MODELING LONG TERM SERIAL CORRELATION IN FLOOD SERIES

YUDE LIN

Modeling Long Term Serial Correlation in Flood Series

by

©Yude Lin, B. Eng., M. S.

A thesis submitted to the School of Graduate
Studies in partial fulfillment of the
requirements for the degree of
Master of Engineering

Faculty of Engineering and Applied Science
Memorial University of Newfoundland

February 1993

St. John's Newfoundland Canada

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

ISBN 0-315-82597-9

TO MY PARENTS

AND

MY WIFE PEIYUN YANG

ABSTRACT

The standard assumption in flood frequency analysis is that annual peak flows are independent events. Hydrologists pay little attention to the serial correlation of flood peak series in flood frequency analysis because standard statistical tests for independence usually do not reject the null hypothesis of serial independence at the 90% or 95% confidence level for most rivers. This study was undertaken to investigate the validity of this assumption with regard to Canadian rivers by statistically analyzing the short-term and long-term dependence of annual peak flows. Ninety stationary flood series of Canadian rivers were tested by eleven tests for short-term dependence and one test for long-term dependence. The results show that about 18%(at 5% level) – 28%(at 10% level) flood series passed the short-term dependence tests but failed the long-term dependence test. The presence of long term serial correlation in flood series is evident and can not be disregarded in flood risk analysis.

Modeling annual flow series by considering the effect of both short-term and long-term dependence was then discussed. This study considered two newly developed models: harmonic analysis of cumulative departures, and the mixed-noise model. These two models were further developed and a comparison between them was made. Finally, the effect of long-term dependence on flood risk analysis is evaluated using Monte Carlo simulations. It was found that neglecting serial correlation could cause considerable uncertainty in the estimated flood risk.

ACKNOWLEDGEMENTS

In my pursuit of the Master degree in Engineering, I have received valuable advice and assistance from a number of people, these are very much appreciated. In particular, I would like to to thank:

(a) My supervisor Dr. L.M. Lye for his technical guidance on the various problems associated with this study and financial support throughout the program;

(b) Dr. J.J. Sharp, Dr. E. Moore, Dr. D. Hansen for their help and assistance in my studies and my other personal problems;

(c) The Faculty of Engineering for financial support through graduate assistantships;

(d) Staff in the Center for Computer Aided Engineering for their help in overcoming many difficulties related to computer work;

(e) My dear wife Peiyun Yang for her understanding and her moral support.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Review of Serial Correlation on Flood Risk Analysis

1.1.1 Serial Correlation of Annual Peak Flows

Many types of hydrologic time series exhibit significant serial correlation, that is, the value of the random variable under consideration at one time period is correlated with the values of the random variable at earlier time periods. But in most analysis of annual peak discharges for the determination of flood frequencies, annual peak flows are assumed to be serially independent events.

The consequences of assuming independent annual peak flows must be considered. Chow [1] cautioned that in actual hydrologic phenomena the variables are likely to be interdependent to an extent, and that the possibility of interdependence should be investigated. The U.S. Water Resources Council Guidelines [2] admitted that a relationship derived to predict future flood activity, if based upon nonrandom data, would have an increased degree of uncertainty. As cited by Carrigan and Huzzen [3], the specific effects could be to underestimate the confidence bands ascribed to a flood frequency distribution and to underestimate the population variance of a peak flow series by neglecting to consider serial correlation. If annual peak flows are dependent, errors in flood frequency analysis procedures due to the assumption of independence should be examined and quantified, and standard procedures revised to account for interdependence. Lye [4] also showed that taking into account the parameter uncertainty which is aggravated by serial correlation, significantly increases the risk associated with future peak flows.

Serial correlation analysis provides a measure of the degree to which the variables in a time series are interdependent. Yevjevich [5] performed serial correlation analysis on sequences of annual river flow and detected positive serial correlation on many rivers. Acknowledging the interdependence of annual flows, Carrigan and Huzzen [3] investigated the serial correlation of annual peak flows. In their analysis of records for 45 streamflow stations throughout the United States, the number of autocorrelation coefficients significantly different from zero for time lags of one and two years was greater than the expected number due to chance. Six of the 45 streams were judged to show signs of dependence. In an analysis of Australian streams by McMahon [6], approximately 17 percent of 133 streams analyzed showed dependence of annual peak discharges. But, for most rivers, standard statistical tests for dependence usually do not reject the null hypothesis of serial independence at the 90% or 95 % confidence level. Wall and Englot [7] concluded that, according to 5 independence tests, annual peak flows are independent for the 57 streams in Pennsylvania. That is why serial correlation is usually disregarded in flood frequency analysis. The obvious exceptions are rivers with large storage features in their drainage basins such as large lakes.

1.1.2 Long Term Persistence in Peak Flow Series

Serial correlation implies persistence. Long-term persistence [1] is the presence in a time series of significant dependence between observations a long time span apart. This is in contrast with the common assumption of time series analysis that observation separated by a long time span are independent or nearly so. Long term persistence may be characterized by the Hurst phenomenon and measured by the Hurst coefficient h.

The basic mathematical expression for the Hurst coefficient can be written as:

$$R_n/S_n \sim n^h \tag{1.1}$$

where R_n and S_n are the sample-adjusted range of cumulative departures from the arithmetical sample mean and the sample standard deviation, respectively, for a given time series of length n; the ratio R_n/S_n is specifically referred to as the adjusted rescaled range; and h denotes the Hurst coefficient. For some 900 geophysical time series, Hurst [8][9] observed that on average h=0.73. However, theoretical attempts to preserve this value of h in synthetic sequences failed as they led to an asymptotic value of 0.5. This discrepancy between the empirical and theoretical values is labelled as the Hurst phenomenon. It 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.

The point that is overlooked in statistical tests for dependence is that these tests are designed to show up only short term serial correlation. They are insensitive

[1] In this thesis, three words, serial correlation, persistence and dependence mean the same thing and are used interchangeably.

to the long term serial correlation structure of the data. This was demonstrated by Booy and Lye [10]. They analyzed the correlation structure of about fifty annual peak flow series from all over Canada, and found that while the high frequency aspects of the serial correlation structure, as measured by the lag-one serial correlation coefficient, are not significantly different from zero, the low frequency aspects as measured by the Hurst coefficient [8], is significant for many of these series. Similar findings for annual flow volumes and annual precipitation were reported by Srikanthan [11] and Hall et al. [12], respectively.

1.1.3 The Modeling of Hydrologic Time Series

For a normally distributed flood series with members that are independent or that follow a simple correlation structure, the distribution of the sample statistics can be obtained analytically in a flood frequency analysis. For flood series with a complicated serial correlation structure, it is difficult to use an analytical approach. One must resort to Monte Carlo techniques. A theoretical time series model that will reproduce the correct correlation structure of the peak flow series is required to obtain the distribution of the estimated parameters.

Many models exist for the modelling of hydrologic time series having a relatively high Hurst coefficient as well as a low lag-one serial correlation coefficient. The better known models are: the Fast Fractional Gaussian Noise process (Mandelbrot, [13]); the Broken Line process (Mejia, [14]); the ARMA(1,1) process (O'Connell, [15]); and the ARMA-Markov process (Lettenmaier and Burges, [16]). A simple and relatively efficient Mixed-Noise process for modelling the "mixed" behaviour of hydrologic series has also been developed (Lye, [4][10]). More recently, a harmonic analysis method developed by Sen [17] has also been shown to be capable of modelling the mixed behaviour.

Srikanthan [11] has compared the performance of some models for modelling annual flow volumes of Australian rivers. But there are a number of issues which have yet to be resolved for the two newly developed models, that is, harmonic analysis of the cumulative departures and Mixed-Noise model. Harmonic analysis method has not been fully developed. Sen [17] did not discuss the suitability of this method, how to model the residuals between the original and HCD (harmonic cumulative departures) curves, and some other problems. For the Mixed-Noise model, the theoretical properties of this model and its performance in relation to the other models have not been investigated in detail. Moreover, how to model skewed flood sequences using Mixed-Noise model is very important in practice and has yet to be developed.

1.1.4 The Effect of Long-Term Dependence on Flood Risk Analysis

The importance of low frequency behaviour on flood risk analysis was previously demonstrated by Booy and Morgan [18]. With a fractional noise model which models low frequency behaviour [13], they have shown that a degree of clustering of high flood years in the the record of annual flood peaks on the Red River in Winnipeg that, statistically speaking, is incompatible with the customary assumption of serial independence in flood frequency analysis. The return period for the flood protection of the City of Winnipeg and the towns in the Red River Valley was reduced to less than half the value estimated with the assumption of serial independence. Long-term fluctuations in soil moisture conditions were also shown to be related significantly to the observed low behaviour in the annual spring peak flows at Emerson, Manitoba (Booy and Lye, [19]). Because of the customary assumption that annual peak flows are serially independent, it is worth making an effort to extend this work to other river flood series and to the application of other models for the modeling long-term dependence in flood risk analysis. Monte Carlo method and mixed-noise model were used for this purpose. Bayesian analysis proposed by Lye et al.(1987, 1988)is another method which is able to evaluate the effect of long-term persistence on flood risk analysis [4][39].

1.2 Objective of Thesis

The first objective of this thesis is to investigate the long term serial correlation in annual peak flows of Canadian rivers. As mentioned above, the consequences of assuming independent annual peak flows should be considered and the statistical tests for dependence which are normally used are insensitive to the long term serial correlation structure of the series. Therefore, it is necessary to know if long term serial correlation exists in annual peak flows before we perform flood risk analysis.

The second objective of this study is to do further studies in two newly developed models for modeling hydrölogic time series: harmonic analysis of cumulative departures and mixed-noise model. These two models were designed to model long term dependence in flood series.

The third objective of this study is to evaluate the effect of long term persistance on the tolerance interval of estimated floods.

1.3 Outline of Thesis

The study begins with performing twelve statistical tests to investigate the serial correlation structure in annual peak flows of 90 selected Canadian rivers. This is given in Chapter Two. Further development on Sen's harmonic analysis method is done in Chapter Three. Chapter Four presents the application of the mixed-noise model in modelling "mixed" behaviour in annual floods, especially in modelling skewed flow series. In Chapter Five, based on mixed-noise model and Monte Carlo simulation, the effect of serial correlation in flood risk is analyzed. Chapter Six provides conclusions and recommendations from this study.

Chapter 2

SERIAL CORRELATION STRUCTURE OF ANNUAL PEAK FLOWS

2.1 General

Statistical frequency analysis assumes that the sample to be analyzed is a reliable set of measurements of independent random events from a stationary population. The validity of this assumption can be verified using statistical significance tests. However, most statistical tests of serial independence are designed to show up only short term serial correlation. They are insensitive to the long term serial correlation structure of the data which can be far more important.

To demonstrate this issue, the serial correlation structure of ninety peak flow series from Canadian rivers was analysed using eleven common tests for short-term dependence and Hurst's K test for long-term dependence. The calculated Hurst's K for each river was tested for statistical significance using bootstrapping and using a table of empirical percentage points developed based on normally distributed independent data.

In the next section, the details of each test of short-term dependence are given. Hurst's K test for long-term dependence is then considered. This is followed by analysis of the results, and conclusions.

2.2 Tests of Short-Term Dependence

In the statistical analysis of short-term dependence of the annual peak flows, the following eleven tests were applied to each time series. The first nine tests are non-parametric tests, the last two tests are parametric tests.

2.2.1 Non-parametric tests

(1) Median crossing test (Fisz,1963) [20]

X replaced by 0 if $x_i < \tilde{x}$ (median) and

X replaced by 1 if $x_i > \tilde{x}$

If the original sequence of X_s has been generated by a purely random process, then m, the number of times 0 is followed by 1 or 1 is followed by 0, is approximately normally distributed, i.e.

$$m \sim N\left(\frac{n-1}{2}, \sqrt{\frac{n-1}{4}}\right) \tag{2.1}$$

where n is the sample size.

(2) Turning points test (Kendall's test) [20]

Kendall's test(Kendall and Stuart, 1976) is also based on a binary series. If $x_{i-1} < x_i > x_{i+1}$ or $x_{i-1} > x_i < x_{i+1}$, then x_i is assigned the value 1; otherwise it is assumed to be 0. The number of 1's, m, is approximately normally distributed,ie,

$$m \sim N\left(\frac{2(n-2)}{3}, \sqrt{\frac{(16n-29)}{90}}\right) \tag{2.2}$$

(3) Length-of-runs test (Gold test, 1929) [20]

A run length s is defined by a set of s consecutive flows either above or below the median. If M_s denotes the total number of runs above and below the median of length s, then for a random process,

$$E(m_s) = \frac{[n+3-s]}{2^{s+1}} \tag{2.3}$$

and

$$\sum_{s=1}^{s'} [m_s - E(m_s)]^2 / E(m_s) \sim \chi^2(s'-1) \tag{2.4}$$

where s' is the maximum run length in the sequence.

(4) Rank difference test (Meacham test, 1968) [20]

Flows are replaced by their relative ranks R_i with the lowest being denoted by rank 1 (R_1).

The U statistic is calculated by:

$$U = \sum_{i=2}^{n} |R_i - R_{i-1}| \tag{2.5}$$

For large sample size n,

$$U \sim N\left(\frac{(n+1)(n-1)}{3}, \sqrt{\frac{(n-2)(n+1)(4n-7)}{90}}\right) \tag{2.6}$$

(5) Cumulative periodogram test (Box-Jenkins, 1970) [20]

The periodogram of a time series is defined as [29]:

$$I(f_j) = \frac{2}{n}[(\sum_{i=1}^{n} x_i \cos 2\pi i f_j)^2 + (\sum_{i=1}^{n} x_i \sin 2\pi i f_j)^2] \tag{2.7}$$

where $f_j = j/n$
j $=$ 1, 2, ..., $(n-2)/2$ for $n =$ even
$=$ 1, 2, ..., $(n-1)/2$ for $n =$ odd
The normalized cumulative periodogram is obtained from

$$C(f_j) = \sum_{i=1}^{j} I(f_i)/ns^2 \tag{2.8}$$

s^2=variance of x_i

for white noise points should fall $\pm K_\sigma/\sqrt{(n-2)/2}$ $n =$ even

$K_\alpha = 1.63(99\%), 1.36(95\%), 1.22(90\%), 1.02(75\%)$

(6) Wald-Wolfowitz test (Wald & Wolfowitz, 1943 [21]

For a sample of size n,

$$R = \sum_{i=1}^{n-1} x_i x_{i-1} + x_1 x_n \tag{2.9}$$

If the elements of the sample are independent,

$$R \sim N\left(\frac{s_1^2 - s_2}{n-1}, \sqrt{\frac{s_2^2 - s_4}{n-1} - (\frac{s_1^2 - s_2}{n-1})^2 + \frac{{s_1}^4 - 4s_1^2 s_2 + 4s_1 s_3 + s_2^2 - 2s_4}{(n-1)(n-2)}}\right) \tag{2.10}$$

where

$$s_r = x_1^r + x_2^r + ... + x_n^r \tag{2.11}$$

If mean is subtracted first, $s_1 = 0$, then

$$R \sim N\left(\frac{-S_2}{n-1}, \sqrt{\frac{S_2^1 - S_4}{n-1} - \frac{S_2^2}{(n-1)^2} + \frac{S_2^2 - 2S_4}{(n-1)(n-2)}}\right) \tag{2.12}$$

(7) Spearman rank order serial correlation coefficient test for dependence [22]

The series x_i is defined by the rank of Q_i (i = 1, ..., n-1)

the series y_i is defined by the rank of Q_i (i = 2, ..., n)

then

$$S_1 = \frac{1}{2} \frac{\sum x_i^2 + \sum y_i^2 - \sum d_i^2}{\sum x_i^2 \sum y_i^2} \tag{2.13}$$

where $\sum x_i^2 = (m^3 - m)/12 - \sum T_x$

$\sum y_i^2 = (m^3 - m)/12 - \sum T_y$

d_i = difference in rank between x_i and y_i

m = n-1

If no ranks are tied (can be assumed if only a small number), then

$$S_1 = 1 - \frac{6 \sum d_i^2}{m^3 - m} \tag{2.14}$$

For tied ranks, $T_x = (t^3 - t)/12$

where t is the number of observations tied at a given rank.

$\sum T_x$ and $\sum T_y$ are defined by extension of the foregoing.

For $n > 10$,

$S_1 \sim t(m-2)$ (one tail test)

$$t = S_1 \sqrt{\frac{m-2}{1 - S_1^2}} \tag{2.15}$$

(8) Runs above and below the median test for general randomness [22]

Data are ranked in chronological order. An A or B is assigned according to whether the corresponding data item is above or "below or equal" to the median.

The number of runs, RUNAB, is determined.

For n_1 A's and n_2 B's with n_1 and n_2 both greater than 20, the sampling distribution of RUNAB tends to normality with

$$z = \frac{|RUNAB - [(2n_1n_2)/(n_1+n_2)+1]|}{2n_1n_2(2n_1n_2 - n_1 - n_2)/[(n_1+n_2)^2(n_1+n_2-1)]^{1/2}} \qquad (2.16)$$

z is an N (0,1) variate and as used in this program, the region of rejection is

z greater than 1.96 for $\alpha = 0.05$

z greater than 1.645 for $\alpha = 0.10$

For n_1 and n_2 both less than 20, the region of rejection is defined by tables.

(9) Rank von Neumann Ratio [24]

Let $r_1, ..., r_n$ denote the ranks associated with the $x_i's$. The rank von Neumann ratio is given by

$$v = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{n(n^2-1)/12} \qquad (2.17)$$

Critical values of $c = [n(n^2-1)/12]v$ and approximate critical values of v are given by Madansky [24]. For large n, v is approximately distributed as $N(2, 4/n)$, though Bartels recommends $20/(5n+7)$ as a better approximation to the variance of v [24].

