AN INTER-SCALE CORRELATION STRUCTURE OF PEAK FLOW SERIES

BOXIAN WU


CENTRE FOR NFLD. STUDIES
MEMORIAL UNIVERSITY
OF NEWFOUNDLAND

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

AN INTER-SCALE CORRELATION STRUCTURE

OF PEAK FLOW SERIES

By

© Boxian Wu

A thesis submitted to the School of Graduate Studies

in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

September 1997

St. John's Newfoundland Canada

Abstract

This thesis deals with the correlation structure of annual peak flow series in detail. The thesis is in divided into four major parts.

Part one takes a closer look at the correlation structure of annual peak flow series at two scales: scale of one that measures short-term behaviour by the lag-one autocorrelation coefficient, r(1), and scale of n, that measures long-term behaviour by Hurst's K. It is shown that there are significant correlation and dependence between Hurst's K and r(1) for both observed data and data from Monte Carlo experiments which imply that short- and long-term behaviour cannot be treated separately as is current practice.

Part two suggests a new approach for quantitatively describing long-term correlation that is rooted in an independent series. The results indicate that long-term correlation rooted in a short-term independent series can be quantitatively estimated, and the simultaneous occurrence of high values of Hurst's K and low values of r(1) is, in fact, not an uncommon phenomenon. A new method of testing for long-term correlation that takes the short-term correlation into account is developed.

Part three further looks at peak flow correlation structure across scales based on the perspective of fractal geometry. A family of probability-scale-threshold curves which contain more information about the correlation structure of peak flows, are constructed and the scaling behaviour of peak flow series is explored.

In order to take serial correlation into account in flood risk analysis, the concept of

scaling plotting positions (SPP), is developed in part four. It takes scaling behaviour of peak flows into account and develops a new plotting position formula in estimation of future floods. The results of Monte Carlo simulation showed that the estimated quantiles of SPP are more efficient and robust when compared with current estimators of flood quantiles.

The study presented in this thesis has provided a view of the correlation structure of peak flows across scales so that flood risk can be better estimated.

To My Father,

Professor Wu Mingyuan

(1919-1992)

Acknowledgements

I would like to express my deep sense of gratitude to my supervisor, Dr. L. M. Lye, for his advice, patience, continual encouragement and guidance throughout the courses and thesis, for reading and critiquing drafts and for arranging financial support enabling me to complete the program.

My thanks also to Dr. J. Sharp for his teaching and encouragement throughout the program, and to Dr. G. Sabin and Dr. A.W. Robertson for their insightful comments and suggestions.

To the staff of the Faculty of Engineering and Applied Science for their continual co-operation.

To my friends the Allureds, R.Yang and Y. Hou and many others for their warm fruitful inputs.

My thanks also extend to my mom Li Weilan, brothers and sisters for their help and support, to my daughter Xiaoli and son Enyin, for their love, support and for reminding me to be ever striving.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context

Hydrology, broadly speaking, is the study of water. One of the definitions of hydrology (Federal Council for Science and Technology, 1962) is:

> "... the science that treats the waters of the Earth, their occurrence, circulation, and distribution, their chemical and physical properties, and their reaction with their environment, including their relation to living things. The domain of hydrology embraces the full life history of water on the earth."

Water is essential for the survival of living organisms but can also foster disaster during periods of flooding. In some parts of the world, floods occur frequently, many resulting in tremendous monetary and emotional loss, physical upheaval, property damage and loss of life. With the advance of civilisation, the demand for, and the need to control, such natural disasters is increasing. For the design of flood mitigation measures, the estimation of the magnitude of flood events according to specified probabilities, is therefore essential. This is normally referred to as flood frequency analysis.

The procedure of classical flood frequency analysis involves four main steps:

(1) Selection of data.

The data analysed are required to be independently and identically distributed. The restriction of homogeneity assures that all observations are from the same population, and the restriction of independence assures that the probability of one hydrologic event does not affect other occurrences. Two data series of peak flows are commonly used for flood frequency analysis: the annual maximum series (AM) (e.g. Fuller, 1914; Chow, 1954; Rossi et al., 1984), and the peaks over threshold series (POT)(e.g. NERC, 1975). AM series takes the single maximum peak flow in each year of record, it is necessary to ensure that the selected annual peaks are independent of one another. The POT series takes all the peaks over a threshold, where there is more chance of the peak flows being correlated and the assumption of true independence is less valid.

For the purpose of probability frequency analysis, the data should be unbiased, independent and homogeneous. Assessment of data quality is usually achieved using statistical testing techniques.

(2) Choosing a probability distribution.

Since nature's distribution is unknown to hydrologists, reasonable 'flood-like' distributions should be chosen to fit the observed flood data. The most widely used method for choosing a probability model is the probability plot. This is a plot of the magnitudes of flood events versus the probabilities that the magnitudes are or are not exceeded. The

selected data are arranged in order of the magnitude and plotted using a suitable plotting position on probability graph paper. Several "flood-like" distributions, such as the Pearson Type III distribution, the log Pearson Type III and the general extreme value distributions are widely used in the estimation of future floods in practice.

(3) Estimating the parameters of the chosen distribution.

Observed peak flows, ignoring the order of their occurrence, are considered as a sample and are used to estimate the parameters for the chosen parent distribution. The estimation may be done either mathematically or graphically. In general, a mathematical estimation can be achieved by the method of maximum likelihood, or the method of moments, or L-moments. By graphical fitting, a subjective graphical curve is simply drawn to fit the plotted data by eye, and although this method is the simplest it involves human error.

(4) Making inferences about future floods with a given probability.

Historically, since the recognition (Fuller, 1914) that there was no such thing as a single design flood but rather a choice of different return period floods depending on circumstances, hydrologists have turned to the statistical analysis of extreme flood events in which observations of peak flows are considered as outcomes of a random experiment in a natural experimental field. Thus, it is now possible for hydrologists to estimate future floods using the above procedure based on statistical theory.

However, in a statistical estimation of future floods as outlined, the latter three steps have to be based on the collected data, which are commonly assumed as independently and

identically distributed. In short, the data must satisfy the assumption that annual peak flows are serially independent. Serially dependent data are usually treated as independent data because the effect of short-term dependence on annual peak flow estimation is very small (Srikanthan and McMahon, 1981). Hence, hydrologists have focused on the study of probabilistic models and parameter estimation methods involved in steps 2 and 3 without seriously worrying about the correlation structure of annual peak flows.

Although statistics work well in hydrologic estimation, some dissatisfaction with statistical hydrologic design has been voiced arising from the use of so called long-term dependent data. Hurst (1951, 1956) found that long-term dependence was inherent in annual flow records of the Nile River and its impact upon the design of the Aswan High Dam. Chow (1964) cautioned that the variables in actual hydrologic phenomena are likely to be interdependent to some extent, and the possibility of this interdependence should be investigated in flood risk analysis. From Carrigan and Huzzen (1967), the specific effects of neglecting to consider serial correlation could underestimate the population variation of a peak flow series. In more recent studies, Lye (1987), Booy and Lye (1989), Lye and Lin (1994) showed that annual maximum flood peak flow series also exhibit long-range dependence, and this information should not be discarded and denied by classical probability analysis. When the series of observations exhibits long-term serial correlation, the variances of the sample statistics are greater than that for either short-term correlation or independent processes. Mandelbrot and Wallis (1969b) showed that with a typical h

value of 0.7, and n of 50, the variance of the sample mean expressed as

$$Var(\bar{x}) = \sigma_x^2 n^{2h-2}$$

is almost twice as large when the data are independent, where Var($\bar{x}$), σ_x^2 and h are variance of sample mean, variance of sample and Hurst coefficient, respectively. The studies by Lye (1987) showed that long-term dependence affects the flood estimation. Taking this into account, the risk associated with future peak flows will significantly increase.

Therefore, although annual peak flow series are assumed to have no correlation "structure" in statistical terms, impacts of long-term dependence, in fact, cannot be ignored in flood risk analysis. Hence long-term behaviour based on short-term independence deserves further investigation.

1.2 Objectives of Thesis

Since the structure of peak flows is directly related to the safety of hydraulic infrastructures, study of the correlation structure of peak flows is an important aspect in flood frequency analysis. This thesis has four objectives related to the issue of the correlation structure of peak flow series.

Objective one is to investigate the simultaneous occurrence of long-term persistence and short-term independence of annual peak flows using actual flow data as well as

simulated data.

It can be demonstrated that independent series can have significantly different levels of long-term behaviours although they have similar short-term behaviour. This means, that the outcomes in the macro scale are possible for the same degree of disorder in the micro scale. However, the historical records we observed are one among many possible outcomes. Of course, the macrocosmic and microcosmic phenomena are related to each other. To describe the simultaneous occurrence of long-term persistence and short-term independence of annual peak flows, the Hurst's K and lag one autocorrelation r(1) which describe the long- and short-term behaviour of peak flows, respectively, are considered as random variables. Statistical tests are carried out to test for a statistical relationship between Hurst's K and r(1) estimated from an analysis of peak flows observed in Canadian and Chinese rivers.

Objective two is to provide a logically consistent probabilistic approach for quantitatively describing the correlation structure of annual peak flows.

The idea is, basically, that because Hurst's K and the lag-one autocorrelation coefficient r(1) are correlated and dependent upon statistical tests, their joint probability exists and can be estimated by classical statistics. In other words, it is possible for us to look for a statistical relationship between the long-term persistence and short-term independence. In a quantitative way, an approximation for the sampling distribution of Hurst's K, which is expressed as the sampling distribution for a given lag-one

autocorrelation coefficient r(1), is developed using Monte Carlo simulation.

Objective three is to explore the correlation structure of peak flows using fractal geometry.

After estimating the "correlation" and "dependence" by the two measures, Hurst's K, and r(1), which measures objects at scales n and one, respectively. The following questions arise: What is the real distribution of peak flows? Can we look at the peak flow structure across scales? To this end, fractal geometry will be used to explore the structure of peak flow series.

Fractal geometry is now widely used in different scientific disciplines to describe the structure of complex self-similar phenomena and scaling behaviour of physical processes. However, most investigations in hydrology concerned the scaling behaviour of spatial rainfall and runoff phenomena (e.g. Venugopal and Foufoula-Georgiou, 1996; Jonas Olsson and Janusz Niemczynowicz, 1996; Paolo Burlando and Renzo Rosso, 1996; Puente, 1996; Haitjema and Kelson, 1996). The scaling behaviour of peak flow points along the time axis has not been investigated.

Logically, since point events can be modelled by the Cantor dust in a fractal world (Mandelbrot, 1977, 1982), the peak flow points on the time axis, which also form a set of point events, can also be described by fractals.

Using fractal geometry, we can describe the structure of peak flows across scales. The method to be employed to investigate the temporal scaling behaviour of peak flow

points is the functional box counting algorithm (Lovejoy et al., 1987) in which two aspects are emphasised: (1) the peak flow points distributed on the time axis related to probabilities; and (2) a set of thresholds defined in advance. From this, a family of probability-scale-threshold curves is constructed to explore the scaling behaviour of temporal peak flow points. Peak flows observed from Canadian and Chinese rivers are collected and analysed in order to show the scaling structure of peak flows in nature.

Due to the fact that the family of curves constructed from $\ln P_\xi$ - $\ln \xi$ - Qs is useful for hydrological engineering studies, where P_ξ is the probability that the time interval of length ξ will include at least one peak flow event, and Qs is a given threshold, hence the final objective is to develop a scaling plotting position formula which takes scaling behavior of peak flows into account for flood risk analysis. The resulting method can be developed using the probability-scale-threshold curves.

Monte Carlo experiments will be carried out to compare the statistical properties of the scaling approach with traditional estimation in terms of efficiency and robustness.

1.3 Outline of Thesis

The thesis consists of eight chapters. The context, objectives and outline of the thesis have been presented in this chapter. In Chapter Two, a closer look at the long- and short-term behaviour of annual peak flow series is presented, and a probabilistic approach dealing with the long- and short-term behaviour of peak flows is proposed in Chapter Three.

Chapter Four provides an introduction to the topic of fractal geometry which will be used in subsequent chapters dealing with a fractal description of peak flow structure. Chapter Five uses fractal geometry to describe the scaling behaviour of peak flow points. A scaling plotting position for flood risk analysis is developed in Chapter Six. Chapter Seven presents conclusions and recommendations, and the statement of originality of the thesis is described in Chapter Eight.

Chapter 2

A Closer Look at Long- and Short-term Behaviour of Annual Peak Flow Series

2.1 General

The objective of a flood risk analysis is to relate the magnitude of extreme events to their frequency through the use of a probability distribution. Thus, annual peak flow data are usually assumed to be independently and identically distributed. That is, there is no correlation in time. This customary assumption in flood risk analysis has been examined by many investigators. Carrigan and Huzzen (1967) found serial correlation in some of the annual peak flows of rivers in the USA. Srikanthan and McMahon (1981) showed that the effect of short-term dependence on annual peak flow estimation is small. Wall and Englot (1985) used five independence tests and found that annual peak flows can be considered independent for the 57 streams in Pennsylvania.

In comparison with the investigation of short-term behaviour of flood peak series, Lye (1987), and Booy and Lye (1989) focused on the study of long-term behaviour, and demonstrated that there is evidence of long-term persistence in the annual peak flood series of many Canadian rivers. In a more recent study, Lye and Lin (1994) performed

statistical tests for short-term and long-term persistence for annual peak flows and concluded that although short-term serial correlation is practically absent for most of the peak flow series, significant long-term persistence is present for a large number of peak flow series tested.

However, previous investigations have only focused on dealing with long-term dependence or short-term independence separately only, studies of a relationship between long-term dependence and short-term independence of peak flow series are absent.

When we are dealing with the long-term and short-term behaviour of peak flows, a useful analogy is to imagine molecular movements, in which individual molecules move incoherently, representing a short-range characteristic, but a huge number of particles can behave in a coherent fashion, a long-range characteristic.

The evolution of peak flows in time appears to resemble molecular chaos. i.e. the long-term behaviour of such flows forcing 'trends' to persist as gradients produced by short-term disorder motion. Figure 2.1 illustrates three independent series with significantly different levels of long-term behaviours but they are outcomes from similar short-term behaviour. This means that the outcomes in the macro scale are possible for the same degree of disorder in the micro scale behaviour. We can observe only one among possible outcomes that gives us an historical record, named "memory" for past evolution.

Among this and the various issues for investigating the serial correlation structure of annual peak flows, a number of important questions arise. Is there correlation and

dependence between long-term and short-term behaviours in annual peak flow series? Is simultaneous occurrence of long-term persistence and short-term independence of annual peak flows a common phenomenon? Is there a new way of describing annual peak flow series?

To answer the above questions, this chapter deals with the statistical relationship between long-term and short-term behaviour of annual peak flow series, which are measured by Hurst's K and the lag-one autocorrelation coefficient r(1), respectively.

Natural annual peak flow series of 258 rivers collected from Canada and China were analysed. The study focuses on dealing with the statistical properties of these two statistics. Parametric and non-parametric statistical hypothesis tests will be carried out to test correlation and dependence between Hurst's K and r(1). Monte Carlo experiments will also be used to show the variation of these two statistics for serially independent series.

The results of this study is to provide a basis for a proposed probability model for dealing with long-term and short-term behaviours of annual peak flow series.

In order to satisfy the customary assumption in flood risk analysis, an independent probability population is argued throughout the chapter.

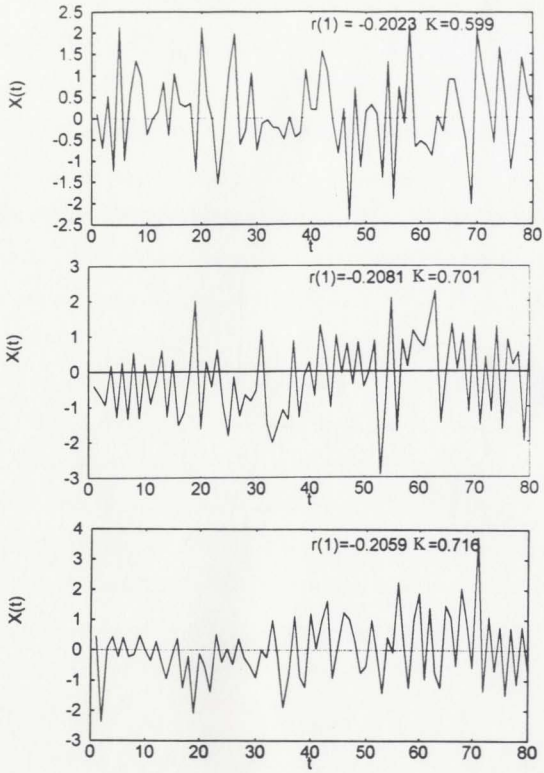

Figure 2.1 Three independent standardized flow time series with different levels of the long- term behaviours, where K and r(1) are Hurst's K and lag-one autocorrelation coefficient, respectively.

2.2 Annual Peak Flows of Canadian and Chinese Rivers

Before investigating the long-term and short-term behaviour of annual peak flow series, an analysis of the statistics of natural annual peak flow series with record length greater than 30 years for 258 Canadian (Environment Canada, 1992) and Chinese rivers (Ministry of Water Resources, 1985) was made. Basic statistics on the selected rivers, including record length, n, mean value, coefficient of variation, Cv, coefficient of skewness, Cs, Hurst's K, and lag-one autocorrelation coefficient, r(1), are presented in Table 2.1. It can be seen from Table 2.1 that most of the rivers have small lag-one serial correlation coefficients, but many rivers have fairly high Hurst's K. The scatter plots, Hurst's K versus lag-one autocorrelation coefficient r(1), for annual peak flow series observed in Canadian and Chinese rivers are shown in Figure 2.2. Examination of the scatter plots in Figures 2.2a - 2.2b, seems to show a correlation between Hurst's K and r(1). In comparison to this, Figures 2.3a - 2.3b show scatter plots of Hurst's K and the coefficient variation, Cv, and Hurst's K and coefficient of skewness of peak flow series, Cs. Apparently, there is no correlation between Hurst's K and these statistics which agrees with a study by Mandelbrot and Wallis (1969b) that the Hurst's statistic is robust against skewness.

Table 2.1a Statistics of natural annual peak flows of Canadian rivers (Source of data: Environment Canada, 1992)

(n - record length in years, r(1) - lag one autocorrelation coefficient, K-Hurst's K, Cv - coefficient of variation, Cs -coefficient of skewness, mean - mean value of annual peak flows in m^3/s)

River Name	n	r(1)	K	Cv	Cs	mean
Adams River Near Squilax	42	.1713	.6321	.30	.20	246.2
Ashnola River Near Keremeds	42	-.3199	.6138	.69	.29	83.3
Athabasca River At Athabasca	47	-.0454	.6011	3.70	2.38	3512.8
Ausable River Near Springbank	43	-.1311	.4882	.64	.16	180.3
Badger Creek Near Cartwright	30	-.0805	.7720	2.04	.80	35.6
Battle River Near Unwin	36	-.0179	.5794	1.41	.67	87.8
Bear River East Branch At Bear River	35	.2360	.6553	.88	.84	39.0
Beaverbank River Near Kinsac	67	-.0957	.7253	.54	.32	29.5
Berens River At Outlet Of Long Lake	31	-.1065	.5686	.71	.24	108.6
Boundary Creek Near Porthill	61	.1189	.7462	.40	.23	47.0
Bowron River Near Wells	33	.0549	.6575	.32	-.02	38.1
Bulkley River At Quick	58	.1750	.6302	.37	.14	587.1
Carrick Creek Near Carlsrume	35	.1858	.7854	.69	.45	28.1
Carrot River Near Smoky Burn	34	-.0225	.5605	1.08	.48	272.7
Castle River Near Beaver Mines	44	.0009	.6694	.87	.92	146.0
Chilliwack River At Vedder Crossing	32	.0015	.6625	.45	.28	304.6
Chilliwack River At Outlet Of Chilliwack Lake	32	-.0221	.6561	.36	.23	72.7
Clearwater River Above Limestone Creek	30	.2277	.7847	.83	.72	89.6
Clearwater River At Outlet Of Clearwater Lake	38	.2459	.7942	.27	.12	630.8
Columbia River At Nicholson	77	-.0769	.7459	.37	.18	437.6
Columbia River At Donald	44	-.2825	.6861	.34	.44	712.6
Cooks Creek Near East Selkirk	32	-.1557	.5364	1.16	.92	45.3
Crow River At Frank	39	.0739	.7650	.68	.20	32.6
Dease River At McDame	30	-.3378	.5372	.34	.04	611.6
Drywood Creek Near Twin Butte	52	.0202	.6876	1.05	1.02	6.6
East River At St. Margarets Bay	63	-.1401	.6961	1.06	1.10	8.0
East Humber River Near Pine Grove	35	-.0390	.7394	.87	.71	23.8
Elbow River Above Glenmore Dam	44	-.0611	.6682	.90	.49	63.2
English River Near Sioux Lookout	60	.0045	.7364	.69	.38	287.3
Fish Creek Near Priddis	33	.0312	.7937	1.54	.75	17.3
Flathead River At Flathead	60	.1865	.7847	.47	.27	208.5
St. Francis River At Outlet Of Glacier Lake	37	.1301	.7681	.55	.14	213.3
Fraser River At Shelley	39	-.1261	.6440	.25	.21	3244.1
Garnish River Near Garnish	30	.1702	.6154	.56	.52	56.6
Gods River Below Allen Rapids	39	.2183	.6920	.47	.23	233.7
Grass River At Wekusko Falls	31	.1871	.6137	.57	.35	22.1
Hall (Riviere) Pres D East Hereford	40	-.0126	.7454	.36	.13	68.9
Harrison River Near Harrison Hot Springs	38	-.0030	.6207	.27	.18	1281.7
Homathko River At The Mouth	32	-.0070	.7379	.52	.66	1242.6
Incomappleux River Near Beaton	37	-.0017	.6655	.39	.62	298.1
Iskut River Below Johnson River	30	-.0675	.5586	.64	1.00	2392.3
Kabinakagami River At Highway No. 11	36	-.1381	.5526	.31	.06	227.2
Kettle River Near Laurier	59	.0668	.7096	.34	.00	591.4
Kluane River At Outlet Of Kluane Lake	36	-.1835	.6746	.25	-.20	275.5
Kootenay River At Newgate	42	.0648	.7744	.45	-.08	1614.5
Lahave River At West Northfield	73	-.0398	.7077	.78	1.52	230.4
McLeod River Above Embarras River	34	-.1685	.4906	1.33	.92	243.6
Liard River At Lower Crossing	42	.2003	.6657	.33	.11	5370.7
Manyberries Creek At Brodin s Farm	45	.0251	.6927	1.11	.57	13.1

Table 2.1a Continued

Southwest Margaree River Near Upper Margaree	70	.1378	.7561	.30	.22	38.7
McEachern Creek At International Boundary	53	-.0489	.5978	1.61	.65	22.6
Middle Brook Near Gambo	30	.1507	.7267	.38	.13	29.0
Mink Creek Near Ethelbert	34	-.0631	.6787	1.46	.68	6.6
Mistaya River Near Saskatchewan Crossing	38	-.0022	.6662	.28	.59	33.6
Moyie River At Eastport	59	.1318	.8481	.45	.07	145.9
Namakan River At Outlet Of Lac La Croix	66	.1544	.7045	.57	.23	318.5
Petite Nation (Riviere De La) Pres De Cote-Saint-Pierre	43	.0350	.5718	.52	.24	69.5
Nith River At New Hamburg	38	.1018	.7443	.58	.17	154.5
Northeast Pond River at Northeast Pond	35	.0532	.7040	.54	.75	2.2
Nottawasaga River Near Baxter	40	.1563	.7035	.69	.48	110.1
East Oakville Creek Near Omagh	32	-.0864	.6365	.56	.26	39.6
Overflowing River At Overflowing River	33	.0682	.6572	.67	.19	53.0
Pembina River Near Entwistle	34	-.1340	.5901	1.41	.80	245.5
Petite Nation (Riviere De La) a Portage-De-La-Nation	46	.0281	.6810	.40	.43	131.4
Pigeon River At Middle Falls	65	-.0071	.6914	.60	.35	128.5
Prairie Creek Near Rocky Mountain House	37	.0340	.5906	1.10	.44	36.3
Quesnel River Near Quesnel	50	.1240	.7227	.30	.19	766.8
Chilko River Near Redstone	62	-.1820	.6070	.26	.30	300.3
Richelieu (Riviere) Aux Rapides Fryers	51	.1780	.6333	.30	-.18	923.8
Rock Creek Below Horse Creek Near International Boundary	32	.0027	.7156	1.26	.54	27.0
Rolph Creek Near Kimball	53	.0779	.7198	1.40	.75	5.0
Roseau River Near Caribou	67	.1994	.6633	.63	.21	47.5
Saint John River At Fort Kent	62	.1503	.7243	.45	.09	2357.8
Salmo River Near Salmo	40	.0751	.5932	.32	.18	243.2
Saugeen River Near Port Elgin	74	-.0062	.6269	.51	.15	500.4
Shekak River At Highway No. 11	37	-.0396	.6622	.36	.21	209.0
Shogomoc Stream Near Trans Canada Highway	45	.0458	.6301	.52	.24	39.6
Similkameen River At Princeton	44	.1125	.6908	.51	.28	237.0
Skootamatta River Near Actinolite	30	.1369	.7937	.44	.22	68.6
South Thompson River At Chase	48	.1321	.6698	.31	.14	996.2
South Nation River At Spencerville	41	.0037	.6213	.60	.27	47.4
St. Mary River Near Marysville	41	-.1097	.6086	.31	.24	303.5
Stikine River At Telegraph Creek	34	-.0565	.7322	.32	.11	2377.1
Stony Creek Near Neepawa	30	-.1325	.6569	1.20	.34	11.3
Sturgeon River Near Barwick	35	.3163	.6585	.71	.32	21.6
Swiftcurrent Creek At Many Glacier	54	.0301	.5831	.71	1.59	28.8
Sydenham River Near Alvinston	40	-.0687	.6494	.58	.29	102.0
Teslin River Near Teslin	41	-.0498	.6632	.40	.19	1052.0
Tetagouche River Near West Bathurst	37	-.1392	.6201	.50	.21	76.7
North Thompson River Near Barriere	44	.0209	.7228	.27	.31	1775.0
Torrent River at Bristol s Pool	30	-.0593	.6065	.49	.44	203.9
Turtle River Near Laurier	40	.1065	.6220	1.26	.86	51.2
Upper Humber River Near Reidville	60	.2080	.6629	.32	.24	578.3
Waterhen River Near Waterhen	34	.5136	.7605	.46	.08	169.0
Whitemouth River Near Whitemouth	42	-.1000	.6767	.87	.36	83.7
Wilson River Near Dauphin	31	.0865	.6598	1.15	.55	44.7
Woody River Near Bowsman	35	.2051	.7275	1.17	.70	66.9
Yukon River Above Frank Creek	36	-.2228	.6463	.17	-.09	688.4
Arrow River Near Arrow River	30	-.1595	.6173	1.53	.76	6.1
Athabasca River Below McMurray	31	-.0522	.6314	.48	.28	2656.1
Atlin River Near Atlin	39	-.0076	.6377	.25	.18	221.0
Babine River At Babine	41	.0753	.6972	.45	.12	125.3
Barnes Creek Near Needles	38	.1769	.7474	.30	.31	34.3
Beaver River At Cold Lake Reserve	33	.1830	.6790	1.21	.86	138.5
Beaurivage (Riviere) A Sainte-Etienne	37	.2450	.7895	.44	.00	182.7
Bell (Riviere) A Senneterre - 2	36	-.1912	.5160	.60	.59	97.8

Table 2.1a Continued

Big Sheep Creek Near Rossland	40	.0555	.6454	.33	.04	48.8
Black River Near Washago	73	.1119	.7306	.34	.07	129.5
Bow River At Banff	80	-.1324	.6512	.36	.19	217.6
Brokenhead River Near Beausejour	46	.0258	.6385	1.10	.83	36.2
Campbell River At Outlet Of Campbell Lake	38	-.0931	.6373	.51	.45	421.5
Cariboo River Below Kangaroo Creek	31	.1668	.6960	.31	.18	383.1
Carrot River Near Armley	34	.0039	.6202	1.29	.55	104.8
Cascade River Near Banff	30	.0917	.7815	.66	.10	36.9
Castor River At Russell	41	.2511	.7164	.50	.00	107.1
Chilko River At Outlet Of Chilko Lake	60	-.0331	.7441	.26	.22	136.8
Clam Harbour River Near Birchtown	31	.1889	.5880	.58	.43	20.2
Clearwater River Near Rocky Mountain House	32	-.0047	.6583	1.00	.42	162.9
Clearwater River Near Clearwater Station	39	.0454	.6190	.26	.16	984.0
Columbia River Near Fairmont Hot Springs	43	-.2917	.6680	.44	.35	46.0
Conjuring Creek Near Russell	30	-.1471	.6246	1.16	.65	3.1
Cottonwood River Near Cinema	34	.0763	.5933	.52	.26	195.0
Cypress Creek Near Clearwater	30	-.2764	.6555	1.72	.60	16.8
Deer Creek At Deer Park	30	-.1069	.7351	.46	.13	7.6
Duncan River Near Howser	33	.1740	.8175	.31	.10	407.8
East Prairie River Near Enilda	30	.0880	.7481	.93	.30	133.1
Elbow River At Bragg Creek	54	-.0467	.6803	.98	.52	60.2
English River At Umfreville	67	-.1359	.6998	.74	.40	158.6
Etomami River Near Bertwell	34	.0448	.6603	.92	.17	59.4
Fish Creek Near Prospect Hill	37	.0165	.7151	.77	.42	44.7
Fraser River At Hansard	36	-.1167	.6774	.25	.22	2046.1
Fraser River At McBride	36	-.0578	.6538	.25	.23	910.3
Gander river at big chute	39	-.0151	.6306	.38	-.13	569.0
Ghost River Near Black Rock Mountain	40	.2299	.6734	.72	.40	22.7
Grand River at Loch Lomond	68	-.0636	.6955	.42	.44	18.9
Harricana (Riviere) A Amos	56	-.1638	.4826	.34	.31	190.1
Highwood River At Diebel s Ranch	38	-.0994	.6790	.76	.42	79.0
Horse Creek At International Boundary	43	-.0937	.6059	1.39	.53	8.9
Icelandic River Near Riverton	30	-.2091	.6543	1.21	.48	56.3
Indian Brook At Indian Falls	34	.1975	.6522	.32	.20	139.3
Island Lake River Near Island Lake	32	-.0773	.7043	.52	.08	167.5
Kettle River Near Ferry	60	.1331	.7382	.34	.07	339.3
Kinojevis (Riviere) En Aval Du Lac Preissac	33	-.1239	.5915	.45	.35	41.6
Kootenay River At Kootenay Crossing	41	-.0933	.5638	.36	.13	33.7
Kootenay River Near Skookumchuck	39	-.2586	.5725	.32	.39	677.4
Lardeau River At Marblehead	43	-.2178	.6478	.27	.23	282.9
Lepreau River At Lepreau	72	-.0206	.5217	.90	1.01	78.8
Lillooet River Near Pemberton	63	.0960	.6400	.37	.63	529.6
Little Saskatchewan River Near Minnedosa	30	.1651	.7277	1.06	.45	27.0
Lobstick River Near Styal	32	-.0855	.6295	1.23	.60	26.9
Lodge Creek Near alberta Boundary	38	.0520	.6966	1.54	.69	25.2
Northeast Margaree River At Margaree Valley	72	.0701	.7323	.54	.62	176.3
St. Mary River At Wycliffe	43	-.0470	.6797	.32	.25	385.3
McKinnon Creek Near McCreary	30	.0978	.6294	1.36	1.14	5.6
Mille Iles (Riviere Des) En Aval Du Lac Des Duex Montagne	35	-.2604	.5804	.53	.02	766.3
Missinaibi River At Mattice	69	.1060	.7285	.43	.18	881.0
Moose River Near Red Pass	34	.0363	.7169	.40	.25	94.3
Nagagami River At Highway No.11	38	.0002	.5602	.37	.15	121.3
Nass River Above Shumal Creek	32	-.1881	.5877	.54	.92	3841.6
Neebing River Near Thunder Bay	35	.1811	.8082	.82	.39	24.8
Nith River Near Canning	42	-.0444	.6818	.60	.06	188.8
North Magnetawan River Near Burk s Falls	73	-.0120	.6147	.47	.30	44.5
North Pine River Near Pine River	35	.0084	.6762	.78	.20	13.4

Table 2.1a Continued

Oldman River Near Waldron s Corner	39	-.0769	.7016	.86	.48	115.4
Pembina River At Jarvie	31	-.2232	.6158	1.10	.63	295.8
Pembina River Below Paddy Creek	33	-.1235	.5580	1.49	.85	193.7
Pigeon River At Outlet Of Round Lake	31	-.1402	.6134	.53	.15	195.5
Poplar River At International Boundary	56	-.0490	.6452	1.71	.73	24.9
Quesnel River At Likely	64	.1300	.7013	.32	.12	394.9
Red Deer River Near The Mouth	33	.0388	.6333	.85	.16	94.9
Richelieu (Riviere) A Saint-Jean	36	-.1122	.6474	.26	-.18	982.1
Roaring River Near Minitonas	30	-.0409	.6227	1.42	.89	35.8
Roseau River Near Dominion City	49	.0377	.5569	.85	.66	64.5
Roseway River At Lower Ohio	71	.0829	.7392	.52	.40	68.6
Salmon River Near Prince George	36	-.0339	.7110	.44	.20	210.4
Saugeen River Near Walkerton	74	.1929	.6744	.58	.33	290.4
Seal river Below Great Island	31	.0528	.5665	.45	.28	948.2
Shell River Near Inglis	38	.1306	.7305	.98	.52	22.4
Sikanni Chief River Near Fort Nelson	44	-.0901	.4247	.73	.44	198.8
Skeena River At USK	41	-.2471	.6171	.35	.43	5053.9
Slocan River Near Crescent Valley	64	.0748	.7062	.36	.18	441.8
Southwest Margaree River Near Upper Margaree	70	.1355	.7565	.30	.23	38.7
Sprague Creek Near Sprague	43	.1043	.6568	1.09	.46	19.7
Stellako River At Glenannan	39	.1617	.8170	.64	.33	74.5
St. Marys River At Stillwater	73	-.0473	.7019	.49	.47	408.6
Stuart River Near Fort St. James	56	.2177	.7512	.39	.20	322.0
Sturgeon River Near Fort Saskatchewan	54	-.1707	.5265	1.08	.77	26.8
Swift Current Creek Below Rock Creek	34	-.1198	.6917	1.12	.46	23.4
Sydenham River Near Owen Sound	43	.0199	.5915	.61	.25	30.3
North Thompson River At McLure	30	.1653	.6285	.25	.21	1867.3
Thompson River Near Spences Bridge	37	.2194	.7378	.26	.05	2833.0
Turtle River Near Mine Centre	58	-.0587	.6682	.63	.27	127.2
Twenty Mile Creek At Balls Falls	32	-.0907	.4990	.56	.16	65.3
Upsalquitch River At Upsalquitch	45	.0309	.6427	.56	.16	367.3
Waterton River Near Waterton Park	41	.0376	.6841	1.01	1.37	142.3
Whitewater Creek Near International Boundary	53	-.0228	.6497	2.11	.81	9.8
Wolf Creek At Highway No. 16A	34	-.1839	.6507	1.67	1.01	64.9
Yukon River Above Frank Creek	36	-.2228	.6463	.17	-.09	688.4

Table 2.1b Statistics of natural annual peak flows of Chinese rivers (Source of data: Ministry of Water Resources, 1985)

(n - record length in years, r(1) - lag one autocorrelation coefficient, K- Hurst's K, Cv- coefficient of variation, Cs - coefficient of skewness, mean - mean value of annual peak flows in m^3/s)

River Name	n	r(1)	K	Cv	Cs	mean
Dadu River at Tongjizhi	40	.0243	.6735	.33	.43	6234.3
Nenjiang River at Ayanqian	72	.0671	.7722	1.01	.38	2168.3
Xinan River at Luotongbu	46	.0283	.7007	.57	.35	8526.1
Fuchenj River at Lucibu	43	-.1056	.5951	.44	.57	13457.7
Yuanshui River at Lanzhiwan	50	-.1099	.5971	.50	.16	18046.8
Yangtze River at Yichan	109	.1443	.6288	.25	-.11	51244.0
Hongshihe River at Duan	41	.0191	.6790	.36	-.07	11989.5
Diersonghua River at Baishan	43	-.0982	.5577	.96	.77	3153.2
Hanjiang River at Shiquan	41	-.1759	.5600	.70	.33	7985.4
Hanjiang River at Ankang	49	.0687	.6803	.69	.20	12287.6
Yellow River at Sanmenxia	47	-.0422	.7287	.62	.44	8731.5
Yangtze River at Chuntan	71	.1324	.6530	.33	.09	52663.4
Yailingjiang River at Beibei	40	.1369	.6854	.43	.06	22585.0
Yangtze River at Changshou	86	-.0474	.7231	.35	.08	28794.2
Yalongjiang River at Xiaolangde	54	-.0045	.7107	.59	.42	8764.3
Jalingjiang River at Dongshiguan	41	-.1661	.6248	.59	.31	13366.6
Hongshuiho River at Yantan	50	-.0287	.6006	.37	-.02	10804.4
Mingjinag River at Zhipingpu	50	-.0884	.6105	.53	.66	2426.2
Hongshui River at Longtan	47	-.0414	.5533	.35	-.05	10216.8
Jiangjieho River at Shilin	45	.0538	.7538	.51	.31	8834.7
Penshui River at Penshui	45	.0465	.6548	.45	.35	10985.8
Yellow River at Daxia	48	-.1800	.5868	.40	.27	3811.7
Xiqiho River at Shaqikou	40	.1970	.5465	.46	.35	9368.0
Taiziho River at Guanyingou	40	-.1802	.5779	1.78	1.28	1447.5
Yalujiang River at Lingjiang	52	-.1242	.6381	1.16	.55	3077.6
Hunho River at Jingkang	48	-.0308	.6259	.98	.62	5415.6
Hunho River at Gaoling	48	-.1158	.6003	.98	.65	5353.1
Daduho River at Pubugou	46	.0666	.7875	.28	.47	4897.2
Jiangjieho River at Goupitan	44	.1032	.7685	.53	.31	7942.3
Yialingjiang River at Shanhuangmiao	40	.1543	.6760	.89	.16	4393.0
Jalingjiang River at Wushen	37	.1197	.6131	.59	.22	12443.0
Yalu River at Shuifeng	39	-.0159	.6373	.74	.23	13008.1
Hunjiang River at Huilong	37	-.0876	.6343	.98	.71	4410.9
Fujiang River at Guanyinchang	32	-.1038	.7381	.66	.16	1351.5
Dongliao River at Erlongshan	38	-.0073	.6950	1.79	1.18	857.7
Chaihe River at Taipingzhai	30	.0565	.6142	1.61	.86	425.1
Jinlongxi River at Yongan	32	.0132	.6696	.53	.45	2662.5
Fujiang River at Xiaoheba	35	-.1439	.7546	.66	.90	9488.6
Qujiang River at Fengtan	32	.1059	.7531	.58	-.01	15563.1
Fujiang River at Tianxiansi	31	-.0367	.7360	.91	.61	2819.7
Xiushui River at Zhelin	30	.1531	.6310	.73	.47	4916.0
Luoshui River at Changshui	30	-.0086	.5468	1.27	.51	1495.8
Hanjiang River at Huangjiagang	34	-.1239	.5607	.84	.55	15681.1
Yujiang River at Henxian	31	.0701	.8203	.46	.25	10557.7
Baohe River at Hedongdian	31	-.0106	.7419	.82	.30	1198.0
Yellow River at Guide	30	.1191	.6466	.32	.29	2399.0
Fujiang River at Taihezhen	35	-.1530	.7306	.61	.44	10016.3
Leishui River at Dongjiang	33	-.1011	.6596	.64	.38	2202.5

Table 2.1b Continued

Yellow River at Lanzhou	31	-.1643	.6118	.39	.24	3913.2
Chaihe River at Taipingzhai	30	.0565	.6142	1.61	.86	425.1
Qujiang River at Donglin	32	-.1362	.5206	.58	.23	5563.1
Dadu River at Tongjizhi	30	.0876	.6173	.33	.32	6618.3
Qujiang River at Luoduxi	33	.0473	.7341	.47	-.14	15300.3
Qujiang River at Qilituo	32	-.0661	.5778	.81	.39	8312.8
Jingshajing River at Pinshan	37	-.0512	.6582	.39	-.01	16854.9
Fujiang River at Fujiangqiao	31	.0174	.6193	.59	.14	5541.6
Fujiang River at Xiaoba	30	-.0458	.6059	.82	.39	513.3
Qujiang River at Mingyuetan	31	-.1713	.5325	.62	.34	484.0
Yangtze River at Wanxian	30	-.0580	.6782	.24	.05	51850.0
Qujiang River at Goudukou	32	.0259	.6194	.82	1.25	17550.9

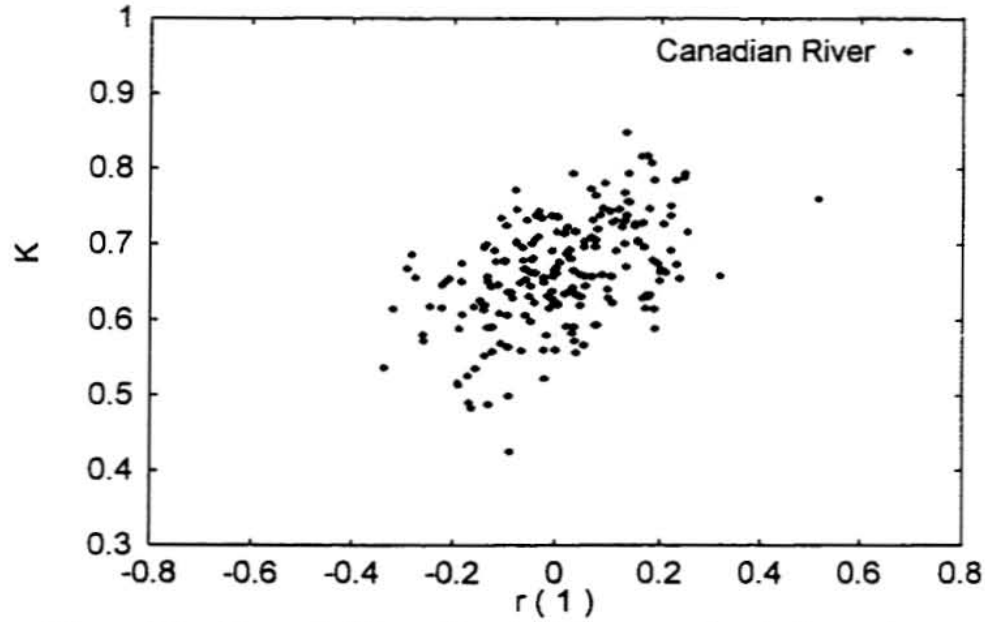

Figure 2.2a Plots of r(1) vs. Hurst's K for annual peak flows observed at Canadian rivers.

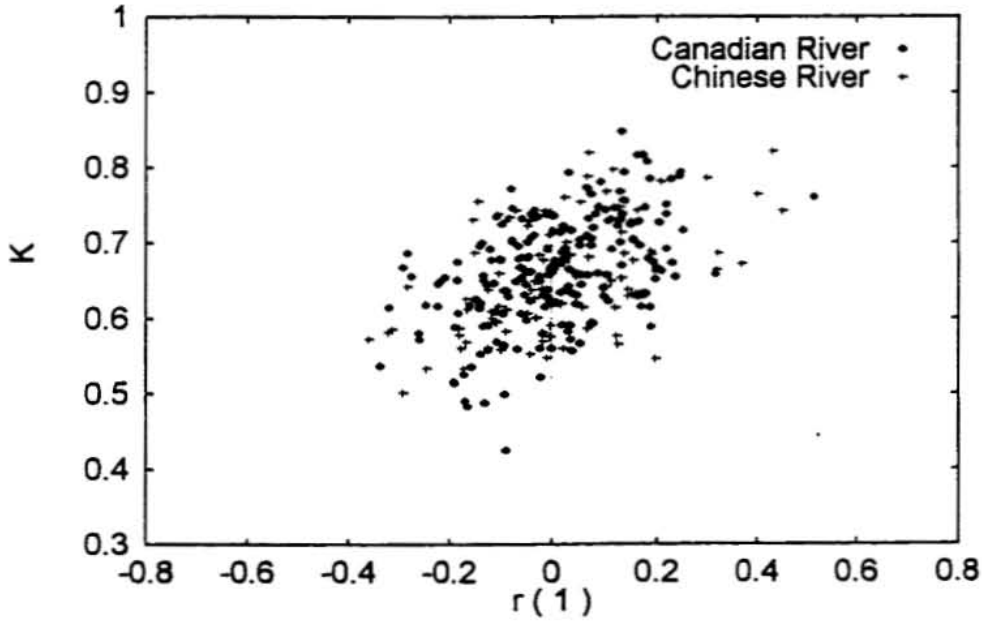

Figure 2.2b Plots of r(1) vs. Hurst's K for annual peak flows observed at Canadian and Chinese rivers.

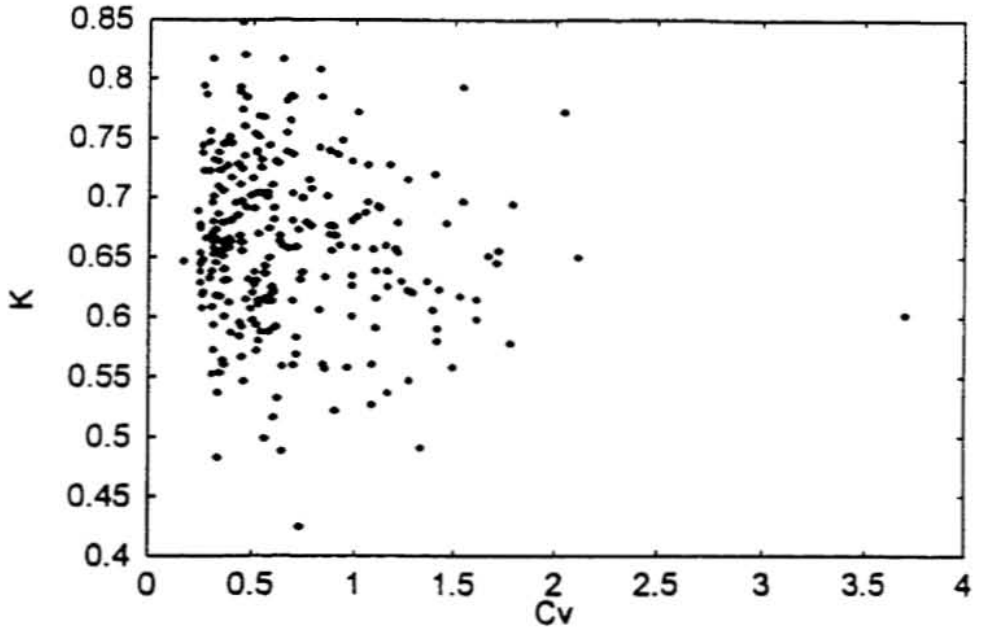

Figure 2.3a Plots of coefficient of variation (Cv) vs. Hurst's K for annual peak flows observed at Canadian and Chinese rivers.

