INFORMATION TO USERS

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

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

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

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

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

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

DISSERTATION

MODELING OF STATIONARY AND NON-STATIONARY HYDROLOGIC PROCESSES

Submitted by

Óli Grétar Blöndal Sveinsson

Civil Engineering

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2002

COLORADO STATE UNIVERSITY

March 18, 2002

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY ÓLI GRÉTAR BLÖNDAL SVEINSSON ENTITLED MODELING OF STATIONARY AND NON-STATIONARY HYDROLOGIC PROCESSES BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Advisor

Co-Advisor

Department Head

ABSTRACT OF DISSERTATION
MODELING OF STATIONARY AND NON-STATIONARY HYDROLOGIC PROCESSES

This dissertation touches on several aspects related to modeling of stationary and non-stationary hydrologic processes. New methods for regional frequency analysis of extreme events are developed, under the concept of the population index flood (PIF). In this method population quantities are used for estimating the index flood, instead of using the sample mean as is done in traditional index flood methods. PIF models are developed for commonly used distributions in hydrology, and procedures for estimating the standard error of at-site quantile estimators are also developed. Extensive simulation experiments are used to test the proposed methods and procedures based on the PIF models. In addition, a Pareto model is developed utilizing only the largest sample order statistics for parameter estimation based on maximum likelihood, and exact formulas for the mean-squared-error of quantile estimators are also derived. Furthermore, shifting mean models are developed for modeling processes that exhibit a type of non-stationarity in the mean, that is represented by sudden shifting patterns. The shifting mean models are formulated under both univariate and multivariate frameworks, and with and without autoregressive AR(1) persistence. Procedures for parameter estimation are explained in detail. The multivariate model is formulated as a contemporaneous shifting mean model and it is further mixed with contemporaneous ARMA models. That is, the multivariate model is capable of modeling mixed systems, where only part of the sites exhibit sudden shifting patterns and the others sites can be represented by a CARMA(p, q) model. The proposed shifting mean models are capable of preserving key

statistical characteristics, and in addition the lag zero spatial correlation in the multivariate models. Numerous examples are presented throughout the dissertation for illustrating the different procedures.

Óli Grétar Blöndal Sveinsson
Civil Engineering Department
Colorado State University
Fort Collins, CO 80523
Spring 2002.

ACKNOWLEDGEMENTS

The author expresses his deepest gratitude to his advisor, Dr. Jose D. Salas, and co-advisor, Dr. Duane C. Boes, for their guidance and encouragement throughout the author's graduate work. Thanks are also due to Dr. Jorge A. Ramirez for his encouragement.

The author is also thankful to other members of his graduate committe: Dr. Peter J. Brockwell for fruitful discussions during competitions on the golf course, and Roger A. Pielke for his assistance.

Support from the Colorado Agricultural Experiment Station project on "Predictability of Extreme Hydrologic Events Related to Colorado's Agriculture" and National Science Foundation grant CMS-9625685 on "Uncertainty and Risk Analysis Under Extreme Hydrologic Events" are gratefully acknowledged. In addition, support of the project "Stochastic Modeling and Simulation of the Great Lakes Net Basin Supplies ", GLERL/NOAA and Hydro-Québec, Canada, is gratefully acknowledged.

PREFACE

This dissertation focuses on two different aspects of modeling of hydrologic processes. Chapters 1–3 focus on frequency analyzes in a regional context, while Chapters 4–7 focus on modeling of processes that can be considered to be non-stationary in the mean.

The different chapters in this dissertation are written in such a way that they can stand independently. That is each chapter has its own abstract, introduction, and conclusions.

In Chapter 1 a new method for regional frequency analysis, dubbed as the population index flood method is introduced. In this method the index flood is taken to be function of the unknown population quantities, while traditional index flood methods have estimated the index flood at each site by the corresponding at-site sample mean. Population index flood models are developed for the most commonly used two- and three-parameter distributions.

In Chapter 2, explicit equations for estimation of standard errors of at-site quantile estimators are derived for the population index flood method with the generalized extreme value distribution as the underlying regional distribution and parameters estimated by maximum likelihood.

In Chapter 3, a Pareto model is used for fitting the upper tail of empirical distributions both in a single-site and regional context. Only the largest sample observations are utilized in estimation of the parameters of the Pareto model using the method of maximum likelihood. In addition, exact formula for the mean-squared-error of quantile estimators is derived.

In Chapter 4 shifting mean models are developed for modeling autocorrelated hydroclimatic processes that show a type of non-stationarity in the mean, and in Chapter 5 the

studies of Chapters 4 are extended to include skewed processes.

In Chapters 6 and 7 an autoregressive AR(1) persistence is added to the shifting mean models in Chapter 4. Both univariate and contemporaneous shifting mean plus persistence models are developed, and the contemporaneous model is further mixed with contemporaneous ARMA models. These persistence models are shown to perform well in modeling of the Great Lakes system.

Contents

Chapter 1

POPULATION INDEX FLOOD METHOD FOR REGIONAL FREQUENCY ANALYSIS

Abstract. Regional frequency analyses based on index flood procedures have been used within the hydrologic community since 1960. It appears that when the index flood method was first suggested the index-flood was taken to be the at-site population mean, which in turn, in the last two or three decades, has been estimated by the at-site sample mean. The objectives of this paper are to investigate the consequences of replacing a population characteristic with its sample counterpart and to propose an analytically correct regional model dubbed as the population index flood (PIF) method. In this method the homogeneity of the region is embedded in the structure of the parameter space of the underlying distribution model. Simulation experiments are conducted to test the proposed PIF method based on the generalized extreme value (GEV) distribution with parameters estimated using the method of maximum likelihood (MLE) and the method of probability weighted moments (PWM). Furthermore, in the simulation experiments the PIF method is compared with the Hosking and Wallis regional estimation scheme (*HW-scheme*). Comparing among all index flood methods investigated herein, the PIF method with parameters estimated using MLE provides the best overall results for the 0.95 and the 0.99 quantiles in terms of both bias and root mean square error for moderate to sufficiently large sample sizes, but for the 0.995 quantile the *HW-scheme* seems to perform best for the investigated sample sizes.

1.1 Introduction

Frequency analysis of multi-site hydrologic data, such as annual maximum floods or annual maximum precipitation, is often used to improve estimation of extreme quantiles. This approach is usually referred to as regional frequency analysis. Two methods that have been widely used in regional frequency analysis are *regional regression* and the *index flood method*. Regional regression equations generally relate event quantiles to physiographic and climatic characteristics using least squares regression techniques, while in the index flood method the event quantile is the product of the at-site index-flood (usually the mean) times a regional quantile estimate. In this paper the index flood method is critically examined.

Dalrymple (1960) introduced the index flood method. The main assumption of the method is that, for a statistically homogeneous region of m sites, data at each site divided by the the index-flood are from the same population. Here a statistically homogeneous region is defined as a region where the sites within the region have identical frequency distribution apart from a scale factor. Dalrymple defined the index-flood as the mean annual flood, which in turn he estimated as the 2.33-yr event from the empirical frequency curve for each site. Note that the population mean of a Gumbel distribution has a return period of 2.33 years, which suggests that Dalrymple must have assumed the Gumbel model as prototypic, although it is not explicitly stated in his 1960 paper.

Several references in the literature followed closely Dalrymple's method by taking the 2.33-yr event as the index-flood. As we are going to see subsequently in this paper, this procedure will work only in Gumbel's world. Several well known references in the literature (see for instance, Linsley et al., 1982; Viessman and Lewis, 1996) followed closely Dalrymple's method of using the 2.33-yr event as the index-flood. Unfortunately, often in the same references, rather than referring to the 2.33-yr event as the index-flood under the Gumbel model assumption, the index-flood was referred to as the "mean annual flood", which in a general context can either be taken as a population property or a sample property. Also very

well known references in the extreme flood and extreme rainfall literature (see for instance, NERC, 1975; Hosking et al., 1985a; Hosking and Wallis, 1997) assumed and popularized the use of the sample mean as the index-flood.

Furthermore, in the last two decades, the so-called *probability weighted moments* (PWMs) (Greenwood et al., 1979) and L-moments (Hosking, 1990) were developed as an efficient procedure for estimating parameters of probabilistic models. Estimators of GEV parameters for single site data using PWMs compared favorably with estimators obtained by the maximum likelihood estimation (MLE) method (Hosking et al., 1985b). In addition, parameter estimation using PMWs or L-moments is often easier than MLE. An alternative index flood estimation method based on PWMs and L-moments has been proposed by Hosking et al. (1985a); Hosking and Wallis (1997). In this method the above mentioned index-flood at each site is estimated by the at-site sample mean.

In this paper, we analyze the validity of the assumptions underlying the index flood method. Furthermore, alternative index flood regional models will be developed based on the concepts first suggested by Boes et al. (1989) utilizing the Weibull distribution. The alternative regional models will be called "Population Index-Flood" (PIF) because they will arise from using certain population quantities in arriving at an index-flood. PIF models will be given for some commonly used two- and three-parameter distributions and simulation studies will be conducted based on the GEV distribution where the PIF model is compared with the currently used Hosking and Wallis index flood regional PWM procedure.

1.2 The Index Flood Regional Procedure

The *index flood method* was introduced by Dalrymple (1960). The assumptions made in this method may be summarized as follows: (a) observations at any given site are independent and identically distributed, (b) observations at different sites are independent, and (c) frequency distributions at different sites are identical apart from a "scale factor" (the

index-flood). Mathematically for m sites in a region, the *qth* quantile at site j is given by

$$\xi_j(q) = \mu_j \cdot \xi_R(q) \qquad , j = 1, \ldots, m \tag{1.1}$$

where μ_j is the *index-flood* for site j, and $\xi_R(q)$ is the regional *qth* quantile. The above equation can be rewritten as

$$\frac{\xi_j(q)}{\mu_j} = \xi_R(q) \qquad , j = 1, \ldots, m \tag{1.2}$$

which says $\xi_j(q)/\mu_j$ should not depend on j. Dalrymple (1960) defined the index-flood as the mean annual flood, which he estimated as the 2.33-*yr* event from the empirical frequency curve for each site. As mentioned earlier, this assumes a Gumbel model as the underlying regional distribution. Dalrymple then defined the empirical regional *qth* quantile as the median of the indexed at-site *qth* quantiles estimated from the empirical at-site growth curves. A regional growth curve was then fitted to the empirical regional quantiles.

As noted previously, it has become a common practice among hydrologists to estimate the *index-flood*, μ_j, by the at-site sample mean, $\hat{\mu}_j = \overline{X}_j$, where the 'hat' represents estimated value. Dalrymple (1960) warned about using the at-site sample mean, which gives the same weight to all observations and is therefore more sensitive to large sampling errors of extreme observations.

1.3 The Effect of Using the Sample Mean to Estimate the Index Flood

1.3.1 Sample Properties

Assume a random sample of size n of independent and identically distributed (*iid*) random variables, $X_1, \ldots, X_n$. Indexing this random sample by the sample mean, $\overline{X}$, results in the random sequence, $X_1/\overline{X}, \ldots, X_n/\overline{X}$. The following proposition shows that any two distinct random variables $X_i/\overline{X}$ and $X_j/\overline{X}$ from the sequence are equicorrelated, and that the correlation coefficient does not depend on the marginal distribution of the X_i's.

Proposition: Assume an *iid* sample $X_1, \ldots, X_n$ with sample mean $\overline{X}$ for which second or-

der moments of $X_1/\overline{X}, \ldots, X_n/\overline{X}$ exist. Any two distinct random variables of $X_1/\overline{X}, \ldots, X_n/\overline{X}$ are equicorrelated with correlation coefficient

$$\rho(X_i/\overline{X}, X_j/\overline{X}) = -\frac{1}{n-1} \quad , \text{ for } i \neq j \tag{1.3}$$

where n is the sample size.

Proof: $X_1/\overline{X}, \ldots, X_n/\overline{X}$ are identically distributed, hence they have a common variance, say τ^2. Also, all distinct pairs $X_i/\overline{X}$ and $X_j/\overline{X}$ are identically distributed and so have a common covariance, say γ. Now,

$$\sum_{i=1}^{n} \frac{X_i}{\overline{X}} = n$$

so

$$0 = Var\left(\sum_{i=1}^{n} \frac{X_i}{\overline{X}}\right) = \sum_{i=1}^{n} Var\left(\frac{X_i}{\overline{X}}\right) + \sum_{i \neq j} Cov\left(\frac{X_i}{\overline{X}}, \frac{X_j}{\overline{X}}\right) = n\tau^2 + n(n-1)\gamma$$

Hence,

$$\begin{aligned} \rho(X_i/\overline{X}, X_j/\overline{X}) &= \frac{Cov(X_i/\overline{X}, X_j/\overline{X})}{\sqrt{Var(X_i/\overline{X}) \cdot Var(X_j/\overline{X})}} \\ &= \frac{\gamma}{\tau^2} = \frac{\gamma}{-(n-1)\gamma} = -\frac{1}{n-1} \end{aligned} \tag{1.4}$$

QED

In addition, if the random variables $X_1, \ldots, X_n$ are from a positive process, then the indexed random variables $X_1/\overline{X}, \ldots, X_n/\overline{X}$ are bounded from above, i.e.

$$\frac{X_i}{\overline{X}} = n\frac{X_i}{X_1 + \cdots + X_n} < n \tag{1.5}$$

Hence, any two random samples of different sizes taken from the same positive population, say $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_m$ where $n \neq m$, and indexed by their sample means will not be identically distributed since their respective ranges

$$0 < \frac{Y_i}{\overline{Y}} < n \qquad , i = 1, \ldots, n \tag{1.6}$$

$$0 < \frac{Z_j}{\overline{Z}} < m \qquad , j = 1, \ldots, m \tag{1.7}$$

are different.

The implication of the foregoing analysis is rather obvious. Using the index flood method with the index-flood as the sample mean leads to a sample that is not *iid* but correlated and to samples that may not be identically distributed. The distortion that results from dividing by the sample mean has also been discussed by Stedinger (1983).

1.3.2 Analytical Example

As a simple analytical example, let's assume that we have a random sample of size n from the one parameter exponential distribution. That is, $X_1, \ldots, X_n \overset{iid}{\sim} exp(\beta)$ with probability density function (PDF) and cumulative distribution function (CDF) given by

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta} I_{(0,\infty)}(x) \tag{1.8}$$

$$F_X(x) = \left(1 - e^{-x/\beta}\right) I_{(0,\infty)}(x) \tag{1.9}$$

respectively, where $\beta > 0$ is a scale parameter and $I_{(a,b)}(x)$ is the indicator function which is equal to one if $x \in (a, b)$ but zero otherwise. The mean and the variance are

$$E[X] = \beta \tag{1.10}$$

$$Var(X) = \beta^2 \tag{1.11}$$

A correct use of the index flood method in which the index-flood, μ, in Eq (1.1) is the population mean (equal to β in our example) yields

$$\frac{X_1}{\mu}, \ldots, \frac{X_n}{\mu} \overset{iid}{\sim} exp(\beta = 1) \tag{1.12}$$

i.e. the distribution of X_i/μ is $exp(1)$ and only differs from Eqs (1.8) and (1.9) by the scale factor β. On the other hand, if the index-flood is taken to be the sample mean, $\overline{X}$, the indexed random variables

$$\frac{X_1}{\overline{X}}, \ldots, \frac{X_n}{\overline{X}} \tag{1.13}$$

are identically distributed but not independent. Furthermore, the distribution of the random variables in (1.13) is not exponential. In fact, the ratio $X_1 / \sum_{i=1}^{n} X_i$ has a beta distribution

with parameters $(1, n-1)$, (see Johnson et al., 1994, pg. 507-508). Hence, the variables $Y_i = X_i/\overline{X}$, $i = 1, \ldots, n$ are identically distributed as

$$Y_1, \ldots, Y_n \sim f_Y(y) = \frac{n-1}{n}\left(1 - \frac{y}{n}\right)^{n-2} I_{(0,n)}(y) \tag{1.14}$$

but not independent. Note that $f_Y(y) \to \exp(-y)\, I_{(0,\infty)}(y)$ as $n \to \infty$. How large does n have to be for $f_Y(y)$ to be a good approximation of the distribution of X_i/μ? Figure 1.1 shows comparison of the PDF's of X_i/μ and $Y_i = X_i/\overline{X}$ for $n \in \{2, 5, 10, 20\}$. The differences of the two PDF's become less as n increases, but for small n's the differences are significant. Perhaps a better comparison can be seen in Fig. 1.2, where the 0.99 quantiles of both distributions are compared for sample size n ranging from 10–500. The 0.99 quantile is underestimated for all n's when the index-flood is estimated by the sample mean, and the underestimation is significant for small n's.

1.4 Simulation Study based on the GEV Distribution and L–moments

In this section a simulation experiment is conducted in order to compare the quantile estimates that can be obtained by the index flood approach assuming that the index-flood μ_j of Eq (1.1) is taken to be the at-site sample mean $\overline{X}_j$ for site j. The simulation study considers a region of three sites, and the generalized extreme value (GEV) distribution with parameters estimated using L–moments as in Hosking and Wallis (1997). The Hosking and Wallis method is summarized in sections 1.4.1-1.4.3, and the simulation results are presented in section 1.4.4.

In this simulation the 0.99 quantile will be estimated for each of the three sites in the region. The random variable for site j is denoted by X_j, $j = 1, 2, 3$, and the distribution of these random variables differs by a scale factor, such that $X_2 = 2X_1$ and $X_3 = 4X_1$. Annual maximum precipitation and flood data often follow a GEV with a negative shape parameter. Thus, the population parameters (refer to Eq (1.20)) considered for site 1 are $[\alpha_1, \beta_1, \kappa_1] = [2.0, 1.0, -0.2]$. The population parameters for the two other sites are derived from the

population parameters for site 1 using the above scale factors. The CDF for site 2 written in terms of the CDF for site 1 is $F_{X_2}(x) = Prob(X_2 \leq x) = Prob(X_1 \leq x/2) = F_{X_1}(x/2)$. Thus, the population parameters for site 2 are $[\alpha_2, \beta_2, \kappa_2] = [4.0, 2.0, -0.2]$. Similarly the population parameters for site 3 are $[\alpha_3, \beta_3, \kappa_3] = [8.0, 4.0, -0.2]$.

1.4.1 Probability Weighted Moments and L–moments

The use of probability weighted moments (PWM) has gained popularity in hydrologic frequency analysis since the late 1970's. The PWM of a random variable X with CDF $F_X(x)$ is defined as (Greenwood et al., 1979)

$$M_{p,r,s} = E[X^p F_X^r(x)(1 - F_X(x))^s] \tag{1.15}$$

where p, r, and s are nonnegative integers. For estimation of upper extreme events the PWM $\beta_r = M_{1,r,0}$ is commonly used. An unbiased estimate of β_r for a sample of size n, $\{X_i\}_{i=1}^n$, is given by

$$\hat{\beta}_r = \frac{1}{n}\binom{n-1}{r}^{-1} \sum_{j=r+1}^{n} \binom{j-1}{r} x_{j:n} \quad , r = 0, 1, \ldots \tag{1.16}$$

where $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$ are the values of the sample order statistics. The $(r+1)$th L–moment, λ_{r+1}, is defined as a linear combination of the PWM's (Hosking, 1990)

$$\lambda_{r+1} = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k}\binom{r+k}{k} \beta_k \quad , r = 0, 1, \ldots \tag{1.17}$$

L–moment-ratios are dimensionless L–moments defined by

$$\tau_2 = \lambda_2/\lambda_1 \tag{1.18}$$

$$\tau_r = \lambda_r/\lambda_2 \quad , r = 3, 4, \ldots \tag{1.19}$$

where τ_2 (*L–CV*), τ_3 (*L–Skewness*), and τ_4 (*L–Kurtosis*) are alternative measures of coefficient of variation, skewness, and kurtosis respectively. In addition, λ_1 is the mean of the distribution and λ_2 is a measure of scale. Unbiased estimates of λ_r and estimates of τ_r are obtained from Eq (1.17)-(1.19) by replacing population quantities with sample quantities.

The sample L-moment-ratios are not unbiased but they are consistent (Hosking and Wallis, 1997).

1.4.2 The GEV Distribution and L-moment Parameter Estimation

The CDF of the GEV distribution with parameters α (location), β (scale) and $\kappa \neq 0$ (shape) is given by

$$F_X(x) = \exp\left\{ -\left[1 - \frac{\kappa}{\beta}(x-\alpha)\right]^{1/\kappa} \right\} \quad , \text{if } \kappa \neq 0 \tag{1.20}$$

where $\alpha + \beta/\kappa \leq x < \infty$ if $\kappa < 0$, and $-\infty < x \leq \alpha + \beta/\kappa$ if $\kappa > 0$. For $\kappa = 0$ the GEV is the Gumbel distribution. The *L-Skewness* of the GEV distribution is a function of κ only:

$$\tau_3 = -3 + 2\,\frac{1-3^{-\kappa}}{1-2^{-\kappa}} \tag{1.21}$$

which can be implicitly inverted to give κ in terms of τ_3. A fifth order polynomial approximation of κ in terms of τ_3 with error in κ less than 8×10^{-6}, for $-1.0 < \kappa \leq 1.0$ is given below:

$$\kappa = \begin{cases} P_1(\tau_3) & \text{if } 1.0 > \tau_3 > 0.1699\ , \ -1.0 < \kappa < 0 \\ P_2(\tau_3) & \text{if } 0.1699 \geq \tau_3 \geq -0.3333\ , \ 0 \leq \kappa \leq 1.0 \end{cases} \tag{1.22}$$

where

$$\begin{aligned} P_1(x) &= 0.2833\,6547 - 1.7898\,9200\,x + 0.7848\,5605\,x^2 \\ &\quad -0.4121\,9549\,x^3 + 0.1683\,9976\,x^4 - 0.0345\,3985\,x^5 \\ P_2(x) &= 0.2837\,7353 - 1.7961\,1310\,x + 0.8225\,7962\,x^2 \\ &\quad -0.5228\,4010\,x^3 + 0.3719\,4789\,x^4 - 0.5247\,8704\,x^5 \end{aligned}$$

Estimates of $(\tau_3, \lambda_2, \lambda_1)$, denoted by $(\hat{\tau}_3, \hat{\lambda}_2, \hat{\lambda}_1)$ are obtained from the data, and then $\hat{\kappa}$ is obtained from $\hat{\tau}_3$ using Eq (1.22), and subsequently,

$$\hat{\beta} = \frac{\hat{\lambda}_2 \hat{\kappa}}{(1-2^{-\hat{\kappa}})\Gamma(1+\hat{\kappa})} \tag{1.23}$$

$$\hat{\alpha} = \hat{\lambda}_1 - \frac{\hat{\beta}}{\hat{\kappa}}(1 - \Gamma(1+\hat{\kappa})) \tag{1.24}$$

1.4.3 Hosking and Wallis Index-Flood Estimation (*HW-scheme*)

For ease of reference the Hosking and Wallis index flood estimation procedure (Hosking and Wallis, 1997) is denoted here as *HW-scheme*. In this estimation scheme the original sample data are indexed by the L-moment $\hat{\lambda}_1$, which is equivalent to the sample mean $\overline{X}$. Then, the regional L-moment-ratios, $\hat{\tau}_r^R$, $r = 2, 3, \ldots$, are estimated as the weighted average of the at-site L-moment-ratios

$$\hat{\tau}_r^R = \frac{1}{\sum_{j=1}^{m} n_j} \sum_{j=1}^{m} n_j \hat{\tau}_r^{(j)} , \quad r = 2, 3, \ldots \tag{1.25}$$

where n_j is the record length for site j, and m is the number of sites in the region. Note that, since in the *HW-scheme* data at each site are scaled by the at-site sample mean, the estimated regional mean must be equal to one, i.e. $\hat{\lambda}_1^R = 1$, and therefore Eq (1.18) yields, $\hat{\tau}_2^R = \hat{\lambda}_2^R$. Given the estimates of the regional L-moment-ratios, then Eqs (1.22)-(1.24) are applied to estimate the GEV parameters of the regional growth curve.

1.4.4 Simulation Results

We will compare the 0.99 quantile estimated by the *HW-scheme* with the quantile estimated by indexing the at-site data by their population means (known values). As opposed to the *HW-scheme* the at-site data indexed by their population means will be treated as one regional sample, i.e. the regional L-moments and L-moment-ratios will be estimated from the regional sample. The argument for using this estimation procedure is: since the at-site data are obtained from *iid* GEV random numbers and since the distributions at different sites are the same apart from a scale factor, then the data at different sites scaled by the true population mean should be *iid*. This procedure will be referred to as "analytic". Although it is not viable in practice it will serve our purposes here for comparison.

The simulation results are shown in Fig. 1.3, where $n_1 = n_2 = n_3$. As expected the results from the analytic procedure agree closely with the at-site 0.99 population quantiles while the *HW-scheme* underestimates the 0.99 quantiles. This underestimation is significant

for small sample sizes but decreases as sample sizes get larger. Hence, these results show that the *HW-scheme* can result in significant underestimation of quantiles. Furthermore, note that the simulation experiments reported in section 1.6 of this paper show that estimating the regional L-moment-ratios using Eq (1.25) results in estimated quantiles with larger bias than if the regional L-moment-ratios were estimated from weighted at-site L-moments.

1.5 Analytical Models for Regional Frequency Analysis

In this section a family of analytical models for use in regional frequency analysis will be suggested. Models of this kind were first introduced by Boes et al. (1989).

Let's assume m sites in a statistically homogeneous region, and denote the population mean at site j by μ_j and the population variance at site j by σ_j^2, $j = 1, \ldots, m$. The CDF and the PDF at site j are denoted by $F_{X_j}(x)$ and $f_{X_j}(x)$, respectively, and the *qth* population quantile at site j is denoted by $\xi_j(q)$. Let $\boldsymbol{\theta}_j$ be the parameter vector of $F_{X_j}(x)$ and $f_{X_j}(x)$, and let Θ_j be the corresponding parameter space. The homogeneity of the region is embedded in the structure built into the parameter space. This structure results from assuming, for instance, that for some indexing function $g(\cdot)$ and $\boldsymbol{\theta}_j \in \Theta_j$, the homogeneity of the region can be represented by assuming that $\xi_j(q)\, g(\boldsymbol{\theta}_j, q)$ is the same for all sites in the region, i.e. is independent of j. Thus homogeneity implies, that at-site population quantiles indexed by the indexing function $g(\cdot)$ are identical. In general the indexing function $g(\boldsymbol{\theta}_j, q)$ is only a function of the population characteristics of the distribution model under consideration. Thus, we denote the resulting regional model as "Population Index-Flood" (PIF) or PIF regional frequency analysis so as to emphasize the essence of the approach (based on population indexing) and to distinguish it from the usual indexing based on the sample mean. Two types of indexing will be considered here: (1) at-site population quantiles divided by the at-site population mean (μ_j), and (2) at-site population quantiles standardized using the at-site population mean (μ_j) and the at-site population standard deviation (σ_j).

The assumption that the at-site population quantiles divided by their population mean are identical implies that any of

$$\frac{\xi_j(q)}{\mu_j} \tag{1.26}$$

$$F_{X_j}(x \cdot \mu_j) \tag{1.27}$$

$$\text{abs}(\mu_j) f_{X_j}(x \cdot \mu_j) \tag{1.28}$$

should not depend on j.

Similarly, the assumption of standardized at-site population quantiles being identical implies that any of

$$\frac{\xi_j(q) - \mu_j}{\sigma_j} \tag{1.29}$$

$$F_{X_j}(x \cdot \sigma_j + \mu_j) \tag{1.30}$$

$$\sigma_j f_{X_j}(x \cdot \sigma_j + \mu_j) \tag{1.31}$$

should not depend on j.

The use of such regional model can be seen in Boes et al. (1989), Sveinsson and Salas (2001), and Chapters 2 and 3.

1.5.1 Two Parameter Distributions

In this section the suggested PIF regional models will be studied for some commonly used two parameter distributions shown in Table 1.1. The distributions will be classified into two groups, where the first group consist of distributions with location and scale parameters and the second group consist of distributions with scale and shape parameters.

The exponential and the Gumbel distribution are in the first group. For a statistically homogeneous region, it is easily shown for these two distributions that $\xi_j(q)/\mu_j$ independent of j implies that the ratio α_j/β_j, say γ, is the same for all sites. Hence, for m sites in a region the parameter space is reduced from $2m$-dimensions to $(m+1)$-dimensions, i.e. if e.g. the estimate $\hat{\alpha}_j$ is obtained at each site j, then the estimate of β_j is obtained by $\hat{\beta}_j = \hat{\alpha}_j/\hat{\gamma}$,

where $\hat{\gamma}$ is the estimate of γ. On the other hand, $(\xi_j(q) - \mu_j)/\sigma_j$ independent of j does not reduce the parameter space, since it does not depend on the parameter vector $\boldsymbol{\theta}_j$. Hence, α_j and β_j have to be estimated individually at each site j in the region.

The gamma, the lognormal, the Pareto, and the Weibull distribution are in the second group. For a statistically homogeneous region, it is easily shown that for all these models $\xi_j(q)/\mu_j$ or $(\xi_j(q) - \mu_j)/\sigma_j$ independent of j implies that the shape parameter κ_j is the same at each site j in the region. Thus, the parameter space reduces from $2m$-dimensions to $(m+1)$-dimensions, where β_j is estimated individually for each of the m sites in the region but the shape parameter, say κ, is estimated commonly for all sites.

Boes et al. (1989) applied the regional Weibull model for estimation of annual flood quantiles. They also derived exact formulas for the Cramer-Rao lower bound for the variance of unbiased quantile estimators. In Chapter 3 the regional Pareto model is used to estimate upper quantiles from the upper order statistics of a Pareto distribution, and exact formulas are derived for the mean-squared-error of quantile estimators. In both Boes et al. (1989) and Chapter 3 parameters are estimated using the method of maximum likelihood.

1.5.2 Three Parameter Distributions

In this section the implications of the suggested regional model on some commonly used three parameter distributions in Table 1.2, will be analysed. The distributions will be classified into groups depending on the implications of the regional model.

The GEV, the Generalized Pareto, and the Pearson Type III distribution are in the first group. For a statistically homogeneous region, $\xi_j(q)/\mu_j$ independent of j implies that the ratio of $\alpha_j/\beta_j = \gamma$ is the same for all sites and that $\kappa_j = \kappa$ is the same for all sites. Hence, the parameter space reduces from $3m$-dimensions to $(m+2)$-dimensions. $(\xi_j(q) - \mu_j)/\sigma_j$ independent of j implies that the shape parameter $\kappa_j = \kappa$ is the same for all sites in the region. Thus, the parameter space reduces from $3m$-dimensions to $(2m+1)$-dimensions.

The lognormal distribution is in the second group. For a statistically homogeneous

region, $\xi_j(q)/\mu_j$ independent of j implies that $\alpha_j e^{-\beta_j} = \gamma$ is the same for all sites and that κ_j is the same for all sites, i.e. the parameter space reduces from $3m$-dimensions to $(m+2)$-dimensions. On the other hand, $(\xi_j(q) - \mu_j)/\sigma_j$ independent of j implies that $\kappa_j = \kappa$ is the same for all sites, reducing the parameter space from $3m$-dimensions to $(2m+1)$-dimensions.

In the last group is the Log-Pearson Type III (or the three parameter log-gamma) distribution. For the case when $\beta < 0$ (not shown) the implications of the PIF method are the same as for $\beta > 0$. For a statistically homogeneous region, $\xi_j(q)/\mu_j$ or $(\xi_j(q) - \mu_j)/\sigma_j$ independent of j implies that $\beta_j = \beta$ is the same for all sites and that $\kappa_j = \kappa$ is the same for all sites. Hence, the parameter space reduces from $3m$-dimensions to $(m+2)$-dimensions.

Sveinsson and Salas (2001) applied the regional GEV and lognormal models for estimation of annual precipitation quantiles for a small region of three sites in northeastern Colorado. They also compared their results with the *HW-scheme*. In Chapter 2 exact formulas are derived for the standard error of at-site quantile estimators of the regional GEV model with maximum likelihood estimation, and the derived formulas are tested using extensive simulation experiments.

1.6 Comparison Between the PIF Regional Model and the HW-Scheme

In this section a comparison between the suggested PIF regional model and the *HW-scheme* (refer to section 1.4.3) will be made using the GEV distribution for the hypothetical region introduced in section 1.4. The 0.95, 0.99, and 0.995 quantiles will be estimated at each site and their biases and root mean squared errors will be compared. The parameters for the PIF regional model will be estimated by the method of probability weighted moments and by the method of maximum likelihood.

1.6.1 Parameter Estimation for the PIF-GEV Regional Model

For the GEV model given in Table 1.2 and Eq (1.20), $\xi_j(q)/\mu_j$ independent of j implies that either the location parameter α_j or the scale parameter β_j should be estimated

for each site j in the region, and the ratio of $\alpha_j/\beta_j = \gamma$ and the shape parameter $\kappa_j = \kappa$ should be estimated commonly for all sites in the region.