2.2.2 Parametric test

(1) Autocorrelation test [23]

Short-term dependence is usually measured by the magnitude of the low-order autocorrelation coefficients. In this thesis the autocorrelation function, r_k, is estimated using:

$$r_k = \left[\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})\right] \Big/ \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \tag{2.18}$$

where k = lag; x_i = annual flow at time i; n = sample size; and

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

The lag-one autocorrelation, r_1, is calculated from Eqn.(2.18) and is normally distributed:

$$r_1 \sim N\left(-\frac{1}{n}, \sqrt{\frac{n^3 - 3n^2 + 4}{n^2(n^2 - 1)}}\right) \tag{2.19}$$

r_1 is checked whether or not it is significantly different from the expected value.

(2) Von-Neumann ratio test [24]

Let

$$V = \frac{\sum_{t=2}^{n}(x_t - x_{t-1})^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \tag{2.20}$$

If data is independent, V is approximately normally distributed with $E(V) = 2$ and $Var(V) = 4(n-2)/(n^2-1)$, that is,

$$Z = \frac{V - 2}{\sqrt{4(n-2)/(n^2-1)}} \tag{2.21}$$

2.3 Long-term Persistence

Long-term dependence is measured by the magnitude of the Hurst coefficient. In this study, the Hurst coefficient is estimated by Hurst's K value since K has a lower variance than other estimators currently in use and its calculation is simple and straightforward. It has, however, a substantial bias in that it overestimates h for values below 0.70 and underestimates h for values over 0.70 (Wallis and Matalas, [25]). Hurst's estimator K (Hurst, [8]) is given by:

$$K = \frac{log(R/s)}{log(n/2)} \tag{2.22}$$

where, R is the range of cumulative departures from the mean, s is the standard deviation, and n the sample length. K is theoretically 0.5 for series of independent data. It increases when there is a greater degree of persistence, and it cannot exceed 1.0. The Hurst coefficient is presently the only measurement available for long-term dependence.

2.3.1 Empirical Percentage Points for Hurst's K

To test the significance of the calculated Hurst's K of a given time series, percentage points of Hurst's K for serially independent data at different probability levels are required. Statistical tables for testing the significance of Hurst's K are not conveniently available. Therefore, in this study Monte Carlo method was used to obtain empirical percentage points for Hurst's K. It is assumed that the null hypothesis is that the flood peak series is normally distributed and serially independent, and the alternate hypothesis is long-term dependence.

The procedure to obtain the empirical percentage points by Monte Carlo simulation is as follows:

1). Generate independent normally distributed data(mean of zero and standard deviation of one used here) of sample size n;

2). Calculate Hurst's K using Eqn.(2.22);

3). Repeat steps (1) and (2) 10,000 times;

4). Obtain the empirical percentage points based on the 10,000 K values.

The assumption of normality should not invalidate the test since Hurst's K is a very robust statistic (Mandelbrot and Wallis, [36]). This will be confirmed later using the bootstrap technique.

As mentioned earlier, bias exists in the estimation of the Hurst coefficient h by Hurst's K. However, since the formula used for calculating K is the same for both the sample and in developing the table of empirical percentage points this bias would not come into play here. Therefore it is reasonable to compare the K value of the flood series with those at given empirical percentage points. The calculation results of the empirical percentage points are shown in Fig. 2.1 and Fig. 2.2. For the convenience of the practising engineer, these percentage points are also shown in Tables 2.1 and 2.2 for sample sizes ranging from 20 to 200.

The test for long-term dependence is based on comparing the observed K value with that which could arise by chance alone from a series of normally distributed independent data. Therefore, if the K value of a flood series is greater than the K given in the Tables at a given significance level for a given sample size, it is concluded that this series is long-term dependent at this probability level. Otherwise it has no long-term dependence. The 5% and 10% levels are used in this study.

2.3.2 Bootstrap method for testing significance of Hurst's K

To check the assumption of using normally distributed data for testing Hurst's K, the non-parameteric bootstrap approach was used. Efron invented the bootstrap method (Efron, [40] [41] [42]) based on the fact that one available sample gives rise to many others. This method can be used here as a non-parametric test for long-term dependence. The bootstrap samples are generated from the data of the original sample as follows:

1. Suppose that the annual flow series x_1, x_2, ... , x_n are independent observations. Each data x_i has the same probability of occurrence which equals to $1/n$;

2. Generate a uniform random data i between 1 and n, then choose x_i as one point in the bootstrap sample. Repeat this step n times to generate a bootstrap sample of the same size n as the original sample;

3. Calculate Hurst's K for the bootstrap sample;

4. Repeat Steps (2) and (3) a large number of times (10,000 in this study);

5. Count the number of times the observed K value of the sample is exceeded by the 10,000 bootstrapped K values.

6. Calculate the p-value given by:

$$p - value = \frac{\#K > K_{obs}}{10000} \qquad (2.23)$$

Therefore, if the p-value is less than the specified significance level, it is concluded that the sample being tested is long-term dependent at the specified level. Otherwise it has no long-term dependence.

2.4 Analysis of the Annual Peak Flows of Canadian Rivers

The twelve tests for dependence described previously, eleven for short-term dependence and one for long-term dependence, were applied to selected unregulated Canadian rivers. The criteria established for data to be analyzed in this study were:

(1) At least 40 years of continuous record are available. This is because the calculation of Hurst's K will become unmeaningful when the length of annual flow series is too short.

(2) The annual peak flow series are from stationary populations. This is to focus the investigation of this paper upon only short-term and long-term dependence.

Data which met the above criteria are available from 90 Canadian rivers. Among these annual peak flow series, 12 are from Alberta, 13 from Atlantic provinces, 32 from B.C., 6 from Manitoba, 17 from Ontario, 5 from Quebec, 4 from Saskatchewan, 1 from Yukon. The 90 rivers range in record length from 40 to 80 years. The rivers analyzed, including length of record, mean of the data, Hurst's K and lag-one correlation, is presented in Table 2.3. It can be seen from the table that most of the rivers have small lag-one serial correlation coefficients, but many rivers have fairly high Hurst's K. Fig. 2.3 shows the distribution of Hurst's K for the ninety Canadian rivers analyzed.

Among the twelve tests used in this study, the parametric tests are designed for normally distributed data. Therefore, if the data is not normal, the Box-Cox transformation (Box & Cox, [43]) is used give approximately normal data. The

Box-Cox transformation is:

$$\begin{aligned} Y_t &= (X_t{}^{\lambda} - 1)/\lambda \quad \text{if } \lambda > 0 \\ Y_t &= ln(X_t) \qquad\quad\ \text{if } \lambda = 0 \end{aligned} \tag{2.24}$$

where Y_t are the transformed values, and λ is obtained using the simple Probability Plot Correlation Coefficient (PPCC) method (Lye, [30]).

For all of the tests in this study, their significance were tested at both five and ten percent levels because these two significance are often used in engineering practice. The results obtained for these two significance levels will be compared. The Table 2.3 shows that most of rivers has small short term serial correlation, but some rivers have high Hurst coefficient h. This difference can be also seen from the results of tests for dependence. (Next section provides more detailed analysis of these results.) The results are summarized in Tables 2.4 and 2.5 in terms of the number of rivers indicating dependence with respect to each of the twelve tests and the numbers of tests indicating dependence at the 5% and 10% levels of significance.

Appendix A provides the detailed test results and shows clearly the serial correlation structure in annual peak flows of 90 Canadian rivers. These results will be further analyzed in the next section.

2.5 Long-term Serial Correlation in Annual Peak Flows of Canadian Rivers

Because the different statistical tests for independence were designed under different assumptions and conditions, they do not have equal power in discriminating between time series which are not truly random; that is, the probability of a type-2 error is not the same for each test. The power of the tests depends somewhat on the nature of the dependence present, and on the length of record (Wallis and Matalas, [26]). Hence, sometimes the various tests employed give different conclusions for the same series. This means that a flood series can fail one test of independence but pass the other tests. This can be clearly seen in the results.

Because of this, it is difficult to say whether a flood series is independent or not just based on the result of one test. It is therefore a good idea to do several tests first before making any judgement. The question that arises is: How many failed tests are needed for a final judgement of short-term dependence? It is not easy to answer such a question. Wall and Englot [7] assumed in their study that at least two of the five tests applied to each data sequence should show signs of dependence. Similarly, Srikanthan et. al [20] assumed in their analysis that at least two of the six tests applied should indicate non-randomness. Therefore, it is reasonable to assume in this analysis that, for short-term dependence, at least four out of the eleven short-term dependence tests applied to each data sequence should indicate dependence. As for long-term dependence, because only one test available, we make our judgements based on the results Hurst's K test. The results of the parametric and non-parametric approaches for this test will be compared.

From the results obtained in Tables 2.4 and 2.5, the test results for short-term dependence and long-term dependence are summarized. This is shown in Table 2.6.

From Table 2.6 the following observations can be made:

1). For the parametric test of Hurst's K, about 17.8% (at 5% level) and 28.9% (at 10% level) of the tested series show long-term dependence. They are quite higher than the corresponding results of the short-term dependence tests, 1.1% (at 5%) and 8.9% (at 10%), respectively. The results from the non-parametric test of Hurst's K are similar to those from the parametric tests, concluding that about 15.6% (at 5% level) and 25.6% (at 10% level) of the tested series show long-term dependence. This means that significant long-term serial correlation structure of the annual peak flows exists in a large number of rivers and should not be ignored, otherwise mistakes will be made in our judgements. The effect of long-term serial correlation in flood risk analysis will be shown in Chapter Five;

2). It can be seen that the short-term dependence tests are insensitive to the long-term serial correlation structure of the data. Most of the series which has long-term dependence can pass most of the tests for short-term dependence. Based on the parametric tests, for example, about 17.8% (at 5% level) and 25.6% (at 10% level) of the tested series show that they are short-term independent but are long-term dependent at the same time;

3). For the ninety annual peak flows tested, the conditional probabilities of the existence of long-term dependence when the series has passed the short-term dependence tests are as follows (based on parametric test results):

At the 5 % level:

P(long-term dep./short-term indep.) = 16/89 * 100% = 18.0%

At the 10 % level:

P(long-term dep./short-term indep.) = 23/82 * 100% = 28.0%

The probabilities are quite high and we have no reason to disregard them. Hence, if a series shows that it is short-term independent, we should still investigate for long-term dependence.

2.6 Summary

The serial correlation structure of annual flood series from ninety Canadian rivers were analysed in this chapter. It was found that significant long-term serial correlation as measured by the Hurst K statistic is present in a large number of rivers. It was found that when a peak flow series shows short-term independence, there is still a fairly high probability of long-term dependence. This long-term dependence cannot be disregarded as in traditional flood frequency analysis; it should be taken into account as this may significantly increase the risk associated with future peak flows.

In the next two chapters, the modeling of long-term serial correlation in annual peak flows is discussed.

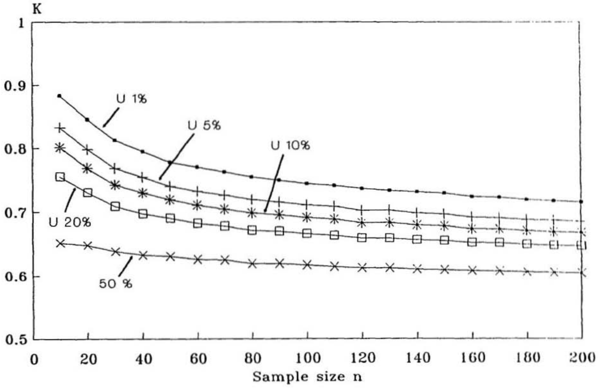


Figure 2.1: Empirical Percentage Points of K for Normally Distributed Independent Data ($n <= 200$)

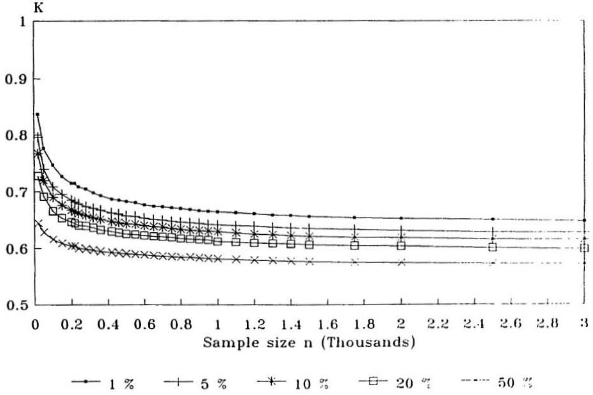


Figure 2.2: Empirical Percentage Points of K for Normally Distributed Independent Data($n <= 3000$)

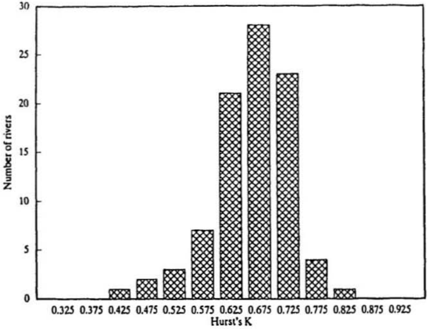


Figure 2.3: Distribution of Hurst's K for Canadian rivers

Table 2.1: Empirical Percentage Points of K for Independent Data (n: 20 - 50)

Sample	*Different Levels*				
Size	1 %	5 %	10 %	20 %	50 %
20	0.8370	0.7961	0.7672	0.7277	0.6449
21	0.8334	0.7930	0.7647	0.7257	0.6442
22	0.8300	0.7900	0.7623	0.7237	0.6435
23	0.8268	0.7871	0.7599	0.7219	0.6428
24	0.8236	0.7844	0.7577	0.7201	0.6422
25	0.8207	0.7817	0.7555	0.7185	0.6415
26	0.8178	0.7792	0.7535	0.7168	0.6409
27	0.8151	0.7767	0.7515	0.7153	0.6403
28	0.8125	0.7744	0.7496	0.7138	0.6397
29	0.8100	0.7721	0.7477	0.7124	0.6391
30	0.8076	0.7700	0.7459	0.7110	0.6385
31	0.8053	0.7679	0.7442	0.7097	0.6379
32	0.8031	0.7659	0.7425	0.7084	0.6374
33	0.8010	0.7640	0.7409	0.7072	0.6368
34	0.7990	0.7621	0.7394	0.7060	0.6362
35	0.7971	0.7603	0.7378	0.7048	0.6357
36	0.7953	0.7586	0.7364	0.7037	0.6352
37	0.7935	0.7570	0.7350	0.7026	0.6346
38	0.7919	0.7554	0.7336	0.7016	0.6341
39	0.7903	0.7538	0.7323	0.7005	0.6336
40	0.7887	0.7524	0.7310	0.6995	0.6331
41	0.7873	0.7510	0.7297	0.6985	0.6326
42	0.7859	0.7496	0.7285	0.6975	0.6322
43	0.7845	0.7483	0.7273	0.6966	0.6317
44	0.7832	0.7470	0.7262	0.6957	0.6312
45	0.7820	0.7457	0.7250	0.6948	0.6308
46	0.7808	0.7445	0.7239	0.6939	0.6303
47	0.7797	0.7434	0.7229	0.6930	0.6299
48	0.7786	0.7423	0.7219	0.6922	0.6295
49	0.7775	0.7412	0.7208	0.6913	0.6291
50	0.7765	0.7401	0.7199	0.6905	0.6287

Table 2.2: Empirical Percentage Points of K for Independent Data (n: 55-200)

Sample	*Different Levels*				
Size	1%	5%	10%	20%	50%
55	0.7720	0.7353	0.7153	0.6867	0.6267
60	0.7682	0.7311	0.7113	0.6833	0.6249
65	0.7649	0.7274	0.7077	0.6802	0.6233
70	0.7620	0.7241	0.7044	0.6775	0.6219
75	0.7593	0.7211	0.7015	0.6750	0.6206
80	0.7567	0.7184	0.6988	0.6728	0.6194
85	0.7542	0.7158	0.6963	0.6708	0.6183
90	0.7518	0.7135	0.6941	0.6690	0.6173
95	0.7494	0.7114	0.6920	0.6673	0.6164
100	0.7470	0.7094	0.6900	0.6657	0.6155
110	0.7422	0.7058	0.6866	0.6627	0.6140
120	0.7377	0.7028	0.6836	0.6601	0.6126
130	0.7336	0.7000	0.6809	0.6578	0.6113
140	0.7301	0.6976	0.6785	0.6559	0.6101
150	0.7270	0.6951	0.6764	0.6543	0.6089
160	0.7245	0.6926	0.6745	0.6527	0.6077
170	0.7223	0.6900	0.6727	0.6510	0.6066
180	0.7201	0.6876	0.6711	0.6492	0.6056
190	0.7177	0.6857	0.6696	0.6475	0.6047
200	0.7148	0.6848	0.6682	0.6468	0.6043