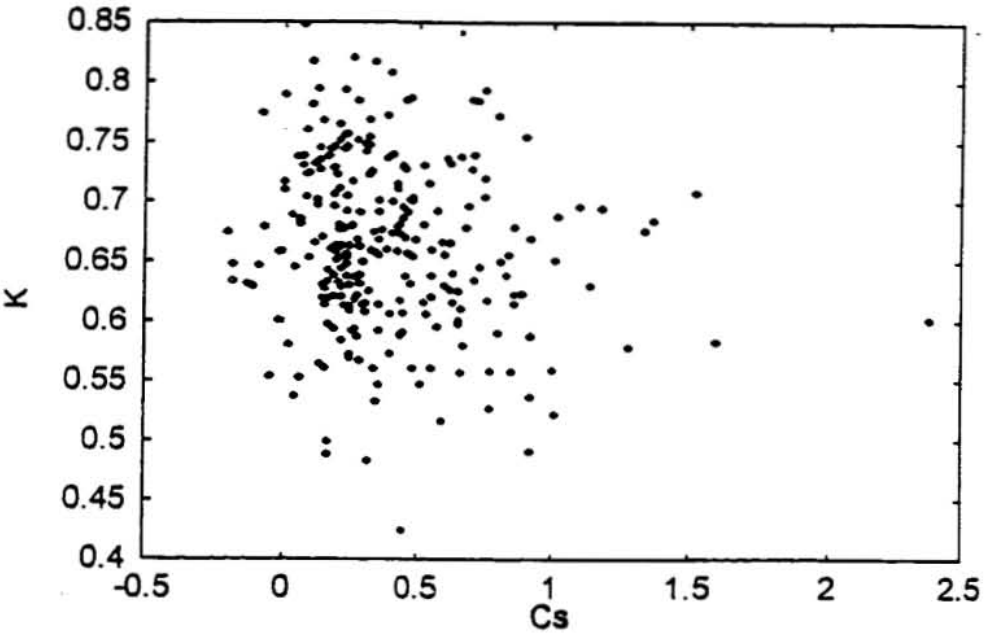

Figure 2.3b Plots of coefficient of skewness (Cs) vs. Hurst's K for annual peak flows observed at Canadian and Chinese rivers.

2.3 Theoretical Values of Lag-one Autocorrelation Coefficient and Hurst Coefficient

2.3.1 Lag-One Autocorrelation Coefficient, $\rho(1)$

In hydrology, short-term independence refers to the belief that an occurrence of a hydrological event, such as a maximum annual flood, has no influence on the probability of occurrence of previous or subsequent flood events.

Short-term persistence can be measured by the magnitude of the lag-one correlation function, $\rho(1)$, given by

$$\rho(1) = \frac{Cov(X(t), X(t+1))}{Var(X(t))} \tag{2.1}$$

where X is the basic random variable, and Cov(X(t),X(t+1)) and Var(X(t)) the covariance and variance, respectively. An estimate r(1) of the autocorrelation function, $\rho(1)$, can be obtained using (Jenkins and Watts, 1968; Box and Jenkins, 1970)

$$r(1) = [\frac{1}{n}\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})]/[\frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})^2] \tag{2.2}$$

where x_i and n are annual flow at time i, and sample size, respectively, and $\bar{x}$ is the sample mean given by:

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

An estimate of $\rho(1)$ obtained through Eq.2.2 is negatively biased and the magnitude of bias being a function of the sample length, n, and of the generating process (Wallis and O'Connell, 1972).

If an annual peak flow series is considered to be a realisation of a stochastic process, the lag-one autocorrelation coefficient r(1), can be employed to measure short-term persistence. It is sufficient to mention here that the magnitude of the lag-one autocorrelation function is theoretically zero for an independent stochastic process.

2.3.2 Hurst Coefficient, h

Long-term dependence is the presence in a time series of significant dependence between observations a long time span apart. In hydrology, it refers to the phenomenon that a quantity (e.g. river flows) can be very large or small and a period of low or high flows can be extremely long indeed (Manderlbrot and Wallis, 1968). The degree of long-term persistence is usually measured by the magnitude of the Hurst coefficient, h (Hurst, 1951, 1956).

The concept of the Hurst coefficient originates from a statistic named the "range of cumulative departure from the mean" by Hurst (1951,1956).

Let S(k) be

$$S(k) = \sum_{i=1}^{k} (x_i - \bar{x}) \tag{2.3}$$

the cumulative departures from the mean value in a discrete-time series. The range of the cumulative departure from the mean is

$$R_n = \text{Max } S(i) - \text{Min } S(j) \qquad i \in (0,1,\ldots,n) \qquad j \in (0,1,\ldots,n) \tag{2.4}$$

where Max and Min are operators defining the largest and smallest values, respectively. Hurst (1951, 1956) studied the range, R_n, both theoretically and empirically. His theoretical study showed that if the series of inflows are independent, with finite mean and variance, i.e. a white noise sequence, then

$$\frac{R_n}{s_n} = 1.25 n^h \tag{2.5}$$

where s_n and n are the standard deviation and sample size, respectively. Hurst coefficient h is theoretically 0.5. A result from Feller (1951) agrees with Hurst's result. In other words, Hurst coefficient h is constant and for a normal independent process $h=0.5$ and $\rho(1)=0.0$, and if $0.5<h<1$ it represents a long-term persistent process. Hurst coefficients in the range 0 to 0.5 represents an anti-persistent process, that is, there is a tendency to show decreases in values following previous increases, and increases following previous decreases (Mandelbrot, 1977,1982).

The Hurst coefficient is usually estimated by Hurst's K (Hurst, 1951, 1956). This estimator has a lower variance than other estimators currently in use and its calculation is simple and straightforward. Hurst's K is given by (Mandelbrot and Wallis, 1969)

$$K = \frac{\log(R_n / s_n)}{\log(n/2)} \tag{2.6}$$

where R_n is the range of cumulative departures from the mean, s_n and n are the standard deviation and sample size, respectively. K is theoretically 0.5 for independent series. K has a substantial bias in that it overestimates h for values below 0.70 and underestimates h for values above 0.70 (Wallis and Matalas, 1970).

2.4 Sampling Distributions of Hurst's K and r(1)

The probability distribution of statistics is known as the sampling distribution. It is obvious that the sampling distributions of K and lag-one autocorrelation coefficient r(1) are functions of random variables and the sample size n.

2.4.1 Relevant Mathematical Expressions

Hurst's K and the lag-one autocorrelation coefficient r(1) as statistics are random variables. A closer look at the expressions of Hurst's K and the lag-one autocorrelation coefficient r(1) given in Eqs.2.2 and 2.6, respectively, show both functions contain a kind of

summation component, that is

$$r(1) \sim \sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})$$

in Eq. 2.2 for the autocorrelation coefficient r(1), and

$$R_n \sim \sum_{i=1}^{n}(x_i - \bar{x})$$

in Eqs. 2.3 to 2.6 for Hurst's K. The similar algebraic structure expressed in the Eqs. 2.2 and 2.6 show that both of these estimates are functions of sequence of occurrence in the series. The theoretical derivation for the relationship between the population h and $\rho(1)$ has been made and used in the models of long-term persistence behaviour. One such model is the fast fractional Gaussian noise model (Mandelbrot and Wallis, 1969a), and the lag-one Markov process (Matalas and Huzzen, 1967).

To illustrate this issue, a simple example of three series of six values of flows in different orders are shown in Table 2.2 where $\bar{x}$, s_x and Cs_x are the mean value, standard deviation and coefficient of skewness, respectively.

The magnitudes of the statistics calculated from the three series, such as the mean value, $\bar{x}$, the standard deviation, s_x, and coefficient of skewness, Cs_x, are the same for the three series except for the lag-one coefficient r(1) and Hurst's K. The reason is simple,

that is, the statistics such as mean value, standard deviation, and skewness, are independent of the order of occurrence but Hurst's K and r(1) are not. This basic outcome is the primary problem in dealing with short-term and long-term behaviours of hydrologic series.

Table 2.2 Statistics for the three series having six values in different orders

Series	$\bar{x}$	s_x	Cs_x	r(1)	K
Series 1: 120 35 91 23 12 58	56.5	59.3	0.26	-0.2092	0.5651
Series 2: 91 23 58 120 35 12	56.5	59.3	0.26	-0.1728	0.4130
Series 3: 35 120 12 58 23 91	56.5	59.3	0.26	-0.6216	0.5474

2.4.2 Sampling Distribution of r(1)

The lag-one autocorrelation coefficient, r(1), calculated from Eq. 2.2 is an approximation of normal distribution if the parent population is normally distributed (Yevjevich, 1971), that is:

$$r(1) \sim N\left(-\frac{1}{n}, \sqrt{\frac{n^3 - 3n^2 + 4}{n^2(n^2 - 1)}}\right) \tag{2.7}$$

2.4.3 Sampling Distribution of Hurst's K

The distribution of Hurst's K is known to be highly skewed when the sample size is small (Mandelbrot and Wallis, 1969b). Many researchers, such as Feller (1951), Anis and Lloyd (1953), Yevjevich (1967) and Wallis and O'Connell (1973) tried to obtain a sampling distribution experimentally and theoretically, but a closed form solution to the sampling distribution of Hurst's K has not yet been obtained due to mathematical difficulties.

2.5 Statistical Relationship Between Hurst's K and r(1)

The statistical properties such as correlation and dependence between Hurst's K and lag-one correlation coefficient r(1) were investigated in order to describe behaviour of annual peak flow series.

2.5.1 Correlation Between Hurst's K and r(1)

The correlation coefficient is a measure of the strength of the linear relationship between variables. The correlation between Hurst's K and lag-one correlation coefficient r(1) can be tested by parametric and nonparametric hypothesis testing.

Before performing the hypothesis test for correlation, a normal transformation of the annual peak flows is needed in order satisfy the statistical test and modelling assumptions in a later Monte Carlo simulation experiment. Both parametric and nonparametric

hypothesis tests for correlation between Hurst's K and r(1) were carried out.

2.5.1.1 Box-Cox Normal Transformation of Observations

The Box-Cox transformation (Box and Cox, 1964) is given by

$$y_{\lambda i} = \frac{x_i^{\lambda} - 1}{\lambda}, \quad \lambda \neq 0$$
$$y_{\lambda i} = \log x_i, \quad \lambda = 0 \tag{2.8}$$

where x_i and y_i are the observations and the transformed values, respectively, and λ is the transformation parameter estimated by the probability plot correlation coefficient (PPCC) method (Lye, 1993).

The statistics of the transformed observations of annual peak flow series from Canadian and Chinese rivers are shown in Tables 2.3 and 2.4. It can be seen from the Tables 2.3 and 2.4 that the bulk of coefficient of skewness of the transformed data are close to zero and PPCC test statistic are greater than the critical value at levels of 5% and 10%. This means that the transformed data can be considered as normally distributed at significance levels of 5% and 10%.

Table 2.3a Statistics of Box-Cox transformed observations of natural annual peak flows of Canadian rivers

(n - record length in years, λ- transformation parameter, r_λ - correlation coefficient, r(1)- lag one autocorrelation coefficient, K - Hurst's K, Cs - coefficient of skewness)

River Name	n	λ	r_λ	r(1)	K	Cs
Adams River Near Squilax	42	.210	.986	.2170	.6542	-.01
Ashnola River Near Keremeds	42	.060	.990	-.3397	.5942	.00
Athabasca River At Athabasca	47	-.960	.991	-.2232	.5963	-.06
Ausable River Near Springbank	43	.330	.990	-.1100	.5427	-.04
Badger Creek Near Cartwright	30	.090	.994	-.1113	.7410	.05
Battle River Near Unwin	36	.000	.993	.0926	.6053	-.01
Bear River East Branch At Bear River	35	-.600	.991	.3331	.6827	-.01
Beaverbank River Near Kinsac	67	-.090	.995	-.0854	.7179	.01
Berens River At Outlet Of Long Lake	31	.390	.991	-.0614	.6156	.05
Boundary Creek Near Porthill	61	.420	.987	.1589	.7540	.06
Bowron River Near Wells	33	1.140	.992	.0594	.6565	.01
Bulkley River At Quick	58	.330	.995	.1917	.6363	-.01
Carrick Creek Near Carlsrume	35	.210	.986	.0895	.7500	.05
Carrot River Near Smoky Burn	34	.240	.990	-.0395	.4783	.04
Castle River Near Beaver Mines	44	-.150	.975	.0488	.7163	.10
Chilliwack River At Vedder Crossing	32	-.150	.987	-.0471	.6565	.05
Chilliwack River At Outlet Of Chilliwack Lake	32	-.240	.996	-.0334	.6200	-.01
Clearwater River Above Limestone Creek	30	-.540	.992	.2542	.7952	.04
Clearwater River At Outlet Of Clearwater Lake	38	.150	.988	.2923	.8013	.02
Columbia River At Nicholson	77	.240	.996	-.0554	.7493	.02
Columbia River At Donald	44	-.570	.992	-.2801	.7071	.00
Cooks Creek Near East Selkirk	32	.330	.974	-.1437	.5896	.11
Crow River At Frank	39	.270	.987	.1275	.7813	-.05
Dease River At McDame	30	.840	.993	-.3365	.5335	.02
Drywood Creek Near Twin Butte	52	-.060	.961	.0816	.7308	.06
East River At St. Margarets Bay	63	-.270	.996	-.0956	.6589	.01
East Humber River Near Pine Grove	35	.180	.984	-.0010	.7766	-.06
Elbow River Above Glenmore Dam	44	-.240	.993	.0524	.6536	-.02
English River Near Sioux Lookout	60	.060	.996	.1282	.7719	-.01
Fish Creek Near Priddis	33	-.360	.992	.2180	.7979	-.01
Flathead River At Flathead	60	.390	.990	.2012	.7839	.06
St. Francis River At Outlet Of Glacier Lake	37	.510	.993	.1490	.7709	-.01
Fraser River At Shelley	39	.060	.982	-.1260	.6745	.00
Garnish River Near Garnish	30	-.600	.993	.3070	.6308	.02
Gods River Below Allen Rapids	39	.300	.996	.2565	.6946	.03
Grass River At Wekusko Falls	31	.120	.994	.1683	.6226	-.02
Hall (Riviere) Pres D East Hereford	40	.510	.990	-.0459	.7421	.00
Harrison River Near Harrison Hot Springs	38	.030	.996	.0058	.5955	.01
Homathko River At The Mouth	32	-1.440	.994	.1930	.8295	.09
Incomappleux River Near Beaton	37	-1.350	.991	.1243	.6467	-.01
Iskut River Below Johnson River	30	-1.470	.991	.0253	.5610	.07
Kabinakagami River At Highway No. 11	36	.660	.996	-.1507	.5527	.00
Kettle River Near Laurier	59	1.050	.994	.0648	.7087	.05
Kluane River At Outlet Of Kluane Lake	36	1.920	.985	-.1710	.6634	-.04
Kootenay River At Newgate	42	1.200	.988	.0593	.7720	.04
Lahave River At West Northfield	73	-.390	.986	-.0164	.7004	-.10
McLeod River Above Embarras River	34	-.450	.989	-.0192	.5228	.07

Table 2.3a Continued

Mink Creek Near Ethelbert	34	.150	.983	.0540	.6904	.04
Mistaya River Near Saskatchewan Crossing	38	-1.410	.990	.0326	.6387	-.01
Moyie River At Eastport	59	.810	.996	.1370	.8486	.00
Namakan River At Outlet Of Lac La Croix	66	.570	.988	.2004	.7155	-.04
Petite Nation (Riviere De La) Pres De Cote-Saint-Pierre	43	.060	.994	.0377	.6334	.00
Nith River At New Hamburg	38	.300	.990	.0359	.7287	.00
Northeast Pond River at Northeast Pond	35	-.360	.973	.0787	.6967	-.06
Nottawasaga River Near Baxter	40	-.090	.993	.0978	.6473	.01
East Oakville Creek Near Omagh	32	.480	.986	-.0804	.6600	.06
Overflowing River At Overflowing River	33	.540	.991	.1200	.6848	.00
Pembina River Near Entwistle	34	-.660	.991	-.1161	.5639	.00
Petite Nation (Riviere De La) a Portage-De-La-Nation	46	-.450	.998	.1131	.6836	-.01
Pigeon River At Middle Falls	65	.180	.992	.0110	.6830	.04
Prairie Creek Near Rocky Mountain House	37	-.120	.988	.1523	.6320	-.03
Quesnel River Near Quesnel	50	-.150	.995	.1792	.7394	-.01
Chilko River Near Redstone	62	-.600	.991	-.1403	.6037	.00
Richelieu (Riviere) Aux Rapides Fryers	51	2.070	.990	.1472	.6226	-.08
Rock Creek Below Horse Creek Near International Boundary	32	.360	.986	-.0260	.7538	.01
Rolph Creek Near Kimball	53	.240	.993	.1821	.7845	-.01
Roseau River Near Caribou	67	.600	.996	.2035	.6646	.02
Saint John River At Fort Kent	62	.690	.995	.1653	.7235	.01
Salmo River Near Salmo	40	.240	.990	.1134	.6056	.02
Saugeen River Near Port Elgin	74	.600	.996	-.0046	.6247	.01
Shekak River At Highway No. 11	37	-.240	.995	-.0891	.6370	.00
Shogomoc Stream Near Trans Canada Highway	45	.330	.996	.0638	.6483	-.01
Similkameen River At Princeton	44	.030	.994	.0724	.7002	.02
Skootamatta River Near Actinolite	30	.450	.986	.0922	.7896	-.03
South Thompson River At Chase	48	.300	.996	.1604	.6885	-.01
South Nation River At Spencerville	41	.300	.993	.0473	.6279	-.01
St. Mary River Near Marysville	41	-.300	.992	-.0791	.6122	-.03
Stikine River At Telegraph Creek	34	.360	.990	-.0537	.7117	-.03
Stony Creek Near Neepawa	30	.360	.990	-.1078	.6700	-.04
Sturgeon River Near Barwick	35	.420	.990	.3577	.6821	-.02
Swiftcurrent Creek At Many Glacier	54	-.540	.973	.0199	.6127	-.22
Sydenham River Near Alvinston	40	.390	.984	-.0408	.6772	.04
Teslin River Near Teslin	41	.480	.991	-.0973	.6503	.02
Tetagouche River Near West Bathurst	37	.360	.995	-.0836	.6474	.01
North Thompson River Near Barriere	44	-.510	.993	.0588	.7477	.02
Torrent River at Bristol s Pool	30	-.540	.994	-.0369	.6180	.03
Turtle River Near Laurier	40	-.090	.992	.0582	.6411	-.04
Upper Humber River Near Reidville	60	.060	.994	.1751	.6475	-.01
Waterhen River Near Waterhen	34	.810	.970	.5363	.7684	-.07
Whitemouth River Near Whitemouth	42	.450	.995	-.0521	.6818	.05
Wilson River Near Dauphin	31	.450	.985	.0698	.7108	.05
Woody River Near Bowsman	35	.300	.986	.2373	.7432	.09
Yukon River Above Frank Creek	36	1.710	.992	-.2282	.6461	-.07
Arrow River Near Arrow River	30	.180	.990	-.1409	.6140	-.01
Athabasca River Below McMurray	31	-.630	.988	-.1082	.6983	-.06
Atlin River Near Atlin	39	-.150	.992	.0131	.6389	-.02
Babine River At Babine	41	.210	.990	.0955	.6983	-.07
Barnes Creek Near Needles	38	-.840	.989	.2070	.7379	-.06
Beaver River At Cold Lake Reserve	33	-.150	.994	.2119	.7001	.00
Beaurivage (Riviere) A Sainte-Etienne	37	1.080	.988	.2466	.7889	.06
Bell (Riviere) A Senneterre - 2	36	-.480	.993	-.1732	.5069	-.01
Big Sheep Creek Near Rossland	40	.780	.996	.0575	.6437	-.01
Black River Near Washago	73	.840	.992	.1078	.7312	.04
Bow River At Banff	80	.030	.996	-.1306	.6394	-.01

Table 2.3a Continued

Brokenhead River Near Beausejour	46	.240	.989	.0944	.6903	.03
Campbell River At Outlet Of Campbell Lake	38	-.690	.992	-.0271	.7252	.04
Cariboo River Below Kangaroo Creek	31	.360	.987	.1870	.7093	.09
Carrot River Near Armley	34	.300	.985	.0538	.5596	.01
Cascade River Near Banff	30	.840	.995	.1207	.7844	.04
Castor River At Russell	41	.960	.993	.2521	.7168	-.04
Chilko River At Outlet Of Chilko Lake	60	-.360	.995	-.0345	.7354	.01
Clam Harbour River Near Birchtown	31	-.510	.984	.1924	.6390	.00
Clearwater River Near Rocky Mountain House	32	-.210	.985	.1336	.7023	-.02
Clearwater River Near Clearwater Station	39	-.210	.991	.0807	.6355	.01
Columbia River Near Fairmont Hot Springs	43	-.360	.987	-.2685	.6882	.02
Conjuring Creek Near Russell	30	.450	.973	-.2180	.5695	.21
Cottonwood River Near Cinema	34	.270	.990	.0711	.6255	.01
Cypress Creek Near Clearwater	30	.210	.987	-.3209	.5421	-.02
Deer Creek At Deer Park	30	.270	.987	-.1101	.7434	.01
Duncan River Near Howser	33	.330	.994	.1876	.8194	-.01
East Prairie River Near Enilda	30	.180	.987	.1558	.7468	-.03
Elbow River At Bragg Creek	54	-.360	.994	.0425	.6774	-.03
English River At Umfreville	67	.120	.997	-.0349	.7237	.00
Etomami River Near Bertwell	34	.660	.989	.0563	.6529	.00
Fish Creek Near Prospect Hill	37	.330	.994	.0601	.7107	.01
Fraser River At Hansard	36	.090	.985	-.1249	.6973	-.02
Fraser River At McBride	36	-.540	.988	-.0779	.6767	-.03
Gander river at big chute	39	1.380	.991	-.0187	.6325	.00
Ghost River Near Black Rock Mountain	40	-.150	.993	.2505	.6823	.03
Grand River at Loch Lomond	68	-.180	.986	-.0464	.7063	-.07
Harricana (Riviere) A Amos	56	.150	.980	-.1530	.5197	.04
Highwood River At Diebel s Ranch	38	.060	.991	-.0260	.7175	.03
Horse Creek At International Boundary	43	.240	.995	-.0193	.6572	-.02
Icelandic River Near Riverton	30	.180	.993	-.2473	.5248	.00
Indian Brook At Indian Falls	34	.360	.991	.2119	.6579	.02
Island Lake River Near Island Lake	32	.660	.990	-.0672	.7057	-.02
Kettle River Near Ferry	60	.720	.996	.1500	.7435	.00
Kinojevis (Riviere) En Aval Du Lac Preissac	33	.000	.990	-.0936	.5843	.03
Kootenay River At Kootenay Crossing	41	.420	.988	-.0955	.5587	.02
Kootenay River Near Skookumchuck	39	-1.170	.996	-.2146	.6103	.02
Lardeau River At Marblehead	43	-.690	.996	-.2206	.6443	-.01
Lepreau River At Lepreau	72	-.750	.996	.0007	.5885	-.01
Lillooet River Near Pemberton	63	-1.290	.994	.0894	.5676	-.01
Little Saskatchewan River Near Minnedosa	30	.180	.993	.1496	.6858	.01
Lobstick River Near Styal	32	-.090	.988	.0671	.7378	-.04
Lodge Creek Near alberta Boundary	38	.240	.989	.0553	.7081	-.01
Northeast Margaree River At Margaree Valley	72	-.360	.994	.1419	.7530	.02
St. Mary River At Wycliffe	43	-.570	.994	-.0057	.6740	-.03
McKinnon Creek Near McCreary	30	-.210	.989	.1800	.7015	.01
Mille Iles (Riviere Des) En Aval Du Lac Des Duex Montagne	35	.960	.984	-.2594	.5824	.03
Missinaibi River At Mattice	69	.420	.996	.1525	.7540	.00
Moose River Near Red Pass	34	-.510	.993	-.0282	.7067	-.01
Nagagami River At Highway No.11	38	.150	.990	-.0348	.5393	-.03
Nass River Above Shumal Creek	32	-1.470	.983	-.3769	.5955	.26
Neebing River Near Thunder Bay	35	.360	.994	.2395	.8061	.02
Nith River Near Canning	42	.660	.987	-.0480	.6748	-.07
North Magnetawan River Near Burk s Falls	73	.090	.996	-.0558	.5846	.03
North Pine River Near Pine River	35	.690	.990	.0693	.6748	.04
Oldman River Near Waldron s Corner	39	-.030	.996	-.0771	.7491	.00
Pembina River At Jarvie	31	-.180	.991	-.1126	.6493	.00

Table 2.3a Continued

Pembina River Below Paddy Creek	33	-.600	.985	-.0061	.5751	.09
Pigeon River At Outlet Of Round Lake	31	.540	.992	-.1056	.6410	.01
Poplar River At International Boundary	56	.090	.995	-.1505	.6617	-.05
Quesnel River At Likely	64	.180	.993	.1753	.7205	-.01
Red Deer River Near The Mouth	33	.540	.989	.1051	.6500	.00
Richelieu (Riviere) A Saint-Jean	36	1.920	.989	-.1166	.6535	.00
Roaring River Near Minitonas	30	.150	.982	.0927	.7030	.06
Roseau River Near Dominion City	49	.180	.991	.0329	.5626	.01
Roseway River At Lower Ohio	71	-.300	.995	.0721	.7377	-.02
Salmon River Near Prince George	36	.210	.994	-.0071	.7057	.01
Saugeen River Near Walkerton	74	.210	.994	.1387	.6445	.02
Seal river Below Great Island	31	.000	.988	.0156	.5743	.01
Shell River Near Inglis	38	.150	.993	.2419	.7823	.06
Sikanni Chief River Near Fort Nelson	44	-.120	.981	-.0505	.5143	.09
Skeena River At USK	41	-.510	.996	-.2462	.6571	-.01
Slocan River Near Crescent Valley	64	.180	.991	.1206	.7239	-.02
Southwest Margaree River Near Upper Margaree	70	-.150	.994	.1340	.7486	-.02
Sprague Creek Near Sprague	43	.390	.996	.1755	.6636	.00
Stellako River At Glenannan	39	.120	.993	.2117	.8288	-.02
St. Marys River At Stillwater	73	-.270	.996	-.0046	.6491	-.01
Stuart River Near Fort St. James	56	.120	.993	.2436	.7507	.00
Sturgeon River Near Fort Saskatchewan	54	.030	.997	-.1468	.6045	.00
Swift Current Creek Below Rock Creek	34	.360	.990	-.1035	.6916	.04
Sydenham River Near Owen Sound	43	.420	.994	-.0160	.5732	.05
North Thompson River At McLure	30	-.360	.996	.1873	.6540	.00
Thompson River Near Spences Bridge	37	.630	.993	.2354	.7396	-.02
Turtle River Near Mine Centre	58	.300	.996	.0335	.6933	.00
Twenty Mile Creek At Balls Falls	32	.570	.989	-.1152	.4915	-.01
Upsalquitch River At Upsalquitch	45	.390	.991	.0550	.6552	-.01
Waterton River Near Waterton Park	41	-.660	.971	.1325	.7311	-.19
Liard River At Lower Crossing	42	.690	.976	.1936	.6641	.04
Manyberries Creek At Brodin s Farm	45	.390	.991	.0232	.6857	.01
Southwest Margaree River Near Upper Margaree	70	-.090	.994	.1355	.7492	-.01
McEachern Creek At International Boundary	53	.300	.994	-.0209	.5592	.06
Middle Brook Near Gambo	30	.420	.991	.0997	.7069	.03
Whitewater Creek Near International Boundary	53	.150	.989	.0663	.6151	.01
Wolf Creek At Highway No. 16A	34	-.240	.996	-.0899	.7317	.03
Yukon River Above Frank Creek	36	1.710	.992	-.2282	.6461	-.07

Table 2.3b Statistics of Box-Cox transformed observations of natural annual peak flows of Chinese rivers

(n - record length, λ- transformation parameter, r_λ-correlation coefficient, r(1)- lag one autocorrelation coefficient, K - Hurst's K, Cs - coefficient of skewness)

River Name	n	λ	r_λ	r(1)	K	Cs
Dadu River at Tongjizhi	40	-1.380	.995	.0529	.7122	-.13
Nenjiang River at Ayanqian	72	.270	.993	.1167	.8050	-.01
Xinan River at Luotongbu	46	.180	.996	.0425	.7092	.01
Fuchenj River at Lucibu	43	-.630	.996	-.1135	.5988	-.02
Yuanshui River at Lanzhiwan	50	.210	.991	-.0843	.5530	-.04
Yangtze River at Yichan	109	1.590	.997	.1419	.6268	-.02
Hongshihe River at Duan	41	1.320	.995	.0457	.6814	.02
Diersonghua River at Baishan	43	.060	.985	-.0239	.6436	.00
Hanjiang River at Shiquan	41	.270	.993	-.1241	.5309	.00
Hanjiang River at Ankang	49	.570	.992	.0998	.6965	-.02
Yellow River at Sanmenxia	47	-.120	.996	.0004	.7508	.00
Yangtze River at Chuntan	71	.480	.994	.1319	.6626	.00
Yailingjiang River at Beibei	40	.750	.995	.1695	.6947	.00
Yangtze River at Changshou	86	.720	.994	-.0538	.7221	.01
Yalongjiang River at Xiaolangde	54	-.030	.996	.0259	.7384	-.01
Jalingjiang River at Dongshiguan	41	.090	.996	-.1387	.5689	.00
Hongshuiho River at Yantan	50	1.080	.992	-.0241	.5998	.02
Mingjinag River at Zhipingpu	50	-1.440	.991	.0587	.6163	.06
Hongshui River at Longtan	47	1.230	.993	-.0263	.5511	.03
Jiangjieho River at Shilin	45	.270	.981	-.0223	.7314	.04
Penshui River at Penshui	45	-.780	.993	-.0974	.5832	.01
Yellow River at Daxia	48	-.120	.997	-.1844	.5693	.01
Xiqiho River at Shaqikou	40	-.210	.993	.1896	.5462	.04
Taiziho River at Guanyingou	40	-.030	.993	-.2724	.5945	.01
Yalujiang River at Lingjiang	52	-.120	.994	-.0666	.6519	.01
Hunho River at Jingkang	48	.030	.995	-.0725	.6097	.02
Hunho River at Gaoling	48	.030	.995	-.1233	.5905	.04
Daduho River at Pubugou	46	-1.380	.995	.0831	.7771	.07
Jiangjieho River at Goupitan	44	.210	.982	.0136	.7428	.00
Yialingjiang River at Shahuagmiao	40	.350	.995	.1093	.6952	.00
Jalingjiang River at Wushen	37	.630	.994	.1588	.6065	.09
Yalu River at Shuifeng	39	.630	.987	-.0080	.6348	.05
Hunjiang River at Huilong	37	.060	.984	-.0081	.6568	.05
Fujiang River at Guanyinchang	32	.570	.993	-.0841	.7390	-.03
Dongliao River at Erlongshan	38	-1.500	.987	-.0828	.5904	-.05
Chaihe River at Taipingzhai	30	.000	.994	-.0243	.5185	.05
Jinlongxi River at Yongan	32	-.120	.990	-.0081	.6331	.04
Fujiang River at Xiaoheba	35	-.390	.990	-.0653	.7782	-.05
Qujiang River at Fengtan	32	.990	.990	.1063	.7530	-.05
Fujiang River at Tianxiansi	31	-.210	.995	.0460	.7191	-.01
Xiushui River at Zhelin	30	.120	.989	.0066	.5583	.03
Luoshui River at Changshui	30	.240	.988	-.1209	.5154	.05
Hanjiang River at Huangjiagang	34	.210	.985	-.0565	.5900	.03
Yujiang River at Henxian	31	.060	.991	.0450	.8004	.01
Fujiang River at Taihezhen	35	.360	.982	-.1340	.7347	.00
Leishui River at Dongjiang	33	.060	.995	-.0288	.6661	.00
Baohe River at Hedongdian	31	-.030	.990	.0456	.7158	-.02

Table 2.3b Continued

Qujiang River at Luoduxi	33	1.530	.992	.0417	.7289	-.02
Qujiang River at Qilituo	32	.390	.988	.0018	.5573	.02
Yellow River at Guide	30	-.480	.997	.1241	.6558	.01
Yellow River at Lanzhou	31	-.240	.993	-.1344	.5964	.02
Jingshajing River at Pinshan	37	1.200	.964	-.0500	.6609	.37
Fujiang River at Fujiangqiao	31	.570	.993	.0000	.6236	-.02
Fujiang River at Xiaoba	30	.180	.995	.0017	.5652	.00
Qujiang River at Mingyuetan	31	.300	.991	-.1622	.4957	.04
Chaihe River at Taipingzhai	30	.000	.994	-.0243	.5185	.05
Qujiang River at Donglin	32	.360	.993	-.1173	.5363	.01
Dadu River at Tongjiezhi	30	-.420	.994	-.0628	.6484	-.04
Yangtze River at Wanxian	30	.600	.990	-.0485	.6906	.02
Qujiang River at Goudukou	32	-.090	.936	.1419	.7256	-.25

Table 2.4 Statistics of Hurst's K and r(1) for Box-Cox transformed data observed in some Canadian and Chinese rivers

Source	Size (n)	Statistics	Mean	Standard error	Skewness	Correlation coefficient
Canadian Rivers	198	Hurst's K	0.6728	0.0728	-0.2021	0.5097
		r(1)	0.0309	0.1483	0.1480	
Chinese Rivers	60	Hurst's K	0.6522	0.0710	0.2922	0.3373
		r(1)	-0.0184	0.0999	0.1395	

2.5.1.2 t-test of Correlation

A parametric hypothesis test, t-test, was used to test the correlation between two random variable, Hurst's K and r(1). An assumption is made in the hypothesis that Hurst's K and r(1) are random variables from a bivariate normal distribution. If population correlation coefficient, ρ, is theoretically zero, and sample estimate of ρ given by r, then the quantity

$$t = r\sqrt{(n-2)/(1-r^2)} \qquad (2.9)$$

has a t-distribution with (n-2) degrees of freedom, where r and n are the estimate of ρ and the sample size, respectively.

Two-tailed test can be used:

H_o: $\rho = 0$

H_1: $\rho \neq 0$

Test statistic: t-statistic has a Student's t distribution with ν = (n-2) degrees of freedom

Rejection region: $| t | > t_{\alpha/2,\, n-2}$

Thus the test statistic, t, is calculated from Eq.2.9 and Ho: $\rho=0$ is rejected at a given significance level α if $| t | > t_{\alpha/2,\, n-2}$. The result of t-tests for correlation between Hurst's K and r(1) calculated from the Box-Cox transformed data is shown in Table 2.5. It indicates a correlation between Hurst's K and lag-one autocorrelation coefficient r(1) at the 5% and 10% levels of significance.

Table 2.5 t-tests for correlation between Hurst's K and r(1) from transformed data

Source	n	t	$t_{\alpha/2,\, n-2}$		Conclusion	
			$\alpha=5\%$	$\alpha=10\%$	$\alpha=5\%$	$\alpha=10\%$
Canadian Rivers	198	8.294	1.960	1.645	reject H_o	Reject H_o
Chinese Rivers	60	2.729	1.960	1.645	reject H_o	Reject H_o

2.5.1.3 Spearman's Nonparametric Test of Correlation

The t-test of correlation is based on the assumption that the variables tested are randomly sampled from a bivariate normal distribution. As mentioned, the lag-one autocorrelation coefficient, r(1), calculated from Eq.2.2 is approximately normally distributed if the parent population is normally distributed, but the distribution of Hurst's K is unknown. Because of the uncertainty concerning the assumption about the form of the population distributions, nonparametric methods often lead to a more efficient decision in hypotheses testing. So Spearman's rank correlation coefficient, r_s (Olds, 1938), a nonparametric statistics in testing correlation between two random variables, Hurst's K and r(1), was also used in this study.

The Spearman's rank correlation coefficient, r_s, is calculated by using the rank as the paired measurements on the two variables, X and Y, in the formula for correlation coefficient r. Thus r_s is given by

$$r_s = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}} \qquad (2.10)$$

where u_i is the rank of x_i (i=1, 2, ..., n) and v_i is the rank of y_i (i=1, 2, ..., n).

The null hypothesis that the population value of Spearman's rank correlation coefficient, $\rho_s = 0$ implies there is no correlation between u and v. The two-tailed test is

carried out:

H_o: $\rho_s = 0$

H_1: $\rho_s \neq 0$

Test statistic: r_s, the sample Spearman's rank correlation expressed by Eq. 2.10

Rejection region: $|r_s| > r_{s\,\alpha/2,\,n}$

where $r_{s\,\alpha/2,\,n}$ is the critical value for Spearman's correlation coefficient.

The Spearman's nonparametric test for correlation shown in Table 2.6 indicates correlation between Hurst's K and r(1) at the 5% and 10% levels of significance.

Both parametric and nonparametric tests identify correlation between the two random variables, Hurst's K and r(1), that is, a linear relationship links the long- and short-term behaviour which are measured by Hurst's K and the lag-one autocorrelation coefficient r(1), respectively. Based on this, an empirical probability approach for dealing with serial long- and short-term behaviours will be developed in the next chapter.

Table 2.6 Spearman's nonparametric test for correlation between Hurst's K and r(1) from transformed data

Source	n	r_s	$r_{s\,\alpha/2,n}$		Conclusion	
			$\alpha=5\%$	$\alpha=10\%$	$\alpha=5\%$	$\alpha=10\%$
Canadian Rivers	198	0.498	0.364	0.305	Reject H_o	Reject H_o
Chinese Rivers	60	0.393	0.364	0.305	Reject H_o	Reject H_o

2.5.2 Dependence Between Hurst's K and r(1)

The dependence or independence between Hurst's K and lag-one correlation coefficient r(1) can be tested by parametric and nonparametric hypothesis testing. If the null hypothesis of independence is rejected, there is dependence between Hurst's K and r(1) at significant levels.

2.5.2.1 ρ-tests of Dependence

A parametric hypothesis test, ρ-test, for dependence was made to test the two variables, Hurst's K and r(1). Assume that the calculated values of Hurst's K and r(1) from Box-Cox transformation data by PPCC method shown in Table 2.3 be n samples drawn from a bivariate normal distribution, $N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ the independence of those two random variables will be tested. The statistical hypothesis is

$$\text{independence: } \rho = 0.$$

Others are

$$\text{positive correlation : } \rho > 0$$

$$\text{negative correlation : } \rho < 0$$

The test is: rejection if

$$|\hat{\rho}| \geq c' \tag{2.11}$$

where $\hat{\rho}$ is the estimate of ρ, and $c'=z/n^{1/2}$, and z is approximately N(0,1). The results of the ρ-test for independence of Hurst's K and r(1) shown in Table 2.7 rejected independence at the 5% and 10% levels of significance. i.e., the Hurst's K and r(1) are dependence at the 5% and 10% levels of significance.

Table 2. 7 ρ-tests for independence between Hurst's K and r(1) from transformed data

Source	n	$\hat{\rho}$	c'		Conclusion	
			$\alpha=5\%$	$\alpha=10\%$	$\alpha=5\%$	$\alpha=10\%$
Canadian Rivers	198	0.5097	0.1393	0.1173	reject H_o	Reject H_o
Chinese Rivers	60	0.3373	0.2530	0.2130	reject H_o	Reject H_o

2.5.2.2 χ^2 - Nonparametric test of Dependence

Hypothesis testing for the independence of two random variables which both come from a joint normal distribution is equivalent to the ρ-test for correlation. That is, if variables X and Y are independent then $\rho_{x,y}=0$. But, the situation here is that lag-one autocorrelation coefficient r(1) is approximately normally distributed (see Yevjevich, 1971), but the Hurst's K distribution is unknown. Hence, a nonparametric test for dependence between Hurst's K and r(1) suggests a more efficient test for independence, is required.

A classical nonparametric test for independence is provided by the ubiquitous χ^2 (see

Gibbons, 1971; Lehmann, 1975). Suppose $(K_1, r(1)_1)$, $(K_2, r(1)_2, \ldots, (K_n, r(1)_n)$ be n samples drawn from the unknown but same distribution. A quantity for the nonparametric hypothesis test of independence is given by

$$\eta = n\sum_{i=1}^{m}\sum_{j=1}^{s}[n_{ij} - \frac{n_{i.}n_{.j}}{n}]^2 / n_{i.}n_{.j} \qquad (2.12)$$

has a χ^2 distribution with (m-1)(s-1) degree of freedom if n_{ij} is large enough for all i, j, where n_{ij} is the accounted number of occurrence and its meaning is shown in Table 2.8a.

Based on the annual peak flows observed at Canadian and Chinese rivers, the χ^2 - test for the independence was made. If the number of observations n is large, the test statistic χ^2 can be shown to possess, approximately, a chi-square distribution. The observed data from Canadian and Chinese rivers were combined for testing. To study the data normally distributed in Chapter 3, transformed data also were tested. The results of both observed and transformed data are shown in Table 2.8b and 2.8c. The tables indicate that the null hypothesis of independence between Hurst's K and r(1) for Canadian and Chinese rivers is rejected at the significance levels of 5% and 10%.

Both parametric and nonparametric tests indicate dependence between the two random variables, Hurst's K and r(1), that is, occurrence of one variable, for example, lag-one autocorrelation r(1), affects the occurrence of Hurst's K and vice versa. Based on the dependence between r(1) and Hurst's K, a theoretical assumption about conditional

probability for long-term persistence given short-term independence will be developed in the next chapter.

Table 2.8a A contingency table

	1	2	...	s	$n_{i.}$
1	n_{11}	n_{12}	n_{13}	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	n_{2s}	$n_{2.}$
:	:	:	:	:	:
m	n_{m1}	n_{m2}	n_{m3}	n_{ms}	$n_{m.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.s}$	n

Table 2.8b A contingency table for testing the independence between Hurst's K and r(1) for the observed peak flow from Canadian and Chinese rivers

	Hurst's K			
r(1)	K>0.658	0.580<K≤0.658	K≤0.580	Total
r(1)≤-0.05	26	41	22	89
-0.05<r(1)≤ 0.05	41	28	8	77
R(1)> 0.05	64	26	2	92
Total	131	95	32	258
$\chi^2_{\alpha=10\%,4}=7.779$		$\chi^2_{\alpha=5\%,4}=9.488$	$\chi^2=37.44$	
$\chi^2>\chi^2_{\alpha=10\%,4}$		$\chi^2>\chi^2_{\alpha=5\%,4}$	Reject H_o	

Table 2.8c A contingency table for testing the independence between Hurst's K and r(1) for the transformed data observed at Canadian and Chinese rivers

	Hurst's K			
r(1)	K>0.658	0.580<K≤0.658	K≤ 0.580	Total
r(1)≤-0.05	27	34	19	80
-0.05<r(1)≤0.05	31	22	10	63
R(1)> 0.05	78	34	3	115
Total	136	90	32	258
$\chi^2_{\alpha=10\%,4}=7.779$		$\chi^2_{\alpha=5\%,4}=9.488$	$\chi^2=30.66$	
$\chi^2>\chi^2_{\alpha=10\%,4}$		$\chi^2>\chi^2_{\alpha=5\%,4}$	Reject H_o	

2.6 Variation of Hurst's K and r(1)

To demonstrate the variation of Hurst's K and the lag-one autocorrelation coefficient r(1), a simple Monte Carlo simulation for a normal distribution N(0,1) was made. The steps are as follows:

1) For a given sample size n, 25,000 realizations are sampled from a normal independent process which has a mean of zero and standard deviation of one, and lag-one autocorrelation of zero.
2) Lag-one autocorrelation coefficient r(1) and Hurst's K were calculated for every replication using Eq.2.2 and Eq.2.6, respectively.

3) Calculate the mean value, standard error, $S_{r(1)}$ and S_k ,and coefficient of skewness of lag-one autocorrelation coefficient r(1) and Hurst's K over the 25,000 realizations.

4) Change sample sizes and repeat Steps (1), (2) and (3).

The lengths of samples vary from 20 to 1,000, covering eight orders of sizes. Results shown in Figs.2.4 and 2.5 and Table 2.9 provide the information about the variability of the mean, standard error and skewness of r(1), and Hurst's K as they vary with sample size.

It can be seen from Figs.2.4 and 2.5 that the means of r(1) and Hurst's K converge to their theoretical values in a similar fashion but at different rates. That is, the rate of convergence to its theoretical value of r(1) is much faster than that of Hurst's K as n increases, even though the magnitudes of standard error of r(1) is much larger than that of Hurst's K. It can also to be seen that in Table 2.9, for simulated data, the coefficient of skewness, Cs, of Hurst's K tends to zero as the sample size increases. This implies that the distribution of Hurst's K like the distribution of r(1) and can be closely approximated by a symmetrical distribution as $n \rightarrow \infty$, even though its exact distribution is unknown.

Table 2.9 Variation of statistics of Hurst's K and r(1) based on 25,000 replications for a normal independent process

Size	r(1)			Hurst's K		
n	Mean	$S_{r(1)}$	Cs	Mean	S_k	Cs
20	-0.0495	0.206	0.050	0.6413	0.096	-0.147
30	-0.0317	0.173	0.010	0.6366	0.083	-0.099
50	-0.0191	0.137	0.018	0.6277	0.070	-0.073
80	-0.0119	0.110	0.035	0.6186	0.061	-0.019
100	-0.0095	0.099	0.042	0.6151	0.058	-0.017
150	-0.0060	0.081	0.028	0.6091	0.053	-0.007
500	-0.0025	0.045	0.025	0.5917	0.040	0.011
1000	-0.0001	0.032	-0.001	0.5835	0.035	0.004

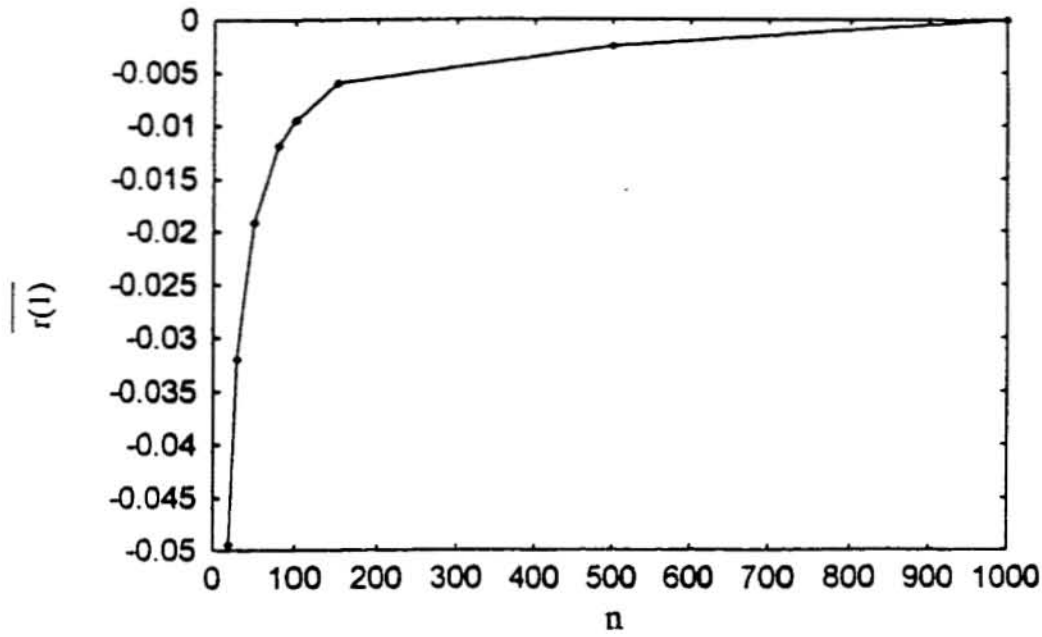

Figure 2.4a Plots of sample size, n, vs. the mean value of r(1) for the normal independent process.