1.6.1.1 Parameter Estimation by Probability Weighted Moments

For m sites in the region the common shape parameter κ is estimated using Eq (1.22). Given $\hat{\kappa}$, an estimate of $\gamma = \alpha_j/\beta_j$ is obtained using Eqs (1.23) and (1.24). It follows

$$\frac{1}{\hat{\gamma}} = \frac{\hat{\tau}_2\,\hat{\kappa}}{(1-2^{-\hat{\kappa}})\Gamma(1+\hat{\kappa}) - \hat{\tau}_2[1-\Gamma(1+\hat{\kappa})]} \tag{1.32}$$

where $\hat{\tau}_2$ is a regional estimate of the *L-CV*. Then, either α_j is estimated at each site j using

$$\hat{\alpha}_j = \hat{\lambda}_1^{(j)} - \hat{\lambda}_2^{(j)}\,\frac{1-\Gamma(1+\hat{\kappa})}{(1-2^{-\hat{\kappa}})\Gamma(1+\hat{\kappa})} \tag{1.33}$$

or an estimate of β_j is obtained at each site j using

$$\hat{\beta}_j = \frac{\hat{\lambda}_2^{(j)}\,\hat{\kappa}}{(1-2^{-\hat{\kappa}})\Gamma(1+\hat{\kappa})} \tag{1.34}$$

where $\hat{\lambda}_1^{(j)}$ and $\hat{\lambda}_2^{(j)}$ are the first and the second sample L-moments at site j, respectively. Given $\hat{\alpha}_j$, then $\hat{\beta}_j = \hat{\alpha}_j/\hat{\gamma}$, and vice versa.

1.6.1.2 Parameter Estimation by the Method of Maximum Likelihood

Assume m sites in the region and denote the sample at site j by $X_{j1},\ldots,X_{jn_j}$ for $j = 1,\ldots,m$, where n_j is the sample size at site j.

For the case when the location parameter α_j is estimated at each site j in the region, and the ratio $\alpha_j/\beta_j = \gamma$ and $\kappa_j = \kappa$ are estimated commonly for all sites in the region. Then for $\theta_j = \alpha_j^{-1}$ the PDF of the GEV (see Table 1.2) at site j can be written as

$$f_{X_j}(x) = \theta_j\,\gamma\,[1-\gamma\,\kappa(\theta_j\,x-1)]^{\frac{1}{\kappa}-1}\cdot\exp\left\{-[1-\gamma\,\kappa(\theta_j\,x-1)]^{1/\kappa}\right\} \tag{1.35}$$

So, the log-likelihood is

$$\begin{aligned}&\ln\mathcal{L}(\theta_1,\ldots,\theta_m,\gamma,\kappa;\mathbf{x}_1,\ldots,\mathbf{x}_m)\\ &\qquad=\sum_{j=1}^{m}\left\{n_j\,(\ln\theta_j+\ln\gamma)+\sum_{i=1}^{n_j}\left[\left(\frac{1}{\kappa}-1\right)\ln\xi_{ji}-\xi_{ji}^{1/\kappa}\right]\right\}\end{aligned} \tag{1.36}$$

where $\mathbf{x}_j = [x_{j1}, \ldots, x_{jn_j}]$ is the sample vector for site j, and $\xi_{ji} = 1 - \gamma\,\kappa(\theta_j\, x_{ji} - 1)$. The partial derivatives of the log-likelihood with respect to the parameters are given by

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_j} = \frac{n_j}{\theta_j} - \gamma \sum_{i=1}^{n_j} x_{ji} \xi_{ji}^{-1} (1 - \kappa - \xi_{ji}^{1/\kappa}) \quad , j = 1, \ldots, m \tag{1.37}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \gamma} = \sum_{j=1}^{m} \left\{ \frac{n_j}{\gamma} - \sum_{i=1}^{n_j} (\theta_j\, x_{ji} - 1) \xi_{ji}^{-1} (1 - \kappa - \xi_{ji}^{1/\kappa}) \right\} \tag{1.38}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \kappa} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left\{ \frac{1}{\kappa^2} (\xi_{ji}^{1/\kappa} - 1) \ln \xi_{ji} - \frac{\gamma}{\kappa} (\theta_j\, x_{ji} - 1) \xi_{ji}^{-1} (1 - \kappa - \xi_{ji}^{1/\kappa}) \right\} \tag{1.39}$$

A similar procedure is followed for the case when the scale parameter β_j is estimated at each site j in the region and the ratio $\alpha_j/\beta_j = \gamma$ and $\kappa_j = \kappa$ are estimated commonly for all sites.

The ML-estimates of the parameters are obtained by setting the partial derivatives equal to zero and solving them simultaneously for the parameters. The above equations can not be solved explicitly for the parameters, so a Newton Rhapson iterative procedure with exact Hessian matrix of the log-likelihood is used to solve the system of equations numerically. The iterative procedure is repeated until the change in all estimated parameters is less than 0.1%.

1.6.2 Results

Recall that the region in section 1.4 was simulated for equal at-site sample sizes $n_1 = n_2 = n_3$, and unequal at-site sample sizes, where $n_2 = n_1 + 10$ and $n_3 = n_1 + 20$. In case of equal sample sizes, the simulation experiment was conducted for $n_1 \in \{10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100\}$, and for unequal sample sizes the simulation experiment was conducted for $n_1 \in \{10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90\}$. For each sample size the region was simulated 10,000 times and the 0.95, 0.99, and 0.995 quantiles were estimated at each site. Then average quantile estimates and root mean square errors (rmse) were determined for each sample size and each estimation method. The results obtained for the PIF regional model will be referred to as "MLE 1" and "PWM 1" when the location

parameter α_j is estimated at each site j using the methods of maximum likelihood and PWM's, respectively, and "MLE 2" and "PWM 2" when the scale parameter β_j is estimated at each site j using the methods of maximum likelihood and PWM's, respectively. In all of these methods the ratio $\alpha_j/\beta_j = \gamma$ and the shape $\kappa_j = \kappa$ are estimated commonly for all sites in the region using the procedures outlined in sections 1.6.1.1 and 1.6.1.2. Note that MLE results for equal sample sizes when $n_1 = 10$ are not shown, because in that case the MLE procedure didn't converge for almost 15% of the samples for MLE 1 and 13% of the samples for MLE 2 (refer to Table 1.3 for proportion of non-converged cases of the MLE procedures). As stated before, the results obtained for the Hosking and Wallis index flood procedure will be referred to as the *HW-scheme*. Also recall that for the *HW-scheme* we analyze the at-site data divided by their sample mean, while for PWM 1, PWM 2, MLE 1, and MLE 2 we analyze the original at-site data.

The results for the 0.99 quantile and corresponding rmse's are shown in Fig. 1.4 for equal at-site sample sizes, and in Fig. 1.5 for unequal at-site sample sizes. For PWM 1, PWM 2, and the *HW-scheme* the regional L-moment-ratios are estimated according to Eq (1.25). Results based on MLE 2 are not shown since they appear to be identical to results based on MLE 1 (refer to Tables 1.4-1.7). Both figures show that quantiles estimated based on PWM 1, PWM 2, and the *HW-scheme* are always negatively biased, but the MLE 1 results in positively biased estimated quantiles for small sample sizes and nearly unbiased estimated quantiles for large sample sizes. In summary, for sufficiently large sample sizes, say of size 30 or greater, the MLE 1 (and MLE 2 not shown in the figures) performs better than the other methods. For the methods based on probability weighted moments (PWM's) the PWM 2 has the smallest bias but also the largest rmse. In general the PWM 1 and the *HW-scheme* perform similarly in terms of bias and rmse.

Because of the significant negative bias of quantile estimators based on probability weighted moments, the simulation experiment was repeated for the case when regional L-moment-ratios ($\hat{\tau}_r^R$) are estimated from Eqs (1.18)-(1.19) in terms of regional L-moments

$(\hat{\lambda}_r^R)$, which in turn are estimated as a weighted average of at-site L-moments $(\hat{\lambda}_r^{(j)})$ as

$$\hat{\lambda}_r^R = \frac{1}{\sum_{j=1}^{m} n_j} \sum_{j=1}^{m} n_j \hat{\lambda}_r^{(j)} , \quad r = 1, 2, \ldots \tag{1.40}$$

where n_j is the sample size at site j. This estimation method was once the recommended one in the *HW-scheme* (Hosking et al., 1985a), but instead Hosking and Wallis (1997) recommended using the method based on Eq (1.25), since in their simulation the method based on Eq (1.25) usually resulted in more accurate quantile estimates in terms of mean squared error. Results of our repeated simulation experiment for the 0.99 quantile are shown in Fig. 1.6 for equal at-site sample sizes, and in Fig. 1.7 for unequal at-site sample sizes. The biases of quantile estimators based on PWM 1, PWM 2, and the *HW-scheme* are significantly reduced (compared to the case of using Eq (1.25)) and the rmse shows a slight (perhaps negligible) increase. Estimated quantiles using the *HW-scheme* are always underestimated but they have smaller rmse compared to quantile estimators based on PWM 1 and PWM 2. For small sample sizes the PWM methods (PWM 1, PWM 2, and *HW-scheme*) have smaller biases than MLE 1 (and MLE 2 not shown). For larger sample sizes, say of size 40 or greater, the MLE procedures perform better than the *HW-scheme* and also better than PWM 1 and PWM 2 if both bias and rmse of quantile estimators are considered. Furthermore, for unequal sample sizes the *HW-scheme* seems to perform worse than the other methods in terms of bias for large sample sizes.

Further comparison of the methods for the 0.95 and the 0.995 quantiles are shown in Tables 1.4-1.7 for selected sample sizes. For all cases shown, the bias of the *HW-scheme* is always negative and it is always reduced when regional L-moment-ratios are estimated using Eqs (1.18)-(1.19) and (1.40) (the numbers shown in parenthesis in the referred tables) as opposed to Eq (1.25). The biases of PWM 1 and PWM 2 are also generally reduced when regional L-moment-ratios are estimated in terms of regional L-moments $(\hat{\lambda}_r^R)$. Regarding the 0.95 quantile the MLE procedures perform better than the PWM methods (PWM 1, PWM 2, and *HW-scheme*) for all sample sizes in terms of both bias and rmse of quantile

estimators. The *HW-scheme* generally has the largest bias among all methods and similar rmse as PWM 1. In addition, PWM 2 seems to be more sensitive to sampling variability than the other methods and has the largest rmse among all methods. On the other hand for the 0.995 quantile the *HW-scheme* (refer to numbers shown in parenthesis in Tables 1.6-1.7) performs best among all methods in terms of bias and rmse for small sample sizes. For larger sample sizes the *HW-scheme* also seems to perform best in terms of bias and almost as well as the MLE procedures in terms of rmse.

1.7 Concluding Remarks

The index flood method for regional frequency analysis was investigated to see the effects of using the sample mean as the index-flood. It was shown that an *iid* sample drawn from a given parent distribution, when indexed by the sample mean, yields a different distribution than the parent distribution, and the indexed sample becomes correlated. For example, when a finite random sample is drawn from the one-parameter exponential distribution, it was shown that the random variables indexed by their sample mean are not exponentially distributed but identically beta distributed and not independent. Further, *iid* samples from the same population indexed by their sample mean were shown to lead to different populations if their sample sizes were different. Although these findings may appear to be "minor mathematical details" or insignificant, it turns out that they can be significant in terms of model structure for regional frequency analysis.

An analytical model for regional frequency analysis denoted as Population Index-Flood (PIF) method was introduced. In the PIF model the homogeneity of the region is embedded in the structure built into the parameter space. Such PIF regional method was developed for several well known probability distributions such as the GEV and Log-Pearson III. For these distributions specific conditions of their parameter space implied by the PIF method were determined. These conditions are key for designing appropriate estimation procedures. For the case of the GEV distribution, MLE and PWM estimation

methods were developed.

Simulation experiments were conducted in order to compare alternative estimation techniques for the proposed PIF method. Likewise, the purpose was to compare the PIF method with the well known index-flood Hosking and Wallis estimation scheme (*HW-scheme*). The results of the simulation study led to the following specific conclusions and recommendations: (1) For PWM estimation, the bias of quantile estimators for PWM 1, PWM 2, and the *HW-scheme* are significantly reduced when regional L-moment-ratios are estimated in terms of regional L-moments as in Eq (1.40) (used prior to 1997) as opposed to using Eq (1.25) (suggested more recently). (2) For the PIF regional method the MLE technique is better than the PWM in terms of both bias and rmse. (3) Comparing among all index flood methods investigated herein, namely the PIF method with MLE and two PWM estimation techniques and the usual *HW-scheme*, the PIF with MLE provides the best overall results for the 0.95 and the 0.99 quantiles in terms of both bias and rmse for moderate to sufficiently large sample sizes, but for the 0.995 quantile the *HW-scheme* seems to perform best for the investigated sample sizes.

The *HW-scheme* is approximate. It is based on treating sample observations indexed by their sample mean as independent and identically distributed for each site in the region. Those indexed samples are also treated as having a common distribution, and the same distribution as the non-indexed samples apart from scale factors (the sample means). Nevertheless the *HW-scheme*, although approximate, is quite robust as the results of the simulation study show.

Table 1.1: Commonly used two parameter distributions in hydrology.

Distribution (pdf/cdf/qth quantile)	Parameter Space	Moments
Exponential		
$f_X(x) = \frac{1}{\beta} e^{-(x-\alpha)/\beta} I_{[\alpha,\infty)}(x)$	$-\infty < \alpha < \infty$	$\mu_X = \alpha + \beta$
$F_X(x) = \left(1 - e^{-(x-\alpha)/\beta}\right) I_{[\alpha,\infty)}(x)$	and $\beta > 0$	$\sigma_X^2 = \beta^2$
$\xi(q) = \alpha - \beta \ln(1-q)$		
Gumbel*		
$f_X(x) = \frac{1}{\beta} \exp\left(-\frac{x-\alpha}{\beta} - e^{-(x-\alpha)/\beta}\right) I_{(-\infty,\infty)}(x)$	$-\infty < \alpha < \infty$	$\mu_X = \alpha - \Gamma'(1)\beta$
$F_X(x) = \exp\left(-e^{-(x-\alpha)/\beta}\right) I_{(-\infty,\infty)}(x)$	and $\beta > 0$	$\sigma_X^2 = \beta^2\pi^2/6$
$\xi(q) = \alpha - \beta \ln(-\ln q)$		
Gamma		
$f_X(x) = \frac{1}{\beta\Gamma(\kappa)} \left(\frac{x}{\beta}\right)^{\kappa-1} e^{-x/\beta} I_{(0,\infty)}(x)$	$\beta > 0$	$\mu_X = \beta\kappa$
$F_X(x) = \frac{1}{\Gamma(\kappa)} \int_0^{x/\beta} u^{\kappa-1} e^{-u} du$	and $\kappa > 0$	$\sigma_X^2 = \beta^2\kappa$
Lognormal†		
$f_X(x) = \frac{1}{\kappa x} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \beta}{\kappa}\right)^2\right] I_{(0,\infty)}(x)$	$-\infty < \beta < \infty$	$\mu_X = e^{\beta+\kappa^2/2}$
$F_X(x) = \Phi\left(\frac{\ln x - \beta}{\kappa}\right) I_{(0,\infty)}(x)$	and $\kappa > 0$	$\sigma_X^2 = (e^{\kappa^2} - 1)e^{2\beta+\kappa^2}$
Pareto		
$f_X(x) = \kappa\beta^\kappa x^{-\kappa-1} I_{(\beta,\infty)}(x)$	$\beta > 0$	$\mu_X = \frac{\beta\kappa}{\kappa-1}$, for $\kappa > 1$
$F_X(x) = \left[1 - \left(\frac{\beta}{x}\right)^\kappa\right] I_{[\beta,\infty)}(x)$	and $\kappa > 0$	$\sigma_X^2 = \frac{\beta^2\kappa}{(\kappa-2)(\kappa-1)^2}$,
$\xi(q) = \beta(1-q)^{-1/\kappa}$		for $\kappa > 2$
Weibull		
$f_X(x) = \frac{\kappa}{\beta} \left(\frac{x}{\beta}\right)^{\kappa-1} e^{-(x/\beta)^\kappa} I_{(0,\infty)}(x)$	$\beta > 0$	$\mu_X = \beta\Gamma(1+\kappa^{-1})$
$F_X(x) = \left[1 - e^{-(x/\beta)^\kappa}\right] I_{[0,\infty)}(x)$	and $\kappa > 0$	$\sigma_X^2 = \beta^2[\Gamma(1+2\kappa^{-1})$
$\xi(q) = \beta[-\ln(1-q)]^{1/\kappa}$		$-\Gamma^2(1+\kappa^{-1})]$

* $\Gamma'(1)$ is the first derivative of the gamma function with argument 1.

† $\Phi(\cdot)$ is the CDF of the standard normal random variable.

Table 1.2: Commonly used three parameter distributions in hydrology.

Distribution (pdf/cdf/qth quantile)	Parameter Space	Moments
Generalized extreme value (GEV)		
$f_X(x) = \frac{1}{\beta}\left[1 - \frac{\kappa}{\beta}(x-\alpha)\right]^{\frac{1}{\kappa}-1} F_X(x)$	$-\infty < \alpha < \infty$,	$\mu_X = \alpha + \beta[1-\Gamma(1+\kappa)]/\kappa$, for $\kappa > -1$
$F_X(x)$ see Eq (1.20)	$\beta > 0$ and	$\sigma_X^2 = \beta^2[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]/\kappa^2$,
$\xi(q) = \alpha + \frac{\beta}{\kappa}[1-(-\ln q)^\kappa]$	$-\infty < \kappa < \infty$	for $\kappa > -1/2$
range: $\alpha + \beta/\kappa < x < \infty$ for $\kappa < 0$, and $-\infty < x < \alpha + \beta/\kappa$ for $\kappa > 0$		
Generalized Pareto (GPA)		
$f_X(x) = \frac{1}{\beta}\left[1 - \frac{\kappa}{\beta}(x-\alpha)\right]^{\frac{1}{\kappa}-1}$	$-\infty < \alpha < \infty$,	$\mu_X = \alpha + \frac{\beta}{1+\kappa}$, for $\kappa > -1$
$F_X(x) = 1 - \left[1 - \frac{\kappa}{\beta}(x-\alpha)\right]^{1/\kappa}$	$\beta > 0$ and	$\sigma_X^2 = \frac{\beta^2}{(1+\kappa)^2(1+2\kappa)}$, for $-1/2 < \kappa < 1/2$
$\xi(q) = \alpha + \frac{\beta}{\kappa}[1-(1-q)^\kappa]$	$-\infty < \kappa < \infty$	
range: $\alpha \le x < \infty$ for $\kappa < 0$, and $\alpha \le x \le \alpha + \beta/\kappa$ for $\kappa > 0$		
Pearson Type III		
$f_X(x) = \frac{1}{\beta\Gamma(\kappa)}\left(\frac{x-\alpha}{\beta}\right)^{\kappa-1} e^{-(x-\alpha)/\beta} I_{(\alpha,\infty)}(x)$	$-\infty < \alpha < \infty$,	$\mu_X = \alpha + \beta\kappa$
$F_X(x) = \frac{1}{\Gamma(\kappa)}\int_0^{(x-\alpha)/\beta} u^{\kappa-1}e^{-u}du$	$\beta > 0$ and $\kappa > 0$	$\sigma_X^2 = \beta^2\kappa$
Lognormal		
$f_X(x) = \frac{1}{\kappa(x-\alpha)}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln(x-\alpha)-\beta}{\kappa}\right)^2\right] I_{(\alpha,\infty)}(x)$	$-\infty < \alpha < \infty$,	$\mu_X = \alpha + e^{\beta+\kappa^2/2}$
$F_X(x) = \Phi\left(\frac{\ln(x-\alpha)-\beta}{\kappa}\right) I_{(\alpha,\infty)}(x)$	$-\infty < \beta < \infty$ and $\kappa > 0$	$\sigma_X^2 = (e^{\kappa^2}-1)e^{2\beta+\kappa^2}$
Log-Pearson Type III		
$f_X(x) = \frac{1}{x\beta\Gamma(\kappa)}\left(\frac{\ln x-\alpha}{\beta}\right)^{\kappa-1} e^{-(\ln x-\alpha)/\beta} I_{(e^\alpha,\infty)}(x)$	$-\infty < \alpha < \infty$,	$\mu_X = e^\alpha(1-\beta)^{-\kappa}$
$F_X(x) = \frac{1}{\Gamma(\kappa)}\int_0^{(\ln x-\alpha)/\beta} u^{\kappa-1}e^{-u}du$	$\beta > 0$ and $\kappa > 0$	$\sigma_X^2 = e^{2\alpha}[(1-2\beta)^{-\kappa} - (1-\beta)^{-2\kappa}]$

Table 1.3: Proportion of non-converged cases of the maximum likelihood procedure for the region simulated in the paper. In MLE 1 the location parameter α_j is estimated at each site j, and in MLE 2 the scale parameter β_j is estimated at each site j. In both MLE 1 and MLE 2 the ratio $\alpha_j/\beta_j = \gamma$ and $\kappa_j = \kappa$ are estimated commonly for all sites in the region.

	Equal sample sizes		Unequal sample sizes	
n_1	MLE 1 [%]	MLE 2 [%]	MLE 1 [%]	MLE 2 [%]
10	14.30	12.92	0.72	0.81
15	3.09	2.88	0.52	0.39
20	1.18	1.16	0.33	0.29
25	0.65	0.55	0.21	0.27
30	0.42	0.33	0.18	0.15
40	0.27	0.20	0.14	0.08
50	0.14	0.14	0.04	0.05
60	0.09	0.06	0.04	0.04
70	0.09	0.08	0.03	0.04
80	0.09	0.06	0.01	0.00
90	0.02	0.02	0.01	0.00
100	0.02	0.02		

Table 1.4: The relative bias and the rmse of the 0.95 quantile when $n_1 = n_2 = n_3$. The region was simulated 10000 times. For the PWM methods, the number that are not in parenthesis show results when the regional L–moment-ratios are estimated from Eq (1.25), while the numbers in parenthesis show results when regional L–moment-ratios are estimated in terms of weighted averages of at-site L–moments (Eq (1.40)).

Sample size n_1		20	40	60	80	100
Site 1, 0.95 quantile, $n_1 = n_2 = n_3$						
rbias	MLE 1	-0.40	-0.48	-0.38	-0.20	-0.16
[%]	MLE 2	-0.41	-0.47	-0.38	-0.20	-0.16
	PWM 1	-1.77 (-0.69)	-1.04 (-0.28)	-0.75 (-0.40)	-0.51 (-0.14)	-0.39 (-0.06)
	PWM 2	0.06 (0.29)	-0.57 (-0.50)	0.07 (0.18)	-0.21 (-0.16)	-0.17 (-0.13)
	HW-scheme	-1.22 (-1.04)	-0.90 (-0.79)	-0.51 (-0.43)	-0.42 (-0.36)	-0.32 (-0.27)
rmse	MLE 1	1.09	0.74	0.61	0.52	0.46
	MLE 2	1.09	0.74	0.61	0.52	0.46
	PWM 1	0.98 (1.08)	0.71 (0.77)	0.60 (0.64)	0.51 (0.56)	0.46 (0.50)
	PWM 2	1.72 (1.82)	1.21 (1.28)	1.01 (1.07)	0.85 (0.90)	0.78 (0.82)
	HW-scheme	1.08 (1.08)	0.77 (0.76)	0.65 (0.65)	0.55 (0.55)	0.50 (0.50)
Site 2, 0.95 quantile, $n_1 = n_2 = n_3$						
rbias	MLE 1	-0.37	-0.51	-0.37	-0.17	-0.24
[%]	MLE 2	-0.37	-0.51	-0.37	-0.17	-0.24
	PWM 1	-1.72 (-0.76)	-1.08 (-0.39)	-0.76 (-0.45)	-0.48 (-0.15)	-0.48 (-0.19)
	PWM 2	-0.56 (-0.65)	-0.39 (-0.47)	-0.25 (-0.24)	-0.11 (-0.13)	-0.20 (-0.21)
	HW-scheme	-1.36 (-1.18)	-0.85 (-0.74)	-0.60 (-0.53)	-0.37 (-0.31)	-0.40 (-0.35)
rmse	MLE 1	2.20	1.48	1.20	1.03	0.92
	MLE 2	2.20	1.48	1.21	1.03	0.92
	PWM 1	1.97 (2.15)	1.43 (1.55)	1.18 (1.27)	1.02 (1.10)	0.91 (0.99)
	PWM 2	3.43 (3.44)	2.49 (2.51)	2.03 (2.06)	1.73 (1.75)	1.54 (1.56)
	HW-scheme	2.15 (2.15)	1.57 (1.56)	1.29 (1.29)	1.10 (1.10)	0.98 (0.98)
Site 3, 0.95 quantile, $n_1 = n_2 = n_3$						
rbias	MLE 1	-0.42	-0.39	-0.29	-0.13	-0.23
[%]	MLE 2	-0.41	-0.38	-0.28	-0.12	-0.23
	PWM 1	-1.75 (-1.03)	-0.95 (-0.39)	-0.65 (-0.44)	-0.43 (-0.16)	-0.47 (-0.23)
	PWM 2	-0.65 (-1.33)	0.09 (-0.30)	-0.49 (-0.69)	0.00 (-0.17)	-0.06 (-0.20)
	HW-scheme	-1.42 (-1.24)	-0.65 (-0.53)	-0.61 (-0.53)	-0.30 (-0.24)	-0.35 (-0.30)
rmse	MLE 1	4.41	2.96	2.41	2.07	1.83
	MLE 2	4.40	2.96	2.41	2.07	1.83
	PWM 1	3.96 (4.23)	2.85 (3.06)	2.36 (2.50)	2.04 (2.18)	1.82 (1.96)
	PWM 2	6.79 (6.17)	4.91 (4.52)	4.01 (3.70)	3.46 (3.23)	3.12 (2.92)
	HW-scheme	4.30 (4.29)	3.09 (3.08)	2.54 (2.53)	2.21 (2.21)	1.97 (1.97)

Table 1.5: The relative bias and the rmse of the 0.95 quantile when $n_2 = n_1 + 10$, and $n_3 = n_1 + 20$. The region was simulated 10000 times. For the PWM methods, the number that are not in parenthesis show results when the regional L-moment-ratios are estimated from Eq (1.25), while the numbers in parenthesis show results when regional L-moment-ratios are estimated in terms of weighted averages of at-site L-moments (Eq (1.40)).

Sample size n_1		15	30	50	70	90
Site 1, 0.95 quantile						
rbias	MLE 1	0.14	-0.12	-0.20	-0.23	-0.15
[%]	MLE 2	0.15	-0.16	-0.20	-0.23	-0.15
	PWM 1	-1.32 (-0.25)	-0.87 (-0.17)	-0.64 (-0.17)	-0.56 (-0.22)	-0.41 (-0.11)
	PWM 2	-0.17 (0.02)	-0.18 (-0.06)	-0.10 (-0.02)	-0.27 (-0.21)	-0.12 (-0.07)
	HW-scheme	-0.96 (-0.81)	-0.67 (-0.56)	-0.49 (-0.41)	-0.48 (-0.42)	-0.32 (-0.27)
rmse	MLE 1	1.11	0.82	0.64	0.54	0.47
	MLE 2	1.11	0.86	0.64	0.54	0.47
	PWM 1	1.04 (1.11)	0.79 (0.84)	0.63 (0.67)	0.53 (0.57)	0.47 (0.50)
	PWM 2	2.07 (2.14)	1.45 (1.50)	1.10 (1.15)	0.93 (0.97)	0.82 (0.86)
	HW-scheme	1.16 (1.16)	0.86 (0.86)	0.68 (0.68)	0.58 (0.58)	0.51 (0.51)
Site 2, 0.95 quantile, $n_2 = n_1 + 10$						
rbias	MLE 1	-0.45	-0.52	-0.17	-0.37	-0.31
[%]	MLE 2	-0.49	-0.51	-0.18	-0.37	-0.32
	PWM 1	-1.42 (-0.44)	-1.10 (-0.47)	-0.56 (-0.13)	-0.67 (-0.36)	-0.55 (-0.28)
	PWM 2	-0.65 (-0.66)	-0.40 (-0.41)	-0.18 (-0.19)	-0.18 (-0.19)	-0.29 (-0.30)
	HW-scheme	-1.20 (-1.04)	-0.89 (-0.78)	-0.44 (-0.36)	-0.52 (-0.46)	-0.47 (-0.42)
rmse	MLE 1	1.91	1.49	1.20	1.03	0.92
	MLE 2	2.08	1.49	1.20	1.03	0.92
	PWM 1	1.79 (1.92)	1.44 (1.54)	1.17 (1.26)	1.02 (1.09)	0.92 (0.99)
	PWM 2	3.00 (3.06)	2.51 (2.55)	2.00 (2.03)	1.73 (1.76)	1.55 (1.57)
	HW-scheme	1.92 (1.92)	1.58 (1.58)	1.27 (1.27)	1.10 (1.10)	0.99 (0.99)
Site 3, 0.95 quantile, $n_3 = n_1 + 20$						
rbias	MLE 1	-0.86	-0.56	-0.27	-0.28	-0.32
[%]	MLE 2	-0.86	-0.56	-0.27	-0.28	-0.32
	PWM 1	-1.60 (-0.79)	-1.04 (-0.52)	-0.60 (-0.25)	-0.56 (-0.31)	-0.53 (-0.32)
	PWM 2	-0.32 (-0.74)	-0.29 (-0.56)	-0.05 (-0.25)	-0.24 (-0.39)	-0.20 (-0.32)
	HW-scheme	-1.20 (-1.05)	-0.81 (-0.70)	-0.43 (-0.35)	-0.47 (-0.40)	-0.44 (-0.39)
rmse	MLE 1	3.53	2.80	2.35	2.00	1.78
	MLE 2	3.53	2.80	2.35	2.00	1.78
	PWM 1	3.29 (3.51)	2.70 (2.88)	2.31 (2.47)	1.98 (2.11)	1.78 (1.90)
	PWM 2	5.07 (4.66)	4.24 (3.94)	3.72 (3.47)	3.20 (2.99)	2.92 (2.73)
	HW-scheme	3.54 (3.52)	2.88 (2.87)	2.48 (2.48)	2.11 (2.11)	1.91 (1.90)

Table 1.6: The relative bias and the rmse of the 0.995 quantile when $n_1 = n_2 = n_3$. The region was simulated 10000 times. For the PWM methods, the number that are not in parenthesis show results when the regional L–moment-ratios are estimated from Eq (1.25), while the numbers in parenthesis show results when regional L–moment-ratios are estimated in terms of weighted averages of at-site L–moments (Eq (1.40)).

Sample size n_1		20	40	60	80	100
Site 1, 0.995 quantile, $n_1 = n_2 = n_3$						
rbias	MLE 1	4.86	0.99	0.53	0.42	0.38
[%]	MLE 2	4.82	1.00	0.54	0.43	0.38
	PWM 1	-3.11 (1.92)	-2.09 (0.96)	-1.44 (0.37)	-1.13 (0.42)	-0.83 (0.48)
	PWM 2	-1.24 (1.02)	-1.63 (-0.31)	-0.59 (0.27)	-0.82 (-0.10)	-0.60 (0.00)
	HW-scheme	-2.54 (-0.06)	-1.96 (-0.57)	-1.19 (-0.22)	-1.04 (-0.32)	-0.76 (-0.16)
rmse	MLE 1	4.66	2.62	2.07	1.75	1.55
	MLE 2	4.65	2.62	2.07	1.75	1.55
	PWM 1	3.26 (4.15)	2.40 (2.96)	2.04 (2.45)	1.74 (2.06)	1.58 (1.84)
	PWM 2	4.34 (4.35)	3.07 (3.01)	2.61 (2.54)	2.19 (2.13)	2.00 (1.93)
	HW-scheme	3.40 (3.57)	2.47 (2.55)	2.11 (2.17)	1.79 (1.82)	1.63 (1.66)
Site 2, 0.995 quantile, $n_1 = n_2 = n_3$						
rbias	MLE 1	4.97	0.95	0.56	0.46	0.28
[%]	MLE 2	4.93	0.96	0.56	0.46	0.28
	PWM 1	-3.07 (1.83)	-2.13 (0.84)	-1.45 (0.32)	-1.09 (0.41)	-0.93 (0.35)
	PWM 2	-1.82 (1.11)	-1.37 (0.31)	-0.92 (0.19)	-0.73 (0.17)	-0.65 (0.11)
	HW-scheme	-2.67 (-0.18)	-1.88 (-0.49)	-1.28 (-0.31)	-0.98 (-0.27)	-0.85 (-0.25)
rmse	MLE 1	9.51	5.22	4.14	3.51	3.08
	MLE 2	9.48	5.22	4.14	3.51	3.08
	PWM 1	6.50 (8.27)	4.82 (5.89)	4.06 (4.88)	3.50 (4.11)	3.14 (3.65)
	PWM 2	8.72 (9.44)	6.38 (6.75)	5.27 (5.50)	4.41 (4.55)	3.93 (4.03)
	HW-scheme	6.81 (7.16)	5.05 (5.22)	4.23 (4.35)	3.59 (3.66)	3.21 (3.25)
Site 3, 0.995 quantile, $n_1 = n_2 = n_3$						
rbias	MLE 1	4.95	1.08	0.63	0.50	0.30
[%]	MLE 2	4.93	1.09	0.63	0.52	0.30
	PWM 1	-3.06 (1.52)	-2.01 (0.80)	-1.36 (0.30)	-1.05 (0.38)	-0.92 (0.30)
	PWM 2	-1.90 (2.11)	-0.96 (1.39)	-1.20 (0.39)	-0.61 (0.64)	-0.51 (0.54)
	HW-scheme	-2.70 (-0.22)	-1.70 (-0.31)	-1.31 (-0.34)	-0.91 (-0.19)	-0.80 (-0.20)
rmse	MLE 1	19.65	10.49	8.28	7.03	6.16
	MLE 2	19.59	10.49	8.29	7.04	6.16
	PWM 1	13.13 (16.43)	9.63 (11.66)	8.13 (9.64)	6.98 (8.17)	6.27 (7.27)
	PWM 2	17.38 (20.84)	12.47 (14.65)	10.33 (12.05)	8.86 (10.14)	7.96 (9.05)
	HW-scheme	13.71 (14.38)	9.95 (10.28)	8.35 (8.58)	7.20 (7.33)	6.45 (6.55)

Table 1.7: The relative bias and the rmse of the 0.995 quantile when $n_2 = n_1 + 10$, and $n_3 = n_1 + 20$. The region was simulated 10000 times. For the PWM methods, the number that are not in parenthesis show results when the regional L–moment-ratios are estimated from Eq (1.25), while the numbers in parenthesis show results when regional L–moment-ratios are estimated in terms of weighted averages of at-site L–moments (Eq (1.40)).