Table 2.3: Canadian Rivers Analyzed

River Name	*Province*	*n(yrs)*	*Mean*	*Hurst's K*	r(1)
Athabasca At Athabasca	Alberta	47	2236.19	0.514	-0.190
Bow	Alberta	80	217.61	0.651	-0.132
Castle Near Beaver Mines	Alberta	44	145.96	0.669	0.001
Drywood Creek*	Alberta	52	6.57	0.688	0.020
Elbow Above Glenmore Dam	Alberta	44	63.25	0.668	-0.061
Elbow At Bragg Creek	Alberta	54	60.25	0.680	-0.047
Ghost	Alberta	40	22.66	0.673	0.230
Manyberries Creek	Alberta	45	13.12	0.693	0.025
Rolph Creek**	Alberta	53	4.96	0.720	0.078
Sturgeon	Alberta	54	26.77	0.526	-0.171
Swiftcurrent Creek	Alberta	54	28.81	0.583	0.030
Waterton	Alberta	41	142.33	0.665	0.032
Upper Humber	Atlantic	60	578.33	0.663	0.208
Lepreau	Atlantic	72	78.75	0.522	-0.021
Saint John*	Atlantic	62	2357.76	0.724	0.150
Shogomoc Stream	Atlantic	45	39.63	0.630	0.046
Upsalguitch	Atlantic	45	367.29	0.643	0.031
Beaverbank*	Atlantic	67	29.52	0.725	-0.096
East	Atlantic	63	8.00	0.696	-0.140
Grand*	Atlantic	68	18.94	0.700	-0.064
Lahave	Atlantic	73	230.36	0.708	-0.040
Northeast Margaree**	Atlantic	72	176.31	0.732	0.070
Roseway**	Atlantic	71	68.61	0.739	0.083
Southwest Margaree**	Atlantic	70	38.68	0.756	0.138
St. Marys(Stillwater)	Atlantic	73	408.62	0.702	-0.047
Adams	B.C.	42	246.17	0.632	0.171
Ashnola	B.C.	42	83.27	0.614	-0.320
Babine	B.C.	41	125.31	0.697	0.075
Big Sheep Creek	B.C.	40	48.81	0.645	0.055
Boundary Creek**	B.C.	61	47.03	0.746	0.119
Bulkley	B.C.	58	587.07	0.630	0.175
Chilko At Outlet of Lake**	B.C.	60	136.8	0.744	-0.033
Chilko Near Redstone	B.C.	62	300.34	0.607	-0.182
Columbia At Nicholson**	B.C.	77	437.60	0.746	-0.077
Columbia At Donald	B.C.	44	712.64	0.686	-0.283
Columbia Near Fairmont Hot	B.C.	43	45.99	0.668	-0.292

continued

River Name	*Province*	*n(yrs)*	*Mean*	*Hurst's K*	*r*(1)
Flathead #*	B.C.	60	208.53	0.785	0.187
Kettle Near Ferry**	B.C.	60	339.33	0.738	0.133
Kettle Near Laurier	B.C.	59	591.36	0.710	0.067
Kootenay At Kootenay Crossing	B.C.	41	33.71	0.564	-0.093
Kootenay At Newgate**	B.C.	42	1614.55	0.774	0.065
Lardeau At Marblehead	B.C.	43	282.93	0.648	-0.218
Liard	B.C.	42	5370.71	0.666	0.200
Lillooet	B.C.	63	529.63	0.640	0.096
Moyie**	B.C.	59	145.90	0.848	0.132
Quesnel At Likely*	B.C.	64	394.86	0.701	0.130
Quesnel Near Quesnel**	B.C.	50	766.84	0.723	0.124
Salmo Near Salmo	B.C.	40	243.23	0.593	0.075
Sikanni Chief	B.C.	44	198.84	0.425	-0.090
Similkameen	B.C.	44	236.95	0.691	0.113
Skeena	B.C.	41	5053.90	0.617	-0.247
Slocan*	B.C.	64	441.77	0.706	0.075
South Thompson	B.C.	48	996.17	0.670	0.132
St. Mary At Wycliffe	B.C.	43	385.28	0.680	-0.047
St. Mary Near Marysville	B.C.	41	303.51	0.609	-0.110
Stuart#*	B.C.	56	322.02	0.751	0.218
North Thompson*	B.C.	44	1775	0.723	0.021
Brokenhead	Manitoba	46	36.17	0.638	0.026
Roseau Near Cariboun	Manitoba	67	47.46	0.663	0.199
Roseau Near Dominion	Manitoba	49	64.54	0.557	0.038
Sprague Creek	Manitoba	43	19.72	0.657	0.104
Turtle Near Laurier	Manitoba	40	51.22	0.622	0.106
Whitemouth	Manitoba	42	83.68	0.677	-0.100
Ausable	Ontario	43	180.27	0.488	-0.131
Black**	Ontario	73	129.49	0.731	0.112
Castor	Ontario	41	107.14	0.716	0.251
English At Umfreville*	Ontario	67	158.59	0.700	-0.136
English (Sioux Lookout)**	Ontario	60	287.28	0.736	0.003
Missinaibi**	Ontario	69	880.99	0.729	0.106
Namakan*	Ontario	66	319.47	0.705	0.154
Nith Near Canning	Ontario	42	188.85	0.682	-0.044

continued

River Name	*Province*	*n(yrs)*	*Mean*	*Hurst's K*	*r*(1)
North Magnetawan	Ontario	73	44.47	0.615	-0.012
Nottawasaga	Ontario	40	110.11	0.703	0.156
Pigeon	Ontario	65	128.53	0.691	-0.007
Saugeen Near Port Elgin	Ontario	74	500.45	0.627	-0.006
Saugeen Near Walkerton	Ontario	74	290..39	0.674	0.193
South Nation	Ontario	41	47.41	0.621	0.004
Sydenham Near Alvinston	Ontario	40	101.95	0.649	-0.069
Sydenham Near Owen Sound	Ontario	43	30.31	0.592	0.020
Turtle Near Mine Center	Ontario	58	127.24	0.668	-0.059
Hall (Riviere)	Quebec	40	67.82	0.735	-0.026
Harricana (Riviere)	Quebec	56	190.13	0.483	-0.164
Petite Nation a Portage-	Quebec	46	131.38	0.681	0.028
Petite Nation Pres De Cote-	Quebec	43	69.51	0.572	0.035
Richelieu Aux Rapides F.	Quebec	51	923.84	0.633	0.178
Horse Creek	Sask.	43	8.88	0.606	-0.094
McEachern Creek	Sask.	53	22.57	0.598	-0.049
Poplar	Sask.	56	24.94	0.645	-0.049
Whitewater Creek	Sask.	53	9.76	0.650	-0.023
Teslin	Yukon	41	1052	0.663	-0.050
Mean		53		0.664	0.015
Standard Deviation		12		0.070	0.125

Note:

1. Short-term independent but Hurst's K significant at 10% (*);
2. Short-term independent but Hurst's K significant at both 10% and 5% (**);
3. Short-term dependent at 10% only, Hurst's K significant at both 10% and 5% (#*).

Table 2.4: Dependence as a Function of Test

Tests	Percentage (and No.) of records indicating dependence at 5% level	Percentage (and No.) of records indicating dependence at 10% level
Median crossing	2.22(2)	8.89(8)
Turning points	3.33(3)	7.78(7)
Length-of-runs	11.11(10)	15.56(14)
Rank difference	4.44(4)	10.00(9)
Cumulative periodogram	0.00(0)	2.22(2)
Wald-Wolfowitz	0.00(0)	10.00(9)
Spearman	0.00(0)	5.56(5)
RUNAB(random)	5.56(5)	11.11(10)
Rank von Neumann	4.44(4)	11.11(10)
Autocorrelation	0(0)	10.00(9)
Von-Neumann	4.44(4)	7.78(7)
Hurst's K test		
Parametric	17.8(16)	28.9(26)
Non-parametric (Bootstrap)	16.7(15)	25.6 (23)

Table 2.5: Number of Rivers Indicating Dependence

Number of Tests	Percentage(and number) of rivers	
Indicating Dependence	5%	10%
11	0.00(0)	1.11(1)
10	0.00(0)	0.00(0)
9	0.00(0)	0.00(0)
8	0.00(0)	1.11(1)
7	0.00(0)	1.11(1)
6	0.00(0)	1.11(1)
5	1.11(1)	2.22(2)
4	1.11(1)	4.44(4)
3	3.33(3)	6.67(6)
2	7.78(7)	10.00(9)
1	17.78(16)	24.44(22)
0	68.89(62)	47.78(43)

* Hurst's K test is based on normally distributed data.

Table 2.6: Comparison of Short-term and Long-term Dependence

	5 % Level		*10 % Level*	
	No. of rivers	*Percentage*	*No. of rivers*	*Percentage*
Short-term dependence	1	1.1%	8	8.9%
Long-term dependence:				
a. Parametric test	16	17.8%	26	28.9%
b. Bootstrap method	15	16.7%	23	25.6%
Only short-term dependence	1	1.1%	5	5.6%
Only long-term dependence				
a. Parametric test	16	17.8%	23	25.6%
b. Bootstrap method	15	16.7%	21	23.3%
Both short- & long-term dep.	0	0%	3	3.33%

Chapter 3

HARMONIC ANALYSIS OF CUMULATIVE DEPARTURES

3.1 General

In previous chapters, the statistical tests of short-term and long-term dependence were used for the flood peak series of 90 Canadian rivers. The results show that the effect of Hurst phenomenon on annual peak flows is evident. Many flood peak series have high Hurst coefficients. Because the corresponding increase in flood risk due to parameter uncertainty can be substantial if the flood peak series have a high Hurst coefficient (Booy and Lye, [10]), how to model long-term persistence (i.e. Hurst's K) in flood peak series is important in flood risk analysis.

As mentioned in Chapter One, a theoretical time series model that will reproduce the correct correlation structure of the peak flow series is required to obtain the distribution of the estimated parameters of a flood distribution. There are several such methods available for modeling series with high Hurst's K and low $\rho(1)$ in flood peak series. For example, ARMA(1,1) model [16], the Broken Line model [14], the Fast Fractional Gaussian Noise process [13], etc. Two recently developed methods are Mixed Noise process (Lye, [4][10]) and the harmonic analysis of the cumulative departures approach(Sen, [17]), and both methods have not been fully developed. In this chapter, some research work is done to investigate the existing problems in the method of harmonic analysis, such as the modeling procedure, the number of harmonics which should be used in the analysis, how to fit a suitable stochastic model of the residuals, and normal transformation of the skew original series. A comparison between Mixed Noise process and harmonic analysis method will be shown in the next chapter.

3.2 Sen's Method

To obtain a mathematical model of the cumulative departures that would preserve the Hurst phenomenon, Sen [17] has performed the harmonic analysis of the cumulative departures of annual flow series from their sample mean values. In his paper, Sen tried to explain the Hurst phenomenon on the basis of the storage-related processes, the sole representation of which is the historic cumulative departures curve. Sen considered seven annual flow series from Europe and the U.S.A. as an example of the proposed procedure. The characteristics of these series are summarized in Table 3.1.

It has been observed that, even though the original annual flow time series is stationary, the cumulative departures curve exhibits strong periodicities with the slowest cycles having periods equal to the total sample length. The cyclic features account for more than 95% of the variability in the Hurst coefficients. However, in the classical simulation studies of the original series which are stationary, such periodicities in the cumulative departures are not considered. So Sen suggested an alternative to direct simulation of annual sequences for preserving the Hurst phenomenon. The suggested procedure is as follow:

(1) Construct the original cumulative departures curve from the observed time series of the hydrological variable.

(2) Apply harmonic analysis to this curve by depicting the first seven to eight harmonic components with fundamental frequency of $1/n$.

(3) Construct the harmonic cumulative departures (HCD) curve from the harmonic components obtained in the previous step.

(4) Find the residuals between the original and HCD curves.

(5) Fit a suitable stochastic model of the residuals.

(6) Generate the synthetic sequence of residuals and add them to the HCD's curve.

Sen [17] presented the results based on the above procedure for seven annual flow series from Europe and the U.S.A., the characteristics of which are reproduced here in the correct units in Table 3.1. It must be pointed out that Sen considered only steps (1) to (3) in his paper. Steps (4) to (6) would be have to be carried out, however, for the generation of synthetic sequences.

In applying Sen's procedure for the generation of synthetic sequences, several problems were encountered. Briefly, it was found that using the first seven to eight harmonic components for fitting the historical cumulative departures curve gave residuals that were anti-persistent and in general require a high order ARMA process to model them. This means that the resulting model based on Sen's procedure would require far too many parameters. In addition, modelling skewed series cannot be handled easily.

In light of the stated problems, this study was conducted to shed more light on the applicability of Sen's procedure for modelling time series. In particular, this study deals with the following:

a) the optimal number of harmonics to be used to obtain the harmonic cumulative departures (HCD),

b) the kind of time series that this method is most suited,

c) the characteristics of the residuals that are most amenable to stochastic modelling, and

d) how skewed time series may be modelled.

In the following section, Sen's method is described and errors in Sen's paper are corrected. This will be followed by a discussion of the stochastic modelling of the residuals using ARMA models. The comparison of the results obtained using different number of harmonics, and conclusions from the study are then presented.

3.3 Cumulative Departures Curve and Harmonic Components for 12 Rivers

3.3.1 Calculation Procedure

In general, the time series of the cumulative departures, S_i, can be represented as:

$$S_i = S_{i-1} + (x_i - \overline{x}) \tag{3.1}$$

where: $i = 1, 2, ..., n$; S_0 and $S_n = 0$; x_i is the original time series which is considered as an input into the reservoir and $\overline{x}$ is its sample mean value which is assumed to be its output. For a given sample size n, Eqn.(3.1) represents a stochastic process with a mean $\overline{S}$ and variance σ_s^2 which can be shown to be given by (G. Sabin, personal communication):

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i = \frac{1}{2n}\sum_{i=1}^{n}(n+1-2i)x_i \tag{3.2}$$

$$\sigma_s^2 = \frac{n^2-1}{12}\bar{x}^2 - \frac{1}{n^2}\sum_{j=1}^{n-1} j(n-j)\sum_{i=1}^{n-j} x_i x_{i+j} \tag{3.3}$$

The values of the mean and variance were incorrectly stated in Sen [17]. Hurst's estimate of the Hurst coefficient h is then obtained from the S_i's and X_i's:

$$K = \frac{log(R/s)}{log(n/2)} \tag{3.4}$$

where: K = Hurst's K; s equals $n^{-1/2}[\sum_{i=1}^{n}(x_i - \overline{x})^2]^{1/2}$, the sample standard deviation; R is the adjusted range and is defined as

$$R = M_n - m_n \tag{3.5}$$

where, M_n equals max $(0, S_1, S_2, ..., S_n)$, and m_n equals min $(0, S_1, S_2, ..., S_n)$.

The cumulative process in Eqn. (3.1) can be represented in two parts, namely, nonstationary and stationary as follows:

$$S_i = P_i + e_i \tag{3.6}$$

where P_i is the periodic component at time instant i and e_i is a stationary zero mean process. Thus, e_i denotes the noise part in the cumulative departures process. Equation(3.6) can be written with its periodic part explicitly as:

$$S_i = \overline{S} + \sum_{k=1}^{m}[A_k sin(\gamma ki) + B_k cos(\gamma ki)] + e_i \tag{3.7}$$

where $\overline{S}$ is the mean of cumulative departures given by Eqn. (3.2). m is the number of significant harmonics; $\gamma = 2\pi/n$ is the cyclic frequency over a base period; A_k and B_k are harmonic coefficients.

In the cumulative departures process the fundamental period can be adopted as equal to the sample length, and therefore, the fundamental frequency is $1/n$. Estimation of harmonic coefficients is achieved by conventional Fourier analysis. Suppose that the number of observations $n = 2m + 1$ is odd, then the least squares estimates of $\overline{S}$ is given by (3.2) and the harmonic coefficients A_k and B_k will be:

$$A_k = \frac{2}{n}\sum_{i=1}^{n} S_i sin(\gamma ki) \tag{3.8}$$

and:

$$B_k = \frac{2}{n}\sum_{i=1}^{n} S_i cos(\gamma ki) \tag{3.9}$$

where $k = 1, 2, ..., m$.

Note: The equations for the Fourier coefficients given in Sen [17] are incorrect.

The periodogram then consists of the $k = (n-1)/2$ values

$$I(f_k) = \frac{n}{2}(A_k^2 + B_k^2) \tag{3.10}$$

where $I(f_k)$ is called the *intensity* at frequency f_k.

When n is even, then set $n = 2m$ and equations (3.2), (3.8) and (3.9), apply for $k = 1, 2, ..., (n-1)$ but

$$B_m = \frac{1}{n}\sum_{i=1}^{n}(-1)^i S_i \tag{3.11}$$

$$A_m = 0 \tag{3.12}$$

and $I(f_m) = nB_m^2$.

The periodogram $I(f_k)$ is also the "sum of squares" associated with the pair of coefficients (A_k, B_k), and hence with the frequency $f_k = k/n$ (Box & Jenkens,

[29]). Thus the proportion of variance, V_k, explained by the k^{th} harmonic can be computed from:

$$\begin{aligned} V_k &= I(f_k) \Big/ \sum_{k=1}^{m} I(f_k) \\ &= I(f_k) \Big/ \sum_{i=1}^{n} (S_i - \overline{S})^2 \end{aligned} \tag{3.13}$$

A plot of V_k against k would give a good pictorial representation of the contribution of the k^{th} harmonic to the explained variance of S_i.

Fisher's test of significance for the intensity is applied in this study [27][28]. Let

$$I_{j1} = max\{I(f_1), I(f_2), ..., I(f_k)\} = I(f_{j1})$$

Suppose that e_i is white noise, the statistic

$$g_i = I_{j1} / \sum_{j=1}^{k} I(f_j) \tag{3.14}$$

follows *Fisher* distribution,

$$p\{g > g_1\} = \sum_{j=0}^{r} (-1)^j C_K^{j+1} [1 - (j+1)g_1]^{K-1} \tag{3.15}$$

where r is the greatest positive integer that make $1 - (r+1)g_1 > 0$.

Given the probability level α, say α=0.005, if

$$p\{g > g_1\} \geq \alpha$$

we conclude that no period item $p(t)$ exists. Otherwise, if $p\{g > g_1\} > \alpha$, we accept that $T_1 = N/j_1$ as the first period item in the original time series.

Let I_{jk} be the k_{th} greatest value, the statistic

$$g_k = \frac{I(f_k)}{\sum_{j=1}^{K} I(f_j)} \tag{3.16}$$

follows *Fisher* distribution

$$p\{g > g_k\} = C_K^{k-1} \sum_{j=0}^{r} C_{K-k+1}^{j+1} \frac{j+1}{j+k}[1-(j+k)g_k]^{K-1} \tag{3.17}$$

where r is the greatest positive integer that let $1-(r+k)g_k > 0$. Given probability level α, if

$$p\{g > g_k\} > \alpha$$

$T_k = N/j_k$ is accepted as a period of the original time series.