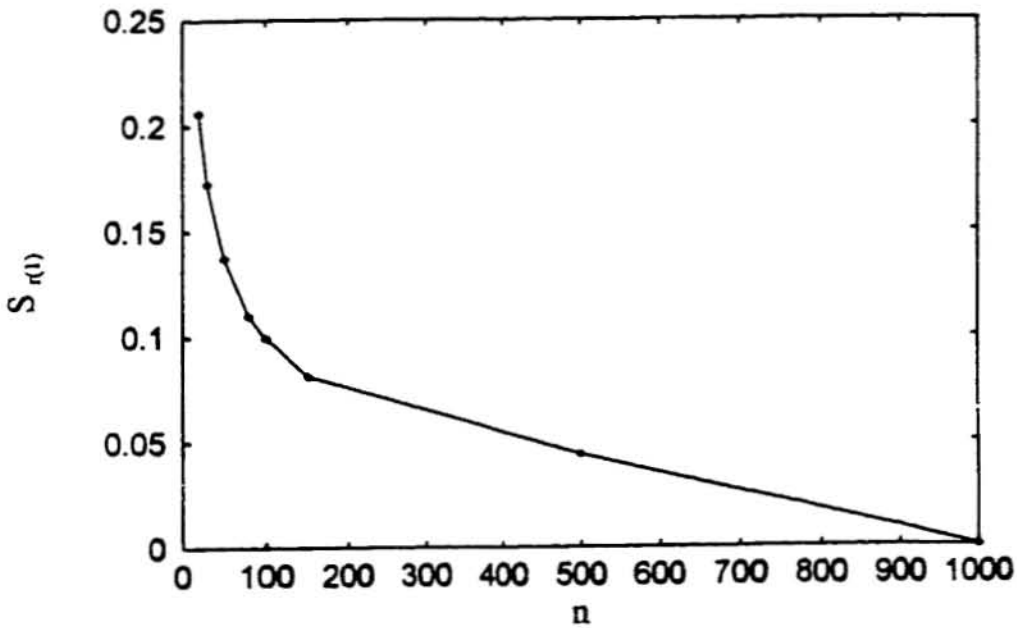

Figure 2.4b Plots of sample size, n, vs. the standard error of r(1), $S_{r(1)}$, for the normal independent process.

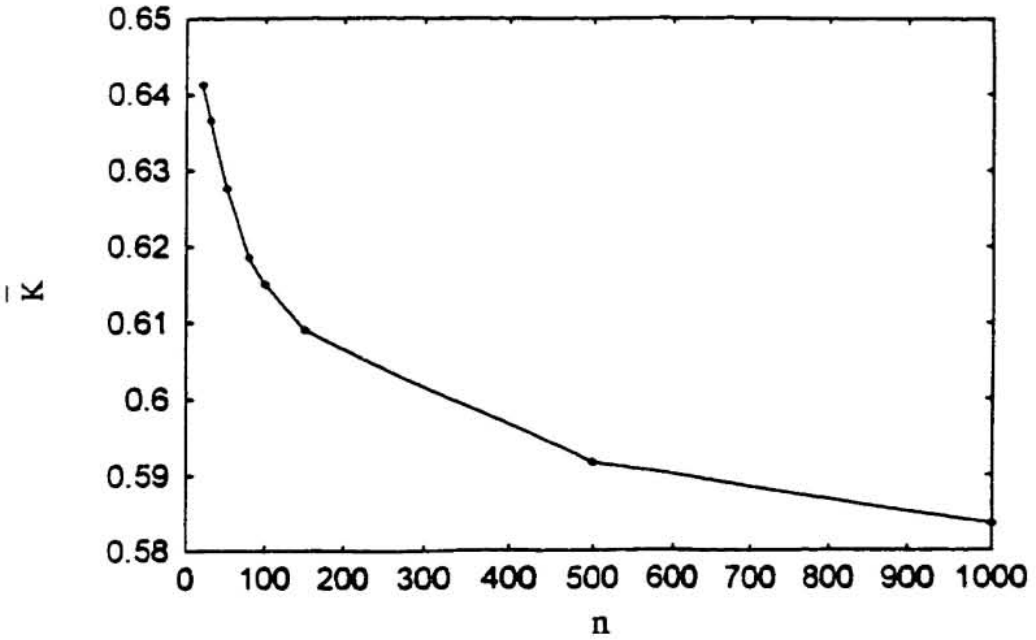

Figure 2.5a Plots of sample size, n, vs. the mean of Hurst's K for the normal independent process.

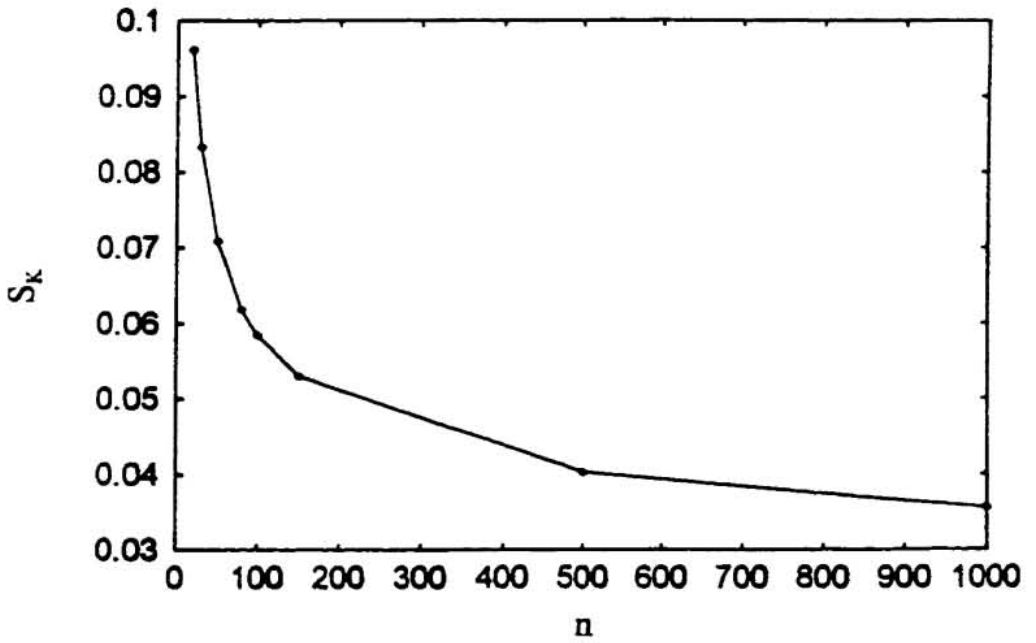

Figure 2.5b Plots of sample size, n, vs. the standard error of Hurst's K, S_k, for the normal independent process.

2.7 Summary

In this chapter, Hurst's K and the lag-one autocorrelation coefficient r(1), which measure the long- and short-term behaviours of annual peak flow series, respectively, have been analysed. Both parametric and nonparametric hypothesis tests show the correlation and dependence between Hurst's K and r(1) of the annual peak flows for Canadian and Chinese rivers.

The significant correlation between these two random variables and from the scatter plots indicated that an approximate linear relationship exists between them. In principle, as r(1) increases, Hurst's K increases, and r(1) decreases and Hurst's K decreases. Hence the long-term behaviour is related to the short-term behaviour of annual peak flows.

The dependence between Hurst's K and r(1) thus provides a strong basis to quantitatively describe the simultaneous occurrence of long- and short-term behaviours of annual peak flows in the next chapter.

Chapter 3

A Probabilistic Approach to Dealing With Long- and Short-term Behaviour of Annual Peak Flow Series

3.1 General

There are few methods of characterising long-term behaviour: an empirical technique based on Hurst's K developed by Lye and Lin (1994), the extremal index method of Leadbetter (1988), and the measure based on the fractional differencing parameter in an ARIMA model (Hosking, 1984).

For the modelling of long-term persistence, many models are available, but generally these models are not useful to use for a peak flow series because of the short sample sizes.

Most studies of the correlation structure of annual peak flow series still focus on discussing the "characterisation of" and "testing for" long-term persistence and short-term independence separately for a peak flow series. Only one study concerned the simultaneous occurrence of short-term independence and long-term dependence for peak flow series observed at Canadian rivers (Lye and Lin, 1994).

Based on the observed correlation and dependence between Hurst's K and lag-one

autocorrelation coefficient r(1) as discussed in the previous chapter, which measures the long- and short-term behaviour of series, respectively, this chapter focuses on quantitatively describing the simultaneous occurrence of the long-term persistence and short-term independence for annual peak flow series.

From a theoretical and practical point of view, it is suggested in this study that a sampling distribution of Hurst's K should be defined as a probability distribution for a given lag-one autocorrelation coefficient r(1). An approximation for the sampling distribution of Hurst's K will be developed using Monte Carlo simulation. Hence, an estimate of the probability for serially independent population, such as the annual peak flow series, to exhibit long-term persistence is also provided. The following sections discuss a probabilistic approach for dealing with long- and short-term behaviour of annual peak flow series that could be useful in flood risk analysis.

3.2 Sampling Distribution of Hurst's K

3.2.1 Standard Error of r(1)

It is clear that it would be rare for the sample autocorrelation coefficient r(1) to be exactly zero, even though the parent autocorrelation function r (1) is strictly zero for a normal independent process. It deviates from zero due to chance. In the analysis of annual peak flow series from Canadian and Chinese rivers, it was found that the standard error of the autocorrelation coefficient r(1) is much larger than that of Hurst's K. Fig.3.1 presents

frequency histograms for the Hurst's K and r(1) calculated from the peak flow series observed in Canadian rivers. The deviation of Hurst's K is much smaller than that of lag-one autocorrelation coefficient r(1) as shown in Fig.3.1. A similar result for generated independent data is shown in Fig.2.2 and Table 2.9.

The t-test for correlation does not reject a linear relationship between Hurst's K and r(1). However, the wide range of r(1) along the horizontal axis in Fig.2.2 is related to the distribution of Hurst's K along the vertical axis. For instance, in Fig.2.2a, for peak flows observed in Canadian rivers, the minimum values of Hurst's K at r(1)=-0.2 and r(1)=0.2, are 0.465 and 0.655, respectively. The difference of Hurst's K here is 0.19. For the maximum and mean values of Hurst's K at r(1)=-0.2 and 0.2 , there is also a wide difference of Hurst's K. It is obvious that the sampling distribution of r(1), which is related to the magnitudes of Hurst's K, should be taken into account. As such, the statistical hypothesis testing of long-term behaviour proposed by Lye and Lin (1994), which ignore those differences, is thus not strictly valid. The concept of r(1) is straightforward and its value is easy to be calculated. In order to emphasise hypothesis testing for long-term behaviour, the information about the distribution of r(1) should be taken into account when the sampling distribution of Hurst's K is considered.

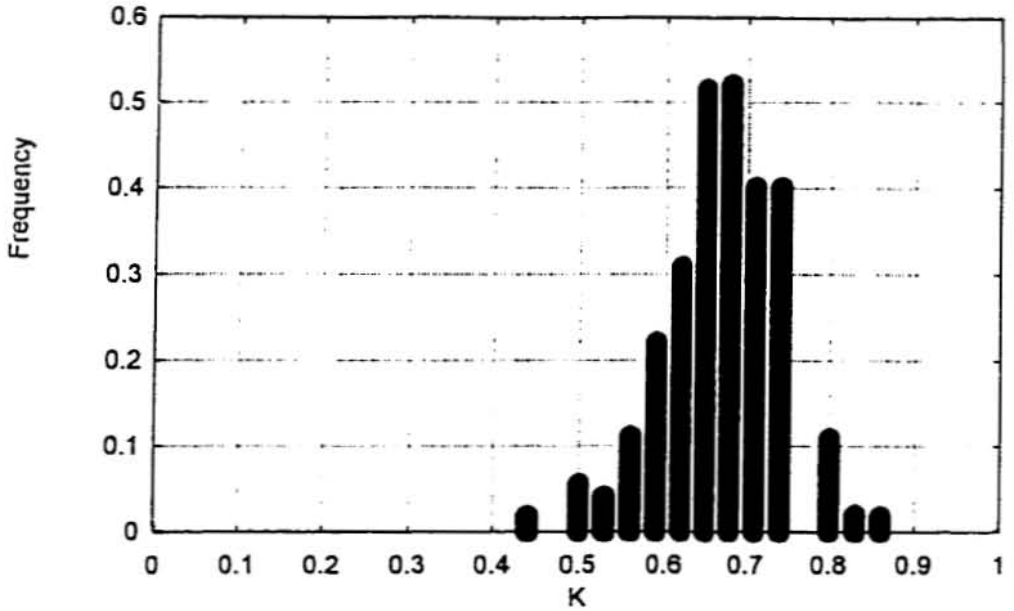

Figure 3.1a Frequency histogram for the Hurst's K calculated from the peak flow series observed in Canadian rivers.

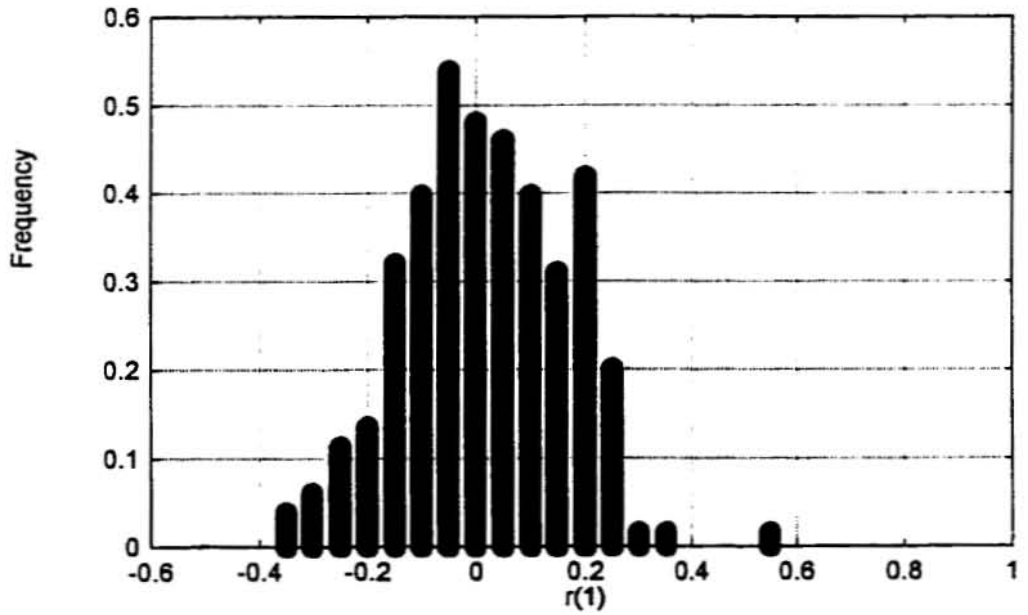

Figure 3.1b Frequency histogram for the r(1) calculated from the peak flow series observed in Canadian rivers.

3.2.2 Sampling Distribution of Hurst's K

It is clear from the previous chapter that Hurst's K and the lag-one autocorrelation coefficient r(1) are correlated and dependent in a probabilistic manner. This provides a theoretical basis to deal with the sampling distribution of Hurst's K which is expressed as the sampling distribution for a given lag-one autocorrelation coefficient r(1). It has also been observed that the standard error of the lag-one autocorrelation coefficient r(1) is much larger than that of Hurst's K. In statistical hypothesis testing, such a wide range of r(1) could increase the test error. Based on these considerations, the sampling distribution of Hurst's K for a given r(1) can be developed by probability theory and an extensive Monte Carlo simulation experiment.

3.2.3 Sampling Distribution of Hurst's K for a Given r(1)

3.2.3.1 Definition of Events of Interest

The probability of an event depends on the occurrence or non-occurrence of one or more related events. If the occurrence of one event, A, for example, is affected by the occurrence of another event, say B, then its probability is conditional. Otherwise, the probability is unconditional. For instance, the probability of the event "flood at a given time" is an unconditional probability. In contrast with this, "a flood occurred yesterday," what is the probability that a flood will occur today? is a conditional probability that projects more accurately than the unconditional one in this case.

The following events of interest dealing with the correlation structure of an annual peak flow series are defined as follows:

Event A: Peak flows exhibiting long-term behaviour measured by Hurst's K

Event B: Peak flows exhibiting short-term behaviour measured by r(1)

In Chapter 2, parametric and nonparametric hypothesis tests show that the two statistics, Hurst's K and lag-one autocorrelation coefficient r(1), are correlated and dependent. The variation of Hurst's K is linearly related to that of the r(1). Due to dependence, the intersection of two events, $A \cap B$, exists and occurs simultaneously in a probabilistic manner. In other words, the long-term (event A) and short-term (event B) behaviours occur simultaneously in a physical and theoretical sense. And, the probability of event A given B has occurred can be quantified using basic probability theory:

$$Prob(A/B) = \frac{Prob\,(A \cap B)}{Prob\,(B)} \tag{3.1}$$

where Prob (B) is the unconditional probability of B and Prob($A \cap B$) is the probability of the intersection of $A \cap B$.

Now, consider events A and B to be the numerical events ($K \leq k$) and ($R1 \leq r_t$). These events A and B can be expressed or measured by the magnitudes of Hurst's K and r(1), respectively. Assume that K and R1 are random variables. The probability of K for the given R1 in the region R_g is

$$Prob(K \le k/R1 \in R_g) = \int_{-\infty}^{k} f_{K/R1}(t/R1 \in R_g)\, dt \qquad (3.2)$$

which is a conditional probability.

In other words, Eq.3.2 describes the conditional distribution of K given R1 in the region R_g, where K and R1 indicate the random variables Hurst's K and lag-one autocorrelation coefficient r(1), respectively. Thus, the sampling distribution of Hurst's K is defined by a conditional distribution, which is the sampling distribution of Hurst's K for a given r(1), or equivalently saying the distribution of Hurst's K given R1 that is in R_g.

When the distribution of K given R1 that is in R_g is considered, the form of the density function in Eq. 3.2 is unknown. The distribution of K given R1 in R_g cannot be found by an analytical approach, thus an extensive Monte Carlo simulation experiment was carried out to obtain the probability distribution of K given R1 that is in R_g.

3.2.3.2 Monte Carlo Simulation Producing the Sampling Distribution of Hurst's K for a Given r(1)

According to the customary assumption in flood risk analysis, a normal distribution n (0,1) is chosen as a population process in the Monte Carlo simulation, and R1 is assumed to take a few regions, R_{gi}, i=1,2,...,m. The Monte Carlo experiment is designed as follows:

1) For a given sample size n, 25,000 samples are generated from a normal independent process with mean zero, standard deviation one, and lag-one autocorrelation zero. The sample sizes vary from 20 to 10,000, eight orders of

magnitude.

2) Calculate r(1) and Hurst's K for every sample using Eqs.2 .2 and 2.6.

3) Divide the total range of calculated r(1) into a several intervals R_{gi}, where i=1,2,..., m. In the present study, seven intervals are considered. There are: r(1)<-0.25, r(1) at [-0.25,-0.15),[-0.15,-0.05), [-0.05, 0.05), [0.05, 0.15), [0.15, 0.25) and r(1) ≥ 0.25. Calculate statistics such as mean value, standard error of Hurst's K, s_K and the coefficient of variation, Cv, and coefficient of skewness, Cs for each region.

4) Rank the data of Hurst's K for the corresponding interval and calculate the frequency, then a simulated probability distribution of K given R1 in R_{gi} is obtained. Clearly, the more data generated the more accurate the cumulative distribution.

The calculated r(1) and Hurst's K from the generated data are illustrated in Fig.3.2 and the cumulative probability distributions of Hurst's K given r(1) in R_{gi} is illustrated in Fig.3.3. The statistics of the cumulative probability distribution of Hurst's K are summarised in Table 3.1.

From Fig.3.3 and Table 3.1 it was found that

a) The mean of Hurst's K increases with an increase of r(1).

b) As sample size increases, as expected, the mean of Hurst's K, as well as its standard error, s_K ,and coefficient of variation, Cv, decreases.

c) The region of the distribution of Hurst's K becomes narrower as the sample size increases.

d) The skewness in the distribution of Hurst's K decreases as the r(1) reaches zero in each sample size, and the distribution of Hurst's K tends to be approximately normal especially for larger sample sizes.

Overall, the sampling distribution of Hurst's K for a given r(1) worked out by Monte Carlo simulation provides a representation of the long-term behaviour based on the short-term properties of the time series; the long- and short-term behaviour being linked by the conditional distribution of K given that R1 is in R_g.

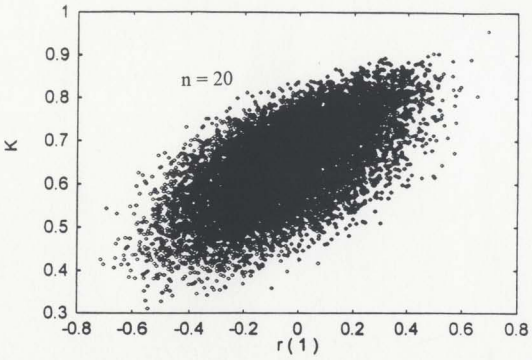

Figure 3.2a Plots of r(1) vs. Hurst's K for generated independent data with n=20.

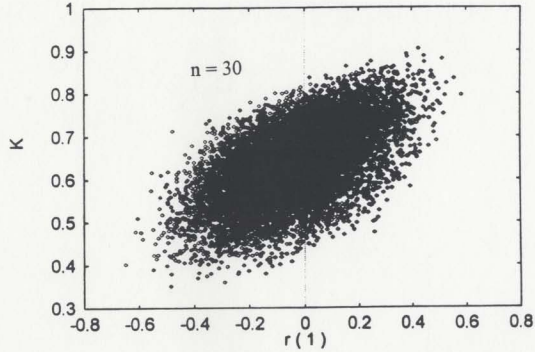

Figure 3.2b Plots of r(1) vs.Hurst's K for generated independent data with n=30.

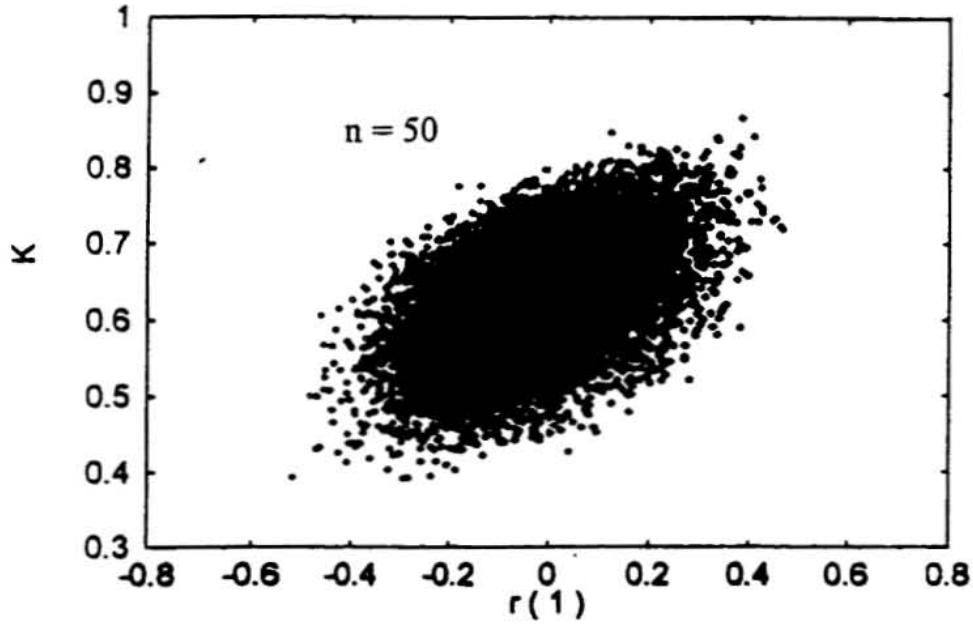

Figure 3.2c Plots of r(1) vs.Hurst's K for generated independent data with n=50.

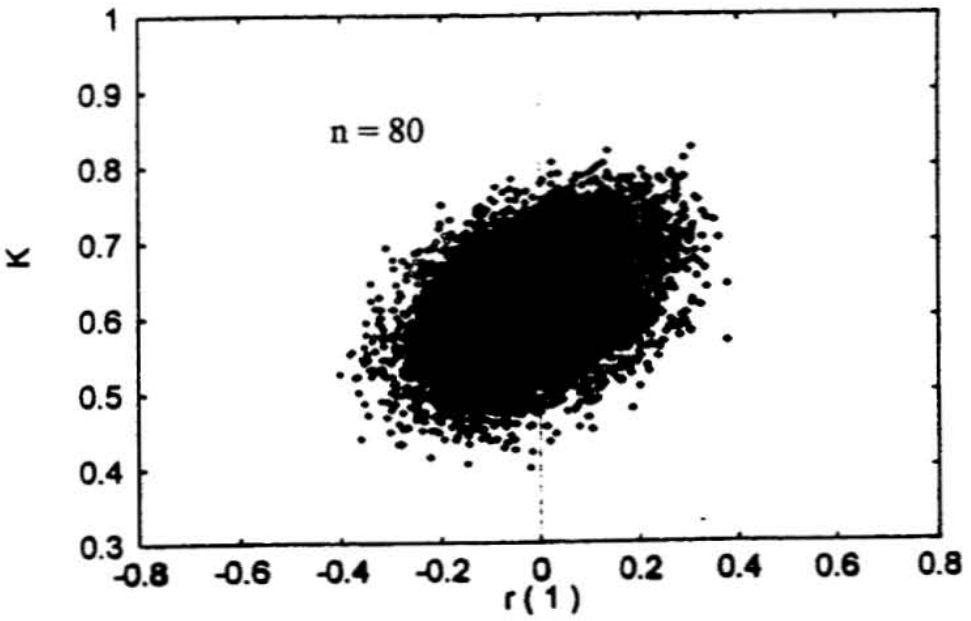

Figure 3.2d Plots of r(1) vs.Hurst's K for generated independent data with n=80.

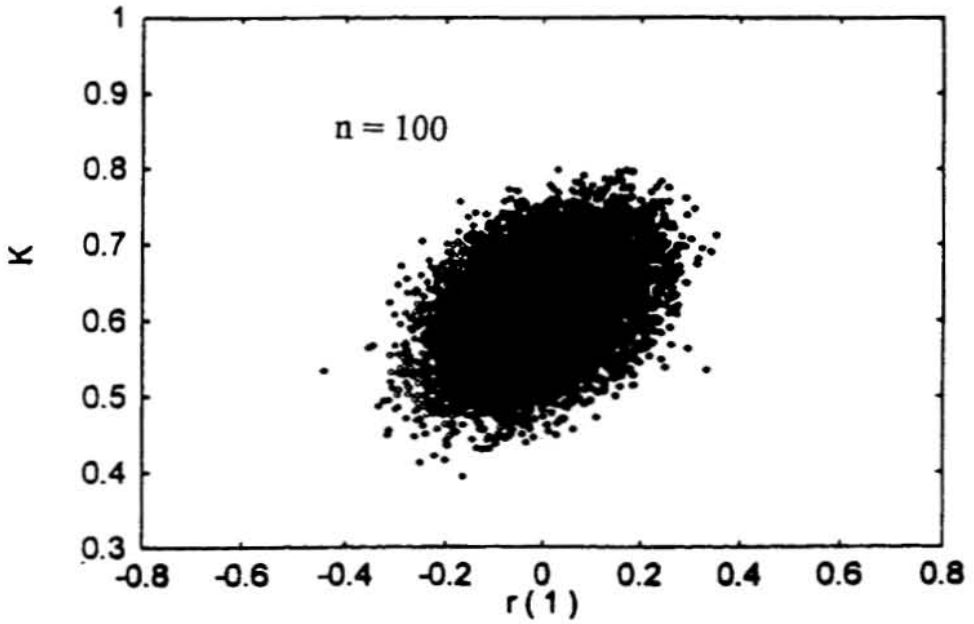

Figure 3.2e Plots of r(1) vs.Hurst's K for generated independent data with n=100.

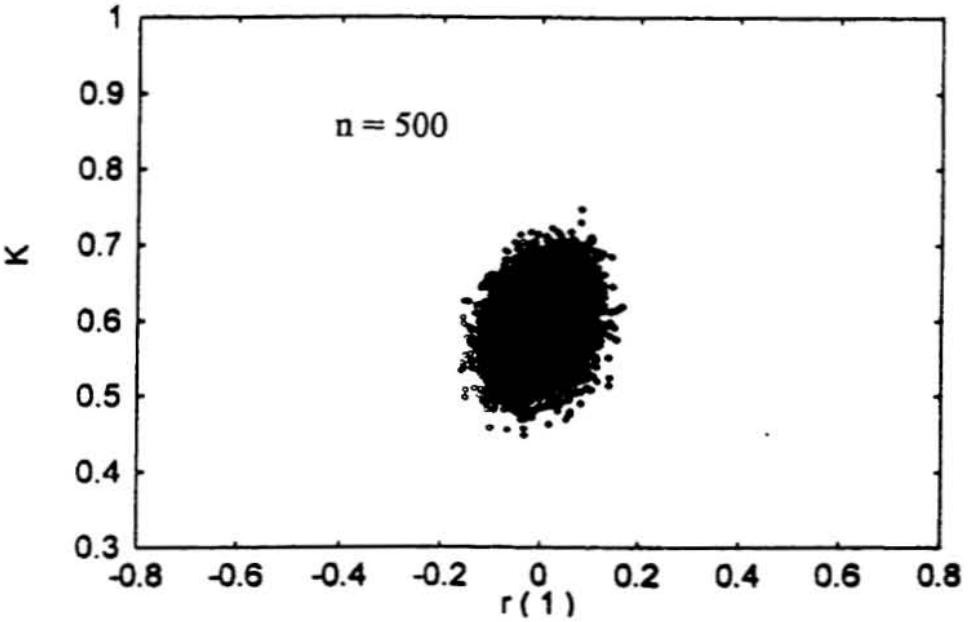

Figure 3.2f Plots of r(1) vs.Hurst's K for generated independent data with n=500.

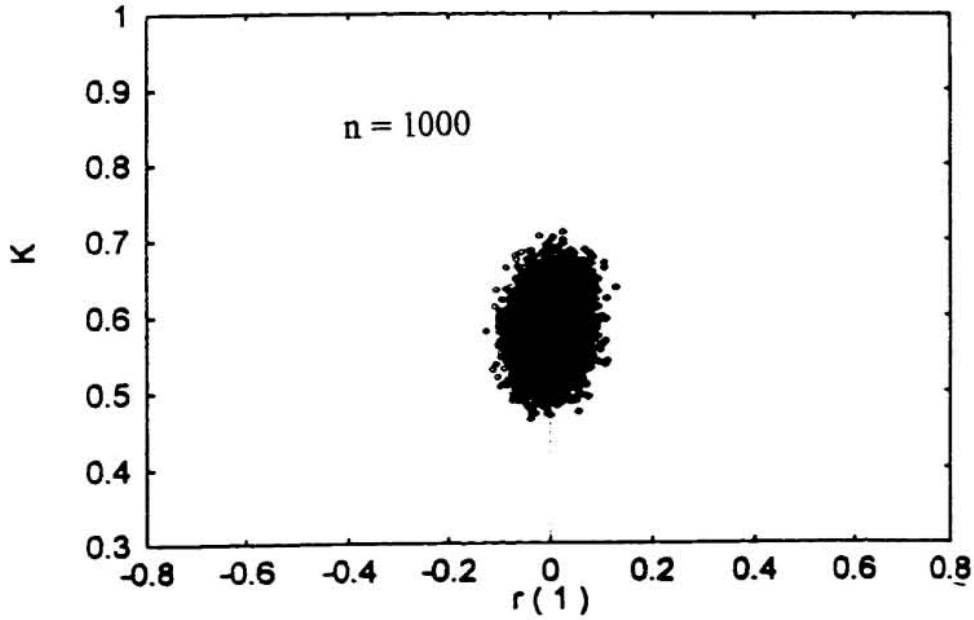

Figure 3.2g Plots of r(1) vs.Hurst's K for generated independent data with n=1000.

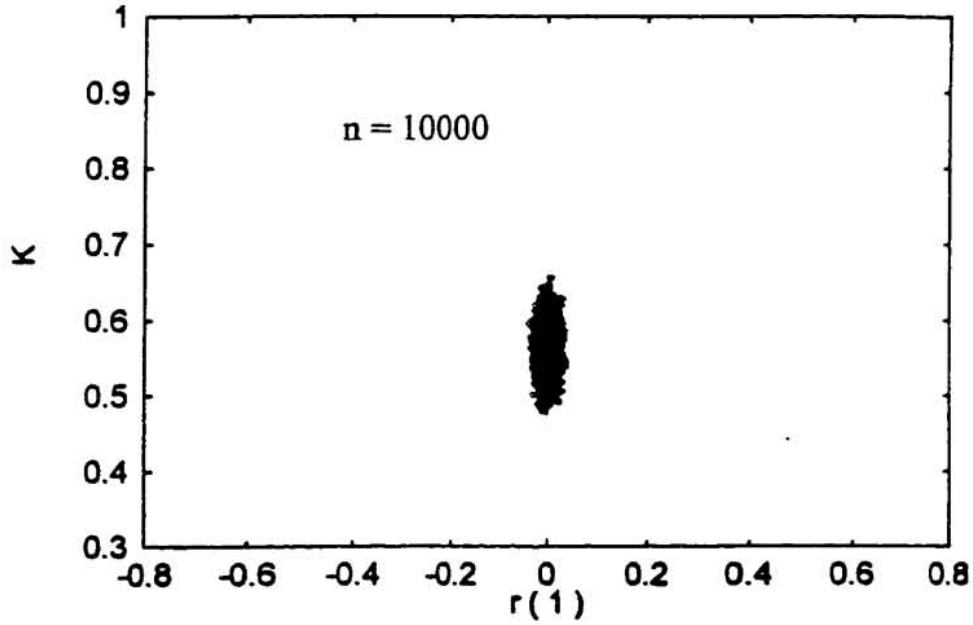

Figure 3.2h Plots of r(1) vs.Hurst's K for generated independent data with n=10000.

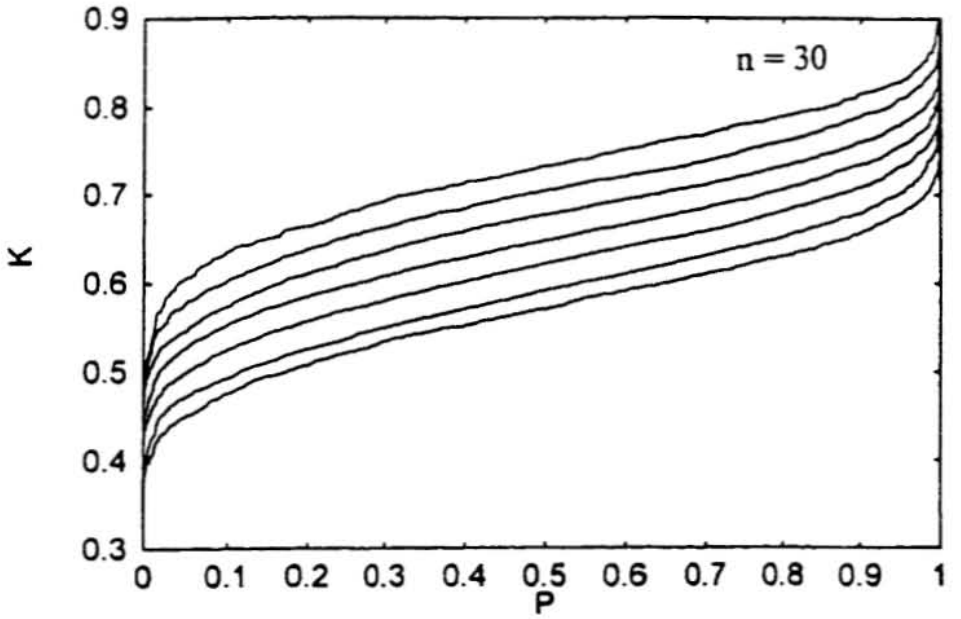

Figure 3.3a The cumulative probability distribution of K given that R1 is in R_z with n=30 where R_z: R1<-.25,R1 at [-.25,-.15), [-.15, -.05), [-.05,.05), [.05, .15), [.15, .25) and r(1)≥.25 from the bottom to the top.

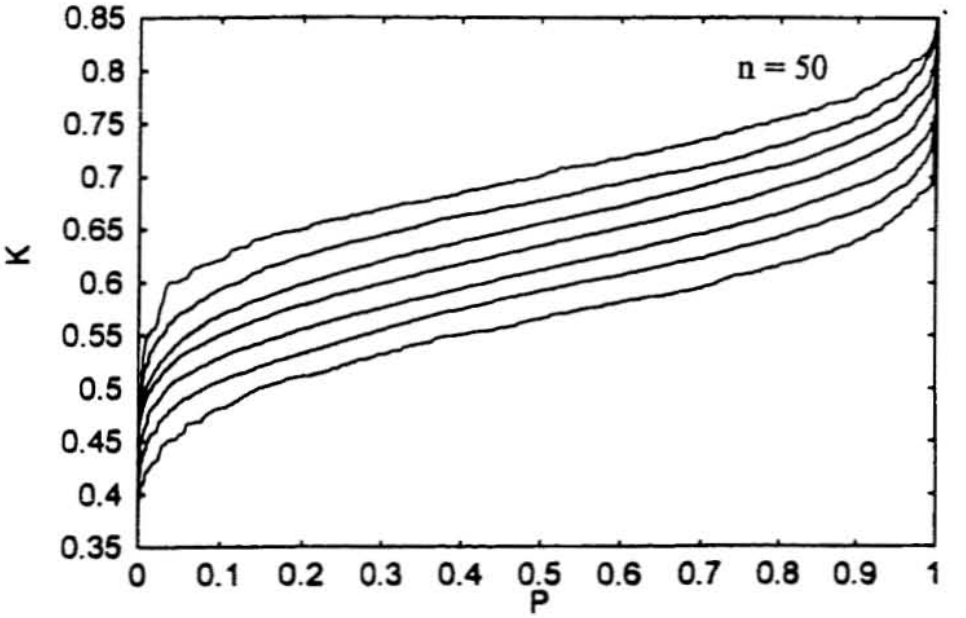

Figure 3.3b The cumulative probability distribution of K given that R1 is in R_z with n=50 where R_z: R1<-.25,R1 at [-.25,-.15), [-.15, -.05), [-.05,.05), [.05, .15), [.15, .25) and r(1)≥.25 from the bottom to the top.

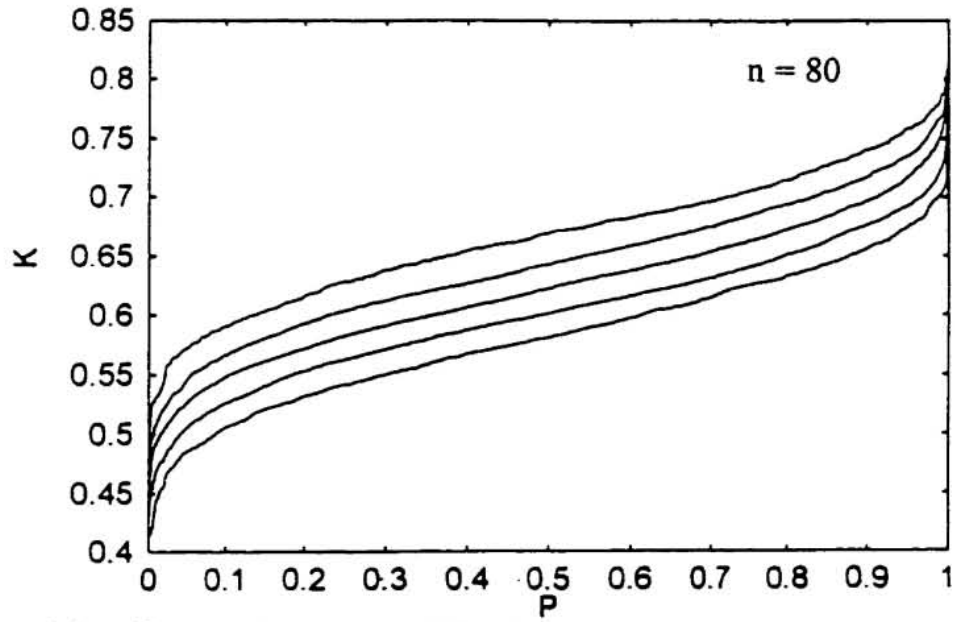

Figure 3.3c The cumulative probability distribution of K given that R1 is in R_g with n=80 where R_g: [-.25,-.15), [-.15, -.05), [-.05,.05), [.05, .15), [.15, .25) from the bottom to the top.

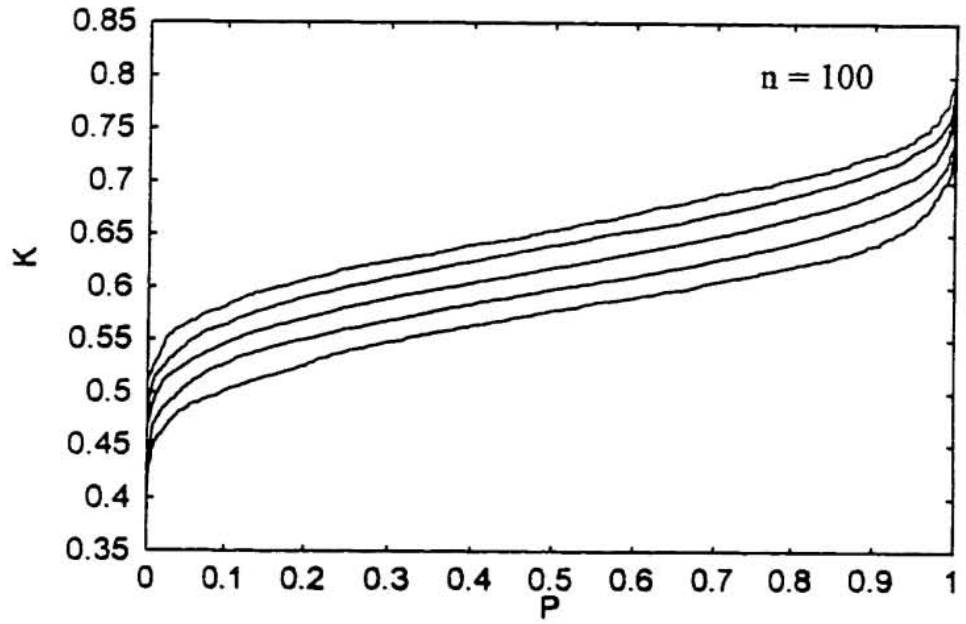

Figure 3.3d The cumulative probability distribution of K given that R1 is in R_g with n=100 where R_g: [-.25,-.15), [-.15, -.05), [-.05,.05), [.05, .15), [.15, .25) from the bottom to the top.

Table 3.1 The Statistics of the sampling distribution of K given that R1 is in R_g

n	R_g	mean	standard	Cv	Cs
30	<-0.25	0.5693	0.0689	0.121	-0.165
	[-0.25, -0.15)	0.5890	0.0714	0.121	-0.093
	[-0.15, -0.05)	0.6187	0.0702	0.113	-0.132
	[-0.05, 0.05)	0.6450	0.0693	0.108	-0.182
	[0.05, 0.15)	0.6715	0.0696	0.104	-0.262
	[0.15, 0.25)	0.6988	0.0702	0.100	-0.275
	≥0.25	0.7265	0.0700	0.096	-0.390
50	<-0.25	0.5629	0.0606	0.108	-0.145
	[-0.25, -0.15)	0.5885	0.0619	0.105	-0.110
	[-0.15, -0.05)	0.6101	0.0621	0.102	-0.103
	[-0.05, 0.05)	0.6331	0.0629	0.099	-0.093
	[0.05, 0.15)	0.6535	0.0645	0.099	-0.147
	[0.15, 0.25)	0.6759	0.0627	0.093	-0.180
	≥0.25	0.7004	0.0599	0.086	-0.305
80	<-0.25	0.5597	0.0538	0.096	-0.130
	[-0.25, -0.15)	0.5812	0.0578	0.099	-0.133
	[-0.15, -0.05)	0.6008	0.0569	0.095	-0.092
	[-0.05, 0.05)	0.6220	0.0577	0.093	-0.013
	[0.05, 0.15)	0.6424	0.0578	0.090	-0.086
	[0.15, 0.25)	0.6660	0.0564	0.085	-0.202
	≥0.25	0.6763	0.0571	0.084	-0.049
100	<-0.25	0.5597	0.0486	0.087	0.215
	[-0.25, -0.15)	0.5748	0.0539	0.094	-0.069
	[-0.15, -0.05)	0.5964	0.0537	0.090	-0.100
	[-0.05, 0.05)	0.6179	0.0545	0.088	-0.039
	[0.05, 0.15)	0.6381	0.0553	0.087	-0.112
	[0.15, 0.25)	0.6545	0.0542	0.083	-0.061
	≥0.25	0.6746	0.0585	0.081	-0.291

3.3 New Empirical Percentage Points for Hurst's K

The direct use of the concept of sampling distribution of Hurst's K for a given r(1) is a new table of empirical percentage points to test for long-term persistence.