Sample size n_1		15	30	50	70	90
Site 1, 0.995 quantile						
rbias	MLE 1	3.42	1.48	0.81	0.43	0.33
[%]	MLE 2	3.44	1.44	0.81	0.43	0.33
	PWM 1	-2.65 (1.80)	-1.88 (1.07)	-1.26 (0.79)	-1.13 (0.27)	-0.88 (0.36)
	PWM 2	-1.45 (0.62)	-1.20 (0.19)	-0.73 (0.23)	-0.85 (-0.23)	-0.59 (-0.01)
	HW-scheme	-2.27 (-0.22)	-1.68 (-0.29)	-1.11 (-0.15)	-1.06 (-0.33)	-0.80 (-0.21)
rmse	MLE 1	3.83	2.78	2.12	1.80	1.56
	MLE 2	3.83	2.83	2.12	1.80	1.56
	PWM 1	3.15 (3.82)	2.55 (3.02)	2.07 (2.45)	1.79 (2.06)	1.58 (1.83)
	PWM 2	4.75 (4.77)	3.46 (3.42)	2.71 (2.65)	2.31 (2.24)	2.06 (1.99)
	HW-scheme	3.34 (3.48)	2.63 (2.72)	2.13 (2.18)	1.84 (1.88)	1.63 (1.66)
Site 2, 0.995 quantile, $n_2 = n_1 + 10$						
rbias	MLE 1	2.83	1.03	0.84	0.27	0.16
[%]	MLE 2	2.79	1.04	0.83	0.27	0.16
	PWM 1	-2.70 (1.61)	-2.14 (0.72)	-1.18 (0.81)	-1.25 (0.12)	-1.02 (0.20)
	PWM 2	-1.96 (0.48)	-1.41 (0.28)	-0.78 (0.39)	-0.75 (0.04)	-0.76 (-0.04)
	HW-scheme	-2.49 (-0.47)	-1.91 (-0.53)	-1.06 (-0.10)	-1.10 (-0.38)	-0.94 (-0.36)
rmse	MLE 1	7.30	5.31	4.15	3.50	3.09
	MLE 2	7.47	5.31	4.15	3.50	3.09
	PWM 1	6.01 (7.34)	4.89 (5.79)	4.05 (4.79)	3.51 (4.06)	3.16 (3.67)
	PWM 2	7.67 (8.00)	6.40 (6.67)	5.15 (5.32)	4.46 (4.55)	3.97 (4.04)
	HW-scheme	6.17 (6.43)	5.08 (5.26)	4.17 (4.27)	3.62 (3.68)	3.25 (3.30)
Site 3, 0.995 quantile, $n_3 = n_1 + 20$						
rbias	MLE 1	2.39	0.95	0.76	0.36	0.15
[%]	MLE 2	2.40	0.96	0.75	0.36	0.15
	PWM 1	-2.89 (1.21)	-2.10 (0.62)	-1.20 (0.71)	-1.15 (0.15)	-1.00 (0.15)
	PWM 2	-1.51 (1.74)	-1.31 (0.92)	-0.62 (0.99)	-0.83 (0.28)	-0.66 (0.33)
	HW-scheme	-2.45 (-0.42)	-1.86 (-0.48)	-1.02 (-0.07)	-1.05 (-0.33)	-0.90 (-0.32)
rmse	MLE 1	14.26	10.35	8.31	6.95	6.10
	MLE 2	14.25	10.35	8.30	6.95	6.10
	PWM 1	11.69 (14.18)	9.52 (11.27)	8.12 (9.57)	6.93 (7.97)	6.26 (7.21)
	PWM 2	14.37 (16.70)	11.66 (13.34)	10.02 (11.47)	8.49 (9.59)	7.73 (8.74)
	HW-scheme	12.17 (12.69)	9.81 (10.10)	8.37 (8.57)	7.10 (7.23)	6.42 (6.52)

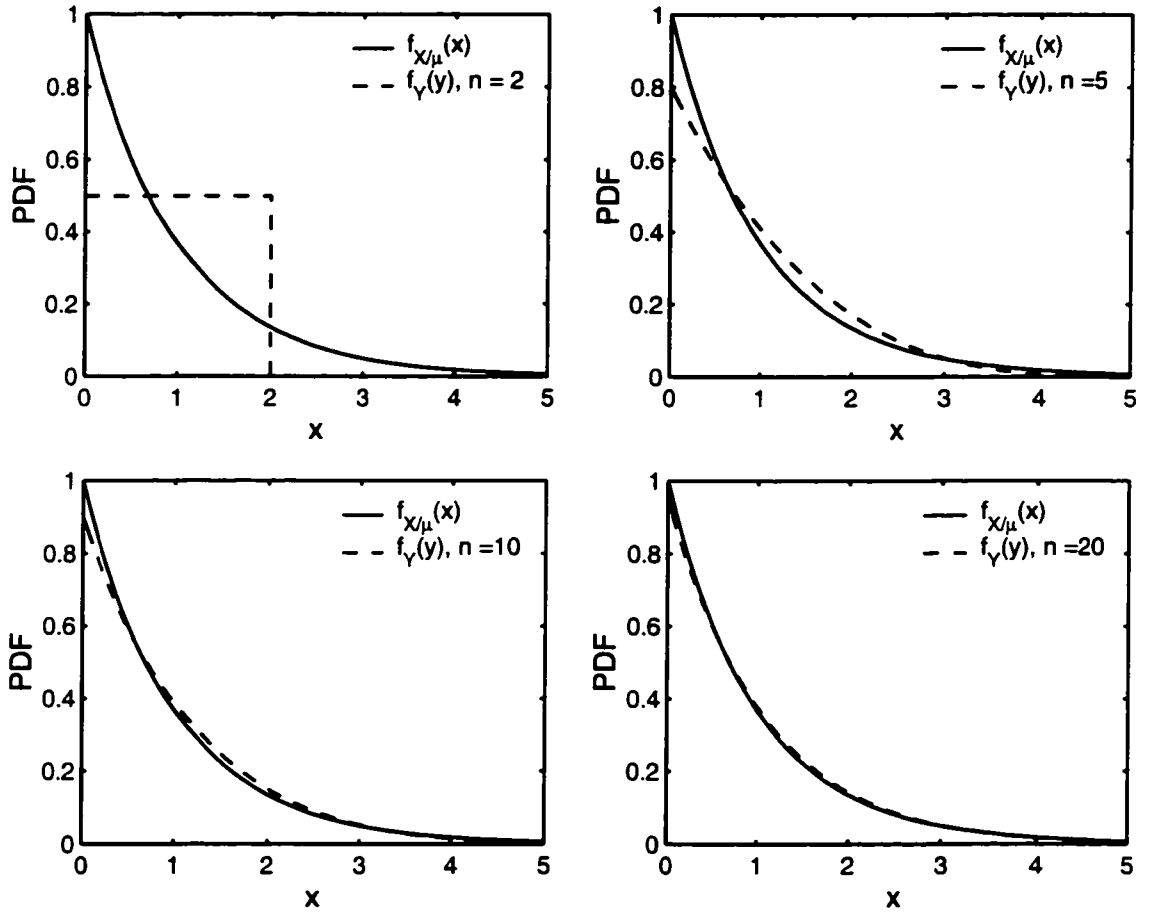

Figure 1.1: Comparison of the PDF's of X/μ and $Y = X/\overline{X}$ for different random sample sizes n, when $X \sim exp(\beta)$.

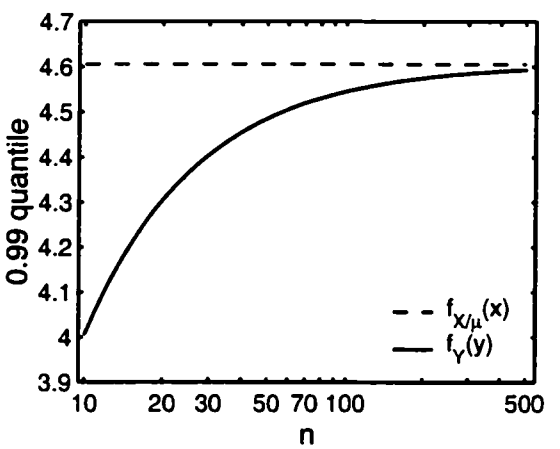

Figure 1.2: Comparison of the 0.99 quantile of the CDF's of X/μ and $Y = X/\overline{X}$ for n from 10–500, when $X \sim exp(\beta)$.

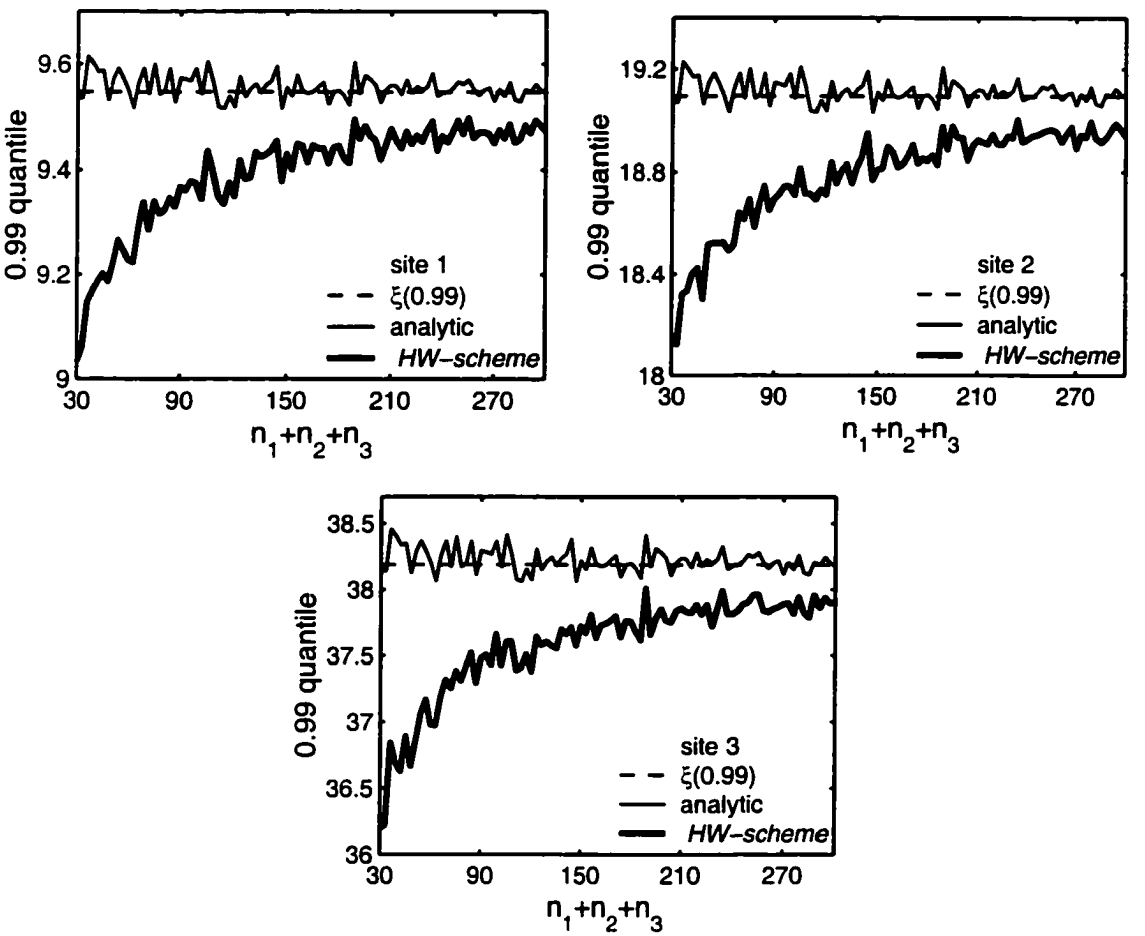

Figure 1.3: Simulation results based on the GEV distribution for a region of three sites, where n_j is the sample size for site j and $n_1 = n_2 = n_3$. For each $n_1 + n_2 + n_3$ the 0.99 quantile is estimated as an average of 5,000 runs. $\xi(0.99)$ is the 0.99 population quantile.

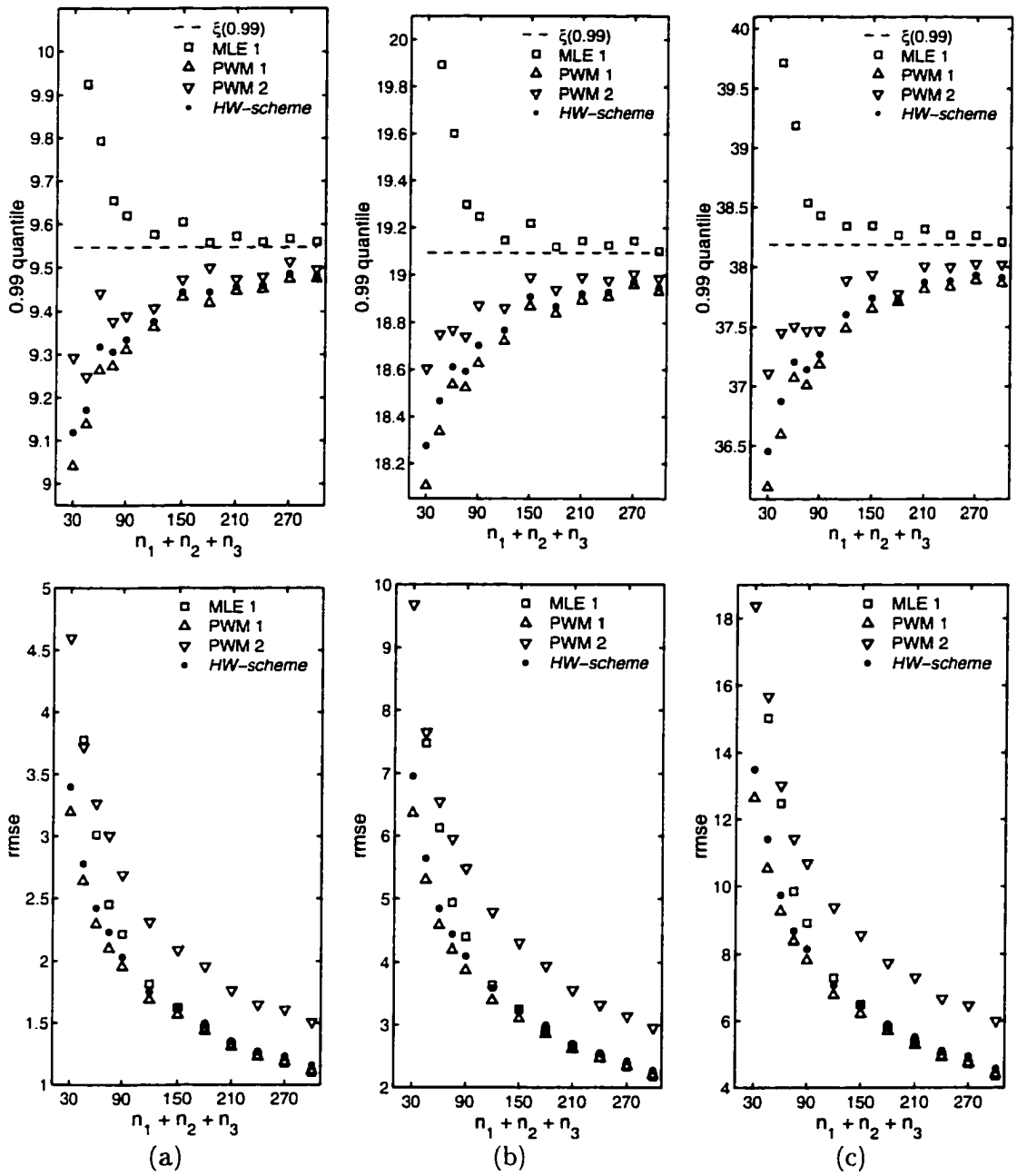

Figure 1.4: Comparison of the suggested regional model and the *HW-scheme* for a region of three sites, where n_j is the sample size for site j and $n_1 = n_2 = n_3$. The regional L-moment-ratios are estimated as a weighted average of at-site L-moment-ratios (Eq (1.25)). For each $n_1 + n_2 + n_3$ the 0.99 quantile is estimated as an average of 10,000 runs. $\xi(0.99)$ is the 0.99 population quantile. (a) site 1, (b) site 2, and (c) site 3.

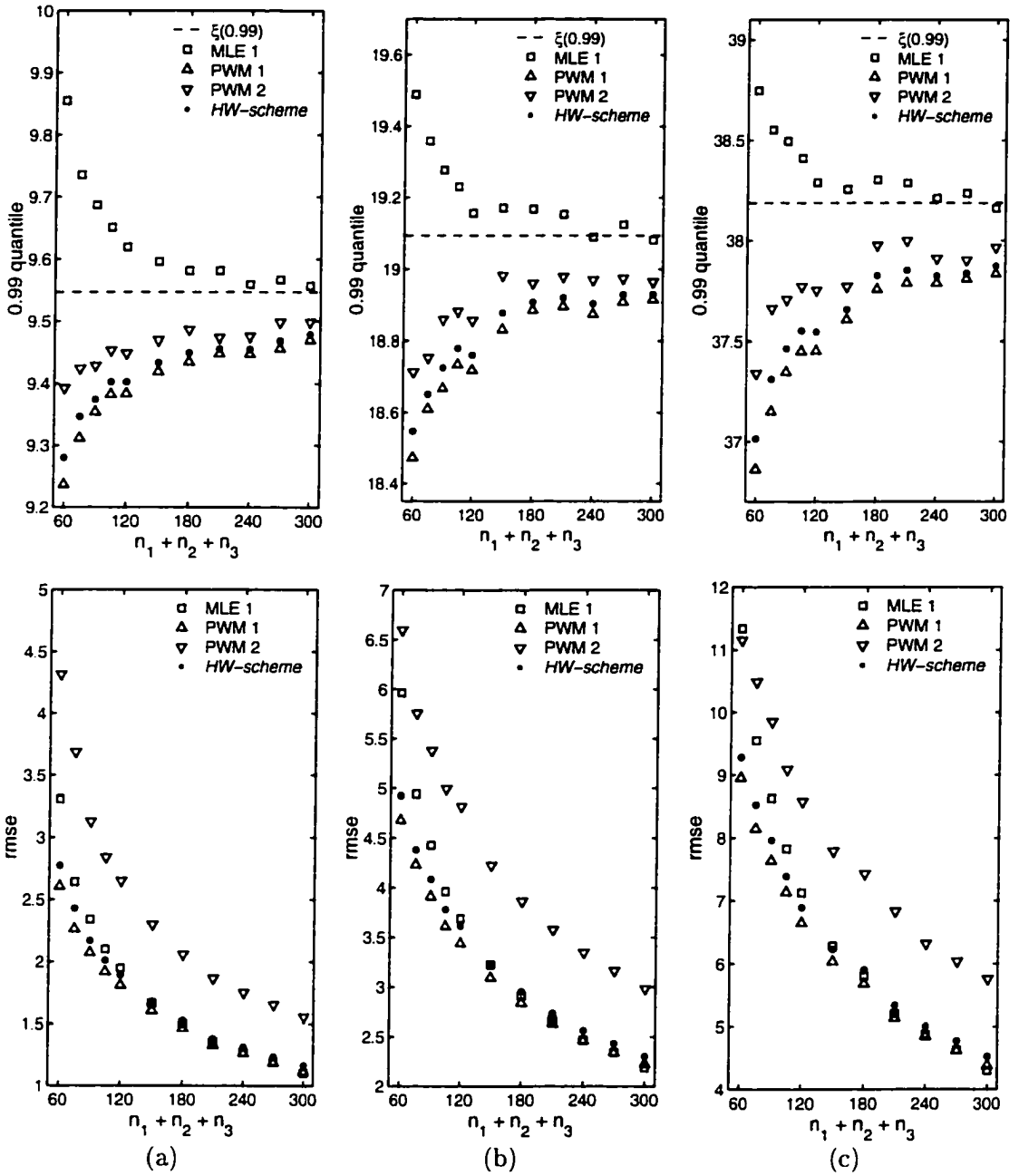

Figure 1.5: Comparison of the suggested regional model and the *HW-scheme* for a region of three sites, where n_j is the sample size for site j, $n_2 = n_1 + 10$, and $n_3 = n_1 + 20$. The regional L-moment-ratios are estimated as a weighted average of at-site L-moment-ratios (Eq (1.25)). For each $n_1 + n_2 + n_3$ the 0.99 quantile is estimated as an average of 10,000 runs. $\xi(0.99)$ is the 0.99 population quantile. (a) site 1, (b) site 2, and (c) site 3.

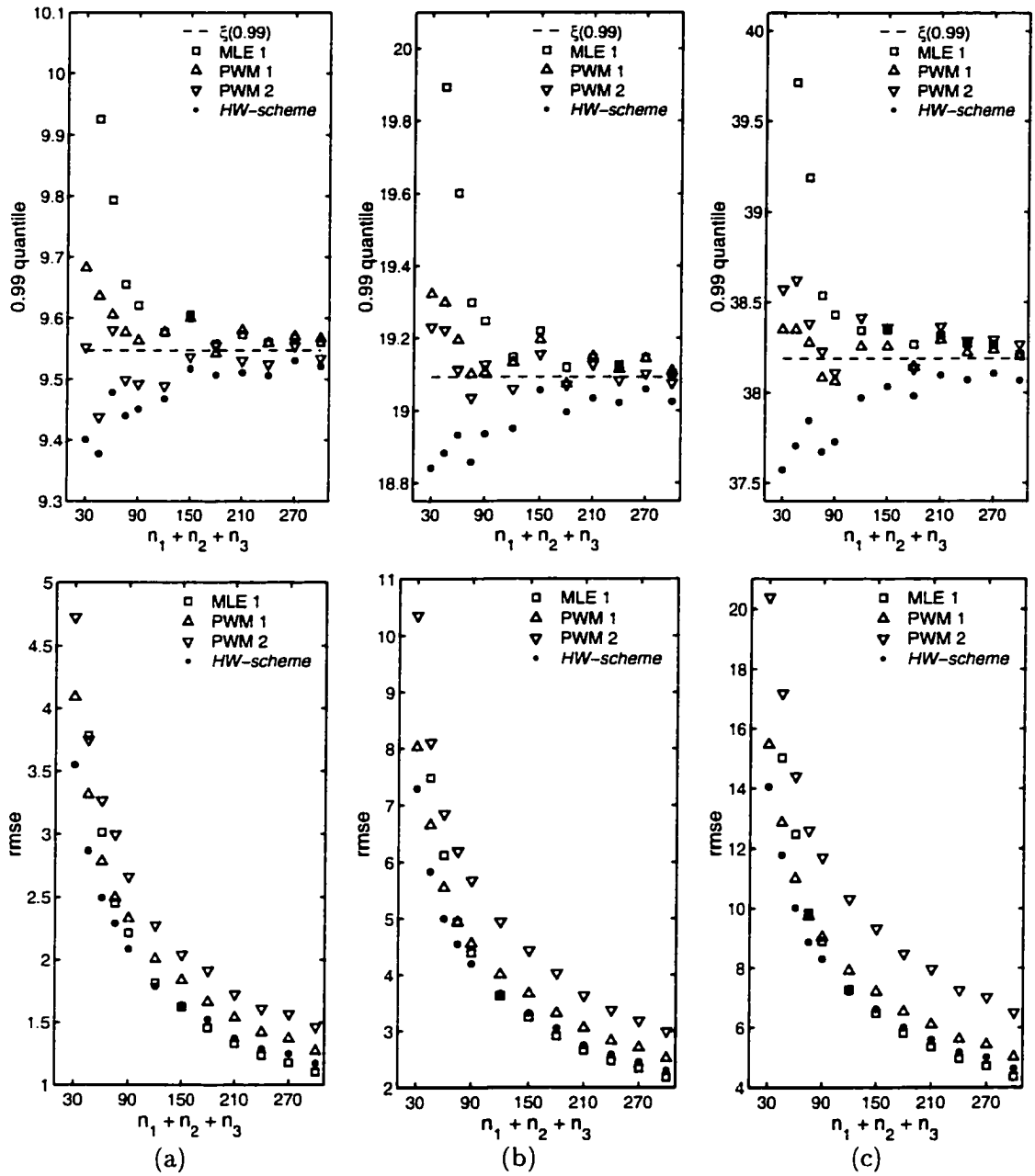

Figure 1.6: Comparison of the suggested regional model and the *HW-scheme* for a region of three sites, where n_j is the sample size for site j and $n_1 = n_2 = n_3$. The regional L–moment-ratios are estimated in terms of weighted averages of at-site L-moments (Eq (1.40)). For each $n_1 + n_2 + n_3$ the 0.99 quantile is estimated as an average of 10,000 runs. $\xi(0.99)$ is the 0.99 population quantile. (a) site 1, (b) site 2, and (c) site 3.

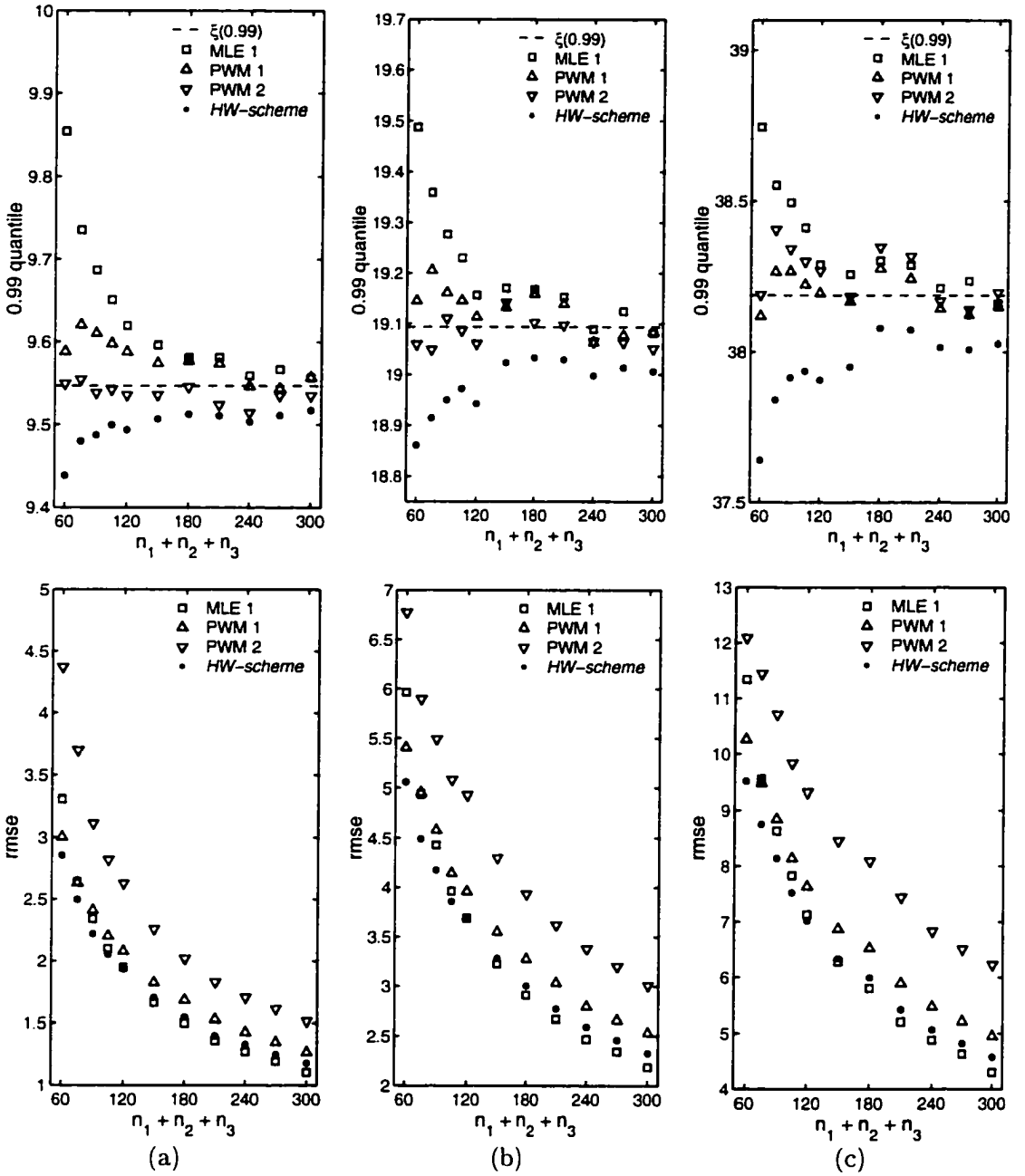

Figure 1.7: Comparison of the suggested regional model and the *HW-scheme* for a region of three sites, where n_j is the sample size for site j, $n_2 = n_1 + 10$, and $n_3 = n_1 + 20$. The regional L–moment-ratios are estimated in terms of weighted averages of at-site L–moments (Eq (1.40)). For each $n_1 + n_2 + n_3$ the 0.99 quantile is estimated as an average of 10,000 runs. $\xi(0.99)$ is the 0.99 population quantile. (a) site 1, (b) site 2, and (c) site 3.

Chapter 2

THE POPULATION INDEX FLOOD METHOD: ESTIMATION OF VARIANCE OF QUANTILE ESTIMATORS; AND COMPARISON WITH THE TRADITIONAL INDEX FLOOD METHOD

Abstract. The population index flood (PIF) method is a new analytical model for regional frequency analysis. In this paper explicit equations based on Fisher's information are derived for estimation of the standard error of at-site quantile estimators for the regional population index flood method utilizing the generalized extreme value distribution with maximum likelihood estimation. Simulation experiments for different sized regions and different values of the shape parameter show, that, the suggested methods for estimating the standard error of at-site quantile estimators result in values close to the actual or true values. In addition, similar simulation experiments are also used to test the accuracy of newly suggested procedures for estimating the standard errors of at-site quantile estimators for the Hosking and Wallis regional index flood method utilizing the generalized extreme value distribution. The results of the simulations indicate that these estimated standard errors can in some cases be very unreliable. In general this study shows that the new PIF models are a useful addition to existing regional frequency analysis models, and that their analytic structure, which is not present in other regional models, have important theoretical and practical implications.

2.1 Introduction

The population index flood (PIF) method has recently been suggested as an alternative to traditional index flood procedures for regional frequency analyzes of extreme hy-

drologic events (refer to Chapter 1). In the PIF method the index flood (or the indexing function) at each site is taken to be a function of the unknown at-site population quantities and, as a result, the homogeneity of the region is embedded in the structure of the parameter space of the underlying distribution model. More precisely, depending on the regional distribution model and the type of the indexing function, some of the distribution parameters are site-specific, while other parameters are common for all sites within the statistically homogeneous region. Because of the analytical framework of the PIF regional method, the method of maximum likelihood can be used for parameter estimation and in addition when regularity conditions are satisfied the variance-covariance matrix of the maximum likelihood estimators can be used to estimate the standard error of estimated quantiles.