It is suggested to use a $\alpha \leq 0.01$, otherwise the probability of selecting some false periods exist [28]. Hence, a α value of 0.005 is used in this study.

3.3.2 Application

Sen [17] suggested that the first 7 to 8 harmonics be used for the harmonic analysis, although many rivers require only the first few harmonics. To investigate the number of harmonics that will contribute significantly to the overall variability of the cumulative departures, different number of harmonic components are considered in this study. The harmonic components used are:

a. Using significant harmonics which are selected by Fisher's probability at $\alpha = 0.005$;

b. 1st harmonic;

c. First 3 harmonics;

d. First 4 harmonics; and

e. First 8 harmonics.

The results are shown in Table 3.2 and Figs 3.1 - 3.5. Table 3.2 shows: the river modelled, the sample size n, Hurst's K and lag-one correlation $r(1)$, whether the data are normally distributed, the K values of the residuals after fitting using significant harmonics, only the first, 3, 4, and 8 harmonics. Figs 3.1-3.5 show for the Thames River the result of fitting the cumulative departure curve using different number of harmonics and the resulting residual series.

From these results, we can see that:

(1). The more harmonics we use, the lower the Hurst's K of the residuals we get. If we use the first harmonic only, the Hurst K of residuals is still very high (sometimes it is higher than the K of original series). That means the function of harmonic analysis here is limited and we have the same difficulty in fitting a suitable stochastic model of the residuals. If we use the first eight harmonics (suggested number by Sen), the K of the residuals is lower than 0.5 which implies anti-persistent behaviour [38]. In the case of antipersistence, an increasing trend in the past implies a decreasing trend in the future, and a decreasing trend in the past makes an increasing trend in the future probable. The process appears very 'noisy' and it is difficult to model. A reasonable number of harmonics to use seems to be 3 or 4, where Hurst's K of the residuals shows no long-term dependence.

(2). For most of the rivers, the harmonic analysis selected many significant harmonics, which means there would be many parameters in our final models if we use significant harmonics. Moreover, selecting too many significant harmonics causes anti-persistent behaviour in the residuals (K lower than 0.5). The only exceptance of this case is the Thames river which has a high Hurst's K and a low r(1). It looks like that using significant harmonics is suitable for a time series with this kind of structure.

To confirm the above conclusion(2), five annual peak flow time series of Canadian rivers with high Hurst's K and low r(1) were selected. The results of similar calculation for these data are shown in Table 3.3.

For these five Canadian rivers with high Hurst's K and low r(1), the numbers of their significant harmonics are few as expected, from 2 to 4. Because they do not have many significant harmonics, anti-persistent behaviour does not exist in their residuals. The K values of the residuals of these rivers range from 0.5 to 0.67. Anti-persistent behaviour exists in both the residuals of using four and eight harmonics. The K values of residuals of using only first harmonics are still high, like those in Table 3.2. Therefore, from the above results, it is reasonable to suggest that using significant harmonics or first three harmonics for those annual peak flow series with high Hurst's K and low r(1).

It should be noted that, although the residuals which have low Hurst's K indicates no long-term persistence is present, it is still possible that short-term persistence is present. Actually, using the tests for dependence in Chapter Two to test the residuals, it was found that most of them were short-term persistent. They are not independent data.

3.4 Stochastic Modeling of Residuals

After obtaining the harmonic cumulative departure (HCD) curves, the problem now is to find a suitable stochastic model for the residuals between the original and HCD curves. Since the residuals are stationary, the ARMA(p,q) class of models (Box & Jenkins, [29]) are used in this study as such stochastic models. The ARMA(p,q) model can be represented in a single equation as

$$\phi(B)(1-B)Y_t = \theta(B)a_t \tag{3.18}$$

where: B is the backshift operator, $\phi(B)$ is the autoregressive (AR) process, $\theta(B)$ is the moving average (MA) process, a_t is the white noise term, and Y_t is the time series modelled.

The main problem here is in trying to decide *which* ARMA(p,q) fits the data best, i.e. in identifying the AR order p and the MA order q. Much of Box and Jenkins is devoted to this so-called "identification" problem. This study identifies the model by considering the following three conditions:

1). Minimizing the Bayesian Information Criterion (BIC) [32]:

$$BIC = -2logL_{(p+q)} + log(n)(p+q) \tag{3.19}$$

where n is the sample size, $(p+q)$ is the number of parameters to be fitted and L is the maximum value of the likelihood function for a $(p+q)$ parameter model.

Actually, AIC is another criterion which is also most commonly employed in model selection. The reasons for prefering BIC to AIC in this study is: BIC is strongly consistent in that it determine the true model asymptotically, whereas for AIC, an overparameterised model will always emerge no matter how long the available realization. Thus it would be appear that the BIC should be used in preference to AIC [33].

2). AR(p) model should also meet the stationary conditions[29]. These conditions are satisfied if the roots (u) of the characteristic equation

$$u^p - \phi_1 u^{p-1} - \phi_2 u^{p-2} - ... - \phi_p = 0 \tag{3.20}$$

lie inside the unit circle. For AR(1) model, the stationary condition becomes:

$$-1 < \phi_1 < 1 \tag{3.21}$$

The stationary condition for AR(2) model is:

$$\phi_1 + \phi_2 < 1$$
$$\phi_2 - \phi_1 < 1 \tag{3.22}$$
$$-1 < \phi_2 < 1$$

3). If

$$\phi_1^2 + 4\phi_2 < 0$$

the second-order difference equation satisfied by the autocorrelation function has complex roots ([29], pp59). This will probably worsen the simulation results, especially when

$$\phi_1^2 + 4\phi_2 < -1.0$$

(from computation experience). In such case, we can try another ARMA model with a higher value of $\phi_1^2 + 4\phi_2$ and a BIC value which is close to the minimum BIC value.

The results are shown in Table 3.4.

3.5 Monte Carlo Simulation For the Comparison of Different Models

3.5.1 Monte Carlo Simulation

Having determined the number of harmonics to represent the cumulative departures and the best ARMA model for the residuals, the next step is the generation of synthetic sequences. If the model is correct, the model should on average be able to reproduce the marginal distribution parameters and the serial correlation structure of the observed time series. That is, the mean, standard deviation and skewness, and both Hurst's K and the lag-one correlation coefficient r(1) should on average be preserved. Since one of the main purpose of this study is to investigate the most suitable number of harmonics to use, the P_i term in Eqn. (3.6) is modelled using the different number of harmonics and its corresponding ARMA model for the residuals. The Monte Carlo method was used and the number of replication used in the Monte Carlo simulation study was 3000. The results are shown in Table 3.5-3.9 for Hurst's K, the lag-one correlation coefficient r(1), mean, standard deviation, and the coefficient of skewness, respectively.

The results show that aside from the coefficient of skewness, the other parameters especially the Hurst K, r(1), and the mean are fairly well reproduced regardless of the number of harmonics and corresponding ARMA model used. The results for the standard deviation is somewhat erratic for some rivers.

From Table 3.8, one can see that the coefficient of skewness are preserved only for those series that are approximately normally distributed or have skew coefficients close to zero. One method of overcoming this problem is to transformed the skewed series to one that is approximately distributed and then do the harmonic analysis and simulation using the transformed series. To recover the original skewed series, the inverse transformation is applied. A convenient transformation to use is the Box-Cox transformation (Box-Cox, [43]). It is given by Eqn.(2.24) in Chapter Two. Similarly, the λ value is obtained using the simple Probability Plot Correlation Coefficient (PPCC) method (Lye, [30]). The simulation results are shown in Table 3.5 - 3.9.

3.5.2 Discussion

Suitability of Sen's Method If Using Significant Harmonics

The Thames River and five Canadian rivers have high Hurst's K and low r(1). The method suggested by Sen seems good for such rivers. There are not too many parameters and from the simulation results the main parameters except the skewness of the original series, Hurst's K, r(1), mean and standard deviation are quite well reproduced.

For those rivers that do not have simultaneously high Hurst's K and low r(1), the simulation results are also reasonable, but many significant harmonics are used resulting in too many parameters in the model. Moreover, simple ARMA models do not always work for the residual series.

The Number of Harmonic Components

Sen suggested in his paper that the first seven to eight harmonic components should be used in harmonic analysis. But as we can see before, using first four or eight harmonics may give a K lower than 0.5 in the residuals. From the simulation results, we can see that for most cases, using the first three harmonics is quite enough to obtain satisfactory results. This is because the first three harmonics account for most of the variance. Therefore, this thesis suggests that using the first three harmonics in the analysis.

From the above results, it seems that the number of significant harmonics in flood series is less for those rivers with a high Hurst's K and a low r(1). In other words, the higher the probability that this series is long-term dependent, the less the number of its significant harmonics of this series. To investigate this interesting phenomenon, eight more Canadian rivers were selected to do harmonic analysis.

These eight flood series are all with high K and low r(1). All of them are only long-term dependent, that is, they passed all tests for short-term dependence but failed in Hurst coefficient test for long-term dependence. The calculation results show that the numbers of their significant harmonics range from 1 to 4. This is shown in Table 3.10.

Hence we can see that harmonic components contain some information of long-term dependence. A further study maybe needed to investigate this interesting relationship.

Normal Transformation of the Original Series

For those rivers with non-normal original series, we can not reproduce the coefficient of skewness if we do not transform the original series first(Table 3.9). The simulation results of coefficient of skewness based on transformed normal original series are shown in Table 3.11.

The above simulated results are not too good, but they are the best results which can be obtained. In this method, reproducing the coefficient of skewness is the most difficult compared with reproducing other parameters.

3.6 Conclusions

From the results and analysis described above, the suggested procedure to simulate storage-related processes proposed by Sen needs to be slightly modified. The following procedure is suggested.

(1) Check whether the observed time series is normally distributed. If it is not normal, use a normal transformation.

(2) Construct the original cumulative departure curve from this original (or

transformed) normal series.

(3) Apply harmonic analysis to this curve by depicting the first three harmonic components with fundamental frequency of $1/n$. For those original series with high Hurst's K and low r(1), significant harmonics selected by Fisher's test can be used directly.

(4) Construct the harmonic cumulative departures curve from the harmonic components obtained in the previous step.

(5) Find the residuals between the original and HCD curves.

(6) Fit a suitable ARMA model for the residuals.

(7) Generate the synthetic sequence of residuals and add them to the HCD's curve. Then, if a transformed normal series is used in step (1), transform this synthetic sequence back to a skewed sequence.

In spite of the modified procedure, there are still some problems that have to be addressed. For instance, the residuals are not easy to model in some cases. Also, for series with a fairly high coefficient of skewness, even Box-Cox transformation may not work. Finally, as pointed out by Sen, the harmonics are dependent on the sample size. Therefore, Sen's method is limited at present to simulating series of the same size as the historical series.

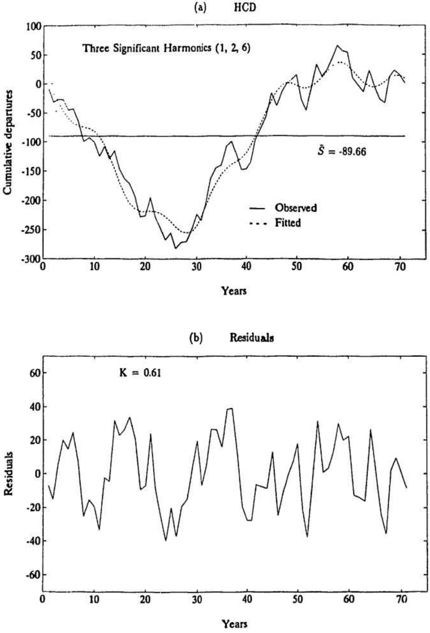
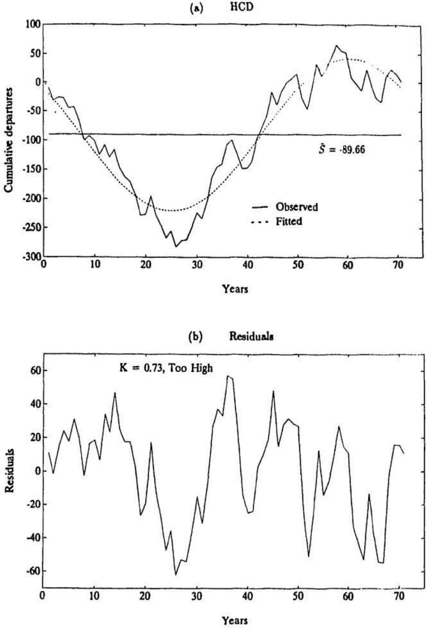


Figure 3.1: HCD and Residuals for the Thames River - Significant Har.



(a) HCD

(b) Residuals

Figure 3.2: HCD and Residuals for the Thames River - First Har.

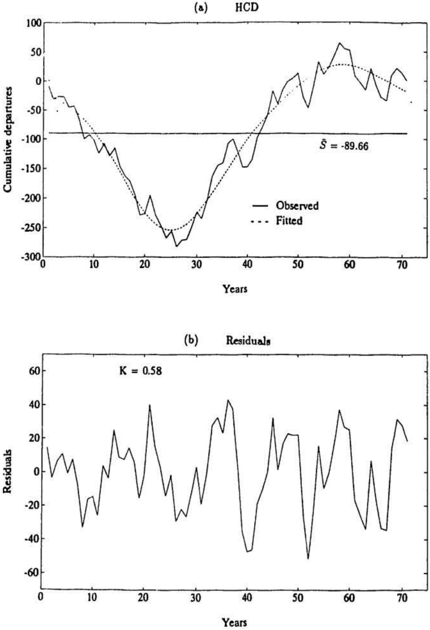


Figure 3.3: HCD and Residuals for the Thames River - First 3 Har.

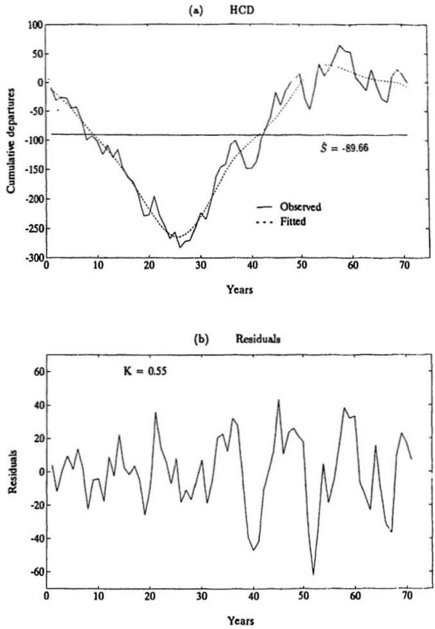


Figure 3.4: HCD and Residuals for the Thames River - First 4 Har.

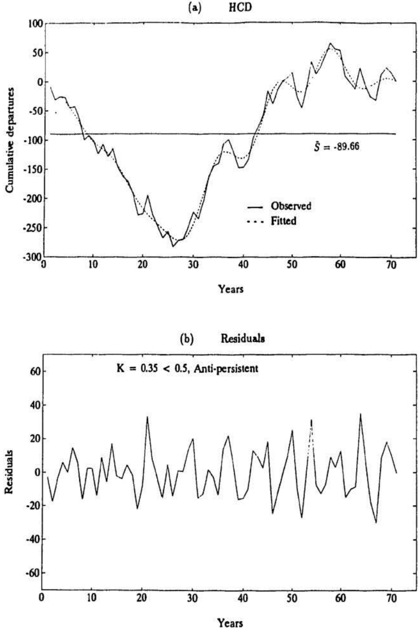


Figure 3.5: HCD and Residuals for the Thames River - First 8 Har.

Table 3.1: Annual Flow Characteristics of Seven Rivers in Sen's Paper

River	Station	No. obs. (year)	Mean (m^3/s)	Standard Deviation (m^3/s)	Lag-one Serial Correlation
St. Lawrence	Ogdesburg	97	6818.64	594.94	0.705
Mississippi	St. Louis	96	4958.62	1482.77	0.295
Mississippi	Keokuk	79	1732.17	511.67	0.415
Munes	Arad	77	167.23	67.01	0.245
Rhine	Basle	150	1026.46	163.46	0.076
Danube	Orshavea	120	5364.18	1027.90	0.094
Thames	Teddington	71	62.95	23.03	0.140

Table 3.2: Harmonic Analysis Results – 1

River	*Original Data*	*Residuals*				
		*Sig-har**	*1-har*	*3-har*	*4-har*	*8-har*
St. Lawrence n = 97	Normal K = 0.89 r(1) = 0.705	Normal K = 0.44 (1-9,13)	Normal K = 0.81	Normal K = 0.66	Normal K = 0.62	Normal K = 0.48
Mississippi,S n = 96	Normal K = 0.65 r(1) = 0.295	Normal K = 0.47 (1-5,8)	Normal K=0.73	Normal K = 0.65	Normal K = 0.58	Normal K = 0.44
Mississippi,K n = 79	Not Nor. K = 0.70 r(1) = 0.415	Normal K = 0.44 (1-4,7)	Normal K=0.76	Normal K = 0.62	Normal K = 0.54	Normal K = 0.45
Mures n = 77	Not Nor. K = 0.68 r(1) = 0.245	Normal K = 0.56 (1,2,3,5)	Normal K=0.78	Normal K=0.58	Normal K = 0.57	Normal K = 0.41
Rhine n = 150	Normal K = 0.61 r(1) = 0.076	Normal K = 0.49 (1-5,7,12,21)	Normal K = 0.78	Normal K = 0.64	Normal K = 0.61	Normal K = 0.47
Danube n = 120	Normal K = 0.63 r(1) = 0.094	Normal K = 0.43 (1-6,8)	Normal K=0.71	Not Nor. K = 0.67	Normal K = 0.59	Normal K = 0.44
Thames n = 71	Normal K = 0.76 r(1) = 0.140	Normal K = 0.61 (1,2,6)	Normal K = 0.73	Normal K = 0.58	Normal K = 0.55	Normal K = 0.35

* The significant harmonics identified are given in parenthesis.