The t-tests shown in Tables 2.5 and 2.6 indicate a correlation between Hurst's K and lag-one autocorrelation coefficient r(1) at the 5% and 10% levels of significance. And also, the observations and Monte Carlo simulation results show that the deviation of r(1) is much greater than that of Hurst's K. It can be seen from Figs. 2.2 and 3.2 that the lag-one autocorrelation coefficient r(1) occupies a much wider range along the horizontal axis than that of Hurst's K along the vertical axis. The information about the correlation between Hurst's K and lag-one autocorrelation coefficient r(1) should be taken into account. Lye and Lin (1994) proposed an empirical percentage points for testing long-term persistence which ignore the information about the distribution of r(1) which is related to the distribution of Hurst's K. It is necessary to expand the basic concept in their proposal and take the information about r(1) into account using the concept of sampling distribution for a given r(1).

The same Monte Carlo procedure shown in the above section to produce an approximation of the sampling distribution of Hurst's K for a given r(1) can be deduced. For a given significant level α, the Hurst's K could be selected from Figs.3.3 according to the different sample sizes and the region of lag-one autocorrelation coefficient r(1). Table 3.2 shows the new table of empirical percentage points at a=5% and 10% for

sample sizes of n=30,50, 80, and 100.

The tests for long-term persistence could be carried out based on the new empirical percentage points for Hurst's K. The observed data can be transformed into normal variables by the Box-Cox transformation (Box and Cox, 1964) if the data are not normal. For the transformed data which are approximately normally distributed, calculate Hurst's K and r(1), then find the critical Hurst's K at the given level in the Table 3.2 using the calculated r(1) and sample size, comparing the observed Hurst's K with the critical Hurst's K. If the K value for the observations is greater than the K value given in the table at the given significance level for the given size, it can be concluded that the observed series is long-term dependent at the given significance level.

Table 3.2 Empirical percentage points for Hurst's K for given r(1) for the normal independent data, where n - the sample size, α - the significant level

		$b_i \le R1 < a_i$						
n	α	<-.25	-.2±.05	-.1±.05	.0±.05	.1±.05	.2±.05	≥0.25
30	5%	0.6781	0.7016	0.7328	0.7554	0.7794	0.8088	0.8277
50		0.6626	0.6873	0.7084	0.7340	0.7550	0.7740	0.7980
80		0.6446	0.6754	0.6930	0.7173	0.7352	0.7583	0.7937
100		0.6431	0.6606	0.6837	0.7054	0.7290	0.7398	0.7831
30	10%	0.6571	0.6794	0.7091	0.7345	0.7599	0.7890	0.8149
50		0.6393	0.6668	0.6911	0.7153	0.7368	0.7554	0.7789
80		0.6186	0.6560	0.6761	0.6964	0.7169	0.7396	0.7683
100		0.6166	0.6400	0.6652	0.6887	0.7099	0.7250	0.7464

3.4 A Useful Index, $P(K \geq k_0)$

The concept of the sampling distribution of Hurst's K for a given r(1) also provides useful information in dealing with the serial correlation of annual peak flows, and the results from the Monte Carlo simulations can shed some light on long- and short-term behaviours of peak flows. This will be described in the next section.

3.4.1 Definition of Events of Interest

The probability for an independent series such as the peak flow series to exhibit long-term persistence can be defined using the following arguments.

Hurst's K and r(1) can serve as numerical outcomes of an experiment, let us define the events of interest once more:

Event A_1: Peak flows exhibiting the long-term persistence identified by $K > k_o$

Event B_i: Peak flows exhibiting the short-term independence identified by r(1),

$$i = 1, 2, \ldots, m$$

where k_0 is a special value of Hurst's K which implies that when the observed Hurst's K is greater than this value, the series exhibit long-term persistence, and B_1, B_2, ... B_m are mutually exclusive and exhaustive.

The relationship between Hurst's K and r(1) for the observations from Canadian and Chinese rivers shown in Figure 2.2a-b, and for the synthetic sequences of n shown in

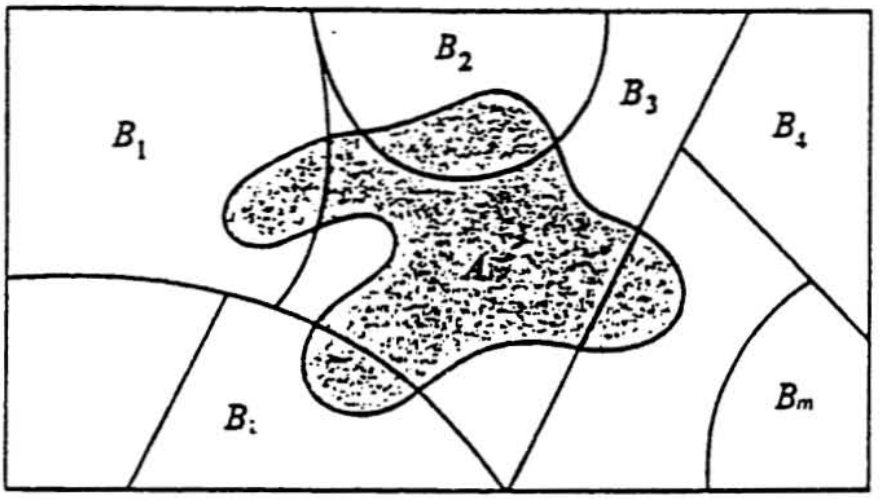

Figure 3.4 Venn diagram for events A_1, B_i.

Figure 3.2a-h can be expressed as 'intersection' and 'union' of the events in the Venn diagram. So the relationship between defined A_1 and B_i, i=1,2,..,m can be expressed on the Venn diagram as shown in Figure 3.4.

Let A_1 be an event of sample space S such that $P(A_1) \neq 0$, and the events $\{B_1, B_2, \ldots, B_m\}$ form a partition of the sample space S, where $P(B_i) \neq 0$, for i=1, 2, ..., m. Events A_1 and B_i are dependent. The event A_1 is seen to be the union of the mutually

exclusive events $B_1 \cap A_1$, $B_2 \cap A_1$, ..., $B_m \cap A_1$, that is,

$$A_1 = (B_1 \cap A_1) \cup (B_2 \cap A_1) \ldots \cup (B_m \cap A_1) \tag{3.3}$$

So, the probability of the event A_1 is defined by the total probability theorem

$$P(A_1) = \sum_{i=1}^{m} P(B_i \cap A_1) \tag{3.4}$$

By the definition of the multiplicative law gives the probability of the intersection $B_i \cap A_1$ as

$$P(B_i \cap A_1) = P(A_1 / B_i) P(B_i) \tag{3.5}$$

where $P(B_i)$ is the unconditional probability and $P(A_1/B_i)$ is the conditional probability, probability of A_1 for a given B_j. So, the probability of the event A_1 is obtained by

$$P(A_1) = \sum_{i=1}^{m} P(A_1 / B_i) \; P(B_i) \tag{3.6}$$

Furthermore, the events A_1 and B_i are considered as numerical events ($K \geq k_o$) and ($R1 \geq r_1$), as mentioned earlier, where K and R1 are two random variables representing the Hurst's K and r(1), while k and r_1 are observations of random variables K and R1, respectively. Hence, a set of definitions can be made according to the above events and probabilities:

$P(K \geq k_0)$: be equivalent to $P(A_1)$, the probability of a peak flow series exhibiting long-term persistence, where k_0 is a 'critical value' and can be chosen arbitrarily at the present time;

$P(b_i \leq R1 < a_i)$: be equivalent to $P(B_i)$, the probability of peak flow series exhibiting short-term independence, where $i = 1, 2, \ldots, n$;

$P(K \geq k_0 / b_i \leq R1 < a_i)$: be equivalent to $P(A_1/B_i)$, the probability of a peak flow series exhibiting long-term persistence for given R1 that is in R_{gi}, $(b_i \leq R1 < a_i)$.

Hence, Eq. 3.6 can be expressed as

$$P(K \geq k_0) = \sum_{i=1}^{m} P(K \geq k_0 / b_i \leq R1 < a_i) P(b_i \leq R1 < a_i) \qquad (3.7)$$

If the critical value of k_0 has been determined in Eq.3.7, the probability $P(b_i \leq R1 < a_i)$ can be obtained directly from Eq. 2.7 based on the distribution of R1 assuming normality or from Monte Carlo simulation if the parent distribution is not normally distributed. The conditional probability $P(K \geq k_0 / b_i \leq R1 < a_i)$ can be obtained from the designed Monte Carlo experiments, so the probability of peak flow series exhibiting long-term persistence, $P(K \geq k_0)$, can be determined.

3.4.2 Estimating $P(K \geq k_0)$

Based on the sampling distribution of Hurst's K for a given r(1), the suggested procedure to estimate the probability of a peak flow series exhibiting long-term persistence, $P(K \geq k_0)$, is as follows:

(1) Determine the lower bound value k_0.

How do we designate a lower bound value, k_0, to represent the behaviour of long-term persistence? This is one of the most important steps in dealing with long-term persistence. This study suspends discussion of this, and takes k_0 arbitrarily. Empirically, Hurst found that Hurst coefficient has a mean of 0.73 and a standard deviation of 0.072 (Hurst, 1951, 1956), hence it is assumed that the magnitude of the lower boundary, k_0, is (0.73-0.072) =0.658, one standard deviation below the mean from Hurst's study.

(2) Determining the probability $P(K \geq k_0 / b_i \leq R1 < a_i)$.

For the given R_{gi} and sample size n, the corresponding cumulative probability distribution curve of Hurst's K shown in Figure 3.3 is selected. The horizontal axis, P, of the selected cumulative probability distribution curve gives the sampling distribution of Hurst's K given r(1) that is in R_{gi}. Once the value of the lower boundary k_0 is designated, the probability $P(K \geq k_0 / b_i \leq R1 < a_i)$, can be obtained from the cumulative probability distribution curves.

Table 3.3 summarises the probability $P(K \geq k_0 / b_i \leq R1 < a_i)$ for the sample size of n=30, 50, 80, 100 based on the Monte Carlo experiments.

(3) Determine the probability $P(b_i \leq R1 < a_i)$.

How do we determine the magnitude of $P(b_i \leq R1 < a_i)$? It can be directly calculated using Eq.2.7 because the parent population is normally distributed, therefore R1 is distributed normally, and otherwise it can be simply computed from the results of the Monte Carlo simulation. Table 3.4 gives the probability, $P(b_i \leq R1 < a_i)$, calculated from Monte Carlo simulation.

(4) Obtain the probability $P(K \geq k_0)$.

Once the conditional probability $P(K \geq k_0 / b_i \leq R1 < a_i)$ and the probability $P(b_i \leq R1 < a_i)$ are obtained from the Monte Carlo simulations, the probability, $P(K \geq k_0)$, that a peak flow exhibiting long-term persistence $P(K \geq k_0)$, can be obtained from Eq.3.7. In other words, the elements in Table 3.3 when multiplied by the corresponding elements in Table 3.4, gives the probability $P(K \geq k_0)$, as required. The results from calculations of the probability, $P(K \geq k_0)$ are given in Table 3.5.

Table 3.3 Probability $P(K \geq k_0 / b_i \leq R1 < a_i)$ calculated from the cumulative probability distribution shown in Figure 3.3, where $k_0 = 0.658$

$P(K \geq k_0 / b_i \leq R1 < a_i)$				
$b_i \leq R1 < a_i$	n = 30	n= 50	n = 80	n = 100
< -0.25	0.0973	0.0559	0.0260	0.0400
-0.20±.05	0.1819	0.1373	0.0955	0.0559
-0.10±.05	0.3054	0.2326	0.1639	0.1280
0.00±.05	0.4462	0.3587	0.2745	0.2426
0.10±.05	0.6021	0.4829	0.4038	0.3694
0.20±.05	0.7270	0.6306	0.5734	0.4768
≥ 0.25	0.8312	0.7721	0.6724	0.6100

Table 3.4 Probability of $P(b_i \leq R1 < a_i)$ calculated from the Monte Carlo simulation

$P(b_i \leq R1 < a_i)$				
$b_i \leq R1 < a_i$	n = 30	n= 50	n = 80	n = 100
< -0.25	0.1050	0.0471	0.0122	0.0069
-0.20±.05	0.1437	0.1269	0.0897	0.0700
-0.10±.05	0.2131	0.2397	0.2661	0.2648
0.00±.05	0.2221	0.2769	0.3446	0.3801
0.10±.05	0.1656	0.1987	0.2177	0.2247
0.20±.05	0.0980	0.0835	0.0828	0.0482
≥ 0.25	0.0524	0.0291	0.0069	0.0054

Table 3.5 Probability of $P(K \geq k_0)$ calculated from Tables 3.3 and 3.4

$P((K \geq k_0) \cap (b_i \leq R1 < a_i)) = P(b_i \leq R1 < a_i)\, P(K \geq k_0 / b_i \leq R1 < a_i)$				
$b_i \leq R1 < a_i$	n =30	n = 50	n = 80	n = 100
< -0.25	0.0102	0.0026	0.0004	0.0003
-0.20±.05	0.0261	0.0174	0.0086	0.0004
-0.10±.05	0.0651	0.0557	0.0436	0.0339
0.00±.05	0.0991	0.0994	0.0946	0.0922
0.10±.05	0.0997	0.0959	0.0879	0.0830
0.20±.05	0.0712	0.0527	0.0475	0.0229
≥ 0.25	0.0436	0.0225	0.0046	0.0033
$P(K \geq k_0)$	**0.3877**	**0.3462**	**0.2872**	**0.2360**

3.4.3 An Estimator for the Population Value of $P(K \geq k_0)$

From Tables 3.3, 3.4 and 3.5 it can be seen that

a) The magnitude of the probability $P(K \geq k_0 / b_i \leq R1 < a_i)$ in Table 3.3 decreases with an increase in sample size, but increases with increasing values of R1.

b) The distribution of $P(b_i \leq R1 < a_i)$ in Table 3.4 shows that random variable R1 appears to be normally distributed as in the theoretical equation of Eq. 2.7. Most observations are located between the interval [-0.1,0.1].

c) In Table 3.5, the probability $P(b_i \leq R1 < a_i)P(K \geq k_0 / b_i \leq R1 < a_i)$ takes the maximum values between interval [-0.05, 0.15]. The magnitudes of probability

$P(K \geq k_0)$, the probability of the peak flow series exhibiting long-term persistence, decreases with increasing sample size.

In view of the above analyses, the probabilities $P(K \geq k_0)$ depends on the variation of both the distributions of r(1) and the conditional probability of Hurst's K given r(1) in R_g. Figure 3.5 shows that the distribution of conditional probability, $P(K \geq k_0 / b_i \leq R1 < a_i)$, varies with r(1), but probability $P(b_i \leq R1 < a_i)$ distributed on the R1-axis indicates R1 is normally distributed. It seems that the probability $P(b_i \leq R1 < a_i)P(K \geq k_0 / b_i \leq R1 < a_i)$ is the conditional probability, $P(K \geq k_0 / b_i \leq R1 < a_i)$, weighted by the probability, $P(b_i \leq R1 < a_i)$. However, by an examination of the estimation procedure, the calculated $P(K \geq k_0)$ depends on the sample size as well as the form of the sampling distribution of Hurst's K for a given r(1) and sampling distribution of r(1), and it can serve as an estimator for the population $P(K \geq k_0)$ which should exist and represent the proportion of the series exhibiting long-term persistence. The expected value of $P(K \geq k_0)$, moreover, can be reached if the sampled realisations for a given sample size n are large enough in the Monte Carlo simulation.

The concept proposed here, an estimate of $P(K \geq k_0)$ representing the probability of an independent series exhibiting the long-term persistence, is meaningful in dealing with serial correlation in peak flows discussed in the next section.

3.5 Practical implications

3.5.1 A Proposed Quantitative Descriptor for Long-term Persistence

Studies of serial correlation of annual peak flows have been ongoing for many years. Until the recently proposed method for testing long-term dependence by Lye and Lin (1994), investigators had discussed the existence of long-term persistence. Their study indicated that although short-term dependence is virtually absent for most of the peak flow series, significant long-term dependence exists for a large number of peak flow series tested.

Based on the correlation and dependence between Hurst's K and r(1) investigated in the previous chapter, a series of definition of events of interest based on probability theory allows us to develop a new method of quantitatively describing long-term persistence rooted in an independent series.

Based on the concept of the sampling distribution of Hurst's K for a given r(1), the estimator for population $P(K \geq k_0)$ and its distribution on the R1 axis, $P(b_i \leq R1 < a_i)P(K \geq k_0 / b_i \leq R1 < a_i)$, shown in Table 3.5 and Figure 3.5 assure that long-term persistence and short-term independence can be quantitatively estimated. In flood risk analysis, we have to assume that the flood record to be analysed is a reliable set of measurements of independent random events from a population. Because of this, the calculation results from the data generated from an normal independent process or transformed normal should be acceptable for the assumed peak flow population.

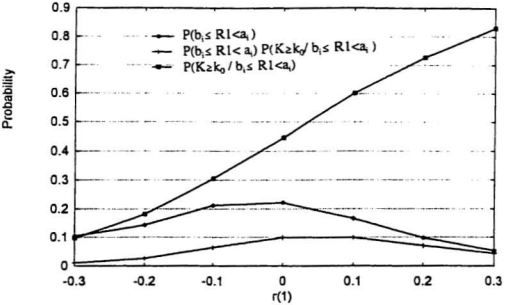

Figure 3.5a $P(b_i \leq R1 < a_i)\ P(K \geq k_o / b_i \leq R1 < a_i)$ varying with r(1) for n=30

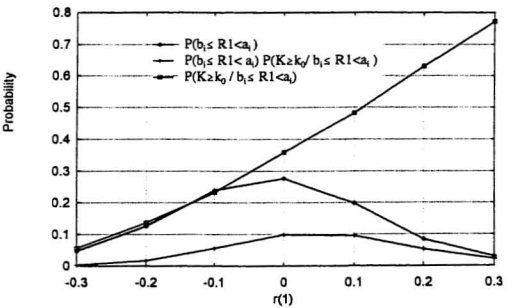

Figure 3.5b $P(b_i \leq R1 < a_i)\ P(K \geq k_o / b_i \leq R1 < a_i)$ varying with r(1) for n=50.

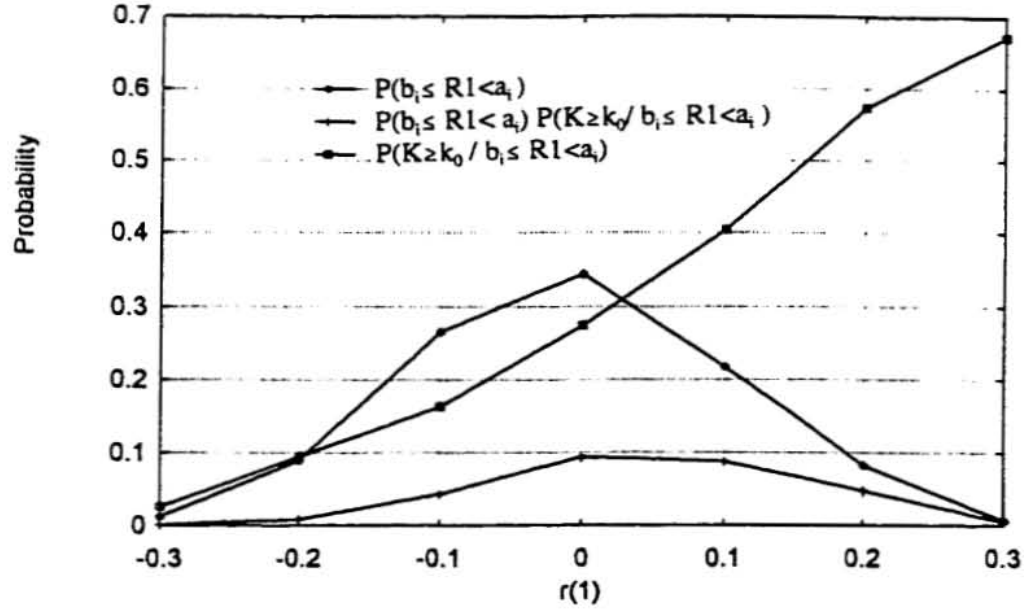

Figure 3.5c $P(b_i \leq R1 < a_i)\ P(K \geq k_0 / b_i \leq R1 < a_i)$ varying with r(1) for n=80

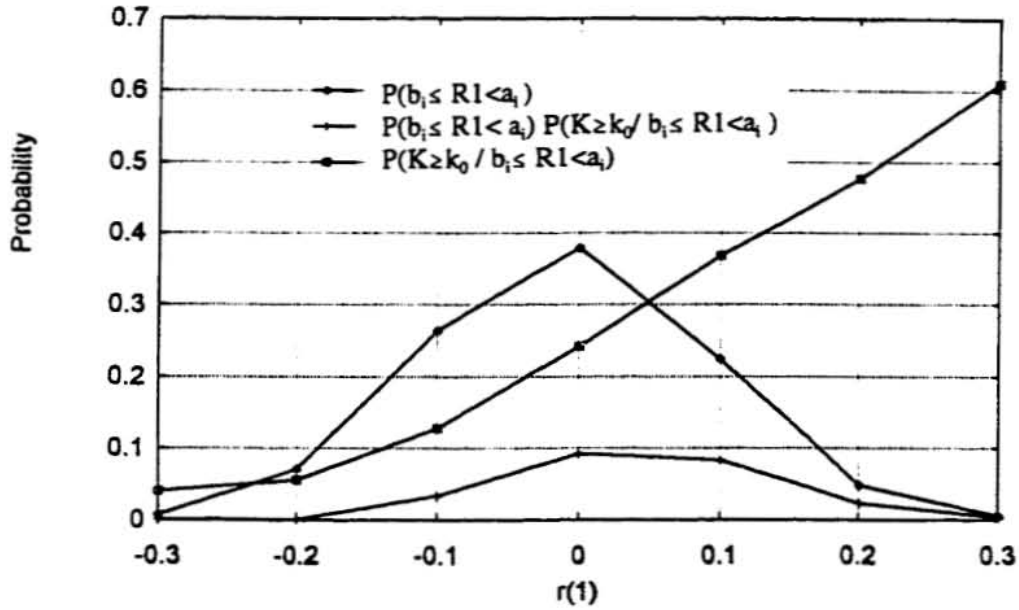

Figure 3.5d $P(b_i \leq R1 < a_i)\ P(K \geq k_0 / b_i \leq R1 < a_i)$ varying with r(1) for n=100.

3.5.2 A Common Phenomenon

"Is the simultaneous occurrence of long-term persistence and short-term independence a common phenomenon in annual peak flows?" _ the question was asked in the beginning of the previous Chapter. The following analysis will answer this.

The probability $P(b_i \le R1 < a_i)\ P(K \ge k_0 / b_i \le R1 < a_i)$ in Eq.3.7, in fact, is the probability of intersection of events A_i and B_i shown in the left of Eq. 3.5, and it is also expressed by

$$P((K \ge k_0) \cap (b_i \le RI < a_i)) = P(K \ge k_0 / b_i \le RI < a_i)\, P(b_i \le RI < a_i) \qquad (3.8)$$

From probability theory, the key word for expressing this intersection is "and" meaning " the event $(K \ge k_0)$ and the event $(b_i \le R1 < a_i)$ occurring simultaneously", and its probability is $P((K \ge k_0) \cap (b_i \le R1 < a_i))$. It is of interest to note that the event $(K \ge k_0)$ represents a peak flow series exhibiting long-term persistence, and event $(b_i \le R1 < a_i)$ represents a peak flow series exhibiting short-term independence in this study. Hence, the intersection of events defines the concept of "simultaneous occurrence" for both events.

Furthermore, Table 3.5 and Figure 3.5 show that the probability $P((K \ge k_0) \cap (b_i \le R1 < a_i))$ for each region of r(1) is greater than zero, and the maximum value for the individual regions and the total value for the given sample size 50 are 9.94% and 34.62%, respectively. Thus, the simultaneous occurrence of long-term persistence and short-term independence seem not to be an uncommon phenomenon. In fact, besides the above

quantitative description, observations and generated data also demonstrate this important conclusion.

The width of scatter in the horizontal axis, denoting r(1), in the scatter plots shown in Figs.2.2a-b and 3.2a-h decreases with an increase in sample size much faster than that in the vertical axis, denoting Hurst's K. As sample size approaches 10,000 in Fig. 3.2h, r(1) approaches zero but Hurst's K still takes a wide range, from 0.49 to 0.67. The same result is found in Table 2.9, that the mean value of Hurst's K is 0.5835 for sample size 1,000, but the mean value of r(1) is at -0.0001 which shows no significant difference from zero. Theoretically, it seems that long-term persistence is rooted in the independent series as explored in the numerical simulation.

3.5.3 A Useful Result

As mentioned previously, the probability P ($K \geq k_0$) and its distribution on the R1 axis, $P((K \geq k_0) \cap (b_i \leq R1 < a_i))$, $i=1,2,\ldots,m$, shown in Figure 3.5 are meaningful in the study of long- and short-term serial correlation. However, in flood risk analysis, it might play an important role in understanding the behaviours of long-term persistence in an independent parent probability distribution.

In chapter 2 we have analysed the annual peak flow series observed at Canadian and Chinese rivers. Suppose the population of observations be EV1 or Pearson Type III distributed, and the data transformed by the Box-Cox method become normally distributed.

The sampling distribution of Hurst's K should be changed with its parent population changing. Consequently, the proportion of the long-term persistence, in fact, is changed after transformation. It is expressed as a change on probability $P(K \geq k_0)$ and its portion $P((K \geq k_0) \cap (b_i \leq R1 < a_i))$, $i=1, 2, \ldots, m$. Table 3.6 shows the differences of statistics of the Hurst's K and r(1) for the non-transformed and transformed data in Canadian and Chinese rivers.

Theoretically, sampling distribution of Hurst's K with EV1 parent population should be different from that of Hurst's K with parent Pearson Type III in the sense. In Table 3.6, the changes of mean and standard error of Hurst's K seem not significant for transformed and original one. However, according to the central limit theory that indicates the nominal significance level is approximately the same, when the sample size becomes quite large. This concept is used to assess suitability of the probability $P(K \geq k_0)$, thus the Kolmogorov-Smirnov test(Lilliefors, 1967; Crutch, 1975) of the null hypothesis that the two large sample distributions are the same is carried out.

Here, we used the sampling distribution of Hurst's K as a population for the hypothesis test. Based on a visual way of comparison, two population samples should be a graph of the two empirical cumulative distributions. If the two empirical cumulative distributions differ greatly, it is expected that the populations being sampled were not the same. If the two curves were quite close each other, the conclusion that the underlying population distributions are essentially the same could be made.

Table 3.6 Statistics of Hurst's K and r(1) for the original and Box-Cox transformed data observed in some Canadian and Chinese rivers

Data	Source	size n	Statistics	mean	standard error	skew-ness	correlation coefficient
Trans-forma-ted data	Canadian Rivers	198	Hurst's K	0.6728	0.0728	-0.2021	0.5097
			r(1)	0.0309	0.1483	0.1480	
	Chinese Rivers	60	Hurst's K	0.6522	0.0710	0.2922	0.3373
			r(1)	-0.0184	0.0999	0.1395	
Obser-ved data	Canadian Rivers	198	Hurst's K	0.6646	0.0715	- 0.0878	0.4801
			r(1)	0.0082	0.1380	0.0214	
	Chinese Rivers	60	Hurst's K	0.6586	0.0708	0.0991	0.3124
			r(1)	-0.0128	0.1120	0.1607	

The Kolmogorov-Smirnov statistic, D, is the maximum absolute difference between two empirical cumulative distribution functions. The distribution of D is only related to the sample size.

The Kolmogorov-Smirnov test consists of: accept the hypothesis if

$$KS = \max_{x} \sqrt{nm/(n+m)}\left|F_n(x) - G_m(x)\right| = \sqrt{nm/(n+m)}D \qquad (3.9)$$

is less than or equal to the given values for the given significant levels

α	0.001	0.01	0.05	0.10
KS	$\leq$ 1.95	$\leq$1.63	$\leq$1.36	$\leq$1.22

if sample sizes are large, say, both 40 or more, where $F_n(x)$ and $G_m(x)$ are empirical cumulative distribution functions, n and m are sample sizes.

Let $F_n(k)$ and $G_n(k)$ be empirical cumulative distribution functions for the Hurst's K of original and transformed data. Figures 3.6a and 3.6b show these empirical cumulative distributions for the data of Hurst's K observed in Canadian and Chinese rivers.

The large sample distribution of D is known. Let n be large, the Kolmogorov-Smirnov test of the null hypothesis that the two large sample distributions, distribution of Hurst's K transformed, $G_n(k)$, and distribution of Hurst's K from non-transformed data, are the same. The sample sizes of Hurst's K are 120 and 60 for the data observed in Canadian and Chinese rivers, respectively. The D-values are 0.1204 and 0.1726 for Canadian and Chinese rivers, respectively. A conclusion of statistical tests summarised in Table 3.7 accepts the null hypothesis, that is, the two distributions of Hurst's K from original and transformed data are the same at the given significant levels.

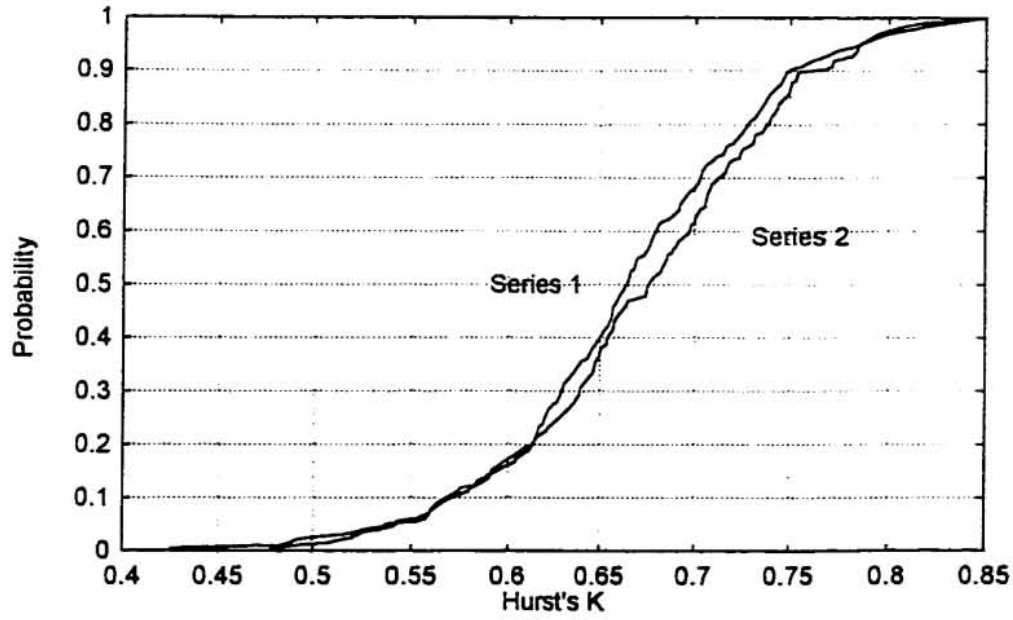

Figure 3.6a Empirical cumulative distributions for the data of Hurst's K, from Canadian rivers, where Series 1 and Series 2 are original and transformed data, respectively.

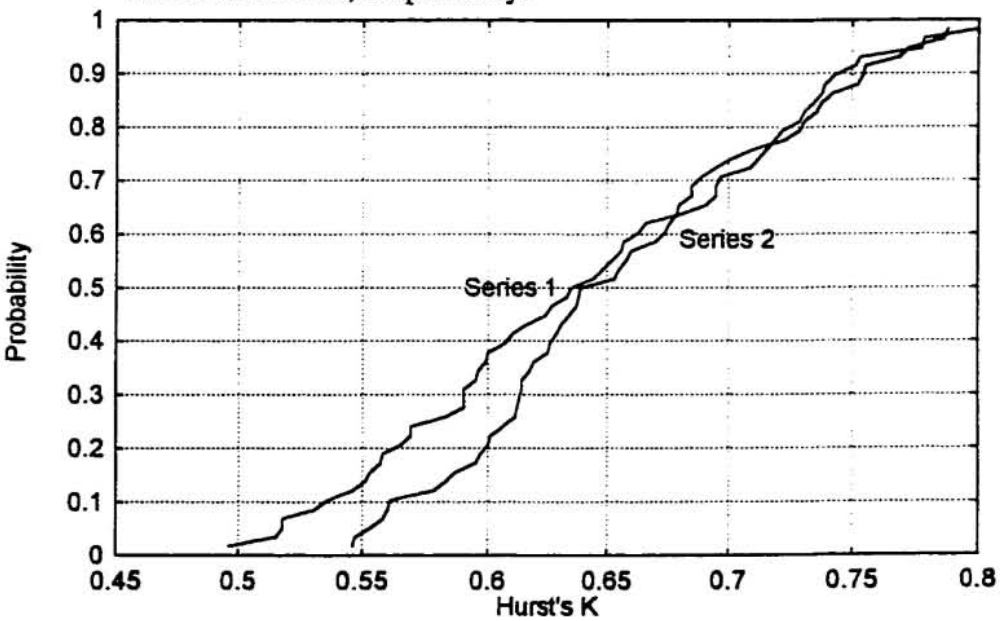

Figure 3.6b Empirical cumulative distributions for the data of Hurst's K, from Chinese rivers, where Series 1 and Series 2 are original and transformed data, respectively.

Table 3.7 Kolmogorov-Smirnov nonparametric test that two distributions of Hurst's K from transformed and non-transformed data are the same

Source	n	KS	Critical Value		Conclusion	
			$\alpha=5\%$	$\alpha=10\%$	$\alpha=5\%$	$\alpha=10\%$
Canadian Rivers	120	0.933	1.36	1.22	Accept H_o	Accept H_o
Chinese Rivers	60	0.945	1.36	1.22	Accept H_o	Accept H_o

The Kolmogorov-Smirnov test employed here is not only a nonparametric test that requires few assumptions for their validity, but also for large sample test. A statistical test is robust for large samples, but not for small samples according to the central limit theory that the normal significance level is approximately the same as the true significance level when the null hypothesis holds. The results of the Kolmogorov-Smirnov hypothesis test here are useful. The conclusion that high probabilities of existence of long-term persistence are involved in a normal independent distributed data may be suitable for the non-normal independent series. However further studies and investigation of this issue should be continued in the future.

3.6 Summary

In this chapter, the serial correlation structure of annual peak flow series has been analysed. Based on probability theory and investigated results in Chapter 2 the Hurst's K and lag-one autocorrelation coefficient r(1) are correlated and dependent, a probabilistic approach for dealing with long- and short-term behaviour of annual peak flow series was proposed.

In this approach, a Monte Carlo simulation was designed to provide a sampling distribution of Hurst's K for a given r(1). The direct use of this result is new empirical percentage points for testing long-term persistence superseding that proposed by Lye and Lin (1994). Also, a useful index, $P(K \geq k_0)$, the estimator of the population probability value of the long-term persistence for independent series such as the peak flow series to exhibit long-term persistence, was proposed and estimated.

The results of the proposed methods have useful practical implications:

1) The proposed estimator for population $P(K \geq k_0)$ and its distribution on the R1 axis, $P(b_i \leq R1 < a_i)$ $P(K \geq k_0 / b_i \leq R1 < a_i)$, shown in Table 3.5 and Figure 3.5 assure that long-term persistence and short-term independence can be quantitatively estimated.

2) The probability $P((K \geq k_0) \cap (b_i \leq R1 < a_i))$ for each region of R1 is greater than zero, and the total value for the small sample sizes range from 23.6% to 38.77%, implying that the simultaneous occurrence of long-term persistence and

short-term independence appear not to be an uncommon phenomenon.

3) Initial study of the properties of the observed and normally transformed peak flow data from Canadian and Chinese rivers indicate that the results of this study seem robust for other distributions if a normal transformation is made.

Chapter 4

Measurement at Scale ξ :

Basic Concepts of Fractal Geometry

4.1 General

From the classical statistical point of view, we have examined the behaviour of annual peak flows at two fixed scales, that is, one at scale of one, measured by the lag-one autocorrelation coefficient r(1) indicating high frequency behaviour and another at scale of n, the sample length, which indicates low-frequency behaviour measured by Hurst's K. The statistical terms "correlation" and "dependence" for expressing the relationship between the two random variables, lag-one autocorrelation coefficient r(1) and Hurst's K have also been discussed and explained. Also, based on basic probability theory, an analysis of Hurst's K and r(1) to show the simultaneous occurrence of long-term persistence and short-term independence in annual peak flows has been performed. However, what has been done previously is to look at the two separate scales for the

features of peak flows. The question that arises is " how about looking at all scales"? Can this approach tell us more about a flood peak series? To this end, fractal geometry can be employed.

Fractal geometry, a new science, lets us see the objects across scales. It may thus give us a more physical and philosophical explanation of the natural behaviour of peak flows. Therefore, the topic is now moved to fractal geometry to investigate the temporal structure of peak flow series in a fractal domain.

4.2 Fractals: Measurement of Coastlines

Fractal geometry originates from the measurement of the length of a coastline (Mandelbrot, 1967). For example, the length of the border between Spain and Portugal has two very different measurements: Spain claims 616 miles, while Portugal quotes 758 miles (Mandelbrot, 1967). Again the length of the coast of Britain in various sources varies between 4,500 and 5,000 miles. What is happening there? However, based on Richardson's empirical data (Richardson, 1961), Mandelbrot (1967, 1982) demonstrated for us that for all practical purposes, typical coastlines do not have a Euclidean length!

Figure 4.1 shows Richardson's empirical data graphically. On the horizontal axis the logarithm of the divider setting, ξ, is indicated. The vertical axis is for the logarithms of the coast length, $L(\xi)$. The log-log plots will show how the length changes when the

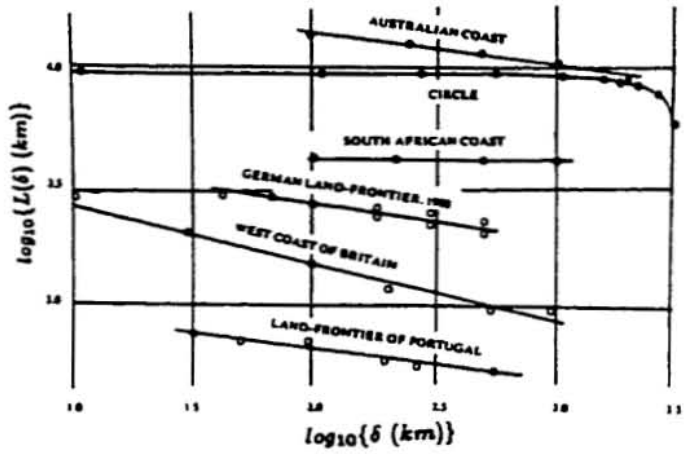

Figure 4.1 Measurement of coastlines (Mandelbrot, 1982).

divider setting is changed. Mandelbrot (1967) paid particular attention to the two constants, b and D, which characterized a power law: length of coast $L(\xi)$ is nicely approximated by a power law being $b\xi^{1-D}$, where ξ is a compass setting. The length of the coastline shows no sign of reaching a fixed value. The length of the coastline would increase without limit if the compass setting gets smaller and smaller. It is obvious that the classical measure of a coastline based on Euclidean geometry, in fact, is not meaningful because the length of the coastline goes to infinity as the ξ tends toward zero. In other words the coast length behaves as a power law that characterizes the complexity of the coastline of Britain for example by expressing how fast the length increases as the scale, ξ, is reduced.

The straight line in the log-log plots shows the behaviour of the coastline that is similar in shape and structure over the range of scale. For instance, large bays contain smaller bays, the small bays contain even smaller bays, and the smaller bays contain even smaller and smaller bays, and their shapes are similar to the whole. Thus, self-similarity is indicated. Therefore, a fractal is defined as a shape made of parts similar to the whole in some way (Mandelbrot, 1977, 1982). The constant D is defined as the fractal dimension which describes the growth law that reflects how rapidly the coastline develops as the measure $\xi \rightarrow 0$ (Falconer, 1990). Furthermore, the magnitudes of the fractal dimension, D, tell us the level of the complexity, such as the coastline of Britain is more convoluted than that of South Africa because the fractal dimension D of the former is greater than that of the latter.

4.3 Related Concepts of Fractal Geometry

From the view of "measurement at scale', a complex object, such as a coastline, leads us to see fractal shapes everywhere: Brownian motion curves, cluster deposited "tree" at electrodes, river networks, the shape of mountains, swift currents in flow and even brain waves of human and the points of earthquakes. How about the annual peak flows of interest herein? So far only one paper has dealt with temporal peak flows in a fractal world (Turcotte and Greene, 1993), but the technique used is flawed as will be explained in Chapter 6.

Before using fractal geometry to search for invariance across scales in peak flows, a brief description of the concepts of fractal geometry is given next.

Self-similarity

A fractal that is invariant under ordinary geometric similarity is called self-similar (Mandelbrot, 1977, 1982). Strict self-similarity over all ranges of scale is found in a classical mathematical fractal, such as the well known Cantor set, the Koch curve, and the Sierpinski gasket etc. Self-similarity over limited ranges of scale is common in nature. For example a cauliflower, the branches when compared with the whole are similar, only smaller. These clusters again can be decomposed into smaller clusters, which again are similar to the whole as well as to the original branches. Are annual peak flows self-similar on time axis? If they are, what does it mean, and what are the implications?

Scaling or scale-invariance

Most fractals are invariant under certain transformations of scales. They are called scaling or scale-invariance (Mandelbrot, 1977). Mathematically, points $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are mapped to the new points $\mathbf{x}_{new} = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n)$ by the same factor λ is scaling or scale-invariance. The scaling property of an object is shown as a straight line in the customary log-log plot in fractal studies.

Self-affinity

In many cases the structure of the objects is invariant with respect to the different scaling ratio λ_i. Thus, a fractal that reproduces itself in some sense under an affine transformation is called self-affine (Mandelbrot, 1982). An affine transformation that transforms from points $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ into new points $\mathbf{x}_{new} = (\lambda_1 x_1, \lambda_2 x_2, \ldots, \lambda_n x_n)$, where the scaling ratios $\lambda_1, \lambda_2, \ldots, \lambda_n$ are not all equal is an antisotropic fractal.

The Brownian random process is a typical self-affine distribution under a transformation that changes the time scale and the length scale by different factors. Based on the self-affine property of Brownian motion, fractional Brownian motion (fBm) was proposed (Mandelbrot and Van Ness, 1968; Mandelbrot and Wallis, 1968, 1969a). It is meaningful that the correlation function of fBm, Cor(t), expressed by $2^{2H-1} - 1$, theoretically implies that Cor(t) is independent of time when $H = ½$! However, for $H \rightarrow ½$ it leads to persistence or antipersistence forever in a time scale! (Mandelbrot, 1977, 1982)

The fractal dimension

The fractal dimension is an important measure in fractal geometry, and its definition is based on the idea of "measurement at scale ξ" (Falconer, 1990). It reflects the degree of irregularity when examined at scale ξ. One of the most widely used fractal dimensions is the box dimension (or box counting dimension or capacity dimension), D_b.

Let S be a subset of R^n, where n=1, 2, or 3. The box-counting dimension of S is:

$$D_b = \lim_{\xi \to 0} \frac{\ln N(\xi)}{\ln (1/\xi)} \tag{4.1}$$

if the limit exists, where N(ξ) is the smallest number of n-dimensional boxes of side length ξ required in order to completely cover S (Falconer, 1990).

The box-counting dimension, D_b, is usually empirically estimated by the gradient of a ln -ln graph of N(ξ) against ξ plotted over a suitable range of ξ.

4.4 Measurement at Scale ξ

Returning to the concept of measurement at scale, Figure 4.2 shows the measurement at scale for Hurst's K, lag-one autocorrelation coefficient r(1) and fractal dimension D_b.

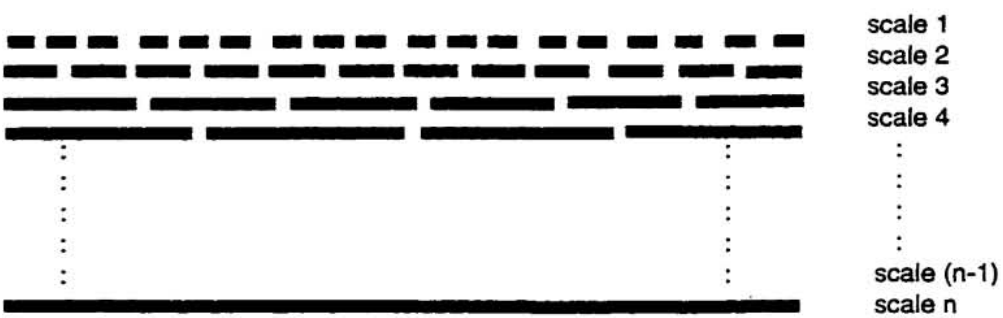

Figure 4.2 Measurement at scale.

Hurst's K is simply considered as the measurement at scale n, a long-term scale, lag-one autocorrelation coefficient r(1) at scale one, a short-term scale. The fractal dimension D_b crosses over various scale related to scales 1 and n. Thus, the fractal dimension D_b plays an important role in looking at the natural behaviour of flood peaks from a non-classical viewpoint.

Recall in Chapters 2 and 3 that we have studied the behaviour of annual peak flows at two scales or seen the properties of annual peak flows at high and low frequencies based on the traditional principle of "taking things apart", even though we have used an artful scheme to connect both, Hurst's K and r(1). But, we viewed this only at the individual scales. Even though a much better understanding of the natural behaviour of annual peak flows was achieved by methods discussed in previous chapters, more detailed features across scales for the peak flows were not possible using the classical stochastic methods.

Statistics emphasize identification, independence, homogeneity and stationarity of the observations, but fractal geometry has partiality for something being fragment, irregular, and disordered , and would display an evolution at scale saying "how rapidly the irregularities develop as $\xi \rightarrow 0$". These concepts will be developed in the next Chapter so that a totally different description of peak flows can be achieved.

4.5 Summary

An initial view of the basic concepts of fractal geometry for studying annual peak flows has been taken. The "measurement at scale" provides a useful tool to study the serial correlation structure of annual peak flows in the following Chapters.

Chapter 5

Scaling Behaviour of Peak Flow Series

5.1 General

Many hydrological time series, such as daily and monthly flow series, which are continuous or discrete can be described by a random function or a stochastic process that can serve as an image of the time series and explains the structure of the observations. Peak flows as a set of points distributed in a time axis are isolated points and completely disordered to the point that classical methods are unable to distinguish differences among various types of peak flow point sets distributed along the time axis.

Figure 5.1 illustrates the distributions of peak flows along time axis at two Canadian rivers. Using conventional statistics, it is difficult to explain how different they are and how well any mathematical paradigm could "translate" them to an "image", because previously there was no suitable methodologies to describe their behaviour in the time axis.