Asymptotic and sample variances of quantile estimators are estimated for the PIF method based on the generalized extreme value (GEV) distribution with maximum likelihood estimation. This is done using a formula for the Cramer Rao lower bound (CRLB) of variance of unbiased estimators and the estimated asymptotic and observed variance-covariance information matrix of the maximum likelihood GEV parameter estimators. The estimated asymptotic and sample variances of the quantile estimators are compared using simulation experiments for different sized regions and two types of indexing functions: (1) sample data at each site are indexed by dividing them by the at-site population mean; and (2) sample data at each site are indexed by standardizing them using at-site population statistics. In addition, the above simulation experiments are repeated to test the accuracy of methods and procedures suggested by De Michele and Rosso (2001) for estimation of the standard error of at-site quantile estimators in the well-known Hosking and Wallis regional estimation scheme (Hosking and Wallis, 1997), where the GEV is the underlying regional distribution. Lastly, the proposed PIF regional method is compared with the Hosking and Wallis regional estimation scheme (*HW* scheme) using extreme precipitation data from northeastern Colorado.

2.2 Notation and Formulation of Algorithms

To get the reader familiar with the notation used herein the generalized extreme value (GEV) distribution for a single site is used with parameters estimated based on maximum likelihood. Furthermore, methods to estimate the uncertainty of quantile estimators are introduced.

2.2.1 The GEV Distribution and Estimation of Parameters by Maximum Likelihood

A random variable X is GEV distributed with parameters α (location), β (scale), and κ (shape) if its probability density function (pdf) is given by

$$f_X(x) = \frac{1}{\beta}\left[1 - \frac{\kappa}{\beta}(x-\alpha)\right]^{\frac{1}{\kappa}-1} \exp\left\{-\left[1 - \frac{\kappa}{\beta}(x-\alpha)\right]^{1/\kappa}\right\} \tag{2.1}$$

where $-\infty < \alpha < \infty$, $\beta > 0$, $-\infty < \kappa < \infty$, and x is the value of X. The range of X is: $\alpha + \beta/\kappa < x < \infty$ for $\kappa < 0$, and $-\infty < x < \alpha + \beta/\kappa$ for $\kappa > 0$. The mean and the variance are $E[X] = \alpha + \beta[1 - \Gamma(1+\kappa)]/\kappa$ for $\kappa > -1$, and $Var(X) = \beta^2[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]/\kappa^2$ for $\kappa > -1/2$. For $\kappa = 0$ the GEV becomes the Gumbel distribution.

The log-likelihood function of a random sample $X_1, \ldots, X_n$ of size n from the GEV distribution is given by

$$\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \ln f_X(x_i) \tag{2.2}$$

where $\mathbf{x} = [x_1, \ldots, x_n]$ is the observed sample vector and $\boldsymbol{\theta} = [\alpha, \beta, \kappa] \in \Theta$ is the parameter vector of $f_X(x)$, and Θ is the corresponding parameter space as specified above. The maximum likelihood estimates (ML-estimates) are found by taking the gradient of the log-likelihood and setting it equal to zero and solving for the parameters

$$\nabla \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = 0 \tag{2.3}$$

where $\nabla = [D_1, D_2, D_3]$ and D_k is the partial derivative with respect to the *kth* element in $\boldsymbol{\theta}$, for example $D_2 = \frac{\partial}{\partial \beta}$. There is no explicit solution for $\boldsymbol{\theta}$ in Eq (2.3), thus an optimization

algorithm such as the Newton-Rhapson iteration can be used to obtain a numerical solution for the ML-estimates $\hat{\boldsymbol{\theta}} = [\hat{\alpha}, \hat{\beta}, \hat{\kappa}]$. In this paper the Newton-Rhapson iteration is utilized. The algorithm evolve on

$$\hat{\boldsymbol{\theta}}_{i+1}^T = \hat{\boldsymbol{\theta}}_i^T + [SI(\mathbf{x}; \hat{\boldsymbol{\theta}}_i)]^{-1} [\nabla \ln \mathcal{L}(\hat{\boldsymbol{\theta}}_i; \mathbf{x})]^T \tag{2.4}$$

where the superscript T denotes transpose, and $SI(\mathbf{x}; \hat{\boldsymbol{\theta}})$ is Fisher's sample information matrix equal to the negative Hessian matrix of the log-likelihood

$$SI(\mathbf{x}; \hat{\boldsymbol{\theta}}) = - \begin{bmatrix} D_{11} \ln \mathcal{L} & D_{21} \ln \mathcal{L} & D_{31} \ln \mathcal{L} \\ D_{21} \ln \mathcal{L} & D_{22} \ln \mathcal{L} & D_{32} \ln \mathcal{L} \\ D_{31} \ln \mathcal{L} & D_{32} \ln \mathcal{L} & D_{33} \ln \mathcal{L} \end{bmatrix} \tag{2.5}$$

with $D_{ij} = D_i D_j$. Furthermore, $[SI(\mathbf{x}; \hat{\boldsymbol{\theta}})]^{-1}$ is the sample variance-covariance matrix of $\hat{\boldsymbol{\theta}}$. Throughout this paper the iterative procedure is repeated until the relative change in all parameters is less than 0.01%, that is $\max |(\hat{\boldsymbol{\theta}}_{i+1} - \hat{\boldsymbol{\theta}}_i)/\hat{\boldsymbol{\theta}}_{i+1}| < 0.0001$.

2.2.2 Uncertainty of GEV Quantile Estimators

The *qth* quantile of the GEV pdf in Eq (2.1) is

$$\xi(q) = \alpha + \frac{\beta}{\kappa}[1 - (-\ln q)^{\kappa}] \tag{2.6}$$

Under regularity conditions the Cramer-Rao lower bound (CRLB) is a lower bound for the variance of unbiased estimators, and the CRLB is also the asymptotic variance (AVar) of maximum likelihood estimators. The CRLB for the variance of unbiased estimators of $\xi(q)$ is

$$\text{CRLB}(\xi(q)) = \nabla \xi(q) \, [EI(\boldsymbol{\theta})]^{-1} \, [\nabla \xi(q)]^T \tag{2.7}$$

where $EI(\boldsymbol{\theta})$ is Fisher's expected information matrix given by

$$EI(\boldsymbol{\theta}) = E[SI(\mathbf{X}; \boldsymbol{\theta})] \tag{2.8}$$

Furthermore $[\boldsymbol{EI}(\boldsymbol{\theta})]^{-1}$ is the asymptotic variance-covariance matrix of $\boldsymbol{\theta}$. In case of the GEV distribution the regularity conditions are satisfied if the diagonal elements of $\boldsymbol{EI}(\boldsymbol{\theta})$ exist and are positive, which is the case when $\kappa < 1/2$.

Given the population parameters $\boldsymbol{\theta}$, the theoretical CRLB for unbiased quantile estimators is given by Eq (2.7). In most practical cases the population parameters $\boldsymbol{\theta}$ are unknown. In that case Eq (2.7), evaluated at $\xi(q) = \hat{\xi}(q;\hat{\boldsymbol{\theta}})$ and $\boldsymbol{EI}(\hat{\boldsymbol{\theta}})$, gives the AVar of the ML-estimator $\hat{\xi}(q)$. Another alternative is to use the sample information matrix $\boldsymbol{SI}(\mathbf{x};\hat{\boldsymbol{\theta}})$ instead of $\boldsymbol{EI}(\hat{\boldsymbol{\theta}})$, in which case Eq (2.7) would give a type of a sample "asymptotic" variance, here dubbed as SVar, of the ML-estimator $\hat{\xi}(q)$. Prescott and Walden (1980, 1983) compared the simulated variance of the shape parameter κ, $\mathrm{Var}(\hat{\kappa})$, with the $\mathrm{CRLB}(\kappa)$, $\mathrm{AVar}(\hat{\kappa})$, and $\mathrm{SVar}(\hat{\kappa})$ for various values of κ and a sample size $n = 100$. Their results showed that $\mathrm{SVar}(\hat{\kappa})$ was the closest to $\mathrm{Var}(\hat{\kappa})$, and that $\mathrm{CRLB}(\kappa)$ and $\mathrm{AVar}(\hat{\kappa})$ tended to underestimate $\mathrm{Var}(\hat{\kappa})$.

2.3 The Population Index Flood Method

The population index flood (PIF) method is an analytical model for regional frequency analysis. A detailed description of the PIF method, and PIF models for many commonly used two- and three-parameter distributions are derived in Chapter 1. Instead of using a sample property as an index, as is commonly done in the traditional index flood approach, in the PIF method the index flood is taken to be a function of the unknown population statistics at each site, and the homogeneity of a region is embedded in the structure of the parameter space of the underlying distribution model. In Chapter 1 two types of indexing functions are considered

$$\frac{\xi_j(q)}{\mu_j} \quad , j = 1, \ldots, m \tag{2.9}$$

and

$$\frac{\xi_j(q) - \mu_j}{\sigma_j} \quad , j = 1, \ldots, m \tag{2.10}$$

where m is the number of independent sites in the region, $\xi_j(q)$ is the qth population quantile at site j, μ_j is the population mean at site j, and σ_j is the population standard deviation at site j. We will refer to indexing based on Eq (2.9) as PIF 1, and indexing based on Eq (2.10) as PIF 2. A statistically homogeneous region is defined as a region where either Eq (2.9) or Eq (2.10) (depending on the method) does not depend on j. That is, for both the PIF 1 and the PIF 2 methods a statistically homogeneous region implies that the skewness, kurtosis, and all higher order moment ratios are the same for all sites in the region. In addition, for the PIF 1 method the coefficient of variation (σ_j/μ_j) is also the same for all sites within the region. Hence, PIF 1 is more restrictive than PIF 2, where PIF 2 can always be used instead of PIF 1 but not vice versa. For a statistically homogeneous region of m sites we will denote the sample at site j by $X_{j1},\ldots,X_{jn_j}$ for $j = 1,\ldots,m$, where n_j is the sample size at site j.

2.3.1 PIF by Indexing by the Population Mean : PIF 1

In Chapter 1 for a statistically homogeneous region of m sites it is demonstrated that Eq (2.9) independent of j implies for the GEV distribution in Eq (2.1) that the ratio $\alpha_j/\beta_j = \gamma$ is the same for all sites, and that the shape parameter κ is the same for all sites. Thus the parameter space has dimension $(m+2)$, where either α_j or β_j is estimated for each site j in the region, and γ and κ are estimated commonly for all sites in the region. By the invariance property of ML-estimators the two cases above are equivalent. For the case when the location parameter α_j is estimated at each site j in the region, then for $\theta_j = \alpha_j^{-1}$ the log-likelihood is

$$\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}_1,\ldots,\mathbf{x}_m) = \sum_{j=1}^{m}\left\{ n_j\left(\ln\theta_j + \ln\gamma\right) + \sum_{i=1}^{n_j}\left[\left(\frac{1}{\kappa}-1\right)\ln\zeta_{ji} - \zeta_{ji}^{1/\kappa}\right]\right\} \qquad (2.11)$$

where $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_m,\gamma,\kappa]$, $\mathbf{x}_j = [x_{j1},\ldots,x_{jn_j}]$ is the sample vector for site j, and $\zeta_{ji} = 1-\gamma\kappa(\theta_j x_{ji}-1)$. The first and the second partial derivatives of the log-likelihood in Eq (2.11), needed for the Newton-Rhapson procedure in Eq (2.4), along with the first partial

derivatives of the *qth* GEV quantile in Eq (2.6), needed for the evaluation of the CRLB in Eq (2.7), are given in appendix A.1.

The elements of Fisher's expected information matrix, $EI(\boldsymbol{\theta})$, are found by taking the expected value of the sample information matrix, whose elements are given in Eq (A.2) in appendix A.1. When $\kappa < 1/2$ regularity conditions are satisfied and the non-zero elements of $EI(\boldsymbol{\theta})$ are given by

$$
\begin{aligned}
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_j^2}\right] &= \frac{n_j}{\theta_j^2 \kappa^2}\{1+\eta\delta[\eta\delta\Gamma(\nu)-2\Gamma(\eta)]\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \gamma \partial \theta_j}\right] &= \frac{n_j}{\theta_j \gamma \kappa^2}\{1+\eta[\eta\delta\Gamma(\nu)-(\delta+1)\Gamma(\eta)]\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa \partial \theta_j}\right] &= \frac{n_j}{\theta_j \kappa^3}\{\kappa[\eta\delta\Gamma'(\eta)-\Gamma'(1)]+\eta[\eta\delta\Gamma(\nu) \\
&\quad -(\eta\delta+1)\Gamma(\eta)+1]\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \gamma^2}\right] &= \frac{n_T}{\gamma^2 \kappa^2}\{1+\eta[\eta\Gamma(\nu)-2\Gamma(\eta)]\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa \partial \gamma}\right] &= \frac{n_T}{\gamma \kappa^3}\{\kappa[\eta\Gamma'(\eta)-\Gamma'(1)]+\eta[\eta\Gamma(\nu) \\
&\quad -(\eta+1)\Gamma(\eta)+1]\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa^2}\right] &= \frac{n_T}{\kappa^4}\{1+2\kappa[\eta\Gamma'(\eta)-\Gamma'(1)-1] \\
&\quad +\kappa^2[\pi^2/6+(1+\Gamma'(1))^2]+\eta^2[\Gamma(\nu)-2\Gamma(\eta)]\}
\end{aligned}
\tag{2.12}
$$

where $\eta = (1-\kappa)$, $\nu = (1-2\kappa)$, $\delta = (\gamma\kappa+1)$, and $n_T = n_1 + \cdots + n_m$. In addition $\Gamma'(y)$ is the first derivative of the gamma function with argument y, and $\Gamma'(1)$ is the negative Euler's constant.

2.3.2 PIF by Standardizing Using Population Statistics : PIF 2

In Chapter 1 for a statistically homogeneous region of m sites it is demonstrated that Eq (2.10) independent of j implies for the GEV distribution in Eq (2.1) that the shape parameter κ is the same for all sites. Thus the parameter space has dimension $(2m+1)$, where α_j and β_j are estimated for each site j in the region, and κ is estimated commonly

for all sites in the region. For $\theta_j = \beta_j^{-1}$ the log-likelihood is given by

$$\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}_1, \ldots, \mathbf{x}_m) = \sum_{j=1}^{m} \left\{ n_j \ln \theta_j + \sum_{i=1}^{n_j} \left[\left(\frac{1}{\kappa} - 1 \right) \ln \zeta_{ji} - \zeta_{ji}^{1/\kappa} \right] \right\} \tag{2.13}$$

where $\boldsymbol{\theta} = [\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_m, \kappa]$, and $\zeta_{ji} = 1 - \theta_j \kappa(x_{ji} - \alpha_j)$. The first and the second partial derivatives of the log-likelihood in Eq (2.11) along with the first partial derivatives of the *qth* GEV quantile in Eq (2.6) are given in appendix A.2.

For $\kappa < 1/2$ the non-zero elements of Fisher's expected information matrix, EI, of $\boldsymbol{\theta}$ are given by

$$\begin{aligned}
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_j^2}\right] &= n_j \theta_j^2 \eta^2 \Gamma(\nu) \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_j \partial \alpha_j}\right] &= \frac{n_j \eta}{\kappa} \{\Gamma(\eta) - \eta\Gamma(\nu)\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa \partial \alpha_j}\right] &= \frac{n_j \theta_j \eta}{\kappa^2} \{\eta[\Gamma(\eta) - \Gamma(\nu)] - \kappa\Gamma'(\eta)\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_j^2}\right] &= \frac{n_j}{\theta_j^2 \kappa^2} \{1 + \eta[\eta\Gamma(\nu) - 2\Gamma(\eta)]\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa \partial \theta_j}\right] &= \frac{n_j}{\theta_j \kappa^3} \{1 + \kappa[\eta\Gamma'(\eta) - \Gamma'(1) - 1] \\
&\quad + \eta[\eta\Gamma(\nu) - (\eta + 1)\Gamma(\eta)]\} \\
E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa^2}\right] &= \frac{n_T}{\kappa^4} \{1 + 2\kappa[\eta\Gamma'(\eta) - \Gamma'(1) - 1] \\
&\quad + \kappa^2[\pi^2/6 + (1 + \Gamma'(1))^2] + \eta^2[\Gamma(\nu) - 2\Gamma(\eta)]\}
\end{aligned} \tag{2.14}$$

where as before $\eta = (1 - \kappa)$, $\nu = (1 - 2\kappa)$, and $n_T = n_1 + \cdots + n_m$.

2.4 Theoretical Results

A measure of efficiency of single site analysis relative to regional analysis for estimation of the *qth* quantile at site j is given by the ratio

$$\text{Reff} = \frac{\text{CRLB}(\xi_j(q) \text{ regional})}{\text{CRLB}(\xi_j(q) \text{ single site})} \tag{2.15}$$

Reff depends on the ratio $m = n_T/n_j$, that is the number of sites within the region, for both the PIF 1 and the PIF 2 models. Furthermore for the PIF 1 method, Reff depends on κ and

the ratio $\gamma = \alpha_j/\beta_j$, but it does not depend on the individual values of α_j and β_j. For the PIF 2 method, Reff depends only on κ. For the case when n_j is common for all sites within the region, then for both the PIF 1 and PIF 2 methods Reff does not depend on j and we refer to

$$\text{Reff}(m_1 : m_2) \quad , m_2 \leq m_1$$

as the efficiency of regional analysis of m_2 sites relative to regional analysis of m_1 sites for estimation of the *qth* quantile at any site j in the region.

Typical ranges for hydrologic data are $\gamma \in (1, 4)$ and $\kappa \in (-0.3, 0.1)$. In Fig. 2.1 the relationship of Reff(6 : 1) versus q is plotted for typical values of γ and κ for both the PIF 1 and the PIF 2 models. The dependence of the PIF 1 model on γ and the independence of the PIF 2 model on γ are clearly seen in the figure. Furthermore, from Fig. 2.1 the advantage of using regional analysis over single site analysis for estimation of lower and upper extreme quantiles are clear. In addition, the gain of using regional analysis based on the PIF 1 model is greater than that based on the PIF 2 model. Recall though that the PIF 1 model may not be suitable in some situations where the PIF 2 model can be used, while the PIF 2 models is applicable in all situation where the PIF 1 model can be used. An interesting property for the PIF 2 method may be observed in the Fig. 2.1, where for every κ there are two quantiles where there is no gain of using regional analysis over single site analysis. Boes et al. (1989) showed that for the Weibull model there is no asymptotic gain of using regional analysis over single site analysis for the 0.783 quantile. Figure 2.2 shows the relationship of Reff(m : 1) for various values of m, $\kappa = -0.1$, and $\gamma = 2$ for both the PIF 1 and the PIF 2 models. Focusing on the upper tail, notice how close the curves based on 6, 12, and 36 sites are, suggesting that in practice a bigger region may not necessarily be better than a smaller region. In addition, in the real world heterogeneity usually increases with the size of the region.

2.5 Simulation Results

In Chapter 1 a statistical homogeneous region of three sites was simulated and the bias and the mean-squared error (MSE) of GEV quantile estimators was investigated using the PIF 1 GEV model with parameters estimated using maximum likelihood and probability weighted moments. The region was simulated for various sample sizes, $\gamma = 2$, and $\kappa = -0.1$. These values of γ and κ appear typical for extreme annual precipitation of short duration in Colorado. Since regions suitable for the PIF 1 model are also suitable for the PIF 2 model, we will use the same region as in Chapter 1, except that in our case the shape parameter κ will vary. Thus the population parameters are $[\alpha_1, \beta_1, \kappa_1] = [2, 1, \kappa]$ for site 1, and $[\alpha_2, \beta_2, \kappa_2] = [4, 2, \kappa]$, and $[\alpha_1, \beta_1, \kappa_3] = [8, 4, \kappa]$ for sites 2 and 3, respectively. That is in distribution $X_2 \stackrel{d}{=} 2X_1$ and $X_3 \stackrel{d}{=} 4X_1$ where X_j is the random variable at site j. To cover the practical range of κ for hydrologic data, our simulation experiments will be made for $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$. The probability for any of the three sites to have negative flows is 6.2×10^{-10}, 2.6×10^{-6}, 9.0×10^{-5}, and .0020, for $\kappa = -0.3$, $\kappa = -0.2$, $\kappa = -0.1$, and $\kappa = 0.1$, respectively. Thus these parameters can be considered realistic for hydrologic data, except perhaps for the case when $\kappa = 0.1$ where there is one in a 500 chance of getting negative values. For these parameter sets the result of our simulation are the same or similar for individual sites, thus we will in most cases unless otherwise indicated only show results based on estimation of quantiles at site 1. In the simulation the $\mathrm{MSE}(\hat{\xi}(q))$ is compared with $\mathrm{CRLB}(\xi(q))$, $\mathrm{AVar}(\hat{\xi}(q))$, and $\mathrm{SVar}(\hat{\xi}(q))$, where the MSE is defined as

$$\mathrm{MSE}(\hat{\xi}(q)) = \mathrm{Var}(\hat{\xi}(q)) + (E[\hat{\xi}(q)] - \xi(q))^2 \tag{2.16}$$

where $E[\hat{\xi}(q)] - \xi(q)$ is the bias of $\hat{\xi}(q)$. Furthermore, the relative bias of quantile estimators is defined by

$$\mathrm{RBIAS} = \frac{E[\hat{\xi}(q)] - \xi(q)}{\xi(q)} \tag{2.17}$$

where RBIAS > 0 indicates overestimation, while RBIAS < 0 indicates underestimation. Simulation results based on the PIF 1 model and $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$ are shown in

Figs. 2.3, 2.4, and 2.5 for the 0.95, 0.99, and 0.998 quantiles respectively. Simulation results based on the PIF 2 model for the 0.95, 0.99, 0.998 quantiles are shown in Figs. 2.6, 2.7, and 2.8 for $\kappa = 0.1$, $\kappa = -0.1$, and $\kappa = -0.2$, respectively. All simulation results in Figs. 2.3–2.8 are based on 10,000 realization for each case shown. Furthermore, note that in each of the figures the same scale is used for each κ, displaying the relative gain resulting from increasing the size of the region, where for $m = 6$ two sets of random samples from each site are used and for $m = 12$ four sets of random samples from each site are used. Overall, the results for both the AVar and the SVar are similar, where both the AVar and the SVar represent the MSE well. That is, either one can be used for estimating the variance of quantile estimators. Going into more detail, then for $m = 3$, SVar appears closest to MSE for $\kappa = 0.1$, and AVar appears closest to MSE for $\kappa \in \{-0.3, -0.2, -0.1\}$. For $m = 6$ and $m = 12$ SVar appears closest to MSE for all κ's, where for say $n_j > 40$ SVar and AVar are almost identical. When comparing the results for the PIF 1 and the PIF 2 models, then for this scenario (underlying model is PIF 1) it should be expected that for the same n_j the AVar, SVar, and MSE are a little higher for the PIF 2 model than the PIF 1 model, since more parameters need to be estimated in the PIF 2 model. This is reflected in Figs. 2.3–2.8, where n_j starts at 25 for the PIF 1 model and at 30 for the PIF 2 model.

2.6 Simulation Results : Variance of Quantile Estimators in the Hosking and Wallis Regional Estimation Scheme

The Hosking and Wallis regional estimation scheme (Hosking and Wallis, 1997), dubbed here as the *HW* scheme, is probably the most widely used regional scheme at present and is used in this paper as a comparison to the PIF methods. In recent studies (De Michele and Rosso, 2001), approximate equations for estimating the variance of at-site quantile estimators in the *HW* scheme have been suggested, where the GEV is assumed as the underlying regional distribution. In this section, the accuracy of these procedures is tested using similar simulation experiment as in section 2.5.

In the *HW* scheme the index flood at site j is estimated by the at-site sample mean, $\hat{\mu}_j = \overline{X}_j$, so that the at-site sample observations are indexed by dividing them by the at-site sample mean, and a single regional growth curve is estimated for the whole region. Once the regional growth curve has been estimated, the *qth* quantile at site j, $\hat{\xi}_j(q)$, is simply

$$\hat{\xi}_j(q) = \overline{X}_j \cdot \hat{\xi}_R(q) \qquad , j = 1, \ldots, m \tag{2.18}$$

where $\hat{\xi}_R(q)$ is the regional *qth* quantile. The practice of estimation the index flood by sample statistics has been questioned in Chapter 1 and Sveinsson et al. (2002a), where certain analytical weaknesses of the *HW* scheme are pointed out in Chapter 1.

The parameters of the regional growth curve in the *HW* scheme are estimated in terms of L-moments (refer to Hosking and Wallis (1997) or Chapter 1). Regional L-moments (λ_r^R) are estimated as

$$\hat{\lambda}_r^R = \frac{1}{n_T} \sum_{j=1}^{m} n_j \hat{\lambda}_r^{(j)} \qquad , r = 1, 2, \ldots \tag{2.19}$$

where $\hat{\lambda}_r^{(j)}$ is the *rth* sample L-moment at site j and $n_T = \sum_{j=1}^{m} n_j$. The regional L-moment-ratios (τ_r^R) are estimated from

$$\hat{\tau}_2^R = \frac{\hat{\lambda}_2^R}{\hat{\lambda}_1^R} \quad , \quad \hat{\tau}_r^R = \frac{\hat{\lambda}_r^R}{\hat{\lambda}_2^R} \qquad , r = 3, 4, \ldots \tag{2.20}$$

In Chapter 1 for the GEV distribution, the bias of the 0.95, 0.99, and 0.995 quantile estimators were shown to be significantly reduced when the regional L-moment-ratios were estimated from Eq (2.20) as opposed to being estimated in terms of weighted averages of at-site L-moment-ratios as in Hosking and Wallis (1997).

De Michele and Rosso (2001) recommend estimating the variance of at-site quantile estimators in the *HW* scheme, with the GEV as the underlying regional distribution, from

$$\operatorname{Var} \hat{\xi}_j(q) = \overline{X}_j^2 \operatorname{Var} \hat{\xi}_R(q) + \hat{\xi}_R^2(q) \operatorname{Var} \overline{X}_j + \operatorname{Var} \hat{\xi}_R(q) \operatorname{Var} \overline{X}_j \tag{2.21}$$

where

$$\operatorname{Var} \hat{\xi}_R(q) = \frac{\beta^2}{n_T} \exp\{-\ln(-\ln q) \exp[-1.823\kappa - 0.165]\} \tag{2.22}$$

is a fitted formula to tabulated values based on simulations in Lu and Stedinger (1992b). Note that Eq (2.21) assumes that $\overline{X}_j$ and $\hat{\xi}_R(q)$ are independent. In this paper estimated variances based on Eq (2.21) will be dubbed as aVar.

The simulation experiment in section 2.5 is repeated here to test the accuracy of the methods in Eqs. (2.21) and (2.22) for estimation of the approximate variance, aVar, of at-site quantile estimators in the *HW* scheme. The results of the simulation experiments are shown in Figs. 2.9, 2.10, and 2.11 for the 0.95, 0.99, and 0.998 quantiles respectively. As for the PIF methods, the MSE should be considered as the true or actual error variance of the quantile estimators in the *HW* scheme. For $\kappa = 0.1$ and $\kappa = -0.1$, the aVar represents the MSE fairly well, except perhaps for $m = 3$, where the aVar tends to overestimate the MSE for both the 0.95 and the 0.998 quantiles. For $\kappa = -0.2$ and $\kappa = -0.3$ the aVar represents the MSE fairly well for all investigated quantiles when $m = 12$, but for $m = 3$ and $m = 6$ the aVar is in some cases unrealistically high, especially for the 0.998 quantile, but also for the 0.99 quantile with $m = 3$. Thus as a general recommendation Eqs. (2.21) and (2.22) should be used cautiously for estimation of the approximate variance of quantile estimators in the *HW* scheme, especially for smaller regions with large negative κ.

The PIF 1 method and the *HW* scheme only differ in the treatment of the index flood, where in the PIF 1 method the index flood at each site is estimated by the at-site population mean, while in the *HW* scheme the index flood at each site is estimated by the at-site sample mean. A comparison of the simulation results of the PIF 1 method and the *HW* scheme indicates, that overall the PIF 1 performs better than the *HW* scheme with respect to the MSE, except for $m = 3$ and smaller sample sizes where the *HW* scheme performs better. In terms of bias, the results of the PIF 1 and the *HW* scheme are similar for the 0.95 quantile. But for the 0.99 and 0.998 quantiles the *HW* scheme is less biased for smaller sample sizes and $m = 3$, while for $m = 6$ and $m = 12$ the relative bias of the two methods appears similar.

2.7 Case Study

In our case study we will analyze annual maximum precipitation of 3-hr duration in the subregion SE from Sveinsson et al. (2002a). In Sveinsson et al. (2002a) this region passes Wilk's multivariate outlier test and the so-called X-10 regional homogeneity test from Lu and Stedinger (1992a). The region consist of 12 sites in the northeastern plains of Colorado with sample sizes ranging from 17 years to 50 years, and elevations ranging from 1133 *meters* to 1635 *meters*. The total sample size for the region is 465 years. The sites and some of their statistical characteristics are shown in Table 2.1. The PIF 1 and PIF 2 regional models are used for estimation of the growth curves at the sites within the region and for comparison the *HW* scheme is also used.

Empirical growth curves and estimated growth curves with approximate 95% confidence bounds for the 12 sites based on the *HW* scheme and the regional PIF 1 and the PIF 2 models are shown in Figs. 2.12 and 2.13, where plots are labeled by (a) for the *HW* scheme, (b) for the PIF 1 model, and (c) for the PIF 2 model. The shape parameter of the estimated growth curves is $\kappa = -0.1042$ for the *HW* scheme, $\kappa = -0.1497$ for the PIF 1 method, and $\kappa = -0.1611$ for the PIF 2 method. Often when different models are compared using empirical data, it can be hard to judge the performance of the different models, since the empirical data are never going to truly reflect the structure of any one model. Overall the results based on the *HW* scheme and the PIF 1 model appear similar, and perhaps it can be argued that overall the PIF 2 model gives the best fit to the data. Note also, that the poorest fit to the empirical data appears to be for sites 4 and 6, where all models appear to overestimate the empirical growth curve for site 4, and underestimate the empirical growth curve for site 6. This is an interesting observation, since the observed coefficient of variation and the observed skewness for site 4 are the lowest among all sites in the region (refer to Table 2.1), while for site 6 they are the highest. Recall that for the PIF 1 and the *HW* scheme, the coefficient of variation, skewness and all higher order moment ratios are

assumed to be the same for all sites within the homogeneous region, while for the PIF 2 the skewness and all higher order moment ratios are assumed to be the same for all sites within the region. Furthermore, only the PIF 2 model appears to fit site 12 well. The skewness for site 12 in Table 2.1 appears typical for the sites in the region, while the coefficient of variation is the second highest of all sites, with only site 6 being higher. This, might explain why the PIF 2 model appears to perform better than the other models for site 12.

2.8 Summary and Conclusions

Formulas for the asymptotic and sample variances of maximum likelihood quantile estimators at each site within a statistically homogeneous region were derived for the the regional PIF model and the three parameter GEV distribution with an assumed independence in space. The formulas were based on the the Cramer Rao lower bound (CRLB) for the variance of unbiased estimators and the observed and expected Fisher's information matrix of the maximum likelihood estimators of the parameters of the GEV distribution. The CRLB was used to calculate the theoretical gain of regionalization, where for extreme upper or lower quantiles there is always significant gain in using regional analysis over single site analysis. Furthermore, for the PIF 1 model there is always gain in regionalization, while for the PIF 2 model there are always two quantiles where there is no gain in regionalization.

Simulation experiments for different sized regions and different values of the GEV shape parameter were used to test the derived formulas for estimating the variance of maximum likelihood quantile estimators of the PIF methods. The formula based on Fisher's expected information matrix (AVar) and the formula based on the Fisher's observed information matrix (SVar) generally give similar results and represented the simulated MSE well. AVar appeared more accurate for a small region (3 sites) while SVar appeared more accurate for larger regions with small sample sizes at each site. For larger sample sizes, say $n_j > 50$, AVar and SVar were not significantly different. Similar simulation experiments were also used to test the accuracy of newly suggested procedures (De Michele and Rosso,

2001) for estimating the variance of at-site quantile estimators for the Hosking and Wallis regional estimation scheme (*HW* scheme) utilizing the generalized extreme value distribution. The results of the simulations indicate that these estimated variances are fairly accurate for $\kappa = 0.1$ and $\kappa = -0.1$, but they can in some cases be very unreliable, especially when $-0.3 < \kappa < -0.2$, and should be used cautiously.

A region of 12 sites in the northeastern plains of Colorado was chosen to compare the two PIF models, PIF 1 and PIF 2, with the Hosking and Wallis regional estimation scheme. The growth curves and confidence bounds based on the PIF 1 model and the *HW* scheme were similar, while overall the more flexible PIF 2 model seemed to give the best fit to the empirical growth curves. As a general conclusion, it appears that the new analytic PIF models can be quite useful addition to existing models for use in regional frequency analysis, and the possibility of using ML-estimation and the Fisher information for estimation of the standard error of quantile estimators is a nice property of the PIF models.

Table 2.1: Sample characteristics of the annual maximum 3-hr duration precipitation data for the sites in the case study.