Table 3.3: Harmonic Analysis Results – 2 (Canadian Rivers)

River	*Original Data*	*Residuals*				
		Sig-har	*1-har*	*3-har*	*4-har*	*8-har*
Columbia	Not Nor.	Normal	Normal	Normal	Normal	Normal
	K = 0.746	K = 0.52	K = 0.78	K = 0.61	K = 0.54	K = 0.34
n = 77	r1 = -0.077	(1,2,4,6)				
Flathead	Not Nor.	Normal	Normal	Normal	Normal	Normal
	K = 0.785	K = 0.62	K = 0.70	K = 0.57	K = 0.53	K = 0.39
n = 60	r1 = 0.187	(1,3,7,9)				
Southeast Mar	Not Nor.	Normal	Normal	Normal	Normal	Normal
	K = 0.756	K = 0.65	K = 0.71	K = 0.60	K = 0.52	K = 0.37
n = 70	r1 = 0.138	(1,3)				
Northeast Mar	Not Nor.	Normal	Not Nor.	Not Nor.	Normal	Normal
	K = 0.732	K = 0.67	K = 0.74	K = 0.52	K = 0.47	K = 0.43
n = 72	r1 = 0.070	(1,3)				
Black	Normal	Normal	Normal	Normal	Normal	Normal
	K = 0.731	K = 0.50	K = 0.73	K = 0.50	K = 0.52	K = 0.39
n = 73	r1 = 0.112	(1,3)				

Table 3.4: Best ARMA Models for Original Series & HCD Residuals

River	*Original Data*	*Residuals*				
		Sig-har	*1-har*	*3-har*	*4-har*	*8-har*
St. Lawrence n = 97	AR(1) ϕ_1=0.694	ARMA(1,2) ϕ_1=0.270 θ_1=0.300 θ_2=0.652	AR(2) ϕ_1=1.476 ϕ_2=-0.499	ARMA(1,1) ϕ_1=0.751 θ_1=-0.535	ARMA(2,1) ϕ_1=1.520 ϕ_2=-0.755 θ_1=0.317	ARMA(2,2) ϕ_1=1.1526 ϕ_2=-0.830 θ_1 =0.684 θ_2=0.286
Mississippi S n = 96	MA(1) θ_1=-0.315	MA(2) θ_1=-0.709 θ_2=-0.234 θ_1=0.465	AR(2) ϕ_1=1.205 ϕ_2=-0.330	AR(2) ϕ_1=1.120 ϕ_2=-0.357	AR(2) ϕ_1=0.990 ϕ_2=-0.375	MA(1) θ_1=-0.538
Mississippi K n = 79	AR(1) ϕ_1=0.412	ARMA(1,2) ϕ_1=0.395 θ_1=0.444 θ_2=0.517	AR(2) ϕ_1=1.233 ϕ_2=-0.322	AR(2) ϕ_1 =1.192 ϕ_2=-0.438	AR(2) ϕ_1=0.962 ϕ_2=-0.444	ARMA(2,1) ϕ_1=1.087 ϕ_2=-0.779 θ_1=0.974
Mures n = 77	AR(1) ϕ_1=0.226	AR(2) ϕ_1=0.769 ϕ_2=-0.323	ARMA(1,1) ϕ_1=0.870 θ_1=-0.242	AR(2) ϕ_1=0.898 ϕ_2=-0.322	AR(2) ϕ_1=0.859 ϕ_2=-0.324	AR(2) ϕ_1=0.459 ϕ_2=-0.375
Rhine n = 150	(Random)	AR(2) ϕ_1=0.592 ϕ_2=-0.289	AR(1) ϕ_1=0.866	AR(1) ϕ_1=0.804	AR(1) ϕ_1=0.776	ARMA(2,2) ϕ_1=1.361 ϕ_2=-0.702 θ_1=0.773 θ_2=0.182
Danube n = 120	(Random)	AR(2) ϕ_1=0.473 ϕ_2=-0.250	AR(1) ϕ_1=0.897	AR(2) ϕ_1=0.966 ϕ_2=-0.142	ARMA(2,1) ϕ_1=1.611 ϕ_2=-0.796 θ_1=0.664	ARMA(2,1) ϕ_1=1.135 ϕ_2=-0.636 θ_1=0.967
Thames n = 71	(Random)	AR(1) ϕ_1=0.524	AR(1) ϕ_1=0.746	AR(2) ϕ_1=0.733 ϕ_2=-0.256	AR(2) ϕ_1=0.679 ϕ_2=-0.269	ARMA(2,1) ϕ_1=0.540 ϕ_2=-0.714 θ_1=0.934

Table 4 continued (Canadian Rivers)

River	*Original Data*	*Residuals*				
		Sig-har	*1-har*	*3-har*	*4-har*	*8-har*
Columbia $n = 77$	(Random)	(Random)	ARMA(1,1) ϕ_1=0.946 θ_1=0.073	AR(3) ϕ_1=0.612 ϕ_2=0.379 ϕ_3=-0.352	AR(3) ϕ_1=0.266 ϕ_1=0.204 ϕ_3=-0.356	ARMA(1,1) ϕ_1=0.371 θ_1=0.914
Flathead $n = 60$	(Random)	AR(2) ϕ_1=0.785 ϕ_2=-0.272	AR(1) ϕ_1=0.742	AR(4) ϕ_1=0.586 ϕ_2=-0.231 ϕ_3=0.048 ϕ_4=-0.419	ARMA(1,1) ϕ_1=0.197 θ_1=-0.526	MA(2) θ_1=0.227 θ_2=0.779
Southeast $n = 70$	(Random)	AR(1) ϕ_1=0.617	AR(1) ϕ_1=0.749	AR(1) ϕ_1=0.586	AR(2) ϕ_1=0.625 ϕ_2=-0.226	ARMA(2,1) ϕ_1=0.430 ϕ_2=-0.786 θ_1=0.922
Northeast $n = 72$	(Random)	AR(1) ϕ_1=0.518	AR(1) ϕ_1=0.698	MA(1) θ_1=-0.428	AR(3) ϕ_1=0.400 ϕ_2=-0.191 ϕ_3=-0.283	ARMA(2,1) ϕ_1=0.678 ϕ_2=-0.752 θ_1=0.930
Black $n = 73$	(Random)	ARMA(2,1) ϕ_1=1.101 ϕ_2=-0.752 θ_1=0.957	AR(1) ϕ_1=0.874	AR(2) ϕ_1=0.470 ϕ_2=-0.296	ARMA(2,1) ϕ_1=1.119 ϕ_2=-0.704 θ_1=0.960	ARMA(2,1) ϕ_1=0.591 ϕ_2=-0.634 θ_1=0.980

Note: 1) ϕ_i & θ_i are parameters of the ARMA(p,q) models.

2) Random means independent data (based on tests of short-term independence at $\alpha = 5\%$).

Table 3.5: Simulation Results of Hurst's K

River	*No. of*	*Original*	*K of Simulated Series*				
	Sig.-H	*K*	*Sig.-H*	*1-H*	*3-H*	*4-H*	*8-H*
St. Lawrence	10	0.892	0.890	0.874	0.882	0.873	0.866
Missi., S	6	0.646	0.616	0.626	0.629	0.615	0.616
Missi., K	5	0.704	0.690	0.718	0.710	0.715	0.701
Mures	4	0.680	0.695	0.636	0.678	0.677	0.692
Rhine	8	0.613	0.606	0.597	0.600	0.596	0.590
Danube	7	0.632	0.629	0.611	0.622	0.632	0.618
Thames	3	0.760	0.739	0.746	0.755	0.758	0.743
Columbia	4	0.746	0.748	0.719	0.738	0.745	0.742
Flathead	4	0.785	0.756	0.717	0.756	0.760	0.777
Southeast Mar	2	0.756	0.765	0.748	0.759	0.767	0.705
Northeast Mar	2	0.732	0.703	0.711	0.705	0.716	0.700
Black	3	0.731	0.678	0.666	0.709	0.698	0.717

Table 3.6: Simulation Results of r(1)

River	*No. of*	*Original*	*r(1) of Simulated Series*				
	Sig.-H	*r(1)*	*Sig.-H*	*1-H*	*3-H*	*4-H*	*8-H*
St. Lawrence	10	0.695	0.652	0.635	0.622	0.654	0.665
Missi., S	6	0.292	0.228	0.264	0.296	0.277	0.169
Missi., K	5	0.411	0.379	0.296	0.363	0.395	0.465
Mures	4	0.246	0.257	0.134	0.237	0.239	0.247
Rhine	8	0.076	0.073	-0.053	-0.074	-0.079	0.139
Danube	7	0.093	0.088	-0.036	0.090	0.302	0.152
Thames	3	0.139	0.035	0.022	0.147	0.146	0.191
Columbia	4	-0.077	-0.034	-0.003	-0.054	-0.059	-0.103
Flathead	4	0.187	0.185	-0.005	0.211	0.165	0.185
Southeast Mar	2	0.138	0.033	0.036	0.021	0.143	0.171
Northeast Mar	2	0.070	-0.068	-0.027	-0.044	0.072	0.184
Black	3	0.112	0.240	-0.017	0118	0.223	0.191

Table 3.7: Simulation Results of the Mean Value of Series

River	*No. of*	*Original*	*Mean Values of Simulated Series*				
	Sig.-H	*Mean*	*Sig.-H*	*1-H*	*3-H*	*4-H*	*8-H*
St. Lawrence	10	6818.64	6821.57	6823.58	6829.68	6831.42	6825.99
Missi., S	6	4958.62	4980.42	4958.49	4972.85	4992.05	4971.06
Missi., K	5	1732.17	1731.54	1743.58	1749.64	1740.72	1735.11
Mures	4	167.23	168.14	170.85	168.97	168.61	167.68
Rhine	8	1026.46	1024.76	1027.28	1026.30	1025.91	1024.52
Danube	7	5364.18	5359.82	5384.55	5371.78	5361.46	5361.20
Thames	3	62.95	63.07	62.80	62.69	62.85	62.96
Columbia	4	437.60	437.75	443.07	438.68	436.95	437.95
Flathead	4	208.53	208.19	209.42	208.26	207.96	208.47
Southeast Mar	2	38.68	38.55	38.55	38.50	38.57	38.54
Northeast Mar	2	176.31	174.85	174.82	175.31	175.49	175.77
Black	3	129.49	129.64	129.75	129.64	129.59	129.37

Table 3.8: Simulation Results of Standard Deviation

River	*No. of*	*Original*	*Std of Simulated Series*				
	Sig.-H	*Std*	*Sig.-H*	*1-H*	*3-H*	*4-H*	*8-H*
St. Lawrence	10	594.94	601.46	570.49	610.47	664.10	690.38
Missi., S	6	1482.77	1421.68	1468.78	1480.72	1517.32	1445.85
Missi., K	5	511.67	516.64	532.21	602.14	540.3	541.59
Mures	4	67.01	68.64	73.14	71.59	70.48	68.09
Rhine	8	163.46	170.61	192.10	182.73	181.09	187.01
Danube	7	1027.90	1040.83	1061.16	1033.95	1116.21	1116.81
Thames	3	23.03	23.08	23.23	23.39	23.16	25.36
Columbia	4	114.29	114.58	123.83	115.26	115.17	117.36
Flathead	4	70.00	71.11	70.02	72.11	70.06	69.34
Southeast Mar	2	8.18	8.08	8.00	8.14	8.11	10.50
Northeast Mar	2	66.89	67.17	67.78	67.79	66.86	72.63
Black	3	31.15	37.08	31.43	31.31	33.56	31.41

Table 3.9: Simulation Results of the Skew Coefficient

River	*No. of*	*Original*	*Skew Coef. of Simulated Series*				
	Sig.-H	*Skew*	*Sig.-H*	*1-H*	*3-H*	*4-H*	*8-H*
St. Lawrence	10	-0.292	-0.325	0.018	0.054	0.230	-0.184
Missi., S	6	0.297	-0.080	0.000	0.010	0.277	-0.041
Missi., K	5	0.480	0.152	0.086	0.237	0.027	0.127
Mures	4	0.925	0.015	*	0.033	0.022	0.068
Rhine	8	0.146	-0.035	0.004	0.003	0.005	-0.015
Danube	7	0.275	-0.012	0.103	0.014	-0.011	-0.013
Thames	3	0.176	0.008	-0.001	-0.003	0.018	0.108
Columbia	4	0.519	-0.022	*	-0.015	-0.039	0.008
Flathead	4	0.754	-0.007	-0.005	-0.001	-0.029	0.023
Southeast Mar	2	0.619	-0.005	-0.002	-0.009	0.027	0.047
Northeast Mar	2	1.747	-0.006	-0.016	0.017	0.042	0.015
Black	3	0.206	-0.020	0.001	-0.041	-0.025	-0.011

* No suitable ARMA model found.

Table 3.10: Numbers of Significant Harmonics in Annual Peak Series with Long-term Dependence

River Name	*Hurst's K*	ρ (1)	*No. of Significant Harmonics*
Roseway	0.739	0.083	4 (1,2,4,5)
Boundary Creek	0.746	0.119	1 (1)
Chilko At Outlet of Lake	0.744	-0.033	2 (1,2)
Kettle Near Ferry	0.738	0.133	4 (1,2,4,5)
Kootenay At Newgate	0.774	0.065	1 (1)
Moyie	0.848	0.132	3(1,2,3)
English(Sioux Lookout)	0.736	0.003	2 (1,3)
Missinaibi	0.729	0.106	3 (1,2,3)

Table 3.11: Simulation Results of Skew Coefficients Based on Transformed Normal Series

River	*No. of*	λ*	*Original*	*Skew Coef. of Simulated Series*				
	Sig.-H	Value	*Skew*	*Sig.-H*	*1-H*	*3-H*	*4-H*	*8-H*
Missi., K	5	0.5	0.480	0.557	0.483	0.798	0.427	0.506
Mures	4	0.11	0.925	1.244	**	1.039	0.987	0.997
Columbia	4	0.235	0.519	0.491	**	0.536	0.493	0.507
Flathead	4	0.395	0.754	0.538	0.528	0.538	0.523	0.532
Southeast Mar	2	0.005	0.619	0.554	0.541	0.546	0.575	0.622
Northeast Mar	2	0.005	1.747	0.949	0.944	0.894	0.910	0.911

* λ = Box-Cox transformation parameter;

** = No suitable ARMA model found.

Chapter 4

MIXED-NOISE MODEL FOR ANNUAL FLOWS

4.1 General

In the previous chapter, the harmonic analysis of the cumulative departure curves, which is one of the newly developed models for modeling long-term persistence in annual flows, was discussed. In this chapter, another newly developed model for annual flows, which is called Mixed-Noise Model and was designed for modeling mixed behaviour in annual flows(Lye, [4][10]), will be discussed and extended to include skewness. A comparison with harmonic analysis method will be also provided.

Among the several well known models mentioned in Chapter One for the modeling of hydrologic time series with a high Hurst coefficient and a low lag-one serial correlation coefficient, fractional noise has been shown useful in reproducing the type of long-term variability time series that are characterized by a Hurst coefficient larger than 0.50 [13]. Several other operational models have been developed that can do the same but it has been shown that these can be regarded as approximations of fractional noise [35] [37].

However, fractional noise confronts the hydrologist with two problems [10]. The first is that exact computer simulation requires an infinite number of operations so that approximations are needed. The second problem is the low lag serial correlation of fractional noise that is much too high for most hydrologic applications. Therefore, an efficient model that is simple to use is needed. The mixed-noise (MN) model is such a model which is capable of reproducing short-term and long-term serial correlation of flood series as well as the relevant marginal distribution properties, i.e., the mean and variance (Lye, [4][10]). However, there are some issues which have yet to be resolved. For example, how to model skewed peak flow series by mixed-noise model? What is the advantage of this model in practice? This chapter focuses the research on these questions.

4.2 The Mixed-Noise Model

Mixed-noise model was developed along the lines of the ARMA-Markov model [16]. In the development of the MN model, the Hurst coefficient, h, and first order serial correlation $\rho(1)$ are used explicitly to estimate the model's parameters which are easily obtained (Lye, [4]).

In principle, the MN model is obtained as the sum of three or four independent autoregressive or AR(1) processes each with a suitable weight so as to reproduce approximately the autocorrelation function characterized by a given lag-one serial correlation coefficient and a long-term correlation structure corresponding to fractional noise with a given Hurst coefficient.

The autocorrelation function of a three-term mixed-noise process is:

$$\rho_{MN}(s) = a^2\rho_H^s + b^2\rho_M^s + c^2\rho_L^s \tag{4.1}$$

where s is the lag, a^2, b^2, c^2, are the variance fractions(or weights) which sum to unity, ρ_H, ρ_M, and ρ_L are the autocorrelation coefficients of the three independent AR(1) processes. The first AR(1) process models the high frequency effects, the second AR(1) process models the intermediate or medium frequency effects, and the third AR(1) process models the low frequency effects of the time series. Hence, essentially the technique is to fit the autocorrelation function of Fractional Gaussian Noise [13] with the given three weighted autocorrelation function of the AR(1) process.

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

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

$$X_t = a(\rho_H X_{t-1}^{(H)} + \epsilon_t^{(H)}) + b(\rho_M X_{t-1}^{(M)} + \epsilon_t^{(M)}) + c(\rho_L X_{t-1}^{(L)} + \epsilon_t^{(L)}) \tag{4.2}$$

where, $\epsilon^{(H)}, \epsilon^{(M)}$, and $\epsilon^{(L)}$ are normal independent process having variance $(1 - \rho_H^2), (1 - \rho_M^2)$, and $(1 - \rho_L^2)$, respectively.