The application of concept of scaling has achieved some success in the field of hydrology (National Research Council, 1991), e.g. dealing with spatial and temporal distribution of rainfall (Zawadzki,1990; Gupta and Waymire, 1990; Kedem and Chiu, 1987; Lovejoy and Mandelbrot, 1985; Lovejoy and Schertzer, 1990; Waymire, 1985;

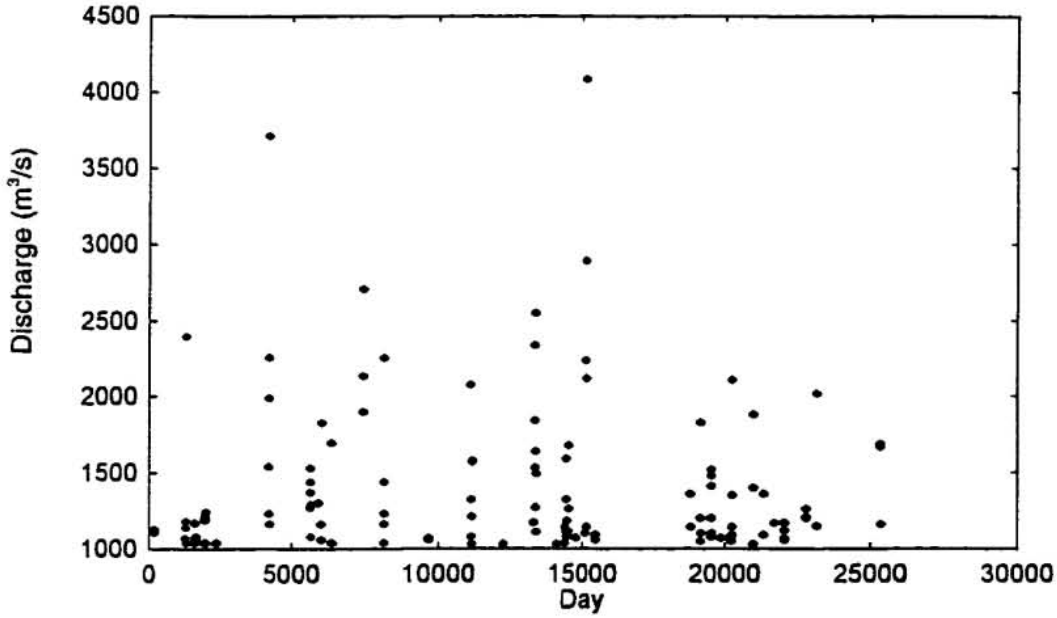

Figure 5.1a Peak flow distribution in the time axis observed at Medicine Hat gauge station, Saskatchewan River, Alta., Canada.

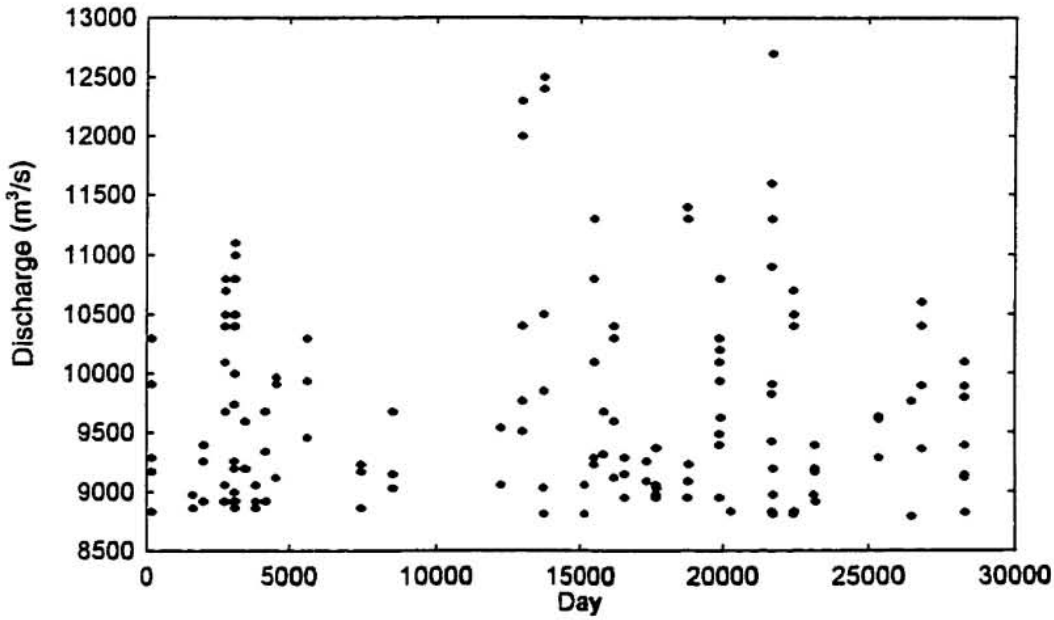

Figure 5.1b Peak flow distribution in the time axis observed at Hope gauge station, Fraser River, B.C., Canada.

Waymire and Gupta, 1987; Olsson et al., 1992; Olsson and Janusz Niemczynowicz, 1996; Venugopal and Foufoula-Georgiou, 1996; Paolo Burlando and Renzo Rosso, 1996; Puente, 1996; Haitjema and Kelson, 1996) and in discussions about spatial and temporal scaling distribution of river flows (Gupta and Waymire, 1990; Wu and Hou, 1991; Jayawardena and Lai, 1994; Gupta et al., 1996). The temporal scaling behaviour of peak flows, however, has not been fully considered up to now except that of Turcotte and Greene (1993) who hypothesized that the annual peak flows are sufficiently scale invariant over time scales from one to one hundred years.

Peak flow observations are considered as a point set distributed on a time axis. These observed points are then related to the probabilities of occurences for given thresholds. Thus, a family of curves can be constructed to explain the feature of peak flow points. A technique in fractal geometry called the functional box counting procedure will be used to construct the family of curves.

A number of peak flows observed from Canadian and Chinese rivers are subsequently analysed. It is hypothesised that this approach will provide new insights into the possible scaling behaviour hidden in peak flow evolution for the river flows investigated.

5.2 Concept and Methodology

5.2.1 Box-Counting Dimension

The fractal dimension is an important tool of fractal geometry, and its definition is based on the idea of "measurement at scale ξ" (Falconer, 1990). One of the more widely used fractal dimensions is the box-counting dimension discussed in Chapter 4.

Broadly speaking, the box-counting dimension D_b, defined in Eq.4.1 says that $N(\xi) \sim \xi^{-Db}$ for small ξ. The box-counting dimension, D_b, is usually empirically estimated as the gradient of a ln - ln graph of $N(\xi)$ against ξ plotted over a suitable range of ξ. Then, using least squares, the regression equation for estimating parameters of the graph line in the logarithmic domain is obtained, i.e.

$$\ln N(\xi) = c - D_b \times \ln \xi \qquad (5.1)$$

where $N(\xi)$ is the smallest counting number of n-dimensional boxes of side length ξ required to completely cover a set, c is the intercept, and D_b is the box counting dimension.

5.2.2 Functional Box Counting Algorithm

The method herein employed to investigate the temporal scaling behaviour of peak flow points is the functional box counting algorithm (Lovejoy et al., 1987). This method

transforms observations into a set of points whose dimension can be estimated by box counting.

The fundamental concept of functional box counting is to consider a function, f(x), which is transformed to an exceedence set:

$$\{A_T \mid f(x) > T\} \tag{5.2}$$

where A_T is defined by threshold T. If A_T exhibits scaling behaviour, the number of boxes to cover A_T , $N_T(\xi)$, can be expressed as

$$N_T(\xi) = \xi^{-D_b(T)} \tag{5.3}$$

Thus, the functional box counting method characterizes a scale invariant set.

5.2.2.1 Two Aspects of Practical Importance

From an engineering viewpoint, two aspects of practical importance will be proposed in the functional box counting algorithm:

A probabilistic approach

From the perspective of fractal geometry, the measure of peak flow structure is to look at the feature of peak flows at varying time scales, ξ. Presupposing peak flow points are measured by decreasing time scales, ξ, or peak flow points are involved varying

time intervals, ξ. Ignoring the difference between occurrence of peak flow points within the same time interval, our interest is in the probability that a step of measure with time scale, ξ, that includes at least one peak flow point. Hypothetically, a fraction or probability of a time interval including at least one occurrence of peak flows, P_ξ, is a function of time interval length ξ, for a given threshold. This situation is similar to the construction of a Cantor set (Mandelbrot, 1977, 1982), in which the probability that a step of length ξ includes a line segment that can be obtained through the construction procedure. However, the relationship between the probability P_ξ and time intervals, ξ, will reveal the structure of peak flow on the time axis.

Thresholds available

For a corresponding data set, thresholds will be selected so that the properties of the temporal scaling behaviour of peak flows which are the events of interest can be investigated.

For a given threshold, such as an annual maximum flow, the peak flow points can be transformed from the observed data, suggesting that a peak flow point-process, a kind of "POT" series (NERC, 1975), i.e. "Peaks Over Threshold", appears in the time axis, then the distribution of the high-level exceedences can be identified.

For example, consider the case of the daily flow observed at the Yichan gauging station, Yantze River, China (Ministry of Water Resources, 1985), as shown in Figure

5.2a. Let the threshold Qs be 50,000m^3/s. Figure 5.2b shows the point process that is transformed from the daily flows at the threshold of 50,000 m^3/s. The flow points transformed from the daily flows are conveniently considered as the instantaneous flows, in comparison with the whole observed duration. Based on the transformed data in Figure 5.2b, a fractal analysis of peak points on the time scale will be performed.

5.2.2.2 Box-Counting Dimension in the Probabilistic Approach

The probability, P_ξ , that an interval of length ξ includes at least one peak flow event, can be obtained through the functional box counting procedure:

Assume probability, P_ξ , is expressed as the following

$$P_\xi = N(\xi) / N \tag{5.4}$$

where $N(\xi)$ and N are the number of an time interval of length ξ and the total number of time intervals, respectively, where $N = L / \xi$ and L is the total length of the studied time.

Further, let

$$P_\xi = N(\xi) / (V / \xi^d) \tag{5.5}$$

where V is the total volume of measured objects, d is the Euclidean dimension of the object, where its magnitude of a point is zero, of a line one, of a square two, and of the cube three. Having drawn the $\ln P_\xi$ - $\ln \xi$ curve and determined the slope of the scaling

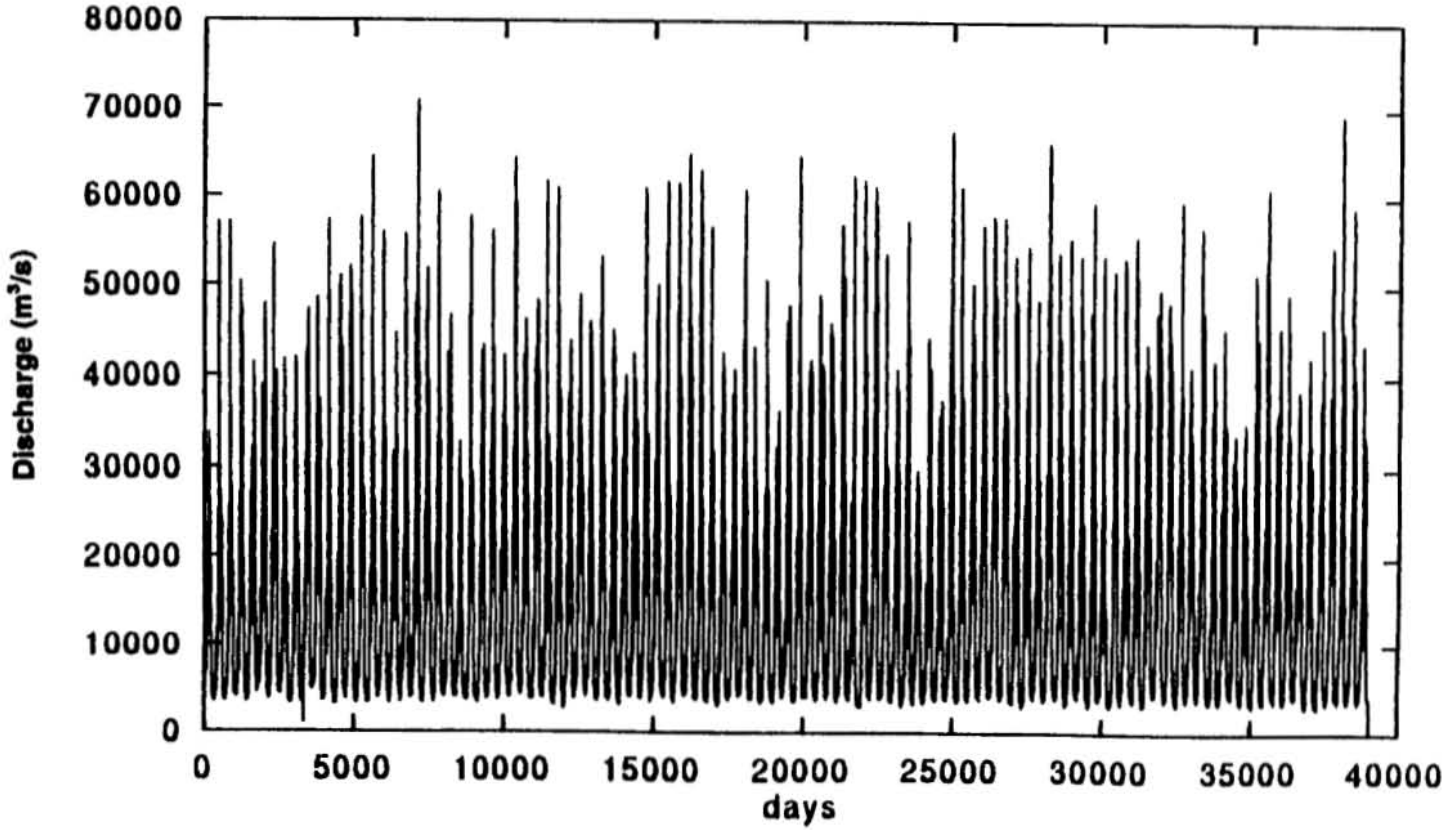

Figure 5.2a Daily flow observed at Yichan gauge station, Yangtze River, China.

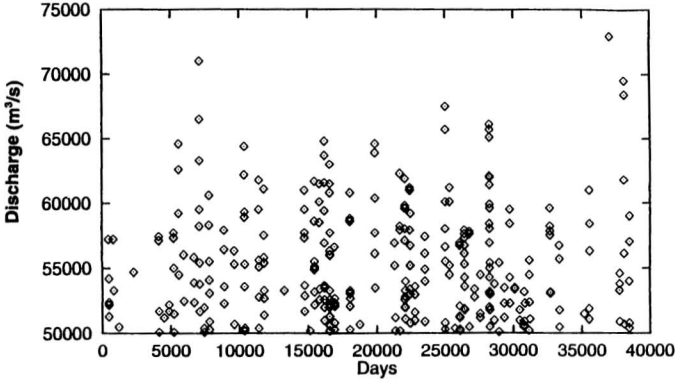

Figure 5.2b Peak flow points of Fig. 5.2a over the threshold, Qs=50,000 m^3/s.

range, an empirical box-counting dimension, D_b, can be found. The regression equation for fitting the straight line in a logarithmic domain is

$$\ln P_\xi = a' + b \ln \xi \tag{5.6}$$

where b is the slope of the straight line.

Inserting Eq. 5.5 into Eq. 5.6, there is

$$\ln N(\xi) = a' + \ln V - (d - b) \ln \xi \tag{5.7}$$

Comparing Eqs. 5.7 and 5.1,

$$D_b = d - b \tag{5.8}$$

the empirical box-counting dimension can be obtained by Eq. 5.8. Thus, the relationship between P_ξ and ξ is

$$P_\xi \sim \xi^{d-D_b} \tag{5.9}$$

5.2.3 Construction of the ln P_ξ - ln ξ - Qs Family of Curves

In order to explore the temporal structure of peak flow points, a family of curves of lnP_ξ - ln ξ - Qs is proposed.

5.2.3.1 A Symbolic Description of Exceedences

Supposing that there are observations of the peak flow points within the total observed time length L shown in Figure 5.2a, for a given threshold, flows exceeding this threshold can be projected on the time axis, located in the corresponding time intervals. With the aid of the idea of symbolic dynamics (Nicolis and Prigogine, 1989) a description of the exceedences can be made.

We use the symbol "_", namely "yes", to denote occurrence of peak flows in time intervals (boxes) of size ξ, and similarly, the symbol " ", namely "no", to indicate the non-exceedences in the time intervals. For a given threshold, Qs, and the time interval $\xi_{i=1}$, the peak flow points in Fig.5.2b are projected into symbols "_" or " ". So the peak flow sequence is changed into a sequence of symbols shown in Fig.5.3 with $\xi_{i=1}$ =730 days as an example.

Changing the length of the time intervals, ξ_i, i= 2, 3, ..., k, the same exceedences of peak flow points are projected in the changed time intervals, ξ_i, and the probability, P_{ξ_i}, that a step of interval length ξ includes at least one peak flow point, can be estimated, in which the number of "_" boxes, $N(\xi)$, contains at least one "_" sign, over the total number of boxes, N, for a given threshold. Under the different time intervals, the features of exceedence points can be described. The procedure proposed is shown in Figure 5.3.

Changing thresholds, with the same procedure performed, the structure of peak flows is described.

5.2.3.2 Steps in Construction of a Family of Curves

The three steps to construct a family of curves of ln P_ξ - ln ξ - Qs are as follows:

1) For a given threshold flood, Qs, calculate P_ξ according to the various scales ξ_i, i= 1, 2, ... k, where k is an integer, N(ξ) is the total number of the symbols of "_", and N is a ratio of L to ξ_i .
2) Draw a curve through the scattered points ln P_ξ - ln ξ over the range of scales. If a straight line exists, its slope, b, is estimated by the least squares method and an estimate of box-counting dimension D_b for a given threshold Qs can be determined by Eq. 5.8, where Euclidean dimension, d, of the problem being considered, is one. If there are several straight line sections, several fractional dimensions can be determined over different ranges of scales.
3) For a given set of thresholds, the above steps are repeated, resulting in a family of curves ln P_ξ - ln ξ - Qs.

If the probability, ln P_ξ, linearly increases with the time scale, ξ, in the log-domain, then a power law exists

$$P_\xi \sim \xi^{1-D_b} \tag{5.10}$$

therefore, it can be concluded that the probabilities of occurrence of peak flows are invariant for a given threshold within the specific scaling range.

As an example, for step one, Figure 5.3 shows the symbolic description of the peak

flow process for a threshold of 50,000 m^3/s; for step two, a constructed curve ln P_ξ - ln ξ - Qs with Qs=50,000 m^3/s is shown in Figure 5.4a , and Figure 5.4b shows a family of curves of ln P_ξ - ln ξ - Qs for the same observations with three different thresholds, Qs_i, i=1, 2, 3.

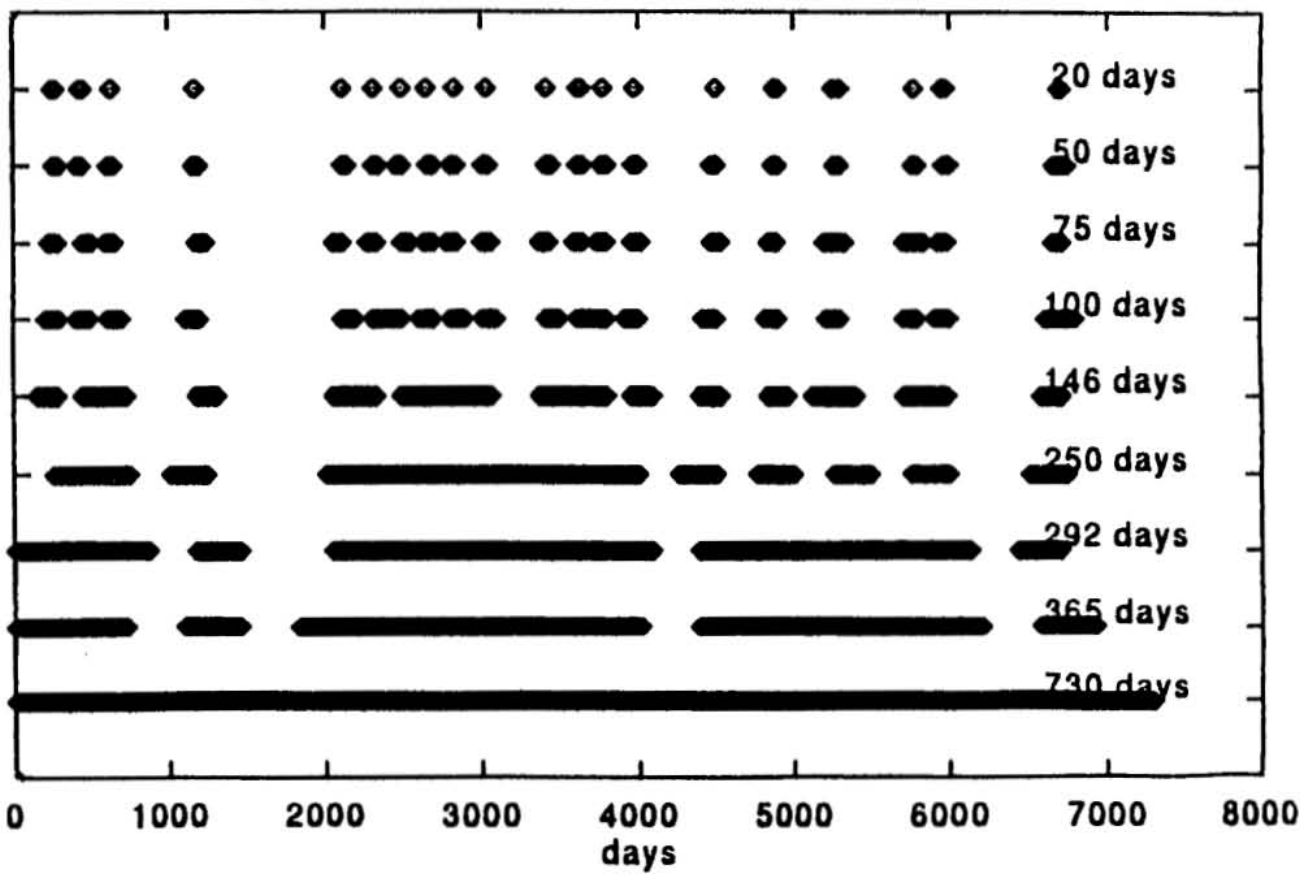

Figure 5.3 Symbolic dynamics describes the procedure of constructing a curve of $\ln P_\xi$ - $\ln \xi$ - Qs for the observations of peak flows shown in Figure 5.2 b

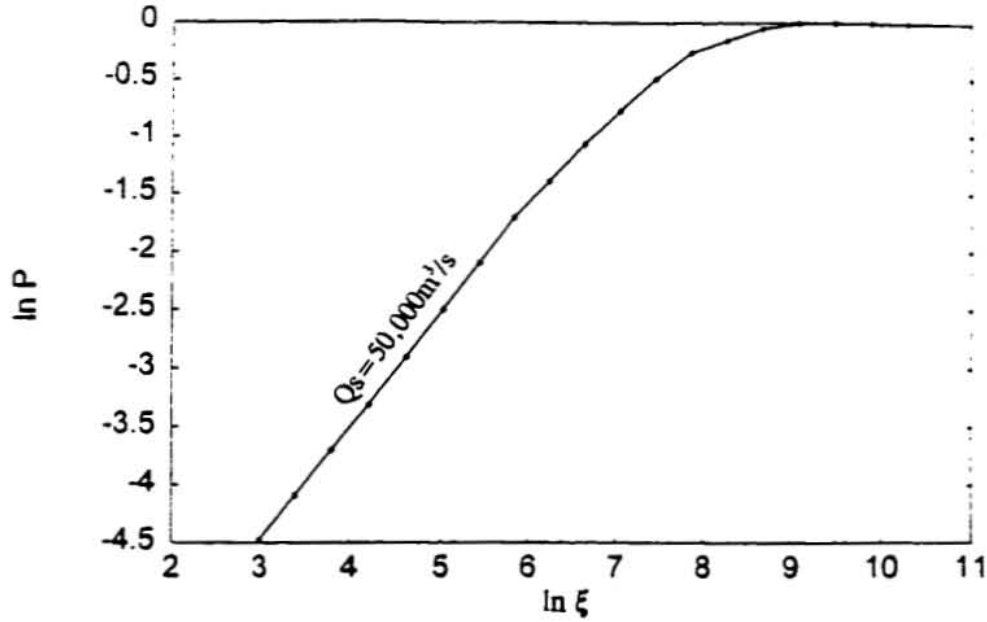

Figure 5.4a Constructed curve ln P_ξ - ln ξ - Qs with Qs=50,000 m^3/s for the observations of peak flows shown in Figure 5.2b.

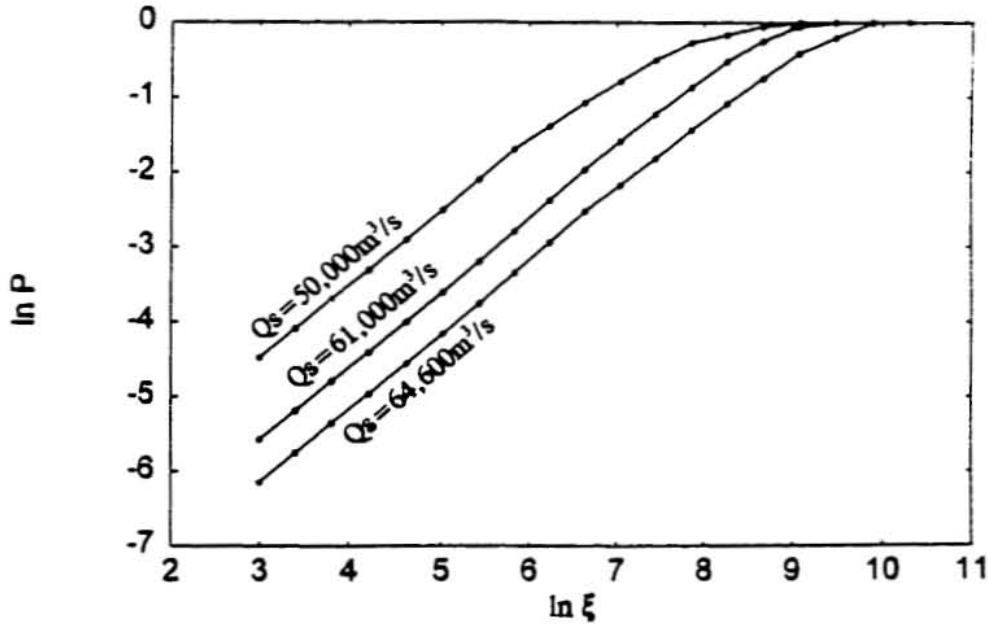

Figure 5.4b Constructed curve ln P_ξ - ln ξ - Qs with various Qs for the observations of peak flows shown in Figure 5.2b.

5.3 Scaling Behaviour of Peak Flows

Family of curves ln P_ξ - ln ξ - Qs can be used to describe features of peak flows across scales. In order to display the scaling behaviour of peak flows, the longest daily flow record observed at Yichan on the Yangtze River, China, and the daily flows collected from a number of Canadian rivers were used as shown in Table 5.1. The daily data are transformed into peak flow points along a time axis using the functional box counting algorithm, in which the probabilities of exceeding a given threshold are related to the scale of measurement in order to give a family of such curves at various thresholds.

The results of a constructed family of curves ln P_ξ - ln ξ - Qs are shown in Figures 5.5 and 5.6a-g. The estimates of the box-counting dimensions, D_b, of observations at Yichan, are illustrated in Table 5.2. As expected, all the box-counting dimensions D_b equal values between 0 and 1. That is, the peak flow points are more than just one point (dimension 0), and much less than the length of a line or curve (dimension 1).

Table 5.1 Characteristics of daily flows collected from Canadian rivers (Environment Canada, 1992)

WSC Number	Prov.	Gauging Station	Drainage (km^2)	years
05AJ001	Alta.	South Saskatchewan River at Medicine Hat	56,400	1913 -1990
05CC002	Alta.	Red Deer River at Red Deer	11,600	1913 -1990
05DF001	Alta.	North Saskatchewan River at Edmonton	28,000	1912 -1990
05HG001	Sask.	South Saskatchewan River at Saskatoon	41,000	1912 -1990
05KJ001	Man.	Saskatchewan River at the Pas	347,000	1913 -1990
08MF005	B.C.	Fraser River at Hope	217,000	1913 -1990
05OC001	Man.	Red River at Emerson	102,000	1913 -1990

Table 5.2 An illustration of box-counting dimension and corresponding scaling range of observations at Yichan, Yangtze River, China.

Qs	D_b	Scaling Range	Qs	D_b	Scaling Range
29800	0.506	60-200 days	50500	0.136	30 days- 1 year
38600	0.350	60 days-1 year	53300	0.095	30 days- 1 year
40200	0.318	60 days-1 year	54600	0.077	30 days- 1 year
41600	0.298	60 days-1 year	55600	0.065	30 days- 1 year
41900	0.293	60 days-1 year	56700	0.064	30 days- 1 year
42100	0.292	60 days-1 year	57800	0.063	30 days- 2 year
43500	0.261	60 days-1 year	59000	0.057	30 days- 3 years
44000	0.258	60 days-1 year	61000	0.021	30 days- 6 years
45300	0.236	60 days-1 year	62300	0.020	30 days- 8 years
46300	0.223	60 days-1 year	64600	0.018	30 days- 15 years
48000	0.187	60 days-1 year	66100	0.014	30 days- 25 years
48500	0.178	60 days-1 year	66600	0.008	30 days- 50 years
49300	0.167	60 days-1 year	77800	0.003	30 days- 55 years

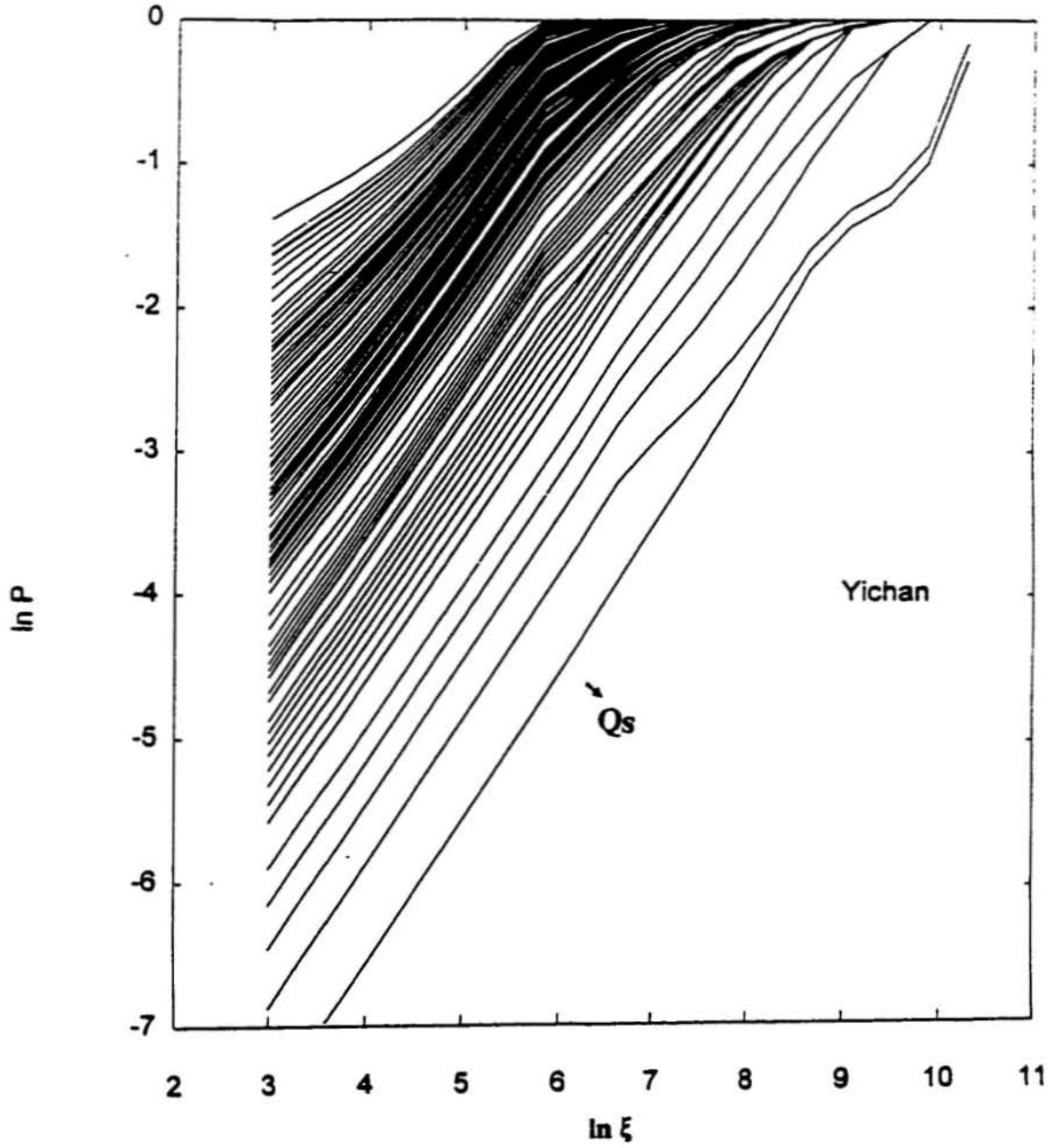

Figure 5.5 A family of curves ln P_ξ - ln ξ - Qs for the peak flows observed at Yichan, Yangtze River, China.

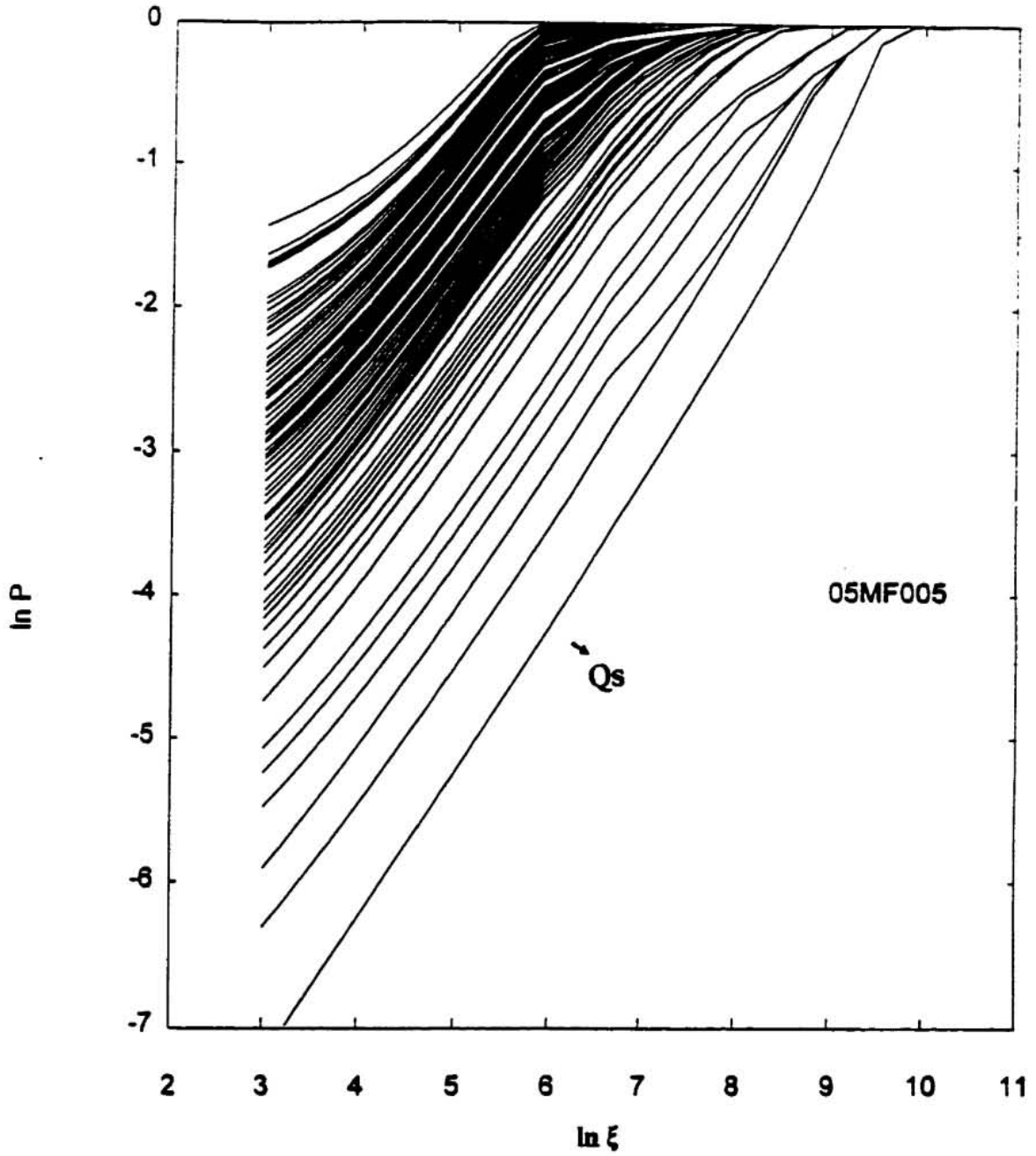

Figure 5.6a A family of curves ln P_ξ - ln ξ - Qs for the peak flows observed at Hope, Fraser River, B.C., Canada.

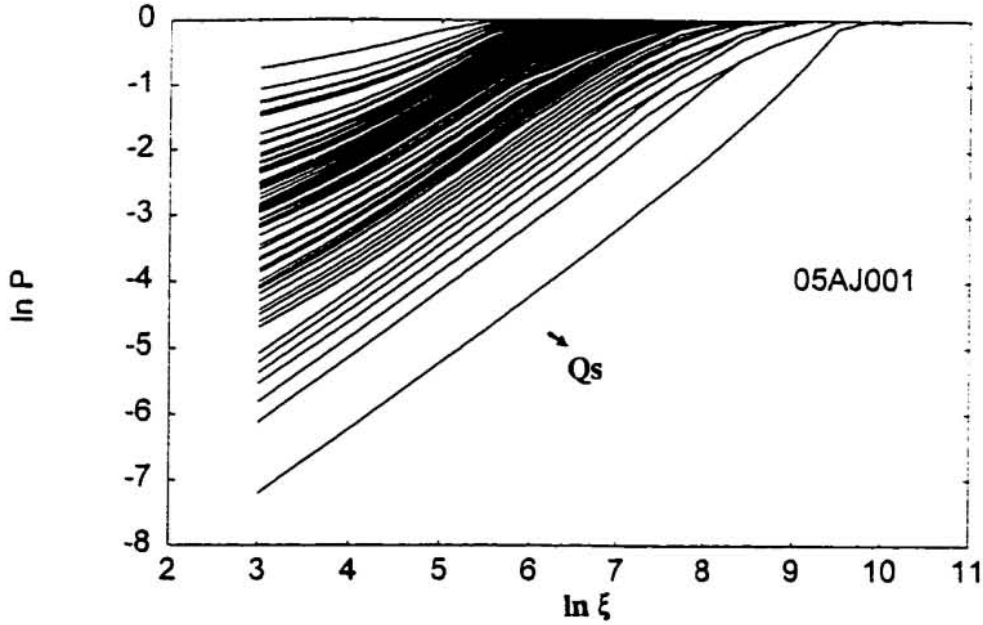

Figure 5.6b A family of curves ln P_ξ-ln ξ-Qs for the peak flows observed at Medicine Hat, South Saskatchewan River, Alta., Canada.

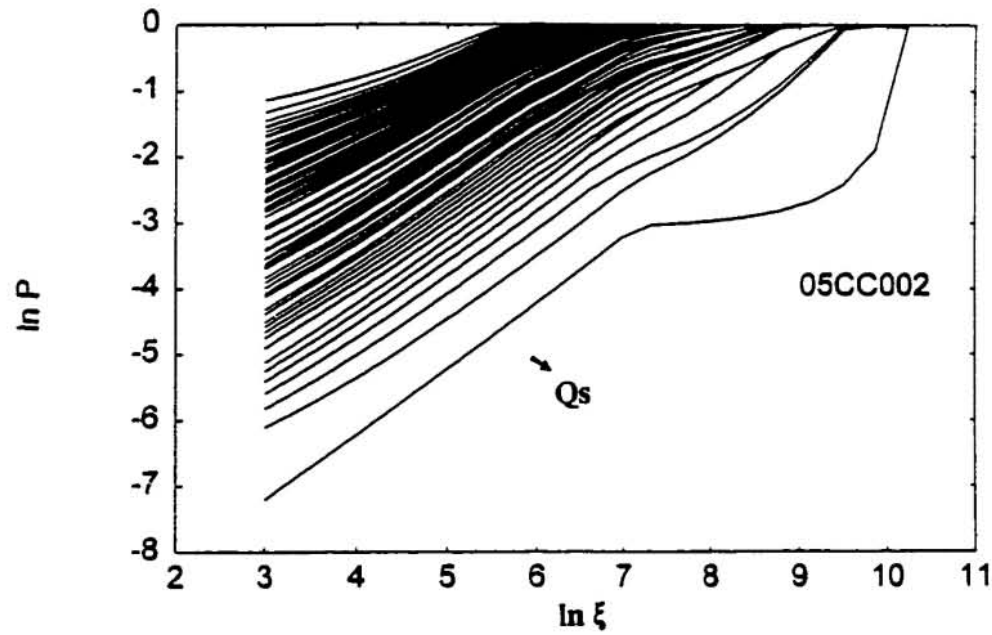

Figure 5.6c A family of curves ln P_ξ - ln ξ - Qs for the peak flows observed at Red Deer, Red Deer River, Alta., Canada.

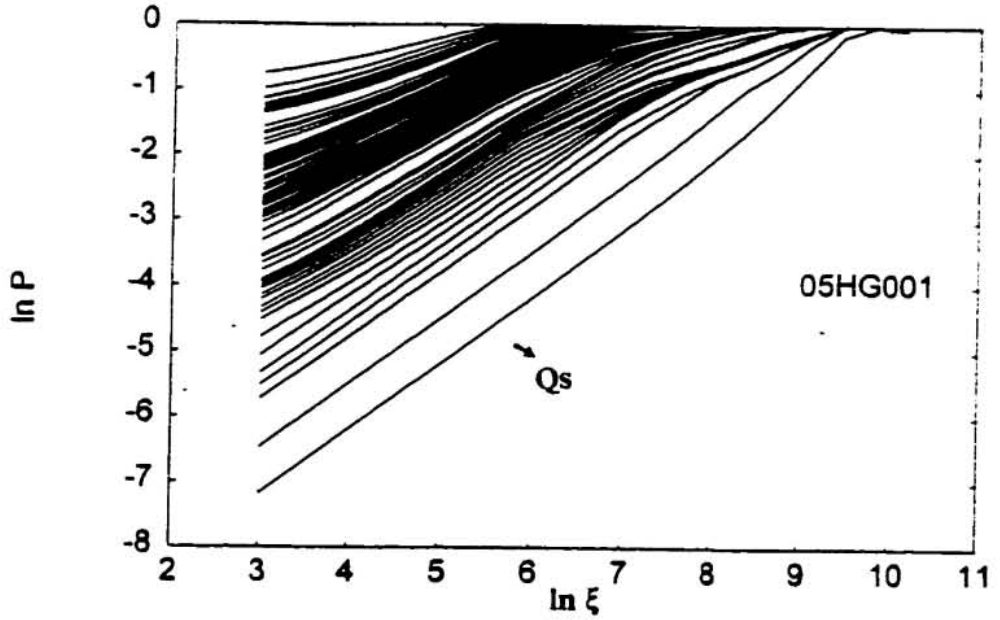

Figure 5.6d A family of curves ln P_ξ - ln ξ - Qs for the peak flows observed at Saskatoon, South Saskatchewan River, Sask., Canada.

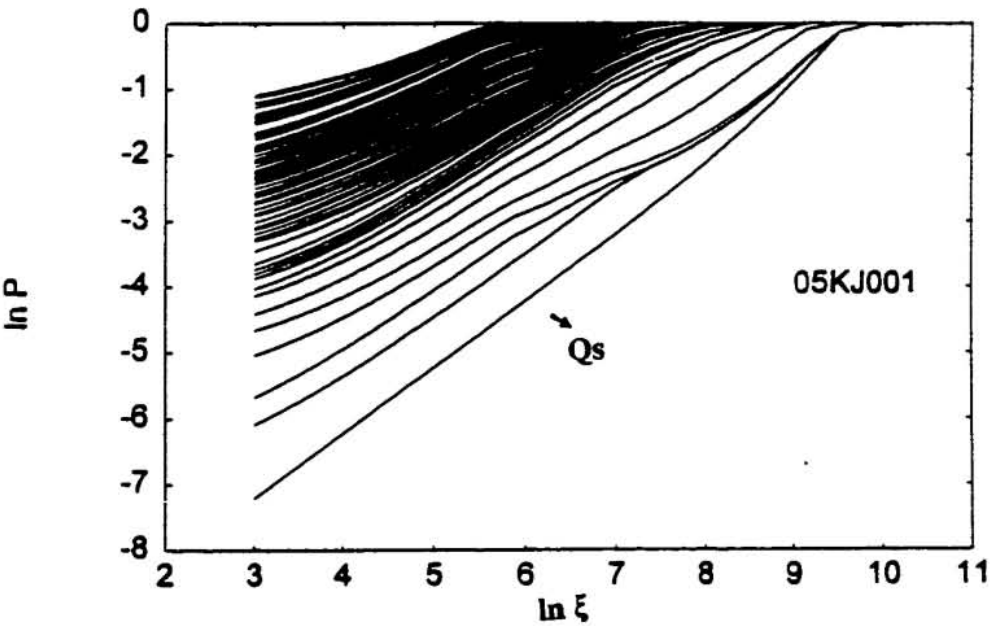

Figure 5.6e A family of curves ln P_ξ - ln ξ - Qs for the peak flows observed at the Pas, Saskatchewan River, Man.,Canada.

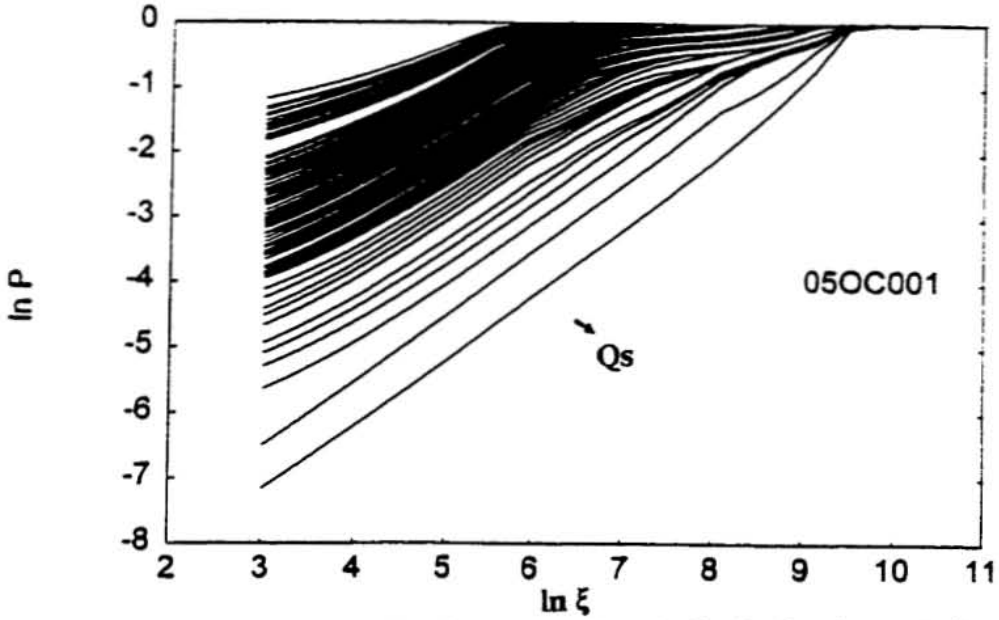

Figure 5.6f A family of curves ln P_ξ - ln ξ - Qs for the peak flows observed at Emerson, Red River, Man.,Canada.

