STDV	C.V.
Untransformed:					
AF	0	0	15.61	1.650	.106
STT	36	7.2	25.55	6.473	.253
FF	0	0	19.00	1.352	.071
W	0	0	16.15	1.529	.095
Log Transform:					
AF	0	0	.0620	.0067	.108
STT	34	6.8	.1035	.0271	.262
FF	0	0	.0771	.0056	.073
W	0	0	.0665	.0066	.099

TABLE 5.12

Smoothing Parameters in Simulation:
Data 1, Sample Size = 50, Replications = 1000

Method	Degenerate Cases	% Degenerate Cases	Smoothing Parameter, h		
			Mean	STDV	C.V.
Untransformed:					
AF	0	0	21.72	3.369	.155
STT	81	8.1	31.12	8.263	.266
FF	0	0	19.49	1.965	.101
W	0	0	23.83	2.948	.124
Log Transform:					
AF	0	0	.0854	.0135	.158
STT	63	6.3	.1260	.0345	.274
FF	0	0	.0791	.0082	.104
W	0	0	.0976	.0126	.129

TABLE 5.13

Smoothing Parameters in Simulation:
Data 2, Sample Size = 100, Replications = 500

Method	Degenerate Cases	% Degenerate Cases	Smoothing Parameter, h		
			Mean	STDV	C.V.
Untransformed:					
AF	0	0	29.24	3.386	.116
STT	50	10.0	38.97	10.74	.276
FF	0	0	40.53	4.108	.101
W	0	0	37.12	5.518	.149
Log Transform:					
AF	0	0	.1165	.0129	.111
STT	34	6.8	.1945	.0514	.264
FF	0	0	.1436	.0107	.075
W	0	0	.1235	.0124	.100

TABLE 5.14

Smoothing Parameters in Simulation:
Data 2, Sample Size = 50, Replications = 1000

Method	Degenerate Cases	% Degenerate Cases	Smoothing Parameter, h		
			Mean	STDV	C.V.
Untransformed:					
AF	0	0	40.89	6.857	.168
STT	110	11.0	49.27	14.08	.286
FF	0	0	41.54	5.811	.140
W	0	0	53.05	9.574	.180
Log Transform:					
AF	0	0	.1593	.0253	.159
STT	82	8.2	.2368	.0641	.271
FF	0	0	.1472	.0154	.105
W	0	0	.1806	.0230	.127

between 6.3 and 11% of the time when using the sample range as h(0).

Using the mean h information in Tables 5.11 to 5.14, Table 5.15 shows the effect of record length on the smoothing parameter. Based on asymptotic expansion theory for the IMSE (eq. 3.25), the theoretical value for the ratio h100/h50 is 0.871. The method STT is closest to this, differing from it by about negative 6%. This is not surprising, as the data-based algorithm of STT is formulated using asymptotic IMSE theory. The same is true for the

TABLE 5.15

Effect of Record Length on Smoothing Parameter

Method	Data 1: Ratio of h100/h50	Data 2: Ratio of h100/h50
Untransformed:		
AF	0.719	0.715
STT	0.821	0.791
FF	0.975	0.976
W	0.678	0.700
Log Transform:		
AF	0.726	0.731
STT	0.821	0.821
FF	0.975	0.976
W	0.681	0.684

method FF; however its difference from the theoretical value is positive 12%. The h value by the method FF is quite invariant to the record length, which is not desirable. The

method AF has a h100/h50 ratio with a difference from 0.871 of between negative 16 to 18%. Note that the method AF is based on an IMSE criterion at the distribution function. For the method W, the % difference from 0.871 is between negative 20 and 22%. The method W is completely heuristic.

The total execution times for the various methods are given in Table 5.16. With sample size 100, 500 replications were done; with sample size 50, 1000 replications.

TABLE 5.16

Computer Times in Simulation with 50,000 Random Values

Data	Sample Size	CPU Seconds (Amdahl Model V8) AF	STT	FF	W
Data 1					
U	100	3050	1074	24	23
L	100	2696	1023	23	23
U	50	1932	714	25	24
L	50	1412	701	24	23
Data 2					
U	100	3197	1616	22	21
L	100	2670	1009	22	22
U	50	2011	1000	26	25
L	50	1365	675	25	25

NOTE: U = Untransformed, L = Log Transformation

There is not much of a difference between the times for Data 1 and Data 2. Generally log transformed data required less time. For the method of STT, if there were more degenerate cases, more computing time was required.

TABLE 5.17

Comparison of Parametric and Nonparametric Simulation Results:
Sample Size = 50, Replications = 1000, LP3 ML 36 Failed, STT 81 Failed

		Mean Estimate							Bias(%)							Root Mean Square Error						
Event	Pop.	LP3 ML	LP3 MOM	LN3 ML	Wak	GEV ML	STT	AF Log	LP3 ML	LP3 MOM	LN3 ML	Wak	GEV ML	STT	AF Log	LP3 ML	LP3 MOM	LN3 ML	Wak	GEV ML	STT	AF Log
Data 1																						
Q50	424	420	426	429	435	422	435	425	-.9	.4	1.1	2.6	-.5	2.6	.2	32.0	29.8	31.1	33.2	27.8	32.3	32.2
Q100	450	446	453	459	467	434	458	450	-.9	.8	2.0	3.9	-3.6	1.7	.1	41.7	37.7	40.0	44.6	37.5	37.8	40.7
Data 2																						
Q50	751	744	757	781	749	717	779	757	-.9	.9	4.0	.2	-4.5	3.8	.9	124	114	127	114	129	159	126
Q100	850	847	865	903	857	761	846	858	-.4	1.7	6.2	.8	-10	-0.5	.9	177	156	176	159	179	185	184

Small sample (N=50) simulation results for the two best nonparametric techniques, AF log and STT untransformed, are compared with those of five parametric ones in Table 5.17. For Data 1, the RMSE of STT untransformed and AF log closely approach those of LP3, which is the correct parametric distribution for the generated data. For Data 2, the RMSE of AF log for Q50 is 2% larger, and for Q100 is 4% larger than the RMSE of LP3 estimated by maximum likelihood. The errors for these events lie within the traditional engineering 5%. However, the important factor is not so much the size of the errors, as the fact that they have been achieved with much weaker assumptions than in parametric analysis. For both Data 1 and Data 2 the bias of AF log is very small, sometimes even smaller than that of LP3.

While the information given in Table 5.17 is limited, it indicates the potential of nonparametric techniques in flood frequency analysis.

Chapter VI

CONCLUSIONS

6.1 MAJOR CONCLUSIONS

Based on the results presented in this thesis, the following conclusions can be made:

1. Nonparametric flood estimates obtained by the kernel method are within the same order of magnitude as the estimates obtained by currently used parametric techniques. Thus the kernel method appears to be a viable alternative with the advantage of not requiring a distributional assumption.

2. The simulation study indicates that nonparametric flood estimates can be equally or more accurate than those of the LPIII distribution used to generate the data.

3. A comparison of flood estimates by the methods AF, STT, FF and W with those assuming LN3, GEV, LPIII and Wakeby distributions shows that the nonparametric design floods determined in this study are less variable than the parametric ones. This suggests that while h is critical, it is perhaps less critical than the computation of parameters in parametric analysis.

4. The kernel estimator requires the selection of only the kernel and the smoothing factor. In this respect it is more parsimonious than parametric models.

5. Of the three kernels used, the rectangular kernel gave the roughest density estimates. Both the quadratic and Gaussian kernels gave smoother estimates.

6. Based on a visual assessment of smoothness in the density function and fit in the distribution function with two sets of real data, determination of the smoothing factor h by the method W gives the best results. However, based on simulation results, the method STT on untransformed data and AF on log transformed data are preferable. It is therefore recommended that all three methods be used in flood frequency analysis applications. Should a particular method be unsuitable for a given set of data (i.e. STT is known to fail 5-10% of the time), the other methods serve as alternates.

7. Flood data for the Margaree River has been shown to be multimodal due to different causative factors for spring and fall floods. Nonparametric density estimation is particularly suitable for multimodal data.

8. Flood data with long tails (eg. Warta River station) present fitting difficulties with the commonly used distributions since more weight is given to the central part. This is especially true when its parameters are estimated by

the method of moments. However, the nonparametric approach allows a greater weight in the tails, and thus provides more reliable flood estimates.

9. With untransformed data, the kernel estimator is not suitable for use outside the range of the data. An estimate too far into the tail would depend heavily on the highest observation and the particular kernel used. Thus with untransformed data the kernel estimator should be used for interpolating rather than extrapolating purposes.

10. Log transformed flood data has also been analysed in this thesis. The log transform decreased the problem of bumps at outlying observations especially evident in untransformed data analysis for the method AF. In the simulation only the method AF showed an improvement in RMSE when used with log transformed data instead of untransformed data.

11. Given long-term stations with record lengths of 50 or more years, flood magnitude values up to the Q100 event can reasonably be estimated by nonparametric methods. Floods with 20 to 100 year return periods are used in many engineering problems, for example irrigation design and the delineation of flood risk areas.

12. The coefficient of skewness could be used to decide whether the data should be transformed. In the simulation,

the nonparametric flood estimates with the more highly skewed data had a larger relative RMSE. For Data 1 the Q50 event is best estimated by the method W on untransformed data. For Data 2 it is preferable to use log transformed data to estimate this event by the method W. Thus it would appear that with increasingly skewed data, a transformation improves the estimates of floods.

13. The optimal histogram and optimal frequency polygon are useful in the graphical presentation of flood data. Adequate graphs can also result by plotting the flood data in terms of % maximum discharge.

14. Methods DHH and NDMPLE appear to be less suitable for use in flood frequency analysis than the other methods investigated. The problem with DHH lies in the nonlinear nature of flood data or in programming difficulty. The problem with NDMPLE is that the user must specify four factors, and there is no clear-cut criterion for balancing the penalty and the likelihood. This makes NDMPLE more of a parametric procedure than a nonparametric one.

15. As nonparametric methods are data-based, a fair amount of computation is required. The algorithms ranked in order of increasing computation time are: W, FF, STT, AF.

In conclusion, the potential usefulness of nonparametric techniques in flood frequency analysis has been demonstrated.

In addition, these techniques can be modified and refined for use in other areas in hydrology.

6.2 RECOMMENDATIONS FOR FURTHER RESEARCH

During this study is has become evident that there are a number of areas where continued research would be beneficial. These include:

1. Further research on the automatic selection of h. In their recent book, Devroye et al. (1985) challenge the widely accepted IMSE density estimation criterion. They note that with this criterion regions with low f values such as tails are less important and misleading estimates can result. They show that L_1 theory, which makes use of the error criterion $J = \int |\hat{f}(x)-f(x)|\, dx$, is in many ways preferable. In L_1 for the kernel estimator, E(J) decreases at the rate $N^{-2/5}$ if h is chosen in proportion to $N^{-1/5}$. The corresponding IMSE rate is $N^{-4/5}$ if h is chosen proportional to $N^{-1/5}$.

2. A comprehensive investigation should be made of the effect and suitability of various transformations. Ideally a transformation should have both theoretical and physical justification. Of particular interest is the transformed kernel estimate suggested by Devroye et al. (1985, p. 244) which is based on triangularizing the data instead of normalizing it. In the method a robust parametric distribution is initially assumed, then transformed to the

triangular density on [0,1]. Using the Epanechnikov kernel, the optimal smoothing factor h is explicitly known as $h = (5/192\pi N)$. Retransformation would provide flood estimates.

3. The nonparametric one-sample procedure of Parzen (1979) could be used in flood frequency analysis. The basis of the method is the fact that every probability family has a characteristic density-quantile function. Empirical and theoretical standardized density-quantile functions are compared in order to select the distribution that best represents the data.

4. Similarly, the Bayesian approach to probability density estimation presented by Ishiguro et al. (1984) could have application in flood frequency analysis. They have shown that their method works efficiently for any data set provided the true density function is properly smooth.

5. Simulations involving multimodal data, outliers, or other skew classes and record lengths could be done. This would be required to reflect various hydrological regimes.

6. Further studies could investigate multimodal data in greater detail, possibly in more than one dimension. Nonparametric confidence intervals are an open question. Finally, nonparametric regression is a related subject with many potential applications in hydrology.

REFERENCES

Adamowski, K. (1981) Plotting Formula for Flood Frequency. Water Resources Bulletin, 17(2), 197-202.

Anderson, G.D. (1969) A Comparison of Probability Density Estimates. Presented at IMS Annual Meeting, 19-22 August, New York.

Bartlett, M.S. (1963) Statistical Estimation of Density Functions. Sankhya (A), 25, 245-54.

Bean, S.J. and C.P. Tsokos (1980) Developments in Nonparametric Density Estimation. International Statistical Review, 48, 267-287.

Benson, M.A. (1950) Use of Historical Data in Flood Frequency Analysis. Transactions, American Geophysical Union, 31(3), 419-22.

Bobee, B. (1975) The Log Pearson Type 3 Distribution and Its Application in Hydrology. Water Resources Research, 11(5), 681-9.

Boneva, L.I., D.G. Kendall and I. Stefanov (1971) Spline Transformations: Three New Diagnostic Aids for the Statistical Data - Analyst. J.R. Statist. Soc. B., 33, 1-70.

Bowman, A.W. (1982) A Comparative Study of some Kernel-Based Nonparametric Density Estimators. Manchester-Sheffield School of Probability and Statistics Research Report No. 84/AWB/1.

Burges, S.J., D.P. Lettenmaier and C.L. Bates (1975) Properties of the Three-Parameter Log Normal Probability Distribution. Water Resources Research, 11(2), 229-35.

Cencov, N.N. (1962) Evaluation of an Unknown Distribution Density From Observations. Translated as Soviet Math. Dokl., 3, 1559-62.

Chow, V.T., ed. (1964) Handbook of Applied Hydrology, McGraw-Hill Book Co., New York.

Condie, R. (1977) The Log Pearson Type 3 Distribution: The T-Year Event and Its Asymptotic Standard Error by Maximum Likelihood Theory. Water Resources Research, 13(6), 987-91.

Condie, R., G.A. Nix and L.G. Boone (1981) Flood Damage Reduction Program Flood Frequency Analysis, Engineering Hydrology Section, Inland Waters Directorate, Environment Canada, Ottawa.

Condie, R. and Lee, K.A. (1982) Flood Frequency Analysis with Historic Information. Journal of Hydrology, 58, 47-61.

Cooper, D.M. and R.T. Clarke (1980) Distribution-Free Methods for Estimating Flood and Streamflow Exceedance Probabilities by Correlation. Water Resources Research, 16(6), 1121-7.

Costa, J.E. (1978) Holocene Stratigraphy in Flood Frequency Analysis. Water Resources Research, 14(4), 626-32.

Cunnane, C. (1978) Unbiased Plotting Positions - A Review. Journal of Hydrology, 37, 205-22.

DeMontricher, G.F., R.A. Tapia and J.R. Thompson (1975) Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods. The Annals of Statistics, 3(6), 1329-48.

Devroye, L. and L. Gyorfi (1985). Nonparametric Density Estimation: The L_1 View, John Wiley & Sons, New York.

Duin, R.P.W. (1976) On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions. IEEE Transactions on Computers, C-25, 1175-9.

Epanechnikov, V.A. (1969) Nonparametric Estimation of a Multivariate Probability Density. Theory of Probability and Its Application, 14, 153-8.

Flanagan, P. (1980) Computer Programs for Kernel Estimates of a Probability Density Function, M.S. Thesis, University of South Carolina.

Fleming, G. (1975) Computer Simulation Techniques in Hydrology, Elsevier Applied Science Publishers, London.

Fryer, M.J. (1977) A Review of Some Nonparametric Methods of Density Estimation. J. Inst. Math. Appl., 20, 335-54.

Good, I.J and R.A. Gaskins (1971) Nonparametric Roughness Penalties for Probability Densities. Biometrika, 58, 255-78.

Good, I.J. and R.A. Gaskins (1980) Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data. J. Amer. Statist. Assoc., 75, 42-56.

Greblicki, W. and A. Krzyżak (1979) Nonparametric Identification of a Memoryless System with a Cascade Structure. Int. J. Systems Sci., 10(11), 1301-10.

Greis, N.P. (1983) Flood Frequency Analysis: A Review of 1979-1982. Reviews of Geophysics and Space Physics, 21(3), 699-706.

Hall, M.J. (1984) Urban Hydrology, Elsevier Applied Science Publishers, London.

Hall, P. (1983) Large Sample Optimality of Least Squares Cross-Validation in Density Estimation. The Annals of Statistics, 11(4), 1156-74.

Hermans, J. and Habbema, J.D.F. (1976) Manual for the ALLOC Discriminant Analysis Programs, University of Leiden, Dept. of Medical Statistics.

Houghton, J.C. (1978) Birth of a Parent: The Wakeby Distribution for Modeling Flood Flows. Water Resources Research, 14(6), 1105-9.

International Mathematical and Statistical Libraries, Inc. (1980), Houston, Texas.

Ishiguro, M. and Yosiynki Sakamoto (1984) A Bayesian Approach to the Probability Density Estimation. Ann. Inst. Statist. Math., 36 (3,B).