Station *Name*	Site *nr.*	Sample size [*years*]	Mean [*cm*]	Coeff. of variation	Skewness coeff.
Bonny Dam 2 NE	1	28	3.510	0.459	2.496
Arapahoe	2	50	3.747	0.434	1.111
Paoli	3	27	3.575	0.402	0.989
Eckley	4	49	3.449	0.382	0.144
Joes 2 SE	5	46	3.383	0.457	1.204
Eads	6	34	3.342	0.632	2.510
Seibert	7	42	3.320	0.441	1.459
Akron 4 E	8	50	3.511	0.449	1.262
Hugo 1 NW	9	50	3.343	0.514	1.915
Ordway 21 N	10	17	2.869	0.402	0.461
Kutch 6 SSE	11	23	2.727	0.449	0.375
New Raymer	12	49	3.508	0.543	1.195

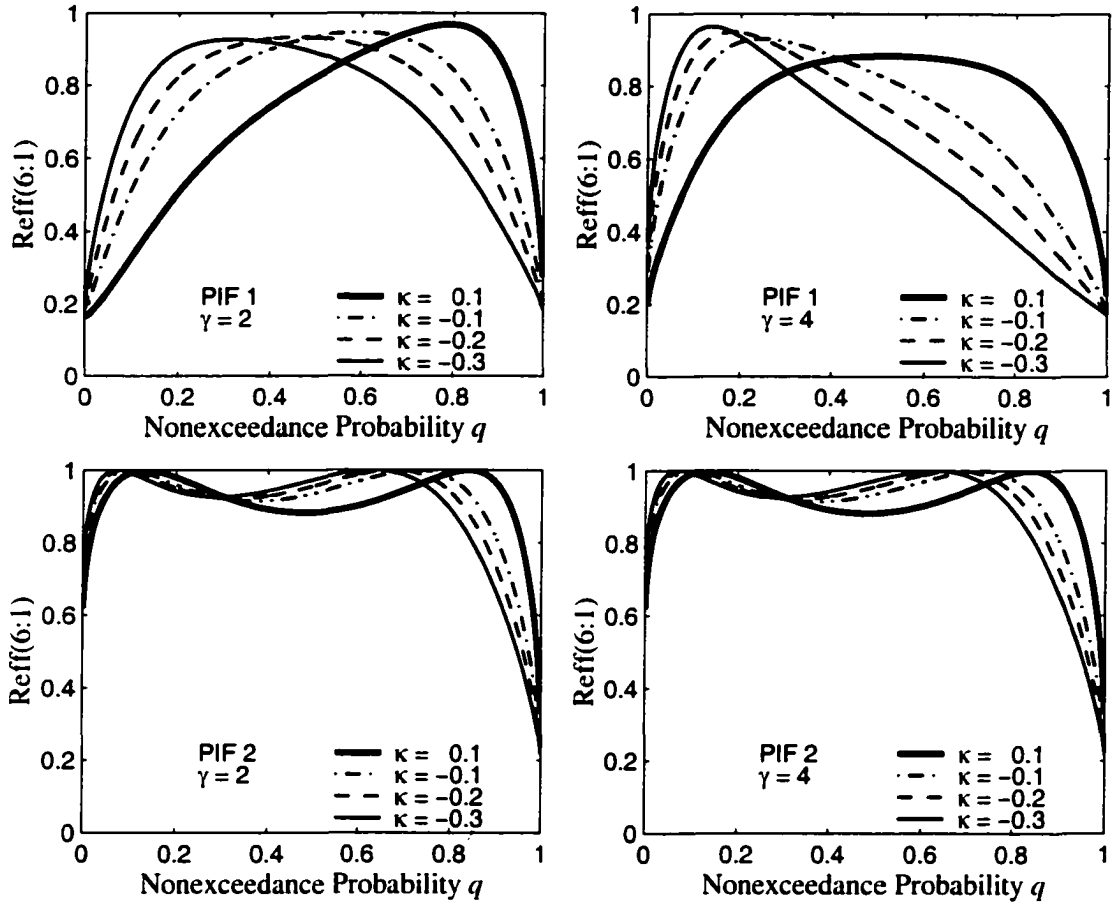

Figure 2.1: Asymptotic efficency of estimating the *qth* quantile at a single site using single site analysis relative to regional analysis of 6 sites, where the sample size n_j is the same at all sites and $\gamma = \alpha_j/\beta_j$ and κ are the same for all sites.

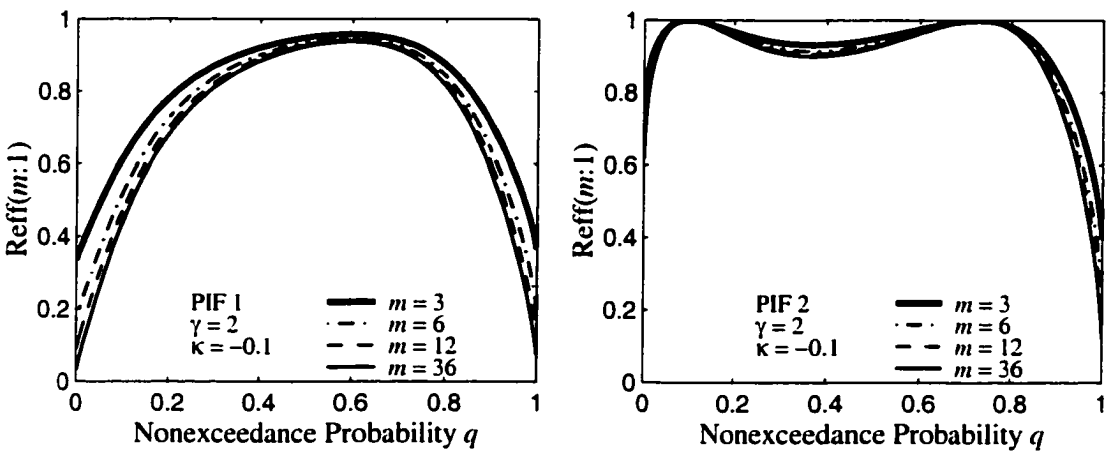

Figure 2.2: Asymptotic efficency of estimating the *qth* quantile at a single site using single site analysis relative to regional analysis of m sites, where the sample size n_j is the same at all sites and $\gamma = \alpha_j/\beta_j$ and κ are the same for all sites.

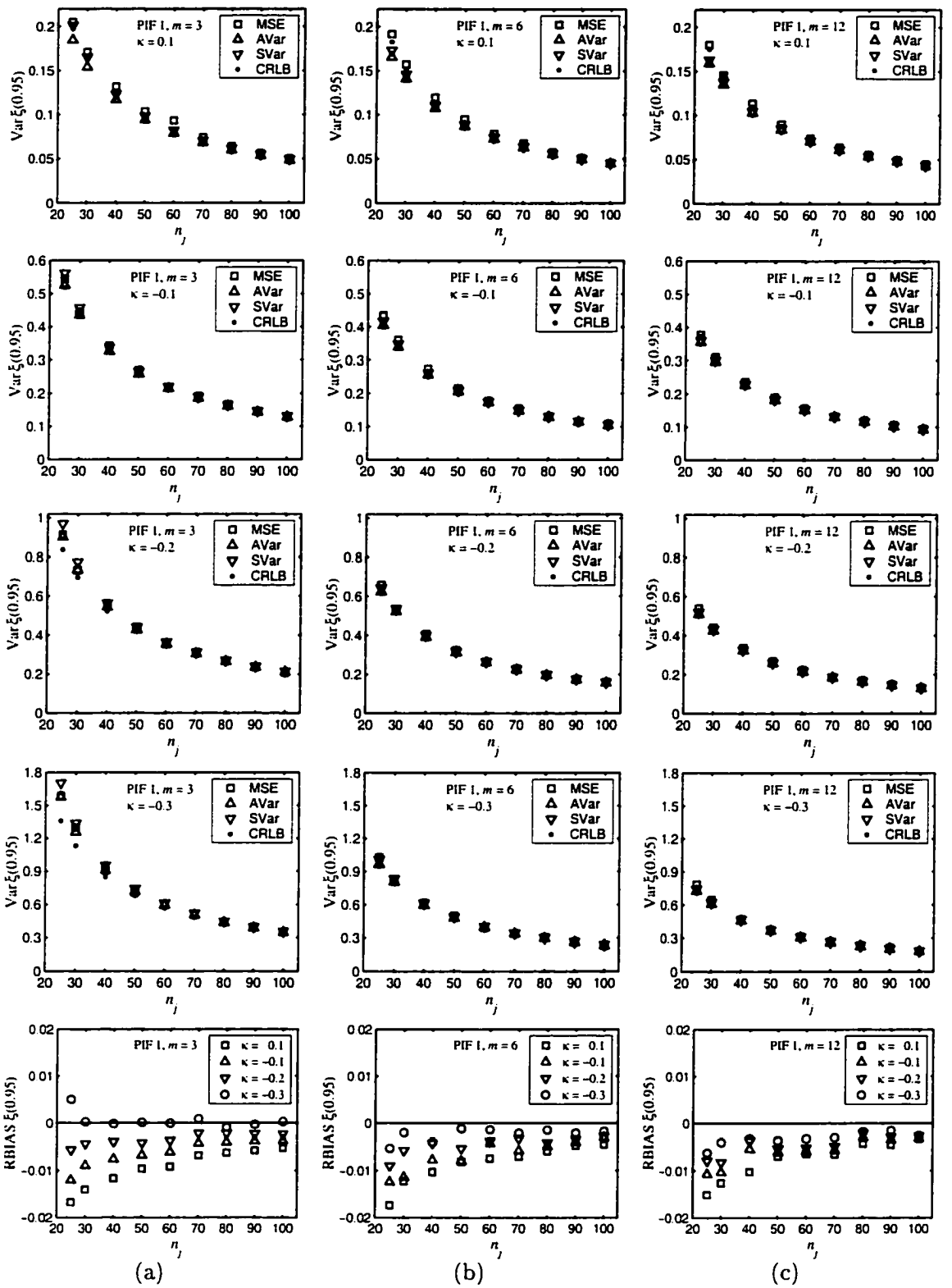

Figure 2.3: Simulation results based on the PIF 1 model and $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$ for the quantile $\xi(0.95)$ at site 1. The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

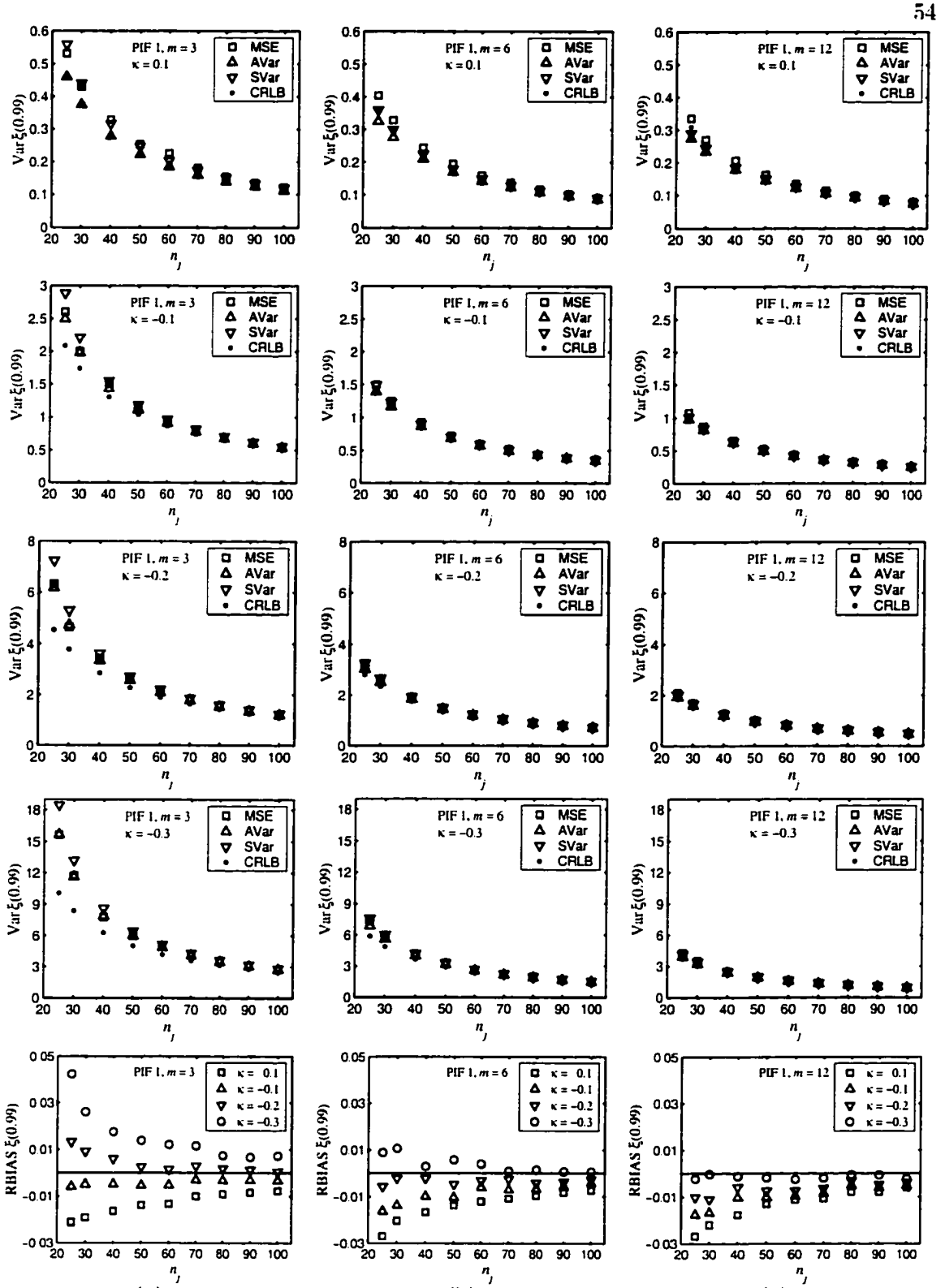

Figure 2.4: Simulation results based on the PIF 1 model and $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$ for the quantile $\xi(0.99)$ at site 1. The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

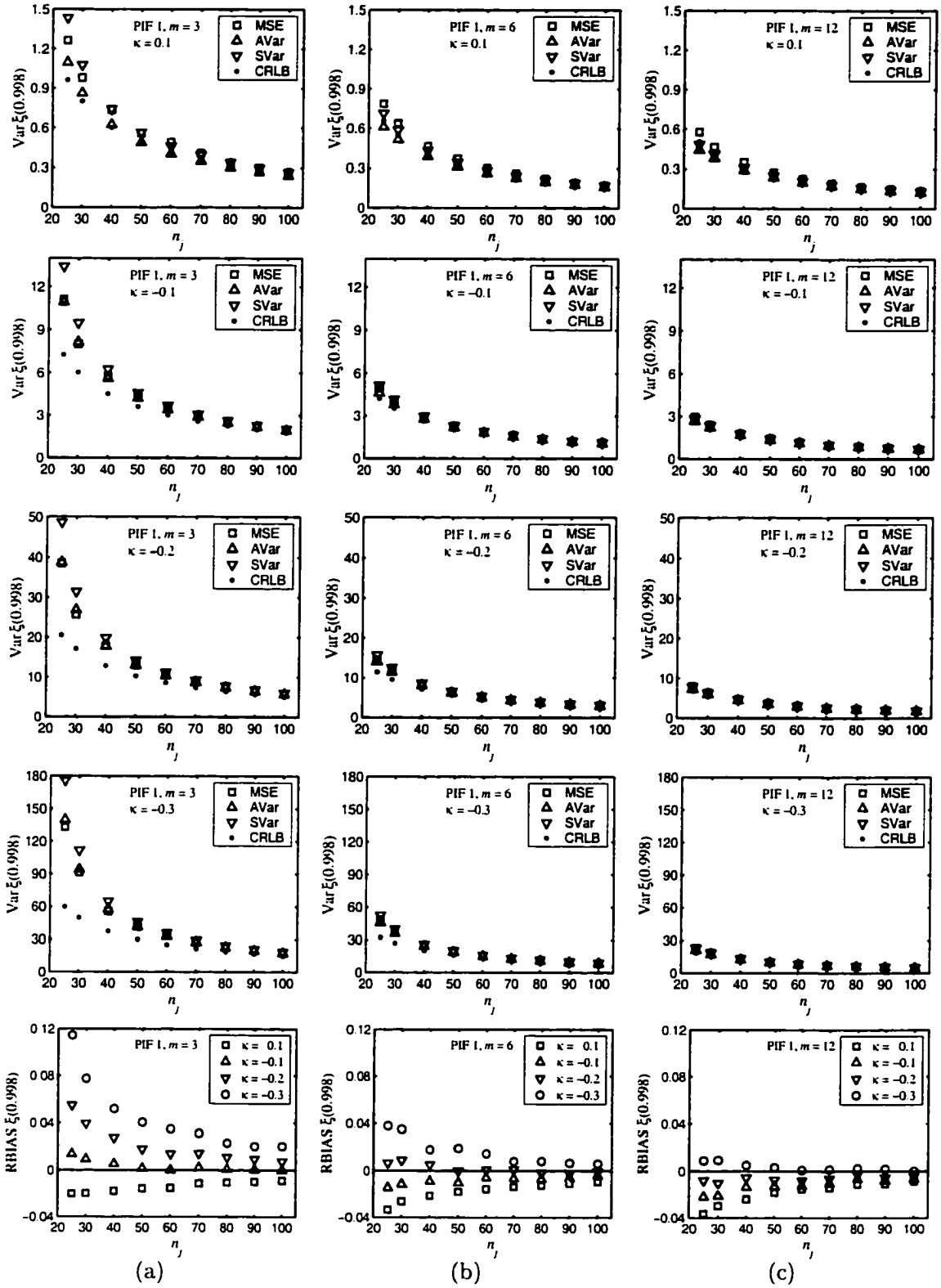

Figure 2.5: Simulation results based on the PIF 1 model and $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$ for the quantile $\xi(0.998)$ at site 1. The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

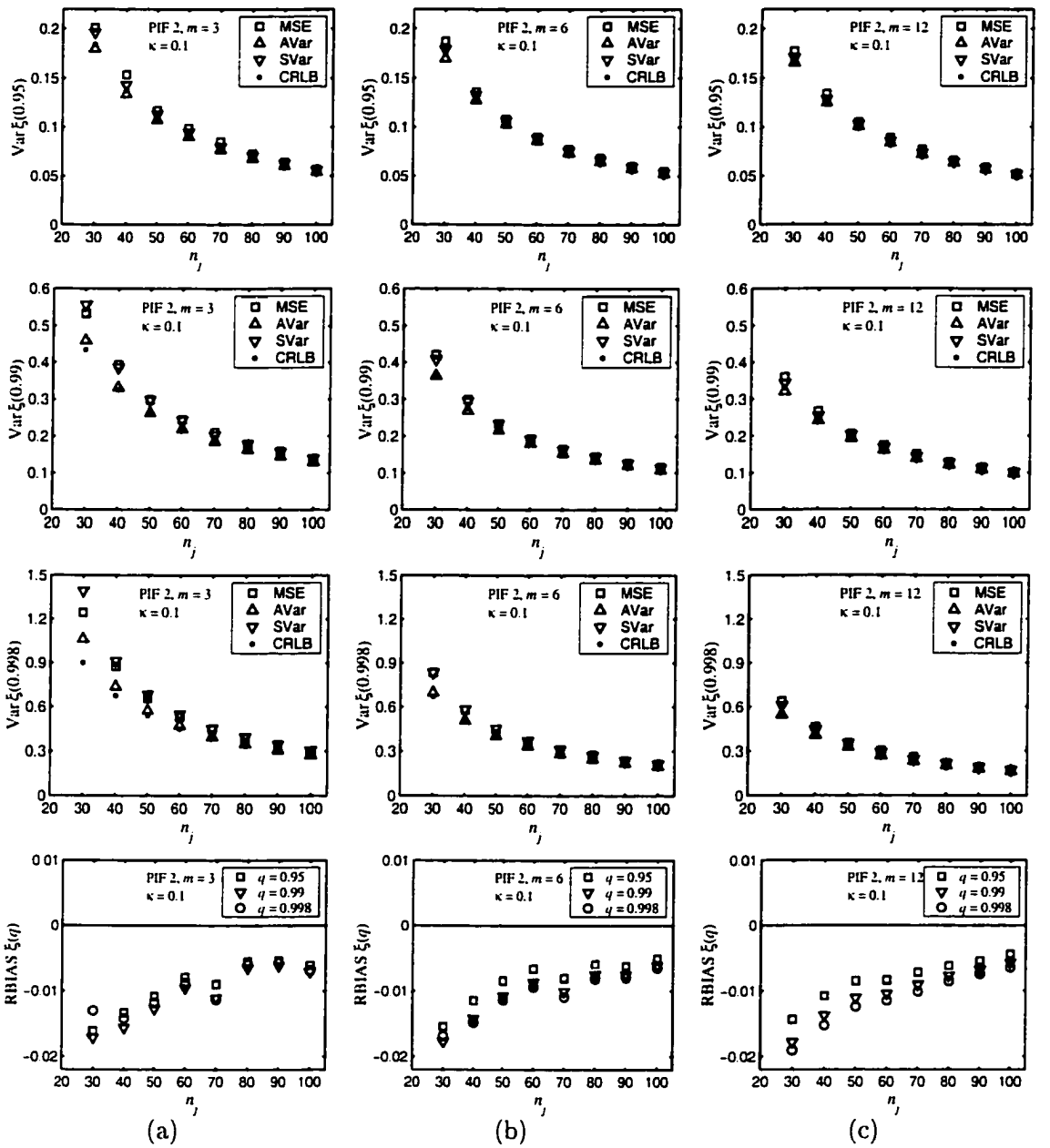

Figure 2.6: Simulation results based on the PIF 2 model with $\kappa = 0.1$ for the quantiles $\xi(0.95)$, $\xi(0.99)$, and $\xi(0.998)$ at site 1. The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

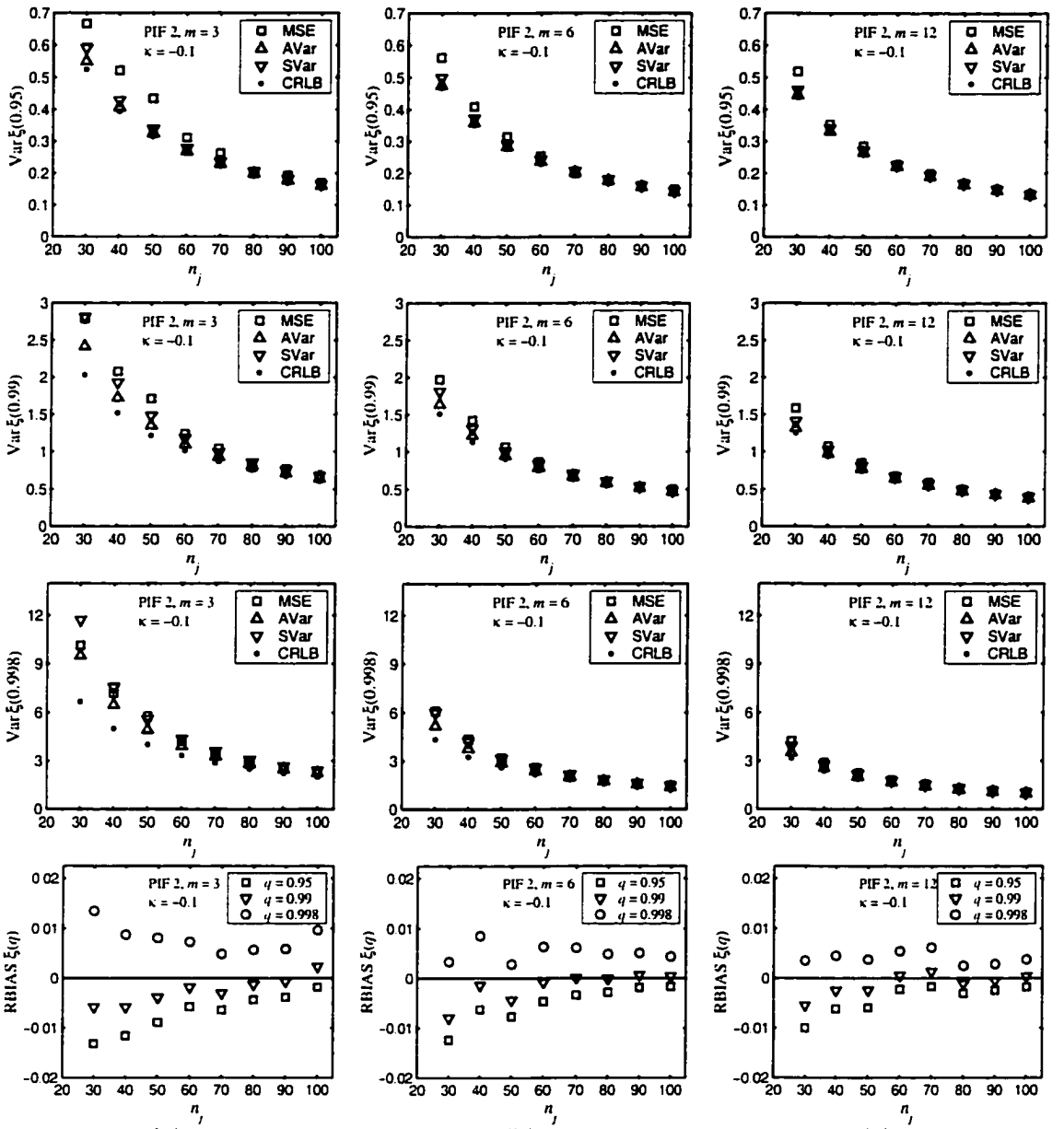

Figure 2.7: Simulation results based on the PIF 2 model with $\kappa = -0.1$ for the quantiles $\xi(0.95)$, $\xi(0.99)$, and $\xi(0.998)$ at site 1. The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

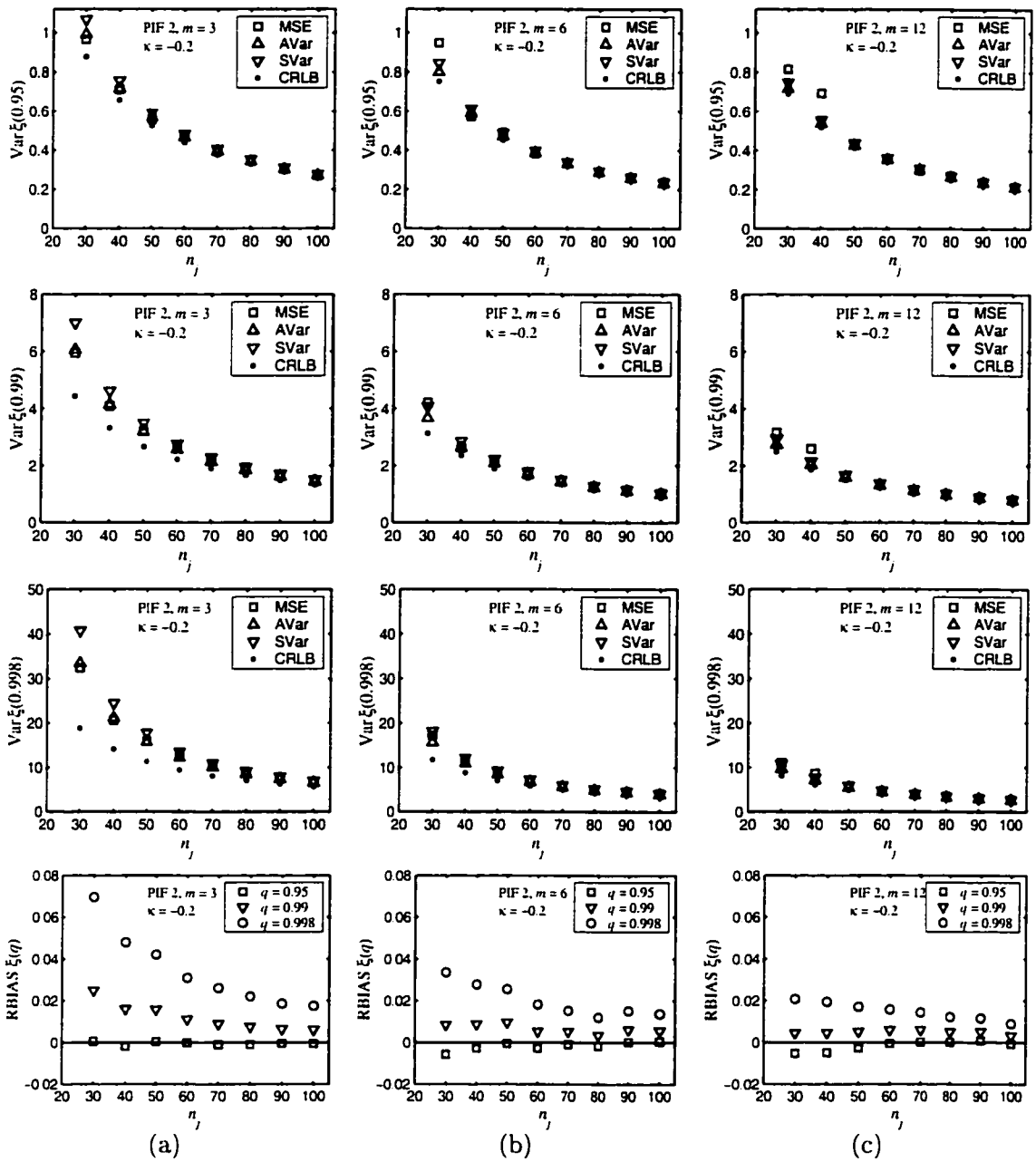

Figure 2.8: Simulation results based on the PIF 2 model with $\kappa = -0.2$ for the quantiles $\xi(0.95)$, $\xi(0.99)$, and $\xi(0.998)$ at site 1. The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

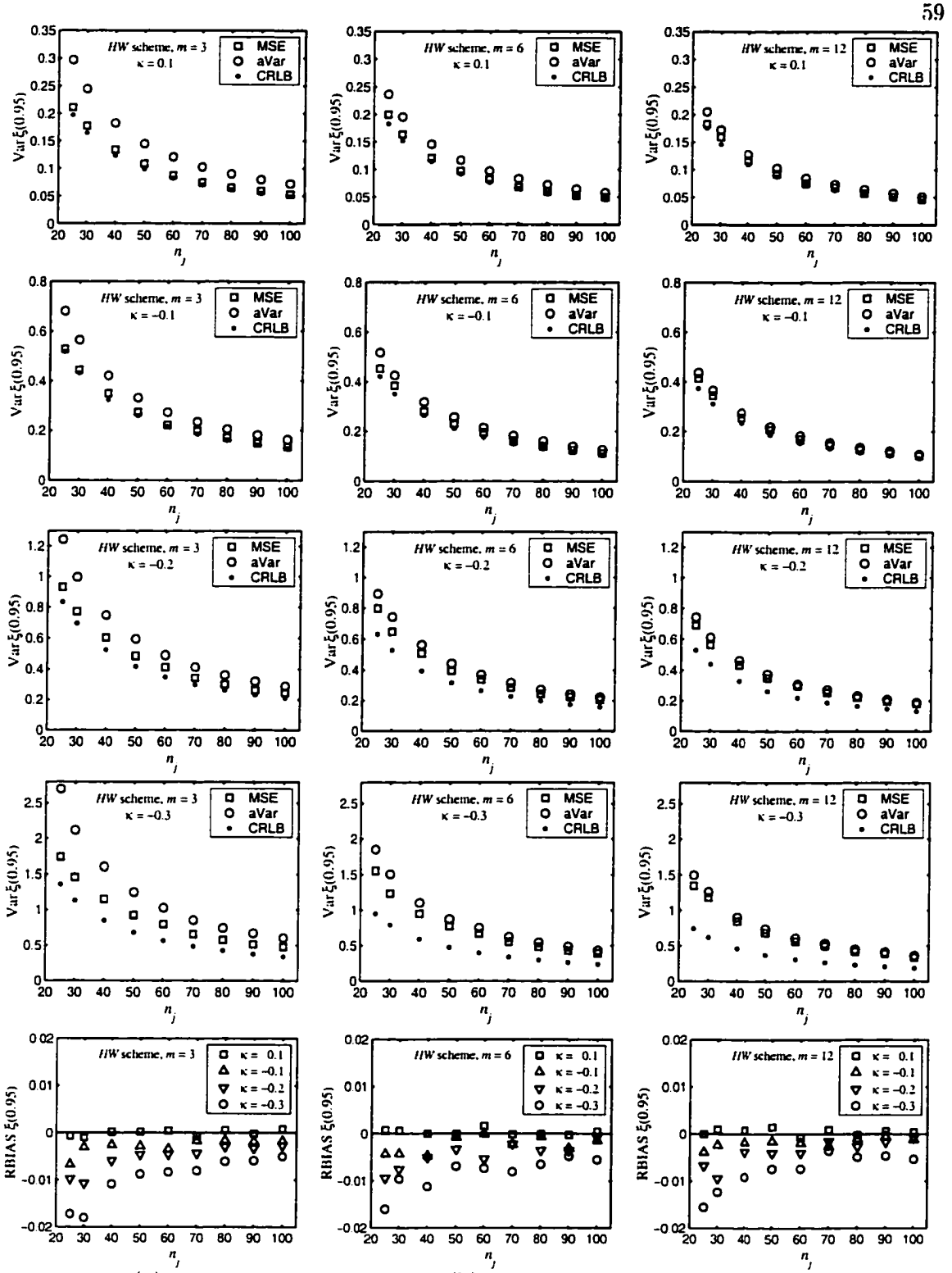

Figure 2.9: Simulation results based on the *HW* scheme and $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$ for the quantile $\xi(0.95)$ at site 1. The approximate variance (aVar) is estimated from Eqs. (2.21) and (2.22). The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