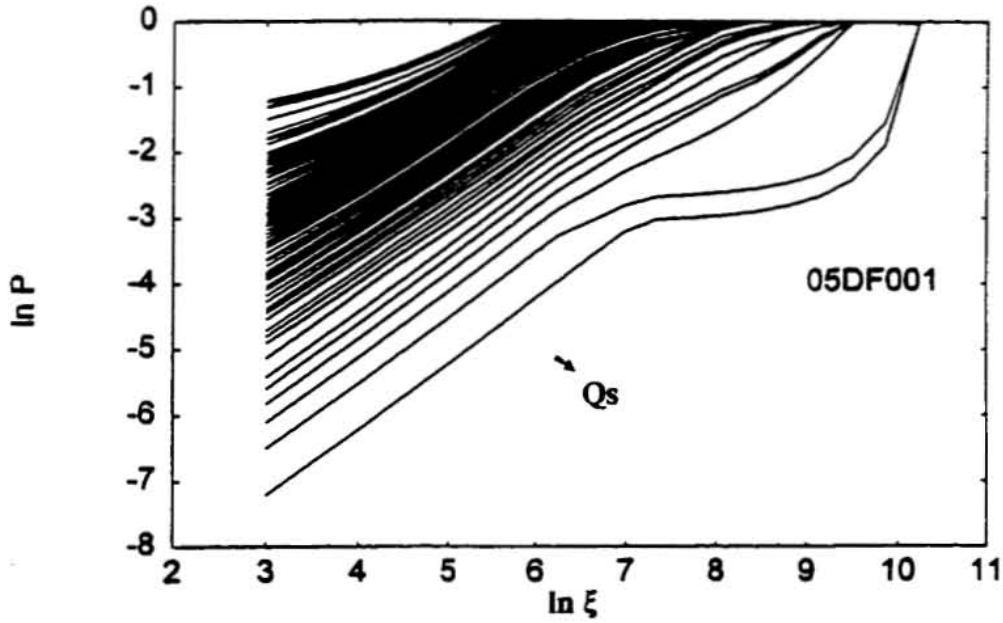

Figure 5.6g A family of curves ln P_ξ - ln ξ - Qs for the peak flows observed at Edmonton, North Saskatchewan River, Alta., Canada.

The family of curves ln P_ξ - ln ξ - Qs in Figures 5.5 and 5.6 contain a set of "message" or "texts" described by:

Probability P_ξ

The family of curves represents the probability distribution of a time interval including at least one occurrence of peak flows for a set of thresholds. The variation of probability P_ξ in the logarithmic domain as shown in curves ln P_ξ - ln ξ - Qs is related to the variation of ξ, the greater the time interval ξ , the greater the probability of exceedence of peak flows. When time intervals ξ increase to a certain size, the probability that the time interval includes at least one occurrence reaches one.

Thresholds Qs:

Threshold Qs as a parameter in the family of curves ln P_ξ - ln ξ - Qs reveals the peak flow structure at different levels of thresholds. The temporal structure of peak flow points shown in the curve depends on the threshold, the higher the threshold, the sparser the pattern, the lower the threshold the more clustered the peak flow structure. According to a specified problem, a suitable set of thresholds is designed, the peak flows correlation structure across scale could be clearly revealed.

Box-counting dimension D_b :

The slopes of the curves shown in Figs. 5.5 and 5.6 are steeper, with increasing thresholds, Qs, which result in decreased box-counting dimensions D_b . Thus, the higher the box-counting dimension, the denser the time structure, and vice versa. In the limiting case, where the slope of the curve is unity for a special threshold, which we call the upper threshold, Qs_u, the box dimension equals zero. That is, they are a set of isolated points. In contrast, when the slope is zero for a lower limit of threshold, Qs_l, $D_b=1$, i.e. a horizontal line. When D_b is between (0,1), some form of scaling of peak flows exists. Thus, the box-counting dimension used here reveals the temporal scaling behaviour of the peak flow point set and is a measure of how the peak flow points will fill the time axis occupied.

Segments of straight lines on the curves:

Each $\ln P_\xi$ - $\ln \xi$ -Qs curve is related to the temporal distribution of the occurrence of peak flows over a threshold. It has a segment of straight line for a certain range of ξ. Thus, the power law shown in Eq. 5.10 is valid and it makes sense to show that an occurrence of exceedences of flows involved in the time intervals, displays invariance or self-affinity within the corresponding scaling range, while ξ gets bigger or smaller outside this range, the power law fails. The wider ranges of straight lines in the family of curves also indicate existence of possible correlation across temporal scales due to similar

variation.

The whole system of peak flows:

If a family of curves ln P_ξ - ln ξ - Qs is considered as a system, this system maps from the peak flow point set to a graphic interpretation which contains all information about the peak flows distributed on time scale and represents inter-scale correlation between these observations for a given watershed.

The significance of those results revealing the existence of scaling behaviour of peak flows is discussed as follows.

5.4 Practical Implications

The proposed family of curves of ln P_ξ - ln ξ - Qs gives us a richer insight into the nature of peak flows. Some practical implications for the family of curves are explained below.

5.4.1 Existence of Scaling Behaviour of Peak Flow

For a wide temporal scale range, a family of curves ln P_ξ - ln ξ - Qs constructed from the peak flows observed in Canadian and Chinese rivers provided a power law between the probability of a time interval of length ξ including at least one occurrence, P_ξ, and the time interval of length ξ. It appears that there is an existence of scaling behaviour of peak flow and the same variation of occurrence of peak flows across scale exists. This phenomenon

resembles the Hurst's phenomenon discussed in Appendix A and Chapter 2.

In classical probability theory, a point event process on the time horizon is usually considered as a Poisson process, $\{\Lambda(t): t>0\}$, in which events occur instantaneously and independently on a time axis, and

$$P\{\Lambda(t) = i\} = \frac{e^{-\lambda t}(\lambda t)^i}{i!}, \quad i = 0, 1, 2, \ldots \tag{5.11a}$$

Corresponding to $\{\Lambda(t)\}$, which also is interpreted as the number of arrival in intervals, there are arrival epochs, $0 \le t_1 \le t_2 \le \ldots$, and inter-arrival times, $\tau_1 = t_1$, $\tau_i = t_i - t_{i-1}$, $i=2,3, \ldots$. The probability $P\{\tau > t\}$ occurs is equal to the probability of no occurrences $P(\Lambda(t)=0\}$, such as

$$\begin{aligned} P\{\tau > t\} &= P\{\Lambda(t) = 0\} \\ &= (\lambda t)^0 \cdot \frac{e^{-\lambda t}}{0!} \\ &= e^{-\lambda t} \end{aligned} \tag{5.11b}$$

where parameter $\lambda > 0$ is the mean rate of occurrences of events and $t>0$. So the probability $P\{\tau \le t\}$ is expressed by

$$\begin{aligned} P\{\tau \le t\} &= 1 - P\{\tau > t\} \\ &= 1 - e^{-\lambda t} \end{aligned} \tag{5.11c}$$

We use time scale ξ instead of t, and in logarithmic domain we have

$$\ln P\{\tau \le \xi\} = \ln (1 - e^{-\lambda \xi}) \tag{5.12}$$

In the logarithmic domain there is a non-linear relationship between probability P and time interval of length ξ in Eq. 5.12.

It is believed that the straight segments appearing in the family of curves in a wide time range are significantly different from the Poisson process. Figure 5.7 shows those differences between the family of curves and Poisson process for the observations at Yichan, Yangtze River, China. The differences appear large on most thresholds except for a few of the highest thresholds.

The proposed family of curves $\ln P_\xi$ - $\ln \xi$ - Qs may serve as a good description of inter-scale structure of peak flows. The difference between the observed distribution of peak flows and the Poisson model indicates that the proposed curves are a better model of occurrence of point events.

5.4.2 A Group of Break Points

By looking at a family of curves (Figs. 5.5 and 5.6), the group of break points on most curves appear regular and suggest some physical meaning.

All points on the family of curves are related to a time interval or to a special duration on the time axis. With this notation, it is not surprising that a group of break

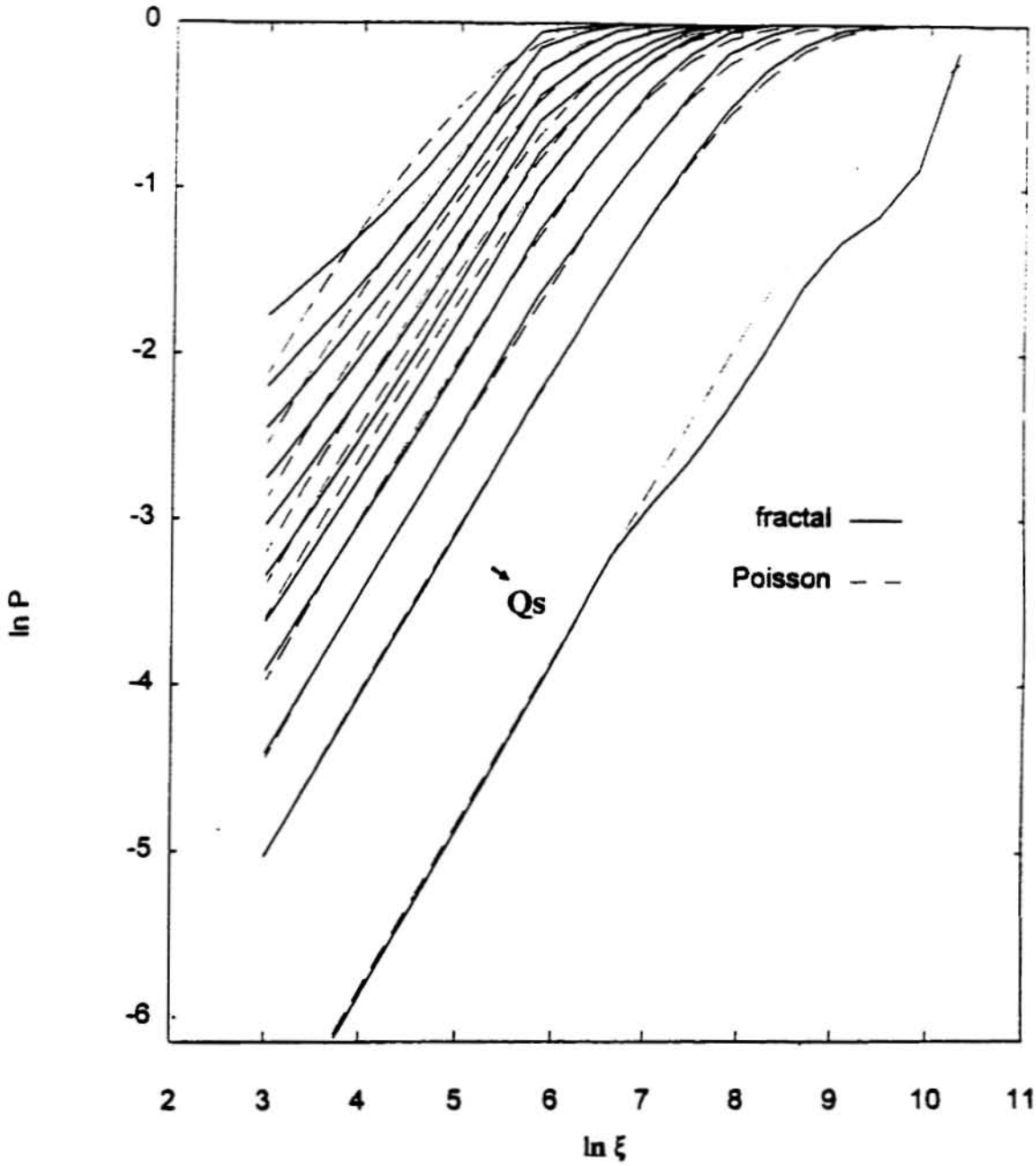

Figure 5.7 A family of curves ln P_ξ - ln ξ - Qs and Poisson process for the observations at Yichan, Yangtze River, China.

points indicate a special duration in which at least one peak flow occurs. One of the most interesting group of break points is the first group which is located from the right of the origin at 5.9 in the ln ξ scale corresponding to about one year as shown in Figs. 5.5 and 5.6a,b,c, d, e, f and g. This indicates a period of one year even though the occurrences of peak flows of small and medium magnitudes have various probabilities within this one year cycle.

Apparently, the break points are inherent in the nature of the data rather than the technique that was applied to generate the curves.

Figures 5.8a and b show the different techniques in choosing the time intervals and thresholds when drawing curves. They are insensitive to the appearance of the first group of break points, for example, for the observations at Hope, Fraser River, B.C., Canada.

Conversely, if synthetic data were used to perform the same procedure, their results will be significantly different.

Here two types of data are generated:

Type 1:

The data are independently and normally distributed. The mean value and standard deviation across time axis are estimated from those observed at Hope, Fraser River, B.C., Canada.

Type 2:

The data generated can adopt any hydrologic stochastic models with seasonal

components such as ARMA(Box and Jenkins, 1970) and Cannonical Expansion model (Spolia and Chander, 1977). The simplified stochastic processes are generated by considering the daily characteristics:

Let Q_{ij} be flow discharge,

$$\begin{matrix} Q_{1,1}\ Q_{1,2} \cdots Q_{1,j} \cdots Q_{1,n} \\ Q_{2,1}\ Q_{2,2} \cdots Q_{2,j} \cdots Q_{2,n} \\ \vdots \\ Q_{i,1}\ Q_{i,2} \cdots Q_{i,j} \cdots Q_{i,n} \\ \vdots \\ Q_{m,1}\ Q_{m,2} \cdots Q_{m,j} \cdots Q_{m,n} \end{matrix} \qquad (5.13)$$

and their mean values and standard deviations each ensemble are

$$\overline{Q_j} = \frac{1}{m}\sum_{i=1}^{m} Q_{i,j}$$

$$S_j = \sqrt{\sum_{i=1}^{m} (Q_{i,j}) - \overline{Q_j})^2 / (m-1)} \qquad (5.14)$$

where the all parameters, such as sample length, n, and the number of realisations, m, mean values $\overline{Q}_j$ and standard deviations, S_j across the ensembles of series, are estimated from the observations at Hope, Fraser River, B.C., Canada.

We are interested in the peak flows over a given threshold, so generated flow discharges, Y_{ij}, are simply obtained by

$$Y_{i,j} = \overline{Q}_j + S_j \times \zeta_{i,j} \tag{5.15}$$

where it is assumed that $\zeta_{i,j}$ is a normal random variable with zero mean and one standard deviation. Thus a simple time series of flows can be obtained.

The generated data Type 2 are closer to the observations while Type 1 is not.

The same procedure of the functional box counting performed on the generated data Types 1 and 2 to produce a family of curves is shown in Figs. 5.9 and 5.10, respectively.

From Figs. 5.9 and 5.10, the first group of break points matches the one year cycle in Fig. 5.10, however, nothing matches a one year cycle in Fig. 5.9.

5.4.3 Saturation Points

Most curves arrive at a horizontal line, i.e. P = 1, which implies a saturation point. Saturation points represent a special time interval in which "yes" intervals for a given threshold Qs occur with probability 1. It is reasonable to regard this point as a kind of

upper limit of empirical or observed return periods. A straight line intersects the horizontal line at a point, in which P=1 can be considered as an upper limit of empirical return period for the corresponding threshold, Qs.

5.5 Engineering Consideration

A family of curves $\ln P_\xi$ - $\ln \xi$ - Qs describes the temporal structure of peak flows and is a parametric set of curves with parameter Qs, $\ln P_\xi$ - $\ln \xi$ which can be transformed into the form

$$\ln P_\xi \sim f(\ln \xi, Q_s) \tag{5.16}$$

In fact, Eq.5.16 concerns two important aspects in engineering hydrology.

5.5.1 Empirical Plotting Positions

For a fixed time scale, say $\xi=365$, the probability, P_ξ, serves as an estimate of exceedence probability corresponding to the threshold Qs. If the annual maximum flows Qs_i, i=1, 2, ..., m, are used as thresholds, a set of the empirical plotting positions can be determined. This is a set of empirical plotting positions which is related to the scaling properties of peak flows and incorporated in flood risk analysis. The details of dealing with this topic will be described in Chapter 6.

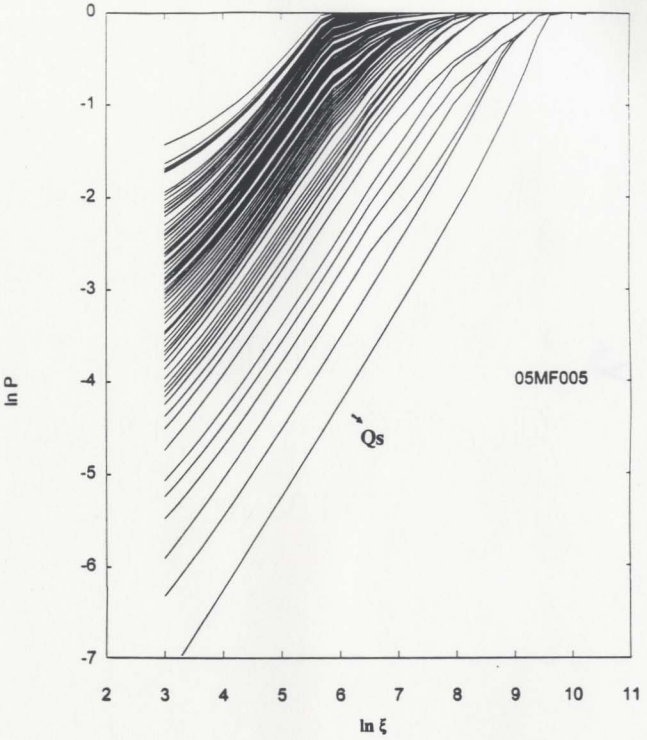

Figure 5.8a A family of curves with special time intervals for the observations at Hope, Fraser River, B.C., Canada.

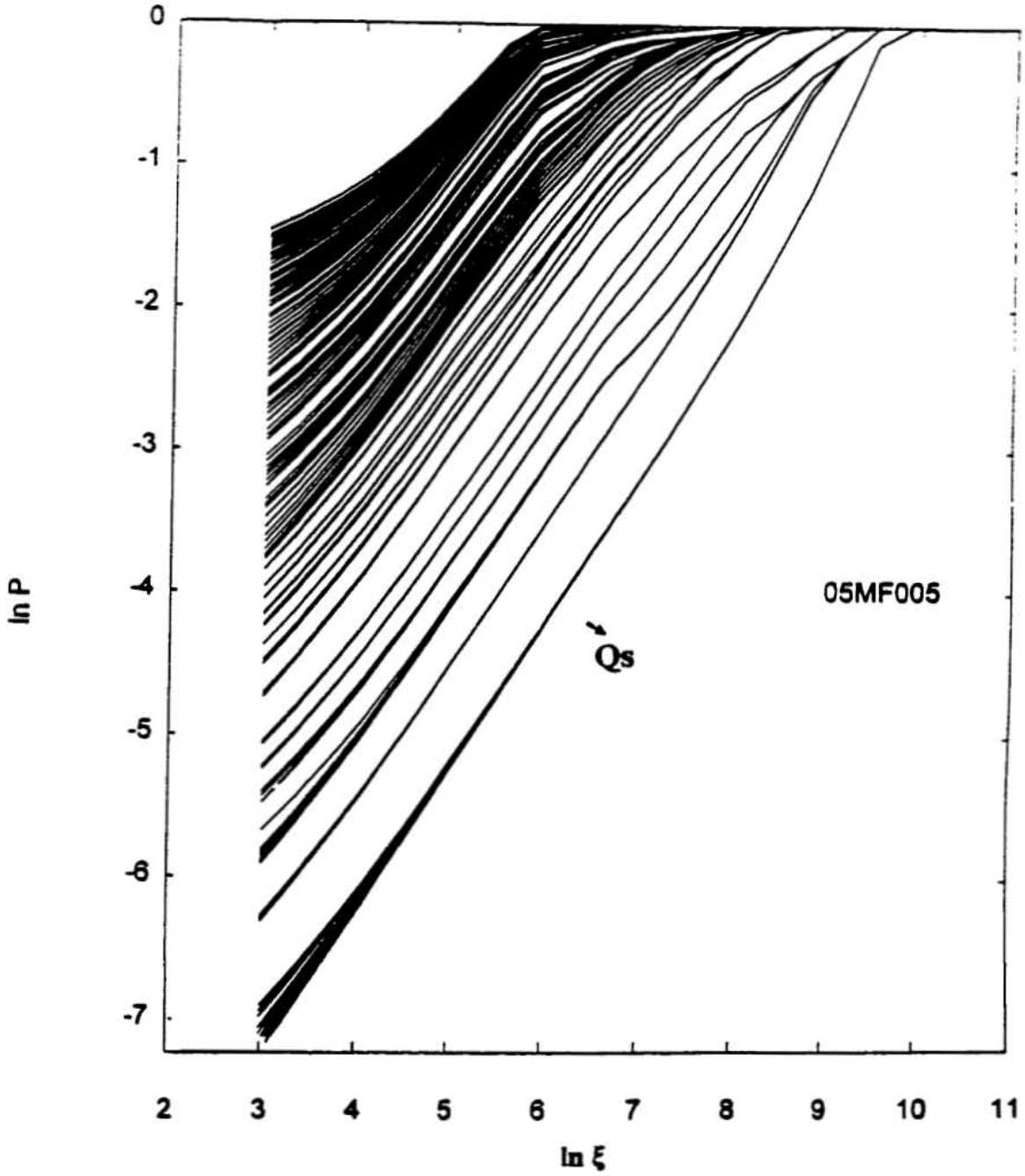

Figure 5.8b A family of curves ln P_ξ - ln ξ - Qs with a special thresholds for the observations at Hope, Fraser River, B.C., Canada.

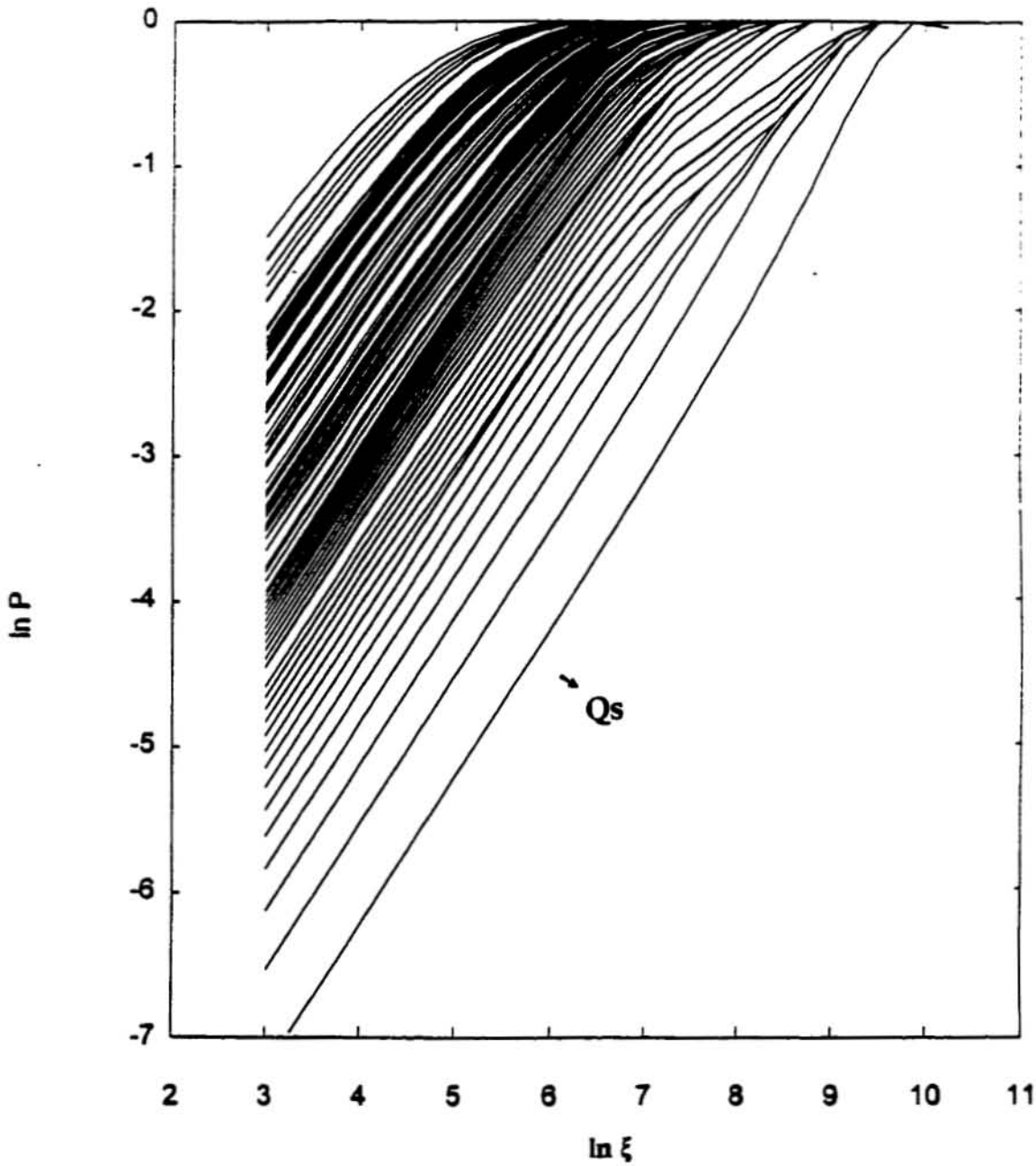

Figure 5.9 A family of curves ln P_ξ - ln ξ - Qs for Type 1 generated data shown in section 5.4.3.

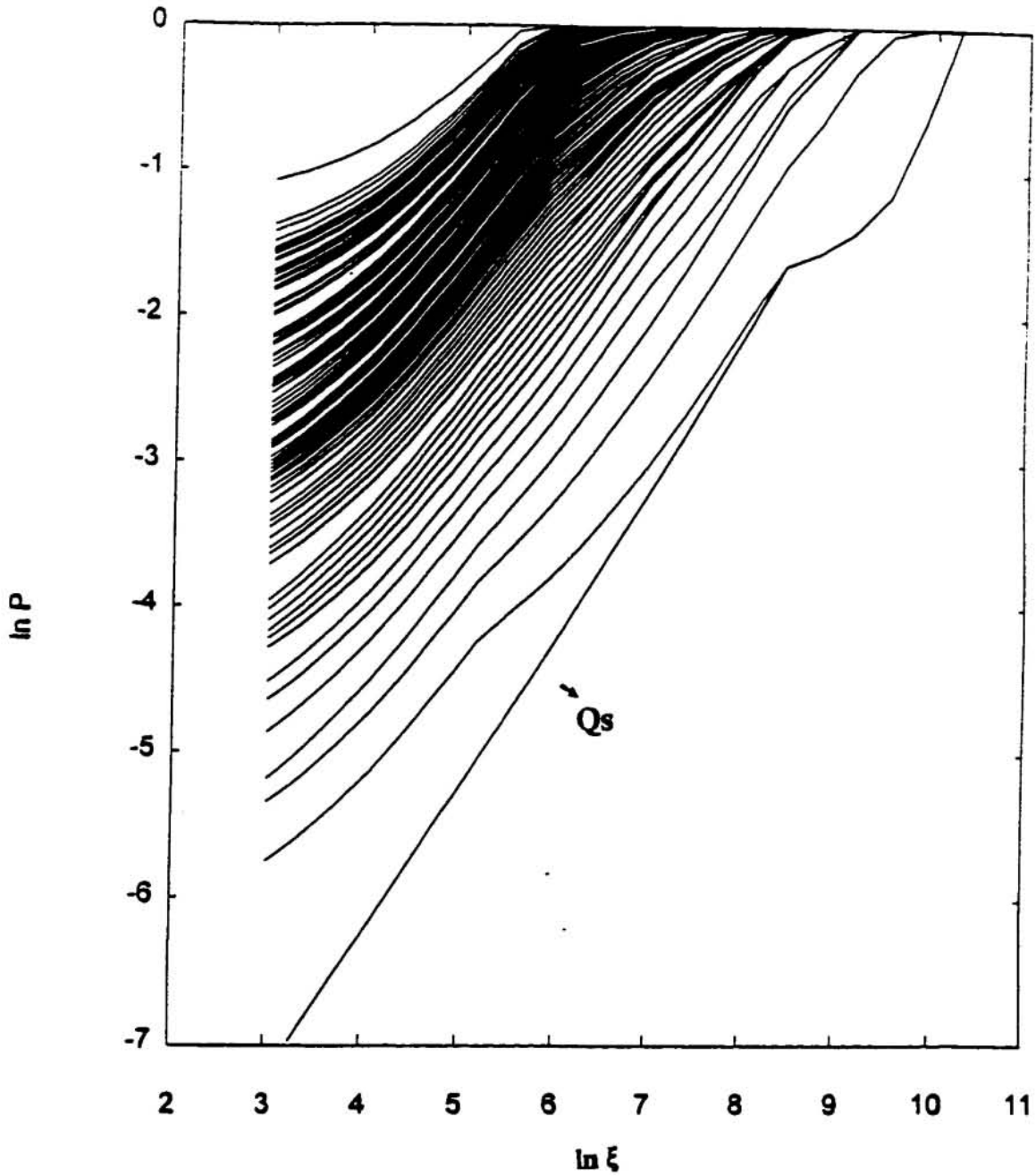

Figure 5.10 A family of curves ln P_ξ - ln ξ - Qs for Type 2 generated data shown in section 5.4.3.

5.5.2 Risk of Failure

What is the probability of a flood exceedence during a design life L_D ? In engineering hydrology, this question is usually answered by using independent processes such as the Binomial or Poisson (Linsley, 1958, 1975; Chow, 1964). However, this approach does not match well with observations as shown in Fig.5.7.

It is understandable that if the time scale, ξ , in a family of curves $\ln P_\xi$ - $\ln \xi$ - Qs is considered as a design life L_D, the probability, P_ξ, of a time interval including at least one occurrence of exceedence, is, in fact, a risk of failure during the design life, denoted as R. Thus, Eq.5.16 can be rewritten as

$$\ln R \sim f\ (\ln L_D, Q_s) \tag{5.17}$$

Furthermore, suppose the threshold Qs is given and defined by a design flood Q_T, the corresponding design return period, T_D, can be determined. Hence, Eq.5.17 can be expressed as

$$\ln R \sim f_1(\ln L_D, T_D) \tag{5.18a}$$

or

$$T_D \sim f_2\ (\ln L_D, \ln R) \tag{5.18b}$$

Hence, the family of curves $\ln P_\xi$ - $\ln \xi$ - Qs can serve as an empirical relation of ln R-

ln L_D - T_D , which is different from the Poisson explanation. When design flood Q_T is fixed, the power law, R- L_D^c, exists, where c is a constant. Figure 5.11 shows the relation ln R- ln L_D - T_D observed at Yichan, on Yangtze River, China, , where T_1, T_2 and T_3 are return periods.

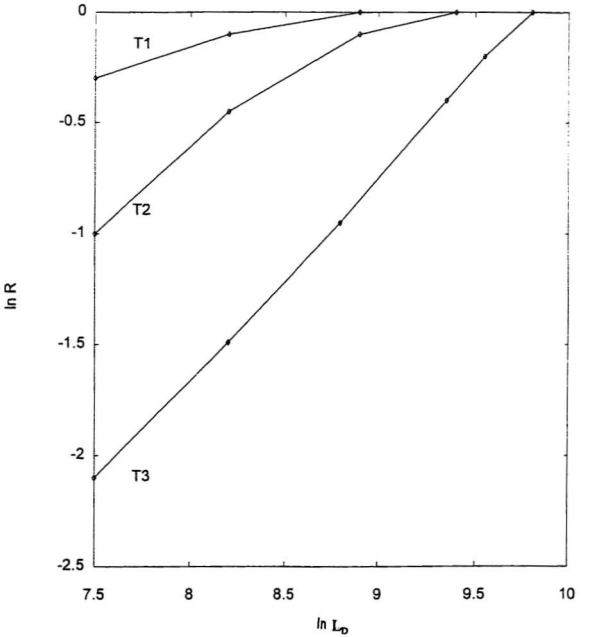

Figure 5.11 The representation of ln R- ln L_D - T_D observed at Yichan, on Yangtze River, China.

5.6 Summary

The variability of peak flow points in the time axis is completely disordered and related to the safety of hydraulic structures. Classical methods do not adequately describe this behaviour. From a macro scale point of view, in fact, peak flows evolution in time certainly have their own special behaviour for a given watershed.

This chapter focuses on studies of peak flow series across scales. A method for transforming observed point processes to a family of curves $\ln P_\xi$ - $\ln \xi$ - Qs was proposed to present the basic characteristics of peak flows along the time axis. The family of curves $\ln P_\xi$ - $\ln \xi$ - Qs thus contains all information about flood temporal characteristics and presents the inter-scale correlation structure for a given watershed. It describes the relationships between various time scales and probabilities of peak flow occurrences. It also explores scaling behaviour and shows some natural behaviour, such as the natural cycles inherent in the peak flow series which cannot be adequately described by classical methods at scale n or at scale one.

The proposed family of curves $\ln P_\xi$ - $\ln \xi$ - Qs is also appropriate for hydrological practice, it meets two important aspects of flood risk analysis, which are discussed in the next Chapter.

Chapter 6

A Scaling Plotting Position for Flood Risk Analysis

6.1 General

Analysis based on fractal geometry in Chapter 5 indicates that peak flow points varying with time intervals follows a power law over wide ranges of time scales, thus a family of curves ln P_ξ-ln ξ-Qs, which transforms observations into a graphic explanation, was developed.

Straight lines in the family of curves indicate observations being invariance across scales, in other words, a correlation structure of peak flows exists. This correlation structure, however, includes the feature of long- term persistence.

According to Hurst's finding, Mandelbrot (1977, 1982) pointed out that an account for the Hurst phenomenon was a symptom of scaling, so "scaling noise" had been defined intuitively. Since the term "scaling" indicates a kind of order, scaling noise describes an order hidden in a fluctuation and can be expressed as a scaling Gaussian random process. Hence, fractional Brownian motion, in which the box dimension is determined by 2-H (Mandelbrot, 1982), plays an important role in describing and modelling the natural feature of observations. The terms "scaling" or "scale-invariance" provide a

broader meaning and a rational basis for describing and exploring the natural behaviour of peak flows.

A family of curves well describes the scaling behaviour of peak flows over broad ranges of time scale, thus the family of curves becomes useful if these long-term behaviours are considered in flood risk analysis.

As we know, the main purpose of flood frequency analysis is to estimate flood magnitudes according to specified probability and the confidence interval of the event associated with selected probability level. Hydrologic frequency analysis statistically finds an optimal quantile and its sampling distribution by means of observed sample. Probability plotting positions which are widely used in flood frequency analysis, are used for the graphical display of observed floods, and serve as estimates of the probability of exceedence of those values (Guo, 1990). However, proposed family of curves $\ln P_\xi$-$\ln \xi$-Qs well presents the scaling behaviour of peak flows, and are directly related to the probabilities of exceedences of those observed data. As was mentioned in section 5.5.1, for a fixed time interval, $\xi=1$ year, the probability, P_ξ, could serve as an estimate of exceedence probability corresponding to the threshold Qs. This gives empirical plotting positions, which are based on the measurement of peak flows across scale.

In this chapter, an empirical probability plotting position formula is proposed based on scaling behaviour of peak flows. Extensive Monte Carlo experiments are carried out to assess the properties of the proposed plotting position formula and existing standard

approaches for flood frequency analysis. Also, practical applications of the proposed method are demonstrated using a number of peak flow observations from Canadian and Chinese rivers.

6.2 Standard Flood Risk Analysis

Flood frequency analysis is concerned with estimating a design flood for a given return period and estimating the probability of exceedence of a given flood within a given time interval. There are three main uncertainties in flood risk analysis (Wood and Rodriquez-Ituzbe, 1975): i.e., natural uncertainty, parameter uncertainty, and model uncertainty. Because the occurrence of floods is a complex process, our limited understanding of the process results in increased model uncertainty. Therefore, to decrease the model uncertainties, it is important to understand the nature of floods.

Two hydrologic data series are commonly used in flood risk analysis: the peaks over threshold series (POT) (eg. NERC, 1975) and the annual maximum series (AM). A POT series is a series of data that takes all the peaks over a selected level, so the number of exceedences in the series is greater that the number of years of the record. An AM series includes the largest values occurring in each of the equally long time intervals of the record. The time interval length is usually taken as one year. The number of exceedences is equal to the number of years of the record (Chow, 1964; Chow et al., 1988). The words "peak flows" in Chapter 5 and "annual peak flows" in Chapter 2 and 3 correspond to the

POT series and AM series, respectively. It is assumed that annual peak flows in AM series are independent of each other. However, there is a greater probability that the peak flows in a POT series are related and less independent.

Based on the assumption that observed annual peak flows are identically and independently distributed (iid), standard flood risk analysis assumes that the annual maximum flows are randomly sampled from an assumed parent probability distribution, $F(\mathbf{Q}/\theta)$, where θ is a set of parameters estimated from the sample and $\mathbf{Q}$ is a flood discharge, a random variable.

Therefore, two aspects of flood risk estimation to be considered are: choice of parent probability function F; and choice of parameter estimation method to estimate θ.

In reality, since nature's distribution is unpredictable, reasonable 'flood-like' distributions have been recommended. For example, the log-Person Type III distribution was recommended by the U. S. Water Resources Council in 1967 for use by U. S. Federal Agencies (Benson, 1968), and the GEV distribution was recommended for use in Britain by NERC (1975).

According to the suggestion from WMO (1989), in which fourteen candidate distributions were recommended to be used with annual maximum, the lognormal distribution (LN) (Kaczmarek, 1957; Stedinger, 1980) and the Pearson Type III distribution (PIII) (Matalas and Wallis, 1973, Bobee, 1973; Hua, 1985) are listed in the first two positions by WMO, these are hence considered in this study. The probability

density functions of Q are expressed as

$$f(q) = \frac{\lambda^{\beta}(q - q_0)^{\beta-1} e^{-\lambda(q-q_0)}}{\Gamma(\beta)} \qquad q \geq q_o \tag{6.1}$$

for Pearson Type III (Foster, 1924) and

$$f(q) = \frac{1}{q\,\sigma_y\sqrt{2\pi}} \exp\left(-\frac{(y - \mu_y)^2}{2\sigma_y^2}\right) \qquad \log q \geq 0 \tag{6.2}$$

$$where \quad y = \log q$$

for lognormal distribution (Chow, 1954), where the parameters a_0, β and α in the Pearson Type III and parameters μ and σ in the lognormal distribution are estimated from observed samples.

The quantile Q_T of floods for a given return period can be expressed as (Chow, 1951)

$$Q_T = \mu + \sigma K_{T_{P\text{-}III}} \tag{6.3}$$

or Pearson Type III and

$$Q_T = \mu_q + \sigma_q K_T \tag{6.4}$$

for lognormal distribution, where K_T and $K_{TP\text{-}III}$ are frequency factors for lognormal and

Pearson Type III respectively, μ and σ in Eq. 6.3 and μ_q and σ_q in Eq.6.4 are population mean and standard deviation estimated from observed samples.

Various methods for estimating the parameters are commonly used in flood risk analysis:

- the maximum likelihood method (ML) which gives unbiased and minimum variance estimates;
- the method of moments (MOM) which is biased and performs poorly for some models (Matalas and Wallis, 1973) but they are efficient for some distributions (Lowery and Nash, 1970) such as Pearson Type III (Song and Ding, 1988; Wu et al., 1991);
- the method of probability weighted moments (PWM)(Greenwood et al., 1979) which is a linear combination of order statistics and is unbiased for small samples; and
- L-moments method (Hosking, 1990) which is a linear combination of PWM's but with a clearer statistical interpretation, and
- graphical curve fitting methods

Among the available methods of parameter estimation, graphical approach (Dalrymple, 1960; Chow, 1964; Chow et al., 1988) which consist of fitting a function visually to the data is favoured by many hydrologists and engineers. It has been widely used both in hydraulic engineering and hydrologic practice.

Graphical estimation yields quantile estimation directly rather than estimates of the individual parameters. The steps are:

(1) Plot data on a probability graph paper;

Rank observed data from the largest to the smallest value

$$q_1 \geq q_2 \cdots \geq q_m \cdots \geq q_n \tag{6.5}$$

where n and m are the total number of values to be plotted and the rank of a value, respectively. Calculate the plotting positions from a selected plotting position formula

$$p_1 \leq p_2 \cdots \leq p_m \cdots \leq p_n \tag{6.6}$$

and then a pair of data is plotted on a specially designed probability paper;

(2) Curve fitting;

Once the data have been plotted on probability paper, an eye-guided line or a curve is drawn through the plotted points;

(3) Estimate quantiles.

Estimated quantiles for various return periods are selected from the 'best-fit' line.

The key step is to determine a plotting position that estimates the probability of future floods. The choice of plotting position formula for use on probability graph paper has been discussed by many authors (Hazen, 1914; Weibull, 1939; Gumbel, 1943, 1947; Blom, 1958; Tukey, 1962; Gringorten, 1963; Cunnane, 1978; Guo, 1990).

6.3 Plotting Position Formulas

A probability plot is a plot of a magnitude of flood versus a probability. Most plotting position formulas expressed as probabilities, are special cases of the general form:

$$P_m = \frac{(m-\alpha)}{(n+1-2\alpha)} \tag{6.7}$$

where P_m is the plotting probability of the m^{th} largest value, n is the sample size and α is a constant. For the Weibull (1939) formula $\alpha=0$, for the Cunnane (1978) formula $\alpha=0.4$, for the Gringorten (1963) formula $\alpha=0.44$, and for the Chegodayev (1955) formula $\alpha=0.3$.

Plotting position formulas are usually associated with theoretical order statistics (see Appendix B). For samples of intermediate size, the expected value of order statistics is dependent on a corresponding quantile with a linear relationship (Harter, 1971). Using this linear relationship, the order statistics of a sample can be used to estimate the sample quantiles.

Let $Q_1, Q_2, \ldots, Q_n$ be a simple random sample from a population with probability density function (pdf), f(q), and cumulative distribution function (cdf), F(q). $Q_1, Q_2, \ldots, Q_n$ are assumed to be statistically independent and identically distributed. When this random sample is ranked as $Q_{(1)} \geq Q_{(2)} \geq \ldots \geq Q_{(m)} \geq \ldots \geq Q_{(n)}$, the m^{th} - order statistic $Q_{(m)}$

which is a random variable, has a pdf, $g_m(q)$, given by

$$g_m(q) = m\binom{n}{m} f(q)\,[1 - F(q)]^{m-1}\,[F(q)]^{n-m} \tag{6.8a}$$

and the cdf, $G_m(q)$,

$$G_m(q) = m\binom{n}{m}\int_{-\infty}^{q}[1 - F(q)]^{m-1}[F(q)]^{n-m} f(q)dq \tag{6.8b}$$

Let $Y_{(m)}=1-F(Q_{(m)})$, where $0 \le Y_{(m)} \le 1$, because $0 \le F(Q_{(m)}) \le 1$. The probability density function of the function $Y_{(m)}$ of the random variable $F(Q_{(m)})$, be $h_m(y)$

$$h_m(y) = m\binom{n}{m} y^{m-1}(1-y)^{n-m} \tag{6.9a}$$

or

$$h_m(y) = [y^{m-1}(1-y)^{n-m}]/[B(m, n-m+1)] \tag{6.9b}$$

where B is a beta function with two parameters m and (n-m+1).

Plotting position formulas based on distribution of $Q_{(m)}$ in Eq. 6.8 should take probability density function of parent population into consideration. Formulas associated with the distribution of $Y_{(m)}$ should use Eq. 6.9 with free-distribution of parent Q. Most of the well-known plotting position formulas are measures of central tendency of the

distributions of either $Y_{(m)}$ or $Q_{(m)}$, thus Eqs. 6.8 and 6.9 are the bases of theoretical plotting position formulas.

However, plotting position formulas are classified into three groups (Ji et al., 1984): Group I, including Weibul (1939), Chegodayev (1955) and Cunnane formulas (1978), is associated with distribution of $Y_{(m)}$ which is distribution free in using Eq. 6.9. Group II, including Gringorten(1963), Weibul (1939), Chegodayev (1955) and Gumbel (1943) formulas, is related to the distribution of $Q_{(m)}$ in Eq. 6.8. Group III is based on the empirical distribution function, it contains the formulas such as Hazen (1914) and California (C.S.D.P.W., 1923). Weibul and Chegodayev formulas can be derived from both Eqs. 6.8 and 6.9 (see Appendix B), so they could be belonged to Groups I and II.

Overall, Eqs. 6.8 and 6.9 as bases of theoretical plotting position formulas, are derived under the condition that $Q_1, Q_2, \ldots, Q_n$ are n independent variables with the same pdf f(q). In other words, the term "short-term independence" discussed in previous Chapters is one of the prerequisite conditions for the theoretical plotting position formulas (see Appendix B).

According to order statistical theory, even through some suggestions of relaxing assumptions and considering nonidentically distributed $Q_1, Q_2, \ldots, Q_n$ as well as various patterns of dependence have been made (David, 1981), statistical long-term properties have not been discussed yet in studies of order statistics.

6.4 Scaling Plotting Positions

6.4.1 Basic Concepts

The basis of the plotting position formulas discussed above suggests that the statistical average of order statistics, such as mean (Weibull formula), median (Gringorten formula) or mode (Chegodayev formula), is linearly varying with corresponding quantiles. Using this linear relationship estimated quantiles are available. However, plotting position formulas involved in Groups I and II are theoretically derived from the probability function of $Y_{(m)}$ or order statistics $Q_{(m)}$. Parent random variable Q_i being independent is the basic assumption for these derivations.

Recall Chapters 2 and 3 dealing with long- and short-term behavior of peak flow series. Studies indicate that long-term dependence is not uncommon phenomena in peak flow series as well as in other independent data. Because of long-term persistence, the variation of parameters and quantiles increases in flood risk analysis, hence, the impacts of this property on design flood estimation cannot be ignored.