Kite, G.W. (1977) Frequency and Risk Analysis in Hydrology, Water Resources Publications, Fort Collins, Colorado.

Klemes, V. (1980) Aspirations and Realities. Canadian Climate Program Water Workshop, University of Alberta, Edmonton.

Kuczera, G. (1982) Robust Flood Frequency Models. Water Resources Research, 18(2), 315-24.

Landwehr, J.M., N.C. Matalas and J.R. Wallis (1978) Some Comparisons of Flood Statistics in Real and Log Space. Water Resources Research, 14(5), 902-20.

Landwehr, J.M., N.C. Matalas and J.R. Wallis (1979) Probability Weighted Moments Compared with some Traditional Techniques in Estimating Gumbel Parameters and Quantiles. Water Resources Research, 15(5), 1055-64.

Lettenmaier, D.P. and S.J. Burges (1978) A Nonparametric Approach to the Analysis of Extreme Hydrologic Events in Proc., Vol. II, Int. Symp. on Risk and Reliability in Water Resources, Waterloo, Ontario, Canada.

Lin, W. and R.A. Bathe (1984) Selection of Probability Distributions. Submitted for publication to Water Resources Research.

Liu, M. (1983) Computerized Density Estimation, M.S. Thesis, University of South Carolina.

Loftsgaarden, D.O. and C.P. Quesenberry (1965) A Nonparametric Estimate of a Multivariate Density Function. Annals of Mathematical Statistics, 38, 1261-5.

Matalas, N.C., J.R. Slack and J.R. Wallis (1975) Regional Skew in Search of a Parent. Water Resources Research, 11(6), 815-26.

McCuen, R.H. and W.J. Rawls (1979) Classification of Evaluation of Flood Flow Frequency Estimation Techniques. Water Resources Bulletin, 15(1), 88-93.

National Research Council (1982), Scientific Basis of Water-Resource Management, Studies in Geophysics, National Academy Press, Washington, D.C.

Nozdryn-Plotnicki, M.J. and W.E. Watt (1979) Assessment of Fitting Techniques for the Log Pearson Type 3 Distribution using Monte Carlo simulation. Water Resources Research, 15(3), 714-8.

Parzen, E. (1962) On Estimation of a Probability Density Function and Mode. Annals of Mathematical Statistics, 33, 1065-1076.

Parzen, E. (1979) Nonparametric Statistical Data Modeling. J. Amer. Statist. Assoc., 74, 105-21.

Pilon, P.J., R. Condie and K.D. Harvey (1985) Consolidated Frequency Analysis Package Version 1, Water Resources Branch, Inland Waters Directorate, Environment Canada, Ottawa.

Rao, D.V. (1980) Log Pearson Type 3 Distribution: Method of Mixed Moments. Proc. Am. Soc. Civ. Engrs., J. Hydraul, Div., 106(HY6), 999-1019.

Rosenblatt, M. (1956) Remarks on Some Nonparametric Estimates of a Density Function. Annals of Mathematical Statistics, 27, 832-7.

Rossi, F., M. Fiorentino and P. Versace (1984) Two-Component Extreme Value Distribution for Flood Frequency Analysis. Water Resources Research, 20(7), 847-56.

Rudemo, M. (1982) Empirical Choice of Histograms and Kernel Density Estimators. Scand. J. Statist., 9, 65-78.

Sagan, C., O.B. Toon and J.B. Pollack (1979) Anthropogenic Albedo Changes and the Earth's Climate. Science, 206(4425), 1363-8.

Sangal, B.P. and A.K. Biswas (1970) The Three-Parameter Lognormal Distribution and its Applications in Hydrology. Water Resources Research, 6(2), 505-15.

Schoenberg, I.J. (1972) Notes on Spline Functions II on the Smoothing of Histograms. MRC Technical Report 1222.

Schuster, E.F. (1982) Considerations in Cross-Validation Type Density Smoothing with a Look at Some Data in Proc. of the NASA Workshop on Density Estimation and Function Smoothing, Texas A&M University.

Scott, D.W. (1976) Nonparametric Probability Density Estimation by Optimization Theoretic Techniques. Technical Report 476-131-1, Rice University, Houston, Texas.

Scott, D.W., R.A. Tapia and J.R. Thompson (1977) Kernel Density Estimation Revisited. Journal of Nonlinear Analysis, Theory, Methods and Applications, 1, 339-72.

Scott, D.W. (1979) On Optimal and Data-Based Histograms. Biometrika, 66(3), 605-10.

Scott, D.W. and L.E. Factor (1981) Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators. J. Amer. Statist. Assoc., 76(373), 9-15.

Scott, D.W. (1984) Frequency Polygons: Theory and Application, Unpublished Report, Rice University.

Silverman, B.W. (1978) Choosing the Window Width When Estimating a Density. Biometrika, 65(1), 1-11.

Silverman, B.W. (1981) Using Kernel Density Estimates to Investigate Multimodality. J.R. Statist. Soc. B., 43(1), 97-99.

Silverman, B.W. (1982) Kernel Density Estimation using the Fast Fourier Transform. Applied Statistics, 31, 93-9.

Singh, K.P. and M. Nakashima (1983) A New Methodology for Flood Frequency Analysis with Objective Detection and Modification of Outliers/Inliers in Frontiers in Hydrology, Water Resources Publications, Littleton, Colorado.

Srikanthan, R. and T.A. McMahon (1981) Log Pearson III Distribution - An Empirically Derived Plotting Position. Journal of Hydrology, 52, 161-3.

Tapia, R.A. and J.R. Thompson (1978) Nonparametric Probability Density Estimation, John Hopkins University Press, Baltimore, Maryland.

Tarter, M.E. and R.A. Kronmal (1976) An Introduction to the Implementation and Theory of Nonparametric Density Estimation. The American Statistician, 30, 105-12.

Unny, T.E., V.S. Panuk, C.D. MacInness and A.K.C. Wong (1981) Pattern Analysis and Synthesis of Time-Dependent Hydrologic Data in Advances in Hydroscience, Vol. 12, Academic Press.

U.S. Water Resources Council (1977) Guidelines for Determining Flood Flow Frequency, Bull. #17A, Washington, D.C.

Vicens, G.J., I. Rodriguez-Iturbe and J.C. Schaake (1975) A Bayesian Framework for the Use of Regional Information in Hydrology. Water Resources Research, 11(3), 405-14.

Wagner, T.J. (1975) Nonparametric Estimates of Probability Density. IEEE Transactions on Information Theory, IT-21(4), 438-40.

Wegman, E.J. (1972a) Nonparametric Probability Density Estimation: I. A Summary of Available Methods. Technometrics, 14(3), 533-46.

Wegman, E.J. (1972b) Nonparametric Probability Density Estimation: II. A Comparison of Density Estimation Methods. J. Statist. Comput. Simul., 1, 225-45.

Wertz, W. and B. Schneider (1979) Statistical Density Estimation: A Bibliography. International Statistical Review, 47, 155-75.

Wood, E.F. and I. Rodriguez-Iturbe (1975) A Bayesian Approach to Analyzing Uncertainty Among Flood Frequency Models. Water Resources Research, 11(6), 839-43.

Woodroofe, M. (1970) On Choosing a Delta Sequence. Annals of Mathematical Statistics, 41, 1665-71.

Yakowitz, S. (1981) Applications of Pattern Recognition Methods to Hydrologic Time Series in Proceedings, 13th Symposium on the Interface of Computer Science and Statistics, Springer-Verlag, New York.

Yakowitz, S. (1985) Markov Flow Models and the Flood Warning Problem, Water Resources Research, 21(1), 81-8.

Yevjevich, V. and J.T.B. Obeysekera (1984) Estimation of Skewness of Hydrologic Variables. Water Resources Research, 20(7), 935-43.

Zhang, Y. (1982) Plotting Positions of Annual Flood Extremes Considering Extraordinary Values. Water Resources Research, 18(4), 859-64.

APPENDIX A

Parametric Flood Frequency Analysis Results

TABLE A.1

WSC STATION NO=01FB001 Listing of Data
WSC STATION NAME=NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY

MONTH	YEAR	DATA	ORDERED	RANK	PROB.	RET. PERIOD
(1)	(2)	(3) (CMS)	(4) (CMS)	(5)	(6) (%)	(7) (YEARS)
11	1917	163.000	355.000	1	0.89	112.000
5	1918	283.000	343.000	2	2.38	42.000
4	1919	124.000	292.000	3	3.87	25.846
3	1920	237.000	283.000	4	5.36	18.667
3	1921	206.000	272.000	5	6.85	14.609
5	1922	175.000	263.000	6	8.33	12.000
10	1923	249.000	249.000	7	9.82	10.182
5	1924	146.000	237.000	8	11.31	8.842
2	1925	137.000	226.000	9	12.80	7.814
5	1926	178.000	225.000	10	14.29	7.000
11	1927	272.000	225.000	11	15.77	6.340
7	1928	125.000	224.000	12	17.26	5.793
5	1929	138.000	216.000	13	18.75	5.333
3	1930	149.000	214.000	14	20.24	4.941
4	1931	80.100	206.000	15	21.73	4.603
1	1932	156.000	206.000	16	23.21	4.308
11	1933	119.000	206.000	17	24.70	4.048
5	1934	226.000	201.000	18	26.19	3.818
1	1935	225.000	191.000	19	27.68	3.613
3	1936	225.000	190.000	20	29.17	3.429
11	1937	187.000	189.000	21	30.65	3.262
4	1938	163.000	187.000	22	32.14	3.111
10	1939	130.000	187.000	23	33.63	2.973
5	1940	168.000	186.000	24	35.12	2.847
5	1941	185.000	185.000	25	36.61	2.732
5	1942	120.000	178.000	26	38.10	2.625
5	1943	167.000	177.000	27	39.58	2.526
5	1944	119.000	175.000	28	41.07	2.435
5	1945	109.000	174.000	29	42.56	2.350
4	1946	177.000	168.000	30	44.05	2.270
5	1947	214.000	167.000	31	45.54	2.196
5	1948	190.000	167.000	32	47.02	2.127
11	1949	161.000	166.000	33	48.51	2.061
4	1950	111.000	163.000	34	50.00	2.000
3	1951	129.000	163.000	35	51.49	1.942
1	1952	125.000	162.000	36	52.98	1.888
2	1953	174.000	161.000	37	54.46	1.836
2	1954	77.000	156.000	38	55.95	1.787
9	1955	118.000	150.000	39	57.44	1.741
1	1956	187.000	149.000	40	58.93	1.697
5	1957	123.000	146.000	41	60.42	1.655
4	1958	131.000	144.000	42	61.90	1.615
11	1959	121.000	138.000	43	63.39	1.577
4	1960	131.000	138.000	44	64.88	1.541
5	1961	167.000	137.000	45	66.37	1.507
4	1962	343.000	137.000	46	67.86	1.474
5	1963	144.000	131.000	47	69.35	1.442
5	1964	137.000	131.000	48	70.83	1.412

Continuation of TABLE A.1

WSC STATION NO=01FB001 Listing of Data
WSC STATION NAME=NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY

MONTH	YEAR	DATA	ORDERED	RANK	PROB.	RET. PERIOD
(1)	(2)	(3) (CMS)	(4) (CMS)	(5)	(6) (%)	(7) (YEARS)
5	1965	108.000	130.000	49	72.32	1.383
5	1966	94.300	129.000	50	73.81	1.355
11	1967	263.000	125.000	51	75.30	1.328
12	1968	206.000	125.000	52	76.79	1.302
5	1969	292.000	124.000	53	78.27	1.278
5	1970	121.000	123.000	54	79.76	1.254
11	1971	206.000	123.000	55	81.25	1.231
5	1972	224.000	121.000	56	82.74	1.209
5	1973	138.000	121.000	57	84.23	1.187
4	1974	123.000	120.000	58	85.71	1.167
12	1975	216.000	119.000	59	87.20	1.147
5	1976	166.000	119.000	60	88.69	1.128
6	1977	201.000	118.000	61	90.18	1.109
1	1978	191.000	111.000	62	91.67	1.091
1	1979	162.000	109.000	63	93.15	1.073
12	1980	186.000	108.000	64	94.64	1.057
4	1981	189.000	94.300	65	96.13	1.040
4	1982	355.000	80.100	66	97.62	1.024
3	1983	150.000	77.000	67	99.11	1.009

TABLE A.2

FREQUENCY ANALYSIS - THREE-PARAMETER LOGNORMAL DISTRIBUTION
01FB001 NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	171.827	57.367	0.334	1.073	4.500
LN X SERIES	5.096	0.319	0.063	0.166	3.135
LN(X-c) SERIES	4.933	0.374	0.076	0.002	3.210

X(MIN)= 77.000
X(MAX)= 355.000
LOWER OUTLIER LIMIT OF X= 65.249

TOTAL SAMPLE SIZE= 67
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

3LN PARAMETERS: c = 23.073 μ_z = 4.933 σ_z = 0.374

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	72.70
1.050	0.952	97.40
1.250	0.800	124.00
2.000	0.500	162.00
5.000	0.200	213.00
10.000	0.100	247.00
25.000	0.040	290.00
50.000	0.020	323.00
100.000	0.010	355.00
200.000	0.005	387.00
500.000	0.002	431.00

TABLE A.3

FREQUENCY ANALYSIS - GENERALIZED EXTREME VALUE DISTRIBUTION
01FB001 NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	171.827	57.367	0.334	1.073	4.500
LN X SERIES	5.096	0.319	0.063	0.166	3.135

X(MIN)= 77.000
X(MAX)= 355.000
LOWER OUTLIER LIMIT OF X= 65.249

TOTAL SAMPLE SIZE= 67
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

GEV PARAMETERS: C = 145.67 A= 43.287 K= -0.025

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	71.10
1.050	0.952	98.10
1.250	0.800	125.00
2.000	0.500	162.00
5.000	0.200	212.00
10.000	0.100	246.00
25.000	0.040	290.00
50.000	0.020	323.00
100.000	0.010	357.00
200.000	0.005	391.00
500.000	0.002	436.00

TABLE A.4

FREQUENCY ANALYSIS - LOG PEARSON TYPE III DISTRIBUTION
01FB001 NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	171.827	57.367	0.334	1.073	4.500
LN X SERIES	5.096	0.319	0.063	0.166	3.135

X(MIN)= 77.000
X(MAX)= 355.000
LOWER OUTLIER LIMIT OF X= 65.249

TOTAL SAMPLE SIZE= 67
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

LP3 PARAMETERS: A= 0.2729E-01 B= 134.6 LOG(C)= 1.422
C = 4.146

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	72.60
1.050	0.952	97.90
1.250	0.800	125.00
2.000	0.500	162.00
5.000	0.200	213.00
10.000	0.100	246.00
25.000	0.040	290.00
50.000	0.020	322.00
100.000	0.010	355.00
200.000	0.005	389.00
500.000	0.002	434.00

TABLE A.5

FREQUENCY ANALYSIS - WAKEBY DISTRIBUTION
01FB001 NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	171.827	57.367	0.334	1.073	4.500
LN X SERIES	5.096	0.319	0.063	0.166	3.135

X(MIN)= 77.000
X(MAX)= 355.000
LOWER OUTLIER LIMIT OF X= 65.249

TOTAL SAMPLE SIZE= 67
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

THE FOLLOWING WAKEBY PARAMETERS WERE OBTAINED BY ASSUMING E TO BE NON-ZERO. THE ITERATION ALGORITHM WAS NOT REQUIRED.