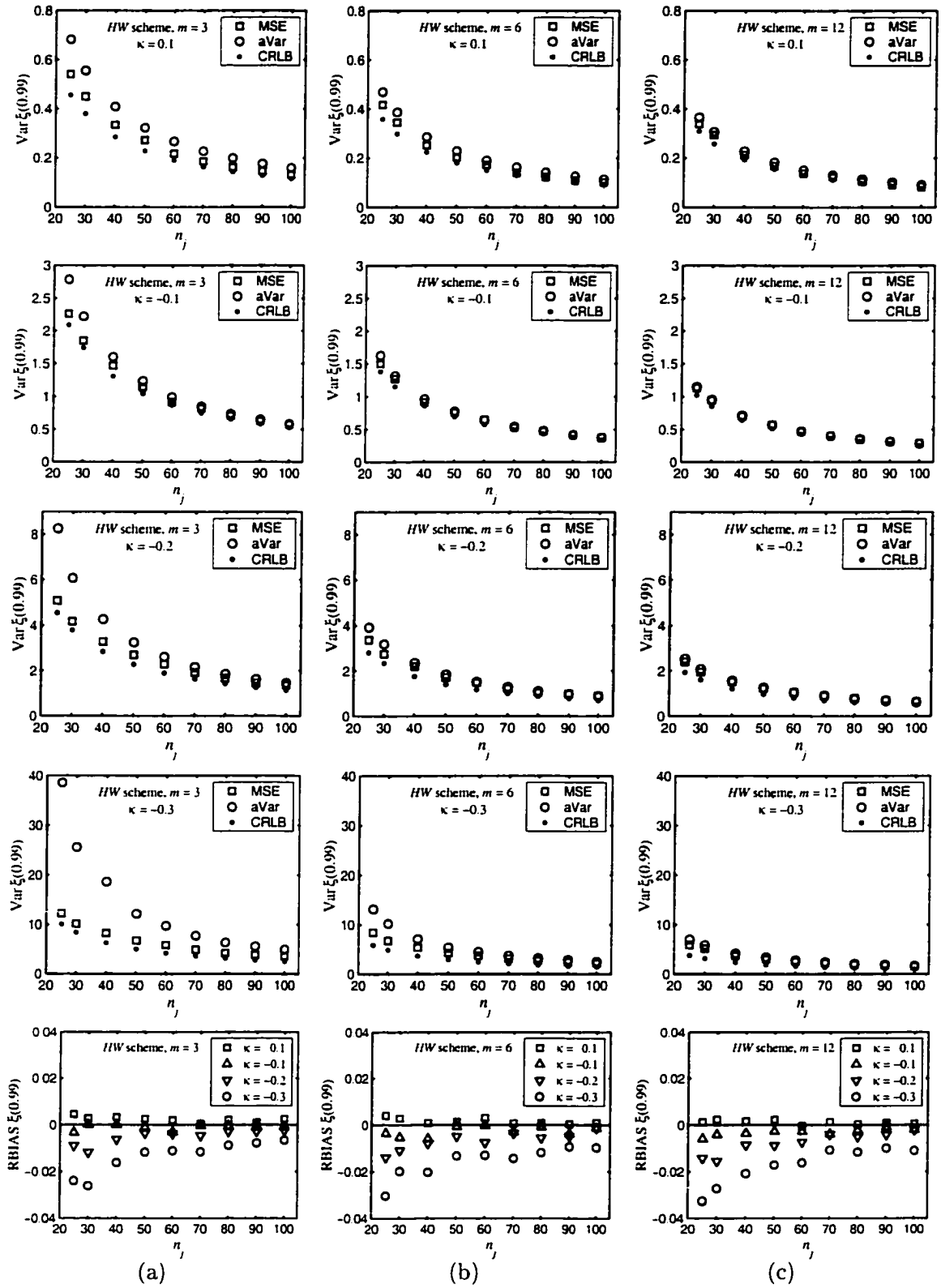

Figure 2.10: Simulation results based on the *HW* scheme and $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$ for the quantile $\xi(0.99)$ at site 1. The approximate variance (aVar) is estimated from Eqs. (2.21) and (2.22). The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

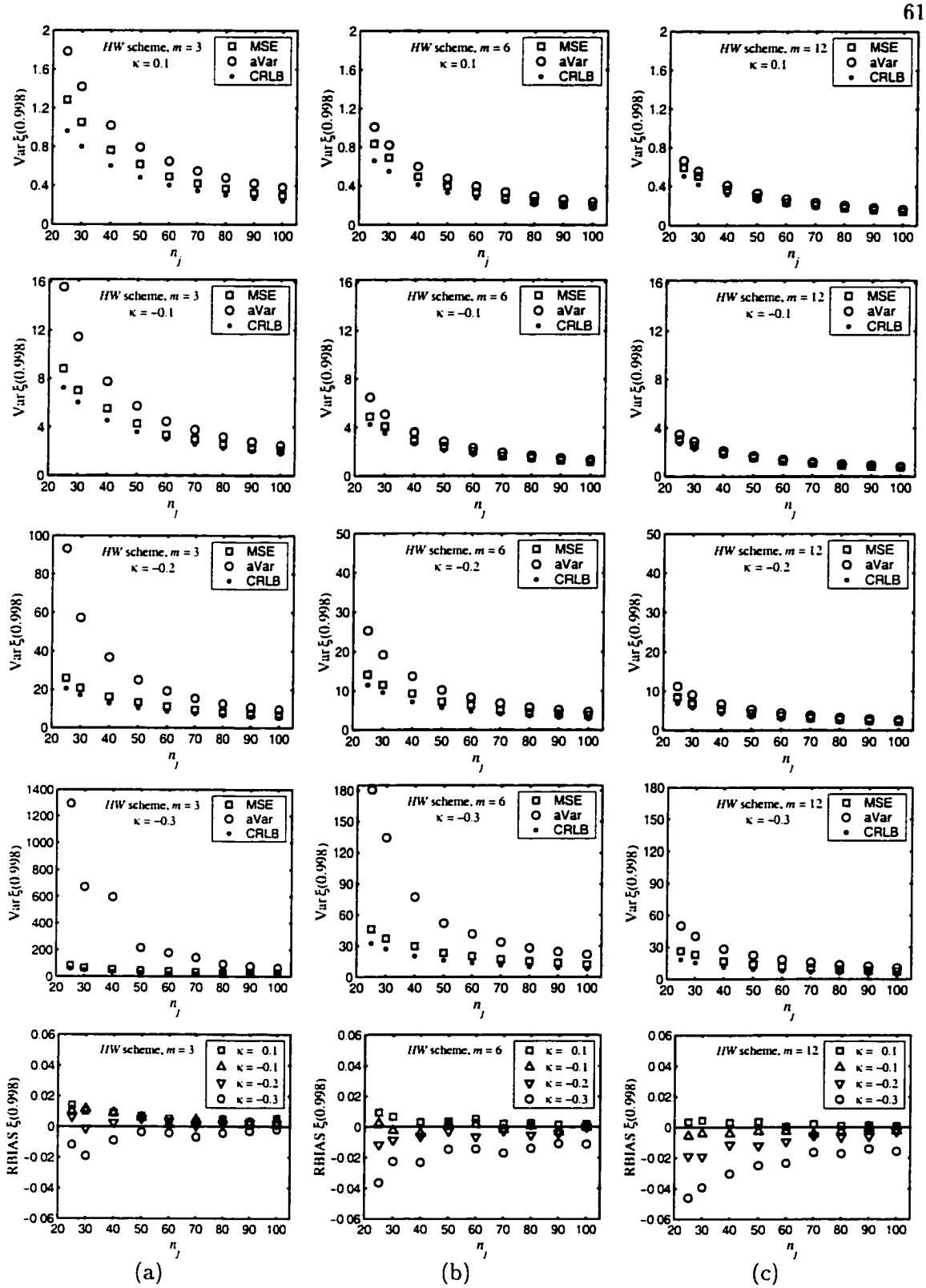

Figure 2.11: Simulation results based on the *HW* scheme and $\kappa \in \{-0.3, -0.2, -0.1, 0.1\}$ for the quantile $\xi(0.998)$ at site 1. The approximate variance (aVar) is estimated from Eqs. (2.21) and (2.22). The results are shown for a region of 3 sites in (a), 6 sites in (b), and 12 sites in (c).

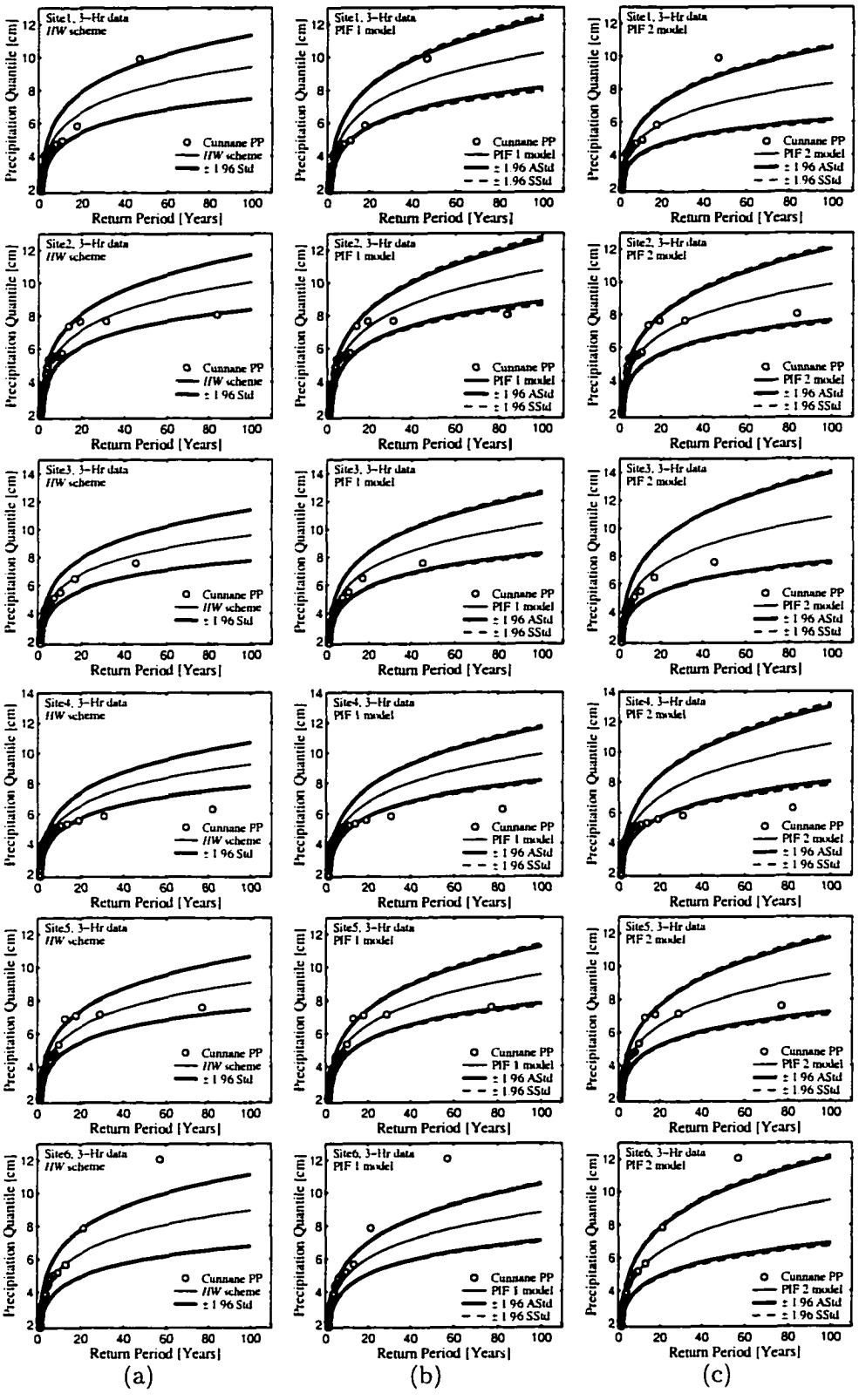

Figure 2.12: Estimated growth curve and approximate 95% confidence limits for sites 1–6 based on the *HW* scheme in (a), PIF 1 model in (b), and PIF 2 model in (c).

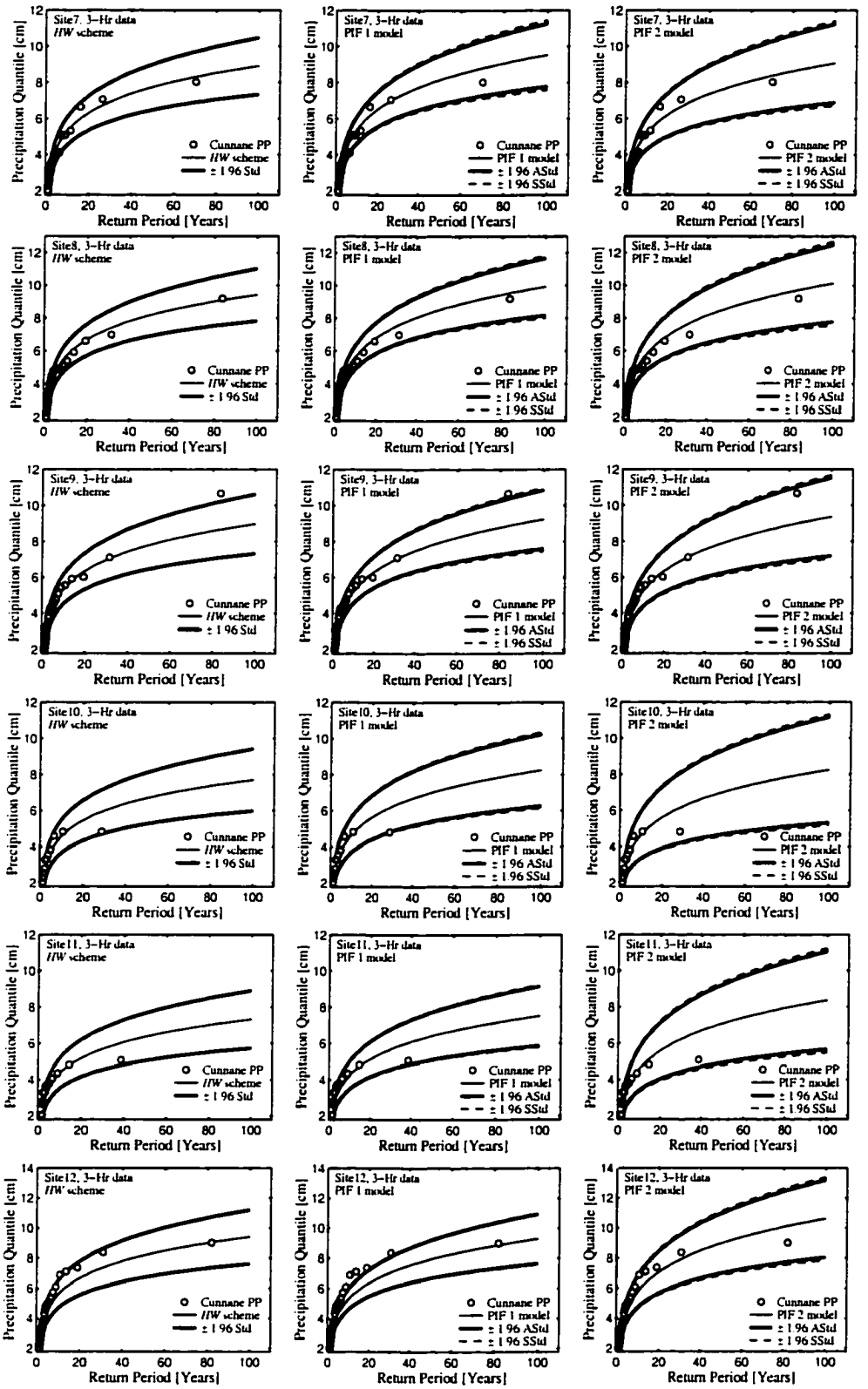

Figure 2.13: Estimated growth curve and approximate 95% confidence limits for sites 7–12 based on the *HW* scheme in (a), PIF 1 model in (b), and PIF 2 model in (c).

Chapter 3

ESTIMATION OF EXTREME-PARETO-QUANTILES USING UPPER ORDER STATISTICS

Abstract A common feature of certain hydrologic data plotted *vs.* their return period in a log-log scale is an apparent straight line fit at the upper end. Such feature is compatible with an assumed two parameter *Pareto* model for the upper end. The objective of this study is to utilize only the largest sample observations in estimation of extreme upper quantiles. Parameters of the *Pareto* are estimated based on the upper order statistics at a single site using maximum likelihood. Also, regional estimates are obtained under an assumed indexing method that yields a type of regional homogeneity. Exact formulas for the mean-squared-error of quantile estimators are given. Extreme precipitation data from the northern front range of Colorado, USA, are used for illustrating the procedures.

3.1 Introduction

In many parts of the world it is evident that an extreme hydrologic time series is a result of several physical mechanism, e.g. thunderstorms *vs.* hurricanes; snowmelt *vs.* rainfall; etc. This suggest that some hydrologic time series come from a mixed population.

Estimation of extreme events of a hydrologic process is often accomplished by fitting a density function to the sample record. The parent distribution of the hydrologic process is never exactly known; thus, distributions fitted to the data may fit well in the middle but not as well in the lower/upper tails.

A good fit in the tails is necessary, if the goal is to obtain accurate estimates of extreme quantiles in either tail. Improved fit in the tails of a distribution is sometimes achieved using probability paper and least-squares methods, where only the sample data that form an acceptable straight line in either tail are used. One could argue that low sample observations ought not be allowed to unduly influence the estimation of upper extreme quantiles, see e.g. Klemes (1987).

Our interest is in the extremes of the upper extreme events of our process. In our study we assume that these extremes are distributed as a *Pareto* with two parameters (scale and shape). Parameters are estimated from minimal sufficient statistics, which are functions of the k largest order statistics of the sample record, using maximum likelihood. A consequence of fitting a *Pareto* is the so called power-law, which yields a straight line fit of event magnitudes and their return periods in log-log space. We select k so as to get a "straight line" in the upper tail.

Two cases are considered: (1) quantile estimates from a single sample using the k largest order statistics, and (2) quantile estimates from regional data using upper order statistics. The accuracy of estimated quantiles is measured by their mean-squared-error. Procedures presented here are tested on extreme precipitation data from the north front range of Colorado, USA. In some cases computer simulations are used to compare different procedures.

3.2 The Pareto Distribution

A random variable X is said to have a *Pareto* distribution with scale parameter b and shape parameter a, denoted by $X \sim Pareto(b, a)$, if

$$f_X(x) = ab^a x^{-a-1} I_{(b,\infty)}(x) \tag{3.1}$$

$$F_X(x) = \left(1 - \left(\frac{b}{x}\right)^a\right) I_{(b,\infty)}(x) \tag{3.2}$$

where $a > 0$ and $b > 0$. $I_{(a,b)}(x)$ is the indicator function defined as

$$I_{(a,b)}(x) = \begin{cases} 1 & \text{if } x \in (a,b) \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

The mean and the variance of the random variable X are

$$E[X] = \frac{ab}{a-1}, \qquad a > 1 \tag{3.4}$$

$$Var(X) = \frac{ab^2}{(a-2)(a-1)^2}, \qquad a > 2 \tag{3.5}$$

Note that $\ln(X)$ is *exponentially* distributed with location parameter $\alpha = \ln b$, and scale parameter $\beta = 1/a$. That is, $Z = \ln(X) \sim exp(\alpha, \beta)$ with

$$f_Z(z) = \frac{1}{\beta} e^{-(z-\alpha)/\beta} I_{(\alpha,\infty)}(z) \tag{3.6}$$

3.3 Estimation of Parameters From k Largest Order Statistics

Assume that we have a random sample of size n, $\{X_i\}_{i=1}^n$, from a *Pareto* distribution. Denote the k largest order statistics $\{X_{n-k+1:n}, \ldots, X_{n:n}\}$ by $Y_1 = X_{n-k+1:n}, \ldots, Y_k = X_{n:n}$. Then the joint distribution of the k largest order statistics is

$$f_{Y_1,\ldots,Y_k}(y_1, \ldots, y_k) = \frac{n!}{(n-k)!} F_X^{n-k}(y_1) \prod_{i=1}^{k} f_X(y_i), \qquad b < y_1 \leq \cdots \leq y_k \tag{3.7}$$

and the log-likelihood is given by

$$\begin{aligned} \ln \mathcal{L}(b, a; \mathbf{y}) = &\ln \frac{n!}{(n-k)!} + (n-k) \ln \left[1 - \left(\frac{b}{y_1} \right)^a \right] \\ &+ k(\ln a + a \ln b) - (a+1) \sum_{i=1}^{k} \ln y_i \end{aligned} \tag{3.8}$$

where $\mathbf{y} = [y_1, \ldots, y_k]$. The partial derivatives with respect to the parameters are

$$\frac{\partial \ln \mathcal{L}}{\partial a} = -\frac{n-k}{1-(b/y_1)^a} \left(\frac{b}{y_1} \right)^a \ln \frac{b}{y_1} + k \left(\frac{1}{a} + \ln b \right) - \sum_{i=1}^{k} \ln y_i \tag{3.9}$$

$$\frac{\partial \ln \mathcal{L}}{\partial b} = -\frac{n-k}{1-(b/y_1)^a} \left(\frac{b}{y_1} \right)^a \frac{a}{b} + k \frac{a}{b} \tag{3.10}$$

Setting the partial derivatives equal to zero and solving them simultaneously for a and b gives the ML-estimates

$$\frac{1}{\hat{a}} = \frac{1}{k}\sum_{i=1}^{k} \ln y_i - \ln y_1 = \frac{1}{k}\sum_{i=2}^{k}(k-i+1)(\ln y_i - \ln y_{i-1}), \quad k > 1 \tag{3.11}$$

$$\hat{b} = \left(\frac{k}{n}\right)^{1/\hat{a}} y_1 \tag{3.12}$$

By equation (3.8), $(Y_1, \sum_{i=1}^{k} \ln Y_i)$ is sufficient via the factorization criterion; hence so is $(Y_1, \sum_{i=2}^{k}(k-i+1)(\ln Y_i - \ln Y_{i-1}))$, and utilizing the lack of memory property of an *exponential*, Y_1 and $\sum_{i=2}^{k}(k-i+1)(\ln Y_i - \ln Y_{i-1})$ are independent. It is easily shown that $(k-i+1)(\ln Y_i - \ln Y_{i-1}) \overset{iid}{\sim} exp(0, a^{-1})$, $i = 2, \ldots, k$. Since, $\hat{A}^{-1}$ is a sum of $(k-1)$ *iid* $exp(0, a^{-1}k^{-1})$ random variables, where $\hat{A}$ is the random variable associated with the estimate $\hat{a}$, it follows that $\hat{A}^{-1} \sim gamma(r = k-1, \beta = a^{-1}k^{-1})$ or

$$f_{\hat{A}^{-1}}(x) = \frac{1}{\Gamma(r)}\frac{1}{\beta}\left(\frac{x}{\beta}\right)^{r-1} e^{-x/\beta} I_{(0,\infty)}(x) \tag{3.13}$$

where $r > 0$, $\beta > 0$, and $\Gamma(\cdot)$ is the gamma function. Hence, $\check{A}^{-1} = (k/(k-1))\hat{A}^{-1}$ is an unbiased estimator of a^{-1} with $\check{A}^{-1} \sim gamma(k-1, a^{-1}(k-1)^{-1})$, and $\tilde{A} = \hat{A}(k-2)/k$ is an unbiased estimator of a with $\tilde{A}^{-1} \sim gamma(k-1, a^{-1}(k-2)^{-1})$.

3.4 Estimation of Parameters for a Region m Sites

Assume a region of m sites, with n_j representing sample size at site j, $j = 1, \ldots, m$. Further assume that observations at different sites are independent. The independent m samples at each site can be represented as: $\{X_{1i}\}_{i=1}^{n_1}, \ldots, \{X_{mi}\}_{i=1}^{n_m}$. In regional frequency analysis a widely used scheme is the *index flood method*, Dalrymple (1960). In this method, data at each site are scaled by dividing them by the at-site sample mean. Another scheme is to use standardized data at each site, see e.g. Elíasson (1997). The indexed data or the standardized data are assumed to be homogenous in the sense of coming from the same population.

From the model perspective the above two schemes imply that $\xi_j(q)/\mu_j$ for the indexed data, and $(\xi_j(q) - \mu_j)/\sigma_j$ for the standardized data, do not depend on j, where $\xi_j(q)$ is the qth quantile at site j and μ_j & σ_j are the population mean & standard deviation at site j. Thus, from equations (3.2), (3.4), and (3.5)

$$\xi_j(q)/\mu_j = \frac{a_j - 1}{a_j}(1-q)^{-1/a_j} \ , \quad \text{or} \qquad a_j > 1 \tag{3.14}$$

$$\frac{\xi_j(q) - \mu_j}{\sigma_j} = \sqrt{a_j - 2}\left(\frac{a_j - 1}{\sqrt{a_j}}(1-q)^{-1/a_j} - \sqrt{a_j}\right) \ , \qquad a_j > 2 \tag{3.15}$$

should not depend on j for the respective schemes. In this case all the shape parameters, a_j, must be the same, and consequently the parameter space is reduced from $2m$-dimensions to $(m+1)$-dimensions. That is the data at site j are distributed as: $X_{j1}, \ldots, X_{jn_j} \sim F_{X_j.}(\cdot; a, b_j)$. As in the previous section, let $\mathbf{Y}_j = [X_{j,n_j-k_j+1:n_j}, \ldots, X_{j,n_j:n_j}]$ be the vector of the k_j largest order statistics from site j. The joint distribution of the m independent vectors $\mathbf{Y}_j$ is

$$\begin{aligned} f_{\mathbf{Y}_1,\ldots,\mathbf{Y}_m}(\mathbf{y}_1, \ldots, \mathbf{y}_m) &= \prod_{j=1}^{m} f_{\mathbf{Y}_j}(\mathbf{y}_j) \\ &= \prod_{j=1}^{m}\left(\frac{n_j!}{(n_j-k_j)!}F_{X_j.}^{n_j-k_j}(y_{j1})\prod_{i=1}^{k_j} f_{X_j.}(y_{ji})\right) \end{aligned} \tag{3.16}$$

where y_{ji} is the ith element of the vector $\mathbf{y}_j$. Then the log-likelihood is

$$\begin{aligned} \ln \mathcal{L}(b_1, \ldots, b_m, a; \mathbf{y}_1, \ldots, \mathbf{y}_m) = \sum_{j=1}^{m}\Bigg(&\ln\frac{n_j!}{(n_j-k_j)!} + (n_j - k_j)\ln\left[1 - \left(\frac{b_j}{y_{j1}}\right)^a\right] \\ &+ k_j(\ln a + a\ln b_j) - (a+1)\sum_{i=1}^{k_j}\ln y_{ji}\Bigg) \end{aligned} \tag{3.17}$$

Taking the partial derivatives of the log-likelihood with respect to the parameters and setting them equal to zero gives the ML-estimates

$$\frac{1}{\hat{a}} = \frac{1}{k_T}\sum_{j=1}^{m}\sum_{i=2}^{k_j}(\ln y_{ji} - \ln y_{j1}) \tag{3.18}$$

$$\hat{b}_j = \left(\frac{k_j}{n_j}\right)^{1/\hat{a}} y_{j1} \ , \qquad j = 1, \ldots, m \tag{3.19}$$

where $k_T = \sum_{j=1}^{m} k_j$. Equation (3.18) can be rewritten as $\hat{a}^{-1} = k_T^{-1} \sum_{j=1}^{m} \sum_{i=2}^{k_j} (k_j - i + 1)(\ln y_{ji} - \ln y_{j(i-1)})$. By equation (3.17), $(Y_{11}, \ldots, Y_{m1}, \sum_{j=1}^{m} \sum_{i=1}^{k_j} \ln Y_{ji})$ is sufficient; hence also $(Y_{11}, \ldots, Y_{m1}, \sum_{j=1}^{m} \sum_{i=2}^{k_j} (k_j - i + 1)(\ln Y_{ji} - \ln Y_{j(i-1)}))$. As in the single site case, utilizing the lack of memory property of an *exponential* Y_{j1}, $j = 1, \ldots, m$, and $\sum_{j=1}^{m} \sum_{i=2}^{k_j} (k_j - i + 1)(\ln Y_{ji} - \ln Y_{j(i-1)})$ are independent. Notice that $\hat{A}^{-1}$ is a sum of $(k_T - m)$ *iid exponential* random variables, hence $\hat{A}^{-1} \sim gamma(r = k_T - m, \beta = a^{-1} k_T^{-1})$. Thus, $\breve{A}^{-1} = (k_T/(k_T - m))\hat{A}^{-1}$ is an unbiased estimator of a^{-1} with $\breve{A}^{-1} \sim gamma(k_T - m, a^{-1}(k_T - m)^{-1})$, and $\tilde{A} = (1 - (m+1)/k_T)\hat{A}$ is an unbiased estimator of a with $\tilde{A}^{-1} \sim gamma(k_T - m, a^{-1}(k_T - m - 1)^{-1})$.

3.5 Mean-Squared-Error of Estimated Quantiles

From equation (3.2) the *qth Pareto* quantile denoted as $\xi(q)$ is

$$\xi(q) = \frac{b}{(1-q)^{1/a}} \tag{3.20}$$

An estimate of $\xi(q)$ is

$$\hat{\xi}(q) = \frac{\hat{b}}{(1-q)^{1/\hat{a}}} \tag{3.21}$$

where $\hat{a}$ & $\hat{b}$ are estimates of a & b, respectively. For a single site, a general estimator of a has the property $\hat{A}^{-1} \sim gamma(r = k - 1, \beta = a^{-1}(k - j)^{-1})$ where for $j = 0$, $\hat{A}$ is ML-estimator of a; $j = 1$, $\hat{A}^{-1}$ is unbiased estimator of a^{-1}; $j = 2$, $\hat{A}$ is unbiased estimator of a. For all those cases an estimate for b is given in equation (3.12). This estimate of b is biased,

$$E[\hat{B}] = b \frac{B(k - 1/a, n - k + 1)}{B(k, n - k + 1)} (1 - \beta \ln(k/n))^{-(k-1)}, \quad k > \max(1, 1/a) \tag{3.22}$$

but much simpler to use in the following derivations than an unbiased estimate of b. $B(\cdot, \cdot)$ is the Beta function. Substituting $\hat{b}$ from equation (3.12) into equation (3.21) gives

$$\hat{\xi}(q) = \left(\frac{k/n}{1-q}\right)^{1/\hat{a}} y_1 = c^{1/\hat{a}} y_1 \tag{3.23}$$

where $c = (k/n)/(1-q)$. From $\hat{A}^{-1} \sim gamma(r, \beta)$ it follows that the ith raw moment of $c^{1/\hat{A}}$ is

$$E[c^{i/\hat{A}}] = (1 - i\beta \ln c)^{-r}, \qquad (1 - i\beta \ln c) > 0, \quad i \in \mathcal{N} \tag{3.24}$$

Recalling that Y_1 is the $(n-k+1)$ smallest order statistic from a random *Pareto* sample, the density function of Y_1 is

$$f_{Y_1}(y) = \frac{n!}{(n-k)!(k-1)!} F_X^{n-k}(y) f_X(y) (1 - F_X(y))^{k-1} \tag{3.25}$$

Using equations (3.1) & (3.2) in (3.25), the ith raw moment of Y_1 is

$$E[Y_1^i] = \frac{n!}{(n-k)!(k-1)!} \int_b^\infty y^i \left(1 - \left(\frac{b}{y}\right)^a\right)^{n-k} a b^a y^{-a-1} \left(\frac{b}{y}\right)^{a(k-1)} dy \tag{3.26}$$

The above integral is easily solved by using the transformation $u = (b/y)^a$, and

$$E[Y_1^i] = b^i \frac{B(k - i/a, n-k+1)}{B(k, n-k+1)} = b^i BB(i), \qquad k > i/a, \quad i \in \mathcal{N} \tag{3.27}$$

where $BB(i) = B(k - i/a, n-k+1)/B(k, n-k+1)$. Since $\hat{A}$ and Y_1 are independent the mean and the variance of $\hat{\Xi}(q)$, the random variable with value $\hat{\xi}(q)$, are

$$E[\hat{\Xi}(q)] = E[c^{1/\hat{A}}]\, E[Y_1] \tag{3.28}$$

$$Var(\hat{\Xi}(q)) = E[c^{2/\hat{A}}]\, E[Y_1^2] - (E[c^{1/\hat{A}}]\, E[Y_1])^2 \tag{3.29}$$

The mean-squared-error of $\hat{\Xi}(q)$ is defined as

$$MSE(\hat{\Xi}(q)) = Var(\hat{\Xi}(q)) + (E[\hat{\Xi}(q)] - \xi(q))^2 \tag{3.30}$$

where $E[\hat{\Xi}(q)] - \xi(q)$ is the bias of the quantile estimator.