Recall Chapters 4 and 5, proposed family of curves well present correlation structure of peak flows based on the measurement at scale. Straight lines over broad ranges in the family of curves indicate existence of scaling feature of peak flows. This scaling behavior is related to the long-term persistence of peak flow series.

The thresholds, Qs, in the family of curves, in fact, are ranked as $Qs_{(1)} \geq Qs_{(2)} \geq \ldots \geq Qs_{(m)} \geq \ldots \geq Qs_{(n)}$ from the right to the left in the family of curves. Furthermore, for

a given time scale, ξ, say $\xi=1$ year or 365 days, probabilities of exceedences, P_ξ, are ranked as $P_{\xi,(1)} \le P_{\xi,(2)} \le \ldots \le P_{\xi,(m)} \le \ldots \le P_{\xi,(n)}$ from the bottom to the top at this fixed time scale. Thus, the ranked $Qs_{(m)}$ and $P_{\xi,(m)}$ could serve as order statistics and corresponding empirical exceedence probability, respectively.

Consider the concepts described previouslly, plotting position formulas are related to the order statistics and corresponding exceedence probabilities (Hirsch, 1987; Hirsch and Stedinger, 1987), and serve as an estimate of the probability of exceedence for observations (Guo, 1990). Theoretical plotting positions in Group I are the expectation of exceedence probability function, $E(Y_{(m)})$ in Eqs.B.9 and B.10 of Appendix B.

If AM series are considered as a set of thresholds ranked in the family of curves, P_ξ are exccedence probabilities for the corresponding thresholds. A statistical model to infer expected value of exceedences is realisable.

AM series are independent and identical distributed within short-term scale. On the other hand, they are possibly correlated over wider scales according to the calculated results of $P(K \ge k_0)$ in Chapter 3, where $P(K \ge k_0)$ is the probability of a peak flow series exhibiting long-term persistence. A family of curves well demonstrates peak flow statistical characteristics including long- term persistence, it should be a basis dealing with plotting position formula which takes long- and short-term behaviors into account.

Thus, an empirical plotting position formula is developed, which serve as an estimate of expectation of exccedence probability, and takes scaling behaviour of peak flows into

account.

Since this plotting position formula is based on the family of curves and related to the scaling behavior of peak flows, it is called from now on a scaling plotting position formula or SPP. The following are the ideas behind the scaling plotting position formula:

1) Let empirical probability of exceedences of flood discharges, P_ξ, be a random variable. Its magnitude is related to other variables, such as the level of exceedence expressed as a threshold Qs ranked in the family of curves, and the time interval of occurrence of exceedence, ξ.
2) The expectation of random variable, P_ξ, can be estimated by a statistical model, thus the problem associated with a statistical model can be expressed as a relationship between the random variable, P_ξ, and the variables, the level of exceedence, Qs, and time interval, ξ.
3) A linear regression model in log domain is assumed to connect the relationship between $\ln P_\xi$, $\ln \xi$ and Qs.
4) The observations of variables, P_ξ , $\ln \xi$ and Qs, are selected from a family of curves which presents scaling behaviour of peak flows.
5) Once the expectation of P_ξ is estimated, a scaling plotting position formula can be determined.

Principally, the family of curves is the basis for the statistical estimation of the proposed plotting positions. Since a family of curves well describes scaling behaviour

including long-term behaviour of peak flows, the plotting position formula based on the family of curves should be more accurate than that of classic plotting position formulas.

Based on the above considerations, the scaling plotting position formula is now developed.

6.4.2 Scaling Plotting Position Formula

The key of the proposed scaling approach plotting position formula is to construct a probabilistic model to infer expectation of probability of exceedences.

6.4.2.1 A Linear Statistical Model

Let a family of curves ln P_ξ-ln ξ be transformed into a mathematical relationship. Assume that the probability of occurrence that can be equivalently expressed as the ordinate in a family of curves, P_ξ, is related to the exceedence level expressed as a threshold, Qs, and the time interval, ξ, in a logarithm domain by:

$$\ln P_\xi = \beta_0 + \beta_1 Q_s + \beta_2 \ln \xi + \ldots + \beta_k Q_s \ln \xi + \eta \qquad (6.10)$$

where P_ξ is a random variable having a mean that is a function of non-random variables, Qs and ξ, and β_0, β_1, β_2 ... β_k are (k+1) unknown parameters, η is a random variable which is normally distributed, $\eta \sim N(0, \sigma)$.

Let $Z = \ln P_\xi$, $X1 = \ln \xi$, $X2 = Q_s$, ..., $Zk = Q_s \ln \xi$, thus a general linear model of the form is

$$\begin{gathered} Z = \beta_0 + \beta_1 X1 + \beta_2 X2 + \ldots + \beta_k Xk + \eta \\ E(\eta) = 0 \\ Var(\eta) = \sigma^2 \\ Cov(\eta_i, \eta_j) = 0, i \neq j \end{gathered} \tag{6.11}$$

where Z is a dependent variable, X1,X2, ... Xk are independent variables in the mathematical sense. The expectation of random variable Z is

$$E(Z) = \beta_0 + \beta_1 X1 + \beta_2 X2 + \ldots + \beta_k Xk \tag{6.12}$$

Thus E(Z) is a linear function of β_0, β_1 ... β_k and represents a plane in the Z, X1, X2, ... Xk space. The unknown parameters β_0, β_1, ... β_k can be estimated by the least squares method. However, there are n-straight lines in a wide range of the family of curves $\ln P_\xi$-$\ln \xi$-Qs, hence n- z_1, z_2 ... z_n on Z and n×k independent observations are available from the family of curves, and the estimate of the expectation of Z from the linear regression equation

$$\hat{Z} = b_0 + b_1 X1 + b_2 X2 + \ldots b_k Xk \tag{6.13}$$

or

$$\hat{P}_\xi = \exp(b_0 + b_1 \ln \xi + b_2 Q_x + \ldots + b_k Q_x \ln \xi) \qquad (6.14)$$

is conveniently estimated by the least squares method, where b_0, b_1, b_2 ...b_k are the estimates of β_0, β_1, β_2 ... β_k.

The individual terms in Eq.6.10 might be numerous, and in fact, a stepwise regression analysis (Klecka, 1980) and corresponding hypotheses testing are carried out. At the end, in a stepwise procedure, the significant terms will be selected for a given selection criteria.

Once parameters b_0, b_1, b_2 ... b_k are estimated, the Eqs. 6.13 or 6.14 can be obtained. Let ξ be 365 days, an empirical plotting position formula can be determined for given Qs.

6.4.2.2 About Logarithmic Transformation

The estimate of expectation of Z in ln-space we deal with is unbiased. This unbiased estimate is useful for investigation of the statistical properties of scaling behaviour of flows. However, P_ξ in Eq. 6.14 can be considered as a biased estimate and does not have the minimum expected error variance, thus correction of bias (Miller, 1984; McCuen and Snyder, 1986; Koch and Smillie, 1986; McCuen et al., 1990) for plotting position or empirical reduction of error needs to be achieved. The method for correcting the bias

suggested by Miller(1984) could used to reduce the bias due to the logarithmic transformation.

6.4.2.3 Steps in SPP Method

Based on the above consideration, the steps of the proposed method to estimate flood quantiles are as follows:

1) Assume peak flow series Q_j, $j=1, 2, \ldots, s$, is a n-year POT series and Qs_i, $i=1, 2, \ldots, n$, denotes n-year AM flood series where $s>n$.
2) The functional box counting procedure is carried out to construct a family of curves $\ln P_\xi$-$\ln \xi$-Qs, where annual maximum floods, Qs_i, $i=1,2, \ldots,n$, are the thresholds.
3) Based on the family of $\ln P_\xi$-$\ln \xi$-Qs curves, necessary information is obtained and parameters in Eq. 6.13 or 6.14 are estimated by a stepwise regression analysis.
4) Let time scale be equal to 365 days in the estimated Eq. 6.13 or 6.14, empirical plotting positions P_ξ are calculated and classical graphical curve fitting procedure is used to fit a curve to the points, and flood quantiles will be estimated from the fitted curve.

6.5 Applications of SPP Method

In this study, the scaling approach plotting position formula has been used to estimate the flood quantiles in Canadian rivers (Environment Canada, 1992) and at Yichan guage, Yangtze River, China (Ministry of Water Resources, 1985) . The daily flow records of Canadian Rivers used are shown in Table 6.1.

Table 6.1 Characteristics of daily flows collected from Canadian rivers (Environment Canada, 1992)

WSC Number	Prov.	Gauging Station	A(km^2)	Years
05AJ001	Alta.	South Saskatchewan River at Medicine Hat	56,400	1913-1987
05DF001	Alta.	North Saskatchewan River at Edmonton	28,000	1912-1986
05HG001	Sask.	South Saskatchewan River at Saskatoon	41,000	1912-1986
05KJ001	Man.	Saskatchewan River at the Pas	347,000	1913-1987
08MF005	B.C.	Fraser River at Hope	217,000	1913-1987

6.5.1 Calculation of SPP

As an example, a stepwise regression procedure and corresponding statistical hypothesis tests for the gauge 08MF005 of Fraser River at Hope are illustrated in Table 6.2, in which an empirical linear regression equation taking the form Eq. 6.13, where the variables X_i, could be chosen subjectively, e.g. for X_i, i=1,2,3, ...5 are given by

Table 6.8 Comparison of SPP with other estimators for the lognormal distribution, where the units of BQ_p and SQ_p are m^3 /s but RQ_p is in dimensionless

P_1=.001 P_2=.005		Estimation Methods of Quantiles						
n	Variable	MOM	PWM	0.0	0.3	0.4	0.5	SPP
30	BQ_{p1}	72768	74463	79015	75482	75093	74828	105872
	BQ_{p2}	67426	68507	71595	69134	68859	68330	87216
	SQ_{p1}	7868	8872	11514	9799	9260	8754	12023
	SQ_{p2}	5609	6146	7463	6550	6255	5976	7877
	RQ_{p1}	10.81	11.92	14.57	12.82	12.27	11.76	11.36
	RQ_{p2}	8.32	8.97	10.42	9.39	9.05	76405	9.03
40	BQ_{p1}	71753	72932	75979	74164	73581	72827	102401
	BQ_{p2}	66676	67473	69616	68326	67901	67365	85044
	SQ_{p1}	6964	7542	9355	8105	7686	7200	9084
	SQ_{p2}	5050	5257	6235	5564	5335	5062	6011
	RQ_{p1}	9.71	10.34	12.31	10.93	10.45	9.89	8.87
	RQ_{p2}	7.57	7.79	8.96	8.14	7.86	7.51	7.07
50	BQ_{p1}	72410	73952	75933	74430	73794	73259	102194
	BQ_{p2}	67239	67857	69665	68581	68136	67741	84870
	SQ_{p1}	6217	6521	7894	6990	6564	6252	8097
	SQ_{p2}	4474	4630	5353	4836	4600	4427	5291
	RQ_{p1}	8.59	8.89	10.40	9.39	8.90	8.53	7.92
	RQ_{p2}	6.65	6.82	7.68	7.05	6.75	6.54	6.23

$$X1 = \ln \xi$$

$$X2 = Q_s \qquad X3 = Q_s \ln \xi$$

$$X4 = (\ln \xi)^{1/2} \qquad X5 = Q_s^{1/2}$$

The criteria for selecting variables to enter into the model and for remaining are:

- Maximum R^2-statistic, where R^2 is sample multiple coefficient of determination;
- C_L-statistic (Mallows, 1973) which is a measure of total squared error where L is the numbers of parameters in regression equation.

If C_L first approaches the number of parameters, L, the model is chosen and the parameter estimates are unbiased (Mallows, 1973; Daniel and Wood, 1980).

A similar procedure was carried out for other Canadian rivers as well as in Yichan, Yangtze River, China. The results of the stepwise regression are shown in Table 6.3. Figure 6.1a-e displays the scaling plotting position formula in which the values of P_ξ have been empirically corrected. In order to limit variation ranges of estimated plotting positions, the upper and lower limits are corrected to be 0.999 and 0.001, if estimated values are greater than 0.999 and less than 0.001, respectively.

Table 6.2 Stepwise procedure of selecting variables for dependent variable Z for the gauge 08MF005 of Fraser River at Hope, B.C., Canada

Step 1 Variable X4 Entered		
R-square = 0.5889347	C_L =21632.544711	F = 1931.28
Variable	Parameter Estimate	Standard Error
INTERCEP	-6.75415176	0.13602869
X4	2.26629944	0.05156964
Step 2 Variable X2 Entered		
R-square = 0.8021814	C_L =9714.0542746	F = 2731.14
Variable	Parameter Estimate	Standard Error
INTERCEPT	-3.99331376	0.11899806
X2	-0.00031371	0.0000082
X4	2.26629944	0.03578767
Step 3 Variable X1 Entered		
R-square = 0.8582471	C_L =6581.9815232	F = 2716.47
Variable	Parameter Estimate	Standard Error
INTERCEP	-16.95549607	0.57075470
X1	-2.00241690	0.08678590
X2	-0.00031371	0.00000697
X4	12.57402310	0.44776942
Step 4 Variable X3 Entered		
R-square = 0.9655799	C_L =584.08201348	F = 9432.76

Variable	Parameter Estimate	Standard Error
INTERCEP	-10.13613237	0.30041152
X1	-2.98252622	0.04537894
X2	-0.00108860	0.00001245
X3	0.00011137	0.00000172
X4	12.57404449	0.22072721

Step 5 Variable X5 Entered

R-square = 0.9759570	C_L = 6.00000000	F = 10911.2
Variable	Parameter Estimate	Standard Error
INTERCEP	-17.56273566	0.39770121
X1	-2.98252608	0.03794056
X2	-0.00191028	0.00003567
X3	0.00011137	0.00000144
X4	12.57404449	0.18454625
X5	0.15700390	0.00651877

Summary of Stepwise Procedure for Dependent Variable Z (significant level α =.15)

Step	Enter	Move	Partial R^2	Model R^2	C_L	F
1	X4		0.5889	0.5889	21632.545	1931.2847
2	X2		0.2132	0.8022	9714.0543	1452.0548
3	X1		0.0561	0.8582	6581.9815	532.3665
4	X3		0.1073	0.9656	584.0820	4194.1377
5	X5		0.0104	0.9760	6.0000	580.0820

Suppose $$Z = \beta_0 + \beta_1 X1 + \beta_2 X2 + \ldots + \beta_k Xk + \eta$$

Table 6.3 Scaling plotting position formulas

WSC Number	b_0	b_1	b_2	b_3	b_4	b_5
05AJ001						
Parameter	-15.8514	-2.2114	-0.0024	0.00024	11.9848	
Error	0.21826	0.03334	0.00002	0.000002	0.17113	
05DF001						
Parameter	-16.4195	-2.1825	-0.0025	0.0002	12.0223	0.0223
Error	0.29536	0.04409	0.00005	0.000003	0.22613	0.00325
05HG001						
Parameter	-16.9645	-2.4215	-0.0031	0.00027	12.7852	0.03348
Error	0.29707	0.04145	0.00003	0.000002	0.22151	0.00211
05KJ001						
Parameter	-12.9662	-2.1965	-0.0072	0.0004	9.3989	0.2484
Error	0.31228	0.31228	0.00011	0.00011	0.19171	0.00893
08MF005						
Parameter	-17.5627	-2.9825	-0.0019	0.00011	12.574	0.1570
Error	0.39770	0.03794	0.00003	0.000001	0.18454	0.00651
Yichan, Yangtze River, China						
Parameter	-34.6590	-3.2655	-0.0007	0.00002	13.6283	0.2172
Error	0.62528	0.03967	0.00001	0.00000	0.19146	0.00509

Scaling plotting position formula:

$$P_{\xi=365} = \exp(b_0 + b_1 \ln \xi + b_2 Q_s + b_3 Q_s \ln \xi + b_4 (\ln \xi)^{1/2} + b_5 Q_s^{1/2})$$

where $\xi = 365$, and error = standard error.

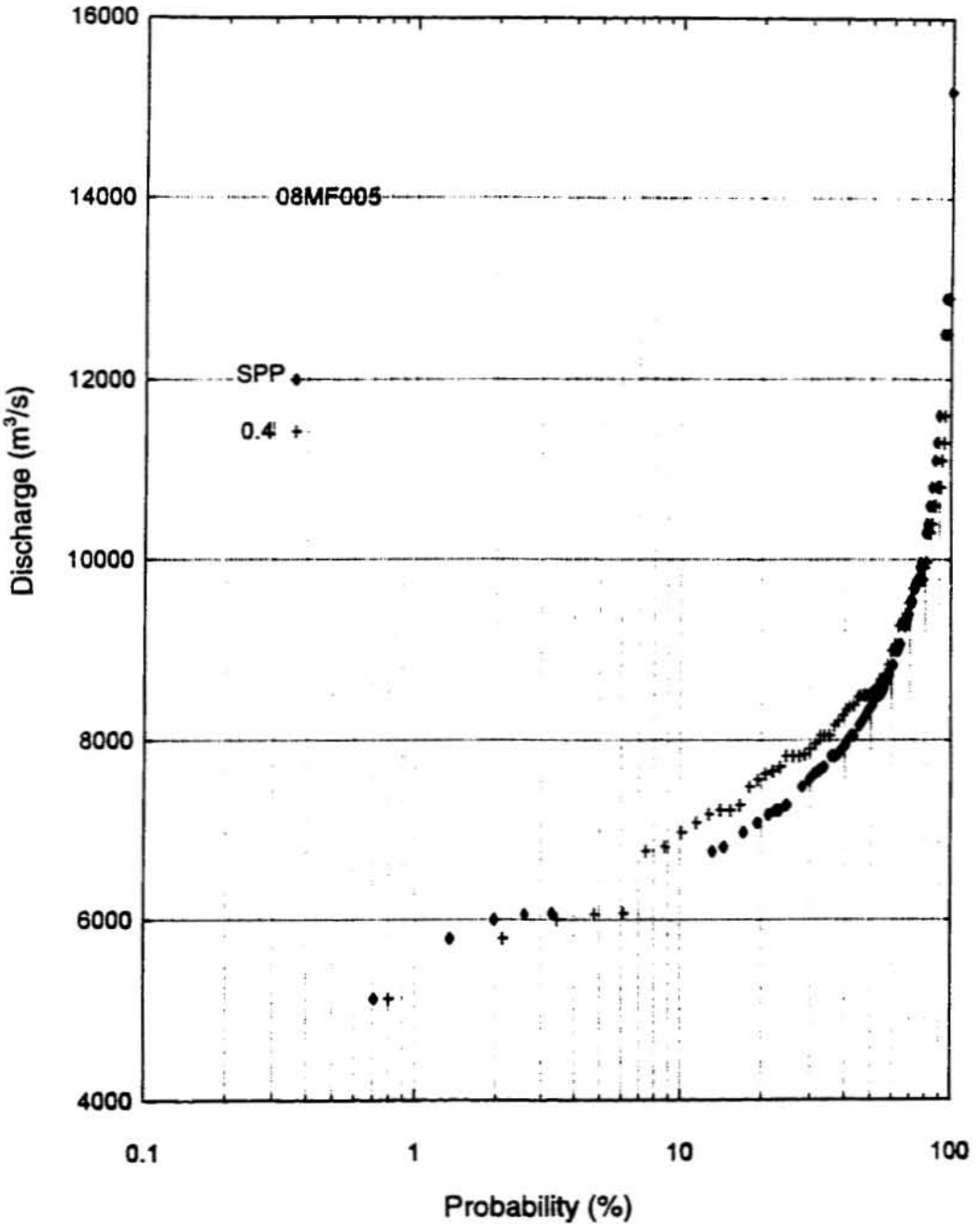

Figure 6.1a Plotting positions using SPP and Cunnane formula for AM series of Fraser River at Hope, B.C., Canada.

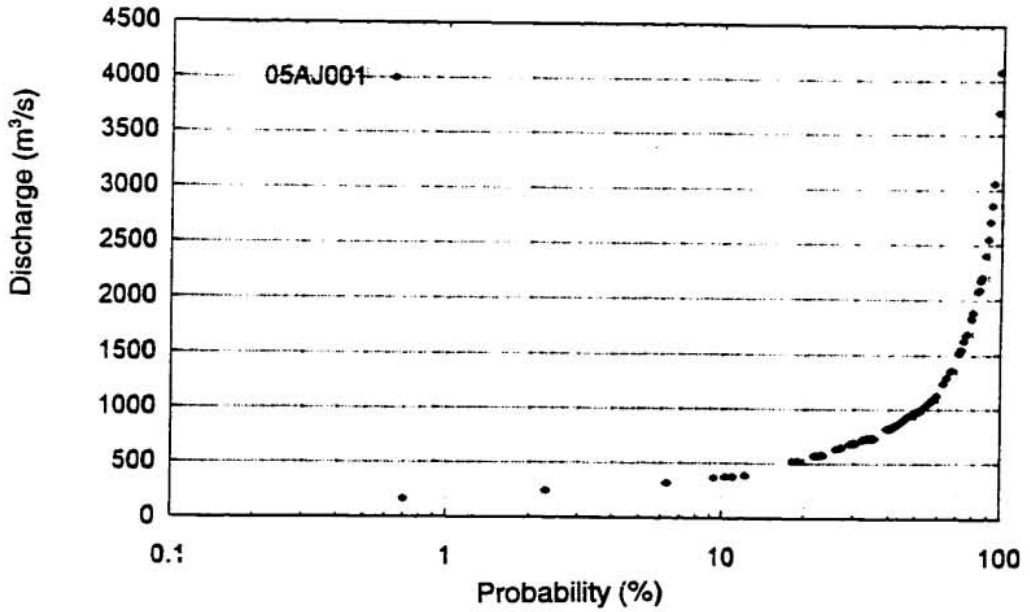

Figure 6.1b Plotting positions using SPP for AM series of South Saskatchewan River at Medicine Hat, Alta., Canada.

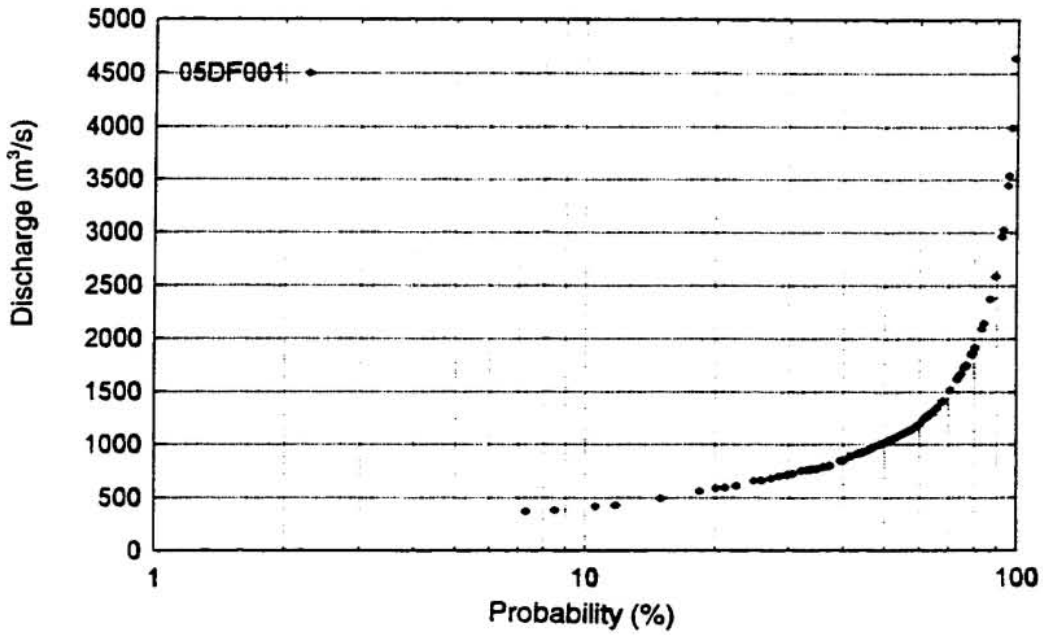

Figure 6.1c Plotting positions using SPP for AM series of North Saskatchewan River at Edmonton, Alta., Canada.

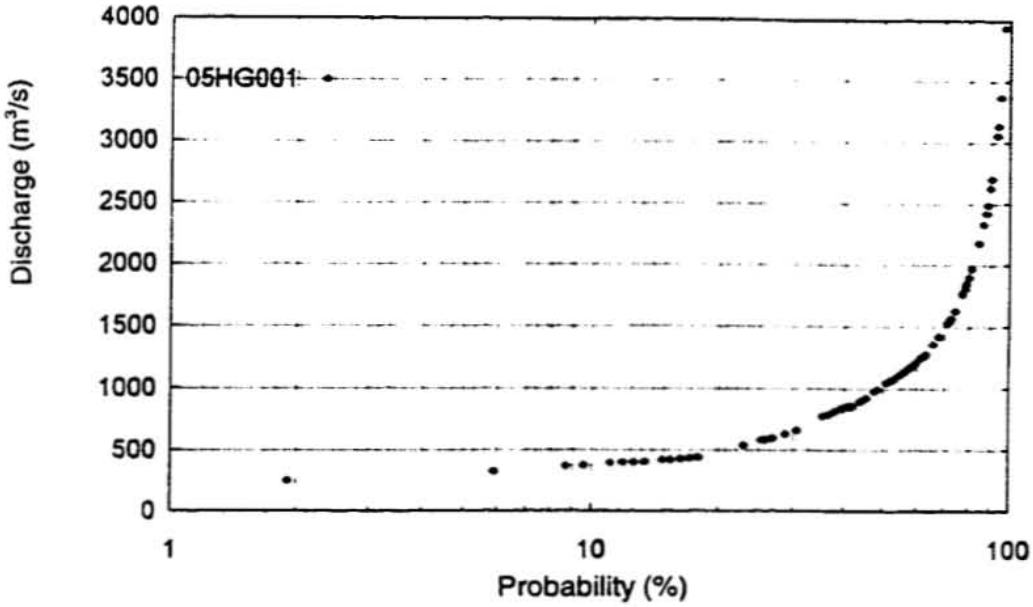

Figure 6.1d Plotting positions using SPP for AM series of South Saskatchewan River at Saskatoon, Sask., Canada.

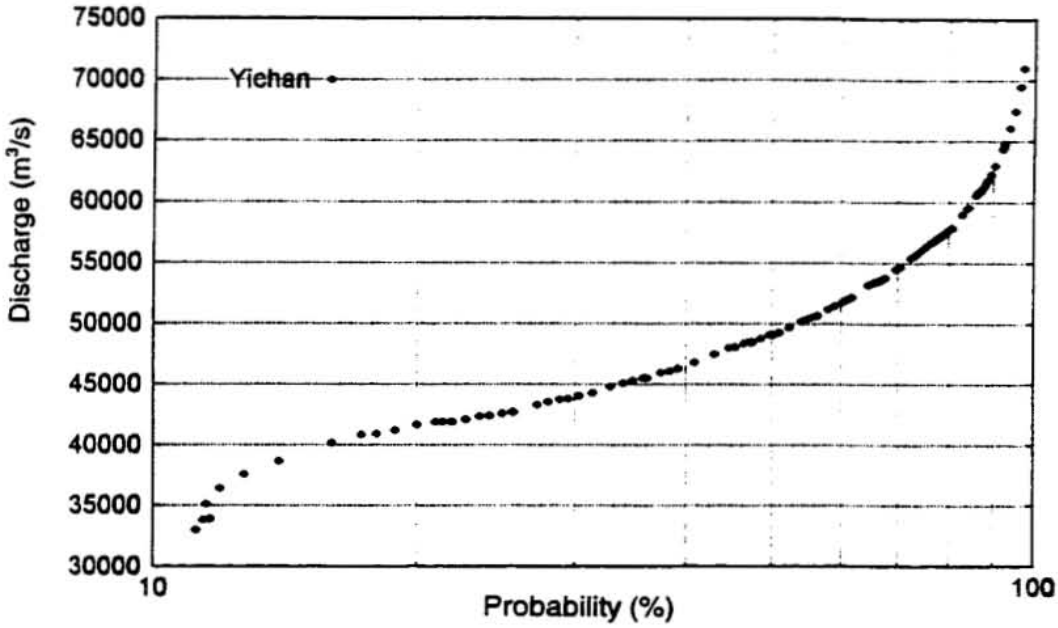

Figure 6.1e Plotting positions using SPP for AM series of Yangtze River at Yichan, China.

6.5.2 Calculation of Flood Quantiles

The Pearson Type III distribution is now assumed as a parent distribution, and the following estimation methods are considered:

1) The conventional moment method (MOM)
2) The PWM method (PWM) (Hosking, 1986, 1990; Song and Ding, 1988)
3) The maximum likelihood method (ML) (Matalas and Wallis, 1973; Cong & Tan, 1979)
4) The graphical curve fitting method (Dalrymple, 1960) where plotting position formulas come from Groups I, II and III.
    a) Weibull plotting position formula (0.0) (Weibull, 1939)
    b) Cunnane plotting position formula (0.4) (Cunnane, 1978)
    c) Hazen plotting position formula (0.5) (Hazen, 1914)
    d) Chegodayev plotting position formula (0.3) (Chegodayev, 1955)
5) Scaling plotting position formula (SPP)

The results of quantiles estimated are illustrated in Table 6.4a-b and Figure 6.2a-b.

Table 6.4a Flood quantiles (units: m^3/s) estimated from various Canadian Rivers

Method	Probability								
	0.001	0.002	0.005	0.01	0.02	0.05	0.100	0.200	0.500
WSC Number: 05AJ001									
MOM	5675	5185	4532	4033	3409	2854	2332	1792	1028
PWM	6132	5561	4806	4234	3662	2904	2330	1753	986
ML	5546	5077	4452	3973	3538	2836	2327	1800	1040
0. 0	6968	6271	5354	4664	3980	3085	2419	1769	958
0. 4	6600	5958	5085	4438	3795	2955	2330	1720	958
SPP	7883	7080	6024	5228	4438	3407	2640	1890	955
WSC Number: 05Hg001									
MOM	5168	4772	4237	3824	3402	2824	2365	1875	1128
PWM	5516	5062	4455	3989	3517	2878	2378	1856	1094
ML	5520	5058	4441	3969	3493	2850	2350	1833	1088
0.0	5871	5372	4707	4198	3684	3684	2452	2452	1092
0.4	5650	5175	4540	4054	3563	2902	2387	1854	1089
SPP	6032	5511	4816	4285	4285	3024	2461	1878	1040
WSC Number: 05DF001									
MOM	6133	5584	4856	4304	3750	3012	2450	1880	1106
PWM	6290	5712	4947	4368	3789	3023	2443	1862	1091
ML	5950	5433	4746	4223	3697	2992	2450	1896	1122
0. 0	7304	6572	5612	4890	4174	4174	2549	1878	1053
0. 4	6927	6240	5337	4658	3986	3109	2459	1829	1053
SPP	7582	6817	5812	5057	4308	3332	2609	1907	1044

Table 6.4b Flood quantiles estimated from real data observed at Yichan, Yangtze River, China

Units of quantiles: m^3/s

Method	Return Period (years)					
	1000	500	200	100	50	20
MOM	76700	75100	72800	70900	68800	65500
PWM	76500	75000	72800	70900	68800	65700
WL	83700	81200	77600	74700	71600	67200
0.0	80000	78000	75200	72900	70300	66500
0.3	79500	77600	74800	72500	70000	66300
0.4	79300	77400	74600	72300	69800	66200
0.5	79100	77200	77200	74400	72200	66000
SPP	100900	95900	89000	83800	78400	71000

Figure 6.2c shows the magnitudes of quantiles estimated by SPP are steeper than the quantiles estimated by other methods. The reasons will be explained in the next section.

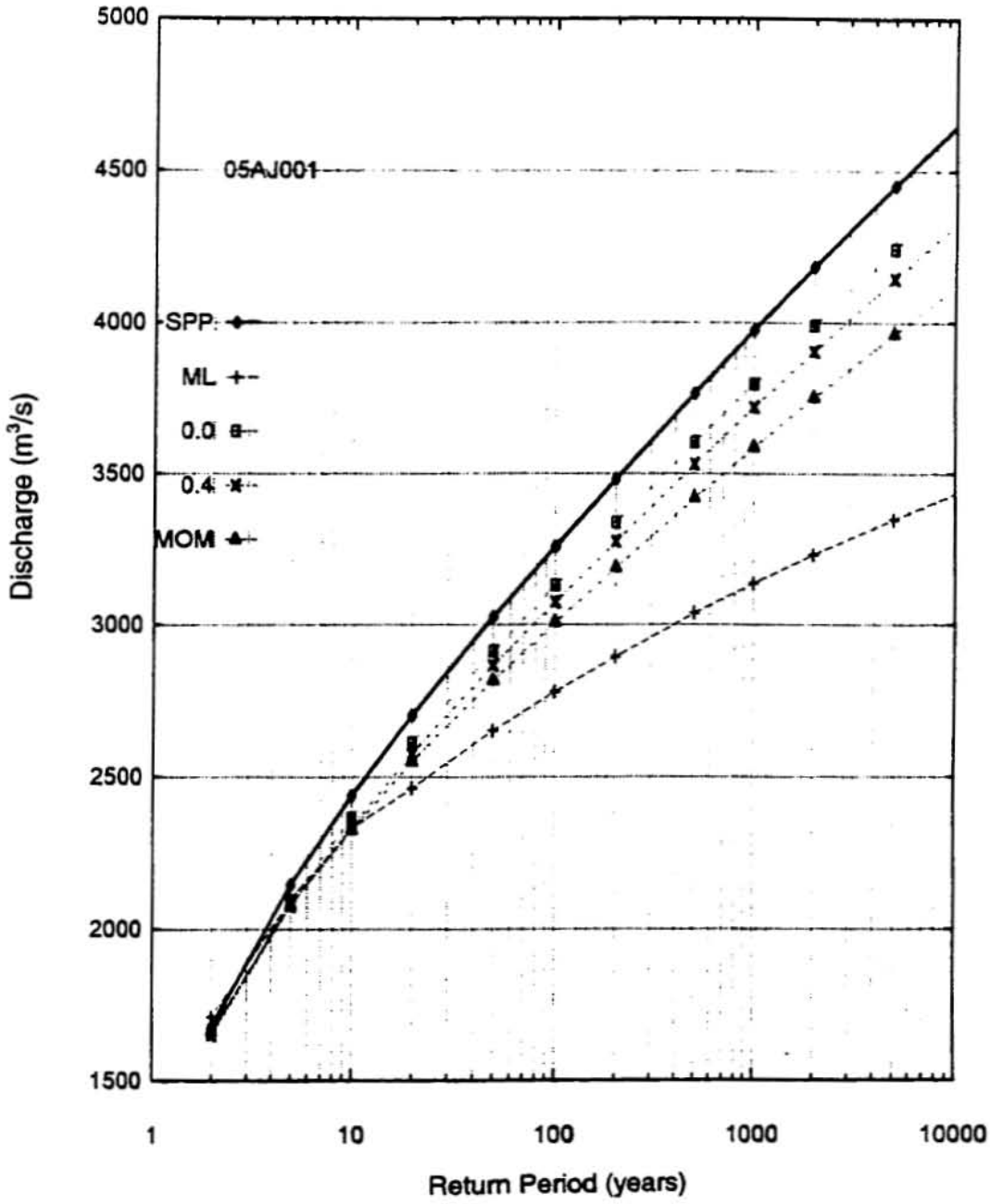

Figure 6.2a Frequency curves of annual flows for South Saskatchewan River at Medicine Hat, Alta., Canada.

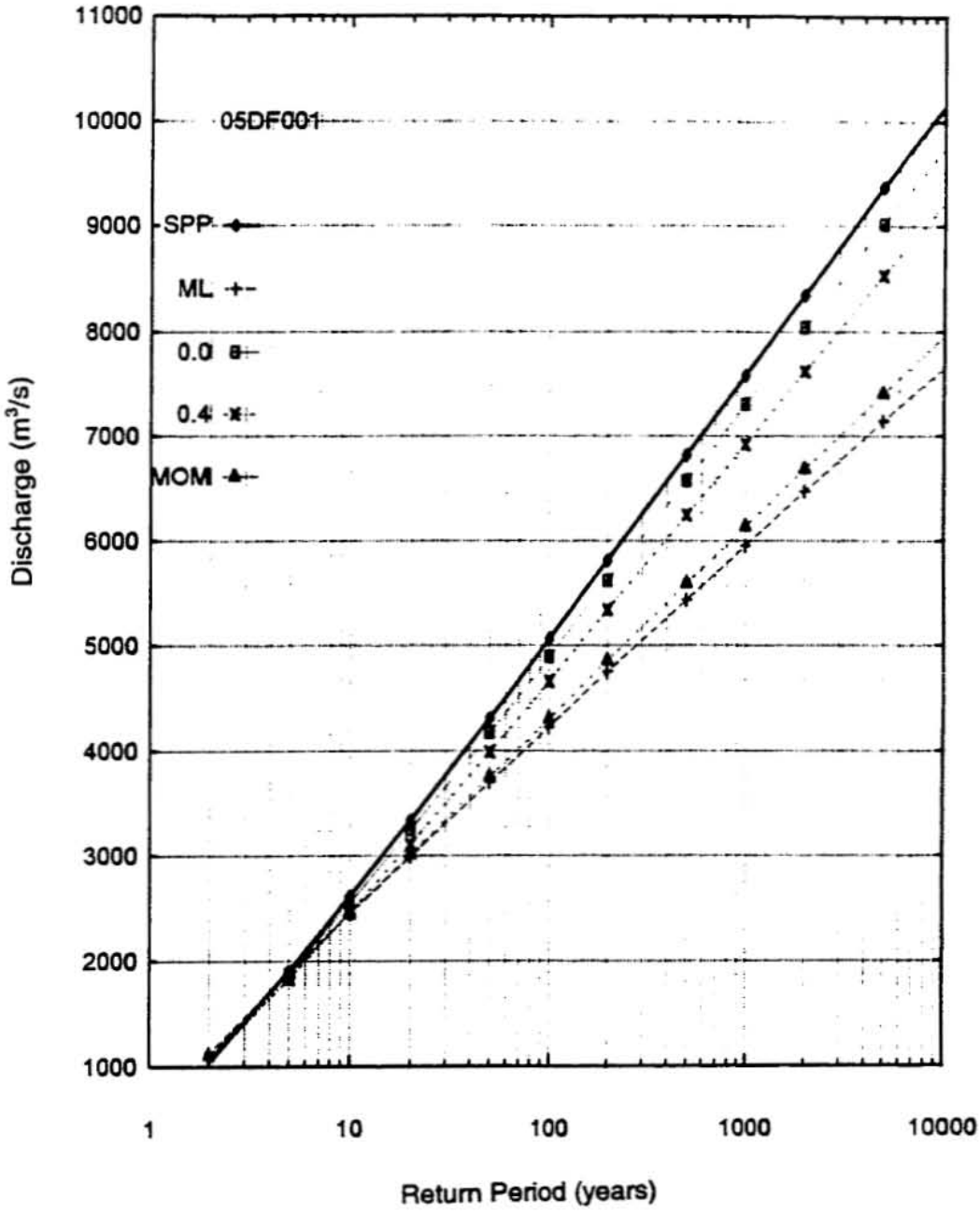

Figure 6.2b Frequency curves of annual flows for North Saskatchewan River at Edmonton, Alta., Canada.

6.6 Comparison Between SPP and Existing Estimators

Statistical experiments will be carried out in order to assess the statistical properties of the SPP estimator and a comparison will be made with existing estimators.

6.6.1 Criterion for Assessment of an Estimator

The choice and appraisal of plotting position formulas have been discussed and the criterion of the optimum of plotting position of unbiasness and efficiency has been widely accepted (Cunnane, 1978). The AM series are sampled from a generated daily flow process and population parameters in this study are not available, so the subject of bias is not discussed. The criterion for the comparison of the desired statistic is to minimise the sum of squares of deviations for the estimated quantiles in this study. Since estimated quantiles are final results of estimation, this criterion is acceptable. Thus, relative root mean square error (RRMSE) of quantiles is used as indices of efficiency for the various quantile estimators.

Let Q_p denote the estimated quantile, the relative root mean square error (RRMSE) is defined by

$$RQ_p = \frac{SQ_p}{BQ_p} \times 100\% \qquad (6.15a)$$

where

$$BQ_p = \frac{1}{K}\sum_{i=1}^{K} Q_p \qquad (6.15b)$$

and

$$SQ_p = \sqrt{\frac{1}{K}\sum_{i=1}^{k}(Q_p - BQ_p)^2} \qquad (6.15c)$$

and K is the number of replications of Monte Carlo experiments. For a flood frequency analysis procedure to be accurate, it should have low RRMSE.

6.6.2 Generation of Flow Time Series

The SPP estimator is related to the temporal structure of observations, daily flow time series are generated and annual maximum series will be obtained from the simulated daily flow series. In order to model long- and short- term dependence the model used here is the mixed-noise model (Lettenmaier and Burges, 1977; Booy and Lye, 1989) to model the daily flow series observed at Yichan, Yangtze River, China.

The mixed noise model is given by:

$$X_t = \sum_{i}^{M} W_i X_t^{(i)} \qquad (6.16a)$$

and

$$X_t^{(i)} = \Phi_i X_{t-1}^{(i)} + \sqrt{(1-\Phi_i^2)}\,\alpha_t^{(i)} \qquad i = 1,2,\ldots,M \qquad (6.16b)$$

where X_t denotes the standard normal series, $X_t^{(i)}$ indicates the standard normal AR(1) series with the first-order correlation coefficient Φ_i, W_i is the weighted coefficient meeting $W_i \geq 0$ and $\sum W_i^2 = 1$, M is the number of terms, and $\alpha_t^{(i)}$ presents the normal independent random variable with mean 0 and variance 1.

The simulated processes of this model are modelling M-α_t series , that is, $\alpha_t^{(1)}$, $\alpha_t^{(2)}$, ..., $\alpha_t^{(M)}$, obtaining M-X_t series, that is, $X_t^{(1)}$, $X_t^{(2)}$,... $X_t^{(M)}$, by Eq.6.16b; then substituting them into Eq.6.16a to obtain X_t; finally obtaining the modelled series of the original flood by inverse standardisation, i.e. Q_t, $Q_t = S_Q X_t + \overline{Q}$, where S_Q , $\overline{Q}$ are the standard deviation and the mean, respectively.

The estimated parameters of the mixed noise models, W and Φ, are shown in Table 6.5 and the corresponding statistics, such as autocorrelation coefficient r_i, i=1,2,3,4, Hurst's K, mean value $\overline{Q}$, and coefficient of variation Cv and coefficient skewness Cs, of simulated model and observations estimates are shown in Table 6.6.

Table 6.5 The estimated parameters of the mixed noise model of daily flows

i	1	2	3	4	5	6	7	8
W_i^2	0.6232	0.2812	0.0562	0.0067	0.0207	0.0037	0.0049	0.0036
Φ_i	0.9200	0.9300	0.9500	0.9915	0.9955	0.9991	0.9996	0.9999

Table 6.6 Comparison of the statistics between the simulated model and observations

Parameter	r_1	r_2	r_3	r_4	Hurst's K	$\overline{Q}$ (m^3/s)	Cv	Cs
Model	0.94	0.54	0.34	0.015	0.735	51500	0.170	-0.15
Observed	0.92	0.66	0.44	0.00	0.742	51600	0.171	0.29

6.6.3 Population Distributions and Comparison Methods

The two population models, Pearson Type III (PIII), and lognormal distribution(LN) expressed by Eqs.6.1 and 6.2, respectively, are used, and five estimation methods are considered in this study. That is:

1) the conventional moment method, PIII/MOM and LN/MOM
2) the PWM methods, PIII/PWM and LN/PWM (Ding et al., 1989)
3) the maximum likelihood method, PIII/WL (Matalas and Wallis, 1973; Cong and Tan, 1979)
4) the graphical curve fitting methods, PIII/FIT and LN/FIT (Dalrymple, 1960) in which probability plotting position formulas include formulas come from Groups I, II and III.
    a) Weibull plotting position formula, PIII/0.0 and LN/0.0 (Weibull, 1939)
    b) Cunnane plotting position formula, PIII/0.4 and LN/0.4 (Cunnane, 1978)
    c) Hazen plotting position formula, PIII/0.5 and LN/0.5 (Hazen, 1914)
    d) Chegodayev plotting position formula, PIII/0.3 and LN/0.3 (Chegodayev,

1955)

5) scaling approach plotting position formula, PIII/SPP and LN/SPP

The quantiles of floods are estimated using Eqs. 6.3 and 6.4 for Pearson Type III and lognormal distributions, respectively.

6.6.4 Monte Carlo Simulation

Monte Carlo experiments consist of the following steps:

1) AM and POT series are sampled from the mixed-noise model with parameters shown in Table 6.5 and Table 6.6. The AM and POT series sampled from this model is used for obtaining the SPP estimator, where AM series only are used in MOM, PWM, ML and graphical curve fitting procedures, AM and POT series are used in SPP procedure.
2) For a given distribution, such as Pearson Type III and lognormal distributions, each selected estimation procedure is performed, and corresponding quantiles are obtained using Eq.6.3 or Eq.6.4.
3) Steps 1-2 are repeated K times and K- quantiles for each procedure are obtained, then the relative root mean square error (RRMSE) of estimators are calculated using Eq.6.15.
4) Compare RRMSE for each estimation procedure.

6.6.5. Results of Monte Carlo Experiments

The results of Monte Carlo simulation are illustrated in Tables 6.7 and 6.8 and Figures 6.3-6.4. Overall the results show:

1) For the PIII and lognormal models, the lines of SPP quantiles varying with return periods are steeper than those of the others.
2) The relative root square errors of the quantiles, RQ_p, are much smaller than those by the alternative estimation procedures assessed whether parametric or nonparametric sampling was used. In other words, SPP quantile estimator is the most efficient among the compared estimators.

In summary, the SPP estimator may be accepted as the most efficient estimator among those estimation procedures studied. According to the good statistical properties of SPP estimator, the author agrees with the remark that "Statistical science can play a role in the future developments of nonlinear science and its possible impact on the future development of statistical science itself." (Chatterjee & Yimaz, 1992), a prediction of the relationship between fractals and statistics.