E= 83.137 A= 38.623 B= 5.80 C= -1483.534 D=-0.039
DISTRIBUTION IS UPPER BOUNDED AT 0.1605E+04

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	84.00
1.050	0.952	95.50
1.250	0.800	124.00
2.000	0.500	161.00
5.000	0.200	212.00
10.000	0.100	249.00
25.000	0.040	297.00
50.000	0.020	332.00
100.000	0.010	366.00
200.000	0.005	399.00
500.000	0.002	441.00

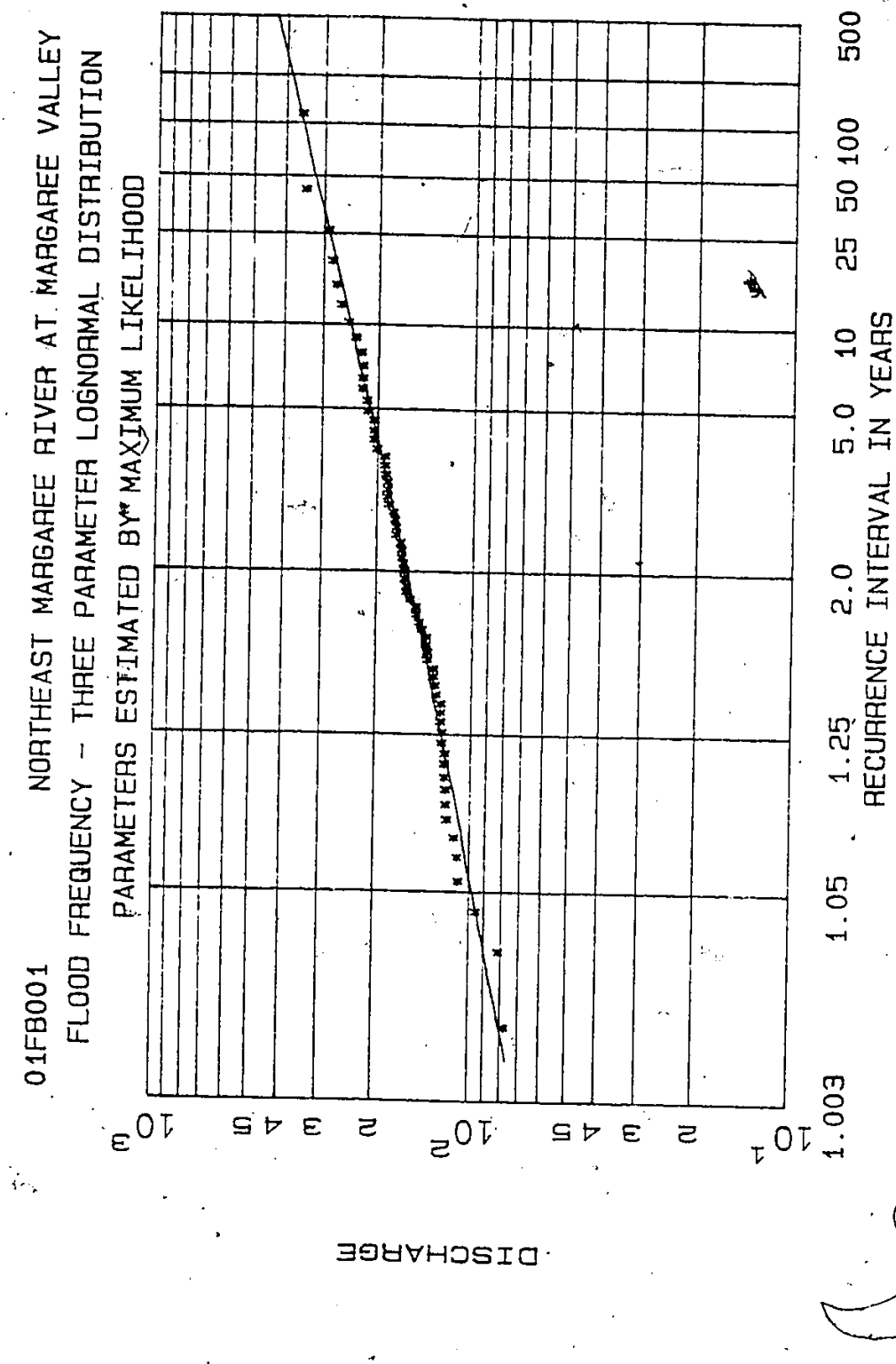

01FB001 NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY
FLOOD FREQUENCY - THREE PARAMETER LOGNORMAL DISTRIBUTION
PARAMETERS ESTIMATED BY MAXIMUM LIKELIHOOD
DISCHARGE
RECURRENCE INTERVAL IN YEARS
1.003
1.05
1.25
2.0
5.0
10
25
50
100
500

FIG. A.1

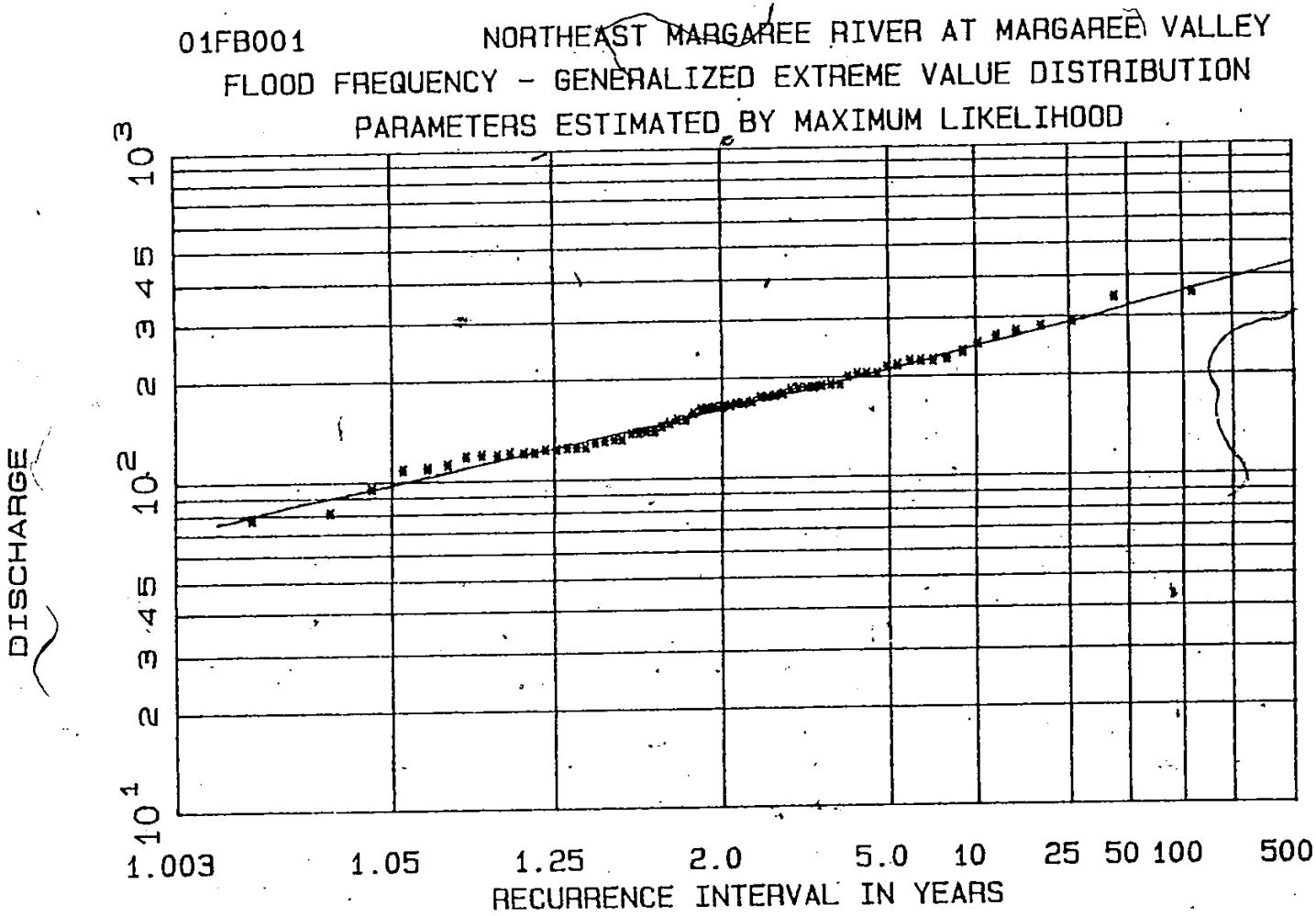

01FB001 NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY
FLOOD FREQUENCY - GENERALIZED EXTREME VALUE DISTRIBUTION
PARAMETERS ESTIMATED BY MAXIMUM LIKELIHOOD
DISCHARGE
10^3
2 3 4 5
10^2
2 3 4 5
10^1
2 3
1.003
1.05
1.25
2.0
5.0
10
25
50
100
500
RECURRENCE INTERVAL IN YEARS

FIG. A.2

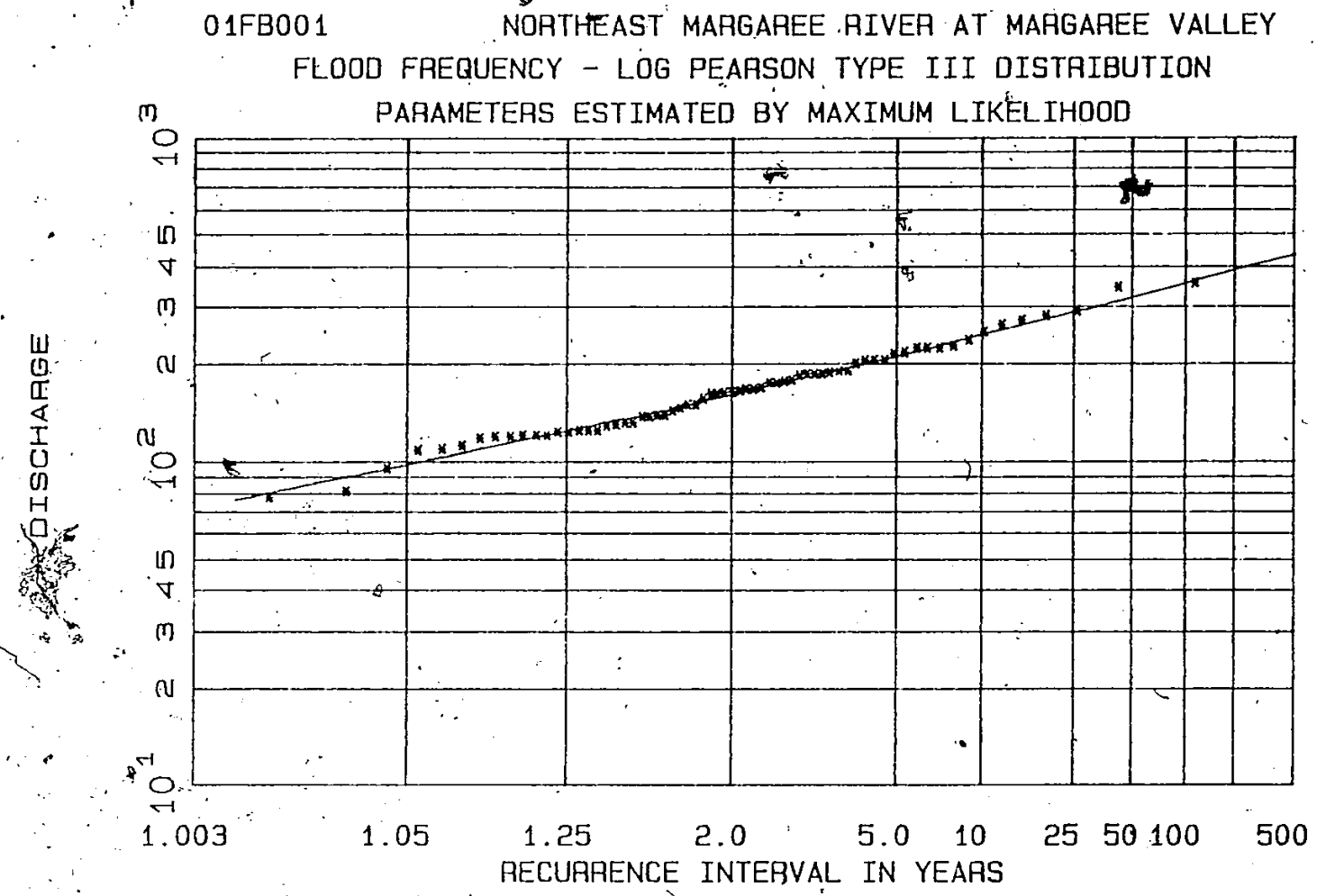

01FB001 NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY
FLOOD FREQUENCY - LOG PEARSON TYPE III DISTRIBUTION
PARAMETERS ESTIMATED BY MAXIMUM LIKELIHOOD
DISCHARGE
10^3
2 3 4 5
10^2
2 3 4 5
2
10^1
1.003
1.05
1.25
2.0
5.0
10
25
50
100
500
RECURRENCE INTERVAL IN YEARS

FIG. A.3

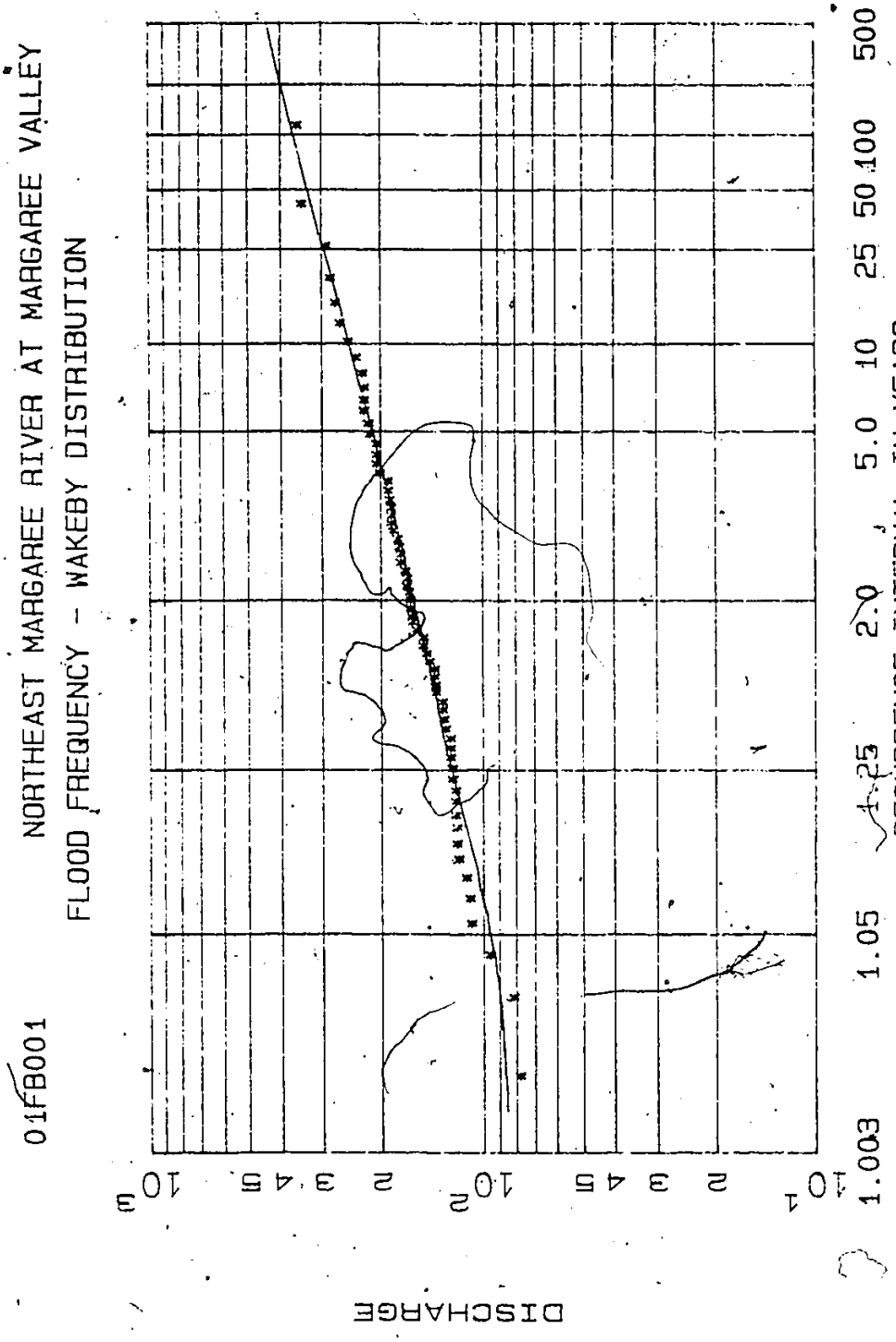

01FB001
NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY
FLOOD FREQUENCY - WAKEBY DISTRIBUTION
DISCHARGE
RECURRENCE INTERVAL IN YEARS
1.003
1.05
1.25
2.0
5.0
10
25
50
100
500

FIG. A.4

TABLE A.6

Listing of Data for
STATION NAME=POLISH RIVER WARTA AT POZNAN

MONTH (1)	YEAR (2)	DATA (3) (CMS)	ORDERED (4) (CMS)	RANK (5)	PROB. (6) (%)	RET. PERIC (7) (YEARS)
1	1826	125.000	1573.000	1	0.43	233.667
1	1827	319.000	1515.000	2	1.14	87.625
1	1828	191.000	1510.000	3	1.85	53.923
1	1829	275.000	1460.000	4	2.57	38.944
1	1830	1035.000	1395.000	5	3.28	30.478
1	1831	393.000	1253.000	6	3.99	25.036
1	1832	206.000	1162.000	7	4.71	21.242
1	1833	382.000	1120.000	8	5.42	18.447
1	1834	692.000	1105.000	9	6.13	16.302
1	1835	105.000	1073.000	10	6.85	14.604
1	1836	303.000	1055.000	11	7.56	13.226
1	1837	548.000	1035.000	12	8.27	12.086
1	1838	725.000	1015.000	13	8.99	11.127
1	1839	512.000	990.000	14	9.70	10.309
1	1840	525.000	775.000	15	10.41	9.603
1	1841	745.000	770.000	16	11.13	8.987
1	1842	118.000	750.000	17	11.84	8.446
1	1843	207.000	745.000	18	12.55	7.966
1	1844	300.000	728.000	19	13.27	7.538
1	1845	1015.000	728.000	20	13.98	7.153
1	1846	728.000	725.000	21	14.69	6.806
1	1847	153.000	718.000	22	15.41	6.491
1	1848	418.000	706.000	23	16.12	6.204
1	1849	457.000	692.000	24	16.83	5.941
1	1850	1395.000	671.000	25	17.55	5.699
1	1851	264.000	600.000	26	18.26	5.477
1	1852	540.000	572.000	27	18.97	5.271
1	1853	718.000	555.000	28	19.69	5.080
1	1854	600.000	548.000	29	20.40	4.902
1	1855	1515.000	548.000	30	21.11	4.736
1	1856	399.000	545.000	31	21.83	4.582
1	1857	122.000	540.000	32	22.54	4.437
1	1858	198.000	539.000	33	23.25	4.301
1	1859	125.000	539.000	34	23.97	4.173
1	1860	555.000	525.000	35	24.68	4.052
1	1861	432.000	518.000	36	25.39	3.938
1	1862	353.000	515.000	37	26.11	3.831
1	1863	100.000	512.000	38	26.82	3.729
1	1864	187.000	511.000	39	27.53	3.632
1	1865	285.000	503.000	40	28.25	3.540
1	1866	252.000	500.000	41	28.96	3.453
1	1867	375.000	475.000	42	29.67	3.370
1	1868	548.000	469.000	43	30.39	3.291
1	1869	295.000	457.000	44	31.10	3.216
1	1870	346.000	455.000	45	31.81	3.143
1	1871	1120.000	455.000	46	32.52	3.075
1	1872	370.000	450.000	47	33.24	3.009
1	1873	213.000	448.000	48	33.95	2.945