For estimation of the parameter a it is not clear if it is more accurate to use ML-estimates or unbiased estimates of $1/a$ or a. Figure 3.1 (a) and (b) show simulation results where these three different estimation procedures of a are compared for typical population values of a and b and sample size n. Using the ML-estimate of a results in smallest bias and smallest rmse, where rmse $= MSE^{0.5}$.

For the special case when the estimates for a and b are the ML-estimates then the mean-squared-error for a single site is

$$MSE(\hat{\Xi}(q)) = b^2 BB(2)(1-2\beta\ln(c))^{-r} - b^2 BB^2(1)(1-\beta\ln(c))^{-2r} + \xi^2(q)[(1-q)^{1/a} BB(1)(1-\beta\ln(c))^{-r} - 1]^2 \quad (3.31)$$
$$for \quad k > \max(1, 2/a), \quad 1-2\beta\ln(c) > 0$$

where $r = k-1$ and $\beta = a^{-1}k^{-1}$. Similarly the mean squared error for site j in a region of m sites is

$$MSE(\hat{\Xi}_j(q)) = b_j^2 BB_j(2)(1-2\beta\ln(c_j))^{-r} - b_j^2 BB_j^2(1)(1-\beta\ln(c_j))^{-2r} + \xi_j^2(q)[(1-q)^{1/a} BB_j(1)(1-\beta\ln(c_j))^{-r} - 1]^2 \quad (3.32)$$
$$for \quad k_j > 2/a, \quad k_T > m, \quad 1-2\beta\ln(c_j) > 0$$

where $r = k_T - m$, $\beta = a^{-1}k_T^{-1}$, $c_j = (k_j/n_j)/(1-q)$, and $BB_j(i) = B(k_j - i/a, n_j - k_j + 1)/B(k_j, n_j - k_j + 1)$.

3.6 Example

Fort Collins, Colorado (USA), experienced an extreme precipitation event on July 28 1997. The flood that followed claimed the lives of five people and caused extensive property damage. The maximum point rainfall was estimated to be around 36.8*cm* southwest of Fort Collins. At Colorado State University gaging station, the *2-hr* storm event maximum was 9.6*cm* and the *6-hr* maximum was 13.5*cm*. These maximums were the largest ever recorded at this gaging station. Previously largest maximums were 7.11*cm* in 1992 for the *2-hr* AMP (annual maximum precipitation) series, and 7.62*cm* in 1951 for the *6-hr* AMP series.

2-hr and *6-hr* extreme precipitation data from Fort Collins are used to test procedures for single site analysis. For regional analysis, data from two nearby sites (Boulder and Longmont) are used along with the Fort Collins data. The hourly AMP series for all stations span 1949-1997. The years 1985 and 1990 for Longmont are excluded in the analysis because of long periods of missing data from those years.

Replacing q in equation (3.20) by $1 - 1/T$ the natural log of the T-year event can be expressed as

$$\ln \xi(T) = \ln b + \frac{1}{a} \ln T \tag{3.33}$$

Hence, a plot of *Pareto* sample events against their estimated return periods should yield a straight line in log-log space. This creates a selection criterion for the number of upper order statistics, k, to use. k should be selected so as to get approximate straight line in the upper tail of a $\ln X$ *vs.* $\ln T$ plot, where X stands for sample event.

Figure 3.1 (c) and (d) show plots of $\ln X$ *vs.* $\ln T$ for *2-hr* and *6-hr* Fort Collins AMP data, where T is estimated using the so called Weibull plotting position formula. Figure 3.1 (e) shows estimated 100-year event along with estimated bias and rmse for *2-hr* Fort Collins data using k between 2 and 19 inclusively; graph (f) shows the same for *6-hr* Fort Collins data using k between 2 and 18 inclusively. An estimate of the 100-yr *2-hr* event could be taken as an average of the estimated events for k between 6 and 12 inclusively. This would result in estimated 100-yr *2-hr* event of $10.9cm$ with an average rmse of $4.1cm$. Using an average for $k \in [8, 18]$ for the *6-hr* data results in estimated 100-yr *6-hr* event of $12.8cm$ with an average rmse of $4.7cm$.

Selected upper order statistics of the regional data are plotted in Fig. 3.1 (g) and (h). Recall that a is estimated for the whole region, while b is estimated individually for each site. For estimation of a one would like the slopes of the individual data sets of the region to be similar. For the *2-hr* regional data it seems appropriate to reduce k for Fort Collins from 19 to 12. Using $k = 12$ for Fort Collins; $k = 12$ for Longmont; and $k = 13$ for Boulder results in an estimated 100-yr *2-hr* event for Fort Collins of $8.9cm$ with rmse $= 1.5cm$. Similarly for the *6-hr* regional data using $k = 18$ for Fort Collins; $k = 15$ for Longmont; and $k = 12$ for Boulder results in an estimated 100-yr *6-hr* event for Fort Collins of $10.7cm$ with rmse $= 1.9cm$.

3.7 Concluding Remarks

The procedures presented in this paper seem useful when the sample data appear to be from a mixed population. In the application example the 100-year *2-hr* and *6-hr* event estimates for Fort Collins were significantly larger when only the Fort Collins data were used as opposed to the regional data. The regional estimates fell well within one rmse of the at-site estimates. For the at-site estimates extrapolation up to about twice the at-site sample size takes place, making the at-site estimates less trustworthy.

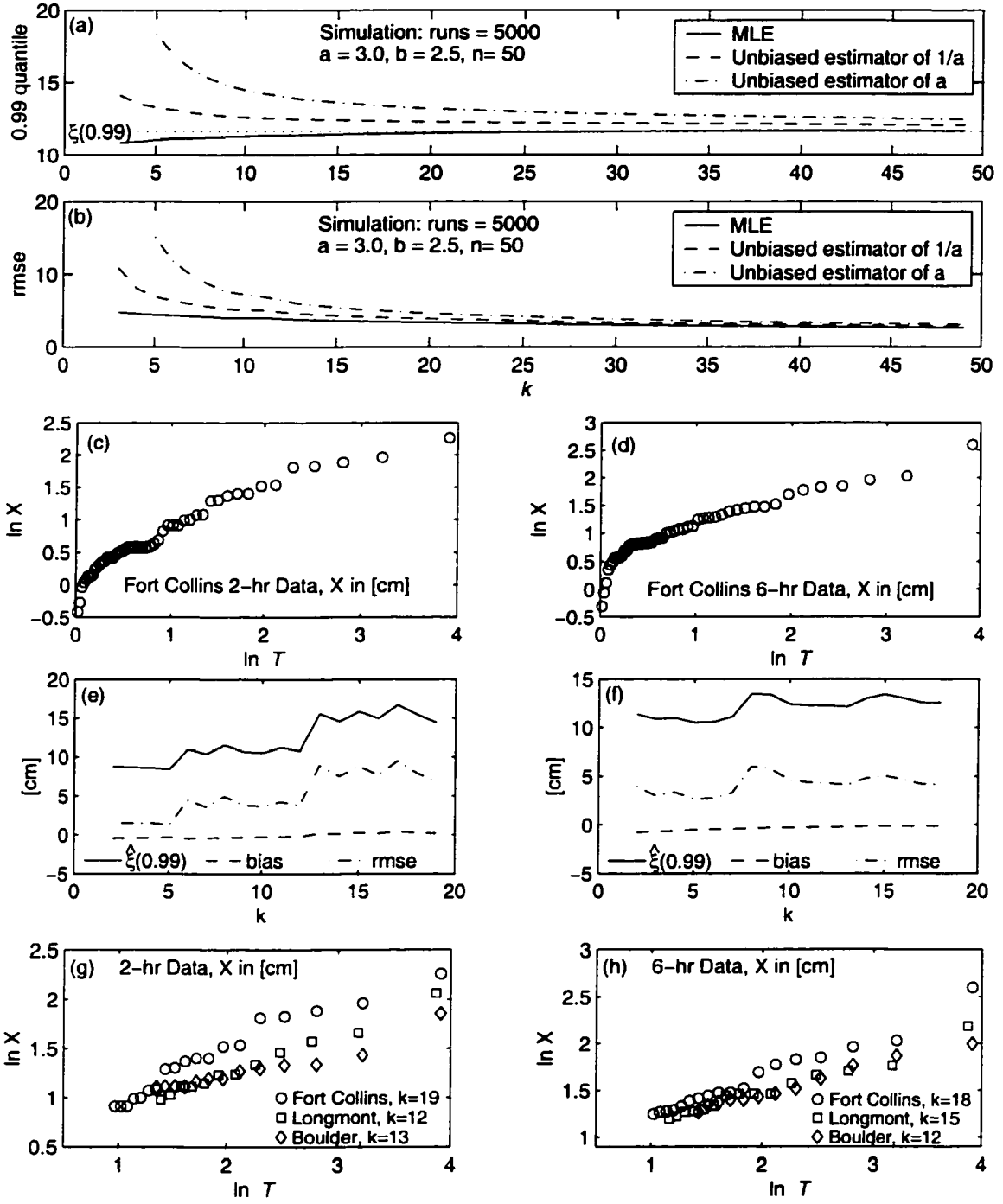

Figure 3.1: (a) and (b) simulation results comparing different parameter estimation procedures. (c) - (f) analysis of *2-hr* and *6-hr* Fort Collins data. (g) and (h) selected upper order statistics of the regional data.

Chapter 4

MODELING THE DYNAMICS OF LONG TERM VARIABILITY OF HYDROCLIMATIC PROCESSES

Abstract The stochastic analysis, modeling, and simulation of climatic and hydrologic processes such as precipitation, streamflow, and sea surface temperature have usually been based on assumed stationarity or randomness of the process under consideration. However, empirical evidence of many hydroclimatic data shows temporal variability involving trends, oscillatory behavior, and sudden shifts. While many studies have been made for detecting and testing the statistical significance of these special characteristics, the probabilistic framework for modeling the temporal dynamics of such processes appears to be lacking. In this paper we propose a family of stochastic models that can be used to capture the dynamics of sudden shifts in hydroclimatic time series. The applicability of such "shifting mean models" are illustrated by using time series data of annual PDO indices and streamflows.

4.1 Introduction

On the annual time-scale, the analysis of climatic and hydrologic processes is often based on assumed stationarity under a time series framework or randomness under a probabilistic framework. While this assumption may be reasonable within a short timeframe, empirical evidence show that most hydroclimatic processes deviate from stationarity in the long term. To some extent the assumption of stationarity has persisted because most historical records have been too short to accurately detect non-stationarity, and because of lack of

mathematical frameworks for analyzing and modeling the dynamics of non-stationary processes. However, as record lengths have increased, trends, oscillatory behavior, and sudden shifts have been observed in sample records.

The main objective of this paper is to study processes that shift abruptly from one stationary state into another. It appears that the first concept behind modeling sudden shifts in hydrologic time series was advanced by Hurst (1957). Subsequently Klemes (1974) and Potter (1976) further argued about the usefulness of algorithms to model shifting behavior. Boes and Salas (1978) developed some special cases of shifting mean models and Salas and Boes (1980) further discussed their applicability to hydrology and their conceptual justification. In this paper, we expand on the earlier concepts and models and develop more versatile models that can be useful for simulating the dynamics of hydroclimatic processes exhibiting sudden shifts. The methods suggested here are intended for simulation and generation of long sample records, rather than for forecasting.

Shifts in hydrologic processes may be related to climate changes (Matalas, 1997). Different indices, such as oscillation indices that measure pressure or temperature differences between two locations, or solar indices that measure sunspot activity, are sometimes used to reflect climatic fluctuations. These indices often appear to change quasi-periodically with time, or shift from one random stationary state to another. Taylor (1999) investigated ice cores from Greenland glacier to get information about climate and climate changes in the past. He concluded, that climate changes large enough to cause significant impacts on the society have occurred in the past over a duration less than 10 years. His ice core measurements also indicated that 11,700 years ago, when the climate in the North Atlantic shifted from a dry and cold ice age to a wetter and warmer climate, most of the shift occurred during a period of only 40 years. Kite (1989) investigated changes in lake levels and streamflow sequences. He used time series analysis and spectral analysis to detect linear trends, jumps, periodicities and other components in the hydrologic sequences. He concluded that any such detected statistical components were not the result of a climate change induced

by the so-called "greenhouse effect". Likewise, Chiew and McMahon (1993) concluded that there was no clear evidence to suggest that trends or change in mean flow volumes of Australian rivers were caused by climate changes. They also commented that their results might be affected by short sample records with high inter-annual variability. Angel and Huff (1997) investigated changes in heavy rainfall in the Midwestern United States. Their results showed, that stations in the entire Midwest are more likely to experience their heaviest rainfall in more recent years.

Others have tried to use oscillation indices such as the Southern Oscillation (SO) often related to El Niño or La Niña depending on its phase, or the Pacific Decadal Oscillation (PDO) to explain observed variability in historical sample series. For example, Waylen and Caviedes (1986) used a three component mixed Gumbel distribution to fit the annual maximum floods on the north Peruvian littoral. The annual maximum flood series were looked at as being produced by three different mechanisms corresponding to different ocean-atmosphere conditions: El Niño; normal; and La Niña. Eltahir (1996) correlated the annual Nile River flows with ENSO (El Niño Southern Oscillation which is based on sea surface temperature). His results suggested that 25% of the variance of the annual Nile River flows was associated with ENSO. Hamlet and Lettenmaier (1999) forecasted Columbia river flows based on a given forecast of ENSO and the phase of PDO. Their results suggested that including such climatic information into streamflow forecasting models could result in an increase in forecast lead time of about six months. These studies indicate the importance of using indices that represent climatic variability in prediction of extreme events, such as floods and droughts, and in forecasting of hydroclimatic processes, such as river flows. Modeling the dynamics of climatic processes or their related processes that exhibit sudden shifts is the main subject of this paper.

4.2 Some Examples of Hydroclimatic Time Series Exhibiting Sudden Shifts

There are many examples of time series of hydroclimatic processes that show evidence of sudden shifts in one or more of their statistical properties. For instance, Mantua et al. (1997), Niebauer (1998), and Hamlet and Lettenmaier (1999) analyzed the time series of annual averages of the PDO index based on sea surface temperature. Their results suggested that several shifts may have occurred since 1900. The latest such shift is suggested to have occurred around 1977. Mantua et al. (1997) defined the PDO to be in a cold phase during 1900–1924 and 1947–1976, and in a warm phase during the periods 1925–1946 and 1977–1996. Hamlet and Lettenmaier (1999) argued that based on the Columbia river flows the phase of the PDO might have shifted again in 1996 or 1997. Figure 4.1 (a) and (b) shows an update of the annual averages of the PDO index from Mantua et al. (1997). The data were downloaded from "ftp://ftp.atmos.washington.edu/mantua/pnw_impacts/INDICES/PDO.latest". In Fig. 4.1 (a) the index time series is shown along with averages of dominant phases as suggested by Mantua et al. (1997). In Fig. 4.1 (b) the index series is shown with alternative choice of phases. The models proposed in this paper are intended for time series that show similar sudden shifting structure as in graph Fig. 4.1 (b). Niebauer (1998) argued that before the regime shift in 1977, El Niño and La Niña conditions were about even, but after the regime shift El Niño conditions were about three times more prevalent, due to a change of location of the Aleutian low.

Gray et al. (2000) use sea surface temperature (SST) in the North Atlantic as one of the predictors for forecasting hurricane activity in the Atlantic Ocean and the Caribbean region. Figure 4.2 reproduced from Gray et al. (2000) shows the annual average North Atlantic SST anomalies over the period 1900–2000 for the region 50-60°N, 10-50°W. Several shifts in the mean of the SST time series are evident in the figure. According to Gray et al., periods of positive anomalies are related to more active hurricane seasons than normal, while periods

of negative anomalies are related to less active hurricane season than normal. The foregoing examples show some empirical evidence that time series have alternating periods of high and low values, where the shift from high to low and vice versa seems to occur abruptly. Although less evident (because of lack of longer data) one could also argue that the duration of the periods of highs and lows are random or at least they are uncertain quantities. The following sections of the paper refer to modeling of this type of processes.

4.3 Shifting Mean Processes

The objective here is to formulate a model that can be used to mimic the behavior of processes that shift randomly from one stationary state to another around their long term mean. The processes considered here are strictly stationary, even though the sudden shifts in the means can be thought of as non-stationary behavior. Boes and Salas (1978) introduced a general shifting mean model, which they used for studying the Hurst phenomenon. Their studies were further developed and explored by Salas and Boes (1980); Obeysekera (1981); Leiva (1983); Ballerini and Boes (1985); Boes (1988); and Saada (1998). The formulation of the shifting mean processes proposed in this paper shares many similarities with the formulation in Boes and Salas (1978).

A general definition of the shifting mean (SM) model proposed in this paper is given by

$$X_t = Y_t + Z_t \tag{4.1}$$

where $\{X_t\}$ is a sequence of random variables representing the climatic or hydrologic process of interest, $\{Y_t\}$ is a sequence of independent and identically distributed (*iid*) variables with mean μ_Y and variance σ_Y^2, and $\{Z_t\}$ is a sequence with mean zero and variance σ_Z^2. The Z_t's represent noise in the mean of the process X_t. This noise is characterized by levels in the sense that $Z_1 = \cdots = Z_{N_1}$, $Z_{N_1+1} = \cdots = Z_{N_1+N_2}$, ..., where N_i is the span or the length of the noise level i for $i = 1, 2, \ldots$. In other words N_i can also be considered as the length

of the stationary state i of the process X_t. This is where the name "shifting mean" comes from, that is at each shift-epoch, $t \in \{1 + N_1, 1 + N_1 + N_2, \ldots\}$, the mean of the process X_t can be thought of as being shifted from one state to another one. In this paper it will be assumed that $N_1, N_2, \ldots$ is a discrete, stationary, delayed-renewal sequence on the positive integers (e.g. see Ballerini and Boes, 1985; Boes, 1988). This implies that $N_2, N_3, \ldots$ are *iid* variables, say with a cdf $F_N(n)$, and N_1 is independent of $\{N_i\}_{i=2}^{\infty}$ with a cdf

$$F_{N_1}(n) = \frac{\sum_{j=0}^{n-1}(1 - F_N(j)}{\sum_{j=0}^{\infty}(1 - F_N(j))} \quad , n = 1, 2, \ldots \tag{4.2}$$

which exists for $E[N_2] < \infty$. For clarification, a delayed-renewal process will arise when the first event in Eq (4.1), X_1, occurs in-between shift-epochs of an ordinary renewal process (that is, X_1 does not mark a beginning of a new stationary state). A schematic representation of a SM model is given in Figure 4.3.

Two types of SM models will be considered based on different treatment of the Z_t's. In the first SM model referred to as SM-1, the sign of the noise levels is random. Thus, $Z_t = M_i$ for $1 + \sum_{j=1}^{i-1} N_j \leq t \leq \sum_{j=1}^{i} N_j$, where M_i is a real valued zero mean random variable (can take both positive and negative values). The SM-1 model is structurally the same as the shifting mean model introduced by Boes and Salas (1978). In the second SM model, denoted by SM-2, $Z_t = Q_i M_i$ for $1 + \sum_{j=1}^{i-1} N_j \leq t \leq \sum_{j=1}^{i} N_j$, where $N_0 = 0$. Here $\{Q_i\}$ is a sequence of variables with values 1 and -1 representing the signs of the noise levels Z_t, and $\{M_i\}$ is a sequence of *iid* positive real valued random variables representing the magnitude of the noise level Z_t. In the SM-2 model two consecutive noise levels $Q_i M_i$ and $Q_{i+1} M_{i+1}$ will always have opposite signs (i.e. $Q_{i+1} = -Q_i$). An example is the process shown in Fig. 4.1 (b).

4.3.1 The SM-1 Model

The SM-1 model represented here is essentially the same as the shifting level model introduced by Boes and Salas (1978) except that in their case the M_i's were modeled by a

nonzero mean process and the Y_t's were modeled by a zero mean process. The opposite is done here. The detailed SM-1 model briefly introduced in Eq (4.1) is given by

$$X_t = Y_t + \sum_{i=1}^{t} M_i I_{(S_{i-1},S_i]}(t) \tag{4.3}$$

where $S_i = N_1 + N_2 + \ldots + N_i$ with $S_0 = 0$, and $I_{(a,b)}(t)$ is the indicator function equal to one if $t \in (a,b)$ and zero otherwise. Furthermore, the noise levels $M_1, M_2, \ldots$ are zero mean random variables (in this paper assumed to be *iid* normal, refer to section 4.3.4), and $N_1, N_2, \ldots$ are positive geometric random variables with parameter p (refer to section 4.3.2). The sequences $\{N_i\}$, $\{M_k\}$, and $\{Y_t\}$ are assumed to be mutually independent. The mean and the variance of the X_t process are

$$E[X_t] = \mu_Y \tag{4.4}$$

$$\text{and} \quad Var(X_t) = \sigma_Y^2 + \sigma_M^2 \tag{4.5}$$

and the autocorrelation function (acf) is

$$\rho_X(h) = \frac{\sigma_M^2(1-p)^h}{\sigma_Y^2 + \sigma_M^2} \quad , h = 1, 2, \ldots \tag{4.6}$$

For the SM-1 model four parameters $\{\mu_Y, \sigma_Y, \sigma_M, p\}$ need to be estimated. To estimate the parameters using the method of moments, the estimated mean and variance of X_t and estimates of $\rho_X(h)$ at up to two different lags greater than zero are needed. The parameter estimates in terms of $\hat{\mu}_X, \hat{\sigma}_X$ and $\hat{\rho}_X(h)$ for $h = 1$ and 2 are

$$\hat{p} = 1 - \frac{\hat{\rho}_X(2)}{\hat{\rho}_X(1)}, \tag{4.7}$$

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 \frac{\hat{\rho}_X^2(1)}{\hat{\rho}_X(2)}, \tag{4.8}$$

$$\hat{\mu}_Y = \hat{\mu}_X, \tag{4.9}$$

$$\text{and} \quad \hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\sigma}_M^2 \tag{4.10}$$

The parameters estimates are feasible if $\hat{\rho}_X(1) > \hat{\rho}_X(2) > \hat{\rho}_X^2(1)$ as illustrated in Fig. 4.4. Because of sample variability of the sample correlogram, infeasible parameter estimates may

result. In such cases the sample correlogram can be fitted by $\hat{\rho}_X(h) = ab^h$ which has the same form as the model correlogram in Eq (4.6) for $0 < b < 1$ and $0 < a < 1$.

4.3.2 The SM-2 Model

Let $\{Q_i\}_{i=1}^{\infty}$ be a simple Markov chain with state space $\{-1, 1\}$ representing the sign of the shift as compared to the long term mean (μ_X) of the process X_t. The transition probability matrix of the Markov chain is

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{4.11}$$

where $P(Q_i = -1 | Q_{i-1} = -1) = 0$, $P(Q_i = 1 | Q_{i-1} = -1) = 1$, etc. The unconditional probabilities are $P(Q_i = -1) = P(Q_i = 1) = 1/2$. Obviously $E[Q] = 0$, $Var(Q) = 1$, and $Skew(Q) = 0$. Let $\{M_i\}_{i=1}^{\infty}$ be a sequence of *iid* positive real variables with mean μ_M and variance σ_M^2. Then the noise level sequence Z_t can be written as

$$\begin{aligned} Z_t &= \begin{cases} Q_1 M_1 & \text{if} \quad t \leq N_1 \\ Q_2 M_2 & \text{if} \quad N_1 < t \leq N_1 + N_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Q_t M_t & \text{if} \quad S_{t-1} < t \leq S_t \end{cases} \\ &= \sum_{i=1}^{t} Q_i M_i I_{(S_{i-1}, S_i]}(t) \end{aligned} \tag{4.12}$$

where as in section 4.3.1, $S_i = N_1 + N_2 + \ldots + N_i$ with $S_0 = 0$.

Let us assume that the sequences $\{N_i\}$, $\{Q_j\}$, $\{M_k\}$, and $\{Y_l\}$ are mutually independent. Then unconditionally the process Z_t is stationary in the mean and the variance with first three raw moments $E[Z_t] = 0$, $E[Z_t^2] = \sigma_M^2 + \mu_M^2$, and $E[Z_t^3] = 0$. It follows that the mean and the variance of X_t in Eq (4.1) are

$$E[X_t] = \mu_Y \tag{4.13}$$

and

$$Var(X_t) = Var(Y_t + Z_t) = \sigma_Y^2 + \sigma_M^2 + \mu_M^2 \tag{4.14}$$

Since the Y_t's are *iid*, the lag-h autocovariance function of X_t, $h = 1, 2, \ldots$, is given by

$$Cov(X_t, X_{t+h}) = Cov(Y_t + Z_t, Y_{t+h} + Z_{t+h}) = Cov(Z_t, Z_{t+h}) = E[Z_t Z_{t+h}] \tag{4.15}$$

Substituting Z_t from Eq (4.12) into (4.15) gives

$$\begin{aligned}
Cov(X_t, X_{t+h}) &= E\left[\sum_{i=1}^{t} Q_i M_i I_{(S_{i-1}, S_i]}(t) \sum_{j=1}^{t+h} Q_j M_j I_{(S_{j-1}, S_j]}(t+h)\right] \\
&= \sum_{i=1}^{t}\sum_{j=1}^{t+h} E[Q_i M_i I_{(S_{i-1}, S_i]}(t) Q_j M_j I_{(S_{j-1}, S_j]}(t+h)] \\
&= \sum_{i=j} E[E[Q_i^2 M_i^2 I_{(S_{i-1}, S_i]}(t) I_{(S_{i-1}, S_i]}(t+h) | S_1, S_2, \ldots]] \\
&\quad + \sum\sum_{i \neq j} E[E[Q_i M_i I_{(S_{i-1}, S_i]}(t) Q_j M_j I_{(S_{j-1}, S_j]}(t+h) | S_1, S_2, \ldots]] \\
&= \sum_{i=1}^{t} E[M^2] P(S_{i-1} < t, S_i \geq t+h) \\
&\quad + \sum_{i=1}^{t}\sum_{j=i+1}^{i+h} E^2[M]\, E[E[Q_i Q_j | Q_1]]\, P(S_{i-1} < t \leq S_i, S_{j-1} < t+h \leq S_j)
\end{aligned}$$

and finally after simplification

$$\begin{aligned}
Cov(X_t, X_{t+h}) =& (\sigma_M^2 + \mu_M^2) \sum_{i=1}^{t} P(S_{i-1} < t, S_i \geq t+h) \\
&+ \mu_M^2 \sum_{i=1}^{t}\sum_{j=i+1}^{i+h} (-1)^{i+j} P(S_{i-1} < t \leq S_i, S_{j-1} < t+h \leq S_j)
\end{aligned} \tag{4.16}$$

Note that in general the autocovariance function of X_t in Eq (4.16) is not stationary, i.e. in general $Cov(X_i, X_{i+h}) \neq Cov(X_j, X_{j+h})$ for $i \neq j$ and $i, j, h \in \{1, 2, \ldots\}$.

Boes and Salas (1978) assumed that $\{N_i\}_{i=1}^{\infty}$ is positive geometric distributed (refer to Eqs (4.17) and (4.18) below). The geometric distribution has a similar shape as the exponential, that is, its mode is at its lowest value and the probability mass function (pmf) falls monotonically towards 0 at infinity. Thus the geometric distribution is useful to model the length of the stationary time spans of processes that shift fairly rapidly from one stationary state to another. More interestingly, if $\{N_i\}_{i=1}^{\infty}$ is assumed to be a stationary, delayed-renewal sequence with $N_2, N_3, \ldots \overset{iid}{\sim}$ posgeom(p) then it may be shown using Eqs (4.2) and

(4.18) that also N_1 is posgeom(p). Random variable N has the positive geometric distribution with parameter p, denoted as $N \sim$ posgeom(p), if the pmf and the cdf of N are given by

$$f_N(n) = P[N = n] = p(1-p)^{n-1} I_{\{1,2,\ldots\}}(n) \tag{4.17}$$

and

$$F_N(x) = \sum_{n=1}^{\infty} [1-(1-p)^n] I_{[n,n+1)}(x) \tag{4.18}$$

respectively, where $0 < p < 1$. The mean and the variance of N are

$$E[N] = 1/p \quad \text{and} \quad Var(N) = \frac{1-p}{p^2} \tag{4.19}$$

The sum of positive geometric random variables is negative binomial distributed. Thus if $N_1, N_2, \ldots \overset{iid}{\sim}$ posgeom(p), then the pmf of $S_j = N_1 + \cdots + N_j$ is given by

$$P(S_j = s) = \binom{s-1}{j-1} p^j (1-p)^{s-j} I_{\{j,j+1,\ldots\}}(s) \quad , j = 1, 2, \ldots \tag{4.20}$$

Using Eqs (4.17) and (4.20) into Eq (4.16), the autocovariance function of X_t can be simplified to the following form

$$\begin{aligned} Cov(X_t, X_{t+h}) &= (\sigma_M^2 + \mu_M^2)(1-p)^h + \mu_M^2 \left(\sum_{j=0}^{h} (-1)^j \binom{h}{j} p^j (1-p)^{h-j} - (1-p)^h \right) \\ &= \sigma_M^2 (1-p)^h + \mu_M^2 (1-2p)^h \quad , h = 1, 2, \ldots \end{aligned} \tag{4.21}$$

which says that under the assumption that $N_1, N_2, \ldots \overset{iid}{\sim}$ posgeom(p) the resulting lag-h autocovariance function of X_t is stationary, that is independent of t. From Eqs (4.14) and (4.21) it follows that the lag-h autocorrelation function of X_t is

$$\rho_X(h) = \frac{\sigma_M^2 (1-p)^h + \mu_M^2 (1-2p)^h}{\sigma_Y^2 + \sigma_M^2 + \mu_M^2} \quad , h = 1, 2, \ldots \tag{4.22}$$

The SM-2 model has five parameters, $\{\mu_Y, \sigma_Y, \mu_M, \sigma_M, p\}$. In order to estimate the parameters using the method of moments, the estimated mean and variance of X_t and

estimates of $\rho_X(h)$ at up to three different lags greater than zero are needed. The following estimation procedure can be used to estimate the parameters in terms of $\hat{\mu}_X, \hat{\sigma}_X$ and $\hat{\rho}_X(h)$ for $h = 1, 2,$ and 3. Solve the quadratic equation

$$2\hat{\rho}_X(1)\hat{p}^2 + 3\hat{p}[\hat{\rho}_X(2) - \hat{\rho}_X(1)] + \hat{\rho}_X(3) - 2\hat{\rho}_X(2) + \hat{\rho}_X(1) = 0 \tag{4.23}$$

for $\hat{p}$. Then the estimates of μ_M and σ_M^2 are obtained from

$$\hat{\mu}_M = \hat{\sigma}_X \sqrt{\frac{(1-\hat{p})\hat{\rho}_X(1) - \hat{\rho}_X(2)}{\hat{p}(1-2\hat{p})}} \tag{4.24}$$

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 \frac{\hat{\rho}_X(3) - (1-2\hat{p})\hat{\rho}_X(2)}{\hat{p}(1-\hat{p})^2} \tag{4.25}$$

and lastly

$$\hat{\mu}_Y = \hat{\mu}_X \tag{4.26}$$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\sigma}_M^2 - \hat{\mu}_M^2 \tag{4.27}$$

The above estimates do not always exist. In fact the tight constraints on the correlogram make it hard to come up with feasible parameter estimates using the sample correlogram. To simplify the parameter estimation procedure, further assumptions can be made about the distribution of the M_i's (refer to section 4.3.3), and/or the sample autocorrelation function of X_t can be smoothed or fitted by a specific functional form to reduce effects of sample variability and other non-model characteristics (refer to section 4.4). In general feasible parameter estimates can only be obtained if $(\hat{\rho}_X(1) - \hat{\rho}_X(2))^2 + 8(\hat{\rho}_X^2(2) - \hat{\rho}_X(1)\hat{\rho}_X(3)) > 0$ and $\{\hat{\rho}_X(2)/(1-p) < \hat{\rho}_X(1) < \hat{\rho}_X(2)/(1-2p)$ for $p < 1/2$, $\hat{\rho}_X(2)/(1-2p) < \hat{\rho}_X(1) < \hat{\rho}_X(2)/(1-p)$ for $p > 1/2\}$ and $(1-2p)\hat{\rho}_X(2)) < \hat{\rho}_X(3) < (1-p)\hat{\rho}_X(2)$ and $\{(1-p)^2\hat{\rho}_X(1) < \hat{\rho}_X(3)$ for $p < 1/2$ and $p > 2/3$, $\hat{\rho}_X(3) < (1-p)^2\hat{\rho}_X(1)$ for $1/2 < p < 2/3$, $(1-2p)^2\hat{\rho}_X(1) < \hat{\rho}_X(3)$ for $p < 2/3$, and $\hat{\rho}_X(3) < (1-2p)^2\hat{\rho}_X(1)$ for $p > 2/3\}$. These are constraints due to estimation of the parameters $\{\mu_M, \sigma_M, p\}$. Additional constraints due to $\hat{\sigma}_Y^2 > 0$ in Eq (4.27) require for example that $\{(2-3p)\hat{\rho}_X(1) - \hat{\rho}_X(2) < (1-2p)(1-p)$ for $p < 1/2$, $(1-2p)(1-p) < (2-3p)\hat{\rho}_X(1) - \hat{\rho}_X(2)$ for $p > 1/2\}$. In Fig. 4.5 the range of $\rho_X(1)$ and $\rho_X(2)$ is plotted for selected values of p.