Table 6.7 Comparison of SPP with other estimators for the Pearson Type III distribution, where the units of BQ_p and SQ_p are m^3 /s but RQ_p is in dimensionless

P_1=.01	P_2=.001	Estimation Methods of Quantiles						
n	Variable	MOM	PWM	0.0	0.3	0.4	0.5	SPP
30	BQ_{p1}	71374	73256	75727	73941	73258	72579	91742
	BQ_{p2}	77564	78680	83615	81247	80336	79441	107601
	SQ_{p1}	6650	7291	8278	7461	7095	6751	8655
	SQ_{p2}	9151	10132	11766	10353	9789	9267	11927
	RQ_{p1}	9.32	10.11	11.06	10.09	9.68	9.30	9.44
	RQ_{p2}	11.80	12.88	14.07	14.74	12.19	11.66	11.08
40	BQ_{p1}	71681	72265	74743	72328	72804	72287	90460
	BQ_{p2}	78023	78866	82208	80337	79641	78960	105910
	SQ_{p1}	6198	6576	7180	6592	6402	6197	7255
	SQ_{p2}	8562	9150	9991	9102	8816	8516	10016
	RQ_{p1}	8.65	9.10	9.61	8.99	8.78	8.57	8.02
	RQ_{p2}	10.97	11.60	12.15	11.33	11.07	10.79	9.46
50	BQ_{p1}	71897	72279	74667	73523	73054	72539	90004
	BQ_{p2}	78325	78877	82158	80654	80028	79340	105327
	SQ_{p1}	5323	5601	6173	3376	5453	5265	6035
	SQ_{p2}	7397	7824	8699	7790	7600	7308	8366
	RQ_{p1}	7.40	7.75	8.27	7.58	7.46	7.26	6.71
	RQ_{p2}	9.44	9.92	10.59	9.66	9.50	9.21	7.94

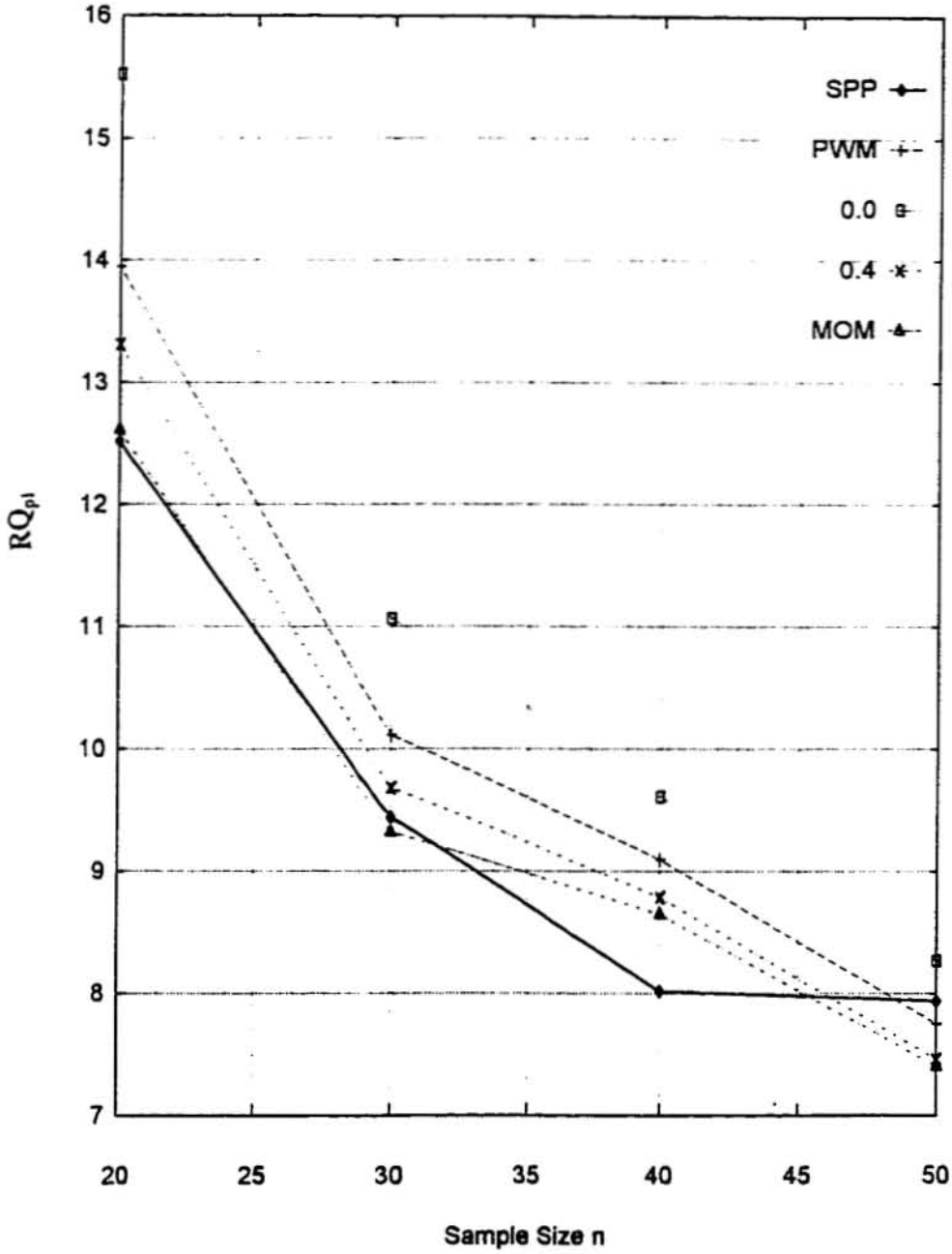

Fig.6.3 SMSE of quantile estimates, RQ_{pl}, varying with the sample size, n, for the Pearson Type III distribution.

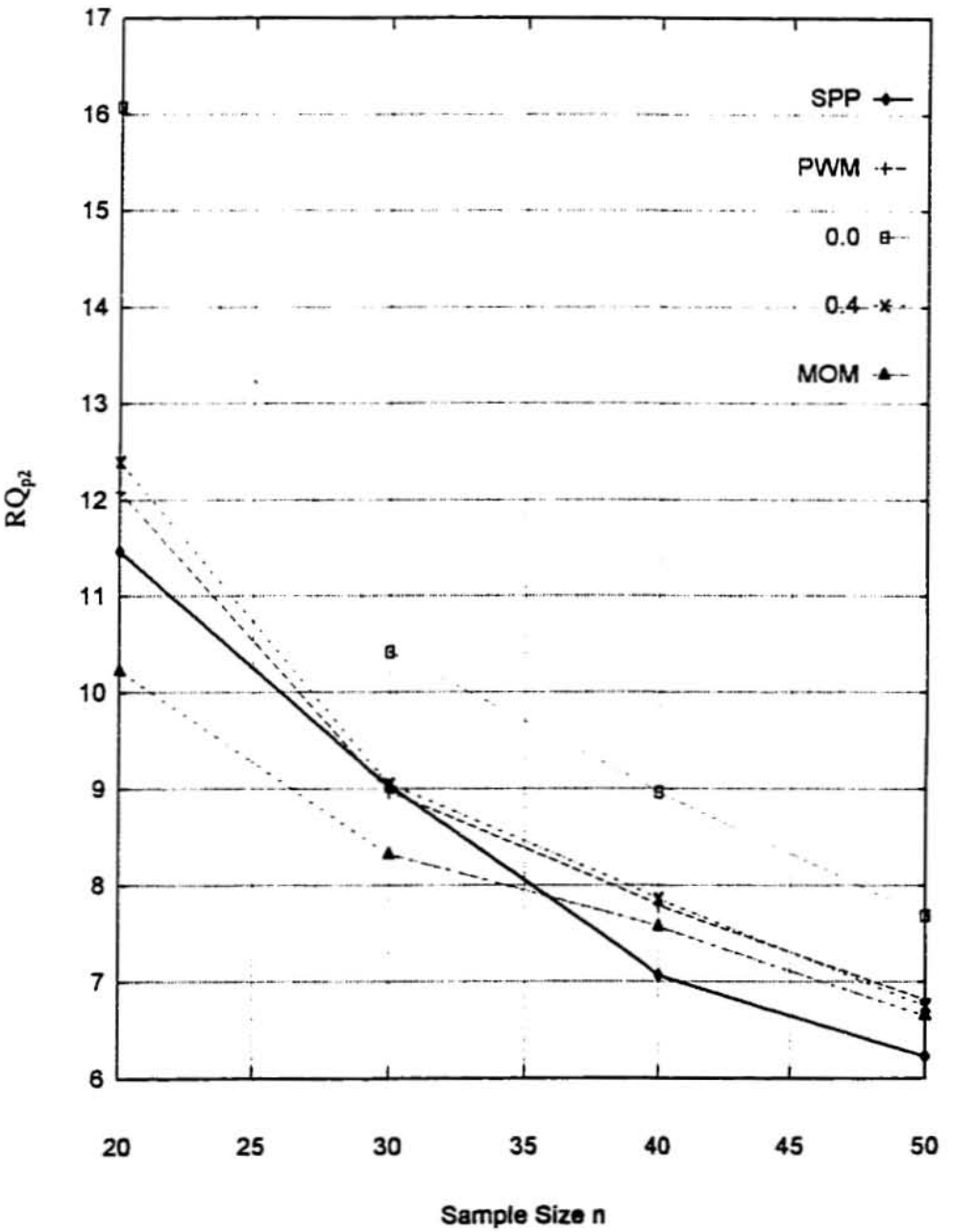

Fig.6.4 SMSE of quantile estimates, RQ_{p2}, varying with the sample size, n, for the lognormal distribution.

6.7 Analysis and Discussion

6.7.1 Background of Development of SPP Formula

Recalling Chapters 2 and 3, high probabilities of long-term persistence exist in the peak flow series observed in Canada and China. And straight lines appeared in the family of curves in Chapter 5 indicate an existence of scale-invariance structure. These statistical features should be incorporated into flood risk analysis.

Classical probability plotting position formulas are derived from independent random variables Q_1, Q_2, ..., Q_n with identical pdf, f(q). The two procedures for theoretically developing plotting position formulas are based on this iid assumption. The first is based on the distribution of the probability of order statistics $Y_{(m)}$ shown in Eq. 6.9 because of iid assumption. The second procedure is based on the distribution of $Q_{(m)}$ using Eq.6.8, it also involves the same assumption. Even through order statistics $Q_{(m)}$ are necessarily dependent because of the inequality relation among them (David, 1984), but the variables Q_1, Q_2, ..., Q_n must be statistically independent and identically distributed in the basic assumptions, otherwise, plotting position formulas cannot be derived from these two procedures.

However, according to classical iid assumption, some mathematical statisticians consider to take short-term dependence of random variables into account (David, 1984), but long-term correlation structure involved in random variables Q_1, Q_2, ..., Q_n, has not been considered yet in the plotting position development. Thus, an empirical plotting

position formula that takes scaling behavior into account is needed to be developed.

6.7.2 Basis of SPP Formula

As Wila et al. (1962) indicated,

> "if the sample is large enough, the sample order statistics when plotted against the corresponding quantiles of the theoretical distribution will tend to yield a set of approximately collinear points, clustering about a line of slope 1 passing through the origin."

For samples of intermediate size, the expectation of order statistics has the same linear relationship with corresponding quantiles. These are the sources most theoretical plotting position formulas are derived from.

Scaling plotting position formula is, in fact, related to this concept. P_ξ is a random variable, a probability of exceedences according to corresponding the ranked threshold, $Qs_{(m)}$, in a family of curves. These ranked threshold, $Qs_{(m)}$, and ranked P_ξ could serve as order statistics and corresponding empirical exceedence probability. If iid assumption is made, the same procedure to derived theoretical plotting position formulas could be processed.

However, SPP is not only agreeing with iid assumption, but, additionally, takes scaling behavior into account.

A family of curves well exhibits peak flow correlation structure across scales. SPP

formula is based on this, the information about scale-invariance including long-term persistence of peak flows could be taken into account.

Suppose the probability of exccedence of peak flows, P_ξ ,is a random variable, it is related to the variable Qs and ξ , where Qs and ξ are the exceedence level expressed as a threshold and time scale, respectively. The expectation of exceedence of peak flows, $E(P_\xi)$ can be estimated using a statistical model.

For a given time scale $\xi_0=1$ year, expected probability of exceedence P_ξ should be an empirical plotting position, and corresponding formula is a formula of probability plotting position. Thus, SPP is an approximation of $E(Y_{(m)})$ shown in Eqs. B.9 and B.10 of Appendix B, where $E(Y_{(m)})$ is the expectation of $Y_{(m)}$.

However, scaling behaviour here is already taken into account in the SPP procedure. It is, therefore, expected that SPP provided an improved capacity of prediction and a more rational and comprehensive way of estimating extreme flood events from the underlying systems. The development of SPP formula is thus reasonable and reliable.

6.7.3 Comparison of Classical Formulas

The SPP has been demonstrated to be a good estimator in terms of efficiency and robustness using Monte Carlo experiments. Additionally, the SPP has the property that it is distribution free even though an assumption that P_{SPP} is normally distributed which has been made for a linear model.

6.7.4 Role of SPP Formula in Flood Risk Analysis

In recent years, scale-invariant (fractal) statistics are being utilised in statistics as a new term. Turcotte and Greene (1993) (hereinafter referred to as T & G) have proposed a scale-invariant approach to flood risk analysis. T & G hypothesised a power law scaling of peak annual discharge and recurrence interval, i.e. the underlying physical processes are sufficiently scale invariant over time scales from one to one hundred years. T & G argued that this hypothesis provided a basis for the application of scale-invariant statistics. The results of their proposal showed that the fractal prediction parallels that of the log Gumbel distribution. T & G also stated that the Hurst exponent cannot be used for flood frequency prediction.

T & G's proposal at first glance seems to provide a rational fractal analysis in flood risk estimation. In fact, the basis of the proposed approach appears weak, and is conceptually flawed, because:

a) Any prediction of floods must be based on an understanding of the behaviour of observations. T & G did not investigate the character of the temporal scaling behaviour of floods. They merely hypothesised to "avoid difficulties with annual variability"!

b) Mandelbrot (1982) has pointed out that the motivation for assuming scaling must not be misinterpreted, nature is not strictly homogeneous or scaling. It is incorrect for T & G to assume that the ratio of ten-year peak discharge to the one-

year peak discharge equals the ratio of the 100-year peak discharge to that of the 10-year peak, since scaling fractals should be limited to an investigated scaling range. The temporal scaling range of floods must be identified before making such an assumption.

c) The equation, $N=C_2Q^{-\alpha}$, given by T & G showing the relationship between the number of exceedences N, and corresponding threshold, Q, is a basis of the proposal, in which the parameters α and C_2 are estimated from observed data. Those curves are not a sufficient proof of an existence of slope α alone. If there should exist more than one slope, the basis of the proposal collapses. Also, most experiences with flood peaks show that there is no single straight line in a log N - log Q plotted curve.

d) T & G's results showed that the Hurst exponent for all ten stations is virtually constant within a range of 0.66 to 0.73. This indicates that there is some persistence inherent in the observed floods. Fractional Brownian noise is a model which reflects long-term character of time series. T & G's assumption of fractional Brownian noise should reflect the persistence inherent in the observed data. T & G claimed that the value of H_1 obtained from the R/S analysis does not correlate with the values of H obtained from the scale-invariant approach. This is evidence that the scale-invariant procedure is both flawed and incomplete.

e) The proposed method by T & G remains as a conventional statistical rubric based

on the assumption that the estimated quantiles by T & G's proposal correlate best with the log Gumbel. The proposed usage is only a method that assumes a linear relationship between log Q_T and log T. The temporal characteristics of the floods are not taken into account.

Hence, fractal statistics that may be used for risk estimation have not yet been developed. SPP estimator may yet play an important role in a new review of flood risk analysis.

6.8 Summary

In this chapter, a scaling plotting position formula has been established. It was shown that the proposed SPP estimator is pertinent both in efficiency or robustness among existing estimation methods by Monte Carlo experiments. Additionally, a number of discussions of the properties of SPP formula have been offered. The proposed SPP procedure extracts more information from underlying systems and takes scaling behaviour into account in flood frequency analysis and, has an enhanced capacity of prediction over those currently available.

Additionally, a few interesting points can be made here:

1) Scaling plotting position formula as an empirical plotting position formula has been developed. It agrees with iid assumption for parent variables and also takes taking scaling behaviour into account.

2) A family of curves best describes peak flow correlation structure across scales. SPP formula is based on this, information about scaling behavior including long-term persistence of peak flows could be taken into account.

3) Monte Carlo experiments show that SPP quantile estimator is the most efficient and robust among the compared estimators.

Chapter 7

Conclusions and Recommendations

7.1 Conclusions

The primary conclusions from this thesis are:

1) A pluralistic view of the correlation structure of peak flows through multiple measurement scales was made and a physical explanation of the natural behaviour of peak flow structure provided a stronger basis for a deeper understanding of the complexity of hydrologic phenomena in nature.

2) The Hurst's K and the lag-one autocorrelation coefficient r(1), which measure the long- and short-term behaviours of annual peak flow series respectively, are significantly correlated and dependent based on parametric and non-parametric hypothesis tests. It indicates that the long-term behaviour of annual peak flows is related to the short-term behaviour statistically.

3) The dependence between Hurst's K and r(1) provides a strong basis to further

investigate the statistical relationship between the long-term persistence and short-term independence. From this, the sampling distribution of Hurst's K was proposed to be expressed as the sampling distribution of Hurst's K for a given r(1).

4) Based on probability theory and empirical statistical results, a probabilistic approach for quantitatively dealing with long-term and short-term behaviour of annual peak flow series was developed. In this approach,
   a) an approximation for the sampling distribution of Hurst's K for a given r(1) was designed and estimated using Monte Carlo experiments;
   b) an estimate of the probability for serially independent population, such as the annual peak flow series, to exhibit long-term persistence was provided and estimated;
   c) empirical percentage points proposed by Lye and Lin (1994) for testing long-term persistence was revised to take short-term behaviour into account.

The results of the proposed approach demonstrate that:

   a) the proposed estimator for population $P(K \geq k_0)$ and its distribution on the R1 axis assure that long-term persistence and short-term independence can be quantitatively estimated;
   b) the magnitudes of the probability $P((K \geq k_0) \cap (b_i \leq R1 < a_i))$ for each region

of R1 imply that the simultaneous occurrence of long-term persistence and short-term independence may not be an uncommon phenomenon.

5) Since Hurst's K and lag-one autocorrelation coefficient r(1) are considered as measurements at individual scales, i.e. scale at n and one respectively, fractal geometry, which makes measurement at scale, is necessary to be used to investigate the feature of peak flow structure.

6) The proposed family of probability-scale-threshold curves, which transforms observed peak flows to a family of curves, well describe the scaling behaviour of peak flows and represent corresponding inter-scale correlation structure of peak flows for a given watershed.

7) A straight line for a certain range ξ of a family of $\ln P_{\xi}$ - $\ln \xi$ - Qs curves implies that the occurrence of exceedances of flows displays invariance within the corresponding scaling range. In other words, a correlation across scales exists for the peak flow points on the time axis.

8) A scaling plotting position formula in which scaling feature of peak flows is taken into account was developed and its quantile estimator is more efficient

and robust compared to current estimators of flood quantiles.

7.2 Recommendations for Further Studies

A number of suggestions are offered for further studies:

1) While the proposed probabilistic approach is a method for quantitatively dealing with long- and short-term behaviour of annual peak flow series, the population distribution this study investigated is limited to the normal distribution. Even though the Kolmogorov-Smirnov hypothesis test shows that high probabilities of existence of long-term persistence are involved in normal independent distributed data may be suitable for the non-normal independent series, further studies and investigation of this issue would enhance the attractiveness of probabilistic approach.

2) According to Hurst's findings (1951, 1954), long-term persistence is related to the order of occurrence. We cannot simply resample from the individual observations, because this would destroy the correlation that we are trying to capture. Design of resampling methods such as block bootstrap or suitable jackknife which can preserve long-term correlation and short-term

independence is desirable in order to increase the accuracy of simulations.

3) Hurst has shown that for many of the natural series he investigated, the Hurst coefficient, h, remains larger than theoretical value of 0.5 even for large sample size n. The failure of natural series to accord with theory is termed the "Hurst phenomenon." The proposed probabilistic approach has quantitatively estimated the probability $P(K \geq k_0)$ for a theoretical independent series. How about the value of $P(K \geq k_0)$ for a natural series? What is the difference between the theoretical and observed series? Further explanation of the Hurst phenomenon would be desirable.

4) We have looked closely at peak flow serial structure at individual scales one and n, and also measured the complexity of peak flows across scales. The correlation among these measurements might be interesting and worth further studying for peak flow point processes.

Chapter 8

Statements of Originality

To the best of the author's knowledge, the following original contributions were made as a result of this study.

1) A set of new approaches, descriptions, and modelling techniques are developed in dealing with long- and short-term behaviour of annual peak flow series based on classical probability and statistical theories. In particular,
   a) sampling distribution of Hurst's K expressed as a sampling distribution of Hurst's K for a given r(1) was proposed for a short-term independent series;
   b) Monte Carlo simulations to produce the sampling distribution of Hurst's K for a given r(1) were designed;
   c) more accurate empirical percentage points for testing long-term persistence were produced;
   d) an approach of quantitatively describing long-term correlation rooted in an independent series was provided;
   e) a quantitative descriptor for long-term persistence, $P(K \geq k_o)$, was defined

and proposed. The calculated results provided a means of determining whether long-term dependence exists in an independent time series; and

f) the conclusion that the simultaneous occurrence of long-term persistence and short-term independence is not an uncommon phenomenon in annual peak flows as well as in normally independent series.

2) A look at the serial correlation structure of peak flows from a fractal view provided a family of curves ln P_ξ - lnξ - Qs which provided a fresh way of understanding and describing the serial structure of natural peak flows. It was shown that:

   a) observed peak flow points can be transformed to a family of curves describing the distributions of peak flow points along time axis that classical methods are unable to identify;

   b) the family of curves is expressed as the relationship between various time scales and probabilities of peak flow occurrences. It explores scaling behaviour of peak flows showing natural behaviour, such as the natural cycles inherent in the peak flow series in which it resides;

   c) the occurrence of peak flows in the time axis has a distribution which differs from an independent Poisson distribution for most time intervals.

3) Based on an increased understanding of the natural behaviour of peak flows, a scaling plotting position formula for flood quantile estimation was proposed in which the corresponding quantile estimator has important statistic characteristics:

   a) it takes the correlation structure of peak flows into account in flood risk analysis;

   b) it is efficient among existing quantile estimators;

   c) because it follows iid assumption in flood frequency analysis and considers the long-term correlation structure of peak flows, it has an enhanced capacity of estimation over those currently available.

Bibliography and References

Anis, A. A. and Lloyd, E.H., 1953. On the Range of Partial Sums of a Finite Number of Independent Normal Variates. Biometrika 40, 35.

Benson, M.A., 1968. Uniform Flood-frequency Estimating Methods for Federal Agencies, Water Resour, Res., Vol 4, no.5, pp 891-908.

Blom, G., 1958. Statistical Estimates and Transformed Beta Variables. Wiley, New York, N.Y., pp. 68-75 and pp.143-146.

Bobee, B., 1973. Sample Error of T Year Events Computed by Fitting a Pearson Type 3 Distribution, Water Resour. Res., 9(5), 1264-1270.

Boes, D.C., and Salas, J.D., 1978. Nonstationarity of the Mean and the Hurst Phenomenon. Water Resources Res., 14(1): 135-43.

Booy, C. and Lye, L.M., 1989. A New Look at Flood Risk Determination. Water Resour. Bull., 25, No. 5.

Box, G.E.P. and Jenkins. G.M., 1970. Time Series Analysis Forecasting and Control Holden-Day, San Franciso, Calif., 553pp.

Box, G.E.P. and Cox, D.R., 1964. An Analysis of Transformation. J.R. Statist. Soc., Ser. B, 26: 211-252.

Carrigan, P.H. and Huzzen, C.S., 1967. Serial Correlation of Annual Floods . Proc. Int. Hydrol. Symp., Fort Collins, CO. Pp. 322-328.

Chatterjee, Sangit and Mustafa R. Yilmaz, 1992. Chaos, Fractals and Statistics. Statistical Science, Vol. 7, No. 1, 49-121.

Chegodayev, 1955. Formulas for the Calculation of the Confidence of Hydrologic Quantities by A.G. Alekseyev. In: V.T. Chow (editor), Handbook of Applied Hydrology, 1964.

Chong, S. and Tan, W., 1979. Statistical Testing Research on the Method of Parameter Estimation in Hydrological Computation. Proc. Int. Symp. Spec. Aspects Hydrol. Comput. Water Proj., UNESCO-IHP.

Chow,V.T., 1951. A General formula for Hydrologic Frequency Analysis, Trans, Am. Geophysical Union, Vol. 32, no. 2, pp. 231-237.

Chow,V.T., 1954. The log Probability Law and Its Engineering Applications. Proc. ASCE, 80(536), 1-25.

Chow, V. T., 1964. Handbook of Applied Hydrology, McGraw-Hill.

Chow, V. T., Maidment, David R. And Mays, Larry W, 1988. Applied Hydrology, McGraw-Hill Book Company.

Crutcher H.L., 1975. A Note on the Possible Misuse of the Kolmogorov-Smirnov Test, Journal of Applied meteorology, 14: 1600-1603.

C.S.D.P.W.(California State Department of Public Works), 1923. Flow in California streams. Bull. 5, Ch.5.

Cunnane, C. 1973. A Particular Comparison of Annual Maximum and Partial Duration Series Methods of Flood Frequency Estimation. J. Hydrol., 18,257-273.

Cunnane, C., 1978. Unbiased Plotting Positions- A Review. J.Hydrl., 37: 205-222.

Cunnane, C., 1985. Hydrological Frequency and Time Series. International Postgraduate Hydrology Course, University College, Galway.

Dalrymple, T., 1960. Flood Frequency Methods. U.S. Geol. Survey, Water Supply Paper 1543 A, pp11-51, Washington.

Daniel, C. And Wood, F., 1980. Fitting Equations to Data, Revised Edition, New York: John Wiley & Sons, Inc.

David, Herbert A., 1981. Order Statistics. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc.

Ding, J., Song, D., Yang, R., and Hou, Y, 1989. Expression Relating Probability Weighted Moments to Parameters of Several Distributions in Expressible in Inverse Form. Journal of Hydrology, No. 110.

Environment Canada, 1992. Hydat CD-Rom, Surface Water and Sediment Data and Water Survey of Canada.

Falconer, K.J., 1990. Fractal Geometry: Mathematical Foundations and Applications. Wiley, Chichester, 288pp.

Federal Council for Science and Technology, 1962. Ad Hoc Panel on Hydrology. Scientific Hydrology. Washington, D.C.

Feller, W., 1951. The Asymptotic Distribution of the Range of Sums of Independent Random Variables. Ann. Math. Stat. 22:427-32.

Forster, H.A., 1924. Theoretical Frequency Curves and Their Application to Engineering Problems, Trans. Am. Soc. Civ. Eng., Vol. 87, pp. 142-173.

Fuller, W.E, 1914. Flood Flows. Trans. Am. Soc. Civil Engrs. 77, 564-617.

Gibbons, J.D., 1971. Nonparametric Statistical Inference. New York: McGraw-Hill.

Greenwood, J. A., Matalas, N.C. and Wallis, J.R., 1979. Probability Weighted Moments: Definition and Relation to Parameters of Distributions Expressible in Inverse Form. Water Reour. Res., 15(5) 1047-1054.

Gringorten, I.I., 1963. A Plotting Rule for Extreme Probability Paper. J. Geophys. Res., 68 (3): 813-814.

Gumbel, E.J., 1943. On the Plotting Rule for Extreme Probability Paper. J. Geophys. Union, 24(2): 699-719.

Gumbel, E.J., 1947. Discussion on Paper by B.F. Kimball, 1946. (Q.v.) Trans, Am. Geophys. Union, 28(6): 951-952.

Guo S. L., 1990. Unbiased Plotting Position Formula for Historical Floods. Journal of Hydrology, 121(1990) 45-61.

Gupta, V. K. and E. Waymire, 1990. Multiscaling Properties of Spatial Rainfall and River Flow Distributions. Journal of Geophysical Research, 95, No. D3, pp. 1999-2009.

Gupta, Vijay K., Sandra L. Castro and Thomas M. Over, 1996. On Scaling Exponents of Spatial Peak Flows From Rainfall and River Network Geometry. Journal of Hydrology, 187(1-2): 81-104.

Haitjema, Henk M. and Kelson, Victor A., 1996. Using the Stream Function for Flow Governed by Poisson's Equation. Journal of Hydrology, 187(3-4): 365-386.

Harter, H.L., 1971. Some Optimization Problems in parameter Estimation. In: J.S. Rustagi (Editor), Optimizing Method of Statistics. Academic Press, New York, N.Y.

Hazen, A., 1914. Storage to be Provided in Impounding Reservoirs for Municipal Water Supply.Trans. ASCE, 77: 1527-1550.

Hirsch, R.M., 1987. Probability Plotting Position Formulas for Flood Records with Historical Information. Presented at U.S.-China Bilateral Symp. On the Analysis of Extraordinary Flood Events, Nanjing, 1985. J. Hydrol., 96: 185-199.

Hirsch, R.M. and Stedinger, J.R., 1987. Plotting Position Formulas for Historical Floods and their Precision. Water Resourc. Res., 23(4): 715-726.

Hosking, J.R.M., 1984. Modelling Persistence in Hydrological Time Series Using Fractional Differencing, Water Resources Research, 20, pp 1898-1908.

Hosking, J.R.M., 1986. The Theory of Probability Weighted Moments.

Hosking, J. R. M., 1990. L-moments: Analysis and Estimation of Distributions Using Linear Combination of Order Statistics. J.R. Statis. Soc., Ser. B, 52(1):105-124.

Hua, Shi-Qian, 1985. A General Survey of Flood Frequency Analysis in Chine. Paper Pres. at US-China Bilateral Symposium on the Analysis of Extraordinary Flood Events, Nanking, Oct. 1985.

Hurst, H.E., 1951. Long Term Storage Capacity of Reservoirs. Trans. Am. Soc. Civ. Eng., 116: 770-808.

Hurst, H.E., 1956. Methods of Using Long-Term Storage in Reservoirs. Proc. Inst. Civ. Eng., 1: 519-543.

Jayawardena, A.W. and Lai, Feizhou, 1994. Analysis and Prediction of Chaos in Rainfall and Stream Flow Time Series, J. Of Hydrology, 153, 23-52.

Jenkins, G.M. and Watts, D.G., 1968. Spectral Analysis and its Applications. Holden–Day, San Franciso, Calif., 525pp.

Ji Xuewu, Ding Jing, H.W. Dhen and J.D.Salas, 1984. Plotting Positions for Pearson Type-III Distribution. Journal of Hydrology, 74 (1984) 1-29.

Jonas Olsson, Janusz Niemczynowicz, Ronny Berndtsson and Magnus Larson, 1992. An

Analysis of the Rainfall Time Structure By Box Counting - Some Practical Implication, Journal of Hydrology, 137, 261-277.

Jonas Olsson and Janusz Niemczynowicz, 1996. Multifractal Analysis of Daily Spatial Rainfall Distributions. Journal of Hydrology, 187(1-2): 29-43.

Kaczmarek, Z., 1957. Sufficiency of the Estimation of Floods with a Given Return Period. Proc. Toronto Sympos., IAHS Pubilc. No.45, 145-159.

Kedem, B. and Chiu, L.S, 1987. Are Rain Rate Processes Self-similar? Water Resour. Res., 23: 1816-1818.

Klecka, W.R., 1980. Discriminant Analysis, Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-019, Beverly Hills: Sage Publications.

Klemes, V. , 1974. The Hurst Phenomenon: A Puzzle?, Water Resources Res., 17(3): 737-51.

Koch, R.W. and G.M. Smillie, 1986. Bias in Hydrologic Prediction Using Log-Transformed Regression Models. Water Resources Bulletin 22(5): 717-724.

Langbein, W.B., 1949. Annual Floods and the Partial Duration Flood Series. Trans. Am. Geophys. Union, 30, 879-881.

Leadbetter, R., 1988. Extremal Theory for Stochastic Processes, Annals of Probability, 16, pp 431-478.

Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks. San Francisco: Holden-Day.

Lettenmaier, D.P. and S.J. Burges, 1977. Operational Assessment of Hydrologic Models of Long-Term Persistence. Water Resources Research 13(1): 113-124.

Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov Test for Normality With Mean and Variance Unknown, J. Am.Stat. Assn.64, 399-402.

Linsley, R. K. J., Max A. Kohler, Joseph L.H. Paulhus, 1958, 1975. Hydrology for Engineers. McGRAW-HILL Book Company.

Lovejoy, S. and Mandelbrot, B., 1985. Fractal Properties for Rain and a Fractal Model. Tellus, 37A: 209-232.

Lovejoy, S. and Schertzer, D., 1990. Multifractals, Universality Classes and Satellite and Radar Measurements of Clouds and Ran Field. J. Geophys. Res., 95: 2021-2034.

Lovejoy, S., Schertzer, D. and Tsonis, A.A., 1987. Functional Box-counting and Multiple Elliptical Dimensions of Rain. Science, 235: 1036-1038.

Lowery, M.D. and Nash, J. E., 1970. A Comparison of Methods of Fitting the Double Exponential Distribution. J. Hydrol., 10(3), 259-275.

Lye, L.M., 1987. Uncertainty in Flood Analysis. Ph.D. Thesis. University of Manitoba.

Lye, L.M., 1993. A Technique for Selecting the Box-Cox Transformation in Flood Frequency Analysis. Can. J. Civ. Eng., 20: 760-766.

Lye, L.M. and and Lin, Y., 1994. Long-term Dependence in Annual Peak Flows of Canadian Rivers. Journal of Hydrology, 160,89-103.

Mallows, C. L., 1973. Choosing a Subset Regression, Unpublished Report, Bell Telephone Laboratories.

Mandelbrot, B.B., 1967. How long is the coast of Britain? Statistical Self-Similarity and Fractal Dimension, Science 155, 636-638.

Mandelbrot, B. B., 1977. Fractals Form, Chance and Dimension, W.H.Freeman, New York.

Mandelbrot, B.B., 1982. Fractal Geometry of Nature, W.H.Freeman, New York.

Mandelbrot, B.B. and Van Ness, J.W., 1968. Fractional Brownian Motions, Fractional Noises and Applications. SIAM Review 10, 422.

Mandelbrot, B.B. and J.R. Wallis, 1968. Noah, Joseph and Operational Hydrology. Wat. Resour. Res. 4, 909.

Mandelbrot, B.B. and J.R. Wallis, 1969a. Computer Experiments with Fractional Gaussian Noises. Part 1: Averages and Variances. Water Resources Res. 5(1):228-241.

Mandelbrot, B.B. and J.R. Wallis, 1969b. Robustness of the Rescaled Range R/S in the Measurement of Noncyclic Long Run Statistical dependence. Water Resources Res. 5(5): 967-988.

Matalas, N.C. and Huzzen, C.S. , 1967. A Property of the Range of Partial Sums. Proc. Fort Collins Symp. Int. Assoc. Of Scientific Hydrol. 4, 252.

Matalas, N.C. and Wallis, J.R., 1973. Eureka! It Fits a Pearson Type 3 Distribution. Water Resour. Res., 11(6), 281-289.

McCuen, R. H. And W. M. Snyder, 1986. Hydrologic Modeling. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

McCuen, R.H., R.B. Leahy, and P.A. Johnson, 1990. Problems With Logarithmic Transformations in Regression. Journal of Hydraulc Engineering 116(3):414-428.

Miller, D.M., 1984. Reducing Transformation Bias in Curve Fitting, The Amer. Statistician, 38(2), pp. 124-126.

Ministry of Water Resources, 1980. Standard Method for Flood Quantile Estimation Procedure. Min. Water Resour. Min. Power Ind., SDJ 22-79, Water Resources Publishing House, Beijing.

Ministry of Water Resources, 1985. Hydrological Survey Data, Water Resources Publishing House, Beijing.

National Environmental Research Council (NERC), 1975. Flood Studies Report, Natur. Environ. Res. Council, London, Vols., 1-5, 1100pp.

National Research Council, 1991. Opportunities in the Hydrologic Sciences, national Academy Press, Washington, D.C.

Nicolis, G. and I. Prigogine, 1989. Exploring Complexity, An Introduction, W.H. Freeman and Company, New York.

Olds, E. G., 1938. Distribution of Sums of Squares of Rank Differences for small Samples. Annals of Mathematical Statistics, Volume 9.

Paolo Burlando and Renzo Rosso, 1996. Scaling and Multiscaling Models of Depth-Duration-Frequency Curves for Storm Precipitation. Journal of Hydrology, 187(1-2): 45-64.

Puente, Carlos E., 1996. A New Approach to Hydrologic Modeling: Derived Distributions Revisited. Journal of Hydrology, 187(1-2): 65-80.

Richardson, L.F., 1961. The Problem of Contignity an Appendix of Statistics of Deadly Quarrels. General Systems Yearbook 6, 139-187.

Rossi, F., M. Fiorentino, and P. Versace, 1984. Two Component Extreme Value Distribution for Flood Frequency Analysis. Water Resour. Res., 20(7), 847-856.

Song Dedun and Ding Jing, 1988. The Application of Probability Weighted Moments in Estimating the Pearson Type Three Distribution. J. of Hydrology, 101, 47-61.

Spolia, S. K. and Chander, S., 1977. Stream Flow Simulation - A Model Based on Cannonical Expansions. J. of Hydrology, Vol. 35, No. 3-4.

Srikanthan, R. and McMahon, T.A., 1981. Log Pearson III Distribution-Effect of dependence, Distribution parameters and Sample Size on Peak Annual Flood Estimates. J. Hydrl., 52: 149-159.

Stedinger, J.R., 1980. Fitting Log Normal Distributions to Hydrological Data. Water Resour. Res., 16(3) 481-490.

Tukey, J.W., 1962. The Future of Data Analysis. Ann. Math. Atat., 33(1): 21-24.

Turcotte, D.L. and Greene, L., 1993. A Scale-Invariant Approach to Flood -Frequency Analysis. Stochastic Hydrology and Hydraulics. Vol. 7, pp33-40.

Venugopal, V. and E. Foufoula-Georgiou, 1996. Energy Decomposition of Rainfall in the Time- Frequency-Scale Domain Using Wavelet Packets. J. of Hydrology, 187(1-2): 3-27.

Wallis, J.R. and O'Connell, P.E., 1972. Small Sample Estimation of r_1. Wat. Resour. Res. 8, 707.

Wallis, J.R. and O'Connell, P.E., 1973. Firm Reservoir Yield-How Reliable are Historic Hydrological Records? Hydrological Sciences Bulletin, XVIII, 39.

Wallis, J.R. and Matalas, N.C., 1970. Small Sample Properties of H and K-Estimators of the Hurst Coefficients h. Water Resources Research 6(6): pp.1583-1594.

Wall, D.J. and Englot, M.E., 1985. Correlation of Annual Peak Flows for Pennsylvania Streams. Water Resour. Bull., 21(3):459-464.

Waymire, E., 1985. Scaling Limits for Precipitation Fields. Water Resour. Res., 21:

1271-1281.

Waymire, E. and Gupta, V.K., 1987. On Long Normality and Scaling in Rainfall. In: S. Lovejoy and D.Schertzer (Editors), Scaling, Fractals and Nonlinear Variability in Geophysics. Reidel, Hingham, MA, 318 pp.

Weibull, W., 1939. A Statistical Theory of Strength of Materials. Ing. Vet. Ak. Handl. 151, Generalstabens Litografiska Anstals Forlag, Stockholm.

Wila, M.B., Granidesikan, R. And Huyett, M.J., 1962. Probability Plots for the Gamma Distribution. Technometrics, 4(1):1-20.

Wood, E. and I. Rodriguez-Iturbe, 1975. Bayesian Inference and Decision Making for Extreme Hydrologic Events. Water Resources Research 11 (4): 533-542.

World Meteorological Organization (WMO), 1989. Statistical Distributions for Flood Frequency Analysis, Operational Hydrology, Report No.33, WMO-No.718, WMO, Geneva, 77 pages.

Wu Boxian, Hou, Y. and Ding, J., 1991. The Method of Lower-Bound to Estimate the Parameters of a Pearson Type 3 Distribution. Hydrological Science Journal, 36, 271-280.

Wu Boxian and Hou,Y., 1991. Fractals and their Predicting Applications in Hydrology. Proceedings of Symposium on Prospects for hydrology in 2000, Nanjing, April 3-8.

Yevjevich, V. 1967. Mean Range of Linearly Dependent Normal Variables with Application to Storage Problems. Wat. Resour. Res.3, 653.

Yevjevich, V.M., 1971. Stochastic Process in Hydrology. Water Resources Publications, Colorado.

Zawadzki, I.I., 1990. Is Rain Fractal? International Workshop on New Uncertainty Concepts in Hydrology and Water Resources, Madralin near Warsaw, Poland.

Appendix A

The Hurst Phenomenon

The concept of Hurst coefficient as a measure of long-term persistence is introduced as follows.

Based on studying the design capacity of reservoirs, H.E. Hurst (1951) observed an unexpected behaviour of natural time series, which has become known as the Hurst phenomenon.

Let $x_1, x_2, \ldots, x_n$ be a sequence of annual inflows into a reservoir over n years. Let the mean flow in the n year period be denoted by

$$\overline{x_n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The accumulated departure of the flows from the mean flow after y years is

$$S_y = \sum_{i=1}^{y} (x_i - \overline{x_n})$$

In the last period, $S_n = 0$. The range of the cumulative departures from the mean is

$$R_n = max\ (S_y) - \min\ (S_y) = S_M - S_m$$

where S_M and S_m are the largest and smallest values in the set (S_y).

Hurst studied how the average value of R_n changes as a function of n and found that the expected value of R_n divided by the standard deviation S_n of the n annual inflows is proportional to n raised to some power h.

$$E(\frac{R_n}{S_n}) \sim n^h \qquad \text{(A.1)}$$

The exponent h which varies between 0 and 1 is called the Hurst statistic. The ratio R_n / S_n is called the rescaled range.

In addition to river discharges, Hurst investigated a host of other natural geophysical time series ranging from tree rings to clay varves. All in all, 75 different phenomenon were used. The total number of series was close to 900 and they vary in length from 40 to 2000 years.

Eq.A.1 implies that the relationship between log E(R/S) and log n is linear with slope h. To determine h, Hurst defined

$$\frac{R_n}{S_n} = (\frac{n}{2})^K$$

where K represents an estimate of h for each of the 900 time series he investigated.

Over all phenomena Hurst's K was found to have an average value of 0.73 with a standard deviation of 0.08. Asymptotically, for independent normal random variables, Hurst (1956) and Feller (1951) showed that

$$E(R_n) = (\frac{\pi}{2})^{\frac{1}{2}} \sigma n^{\frac{1}{2}}$$

In other words, h $\rightarrow$0.5 as n becomes large. But Hurst also made the far reaching discovery that for many of the natural series he investigated, the slope h remains much steeper than 0.5 even for large values of n. The failure of natural series to accord with theory is termed the "Hurst phenomenon". This so called phenomenon generated considerable interest among hydrologists and mathematicians alike since it indicates a puzzling long term "memory" or "persistence" in the random process that generated the series.

Conversely, anti-persistent processes, that is $h<0.5$, on the other hand, tend to show a decrease in values following previous increases, and show increases following previous decreases. The record of an anti-persistent process, such as the $h=0.1$ curve, appears very "noisy". They have local noise of the same order of magnitude as the total excursions of the record.

In the literature, there are three main lines of thought explaining the Hurst phenomenon:

1) The Hurst phenomenon is a transitory behaviour. The argument is that our series are simply not long enough to test the steady-state behaviour of R, which according to the argument is the square-root law. This period of transition can be reproduced by Markov-autoregressive models. On the basis of a very long time series, Mandelbrot and Wallis (1968) effectively argue against this explanation.

2) The Hurst phenomenon is due to nonstationarities in the underlying mean of the process. This argument claims that a low-frequency, slowly time-varying mean explains the Hurst behaviour (Klemes, 1974; Boes and Salas, 1978).

3) The Hurst phenomenon is due to stationary processes with very large memory. That is, stationary processes that have correlation functions that decay very slowly in time, much slower that Markov-Gaussian-autoregressive processes. In the limit, this argument claims infinite memory for natural processes.

Appendix B

Derivation of Plotting Position of Formulas

As an example to show the basic concepts of plotting position formulas of Group I, Weibull (1939) formula is derived as following:

If the random variables $X_1, X_2, \ldots, X_n$ are arranged in the order of magnitude as

$$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(m)} \geq \ldots \geq X_{(n)}$$

$X_{(m)}$ are called the m^{th} order statistic (i=1,2, ...,n). X_i are assumed to be statistically independent and identically distributed with probability density function, f(x), and cumulative distribution function, F(x). Then the probability density function of the m^{th} order statistic $X_{(m)}$ is given by

$$g_m(x) = m\binom{n}{m} f(x)[1 - F(x)]^{(m-1)}[F(x)]^{(n-m)} \qquad \text{(B.1)}$$

Let $Y_{(m)}$ be a function of $X_{(m)}$, in the form,

$$Y_{(m)} = 1 - F(X_{(m)}) \qquad \text{(B.2)}$$

where F has the same meaning as above, $0 \leq Y_{(m)} \leq 1$ due to the fact that $0 \leq F(x) \leq 1$. Hence the probability density function of the random variable $Y_{(m)}$, $h_m(y)$, should be

$$h_m(y) = g_m(x) / \left| dy/dx \right| \tag{B.3}$$

where

$$\frac{dy}{dx} = \frac{d}{dx}(1 - F(x)) = -\frac{df(x)}{dx} = -f(x) \tag{B.4}$$

i.e.

$$h_m(y) = m\binom{n}{m} f(x)[1 - F(x)]^{m-1}[F(x)]^{n-m} \frac{1}{f(x)} = m\binom{n}{m} y^{m-1}(1-y)^{n-m} \tag{B.5}$$

where $0 \leq y \leq 1$, m= 1~n. In terms of the beta function, because

$$B(m, n-m+1) = [\Gamma(m)\Gamma(n-m+1)]/[\Gamma(n+1)] \tag{B.6}$$

and

$$B(m, n-m+1) = \int_0^1 z^{m-1}(1-z)^{n-m}\, dz \tag{B.7}$$

then, Eq.B.5 is reduced as

$$h_m(y) = [y^{m-1}(1-y)^{n-m}]/[B(m, n-m+1)] \tag{B.8}$$

The expectation of $Y_{(m)}$ is

$$E[Y_m] = \int_0^1 y h(y)\, dy \tag{B.9}$$

i.e.

$$E[Y_{(m)}] = 1/B(m, n-m+1)\int_0^1 y^m (1-y)^{n-m} dy \qquad \text{(B.10)}$$

Because of Eqs. B.6 and B.7, $E(Y_{(m)})$ should be

$$E[Y_m] = B(m+1, n-m+1)/B(m, n-m+1) = m/(n+1) \qquad \text{(B.11)}$$

Thus, we obtain the Weibull formula which is distribution-free that is widely used in engineering hydrology.

Based on the concept of expectation $E(X_{(m)})$ of the order statistic, Weibull formula can be also derived using Eq.B.1. In this case, probability density function of parent population, the uniform distribution, should be taken into consideration.

Hence Weibull formula can be derived from two procedures. The first procedure is based on the $Y_{(m)}$, the distribution of the probability of order statistics, and is to be distribution free of X. The second is based on the distribution of $X_{(m)}$ and related to the uniform distribution.