Continuation of TABLE A.6

Listing of Data for
STATION NAME=POLISH RIVER WARTA AT POZNAN

MONTH	YEAR	DATA	ORDERED	RANK	PROB.	RET. PERIOD
(1)	(2)	(3) (CMS)	(4) (CMS)	(5)	(6) (%)	(7) (YEARS)
1	1922	301.000	257.000	97	68.90	1.451
1	1923	518.000	255.000	98	69.61	1.436
1	1924	1573.000	252.000	99	70.33	1.422
1	1925	199.000	250.000	100	71.04	1.408
1	1926	413.000	247.000	101	71.75	1.394
1	1927	545.000	241.000	102	72.47	1.380
1	1928	539.000	214.000	103	73.18	1.366
1	1929	299.000	214.000	104	73.89	1.353
1	1930	109.000	213.000	105	74.61	1.340
1	1931	362.000	207.000	106	75.32	1.328
1	1932	192.000	207.000	107	76.03	1.315
1	1933	168.000	206.000	108	76.75	1.303
1	1934	116.000	205.000	109	77.46	1.291
1	1935	250.000	202.000	110	78.17	1.279
1	1936	150.000	200.000	111	78.89	1.268
1	1937	330.000	199.000	112	79.60	1.256
1	1938	450.000	199.000	113	80.31	1.245
1	1939	292.000	198.000	114	81.03	1.234
1	1940	990.000	195.000	115	81.74	1.223
1	1941	750.000	193.000	116	82.45	1.213
1	1942	671.000	192.000	117	83.17	1.202
1	1943	148.000	191.000	118	83.88	1.192
1	1944	195.000	187.000	119	84.59	1.182
1	1945	455.000	185.000	120	85.31	1.172
1	1946	386.000	185.000	121	86.02	1.163
1	1947	1055.000	182.000	122	86.73	1.153
1	1948	511.000	182.000	123	87.45	1.144
1	1949	255.000	168.000	124	88.16	1.134
1	1950	193.000	153.000	125	88.87	1.125
1	1951	150.000	150.000	126	89.59	1.116
1	1952	148.000	150.000	127	90.30	1.107
1	1953	706.000	148.000	128	91.01	1.099
1	1954	207.000	148.000	129	91.73	1.090
1	1955	182.000	141.000	130	92.44	1.082
1	1956	214.000	134.000	131	93.15	1.074
1	1957	299.000	125.000	132	93.87	1.065
1	1958	437.000	125.000	133	94.58	1.057
1	1959	185.000	122.000	134	95.29	1.049
1	1960	182.000	118.000	135	96.01	1.042
1	1961	293.000	116.000	136	96.72	1.034
1	1962	350.000	109.000	137	97.43	1.026
1	1963	350.000	105.000	138	98.15	1.019
1	1964	305.000	100.000	139	98.86	1.012
1	1965	469.000	89.000	140	99.57	1.004

Continuation of TABLE A.6

Listing of Data for
STATION NAME=POLISH RIVER WARTA AT POZNAN

MONTH	YEAR	DATA	ORDERED	RANK	PROB.	RET. PERIOD
(1)	(2)	(3) (CMS)	(4) (CMS)	(5)	(6) (%)	(7) (YEARS)
1	1874	275.000	441.000	49	34.66	2.885
1	1875	280.000	437.000	50	35.38	2.827
1	1876	1105.000	432.000	51	36.09	2.771
1	1877	365.000	431.000	52	36.80	2.717
1	1878	380.000	418.000	53	37.52	2.665
1	1879	333.000	418.000	54	38.23	2.616
1	1880	572.000	413.000	55	38.94	2.568
1	1881	475.000	399.000	56	39.66	2.522
1	1882	134.000	393.000	57	40.37	2.477
1	1883	269.000	386.000	58	41.08	2.434
1	1884	259.000	382.000	59	41.80	2.392
1	1885	200.000	380.000	60	42.51	2.352
1	1886	775.000	376.000	61	43.22	2.314
1	1887	260.000	375.000	62	43.94	2.276
1	1888	1510.000	370.000	63	44.65	2.240
1	1889	1460.000	365.000	64	45.36	2.204
1	1890	312.000	362.000	65	46.08	2.170
1	1891	1162.000	362.000	66	46.79	2.137
1	1892	515.000	353.000	67	47.50	2.105
1	1893	500.000	350.000	68	48.22	2.074
1	1894	335.000	350.000	69	48.93	2.044
1	1895	728.000	346.000	70	49.64	2.014
1	1896	199.000	335.000	71	50.36	1.986
1	1897	503.000	333.000	72	51.07	1.958
1	1898	247.000	330.000	73	51.78	1.931
1	1899	257.000	319.000	74	52.50	1.905
1	1900	418.000	312.000	75	53.21	1.879
1	1901	448.000	305.000	76	53.92	1.854
1	1902	441.000	303.000	77	54.64	1.830
1	1903	297.000	301.000	78	55.35	1.807
1	1904	297.000	300.000	79	56.06	1.784
1	1905	89.000	299.000	80	56.78	1.761
1	1906	214.000	299.000	81	57.49	1.739
1	1907	362.000	297.000	82	58.20	1.718
1	1908	455.000	297.000	83	58.92	1.697
1	1909	1253.000	295.000	84	59.63	1.677
1	1910	185.000	294.000	85	60.34	1.657
1	1911	539.000	293.000	86	61.06	1.638
1	1912	294.000	292.000	87	61.77	1.619
1	1913	205.000	287.000	88	62.48	1.600
1	1914	287.000	285.000	89	63.20	1.582
1	1915	202.000	280.000	90	63.91	1.565
1	1916	431.000	275.000	91	64.62	1.547
1	1917	1073.000	275.000	92	65.34	1.531
1	1918	376.000	269.000	93	66.05	1.514
1	1919	141.000	264.000	94	66.76	1.498
1	1920	770.000	260.000	95	67.48	1.482
1	1921	241.000	259.000	96	68.19	1.467

TABLE A.7

FREQUENCY ANALYSIS - THREE-PARAMETER LOGNORMAL DISTRIBUTION
POLISH RIVER WARTA AT POZNAN

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	436.457	319.690	0.732	1.770	6.046
LN X SERIES	5.863	0.646	0.110	0.272	2.759
LN(X-c) SERIES	5.654	0.787	0.139	-0.067	2.870

X(MIN)= 89.000
X(MAX)= 1573.000
LOWER OUTLIER LIMIT OF X= 46.572

TOTAL SAMPLE SIZE= 140
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

3LN PARAMETERS: $c = 52.042$ $\mu_z = 5.654$ $\sigma_z = 0.787$

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	84.90
1.050	0.952	129.00
1.250	0.800	199.00
2.000	0.500	337.00
5.000	0.200	605.00
10.000	0.100	834.00
25.000	0.040	1180.00
50.000	0.020	1490.00
100.000	0.010	1830.00
200.000	0.005	2220.00
500.000	0.002	2800.00

TABLE A.8

FREQUENCY ANALYSIS - GENERALIZED EXTREME VALUE DISTRIBUTION
POLISH RIVER WARTA AT POZNAN

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	436.457	319.690	0.732	1.770	6.046
LN X SERIES	5.863	0.646	0.110	0.272	2.759

X(MIN)= 89.000
X(MAX)= 1573.000
LOWER OUTLIER LIMIT OF X= 46.572

TOTAL SAMPLE SIZE= 140
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

GEV PARAMETERS: C = 273.61 A= 157.087 K= -0.361

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	69.00
1.050	0.952	130.00
1.250	0.800	205.00
2.000	0.500	335.00
5.000	0.200	586.00
10.000	0.100	819.00
25.000	0.040	1220.00
50.000	0.020	1620.00
100.000	0.010	2130.00
200.000	0.005	2780.00
500.000	0.002	3940.00

TABLE A.9

FREQUENCY ANALYSIS - LOG PEARSON TYPE III DISTRIBUTION
POLISH RIVER WARTA AT POZNAN

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	436.457	319.690	0.732	1.770	6.046
X SERIES	5.863	0.646	0.110	0.272	2.759

X(MIN)= 89.000
X(MAX)= 1573.000
LOWER OUTLIER LIMIT OF X= 46.572

TOTAL SAMPLE SIZE= 140
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

LP3 PARAMETERS: A= 0.1223 B= 27.91 LOG(c)= 2.449
c = 11.57

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	77.40
1.050	0.952	129.00
1.250	0.800	203.00
2.000	0.500	338.00
5.000	0.200	596.00
10.000	0.100	822.00
25.000	0.040	1180.00
50.000	0.020	1510.00
100.000	0.010	1890.00
200.000	0.005	2340.00
500.000	0.002	3050.00

TABLE A.10

FREQUENCY ANALYSIS - WAKEBY DISTRIBUTION
POLISH RIVER WARTA AT POZNAN

SAMPLE STATISTICS

	MEAN	S.D.	C.V.	C.S.	C.K.
X SERIES	436.457	319.690	0.732	1.770	6.046
LN X SERIES	5.863	0.646	0.110	0.272	2.759

X(MIN)= 89.000
X(MAX)= 1573.000
LOWER OUTLIER LIMIT OF X= 46.572

TOTAL SAMPLE SIZE= 140
NO. OF LOW OUTLIERS= 0
NO. OF ZERO FLOWS= 0

THE FOLLOWING WAKEBY PARAMETERS WERE OBTAINED BY ASSUMING E TO BE ZERO. THE ITERATION ALGORITHM WAS NOT REQUIRED.

E = 0.000 A= 141.096 B= 44.11 C= 2882.565 D= 0.094

FLOOD FREQUENCY REGIME

RETURN PERIOD	EXCEEDANCE PROBABILITY	FLOOD
1.003	0.997	18.30
1.050	0.952	138.00
1.250	0.800	202.00
2.000	0.500	335.00
5.000	0.200	611.00
10.000	0.100	836.00
25.000	0.040	1160.00
50.000	0.020	1420.00
100.000	0.010	1700.00
200.000	0.005	2000.00
500.000	0.002	2420.00

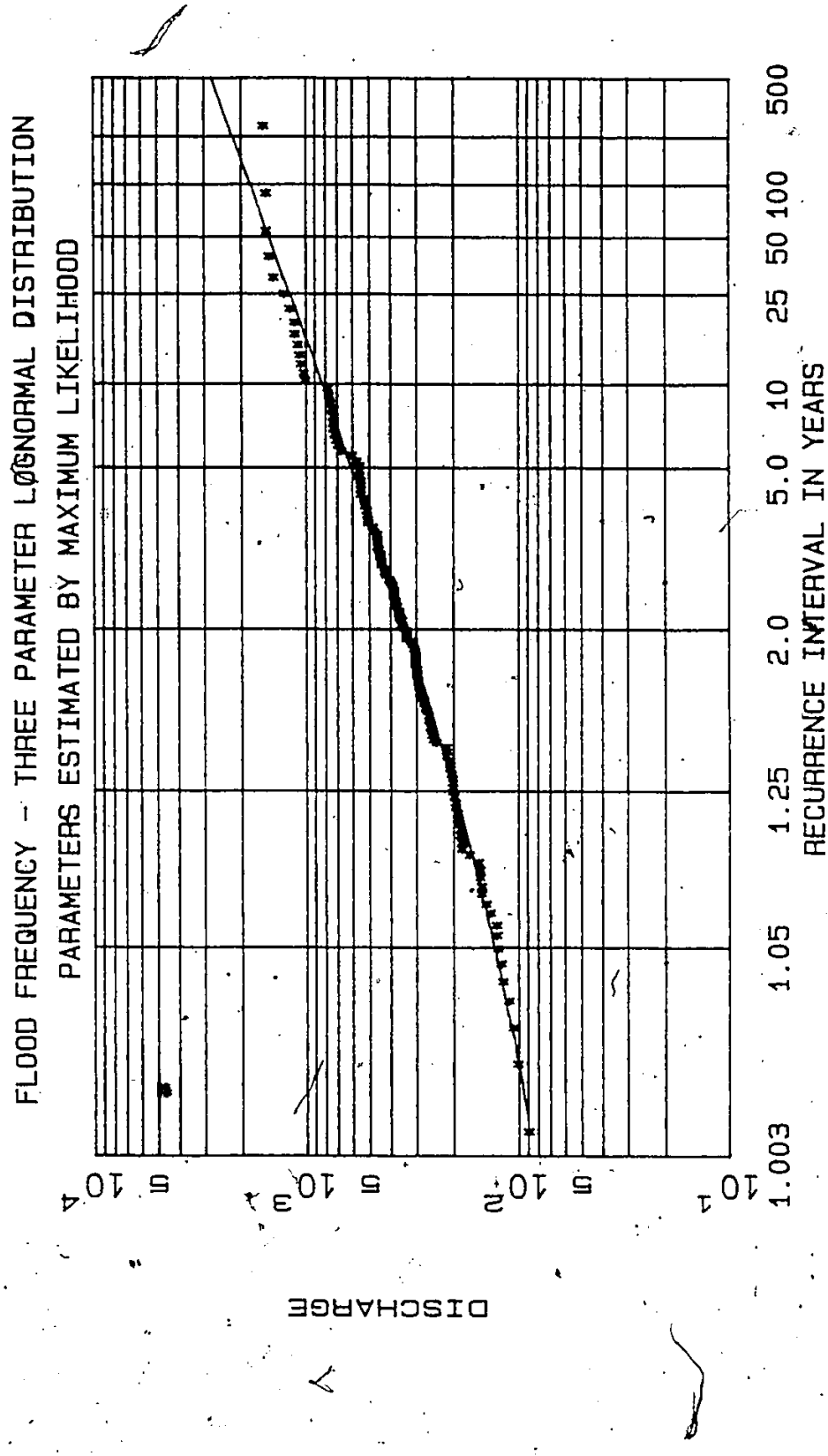

POLISH RIVER WARTA AT POZNAN
FLOOD FREQUENCY - THREE PARAMETER LOGNORMAL DISTRIBUTION
PARAMETERS ESTIMATED BY MAXIMUM LIKELIHOOD
DISCHARGE
5 10⁴
5 10³
5 10²
10¹
1.003
1.05
1.25
2.0
5.0
10
25
50
100
500
RECURRENCE INTERVAL IN YEARS