The autocorrelation function of this process is fitted to the theoretical autocorrelation function of FGN at four specified lags, s_1, s_2, s_3, and s_4. The lag-one serial correlation coefficient $\rho(1)$ may be arbitrarily specified. To obtain the parameters of the model requires the solution of the following system of equations:

$$a^2 + b^2 + c^2 = 1 \tag{4.3}$$

$$a^2\rho_H + b^2\rho_M + c^2\rho_L = \rho(1) \tag{4.4}$$

$$a^2\rho_H^{s1} + b^2\rho_M^{s1} + c^2\rho_L^{s1} = C(s_1;h) \tag{4.5}$$

$$a^2\rho_H^{s2} + b^2\rho_M^{s2} + c^2\rho_L^{s2} = C(s_2;h) \tag{4.6}$$

$$a^2\rho_H^{s3} + b^2\rho_M^{s3} + c^2\rho_L^{s3} = C(s_3;h) \tag{4.7}$$

$$a^2\rho_H^{s4} + b^2\rho_M^{s4} + c^2\rho_L^{s4} = C(s_4;h) \tag{4.8}$$

where, a^2, b^2, c^2, ρ_M, and ρ_L are constrained to lie between 0 and 1, and $\rho(1)$ is the desired first order serial correlation coefficient. The coefficient ρ_H is allowed to be negative or can be set to zero to match the desired first order serial correlation coefficient. $C(s; h)$ is the theoretical autocorrelation function of Fractional Gaussian Noise (FGN) given by:

$$C(s;h) = \frac{1}{2}[|s+1|^{2h} - 2|s|^{2h} + |s-1|^{2h}] \tag{4.9}$$

For large s, the function is approximately given by

$$C(s;h) \approx h(2h-1)s^{2h-2} \tag{4.10}$$

Lye [4] found it was convenient to take $s_1 = 4, s_2 = 15, s_3 = 54$, and $s_4 = 200$. On a logarithmic scale, these chosen lags are equally spaced. The value of $s_4 = 200$ is chosen to reflect the planning period of most water resources projects. Also, the chosen spacings make it easy to estimate the model parameters. Since the autocorrelation function of an AR(1) process diminishes rapidly with increasing lags, the system of equations can be evaluated sequentially rather than simultaneously starting from the low frequency end.

Dividing (4.7) by (4.8), and assuming $\rho_H^{s3}, \rho_M^{s3}, \rho_H^{s4}$ and ρ_M^{s4} to be negligible at lags s_3 and s_4 leads to:

$$\frac{C(s_3;h)}{C(s_4;h)} = \frac{c^2\rho_L^{s3}}{c^2\rho_L^{s4}} \tag{4.11}$$

For a given h value and, s_3 and s_4, the left hand side of (4.11) is defined. Therefore ρ_L can be calculated. Substituting into (4.7) and ignoring the high and medium frequency terms, c^2 is obtained.

From (4.6), and assuming ρ_H^{*2} to be negligible, one gets:

$$b^2 = \frac{C(s_2; h) - c^2\rho_L^{*2}}{\rho_M^{*2}} \tag{4.12}$$

Substituting into (4.5), ρ_M is obtained.

Then from (4.12), one obtains b^2; and from (4.3):

$$a^2 = 1 - b^2 - c^2 \tag{4.13}$$

Finally from (4.4),

$$\rho_H = [\rho(1) - b^2\rho_M - c^2\rho_L]/a^2 \tag{4.14}$$

4.3 Skewed Mixed-Noise Process

The main work on Mixed-Noise model in this study is extending this model to include skewness. Skewed MN sequences may be generated by using a suitable transformation so that the transformed flows are assumed to be normally distributed. Box-Cox transformation is used as such a transformation in this thesis. The procedure for the synthetic generation of skewed flood sequences X_t based on MN model with Box-cox transformation is as follows:

1) Transform the skewed X_t to normal sequences $X_{N,t}$ by Box-Cox transformation:

$$X_{N,t} = \frac{{X_t}^{\lambda} - 1}{\lambda} \tag{4.15}$$

2) Calculate the M(mean), Std(standard deviation), and the parameters (a, b, c, ρ_H, ρ_M, ρ_L) of $X_{N,t}$;

3) Generate three normal independent process $\epsilon^{(H)}$, $\epsilon^{(M)}$, and $\epsilon^{(L)}$, having variance $(1-\rho_H^2)$, $(1-\rho_M^2)$, and $(1-\rho_L^2)$ respectively;

4) Obtain the sequences $Z_{N,t}$

$$Z_{N,t} = a(\rho_H Z_{t-1}^{(H)} + \epsilon_t^{(H)}) + b(\rho_M Z_{t-1}^{(M)} + \epsilon_t^{(M)}) + c(\rho_L Z_{t-1}^{(L)} + \epsilon_t^{(L)}) \tag{4.16}$$

Actually, this is Equ.(4.2), where $Z_t^{(H)}$, $Z_t^{(M)}$ and $Z_t^{(L)}$ are AR(1) process given by:

$$Z_t^{()} = \rho_{()}^3 Z_{t-1}^{()} + \epsilon_{(),t} \tag{4.17}$$

5)Add the mean and standard deviation

$$Z_t = M + Std * Z_{N,t} \tag{4.18}$$

6) Transform Z_t back to skewed squence

$$Y_t = (Z_t{}^{\lambda} + 1)^{1/\lambda} \tag{4.19}$$

Y_t is the needed synthetic skewed MN sequence.

Another way to generate skewed MN variates is by modifying the random numbers used in the generation process [16]. The necessary skewness in the mixed-noise variate may be obtained in different ways (Lye, [4]). The mixed-noise process (4.2) can be written as:

$$X_t = aX_t^{(H)} + bX_t^{(M)} + cX_t^{(L)} \tag{4.20}$$

Cubing both sides and taking expectations,

$$E(X_t^3) = a^3 E(X_t^{(H)3}) + b^3 E(X_t^{(M)3}) + c^3 E(X_t^{(L)3}) \tag{4.21}$$

Since $X_t^{(H)}$, $X_t^{(M)}$, and $X_t^{(L)}$ are independent of each other and have zero mean, the expected values of the cross-product terms are all zero. Also, $X_t^{(H)}$, $X_t^{(M)}$, and $X_t^{(L)}$ are AR(1) processes given by:

$$X_t^{()} = \rho_{()} X_{t-1}^{()} + (1 - \rho_{()}^2)^{1/2} \epsilon_{(),t} \tag{4.22}$$

Cubing both sides and taking expectations,

$$E(X_{(),t}^3) = \rho_{()}^3 E(X_{(),t-1}^3) + (1 - \rho_{()}^2)^{3/2} E(\epsilon_{(),t}^3) \tag{4.23}$$

That is,

$$\gamma_{X_{()}} = \frac{(1 - \rho_{()}^2)^{3/2}}{1 - \rho_{()}^3} \gamma_{\epsilon,()} \tag{4.24}$$

where $\gamma_{X_{()}}$ and $\gamma_{\epsilon,()}$ are the coefficient of skewness of $X_{()}$, and $\epsilon_{()}$ respectively. Substituting into (4.21) one gets:

$$\gamma_X = a^3 \frac{(1 - \rho_{(H)}^2)^{3/2}}{1 - \rho_{(H)}^3} \gamma_{\epsilon,(H)} + b^3 \frac{(1 - \rho_{(M)}^2)^{3/2}}{1 - \rho_{(M)}^3} \gamma_{\epsilon,(M)} + c^3 \frac{(1 - \rho_{(L)}^2)^{3/2}}{1 - \rho_{(L)}^3} \gamma_{\epsilon,(L)} \tag{4.25}$$

From (4.25) there are several possible ways of obtaining the required skewness γ_X:

1. Modify only the high frequency term. Here, $\gamma_{\epsilon,M}$ and $\gamma_{\epsilon,L} = 0$, and the required skewness of the random numbers in the high frequency component is given by:

$$\gamma_{\epsilon,H} = \frac{1 - \rho_{(H)}^3}{a^3 (1 - \rho_{(H)}^2)^{3/2}} \gamma_X \tag{4.26}$$

The Wilson-Hilferty transformation can then be used ' o obtain the required skewed random variate. The transform is given by:

$$\eta_t = \frac{2}{\gamma_{\epsilon,H}} [1 + \frac{\gamma_{\epsilon,H} \epsilon_t}{6} - \frac{\gamma_{\epsilon,H}^2}{36}]^3 - \frac{2}{\gamma_{\epsilon,H}} \tag{4.27}$$

where, η_t is approximately gamma distributed with a mean of zero, unit variance and skewness $\gamma_{\epsilon,H}$; $\gamma_{\epsilon,H}$ is the skewness of the random deviates required; and ϵ_t is a normally distributed random deviate with zero mean and unit variance.

2. Modify only the medium frequency term. In this case, $\gamma_{\epsilon,H}$ and $\gamma_{\epsilon,L} = 0$, and the required skewed deviates are from (4.26) and (4.27) with $\gamma_{\epsilon,H}$ replaced by $\gamma_{\epsilon,M}$.

3. Modify only the low frequency term. In this case, $\gamma_{\epsilon,H}$ and $\gamma_{\epsilon,M} = 0$, and the required skewed deviates are from (4.26) and (4.27) with $\gamma_{\epsilon,H}$ replaced by $\gamma_{\epsilon,L}$.

4. Skewed random deviates can also be obtained by assuming the same skewness for each component. That is, $\gamma_\epsilon = \gamma_{\epsilon,H} = \gamma_{\epsilon,M} = \gamma_{\epsilon,L}$. From (4.25),

$$\gamma_\epsilon = \gamma x [a^3 \frac{{1-\rho_{(H)}^2}^{3/2}}{1-\rho_{(H)}^3} + b^3 \frac{{1-\rho_{(M)}^2}^{3/2}}{1-\rho_{(M)}^3} + c^3 \frac{{1-\rho_{(L)}^2}^{3/2}}{1-\rho_{(L)}^3}]^{-1} \qquad (4.28)$$

and the required skewed random variates is obtained from (4.27) by replacing $\gamma_{\epsilon,H}$ with γ_ϵ.

Second and third methods do not work because b and c are usually very small. From (4.26), $\gamma_{\epsilon,M}$ and $\gamma_{\epsilon,L}$ become very large. But in (4.27), because this large number ($\gamma_{\epsilon,M}$ or $\gamma_{\epsilon,H}$) should be squared and then be cubed, the corresponding η_t becomes very huge. Therefore only the results of methods 1 and 4 are shown in the next section.

4.4 MN Model Applied to Annual Peak Flows

Monte Carlo simulations are used to test the suitability of Mixed-Noise model in modeling the annual peak flows. The data of the twelve rivers discussed in previous chapters are used here as examples. The number of replications equals 3,000.

Two ways of modeling skewed MN sequences are all considered in the calculation. A comparison of these results is provided. The simulation results are shown in Table 4.1 - 4.3

The results in these tables show that, in most cases, using Box-Cox transformation and using Wilson-Hilferty transformation in MN skewed sequences obtain almost the same results in reproducing several important parameters, like the mean, standard deviation, R_1, and Hurst's K. But, for modeling skew coefficient, Wilson-Hilferty transformation gets better results, especially when the skew coefficient of original data is very high. It seems that it is difficult to use Box-Cox transformation to model high skew coefficient.

The bias in reproducing several parameters are quite small. A little larger bias exists in reproducing Hurst's K. Usually, the simulated Hurst's K is less than the original K if original $K > 0.70$. If original $K < 0.70$, the simulated K becomes greater than the original K. This bias pattern of simulated K here is therefore quite normal [25].

To produce generated sequences from the modified mixed-noise model that "on average" reproduce statistics equal to the historical values, the parameters used in the model must be corrected for bias. Analytical expression for bias correlation for the modified process maybe possible. However, bias correction derived from the Monte Carlo method is sufficient for most practical purposes. Lye [4] gave some simple curves which can be easily used to find suitable inputs of K and $\rho(1)$ in simulation for bias correction [4]. Some improved results based on this bias correction method are shown in Table 4.4.

Compared with the simulated results of harmonic analysis in Chapter Three, the mixed-noise model with Wilson-Hilferty transformation is more effective in reproducing skew coefficients of the original sequences. Its advantage in this aspect is very obvious for the high skew sequences(e.g. the results for Northeast River). The harmonic analysis method gave better results in reproducing Hurst's K if mixed-noise model is used without bias correction. This is because the harmonic analysis constructs the original cumulative departures curve firstly from the observed time series, and the separated cumulative departures curve accounts for more than 95% of the variability in the Hurst coefficient.

4.5 Summary

Long-term dependence can not be ignored in modeling flood time series, that is, the effects of medium and low frequencies should be considered. Mixed-noise model is quite effective in this way. It uses three AR(1) processes which are able to model the effects of high frequency, medium frequency, and low frequency respectively. Hence it can be used to model those flood series with long-term dependence, and its physical idea is easy to understand.

Mixed-noise model with a suitable transformation can be used to model skewed series efficiently. This advantage is evident for those high skew series, compared with harmonic analysis method. Wilson-Hilferty transformation is recommended to be used as such a transformation by this thesis because it obtains better results than using Box-Cox transformation, although it is more difficult to use.

Bias in reproducing statistical parameters exists in mixed-noise model. But this can be corrected by changing the inputs of Hurst's K and $\rho(1)$ in this model.

The inputing values of K and $\rho(1)$ can be easily obtained from curves which are obtained based on the Monte Carlo method.

The mixed-noise process has several advantages. Firstly, it uses both h and $\rho(1)$ explicitly to derive its parameters. Secondly, the parameters are easy to estimate. Finally, because of its simple structure, it is relatively efficient when compared to the computational time of present models. It remains, however, to determine the optimum values of the lags (s_1, s_2, s_3, and s_4) where the MN correlation function is forced to match the FGN correlation function. More curves for bias correction for different sample lengths should be obtained by Monte Carlo Method. In addition, a comprehensive comparison with the ARMA-Markov model and other contending models in terms of small sample biases, generation of skewed variates, and extension to the multivariate case remains to be carried out.

Table 4.1: Simulation Results of MN Skewed Sequences By Box-Cox Transformation(considered all H, M, L terms)

River	λ	*Mean*		*Std*		*R1*		*Skew*		*Hurst's K*	
		Ori.	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*
St. Lawrence*	2.360	6818.64	6821.60	594.94	470.74	0.695	0.443	-0.292	-0.269	0.892	0.793
Missi., S*	0.600	4958.62	4953.98	1482.77	1452.34	0.292	0.255	0.297	0.329	0.646	0.703
Missi., K	0.500	1732.17	1729.79	511.67	490.76	0.411	0.332	0.480	0.358	0.704	0.731
Mures	0.110	167.23	167.42	67.01	65.73	0.246	0.119	0.925	0.973	0.680	0.687
Rhine*	0.680	1026.46	1025.93	163.46	162.08	0.076	0.060	0.146	0.144	0.613	0.657
Danube*	0.170	5364.18	5361.91	1027.90	1021.39	0.093	0.058	0.275	0.439	0.632	0.665
Thames*	0.580	62.95	62.91	23.03	21.07	0.139	0.010	0.176	0.407	0.760	0.703
Columbia	0.235	437.60	437.75	114.29	108.88	-0.077	-0.169	0.519	0.534	0.746	0.684
Flathead	0.395	208.53	208.78	70.00	64.27	0.187	0.039	0.754	0.585	0.785	0.713
Southeast Mar	0.005	38.68	38.75	8.18	7.76	0.138	0.021	0.619	0.589	0.756	0.701
Northeast Mar	0.005	176.31	176.41	66.89	62.77	0.070	0.005	1.747	1.464	0.732	0.690
Black*	0.830	129.49	129.46	31.15	29.79	0.112	0.019	0.206	0.108	0.731	0.698

* Originally normal data

Table 4.2: Simulation Results of MN Skewed Sequences By Wilson-Hilferty Transformation(γ_x from high freq. terms)

River	*Mean*		*Std*		*R1*		*Skew*		*Hurst's K*	
	Ori.	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*
St. Lawrence	6818.64	6817.46	594.94	468.95	0.695	0.490	-0.292	-0.525	0.892	0.796
Missi., S	4958.62	4955.67	1482.77	1456.60	0.292	0.253	0.297	0.286	0.646	0.701
Missi., K	1732.17	1729.21	511.67	491.75	0.411	0.344	0.480	0.472	0.704	0.732
Mures	167.23	167.20	67.01	64.89	0.246	0.189	0.925	0.881	0.680	0.702
Rhine	1026.46	1025.93	163.46	162.26	0.076	0.055	0.146	0.141	0.613	0.657
Danube	5364.18	5361.96	1027.90	1016.62	0.093	0.065	0.275	0.266	0.632	0.667
Thames	62.95	62.95	23.03	21.65	0.139	0.014	0.176	0.197	0.760	0.705
Columbia	437.60	437.51	114.29	108.52	-0.077	-0.186	0.519	0.525	0.746	0.683
Flathead	208.53	208.62	70.00	64.08	0.187	0.031	0.754	0.840	0.785	0.715
Southeast Mar	38.68	38.74	8.18	7.69	0.138	0.020	0.619	0.653	0.756	0.704
Northeast Mar	176.31	175.96	66.89	62.20	0.070	-0.006	1.747	1.618	0.732	0.696
Black	129.49	129.42	31.15	29.81	0.112	0.018	0.206	0.214	0.731	0.695

Table 4.3: Simulation Results of MN Skewed Sequences By Wilson- Hilferty Transformation(γ_r from all H,M,L freq. terms)

River	*Mean*		*Std*		*R1*		*Skew*		*Hurst's K*	
	Ori.	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*
St. Lawrence	6818.64	6817.30	594.94	469.61	0.695	0.489	-0.292	-0.475	0.892	0.795
Missi., S	4958.62	4955.70	1482.77	1456.40	0.292	0.253	0.297	0.285	0.646	0.701
Missi., K	1732.17	1729.03	511.67	491.58	0.411	0.344	0.480	0.467	0.704	0.731
Mures	167.23	167.19	67.01	64.86	0.246	0.187	0.925	0.877	0.680	0.700
Rhine	1026.46	1025.93	163.46	162.26	0.076	0.055	0.146	0.141	0.613	0.657
Danube	5364.18	5361.95	1027.90	1016.50	0.093	0.064	0.275	0.265	0.632	0.667
Thames	62.95	62.95	23.03	21.65	0.139	0.014	0.176	0.192	0.760	0.705
Columbia	437.60	437.48	114.29	108.47	-0.077	-0.189	0.519	0.515	0.746	0.682
Flathead	208.53	208.69	70.00	64.03	0.187	0.027	0.754	0.819	0.785	0.714
Southeast Mar	38.68	38.73	8.18	7.67	0.138	0.018	0.619	0.641	0.756	0.703
Northeast Mar	176.31	175.20	66.89	62.09	0.070	-0.017	1.747	1.633	0.732	0.692
Black	129.49	129.42	31.15	29.80	0.112	0.017	0.206	0.211	0.731	0.699

Table 4.4: MN Skewed Sequences with Bias Correction (Wilson- Hilferty Transformationγ_r from all H,M,L freq. terms)

River	*Mean*		*Std*		*R1*		*Skew*		*Hurst's K*	
	Ori.	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*	*Ori.*	*Simu.*
St. Lawrence	6818.64	6816.58	594.94	527.73	0.695	0.729	-0.292	-0.075	0.892	0.869
Missi., S	4958.62	4957.25	1482.77	1472.95	0.292	0.250	0.297	0.275	0.646	0.675
Missi., K	1732.17	1730.60	511.67	520.86	0.411	0.348	0.480	0.443	0.704	0.714
Thames	62.95	62.94	23.03	23.00	0.139	0.155	0.176	0.173	0.760	0.760

Chapter 5

CONSIDERATION OF THE HURST PHENOMENON IN FLOOD RISK ANALYSIS

5.1 General

In the previous chapters, the existence of long-term dependence in flood peak series and how to model this kind of dependence were discussed. But there is still another question left to be answered, that is, why do we need to model long-term dependence in hydrologic series? In other words, does it make sense in engineering practice that we consider the Hurst phenomenon in flood risk analysis? In this chapter, the effect of serial correlation on flood risk will be discussed. Mixed-noise model and Monte Carlo method will be used in the analysis.