FIG. A.5

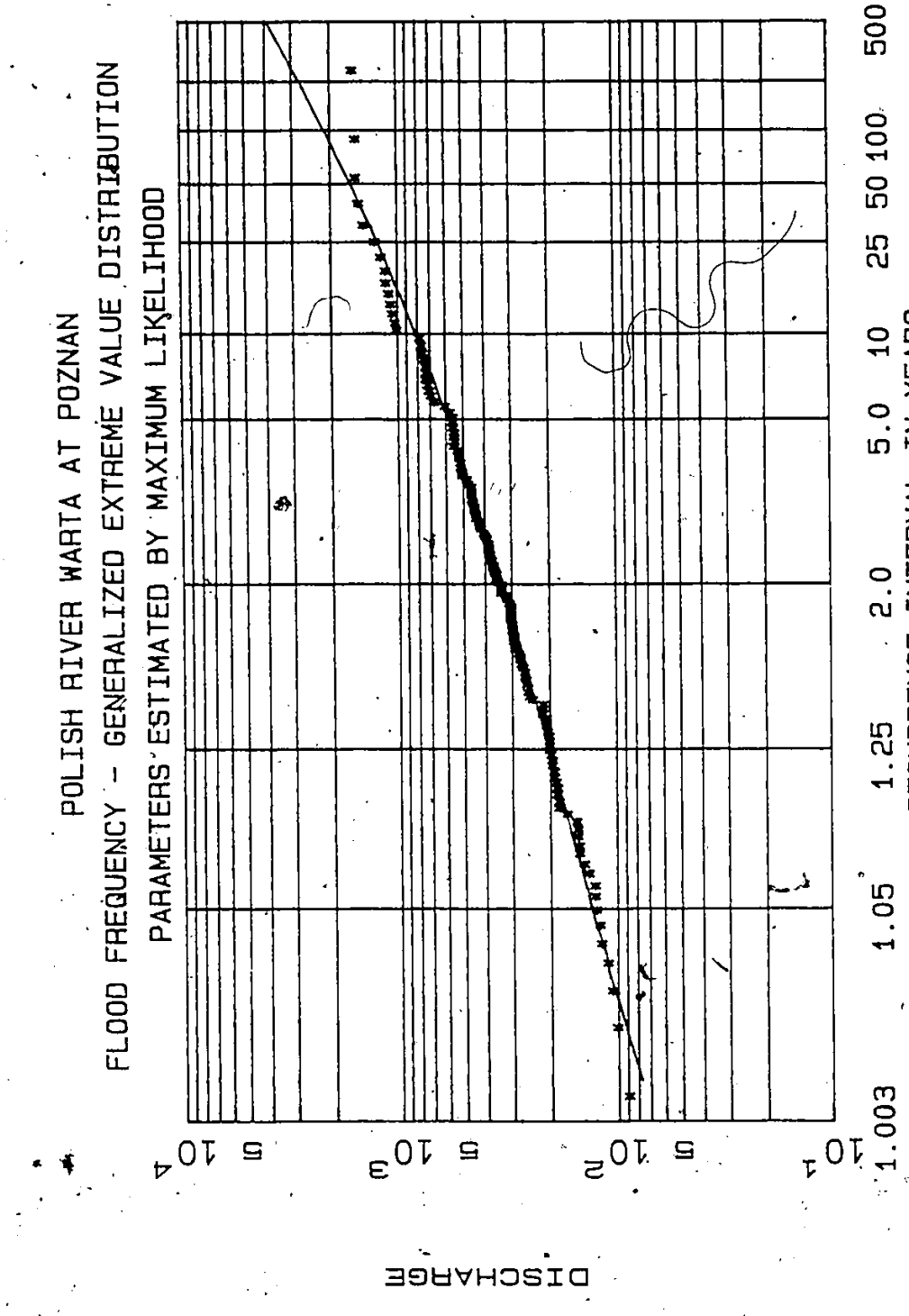

POLISH RIVER WARTA AT POZNAN
FLOOD FREQUENCY - GENERALIZED EXTREME VALUE DISTRIBUTION
PARAMETERS ESTIMATED BY MAXIMUM LIKELIHOOD
DISCHARGE
5 10^4
5 10^3
5 10^2
10^1
1.003
1.05
1.25
2.0
5.0
10
25
50
100
500
RECURRENCE INTERVAL IN YEARS

FIG. A.6

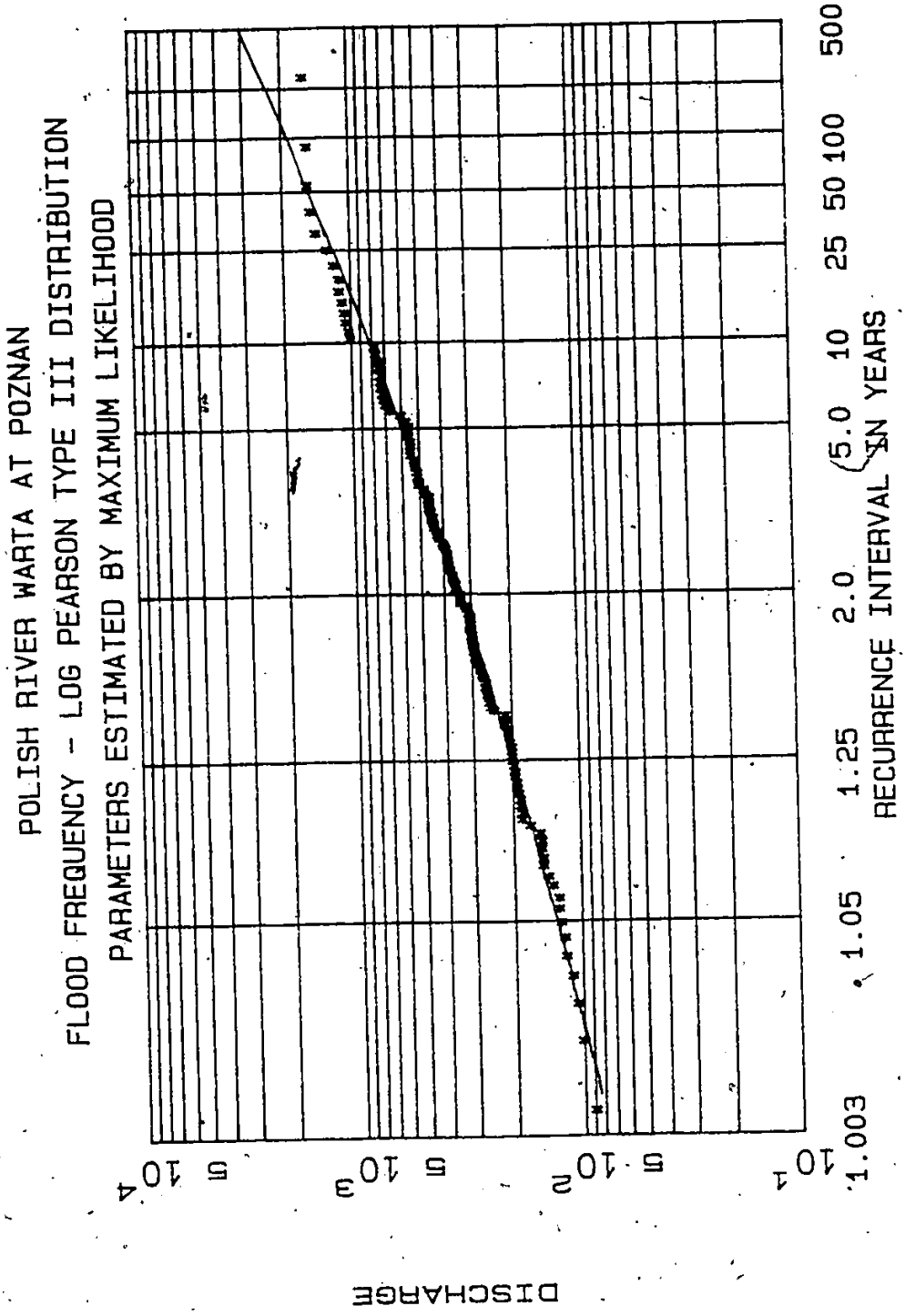

POLISH RIVER WARTA AT POZNAN
FLOOD FREQUENCY - LOG PEARSON TYPE III DISTRIBUTION
PARAMETERS ESTIMATED BY MAXIMUM LIKELIHOOD
DISCHARGE
5 10⁴
5 10³
5 10²
10⁴
1.003
1.05
1.25
2.0
5.0
10
25
50
100
500
RECURRENCE INTERVAL IN YEARS

FIG. A.7

POLISH RIVER WARTA AT POZNAN
FLOOD FREQUENCY - WAKEBY DISTRIBUTION

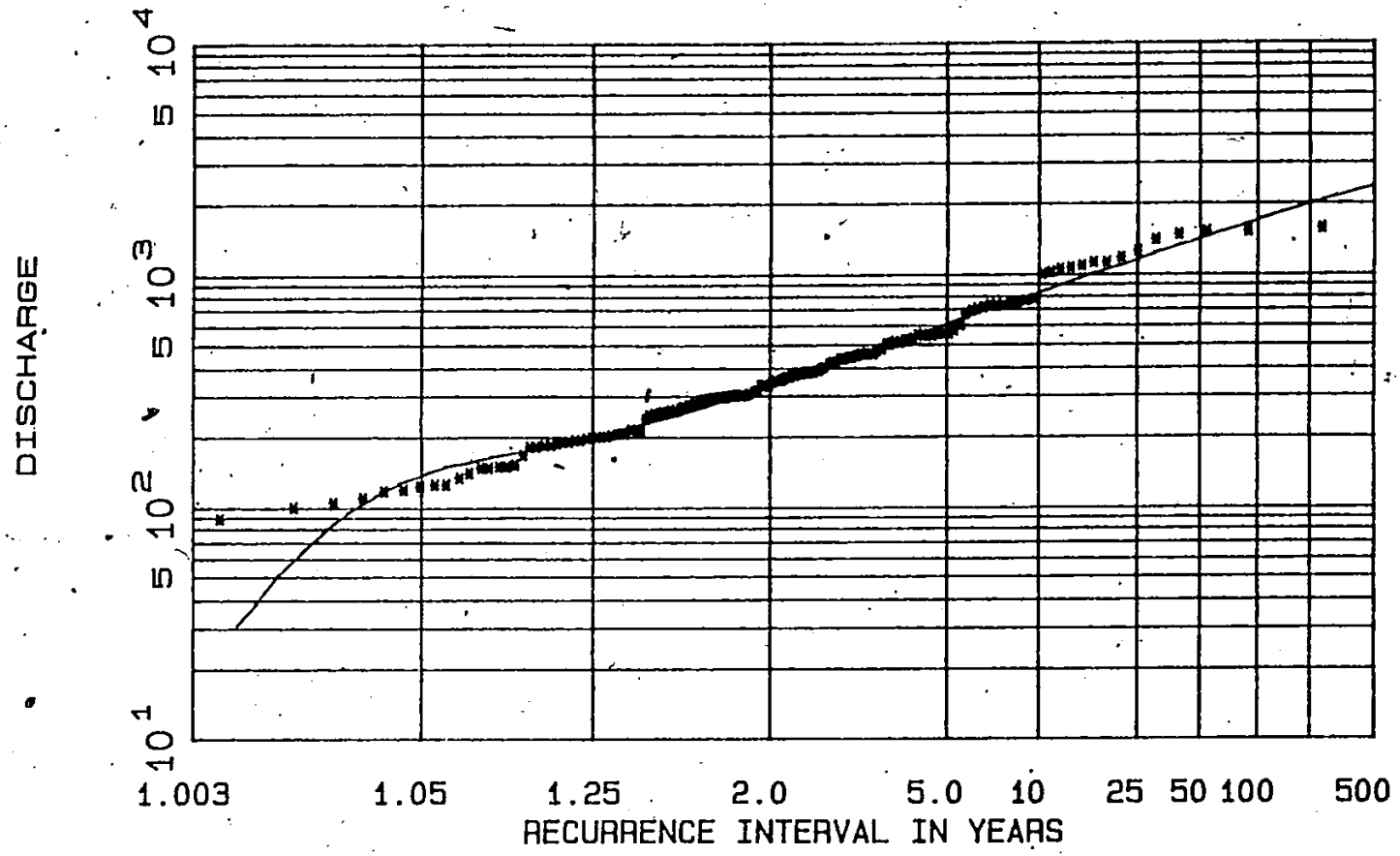


FIG. A.8

APPENDIX B

Computer Programs

```
      ADAMOWSKI/FELUCH NUMERICAL PROCEDURE - RECTANGULAR KERNEL
C
      PROGRAMMERS: DR. W. FELUCH & C. LABATIUK - 1984
C
C     THIS PROGRAM ACCEPTS DATA POINTS AND COMPUTES FLOOD ESTIMATES FOR
C     10,50,100 AND 200 YEAR RETURN PERIODS USING NUMERICAL PROCEDURE
C     SUGGESTED BY DRS. ADAMOWSKI AND FELUCH. THE NUMERICAL PROCEDURE IS
C     BASED ON THE ROSENBLATT/PARZEN ESTIMATOR AND RECTANGULAR KERNEL.
C     THE WINDOW WIDTH IS DETERMINED BY MINIMIZING A MEAN-SQUARE ERROR
C     EXPRESSION WHICH MAKES USE OF AN EMPIRICAL PLOTTING FORMULA. ON
C     OPTION, THE GRAPH OF THE DENSITY FUNCTION IS PRINTED.
C
C     INPUT EXPECTED
C     (1)THE STATION INVOLVED (CHARACTER VARIABLE, 16A4 FORMAT)
C     (2)RECORD LENGTH,KPP (INTEGER VARIABLES, I3,I2 FORMAT)
C     (3)ANNUAL MAXIMUM FLOOD DATA (REAL VARIABLES, 7F10.3 FORMAT)
C     (4)EPS,RED, (REAL VARIABLES, 2F10.5 FORMAT)
C
C     IMPORTANT VARIABLES:
C        X=INPUT VECTOR OF LENGTH N CONTAINING THE FLOOD DATA;
C        N=RECORD LENGTH INTEGER;
C        EM=ESTIMATE MEAN;
C        ES=ESTIMATE STANDARD DEVIATION;
C        A,B=INITIAL LIMITS FOR METHOD OF BISECTION;
C        EPS=ACCURACY LEVEL, METHOD OF BISECTION;
C        RED=NUMERICAL REDUCTOR, PERMITS BETTER COMPUTATIONAL ACCURACY;
C        B=VECTOR CONTAINING THE ABSCISSAE AT WHICH DENSITY ESTIMATES
C          ARE DESIRED;
C        F=VECTOR CONTAINING THE ESTIMATES OF THE DENSITY AT THE
C          POINTS SPECIFIED IN B;
C        H=THE WINDOW WIDTH;
C        T=RETURN PERIOD;
C        XP=FLOOD ESTIMATE;
C        KEY=CONTROL INTEGER, LETS POLO SOLVE MORE THAN ONE EQUATION;
C        KPR=OUTPUT CONTROL, IF KPR=-1, ONLY T & XP ARE PRINTED
C                            IF KPR=1, DENSITY FUNCTION IS ALSO PLOTTED.
C
C     SUBPROGRAMS:
C        P     -THIS FUNCTION COMPUTES THE PROBABILITY P(XJ);
C        FCN3  -THIS FUNCTION IS EQUATED TO ZERO AND SOLVED FOR XP;
C        POLO  -THIS FUNCTION COMPUTES ROOT BY METHOD OF BISECTION;
C        FCN   -THIS FUNCTION LETS POLO SOLVE MORE THAN ONE EQUATION;
C        OUT   -THIS SUBROUTINE PRINTS FLOOD DATA IN SEVEN COLUMNS;
C        PEM   -THIS FUNCTION COMPUTES EMPIRICAL PROBABILITY PJ;
C        SORT  -THIS SUBROUTINE SORTS THE DATA AND ASSIGNS EACH
C               OBSERVATION AN EMPIRICAL PROBABILITY;
C        FCN1  -THIS FUNCTION IS EQUATED TO ZERO AND SOLVED FOR H;
C        FF    -THIS FUNCTION COMPUTES THE DENSITY FUNCTION;
C        C     -THIS FUNCTION GIVES EXCEEDANCE PROBABILITY;
C        DC    -THIS FUNCTION IS DC/DH;
C        RK    -THIS FUNCTION GIVES EQUATION FOR RECTANGULAR KERNEL.
C
      DOUBLE PRECISION SUM1,SUM2,SUM3
      DIMENSION TITLE(16),T(4),ITAB(50),F(101),BEE(101)
      COMMON /EST/XOBS(200),N,FI/PAT/KEY,PALFA,H,XP/TIT/EPS
```



```
      DATA T/10.,50.,100.,200./,ISPAC/1H /,IS/1H*/,ILEVEL/1HI/
      DATA ZERO,HALF,THREE,SIX,TWELVE/0.,0.5,3.,6.,12./
C
C     READ, SORT, REDUCE AND PRINT DATA
C
      READ(5,10)(TITLE(I),I=1,16)
      WRITE(6,20)(TITLE(I),I=1,16)
      READ(5,71)N,KPR
      READ(5,72)(XOBS(I),I=1,N)
      CALL OUT(N,XOBS)
      NP=2
      WRITE(6,40)NP
      CALL SORT(N,XOBS)
      READ(5,77)EPS,RED
      XMAX=-99999999.
      XMIN=99999999.
      DO 90 I=1,N
         X=XOBS(I)
         IF(X.GT.XMAX)XMAX=X
         IF(X.LT.XMIN)XMIN=X
   90 CONTINUE
      A=10.E-9
      B=3.*XMAX
      ID=0
      WRITE(6,23)A,B,EPS,RED,KPR
C
C     CALCULATE AND PRINT H,T,XP
C
      KEY=1
      BOUND=XMAX
      H=POLO(1.E-08001,BOUND,EPS)
      WRITE(6,22)H
      WRITE(6,24)
      DO 4 I=1,4
         FI=1./T(I)
         KE=3
         AA=A
         BB=B
         XP=POLO(AA,BB,EPS)
         WRITE(6,29)T(I),XP
    4 CONTINUE
      WRITE(6,26)
  909 IF(KPR.LE.-1)GO TO 9005
      WRITE(6,24)
C
C     GENERATE AND PRINT DENSITY FUNCTION GRAPH
C
      SCALE=1.0
      XMIN=XMIN/RED
      DO 5 I=1,50
         ITAB(I)=ISPAC
    5 CONTINUE
      DO 80 I=1,N
         XOBS(I)=XOBS(I)/RED
   80 CONTINUE
```

```
      H=H/RED
      DO 2 I=1,50
         XP=3.19-FLOAT(I-1)*0.072
         BEE(I)=XP
         FXP=FF(N,XP,XOBS,H)
         E(I)=FXP
         FFI=1.0-P(N,XP,XOBS,H)
         YY1=FXP*50.*XMIN*FLOAT(N)/1.50+0.5
         NY1=INT(YY1)
         DO 198 L=1,NY1
            M=51-L
            ITAB(M)=IS
  198    CONTINUE
  254    WRITE(6,9)ILEVEL,(ITAB(L),L=1,50),ILEVEL,XP,FXP,FFI
         DO 7 L=1,50
            ITAB(L)=ISPAC
    7    CONTINUE
    2 CONTINUE
      WRITE(6,32)
      WRITE(6,26)
      SUM1 = ZERO
      SUM2 = ZERO
      SUM3 = ZERO
      DO 150 K=1,50
         FK = F(K)
         BK = BEE(K)
         SUM3 = SUM3+FK
         SUM1 = SUM1+FK*BK
         SUM2 = SUM2+FK*BK*BK
  150 CONTINUE
      EM=SUM1/SUM3
      WRITE(6,45)EM
      VAR=(SUM2-SUM3*EM**2)/(SUM3-1.)
      ES=SQRT(VAR)
      WRITE(6,46)ES
 9005 STOP
    9 FORMAT(' ',A1,50A1,A1,F7.3,1X,F8.6,1X,F6.4)
   10 FORMAT(16A4)
   20 FORMAT(1H ,///4X,68(1H*)/4X,2H* ,16A4,2H *,/4X,68(1H*)///7H XOBS :
     *)
   23 FORMAT(///,' RESULTS FOR RECTANGULAR KERNEL: ',//4H A =,F9.3,3X,4H
     * B =,F12.3,3X,6H EPS =,F10.8/21H NUMERICAL REDUCTOR =,F11.3,3X,16H
     * KEY OF OUTPUT =,I4)
   22 FORMAT(/4H H =,E14.7///)
   24 FORMAT(1H ,51X,'ABSCISSA',1X,'ORDINATE'/,1X,52(1H-),1X,'FLOOD',2X,
     *'*DENSITY',1X,'DIST.')
   26 FORMAT(////)
   28 FORMAT(1H /,5X,15HRETURN    FLOOD,/,5X,17HPERIOD   ESTIMATE/)
   29 FORMAT(2X,F9.3,3X,F7.0)
   32 FORMAT(1X,52(1H-))
   40 FORMAT(/' NUMBER OF EMPIRICAL PROBABILITY EQUATION =',I2/)
   45 FORMAT(' ',///,' ESTIMATE MEAN     =',F10.2)
   46 FORMAT(' ',('ESTIMATE STAN.DEV.=',F10.2,/))
   71 FORMAT(I3,I2)
   72 FORMAT(8F8.1)
```



```
   77 FORMAT(2F10.5)
      END

C     P SUBPROGRAM--
C     THIS FUNCTION COMPUTES THE PROBABILITY P(XJ)

      FUNCTION P(N,X,XOBS,HN)
      DIMENSION XOBS(N)
      PI=3.1415926535
      S=0.
      FN=FLOAT(N)
      DO 3 I=1,N
         S=S+C(I,X,XOBS,HN)
    3 CONTINUE
      P=S/FN/HN
      RETURN
      END

C     PEM SUBPROGRAM--
C     THIS FUNCTION COMPUTES EMPIRICAL PROBABILITY PJ

      FUNCTION PEM(N,I)
      FI=FLOAT(I)
      FN=FLOAT(N)
      NP=2
      GO TO (1,2,3,4,5,6) ,NP
    1 PEM=FI/(FN+1.)
      RETURN
    2 PEM=(FI-0.25)/(FN+0.5)
      RETURN
    3 F=FN+1.
      S=SQRT(F*FN)
      PEM=(2.*FI-F+S)/S/2.
      RETURN
    4 PEM=(FI-0.3)/(FN+0.4)
      RETURN
    5 PEM=(FI-0.4)/(FN+0.2)
      RETURN
    6 PEM=(FI-0.5)/FN
      RETURN
      END

C     SORT SUBPROGRAM--
C     THIS SUBROUTINE SORTS THE DATA FROM HIGHEST TO LOWEST AND ASSIGNS
C     EACH OBSERVATION AN EMPIRICAL PROBABILITY USING PEM.

      SUBROUTINE SORT(N,A)
      DIMENSION A(N)
      N1=N-1
      DO 2 I=1,N1
         I1=I+1
         AI=A(I)
         ID=0
         DO 2 J=I1,N
            AJ=A(J)
```

```
            IF(AI.GT.AJ.AND.ID.EQ.0) GO TO 4
            A(J)=AI
            A(I)=AJ
            AI=AJ
    4     CONTINUE
    2   CONTINUE
        N3=N-3
        N2=N-2
        N1=N-1
        WRITE(6,23)
        DO 12 I=1,N,4
           I4=I+3
           I3=I+2
           I2=I+1
           IF(I.GT.N3) GO TO 33
           P1=PEM(N,I)
           P2=PEM(N,I2)
           P3=PEM(N,I3)
           P4=PEM(N,I4)
           WRITE(6,22)P1,A(I),P2,A(I2),P3,A(I3),P4,A(I4)
           GO TO 5
   33      IF(I.GT.N2) GO TO 24
           P1=PEM(N,I)
           P2=PEM(N,I2)
           P3=PEM(N,I3)
           WRITE(6,22)P1,A(I),P2,A(I2),P3,A(I3)
           GO TO 5
   24      IF(I.GT.N1) GO TO 13
           P1=PEM(N,I)
           P2=PEM(N,I2)
           WRITE(6,22)P1,A(I),P2,A(I2)
           GO TO 5
   13      P1=PEM(N,I)
           WRITE(6,22)P1,A(I)
    5      CONTINUE
   12   CONTINUE
   22   FORMAT(1H ,4(F7.4,1X,F8.4,3X))
   23   FORMAT(1H ,/17H EMPIRICAL DATA : ,///4(4X,1HP,7X,4HXOBS,3X)/)
        RETURN
        END
C
C       FCN1 SUBPROGRAM--
C       THIS FUNCTION IS EQUATED TO ZERO AND SOLVED FOR H USING POLO
C
        FUNCTION FCN1(H)
        COMMON /EST/XOBS(200),N,FI
        SJ=0.
        DO 4 J=1,N
           SI=0.
           XJ=XOBS(J)
           DO 5 I=1,N
              SI=SI-C(I,XJ,XOBS,H)/H+DC(I,XJ,XOBS,H)
    5      CONTINUE
           SJ=SJ+(PEM(N,J)-P(N,XJ,XOBS,H))*SI
    4   CONTINUE
```



```
        FCN1=SJ
        RETURN
        END
C
C       FF SUBPROGRAM--
C       THIS FUNCTION COMPUTES DENSITY FUNCTION
C
        FUNCTION FF(N,X,XOBS,HN)
        DIMENSION XOBS(N)
        S=0.
        FN=FLOAT(N)
        DO 3 I=1,N
           S=S+RK(I,X,XOBS,HN)
    3   CONTINUE
        FF=S/FN/HN
        RETURN
        END
C
C       C SUBPROGRAM--
C       THIS FUNCTION GIVES EXCEEDANCE PROBABILITY
C
        FUNCTION C(I,X,XOBS,HN)
        DIMENSION XOBS(I)
        Z=XOBS(I)
        D=Z-HN
        S=Z+HN
        IF(X.LE.D) C=HN
        IF(X.GT.D.AND.X.LT.S) C=(S-X)/2.
        IF(X.GE.S) C=0.
        RETURN
        END
C
C       DC SUBPROGRAM--
C       THIS FUNCTION IS DC/DH
C
        FUNCTION DC(I,X,XOBS,HN)
        DIMENSION XOBS(I)
        Z=XOBS(I)
        D=Z-HN
        S=Z+HN
        IF(X.LE.D) DC=1.
        IF(X.GT.D.AND.X.LT.S) DC=0.5
        RETURN
        END
C
C       RK SUBPROGRAM--
C       THIS FUNCTION GIVES EQUATION FOR RECTANGULAR KERNEL
C
        FUNCTION RK(I,X,XOBS,HN)
        DIMENSION XOBS(I)
        Z=XOBS(I)
        D=Z-HN
        S=Z+HN
        IF(X.LE.D) RK=0.
        IF(X.GT.D.AND.X.LT.S) RK=0.5
```

```
      IF(X.GT.5) RK=3.
      RETURN
      END

C
C     FCN3 SUBPROGRAM--
C     FOR A GIVEN RETURN PERIOD, THIS FUNCTION IS EQUATED TO ZERO AND
C     SOLVED FOR XP USING POLO.
C

      FUNCTION FCN3(X)
      COMMON /EST/XOBS(200),N,FI/PAT/KEY,PALFA,H
      FCN3=P(N,X,XOBS,H)-FI
      RETURN
      END
C
C     POLO SUBPROGRAM--
C     THIS FUNCTION COMPUTES ROOT BY METHOD OF BISECTION.
C

      FUNCTION POLO(A,B,EPS)
      FA=FCN(A)
 10   X=(A+B)/2.0
      FM=FCN(X)
      FAMA=FA*FM
      IF(FAMA) 2,3,4
 2    B=X
      GO TO 5
 4    FA=FM
      A=X
 5    IF(ABS(A-B).GT.EPS) GO TO 10
 3    POLO=(A+B)/2.
      RETURN
      END
C
C     FCN SUBPROGRAM--
C     THIS FUNCTION ALLOWS POLO TO SOLVE BOTH FCN3 (KEY=3) AND
C     FCN1 (KEY=1).
C

      FUNCTION FCN(X)
      COMMON /PAT/ KEY,PALFA,H
      IF(KEY.EQ.1) FCN=FCN1(X)
      IF(KEY.EQ.3) FCN=FCN3(X)
      RETURN
      END
C
C     OUT SUBPROGRAM--
C     THIS SUBROUTINE PRINTS FLOOD DATA IN SEVEN COLUMNS.
C

      SUBROUTINE OUT(N,Y)
      DIMENSION Y(N)
      N6=N-6
      N5=N-5
      N4=N-4
      N3=N-3
      N2=N-2
      N1=N-1
      DO 12 I=1,N,7
```



```
      IF(I.GT.N5) GO TO 6
      WRITE(6,22) Y(I),Y(I+1),Y(I+2),Y(I+3),Y(I+4),Y(I+5),Y(I+6)
      GO TO 14
 6    IF(I.GT.N5) GO TO 5
      WRITE(6,22) Y(I),Y(I+1),Y(I+2),Y(I+3),Y(I+4),Y(I+5)
      GO TO 14
 5    IF(I.GT.N4) GO TO 4
      WRITE(6,22) Y(I),Y(I+1),Y(I+2),Y(I+3),Y(I+4)
      GO TO 14
 4    IF(I.GT.N3) GO TO 3
      WRITE(6,22) Y(I),Y(I+1),Y(I+2),Y(I+3)
      GO TO 14
 3    IF (I.GT.N2) GO TO 2
      WRITE(6,22) Y(I),Y(I+1),Y(I+2)
      GO TO 14
 2    IF(I.GT.N1) GO TO 1
      WRITE(6,22) Y(I),Y(I+1)
      GO TO 14
 1    WRITE(6,22) Y(I)
 14   CONTINUE
 12   CONTINUE
 22   FORMAT(1H ,7(2X,F8.4))
      RETURN
      END
```

```
C     KERNEL DENSITY ESTIMATOR  - H BY METHOD OF SCOTT, TAPIA & THOMPSON
C
C     PROGRAMMER: C. LABATIUK CVG MASTER'S THESIS-1984
C     ALGORITHM FOR DETERMINING H SUPPLIED BY DR. SCOTT, RICE UNIVERSITY
C
C     THIS PROGRAM ACCEPTS DATA POINTS AND COMPUTES AN APPROXIMATION
C     OF THE PROBABILITY DENSITY FUNCTION USING THE KERNEL METHOD WITH A
C     NORMAL KERNEL (K=EXP(-Y**Y/2.)/2.506628).  USE IS MADE OF IMSL
C     ROUTINE NDKER AND THE BANDWIDTH IS CALCULATED BY THE METHOD OF
C     SCOTT, TAPIA & THOMPSON (STT).  DATA INPUTTED IS ARRANGED IN
C     INCREASING ORDER AND SCALED DOWN (OR UP), IF NECESSARY; DENSITY
C     ESTIMATIONS ARE CALCULATED; AND THE GRAPH OF THE DENSITY FUNCTION
C     IS PRINTED.  THE DISTRIBUTION FUNCTION IS CALCULATED USING
C     SIMPSON'S 1/3 RULE AND USED IN FLOOD FREQUENCY ANALYSIS.  FLOODS
C     ARE ESTIMATED FOR 10,50,100, AND 200 YEAR RETURN PERIODS.
C
C     INPUT EXPECTED
C     (1)THE UNITS INVOLVED (CHARACTER VARIABLE <= 12 LETTERS)
C     (2)RECORD LENGTH (INTEGER VARIABLE, I3 FORMAT)
C     (3)ANNUAL MAXIMUM FLOOD DATA (REAL VARIABLES, 8F8.1 FORMAT)
C     (4)INITIAL CHOICE FOR H (DOUBLE PRECISION VARIABLE, D15.8 FORMAT)
C        H SHOULD BE LARGE IF NO PRIOR KNOWLEDGE.  1.0D0 USUALLY WORKS.
C
C     IMPORTANT VARIABLES:
C        X=INPUT VECTOR OF LENGTH N CONTAINING THE FLOOD DATA;
C        N=RECORD LENGTH INTEGER;
C        BWIDTH=THE BANDWIDTH (OR WINDOW WIDTH);
C        DEL=INPUT PARAMETER USED IN NDKER.  IF THE KERNEL HAS FINITE
C            SUPPORT AND IF THE USER WISHES TO USE THIS PROPERTY FOR
C            GREATER EFFICIENCY, DEL IS A POSITIVE NUMBER SUCH THAT
C     *      XKER(Y)=0.0 FOR ALL Y FOR WHICH ABS(Y).GT.DEL.  IF NO SUCH
C            DEL EXISTS OR IF THE USER DOES NOT WISH TO MAKE USE OF THIS
C            PROPERTY, DEL SHOULD BE ASSIGNED ANY NONPOSITIVE VALUE;
C        B=INPUT VECTOR USED IN NDKER OF LENGTH NPT CONTAINING THE
C            ABSCISSAE AT WHICH DENSITY ESTIMATES ARE DESIRED.  B MUST
C            BE IN ASCENDING ORDER IF DEL.GT.0.;
C        NPT=INPUT INTEGER USED IN NDKER.  THE NUMBER OF POINTS AT WHICH
C            IT IS DESIRED TO ESTIMATE THE DENSITY;
C        FF=OUTPUT VECTOR FROM NDKER OF LENGTH NPT CONTAINING THE
C            ESTIMATES OF THE DENSITY AT THE POINTS SPECIFIED IN B;
C        H=INITIAL CHOICE FOR SCALING FACTOR IN METHOD OF STT.  ON
C            RETURN FROM FUNCTION HN, H=BWIDTH;
C        XBAR=THE SAMPLE MEAN;
C        XSTD=THE SAMPLE STANDARD DEVIATION;
C        EM=ESTIMATE MEAN;
C        ES=ESTIMATE STANDARD DEVIATION;
C        XINC=THE INCREMENT BETWEEN LABELS ON THE HORIZONTAL AXIS;
C        G(,)=THE GRAPH CHARACTERS;
C        F(,)=THE NUMERICAL GRAPH COORDINATES;
C
C     SUBPROGRAMS:
C        GRAPH - THIS SUBROUTINE GENERATES THE GRAPH ITSELF;
C        HN    - THIS FUNCTION CALCULATES THE BWIDTH BY METHOD OF STT;
C        ERROR - THIS SUBROUTINE OUTPUTS VARIOUS TERMINAL ERRORS;
C        XKER  - THIS FUNCTION IS THE NORMAL KERNEL.
C
      IMPLICIT REAL*8 (A-H,O-Z)
      INTEGER N,NPT
      REAL EN,FI
      DIMENSION T(4),Y(4),FLD(4),GR(13),CUMDEN(8),
     *X(200),B(244),FF(244),UNITS(3),F(122,3)
      DATA XSUM/0.0D0/,XSUMSQ/0.0D0/,FMAX/0.0D0/,
     *T/10.D0,50.D0,100.D0,200.D0/
      DATA ZERO,HALF,THREE,SIX,TWELVE/0.D0,0.5D0,3.D0,6.D0,12.D0/
      COMMON X,F
      RND(X)=DBLE(AINT(SNGL((X*2.D0+DSIGN(1.D0,X))/2.D0)))
      ALOW(X)=DBLE(AINT(SNGL(X-(1.D0-DSIGN(1.D0,X))/2.D0)))
      EXTERNAL XKER
      CALL ERRSET(208,256,-1,1)
C
C     READ, TRANSFORM, SORT AND PRINT DATA
C
      READ(5,35)UNITS
   35 FORMAT(3A4)
      READ(5,12)N
   12 FORMAT(I3)
      READ(5,13)(X(I),I=1,N)
   13 FORMAT(8F8.1)
      DO 505 I=1,N
         XSUM=XSUM+X(I)
         XSUMSQ=XSUMSQ+X(I)**2
   20    CONTINUE
         IHIGH=I
         ILOW=1
   22    CONTINUE
         IF(IHIGH.EQ.ILOW)GO TO 50
         IHALF=(ILOW+IHIGH)/2
         IF(X(I)-X(IHALF))30,25,40
   25    CONTINUE
         INDEX=IHALF+1
         GO TO 60
   30    CONTINUE
         IHIGH=IHALF
         GO TO 22
   40    CONTINUE
         ILOW=IHALF+1
         GO TO 22
   50    CONTINUE
         INDEX=ILOW
   60    CONTINUE
         SAVE=X(I)
         ISORT=I-INDEX
         DO 510 I2=1,ISORT
            I3=I+1-I2
            X(I3)=X(I3-1)
  510    CONTINUE
         X(INDEX)=SAVE
  505 CONTINUE
      WRITE(6,110)N
  110 FORMAT('1',4X,'NUMBER OF DATA POINTS READ IN = ',I4,//)
```

```
FILE: NOR      FORTRAN  A          VM/SP CONVERSATIONAL MONITOR S
      EN=FLOAT(N)
      DEN=DBLE(EN)
      XBAR=XSUM/DEN
      XVAR=(XSUMSQ-DEN*XBAR**2)/(DEN-1.D0)
      XSTD=DSQRT(XVAR)
      XLAR=DMAX1(DABS(X(N)),DABS(X(1)))
      XSP=X(N)-X(1)
      IF((XLAR.LT.100.D0).AND.(XSP.GT.1.0D0).AND.(XLAR.GT.1.0D0))GOTO175
      DO 515 I=1,N
         X(I)=(X(I)-XBAR)/XSTD
  515 CONTINUE
      WRITE(6,136)(X(I4),I4=1,N)
      XLAR=(XLAR-XBAR)/XSTD
      WRITE(6,173)XBAR,XSTD
  173 FORMAT(///,4X,'DATA ALTERATION HAS BEEN EMPLOYED DUE TO THE',/,
     *5X,'MAGNITUDE OF THE LARGEST DATA POINT.EACH VALUE HAS BEEN',/,5X
     *,'REDUCED BY THE SAMPLE MEAN = ',F12.5,/,5X,'AND DIVIDED BY
     *THE SAMPLE STANDARD DEVIATION = ',F12.5,//)
  177 FORMAT(/12X,'SAMPLE MEAN =',F10.5/5X,'STANDARD DEVIATION =',F10.5/
     *)
  130 FORMAT(58X,'SORTED DATA',//,6(6X,E14.6))
      GO TO 176
  175 WRITE(6,130)(X(I4),I4=1,N)
      WRITE(6,177)XBAR,XSTD
C
C     CALCULATE BANDWIDTH BY METHOD OF SCOTT, TAPIA & THOMPSON (STT)
C
  176 BWIDTH=HN(N,X)
      WRITE(6,21)BWIDTH
   21 FORMAT(/5X,'H BY STT =',E12.7//)
C
C     APPROPRIATE HORIZONTAL AXIS LABEL COEFFICIENTS ARE CALCULATED
C     FOR THE GRAPH AND FOR USE IN NDKER. IMSL ROUTINE NDKER IS CALLED
C     TO ESTIMATE THE DENSITY BY THE KERNEL METHOD. SUBROUTINE GRAPH IS
C     THEN CALLED TO PRODUCE THE PLOT.
C
      H=BWIDTH
      XINC=RND(X(N)-X(1)+2.D0*BWIDTH*3.0D0)/10.D0
      XL=X(1)-BWIDTH*3.0D0
      XLO=ALOW(XL)
      DO 520 I=1,13
         FI=FLOAT(I)
         GR(I)=XLO+(DBLE(FI)-1.0D0)*XINC
  520 CONTINUE
      DO 530 K=1,4
         Y(K)=1.D0-(1.D0/T(K))
  530 CONTINUE
      K=1
      ZERO=0.0D0
      SPACE=RND((XL-XLO)*10.D0)/XINC)
      XINC2=XINC/10.D0
      F(1,3)=0.0D0
      DEL=-1.0D0
      NPT=220
      B(1)=XL
```



```