The effect of short and long term serial correlation on the variability of sample statistics were discussed by some researchers. Loucks et al. [34] showed that short-term dependence increased the variances of sample mean and sample variance [34]. Lye [4] showed that when the series of observations exhibits long term serial correlation, the variance of the sample statistics are greater than that for ei'her short term correlated or independent processes [4]. This chapter will illustrate that, when Hurst's K is high, how much it will affect flood tolerance limits in risk analysis if we ignore the serial correlation.

5.2 Method of Flood Risk Analysis

As a rule, hydrologists pay little attention to the serial correlation of flood peak series. This is inherent in the way a flood frequency analysis is performed. The flood data are arranged in order of magnitude. Then they are plotted on probability graph paper, and finally, a probability distribution or a curve is fitted through the plotted points. It is evident that in this process, the order in which the data occured in the time series, and therefore the serial correlation structure, is considered to be irrelevant.

When we use a flood frequency curve, we set a magnitude of flood to occur in P percent of some long future record of floods. Usually, we think of the return period (T) rather than the probability. A 25-year flood would be found to be exceeded, *on the average*, four times in each 100-year period of a large number of 100-year records. By using the return period as a design criterion, we are implying that we expect average conditions to apply over some long future.

Risk estimation is an alternative to the return-period concept. Risk can be introduced using the annual maximum flood as the random variable. Thus the time period is 1 year. We will assume that in any one year, a flood either occurs or it does not occur and that no more than one flood of a certain magnitude will occur in any one year. If we also assume that the probability of that flood occuring remains constant from year to year, we have satisfied the four assumptions underlying the binomial distribution. Thus, if we define the risk as being the probability of one or more floods having probability p occuring in n years, we get the risk as

$$risk = 1 - \binom{n}{0} p^0(1-p)^{n-0} = 1 - (1-p)^n \tag{5.1}$$

We can develop the concept of risk in a different way. If p is the probability that a flood will occur in any year, $1-p$ is the probability that it will not occur. If, further, we need n years for construction, $(1-p)^n$ is the probability that the flood will not occur in these n years. Conversely, $1-(1-p)^n$ is the probability that the n-year period will not be flood free. In other words, it is the chance of at least one flood equal to or greater than the flood corresponding to p, and represents risk.

The above methods are correct procedures only when we have long-lived structures, or economic benefits accruing over very long periods and the floods are serially *independent*. But what risks do we run when the floods are serially dependent or there is no opportunity for long-time averaging? We can use confidence intervals to make probabilistic predictions about possible future values of the mean of the sample. Also, we can use tolerance limits to make probabilistic predictions about possible future values of specified proportions of our samples [44]. When doing this, we must attach a confidence level to the prediction. The probability of 0.5% is a sample proportion of 1/200, and we must estimate how this proportion might range.

If we had performed the simulation of m samples with the same sample size n, we would be in a position to answer several questions. What are the range and distribution of the largest x in each of the samples? What are the range and distribution of the kth largest value? What are the range and distribution of the middle value of each sample? Answers to such questions are illustrated in the next section, and a comparison of these answers between two corresponding simulated series(dependent series and independent series) is also provided to show the increase in the uncertainty of flood risk assessment when the serial correlation is not considered in a high Hurst's K flood series.

5.3 Results and Analysis

The Thames river is chosen as an example in this analysis. The mixed series data passed all tests for short-term dependence considered in this study, but the data failed Hurst's K test for long-term dependence. That is, it is an independent series under the traditional consideration but a dependent series under the viewpoints in this study. Hence, it is a good example for our comparing the their difference on flood risk.

Monte Carlo method is used for this analysis. Two thousand samples of independent sequences, with the same sample size as Thames' flow series(n=71), are generated first. Their means, standard deviations, and coefficients of skewness are statistically equal to those of Thames' flow series. Then, two thousand samples of dependent sequences, also with the same sample size, are generated by mixed-noise model. Following the fitting procedure with bias corrections for this model suggested by Lye [4], the synthetic dependent sequences on average reproduce the required sample statistics(mean, standard deviation, R_1, coefficient of skewness, and Hurst's K). These are shown in the following Table 5.1.

To analyze the simulation results, we are specially interested in the probability distributions of largest events. This is because in engineering practice, only those results with highest sample ranks affect our selecting design flood value. For example, rank number 1 corresponds to the 1/71 event, rank number 2 corresponds to the 2/71 event, and so on. The simulation results of 5 highest sample ranks and the median rank (rank 36 for Thames flood series with n=71) are shown in Table 5.2.

The distributions of these events are shown in Fig. 5.1 using boxplots. Here the lower hinge H_L is the first quartile, the upper hinge H_U is the third quartile. The inner fences is between $H_L - 1.5(H_U - H_L)$ and $H_U + 1.5(H_U - H_L)$, and the outer fences is between $H_L - 3.0(H_U - H_L)$ and $H_U + 3.0(H_U - H_L)$. " * " represents possible outlier which is between the inner and outer fences. " 0 " represents the probable outlier which is beyond the outer fences.

The above results illustrate clearly that the effect of serial correlation on flood risk. For dependent series, the median values of ranks numbers ("+" in the box) are almost the same as those of independent series. In other words, consideration of serial correlation (or Hurst phenomenon) does not change the results of frequency analysis. It obtains the same answers in estimated median flood values for certain return periods. However, it gave much larger variance in flood values for every rank. The larger variance means larger uncertainty and the decrease in the reliability of the results of frequency analysis. Therefore, uncertainty in the estimated flood risk increases if the effect of Hurst phenomenon is taken into account.

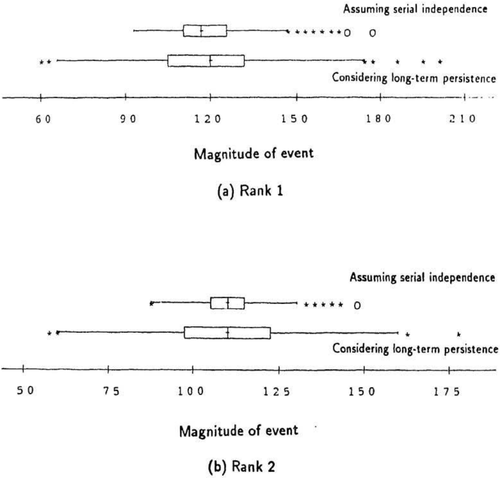
To investigate how much is exactly the increase on flood risk, a further analysis is needed. Either for dependent series or independent series, all 2000 samples have their events arranged in order of magnitude. As a second step, each of these rank numbers may also be arranged in order of magnitude. We thus have ranks 1 to 71 of the individual samples ranked across the 2000 samples. Selected tolerance limits numbers, the means and the median values of the 15 highest sample ranks are plotted in Fig. 5.2 and shown in Table 5.3.

From the above demonstrations one can see that the variance of the ordered statistics is substantially higher for flood series that exhibit the Hurst phenomenon. Thus, neglecting uncertainty in these cases may cause serious underestimation of the future flood risk. For example, under the traditional viewpoints, the flow series of Thames river is an independent series because it passed all statistical tests for short-term dependence, and hence the serial correlation can be disregarded in flood frequency analysis. But from the results in Table 5.3, this causes serious underestimation of the flood risk. If we want to estimate the possible 2/71 event (rank 2 number) at a significance level of $\alpha = 5\%$, the error of underestimation for the upper limit would be 11.01%. For an estimation of possible 14/71 event (about rank 14 number) at the same significance level, this error could be up to about 20%.

5.4 Summary

From the above discussion, it is evident that the serial correlation of annual peak flows, and by implication the nature of the variability of these peak flows, should not be taken for granted. It should be deduced from observations and every effort should be made to determine the hydrometeorological conditions that may cause the serial correlation. In addition to the standard statistical tests for short-term dependence, flood peak flows should also be tested by Hurst's K for its possible long-term dependence.

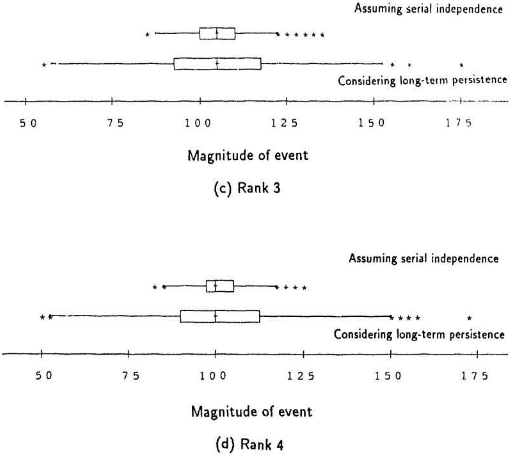
If flood peak flows fail in Hurst coefficient test, the serial correlation should be taken into account in flood risk analysis, otherwise the underestimation of flood risk could be serious. Mixed-noise model with bias corrections and some other models like harmonic analysis can be used to model flood sequences with long-term persistence. If necessary, the uncertainty in the distributions of the highest ranks numbers can be quantified using Monte Carlo methods, like the analysis methods used in this chapter. The most important aspect of this flood risk analysis method may be that generated synthetic flood sequences should on average reproduce the required sample statistics of original flood sequence.



(a) Rank 1

(b) Rank 2

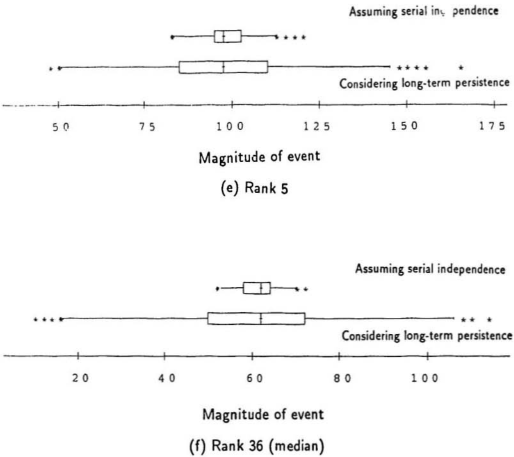
Figure 5.1: Boxplots of selected events



(c) Rank 3

(d) Rank 4

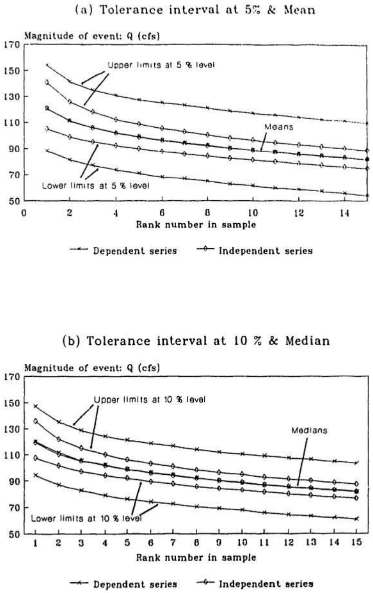
Figure 5.1: Boxplots of selected events



(e) Rank 5

(f) Rank 36 (median)

Figure 5.1: Boxplots of selected events



(a) Tolerance interval at 5% & Mean

(b) Tolerance interval at 10 % & Median

Figure 5.2: Tolerance intervals for the 15 highest ranks

Table 5.1: Simulation Results of Thames' Annual Flow Series (Replications in Monte Carlo method: 2000)

Parameter	*Mean*	*Std*	R_1	*Coef. Skew*	*K*
Original Series	62.947	23.025	0.139	0.176	0.760
Dependent Series	62.938	23.004	0.155	0.173	0.760
Independent Series	62.860	22.944	-0.013	0.175	0.623

Table 5.2: Simulation results of five highest ranks and median rank

Rank		1	2	3	4	5	36(median)
Dependent	*Mean*	120.46	111.41	105.95	102.15	99.07	62.27
Series	*Std*	20.42	18.52	17.88	17.61	17.38	16.74
Independent	*Mean*	120.61	111.04	105.80	101.91	98.87	62.11
Series	*Std*	11.08	8.24	6.96	6.14	5.68	3.43

Table 5.3: Tolerance limits for the 15 highest sample ranks

(a) Dependent series

Rank	U 5%	U 10%	L 90%	L 95%	*Median*	*Mean*
1	154.04	147.24	94.44	88.33	120.07	120.46
2	141.32	135.03	86.94	81.17	111.53	111.41
3	134.74	128.69	82.70	77.05	105.53	105.95
4	130.83	123.93	78.95	73.93	102.01	102.15
5	127.25	121.07	76.04	70.94	98.80	99.08
6	125.03	118.59	74.07	68.29	96.38	96.55
7	123.23	116.37	72.28	67.05	94.31	94.35
8	120.65	113.94	70.45	64.97	92.22	92.36
9	118.75	111.68	68.95	62.84	90.17	90.43
10	117.05	110.14	67.59	61.57	88.41	88.77
11	115.65	109.07	65.69	59.82	86.81	87.23
12	114.16	107.56	64.41	58.82	85.44	85.78
13	112.41	106.34	62.97	57.38	84.01	84.41
14	111.36	104.61	61.88	55.62	82.60	83.10
15	109.61	103.28	60.76	54.39	81.34	81.84

(b) Independent series

Rank	U 5%	*Err at 5%	U 10%	L 90%	L 95%	*Median*	*Mean*
1	140.77	-8.62%	135.77	107.76	105.02	119.32	120.61
2	125.76	-11.01%	121.79	101.44	98.77	110.30	111.04
3	117.93	-12.48%	115.01	97.29	95.10	105.32	105.80
4	112.41	-14.08%	109.94	94.18	92.43	101.67	101.91
5	108.71	-14.57%	106.09	91.96	89.94	98.76	98.87
6	105.42	-15.69%	103.22	89.48	87.70	96.14	96.28
7	102.80	-16.58%	100.74	87.56	85.86	93.96	94.04
8	100.06	-17.06%	98.31	85.62	84.09	92.03	92.07
9	98.06	-17.43%	96.34	84.15	82.66	90.31	90.29
10	96.41	-17.63%	94.66	82.70	81.30	88.69	88.65
11	94.45	-18.33%	92.89	81.46	79.87	87.02	87.12
12	92.95	-18.58%	91.38	80.12	78.57	85.60	85.70
13	91.66	-18.45%	90.08	78.66	77.30	84.22	84.31
14	90.04	-19.15%	88.57	77.53	76.06	82.90	82.98
15	88.82	-18.97%	87.32	76.50	74.94	81.60	81.74

*** Note: "Err at 5%" represents the underestimation in the risk assessment because of the assumption that the flow series is independent.**

Chapter 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

There are several conclusions which can be drawn from this thesis.

1. Many annual peak flows exhibit the Hurst phenomenon, but standard statistical tests for independence are insensitive to the long-term dependence of the peak flow series. Significant long-term serial correlation as measured by the Hurst coefficient is present in a large number of the peak flow series which passed standard statistical tests for short-term independence. Therefore, all annual peak series should be examined by Hurst coefficient test for their long-term dependence.

2. Generally, harmonic analysis of the cumulative departures of annual flow series is a good method to simulate storage-related process. The first advantage of this method is in reproducing required Hurst's K in simulated series. But the difficulties come out in reproducing the coefficient of skewness when the original flood series is highly skewed. Because the residuals are usually not

independent data, sometimes it is difficult to find the best ARMA model for the residuals.

3. Mixed-noise model is an new effective model which is able to model the effects of high frequency, medium frequency, and low frequency in flood series respectively. Using Wilson-Hilferty transformation, this model can easily reproduce the the coefficient of skewness of a high skew flood series. Bias in reproducing the statistical parameters exists for this model. But, with some curves prepared by Monte Carlo method, this bias is easily corrected.