FILE: NOR      FORTRAN  A -        VM/SP CONVERSATIONAL MONITOR S
      F(1,1)=XL
      DO 535 I=2,NPT
         B(I)=XL+0.5D0*XINC2*(I-1)
  535 CONTINUE
      CALL NDKER(X,N,H,DEL,XKER,B,NPT,FF)
      F(1,2)=FF(1)
      L=2
      DO 540 J=2,210,2
         F(L,1)=B(J)
         F(L,2)=FF(J)
         FEXM=FF(J-1)
         F(L,3)=F(L-1,3)+(F(L-1,2)+4.D0*FEXM+F(L,2))*(XINC2/6.D0)
         IF(F(L,2).GT.FMAX)FMAX=F(L,2)
         TEST1=Y(K)-F(L-1,3)
         TEST2=Y(K)-F(L,3)
         IF(TEST1.GE.ZERO.AND.TEST2.LE.ZERO)GO TO 504
         GO TO 540
  504    M=K+2
         FLD(K)=F(L-1,1)
         CUMDEN(M-1)=F(L-1,3)
         CUMDEN(M)=F(L,3)
         K=K+1
  540 L=L+1
      WRITE(6,205)FMAX
  205 FORMAT('1',1X,'FMAX=',F5.3)
      CALL GRAPH(GR,UNITS,SPACE,FMAX)
      WRITE(6,207)BWIDTH
  207 FORMAT('0',24X,'BANDWIDTH = ',F9.6,5X,'KERNEL = NORMAL')
      WRITE(6,208)
  208 FORMAT('0',24X,'PLOT OF KERNEL DENSITY  - H BY METHOD OF SCOTT, T
     *PIA & THOMPSON')
      WRITE(6,210)((FI,J),J=1,3),I=2,110)
  210 FORMAT('1',42X,'GRAPH COORDINATES',/,43X,'----- -----------',///,
     *3(11X,'X',8X,'DENSITY',2X,'DIST.'),///,(1X,F8.3,1X,F8.6,1X,F8.4))
      WRITE(6,506)
  506 FORMAT(1H //////,5X,15HRETURN    FLOOD,/,5X,17HPERIOD   ESTIMATE/)
      DO 545 K=1,4
         M=K+2
         XP=FLD(K)+((Y(K)-CUMDEN(M-1))/(CUMDEN(M)-CUMDEN(M-1)))*XINC2
         WRITE(6,507)T(K),XP
  545 CONTINUE
  507 FORMAT(4X,F7.3,3X,F7.0)
      WRITE(6,220)
  220 FORMAT('1')
   46 FORMAT(' ',('ESTIMATE STAN.DEV.=',F10.2,/))
      SUM1 = ZERO
      SUM2 = ZERO
      SUM3 = ZERO
      DO 150 I=2,122
         FK = F(I,2)
         BK = F(I,1)
         SUM3 = SUM3+FK
         SUM1 = SUM1+FK*BK
         SUM2 = SUM2+FK*BK*BK
  150 CONTINUE
```

```
      EM=SUM1/SUM3
      WRITE(6,45)EM
   45 FORMAT(' ',///,' ESTIMATE MEAN      =',F10.2)
      VAR=(SUM2-SUM3*EM**2)/(SUM3-1.D0)
      ES=DSQRT(VAR)
      WRITE(6,46)ES
      STOP
      END
C
C     GRAPH SUBPROGRAM--
C     THE GRAPH SUBROUTINE TAKES THE COORDINATES OF THE ESTIMATE AND
C     CONSTRUCTS THE POINTS OF THE GRAPHICAL REPRESENTATION INCLUDING
C     ALL GRAPH LABELS.  THE HEIGHT OF THE GRAPH IS BASED ON THE
C     LARGEST COORDINATE VALUE AND IS SET AT 0.25,0.50,0.75, OR 1.00
C
      SUBROUTINE GRAPH(GR,UNITS,SPACE,FMAX)
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL FI3
      COMMON X,F
      DIMENSION GR(13),F(110,3),G(51,125),UNITS(3),X(200)
      DATA G/6375*' '/,S/'*'/
      RND(X)=DBLE(AINT(SNGL((X*2.D0+DSIGN(1.D0,X))/2.D0)))
      IF(FMAX.LE.1.5D0) GO TO 205
      DY=FMAX/50.D0
      DO 210 J=2,110
         IG:=SPACE+J
         ITG=IDINT((F(J,2)/DY)+0.D05)
         IG=51-ITG
         DO 215 JR=IG,50
            G(JR,IG2)=S
  215    CONTINUE
  210 CONTINUE
      IP=0
      DO 220 I=1,51
         I4=51-I
         IF(MOD(I4,5).NE.0.D0)      GO TO 225
         FN=FMA)-(FMAX/10.D0)*IP
         IP=IP+1
         WRITE(6,230)FN
  230    FORMAT(1X,F4.2,'+')
         GO TO 250
  225    CONTINUE
         WRITE(6,235)
  235    FORMAT(5X,'I')
  250    CONTINUE
         WRITE(6,240)(G(I,IT),IT=1,125)
  240    FORMAT(1H+,5X,125A1)
  220 CONTINUE
      GO TO 245
  205 CONTINUE
      IGHIGH=IDINT((FMAX*100.D0/25.D0)+1.D0)*25
      IF(IGHIGH.GT.100)IGHIGH=100
      DO 200 J=2,110
         IG2=SPACE+J
         ITG=RND((F(J,2)*5000.D0)/IGHIGH)
```



```
         IG=51-ITG
         IF(I0.EQ.0.0D0) GO TO 200
         DO 198 JR=IG,50
            G(JR,IG2)=S
  198    CONTINUE
  200 CONTINUE
      NU=-1
      DO 100 I=1,51
         I2=(I-1)/5
         I3=51-I
         FI3=FLOAT(I3)
         FN=(DBLE(FI3)*IGHIGH)/5000.0D0
         IF(I2.EQ.NU)GO TO 25
         WRITE(6,20)FN
   20    FORMAT(1X,F4.2,'+')
         NU=I2
         GO TO 40
   25    CONTINUE
         WRITE(6,30)
   30    FORMAT(5X,'I')
   40    CONTINUE
         WRITE(6,50)(G(I,IT),IT=1,125)
   50    FORMAT(1H+,5X,125A1)
  100 CONTINUE
  245 WRITE(6,120)
  120 FORMAT(1H+,5X,'+',12('---------+'),'-----')
      WRITE(6,130)(GR(J),J=1,13)
  130 FORMAT(4X,F5.2,12(5X,F5.2))
      WRITE(6,140)UNITS
  140 FORMAT('0',50X,3A4)
      RETURN
      END
C
C     HN SUBPROGRAM--
C     FUNCTION HN USES THE QUASI-OPTIMAL PROCEDURE OF SCOTT, TAPIA &
C     THOMPSON (STT) BASED ON FUNCTIONAL ITERATION TO DETERMINE THE
C     BANDWIDTH
C
      FUNCTION HN(N,X)
      IMPLICIT REAL*8 (A-H,O-Y)
      DIMENSION  X(200)
      CALL ERRSET(208,256,-1,1)
      GO TO 10
   11 WRITE(6,4)
    4 FORMAT(6X,'XXXX',9X,'NEGATIVE TWICE OR BAD SLOPE')
   10 IF(N.LE.0) GO TO 9006
      WRITE(6,3)
    3 FORMAT(///,5X,'CALCULATION OF BANDWIDTH BY METHOD OF SCOTT, TAPI
     *A THOMPSON (STT)')
      XN=N
      IF (N.GT.200)CALL ERROR(1)
      PISQRT = DSQRT(3.141592654D0)
      SIGH = (2.0D0*PISQRT*XN)**(-0.2D0)
      CONS = 3.0D0/(8.0D0*PISQRT*XN*XN)
      H = 1.0D0
```

```
FILE: NOR      FORTRAN  A                VM/SP CONVERSATIONAL MONITOR SY
      READ (5,203) H
  203 FORMAT(D15.8)
      IER = 0
      ITER = 0
      WRITE(6,201)
  201 FORMAT(//6X,'ITER',9X,'H',23X,'SUM',23X,'TOT',20X,'GH',
     *18X,'GPRIME'//)
   30 IF (ITER.GT.12) GO TO 11
      H2 = H*H
      H4 = H2*H2
      H6 = H2*H4
      HINV = -1.0D0/(4.0D0*H2)
      SUM = 0.0D0
      TOT = 0.0D0
      DO 40 I = 2,N
         IM1 = I-1
         XI = X(I)
         DO 35 J=1,IM1
            C2 = (X(J)-XI)**2
            C3 = HINV*C2
            IF (C3.LT.-150.0D0) GO TO 35
            C4 = C2*C2
            C3 = DEXP(C3)
            SUM = SUM+(H4-C2*H2+C4/12.0D0)*C3
            TOT = TOT+(-5.0D0*H6+7.5D0*C2*H4-1.25D0*C4*H2+C4*C2/24.D0)*C3
   35    CONTINUE
   40 CONTINUE
      SUM = (XN*H4+2.0D0*SUM)*CONS/(H4*H4*H)
      TOT = (-5.0D0*XN*H6+2.5D0*TOT)*CONS/(H6*H6)
      GH = H-SIGN*SUM**(-0.200)
      GPRIME = 1.0D0+SIGN*SUM**(-1.2D0)*TOT/5.0D0
      WRITE(6,45) ITER,H,SUM,TOT,GH,GPRIME
   45 FORMAT(I10,3F25.15,2D20.8)
      TEMP = GH/GPRIME
      IF (DABS(TEMP).GT.0.5D0*H) TEMP = GH
      H = H-TEMP
      ITER = ITER+1
      IF (DABS(GH).LT.1.0D-5) GO TO 47
      IF (H.GT.0.0D0) GO TO 30
      IF (IER.NE.0) GO TO 11
      H = 0.1D0
      IER = 129
      GO TO 30
   47 WRITE(6,45) ITER,H
      IF (DABS(GPRIME-1.0D0).GE.1.0D0) GO TO 11
      HN=H
 9006 RETURN
      END
C
C     ERROR SUBPROGRAM--
C     SUBROUTINE ERROR IS CALLED BY FUNCTION HN IF THERE IS AN ERROR
C
      SUBROUTINE ERROR(K)
      WRITE(6,5) K
    5 FORMAT(' ***** TERMINAL ERROR',I2,' *****')
```



```
FILE: NOR      FORTRAN  A                VM/SP CONVERSATIONAL MONITOR
 9005 STOP
 9006 RETURN
      END
C
C     XKER SUBPROGRAM--
C     THIS FUNCTION IS THE NORMAL KERNEL AND IS CALLED BY NDKER
C
      REAL FUNCTION XKER(Y)
      DOUBLE PRECISION Y,RT2PI
      DATA RT2PI/2.506628275D0/
      XKER=DEXP(-Y*Y/2.D0)/RT2PI
      RETURN
      END
```

```
C
C     MODIFIED VARIABLE BANDWIDTH KERNEL DENSITY ESTIMATOR
C
C     PROGRAMMER: P. FLANAGAN STAT MASTER'S THESIS-1980
C     MODIFIED BY MING-CHUNG LIU STAT MASTER'S THESIS-1983
C     MODIFIED BY C. LABATIUK CVG MASTER'S THESIS-1984
C
C     THIS PROGRAM ACCEPTS DATA POINTS AND COMPUTES AN APPROXIMATION
C     OF THE PROBABILITY DENSITY FUNCTION USING THE MODIFIED VARIABLE
C     BANDWIDTH KERNEL DENSITY ESTIMATOR WITH A QUADRATIC KERNEL
C     (K=3/(4*5**.5)*(1-Y**2/5)).  DATA INPUTTED IS ARRANGED IN
C     INCREASING ORDER AND SCALED DOWN (OR UP), IF NECESSARY; DENSITY
C     ESTIMATIONS ARE CALCULATED; AND THE GRAPH OF THE DENSITY FUNCTION
C     IS PRINTED.  THE DISTRIBUTION FUNCTION IS CALCULATED USING
C     SIMPSON'S 1/3 RULE AND USED IN FLOOD FREQUENCY ANALYSIS.  FLOODS
C     ARE ESTIMATED FOR 10,50,100, AND 200 YEAR RETURN PERIODS.
C
C     INPUT EXPECTED
C     (1)THE UNITS INVOLVED (CHARACTER VARIABLE <= 12 LETTERS)
C     (2)RECORD LENGTH (INTEGER VARIABLE, I3 FORMAT)
C     (3)ANNUAL MAXIMUM FLOOD DATA (REAL VARIABLES, 8F8.1 FORMAT)
C
C     IMPORTANT VARIABLES:
C        X=INPUT VECTOR OF LENGTH N CONTAINING THE FLOOD DATA;
C        N=RECORD LENGTH INTEGER;
C        BANDWIDTH=THE BAND WIDTH (OR WINDOW WIDTH);
C        XBAR=THE SAMPLE MEAN;
C        XSTD=THE SAMPLE STANDARD DEVIATION;
C        XINC=THE INCREMENT BETWEEN LABELS ON THE HORIZONTAL AXIS;
C        G(,)=THE GRAPH CHARACTERS;
C        F(,)=THE NUMERICAL GRAPH COORDINATES;
C        IHOLD=THE INTEGER INDEX FOR LIMITING SUMMATION IN THE
C              SUBROUTINE FKER1 TO JUST THOSE X VALUES APPLICABLE.
C
C     SUBPROGRAMS:
C        GRAPH - THIS SUBROUTINE GENERATES THE GRAPH ITSELF;
C        FKER1 - THIS FUNCTION COMPUTES THE VALUE OF THE ESTIMATOR
C                FOR EACH POINT ON THE GRAPH HORIZONTAL AXIS;
C        HN    - THIS FUNCTION GENERATES THE VARIABLE BANDWIDTH
C                VALUE BASED ON THE DATA.
C
      DOUBLE PRECISION SUM1,SUM2,SUM3
      DIMENSION X(200),F(118,3),GR(13),UNITS(3),
     *A(35),B(35),T(4),Y(4),FLD(4),CUMDEN(8)
      DATA XSUM/0.0/,XSUMSQ/0.0/,FMAX/0.0/,
     *T/10.,50.,100.,200./,RT5/2.236068/
      DATA ZERO,HALF,THREE,SIX,TWELVE/0.,0.5,3.,6.,12./
      COMMON X,F
      RND(X)=AINT((X+2*SIGN(1.,X))/2.0)
      ALOW(X)=AINT(X-(1.-SIGN(1.,X))/2.)
C
C     READ AND SORT SEGMENT--THIS SEGMENT READS IN EACH DATA POINT AND
C     DOES A BINARY SEARCH FOR ITS POSITION IN THE PREVIOUSLY SORTED
C     DATA.  THE DIFFERENCE BETWEEN TWO INDEXES EQUAL TO THE POSSIBLE
C     SUBSCRIPTS OF THE X VARIABLE ARE GRADUALLY NARROWED, WHEN THEY
C     ARE EQUAL, THE SEARCH IS COMPLETED AND THE NEW VALUE ENTERED.
C
      READ(5,8)UNITS
    8 FORMAT(3A4)
      READ(5,10)N
   10 FORMAT(I3)
      IF(N.LE.200)GO TO 70
      WRITE(6,12)
   12 FORMAT(1X,'THE RECORD LENGTH HAS EXCEEDED 200')
      WRITE(6,14)
   14 FORMAT(1X,'THIS PROGRAM IS TERMINATED')
      WRITE(6,16)
   16 FORMAT(1X,'IF YOU INTEND TO RUN AGAIN USE 200 RANDOM VALUES')
      GO TO 9005
   70 CONTINUE
      READ(5,18)(X(I),I=1,N)
   18 FORMAT(8F8.1)
      DO 505 I=1,N
         XSUM=XSUM+X(I)
         XSUMSQ=XSUMSQ+X(I)**2
   20    CONTINUE
         IHIGH=I
         ILOW=1
   22    CONTINUE
         IF(IHIGH.EQ.ILOW)GO TO 50
         IHALF=(ILOW+IHIGH)/2
         IF(X(I)-X(IHALF))30,25,40
   25    CONTINUE
         INDEX=IHALF+1
         GO TO 60
   30    CONTINUE
         IHIGH=IHALF
         GO TO 22
   40    CONTINUE
         ILOW=IHALF+1
         GO TO 22
   50    CONTINUE
         INDEX=ILOW
   60    CONTINUE
         SAVE=X(I)
         ISORT=I-INDEX
         DO 510 I2=1,ISORT
            I3=I+1-I2
            X(I3)=X(I3-1)
  510    CONTINUE
         X(INDEX)=SAVE
  505 CONTINUE
      WRITE(6,24)N
   24 FORMAT('1',4X,'NUMBER OF DATA POINTS READ IN = ',I4,//)
      EN=FLOAT(N)
      XBAR=XSUM/EN
      XVAR=(XSUMSQ-EN*XBAR**2)/(EN-1.)
      XSTD=SQRT(XVAR)
      XLAR=AMAX1(ABS(X(N)),ABS(X(1)))
      XSPAN=X(N)-X(1)
```

```
C
C     SCALE CALCULATION SEGMENT--THIS SECTION EXAMINES THE DATA FOR
C     SCALE SIZE AND, IF NECESSARY, IT PERFORMS A SCALE CHANGE BY SUB-
C     TRACTING THE SAMPLE MEAN FROM EACH DATA POINT AND DIVIDING BY THE
C     SAMPLE STANDARD DEVIATION.  IF SUCH AN ALTERATION IS NECESSARY, A
C     MESSAGE TO THIS EFFECT IS PRINTED.  FINALLY, THE APPROPRIATE HORI-
C     ZONTAL AXIS LABEL COEFFICIENTS ARE CALCULATED.
C
      IF((XLAR.LT.100.).AND.(XSPAN.GT.1.0).AND.(XLAR.GT.1.0))GOTO80
      DO 515 I=1,N
         X(I)=(X(I)-XBAR)/XSTD
  515 CONTINUE
      XLAR=(XLAR-XBAR)/XSTD
      BWIDTH=BWIDTH/XSTD
      WRITE(6,76)(X(I4),I4=1,N)
      WRITE(6,72)XBAR,XSTD
   72 FORMAT(///,4X,'DATA ALTERATION HAS BEEN EMPLOYED DUE TO THE',/,
     *5X,'MAGNITUDE OF THE LARGEST DATA POINT.EACH VALUE HAS BEEN',/,5X,
     *'REDUCED BY THE SAMPLE MEAN = ',F12.5,/,5X,'AND DIVIDED BY
     *THE SAMPLE STANDARD DEVIATION = ',F12.5,//)
   74 FORMAT(/12X,'SAMPLE MEAN =',F10.5/5X,'STANDARD DEVIATION =',F10.5/
     *)
   76 FORMAT(28X,'SORTED DATA',//,3(6X,E14.6))
      GO TO 90
   80 WRITE(6,76)(X(I4),I4=1,N)
      WRITE(6,74)XBAR,XSTD
   90 BWIDTH=HN(N)
      WRITE(6,78)
   78 FORMAT(///,5X,'MODIFIED BANDWIDTH CALCULATIONS FOR THE QUADRATIC K
     *ERNEL HAVE BEEN EMPLOYED',/)
      XINC=RND(X(N)-X(1)+2.*BWIDTH*RT5)/10.
      XL=X(1)-BWIDTH*RT5
      XLO=ALOW(XL)
      DO 520 I=1,13
         GR(I)=XLO+(FLOAT(I)-1.0)*XINC
  520 CONTINUE
C
C     POINT CALCULATION SEGMENT--THIS SECTION CALLS THE FKER1 FUNCTION
C     FOR EACH POINT ON THE GRAPH'S X-AXIS, DETERMINING ITS ESTIMATOR.
C     SUBROUTINE GRAPH IS THEN CALLED TO PRODUCE THE PLOT.
C
      DO 525 K=1,4
         Y(K)=1.-(1./T(K))
  525 CONTINUE
      K=1
      ZERO=0.0
      IHOLD=1
      SPACE=RND(((XL-XLO)*10.)/XINC)
      XINC2=XINC/18.
      BWID2=BWIDTH*RT5
      EX0=XL
      F(1,1)=EX0
      F(1,2)=FKER1(N,IHOLD,BWIDTH,EX0)
      F(1,3)=0.0
      DO 530 J=2,101
         EX=XL+XINC2*(J-1)
```



```
         EXM=EX-XINC2/2.
         GO TO 195
  190    CONTINUE
         IHOLD=IHOLD+1
  195    CONTINUE
         IF((X(IHOLD).LE.(EX-BWID2)).AND.(IHOLD.LT.N))GO TO 190
         F(J,1)=EX
         F(J,2)=FKER1(N,IHOLD,BWIDTH,EX)
         FEXM=FKER1(N,IHOLD,BWIDTH,EXM)
         F(J,3)=F(J-1,3)+(F(J-1,2)+4.*FEXM+F(J,2))*(XINC2/6.)
         IF(F(J,2).GT.FMAX)FMAX=F(J,2)
         TEST1=Y(K)-F(J-1,3)
         TEST2=Y(K)-F(J,3)
         IF(TEST1.GE.ZERO.AND.TEST2.LE.ZERO)GO TO 200
         GO TO 530
  200    M=K*2
         FLD(K)=F(J-1,1)
         CUMDEN(M-1)=F(J-1,3)
         CUMDEN(M)=F(J,3)
         K=K+1
  530 CONTINUE
      WRITE(6,205)FMAX
  205 FORMAT('1',1X,'FMAX=',F5.3)
      CALL GRAPH(GR,UNITS,SPACE,FMAX)
      WRITE(6,207)BWIDTH
  207 FORMAT('0',27X,'BANDWIDTH = ',F9.6,5X,'KERNEL = QUADRATIC')
      WRITE(6,209)
  209 FORMAT('0',27X,'PLOT OF MODIFIED VARIABLE BANDWIDTH KERNEL DENSITY
     *')
      WRITE(6,210)((F(I,J),J=1,3),I=1,101)
  210 FORMAT('1',42X,'GRAPH COORDINATES',/,43X,'----- ----------',///,
     *3(11X,'X',8X,'DENSITY',2X,'DIST.'),//,(1X,F8.3,1X,F8.6,1X,F8.4))
      WRITE(6,212)
  212 FORMAT(1H ///////,5X,15HRETURN    FLOOD,/,5X,17HPERIOD   ESTIMATE/)
      DO 535 K=1,4
         M=K*2
         XP=FLD(K)+(((Y(K)-CUMDEN(M-1))/(CUMDEN(M)-CUMDEN(M-1)))*XINC2)
         WRITE(6,214)T(K),XP
  535 CONTINUE
  214 FORMAT(4X,F7.3,3X,F7.0)
      WRITE(6,216)
  216 FORMAT('1')
      SUM1 = ZERO
      SUM2 = ZERO
      SUM3 = ZERO
      DO 150 K=1,118
         FK = F(K,2)
         BK = F(K,1)
         SUM3 = SUM3+FK
         SUM1 = SUM1+FK*BK
         SUM2 = SUM2+FK*BK*BK
  150 CONTINUE
      EM=SUM1/SUM3
      WRITE(6,45)CEM
   45 FORMAT(' ',(//,' ESTIMATE MEAN      =',F10.2))
```

```
      VAR=(SUM2-SUM3*EM**2)/(SUM3-1.)
      ES=SQRT(VAR)
      WRITE(6,46)ES
   46 FORMAT(' ',('ESTIMATE STAN.DEV.=',F10.2,/))
 0005 STOP
      END
C
C     GRAPH SUBPROGRAM--
C     THE GRAPH SUBROUTINE TAKES THE COORDINATES OF THE ESTIMATE AND
C     CONSTRUCTS THE POINTS OF THE GRAPHICAL REPRESENTATION INCLUDING
C     ALL GRAPH LABELS.  THE HEIGHT OF THE GRAPH IS BASED ON THE
C     LARGEST COORDINATE VALUE AND IS SET AT 0.25,0.50,0.75, OR 1.00.
C
      SUBROUTINE GRAPH(GR,UNITS,SPACE,FMAX)
      DIMENSION GR(13),F(110,3),G(51,125),UNITS(3),X(200)
      DATA G/6375*' '/,S/'*'/
      COMMON X,F
      RND(X)=AINT((X*2+SIGN(1.,X))/2.0)
      IF(FMAX.LE.1.0) GO TO 205
      DY=FMAX/50.
      DO 210 J=2,101
         IG2=SPACE+J
         ITG=IFIX((F(J,2)/DY)+0.5)
         IG=51-ITG
         DO 215 JR=IG,50
            G(JR,IG2)=S
  215    CONTINUE
  210 CONTINUE
      IP=0
      DO 220 I=1,51
         I4=51-I
         IF(MOD(I4,5).NE.0.)GO TO 225
         FN=FMAX-(FMAX/10.)*IP
         IP=IP+1
         WRITE(6,230)FN
  230    FORMAT(1X,F4.2,'+')
         GO TO 250
  225    CONTINUE
         WRITE(6,235)
  235    FORMAT(5X,'I')
  250    CONTINUE
         WRITE(6,240)(G(I,IT),IT=1,125)
  240    FORMAT(1H+,5X,125A1)
  220 CONTINUE
      GO TO 245
  205 CONTINUE
      IGHIGH=IFIX((FMAX*100./25.)+1.)*25
      IF(IGHIGH.GT.100)IGHIGH=100
      DO 200 J=1,101
         IG2=SPACE+J
         ITG=RND((F(J,2)*5000.)/IGHIGH)
         IG=51-ITG
         IF(IG.EQ.0.0)GO TO 200
         DO 198 JR=IG,50
            G(JR,IG2)=S
```



```
  198    CONTINUE
  200 CONTINUE
      NU=-1
      DO 100 I=1,51
         I2=(I-1)/5
         I3=51-I
         FN=(FLOAT(I3)*IGHIGH)/5000.0
         IF(I2.EQ.NU)GO TO 25
         WRITE(6,20)FN
   20    FORMAT(1X,F4.2,'+')
         NU=I2
         GO TO 40
   25    CONTINUE
         WRITE(6,30)
   30    FORMAT(5X,'I')
   40    CONTINUE
         WRITE(6,50)(G(I,IT),IT=1,125)
   50    FORMAT(1H+,5X,125A1)
  100 CONTINUE
  245 WRITE(6,120)
  120 FORMAT(1H+,5X,'+',12('----------+'),'-----')
      WRITE(6,130)(GR(J),J=1,13)
  130 FORMAT(4X,F5.2,12(5X,F5.2))
      WRITE(6,140)UNITS
  140 FORMAT('0',50X,3A4)
      RETURN
      END
C
C     FKER1 SUBPROGRAM--
C     THIS FUNCTION CALCULATES THE PROBABILITY DENSITY FUNCTION
C     ESTIMATOR OF THE DATA INPUTTED SCALED AS NECESSARY.
C
      FUNCTION FKER1(N,IHOLD,BWIDTH,EX)
      DIMENSION X(200),F(105,3)
      COMMON X,F
      DATA RT5/2.236068/
      FKER1=0.0
      DO 150 I=IHOLD,N
         VAL=(EX-X(I))/BWIDTH
         ANEG=-RT5
         IF(VAL.GT.RT5)GO TO 150
         IF(VAL.LE.ANEG)GO TO 160
         FKER1=FKER1+0.33541*(1.-VAL**2/5.)
  150 CONTINUE
  160 CONTINUE
      FKER1=FKER1/(N*BWIDTH)
      RETURN
      END
C
C     HN SUBPROGRAM--
C     THIS FUNCTION CALCULATES THE VARIABLE BANDWIDTH USING METHOD
C     SUGGESTED BY WAGNER(1975).
C     HN EQUALS AVERAGE OF D(JN), WHERE D(JN) IS THE DISTANCE FROM
C     EVERY X(J) TO ITS K(N) NEAREST NEIGHBOR FROM X(1),X(2),...,X(N)
C     WITH X(J) OMITTED.
```

FILE: MVB .FORTRAN A VM/SP CONVERS.

```
C
      FUNCTION HN(N)
      DIMENSION X(200),A(35),B(35),F(105,3)
      COMMON X,F
      DSUM=0.
      KN=INT(N**0.5)
      DO 100 I=1,N
         IF(I.EQ.1) GO TO 15
         IF(I.LE.KN) GO TO 5
         IF(I.GT.KN) GO TO 20
    5    K1=I-1
         DO 10 M=1,K1
            B(I-M)=X(I)-X(M)
   10    CONTINUE
         B(I)=10000.
         GO TO 16
   20    K2=I-KN
         K1=I-1
         DO 25 M=K2,K1
            B(I-M)=X(I)-X(M)
   25    CONTINUE
         GO TO 16
   15    B(I)=10000.
   16    IF(I.EQ.N) GO TO 30
         IF(I.LE.N-KN) GO TO 35
         IF(I.GT.N-KN) GO TO 45
   35    K3=I+1
         K4=I+KN
         DO 40 MM=K3,K4
            A(MM-I)=X(MM)-X(I)
   40    CONTINUE
         GO TO 88
   45    K3=I+1
         DO 50 MM=K3,N
            A(MM-I)=X(MM)-X(I)
   50    CONTINUE
   30    A(N-I+1)=10000.
   88    L=1
         LL=1
   55    IF(L+LL-KN)60,60,65
   60    IF(A(L)-B(LL))70,70,75
   70    L=L+1
         GO TO 55
   75    LL=LL+1
         GO TO 55
   65    IF(A(L)-B(LL))85,85,86
   85    D=A(L)
         GO TO 105
   86    D=B(LL)
  105    DSUM=DSUM+D
  100 CONTINUE
      HN=DSUM/N
      RETURN
      END
```