4. The effect of serial correlation on flood risk analysis can be substantial. It should be considered for those rivers exhibiting significant long-term dependence, otherwise it may cause serious underestimation of the future flood risk.

6.2 Recommendations

From the results of this study, the following issues should be considered for further research:

1. Since parameter uncertainty caused by long term serial correlation is quite substantial leading to a upward assesment of flood risk, physical reasons for the long term behaviour should be investigated for each river basin where this phenomenon is observed.

2. It seems that there is some kind of relationship between the number of significant harmonics of flood series and long-term dependence. It may be inter-

esting to do further studies on this relationship and its physical explanations.

3. There are several models available which are capable of simultaneously reproducing high and low frequency effects. It is necessary to develop the procedures for ease of selection of models and input parameters for any desired output characteristics.

References

[1] Chow, V.T.,1964. Handbook of Applied Hydrology. McGraw-Hill, New York, New York.

[2] U.S. Water Resources Council, 1981. Guidelines for Determining Flood Flow Frequency. Bulletin No. 17B of the Hydrology Committee.

[3] Carrigan, P.H. and Huzzen, C.S., 1967. Serial Correlation of Annual Floods. Proceedings, the International Hydrology Symposium, Ft. Collins, Colorado, pp. 322-328.

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

[5] Yevjevich, V.M., 1964. Fluctuations of Wet and Dry Years. Part II. Analysis by Serial Correlation. Hydrology Papers, No.4, Colorado State University, Ft.Collins, Colorado.

[6] McMahon, T.A., 1979. Hydrologic Characteristics of Australian Streams. Civil Engineering Reports, Monash University, Australia.

[7] Wall, D.J. and Englot, M.E., 1985. Correlation of Annual Peak Flows for Pennsylvania Streams. Water Resources Bulletin, 21(3), pp. 459-464.

[8] Hurst, H.E., 1951. Long Term Storage Capacity of Reservoirs. Trans. Am. Soc. Civ. Engrs., 116, pp. 770-808.

[9] Hurst, H.E., 1956. Methods of Using Long-term Storage in Reservoirs. Proc. Inst. Civil Engineers, 1, pp. 519-543.

[10] Booy, C. and Lye, L.M., 1989. A New Look at Flood Risk Determination. Water Resources Bulletin, 25(5), pp. 933-942.

[11] Srikanthan, R., 1979. Stochastic Generation of Annual and Monthly Flow Volumes. Ph.D. Thesis, Monash University, Australia.

[12] Hall, W.A., Askew, A.J. and Yeh, W.W.-G., 1969. Use of Critical Period in Reservoir Analysis. Water Resources Research, 5(6), pp. 1205-1215.

[13] Mandelbrot, B.B., 1971. A Fast Fractional Gaussian Noise Generator. Water Resources Research, 7(3), pp. 543-553.

[14] Mejia, J.M., Rodriguez-Iturbe, I. and Dawdy, D.R., 1972. The Broken Line Process as a Potential Model for Hydrologic Simulation. Water Resources Research, 8(4), pp. 931-941.

[15] O'Connell, P.E., 1974. Stochastic Modelling of Long Term Persistence on Streamflow Sequences. Ph.D. Thesis, Imperial College, University of London, pp. 284.

[16] Lettenmaier, D.P. and Burges, S.J., 1977. Operational Assessment of Hydrologic Models of Long Term Persistence. Water Resources Research, 13(2), pp. 281-290.

[17] Sen, Z., 1991. Note On The Cyclic Features in Cumulative Departures of Annual Flow Series. Journal of Hydrology, 125: pp. 47-54.

[18] Booy, C. and Morgan, D.R., 1985. The Effect of Clustering of Flood Peaks on a Flood Risk Analysis for the Red River. Canadian Journal of Civil Engineering, 12(1), pp. 150-165.

[19] Booy, C. and Lye, L.M., 1986. Accumulated Basin Storage as a Factor in the Correlation Structure of Annual Peak Flows on the Red River. Canadian Journal of Civil Engineering, 13(3), pp. 365-374.

[20] Srikanthan, R., McMahon, T.A. and Irish, J.L., 1983. Time Series Analysis of Annual Flows of Australian Streams. J. Hydrology, 66: pp. 213-226.

[21] Wald, A. and Wolfowitz, J., 1943. An Exact Test for Randomness in the Nonparametric Case Based on Serial Correlation. Ann. Math. Statistics, 14, pp. 378-388.

[22] Shiau, S.Y. and Condie, R., 1980. Statistical Tests for Independence, Trend, Homogeneity and Randomness. Hydrologic Applications Division, Water Resources Branch, Inland Waters Directorate, Environment Canada, Ottawa, Ontario.

[23] Yevjevich, V., 1971. Stochastic Processes in Hydrology. Water Resources Publications, Fort Collins, Colorado.

[24] Madansky, A., 1988. Prescriptions for Working Statisticians. Springer-Verlag, New York. pp. 93-118.

[25] Wallis, J.R. and Matalas, N.C., 1971. Small Sample Properties of H and K–Estimators of the Hurst Coefficients h. Water Resources Research 6(6): pp. 1583-1594.

[26] Wallis, J.R. and Matalas, N.C., 1971. Correlogram Analysis Revisited. Water Resources Research, 7(6): pp. 1448-1459.

[27] Yevjevich, V., 1972. Structural Analysis of Hydrologic Time Series. Colorado State University Hydrology Papers, No.56, Fort Collins, Colorado.

[28] Probability and Statistics Group, Computation Center of the Chinese Academy of Science, 1979. Probability and Statistics Calculation. Science Publications, Beijing.

[29] Box, G.E.P. and Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, Calif.

[30] Lye, L.M., 1993. Selecting An Appropriate Box-Cox Transformation in Flood Frequency Analysis, to appear in Canadian Journal of Civil Engineering, Oct.

[31] Booy, C. and Lye, L.M., 1987. Uncertainty in Flood Risk Analysis. Proceedings, 8th Canadian Hydrotechnical Conference, Montreal, pp. 401-418.

[32] Schwarz, G., 1978. Estimating the Dimension of a Model. Annals of Statistics, 6, 2, pp. 461-464.

[33] Mills, T.C., 1990. Time Series Techniques for Economists. Cambridge University Press, Cambridge. pp. 139.

[34] Loucks, D.P., Stedinger, J.R. and Haith, D.A., 1981. Water Resources System Planning and Analysis. Prentice Hall, Englewood Cliffs, N.J. pp. 559.

[35] Mandelbrot, B.B., 1972. Broken Line Process Derived as an Approximation to Fractional Noise. Water Resources Research 8(5): pp. 1354-1356.

[36] Mandelbrot, B.B. and Wallis, J.R., 1969. Robustness of the Rescaled Range R/S in the Measurement of Non-cyclic Long-Run Statistical Dependence. Water Resources Research, 5(5), pp. 967-988.

[37] Wallis, J.R. and O'Connell, P.E., 1973. Firm Reservoir Yield: How Reliable Are Historic Hydrological Records. Bulletin of the International Association for Hydrological Science 18(3): pp. 347-365.

[38] Feder, J., 1988. Fractals. Plenum Press. pp. 171 & 181.

[39] Lye, L.M., Sinha, S.K., and Booy, C., 1988. Bayesian analysis of the T-year events for flood data fitted by a three-parameter lognormal distribution. Civ. Engng Syst. vol 5, June: pp. 81-86.

[40] Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7, pp. 1-26.

[41] Efron, B., 1981. Nonparametric Standard Errors and Confidence Intervals. Canadian Journal of Statistics, 9, pp. 139-172.

[42] Efron, B., 1982. The Jackknife, the Bootstrap, and Other Resampling Plans. SIAM, monograph #38, CBMS-NSF.

[43] Box, G.E.P. and Cox, D.R., 1964. An Analysis of Transformation. J. R. Statistical Society, Series B. Vol. 26, pp 211-252.

[44] McCuen, R.H. and Snyder, W.M., 1986. Hydrologic Modeling: Statistical Methods and Applications. Prentice-Hall. pp. 134-142.

APPENDIX A

Note: In the following tables, the value '0' means the series passed that test (or: the series is independent);
the value '1' means the series failed that test (or: the series is dependent).

Results of Statistical Tests for Dependence at 5 % Level

Pro-vince	River Name	n	Nonparametric Tests										Parametric Tests			No. of failed tests
			Median crossing	Turning points	Length-of-runs	Rank difference	Cumulative periodgam	Wald-Wolfowitz	Spear-man	RUNAB (random)	RVN Ratio	Hurst K	Autocor-relation	Von-Neumann	Hurst K	
	Athabasca(Athabasca)	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bow	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Castle	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	Drywood Creek	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	Elbow(Glenmore Dam)	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Elbow(Bragg Creek)	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Ghost	40	0	0	0	0	0	0	0	0	1	0	0	1	0	2
	Manyberries Creek	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Rolph Creek	53	0	0	1	0	0	0	0	1	0	0	0	0	1	3
	Sturgeon	54	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	Switcurrent Creek	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Waterton	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Upper Humber	60	0	0	1	1	0	0	0	0	0	0	0	0	0	2
	Lepreau	72	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	Saint John	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	Shogomoc Stream	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	Upsalquitch	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	Beaverbank	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	East	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	Grand	68	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	Lahave	73	0	0	0	0	0	0	0	0	0	1	0	0	0	1
C	Northeast Margaree	72	0	0	1	0	0	0	0	0	0	1	0	0	1	3
	Roseway	71	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Southwest Margaree	70	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	St. Marys(Stillwater)	73	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	Adams	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Ashnola	42	0	0	0	0	0	0	0	0	0	0	0	1	1	2
	Babine	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	Big Sheep Creek	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.	Boundary Creek	61	0	0	0	0	0	0	0	0	0	1	0	0	1	2
C	Bulkley	58	0	1	1	0	0	0	0	0	0	0	0	0	0	2
.	Chilko(Chilko Lake)	60	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Chilko(Redstone)	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Columbia(Donald)	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Columbia(Nicholson)	77	0	1	0	0	0	0	0	0	0	1	0	0	1	3
	Columbia(Fairmont)	43	0	1	1	1	0	0	0	1	1	0	0	0	0	5
	Flathead	60	1	0	0	1	0	0	0	0	1	1	0	0	1	5
	Kettle (Ferry)	60	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Kettle (Laurier)	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Kootenay (Crossing)	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Kootenay (Newgate)	42	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Lardeau	43	1	0	1	1	0	0	0	0	0	0	0	0	0	3

Results of Statistical Tests for Dependence at 5 % Level *(continued)*

	Liard	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Lillooet	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Moyie	59	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Quesnel (Likely)	64	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	Quesnel (Quesnel)	50	0	0	0	0	0	0	0	1	0	0	0	0	0	1
.	Salmo (Salmo)	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	Sikanni Chief	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.	Similkameen	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Skeena	41	0	0	1	0	0	0	0	1	0	0	0	0	0	2
	Slocan	64	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	South Thompson	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	St. Mary (Wycliffe)	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	St. Mary (Marysville)	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Stuart	56	0	0	0	0	0	0	0	0	1	1	0	1	1	4
	North Thompson (Bar.)	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Brokenhead	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Roseau (Cariboun)	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	Roseau (Dominion)	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	Sprague Creek	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Turtle	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Whitemouth	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Ausable	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Black	73	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Castor	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	English (Umfreville)	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	English (Sioux L.)	60	0	0	0	0	0	0	0	0	0	1	0	0	1	2
O	Missinaibi	69	0	0	0	0	0	0	0	0	0	1	0	0	1	2
N	Namakan	66	0	0	0	0	0	0	0	0	0	0	0	1	0	1
T	Nith (Canning)	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	North Magnetawan	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nottawasaga	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pigeon	65	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Saugeen (Port Elgin)	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Saugeen (Walker)	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	South Nation	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sydenham (Alvinston)	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sydenham (Owen Sound)	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Turtle (Mine Centre)	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hall	40	0	0	1	0	0	0	0	0	0	0	0	0	0	1
P	Harricana	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	Petite Nation (Portage–)	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Petite Nation (Pres De)	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Richelieu (Aux R.)	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Horse Creek	43	0	0	1	0	0	0	0	0	0	0	0	0	0	1
S	McEachern Creek	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	Poplar	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Whitewater Creek	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
YK	Teslin	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
No. of Rivers failed in one test			2	3	10	4	0	0	0	5	4	15	0	4	16	63

Results of Statistical Tests for Dependence at 10 % Level

Pro-vince	River Name	n	Nonparametric Tests										Parametric Tests			No. of failed tests
			Median crossing	Turning points	Length-of-runs	Rank difference	Cumulative periodgam	Wald-Wolfowitz	Spear-man	RUNAB (random)	RvN Ratio	Hurst K	Autocor-relation	Von-Neumann	Hurst K	
	Athabasca(Athabasca)	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bow	80	0	1	0	0	0	0	0	0	0	0	0	0	0	1
	Castle	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	Drywood Creek	52	0	0	0	0	0	0	0	0	0	0	0	0	1	1
B	Elbow(Glenmore Dam)	44	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	Elbow(Bragg Creek)	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Ghost	40	1	0	0	1	0	1	1	1	1	0	1	1	0	8
	Manyberries Creek	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Rolph Creek	53	1	0	1	0	0	0	0	1	0	1	0	0	1	5
	Sturgeon	54	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	Swiftcurrent Creek	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Waterton	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Upper Humber	60	0	0	1	1	0	1	1	0	1	0	0	0	0	5
	Lepreau	72	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	Saint John	62	0	0	0	0	0	0	0	0	0	1	0	0	1	2
T	Shogomoc Stream	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	Upsalguitch	45	0	1	0	0	0	0	0	0	0	0	0	0	0	1
A	Beaverbank	67	0	0	0	0	0	0	0	0	0	1	0	0	1	2
N	East	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	Grand	68	0	0	0	0	0	0	0	0	0	0	0	0	1	1
I	Lahave	73	0	0	0	0	0	0	0	0	0	1	0	0	0	1
C	Northeast Margaree	72	0	0	1	0	0	0	0	1	0	1	0	0	1	4
	Roseway	71	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Southwest Margaree	70	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	St. Marys(Stilwater)	73	0	0	1	0	0	0	0	1	0	1	0	0	0	3
	Adams	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Ashncla	42	0	0	0	0	0	1	0	0	1	0	1	1	0	4
	Babine	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	Big Sheep Creek	40	0	0	0	0	0	0	0	C	0	0	0	0	0	0
.	Boundary Creek	61	0	0	0	1	0	0	1	0	1	1	0	0	1	5
C	Bulkley	58	0	1	1	0	0	0	0	0	0	0	0	0	0	2
.	Chilko(Chilko Lake)	60	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Chilko(Redstone)	62	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	Columbia(Donald)	44	0	0	0	0	0	1	0	0	0	0	1	0	0	2
	Columbia(Nicholson)	77	0	1	0	0	0	0	0	0	0	1	0	0	1	3
	Columbia(Fairmont)	43	1	1	1	1	1	1	0	1	1	0	1	1	1	11
	Flathead	60	1	0	0	1	0	0	1	0	1	1	1	0	1	7
	Kettle (Ferry)	60	0	0	0	1	0	0	0	0	0	1	0	0	1	3
	Kettle (Laurier)	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Kootenay (Crossing)	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Kootenay (Newgate)	42	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Lardeau	43	1	0	1	1	0	0	0	1	1	0	0	0	0	5

Results of Statistical Tests for Dependence at 10 % Level *(continued)*

	Liard	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Lillooet	63	0	1	0	0	0	0	0	0	0	0	0	0	0	1
	Moyie	59	0	0	0	0	0	0	0	0	1	1	0	0	1	3
	Quesnel (Likely)	64	0	0	0	0	0	0	0	0	0	0	0	0	1	1
B	Quesnel (Quesnel)	50	1	0	0	0	0	0	0	1	0	1	0	0	1	4
.	Salmo (Salmo)	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	Sikanni Chief	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.	Similkameen	44	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	Skeena	41	1	0	1	0	0	0	0	1	0	0	0	0	0	3
	Slocan	64	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	South Thompson	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	St. Mary (Wycliffe)	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	St. Mary (Marysville)	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Stuart	56	0	0	0	1	0	1	1	0	1	1	1	1	1	8
	North Thompson (Bar.)	44	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Brokenhead	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Roseau (Cariboun)	67	0	0	0	0	0	1	0	0	1	0	1	1	0	4
M	Roseau (Dominion)	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	Sprague Creek	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Turtle	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Whitemouth	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Ausable	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Black	73	0	0	0	0	0	0	0	0	0	1	0	0	1	2
	Castor	41	0	0	0	0	0	1	0	0	0	0	1	1	0	3
	English (Umfreville)	67	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	English (Sioux L.)	60	0	0	0	0	0	0	0	0	0	1	0	0	1	2
O	Missinaibi	69	0	0	0	0	1	0	0	0	0	1	0	0	1	3
N	Namakan	66	0	0	0	0	0	0	0	0	0	1	1	1	1	4
T	Nith (Canning)	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	North Magnetawan	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nottawasaga	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pigeon	65	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Saugeen (Port Elgin)	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Saugeen (Walker)	74	0	0	0	0	0	1	0	0	0	0	0	0	0	1
	South Nation	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sydenham (Alvinston)	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sydenham (Owen Sound)	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Turtle (Mine Centre)	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hall	40	1	0	1	0	0	0	0	0	0	1	0	0	0	3
P	Harricana	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	Petite Nation (Portage–)	46	0	0	1	1	0	0	0	1	0	0	0	0	0	3
	Petite Nation (Pres De)	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Richelieu (Aux R.)	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Horse Creek	43	0	1	1	0	0	0	0	0	0	0	0	0	0	2
S	McEachern Creek	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	Poplar	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Whitewater Creek	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
YK	Teslin	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
No. of Rivers failed in one test			8	7	14	9	2	9	5	10	10	23	9	7	26	139