4.3.3 Simplified SM-2 Model

In the general SM-2 model, introduced in section 4.3.2, $M_1, M_2, \ldots$ are assumed to be positive *iid* variables representing the absolute value of the departure of the shifting mean process in each stationary state from its long term mean. In most cases it should be sufficient to model the M_i's by a one parameter distribution. Such a distribution could be for example the exponential distribution. The concave shape of the exponential distribution is useful for generation of noise levels that are characterized by few extreme values and a high concentration of values close to zero. A more convex or bell shaped curve (like the normal density function above the 0.5 quantile) can also be used, where the value of the probability density function in the lower tail does not vary as much as of the exponential density function. In the SM-1 model, presented in section 4.3.1, the M_i's will be assumed to be zero mean normal *iid* variables (not a necessary choice). Since one of the purposes here is to compare the SM-2 model with the SM-1 model, then we will assume for the SM-2 model that the values of the M_i's are the absolute values of zero mean normal *iid* variables. More precisely, if $W \sim N(\mu = 0, \sigma^2 = \beta^2)$, then $M = |W|$ and the probability density function of M is

$$f_M(m) = \sqrt{\frac{2}{\pi}}\beta^{-1}\exp\left(-\frac{m^2}{2\beta^2}\right)I_{[0,\infty)}(m) \tag{4.28}$$

with mean $E[M] = \sqrt{2/\pi}\,\beta$ and variance $Var(M) = (1 - 2/\pi)\beta^2$, respectively. Thus, the number of parameters of the SM-2 model reduces from five to four and the acf in Eq (4.22) simplifies to

$$\rho_X(h) = \frac{\beta^2}{\pi(\sigma_Y^2 + \beta^2)}\left[(\pi - 2)(1-p)^h + 2(1-2p)^h\right] \quad , h = 1, 2, \ldots \tag{4.29}$$

The following estimation procedure can be used to estimate the parameters $\{\mu_Y, \sigma_Y, \beta, p\}$ in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$, and $\hat{\rho}_X(h)$ for $h = 1$ and 2. The quadratic equation

$$\hat{p}^2\hat{\rho}_X(1)(\pi + 6) - \hat{p}(2\hat{\rho}_X(1) - \hat{\rho}_X(2))(\pi + 2) + (\hat{\rho}_X(1) - \hat{\rho}_X(2))\pi = 0 \tag{4.30}$$

is solved for $\hat{p}$, and then the estimates of β, μ_Y, and σ_Y^2 are obtained from

$$\hat{\beta} = \hat{\sigma}_X \sqrt{\frac{\pi \hat{\rho}_X(1)}{\pi - \hat{p}(\pi + 2)}}, \tag{4.31}$$

$$\hat{\mu}_Y = \hat{\mu}_X, \tag{4.32}$$

$$\text{and} \quad \hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\beta}^2 \tag{4.33}$$

respectively. Equation (4.30) will in some cases give two feasible estimates of $\hat{p}$, but usually only one of those estimates will yield both $\hat{\beta} > 0$ and $\hat{\sigma}_Y^2 > 0$ in Eqs (4.31) and (4.33). The parameter space for $\rho_X(1)$ in terms of $\rho_X(2)$ can be constructed from the following relation

$$\rho_X(2) = \frac{p^2(\pi + 6) - 2p(\pi + 2) + \pi}{-p(\pi + 2) + \pi} \rho_X(1) \tag{4.34}$$

where $\{0 < \rho_X(1) < 1 - p(1 + 2/\pi)$ for $0 < p < \pi/(2 + \pi)\}$, and $\{1 - p(1 + 2/\pi) < \rho_X(1) < 0$ for $\pi/(2 + \pi) < p < 1\}$. The range of $\rho_X(1)$ and $\rho_X(2)$ is plotted in Fig. 4.6. Due to sample variability or other factors it is possible that the sample acf falls outside of the parameter space in Fig. 4.6. In such cases the sample acf can be smoothed or fitted as is done in the examples in section 4.4.

4.3.4 Choice of Distributions to Model the Y_t's and the M_i's

The procedures for parameter (or moment) estimation of the SM-1 and SM-2 models presented in previous sections are independent of the choice or type of distributions to model the Y_t's and the M_i's of the referred models. In general it has been assumed that the Y_t's follow a distribution with two unknown parameters and that the M_i's follow a distribution with one unknown parameter. The fourth parameter is the parameter p of the geometric distribution. We will assume that the X_t process has zero skewness ($\gamma_X = 0$) and that $Y_1, Y_2, \ldots \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ for both the SM-1 and SM-2 models. Furthermore, for the SM-1 model it is assumed that the noise levels $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$. The parameters of the SM-1 model are estimated using Eqs (4.7)–(4.10) and the parameters of the simplified SM-2 model are estimated using Eq (4.30)-(4.33).

On the other hand, if the X_t process has non-zero skewness and one would like preserve it, then skewed distributions can be used to model the Y_t or/and the M_i process of the SM-1 model, and the Y_t process of the SM-2 model. Procedures how to incorporate skewed distributions in the modeling process will be discussed elsewhere. In order to preserve the skewness of X_t either two or three parameter skewed distributions can be used to model the components of the SM-1 and SM-2 models.

4.3.5 Properties of the Autocorrelation Functions of the SM-1 and SM-2 Models

As stated in Boes and Salas (1978) the acf of the SM-1 model in Eq (4.6) has the same form as the acf of an ARMA(1, 1) process, i.e. $\rho_X(h) = \phi^{h-1}\rho_X(1)$ for $h = 1, 2, \ldots$. For the SM-1 model the acf is always positive and falls exponentially towards zero. The acf of the SM-2 model in Eq (4.29) behaves similarly for $p < \pi/(\pi + 2)$ but can take negative values if h is odd and p is relatively large. In general $\rho_X(h) < 0$ in Eq (4.29) if and only if $p > (2^{1/h} + (\pi - 2)^{1/h})/(2 \cdot 2^{1/h} + (\pi - 2)^{1/h})$ and h is odd. Futhermore, if $p > 2/3$ then $\rho_X(h) < 0$ for all odd h's. For illustration and comparison, the autocorrelation functions of Eqs (4.6) and (4.29) are plotted in Fig. 4.7 for $p \in \{0.02, 0.2, 0.9\}$. For the SM-1 process $\sigma_Y^2 = \sigma_M^2$ in Eq (4.6), and for the SM-2 process $\sigma_Y^2 = \beta^2$ in Eq (4.29). Thus for the particular case shown in the figure, the random variables M_i's of the SM-2 process are equivalent in distribution to the absolute random variables M_i's of the SM-1 process.

Obviously the choice of using the geometric distribution to model the length of the random time spans (N_i's) may not be appropriate in all cases. For example if a given time series shows signs of periodic behavior, then that periodic behavior should be reflected in the sample correlogram. The use of the geometric distribution will result in a fitted model with an acf that has no signs of periodicities. In such situation a different distribution that could reproduce such periodic characteristics would be more suitable for modeling the lengths of the random time spans. Such a distribution could be for example the binomial,

the Poisson, the discrete triangular, or the discrete uniform distribution. On the other hand, using a different distribution than the geometric would in most cases complicate parameter estimation if all the processes are assumed to be coupled together as is done here. For simplification and for the purpose of this study we will stick with the geometric distribution, but we do intend to study the effects of choosing different discrete distributions for modeling the lengths of the random time spans, and the possibility of uncoupling of the processes that make up the SM models to simplify parameter estimation. The results of such a study will be represented elsewhere.

4.4 Examples

In this section the use of the SM-1 and SM-2 models for modeling of climatic and hydrologic time series will be demonstrated under different scenarios. These scenarios can be simple generation of time series of the same length as the historical record, or frequency analysis of extreme events such as drought lengths using simulation experiments.

The parameters estimates of the SM-1 and the SM-2 model are not always feasible. The sample variability of the correlogram can result in estimated parameters that are outside the parameter space. To reduce effects of sample variability and periodic behavior on the parameter estimates, the sample correlograms will in most cases be fitted by an exponentially decaying function of the form $\rho_X(h) = ab^h$. Such fitted correlogram may not completely capture the sample correlogram, but on the other hand the correlogram of the fitted SM model will closely resemble the fitted sample correlogram.

4.4.1 The PDO Data

The acf of the annual PDO index (see Fig. 4.1) up to lag 15 is shown in Fig. 4.8 along with approximate 95% confidence bounds ($\pm 1.96/\sqrt{n}$) for an *iid* sequence. The sample correlogram for lags 1-13 has been fitted by $\hat{\rho}_X(h) = ab^h$ using the method of least squares. The fitted correlogram has somewhat different shape than the sample correlogram, where

the sample correlogram seems to indicate a periodic behavior with a period around five to six years. The SM-1 and SM-2 models are fitted assuming that the Y_t's of the SM-1 and the SM-2 models are normally distributed and that the M_i's in the SM-1 process are normally distributed (refer to section 4.3.4). The estimated parameters for both models in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$ and the lag 1 and 2 acf from the fitted correlogram are: for the SM-1 model, $\{\hat{p} = 0.2703,\ \hat{\sigma}^2_M = 0.5371,\ \hat{\mu}_Y = 0.04769,\ \hat{\sigma}^2_Y = 0.07000\}$, and for the SM-2 model, $\{\hat{p} = 0.1709,\ \hat{\beta} = 0.7376,\ \hat{\mu}_Y = 0.04769,\ \hat{\sigma}^2_Y = 0.06300\}$.

In order to see if some of the shifting behavior of the PDO data in Fig. 4.1 can be captured by the models, PDO samples of the same length as the historical record length were simulated. The generated sequences are plotted in Fig. 4.9 (a) based on the assumed SM-1 model, and in Fig. 4.9 (b) based on the assumed SM-2 model. From the figure it can be concluded that the generated sequences do show somewhat similar shifting behavior as the sample data in Fig. 4.1 does. Furthermore, 1,000 realizations of the PDO index of the same length as the historical record (n) were generated and the average mean, variance, and skewness were computed based on these 1,000 realizations. The SM-1 model gave $\{\hat{\mu}_X, \hat{\sigma}^2_X, \hat{\gamma}_X\} = \{0.0555, 0.5875, -0.0022\}$, and the SM-2 model gave $\{\hat{\mu}_X, \hat{\sigma}^2_X, \hat{\gamma}_X\} = \{0.0440, 0.5726, 0.0098\}$. For comparison the respective statistics of the historical sample are $\{0.0477, 0.6071, 0.0793\}$. Thus in general it can be concluded that the SM-1 and SM-2 models preserve the mean and the variance quite well. Note that since the sample skewness is near zero no attempt was made to preserve it, i.e. the skewness of the fitted SM-1 model and the fitted SM-2 model is zero.

Correlograms based on the fitted SM-1 and SM-2 models are plotted in Fig. 4.10, where for both models a correlogram is estimated based on one generated sample of size $1,000n$ and based on the average acf's of 1,000 generated samples of the same size as the historical PDO record (n). The estimated correlograms based on one sample of size $1,000n$ can be considered the same as the actual model correlograms. As often is the case, when average correlograms are estimated based on a number of generated samples of small sizes,

the average correlograms (uncorrected for bias) based on 1,000 generated sequences of the same size as the historical PDO record underestimate the true model correlograms for both the SM-1 and the SM-2 models.

4.4.2 Mean Annual Flows of the Niger River at Koulikoro

In Fig. 4.11 the mean annual flows (1907-1999) for the Niger River at Koulikoro (Mali, Africa), are plotted along with its acf up to lag-15. The sample statistics of the 93-year long historical sample are $\hat{\mu}_X = 1,374\ cms$, $\hat{\sigma}_X = 398.0\ cms$, and $\hat{\gamma}_X = 0.1746$. The fitted acf at lags 1 and 2 has the values 0.7092 and 0.6224.

The estimated parameters for the SM-1 and SM-2 models in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$ and the lag 1 and 2 acf of the fitted correlogram are: for the SM-1 model, $\{\hat{p} = 0.1223, \hat{\sigma}_M = 357.8\ cms, \hat{\mu}_Y = 1,374\ cms, \hat{\sigma}_Y = 174.4\ cms\}$, and for the SM-2 model, $\{\hat{p} = 0.07566, \hat{\beta} = 358.1\ cms, \hat{\mu}_Y = 1,374\ cms, \hat{\sigma}_Y = 173.7\ cms\}$. The estimated skewness of the historical sample is not significantly different from zero so it is reasonable to assume that the Y_t's of both the SM-1 and the SM-2 models are normally distributed, and that the M_i's of the SM-1 model are normally distributed. Based on the fitted models, 1,000 realizations of the Niger River flows were generated. The mean and the variance were relatively well preserved: $\{\hat{\mu}_X, \hat{\sigma}_X, \hat{\gamma}_X\} = \{1,379\ cms, 371.7\ cms, -0.0298\}$ for the SM-1 model, and $\{\hat{\mu}_X, \hat{\sigma}_X, \hat{\gamma}_X\} = \{1,378\ cms, 369.8\ cms, -0.0063\}$ for the SM-2 model. In Fig. 4.12 estimated correlograms of both the SM-1 and SM-2 models are plotted based on one generated sample of length 1,000 n. The estimated correlograms preserve the sample correlogram quite well for both models. Furthermore using both models, 1,000 samples of the same length as the historical Niger River record were generated, and the average acf was calculated. As was the case for the PDO annual index, the average correlograms based on these 1,000 generated sequences underestimated the true model correlogram for both the SM-1 and SM-2 models.

For a given demand level, a drought duration of length L is represented by L consecutive years with flows less than a certain demand level. Simulations were used to estimate

the return period of drought durations of lengths 5 to 25 years, based on the fitted SM-1 and SM-2 models. The demand level was assumed equal to the mean annual flow. Average return periods of 2,000 occurrences of each drought length are shown in Fig. 4.13. The drought frequency curves based on the fitted SM-1 and SM-2 models are almost identical. Furthermore, note that the longest drought in the 93 year long historical Niger River record is 14 years. The return period of a 14 year drought is 96.0 years based on the fitted SM-1 model and 95.9 years based on the fitted SM-2 model.

4.5 Interpretation of Shifting Statistics

The changes in statistical behavior shown in the previous sections are examples of what is typically associated with nonlinear systems. Figure 4.14 reproduced from Kabat (2002) illustrates schematic examples of nonlinear behavior, which can be related to the examples presented in this paper. Figure 4.15 illustrates tolerance band with respect to intensity of a hazard and exposure. In Figure 4.14 (a), for instance, there is a shift in the long term mean while the (short term) variability remains about the same. Figure 4.1 provides an example of this behavior. In Fig. 4.14 (c) the long term mean changes little over time, but the band of tolerance decreases. Figure 4.14 (b) is a schematic of when the mean remains about the same but the variability increases.

If the examples presented in sections 4.2-4.4 in this paper represent nonlinear systems, their prediction into the future is inherently difficult, if not impossible. An alternate approach, reported in Kabat (2002), is to identify the vulnerability space of specified resources. The tolerance region is determined as the parameter space beyond which a significant negative impact would occur. The limit of the band of tolerance represents a vulnerability threshold. An example of a negative threshold is the occurrence of freezing conditions in the fall (that is, the end of the growing season). In general, there are multiple environmental influences that determine the band of tolerance, and when a threshold occurs. Changing statistics, as evident in Fig. 4.14, can result in greater or less probability for a threshold to

occur. The interpretation and stochastic modeling of hydroclimatic processes that exhibit shifting patterns as discusses in this paper may be helpful in examining the vulnerability space as suggested herein (or as suggested by Smith (1996) and Kabat (2002)).

4.6 Concluding Remarks

Empirical evidence show that some hydroclimatic processes exhibit shifting patterns in addition to autocorrelation. Two types of shifting mean (SM) models were proposed to analyse climatic and hydrologic processes under a probabilistic framework. The proposed shifting mean models were considered to be non-stationary in the mean, in the sense that they were allowed to shift from one stationary state to another around a long term mean. The process of interest was written as a sum of two independent random variables Y_t and Z_t, where the Y_t's were assumed to be *iid* variables and the Z_t's were assumed to represent departure of each stationary state from the long term mean of the process. That is during each stationary state the Z_t's remained fixed at a value referred to as a noise level. In the SM-1 model the noise levels were allowed to fluctuate in random manner, while in the SM-2 model two consecutive stationary states always had noise levels of opposite signs. In this paper only the positive geometric distribution was considered for modeling of the length the process spent in each stationary state. As a result the correlograms of the SM models under consideration were restricted to have certain shapes.

The applicability of the two SM models to simulate hydroclimatic time series exhibiting sudden shifts was demonstrated. In the examples both climatic and hydrologic time series were fitted by the SM models. In general the SM models were capable of preserving the mean, variance, and the autocorrelation function of the sample series. No attempt was to preserve the sample skewness. The SM models were shown to be useful for generation of long sample series and for frequency analysis of droughts. The SM models are capable of modeling time series that are correlated and have shifts in the mean. On the other hand the referred SM models do not model trends or time series that show changes in the sample

variability. However, these additional features can be incorporated by the modifying the referred SM models. For example, we are currently studying the effects of inducing a persistence into the Y_t process. In general, the proposed SM models appear to have a wide range of applicability for modeling of any type of climatic, hydrologic, and geophysical process.

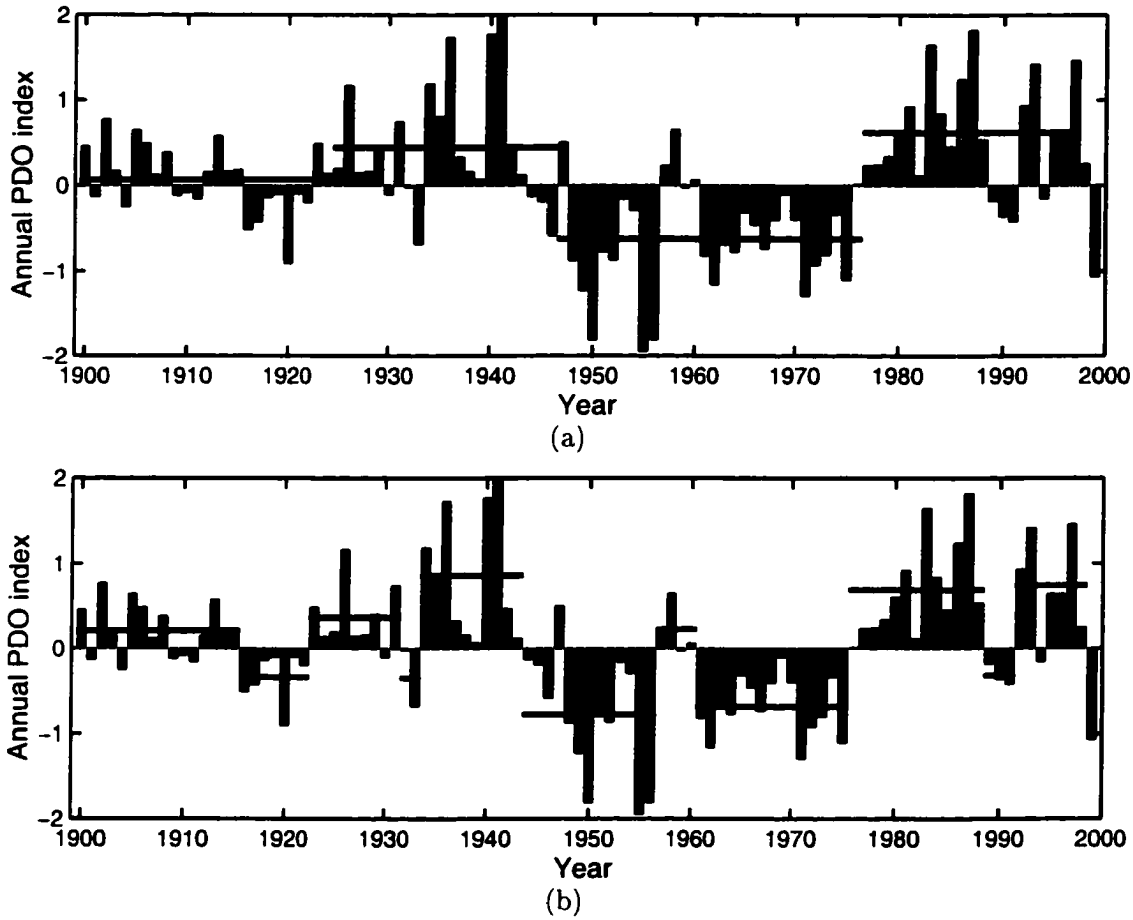

Figure 4.1: Time series of annual averages of the PDO index based on sea surface temperature, annual averages 1900-1999 from Mantua et al. (1997). (a) solid lines show averages of dominant phases of the PDO as suggested by Mantua et al. (1997). (b) solid lines show alternative shifts.

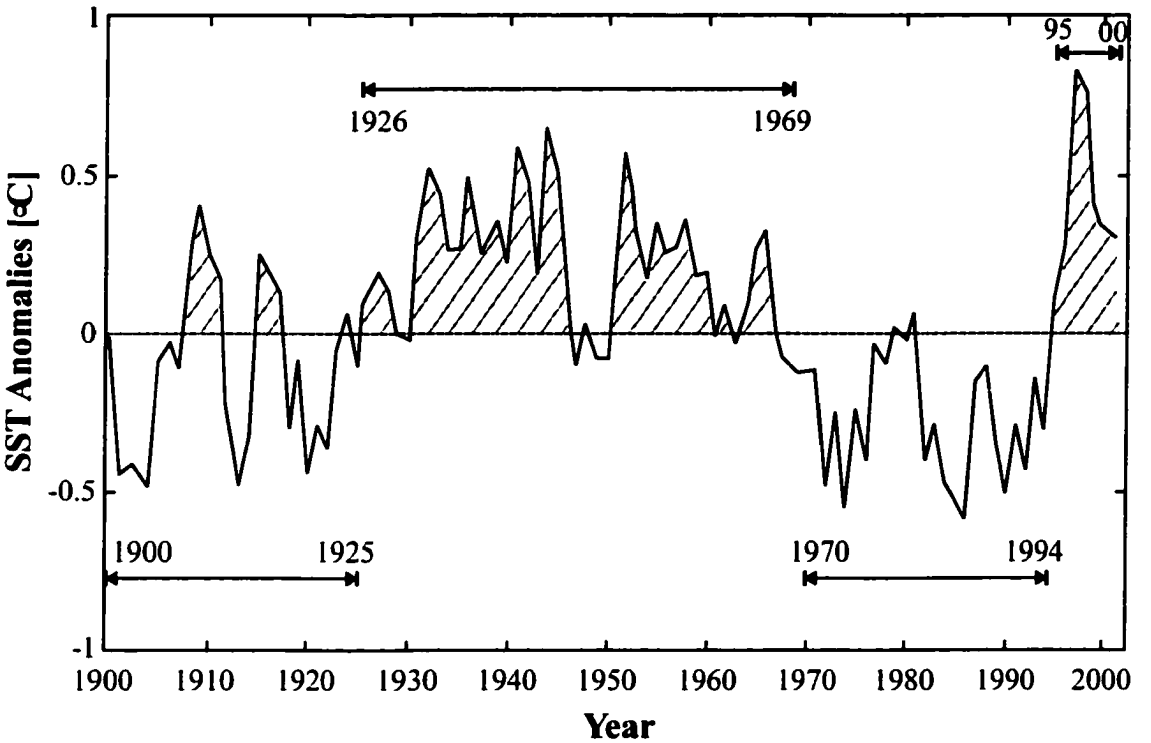

Figure 4.2: Annual average North Atlantic SST anomalies in °C in the area beween 50-60°N, 10-50°W for 1900–2000 (reproduced from Gray et al., 2000).

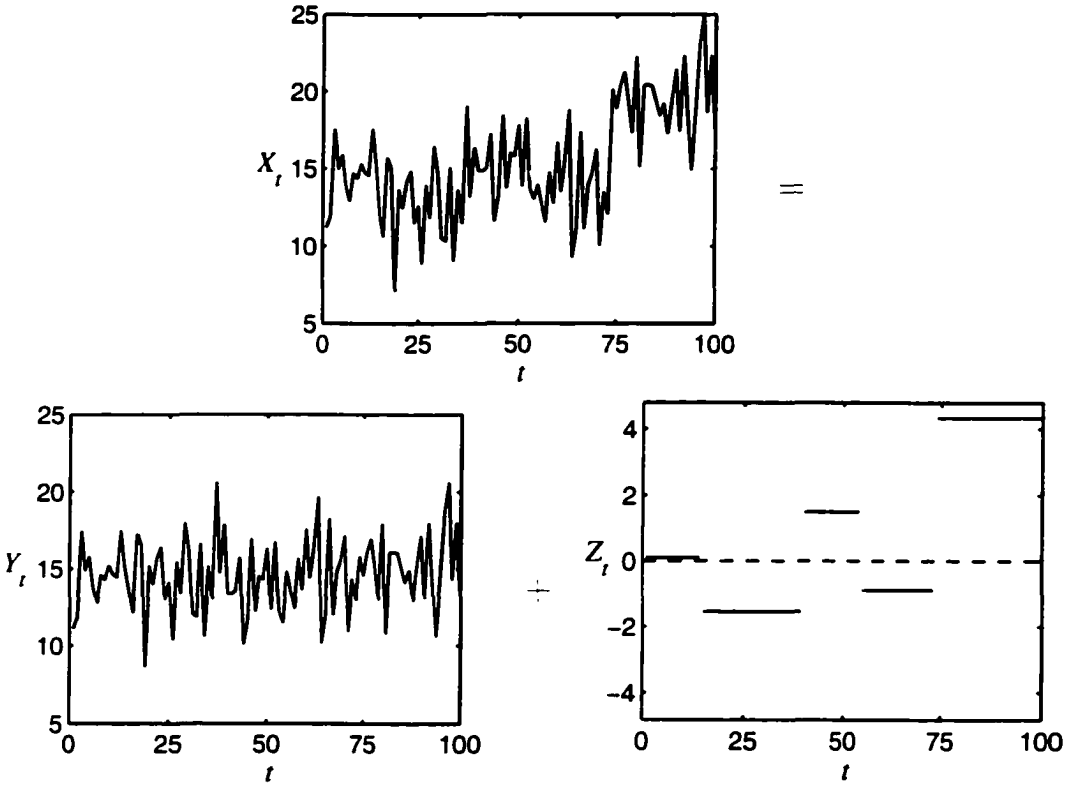

Figure 4.3: A schematic representation of the shifting mean process in (4.1).

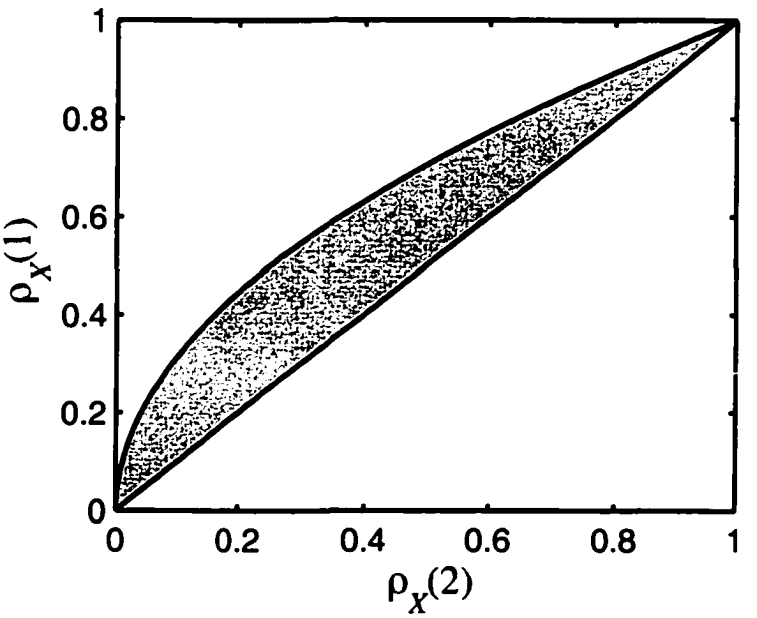

Figure 4.4: The range of $\rho_X(1)$ and $\rho_X(2)$ of the SM-1 model for the case when $N_1, N_2, \ldots \overset{iid}{\sim}$ posgeom(p).

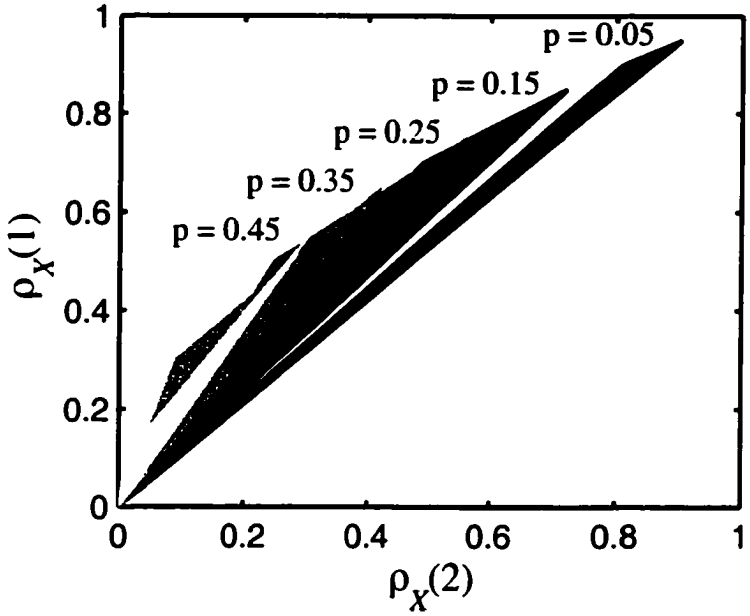

Figure 4.5: The range of $\rho_X(1)$ and $\rho_X(2)$ of the SM-2 model for selected values of p, where $N_1, N_2, \ldots \overset{iid}{\sim}$ posgeom(p).

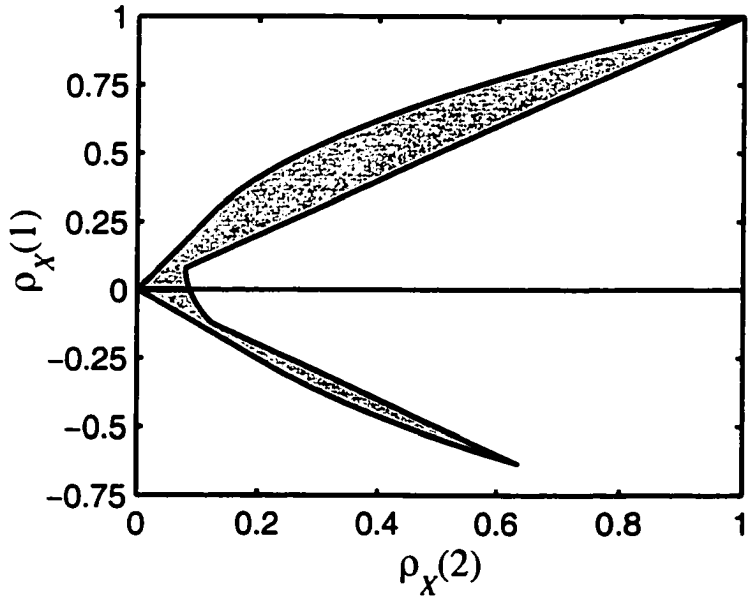

Figure 4.6: The range of $\rho_X(1)$ and $\rho_X(2)$ of the SM-2 model for the case when $N_1, N_2, \ldots \overset{iid}{\sim}$ posgeom(p) and $M_1, M_2, \ldots$ are *iid* zero mean absolute normal variables.

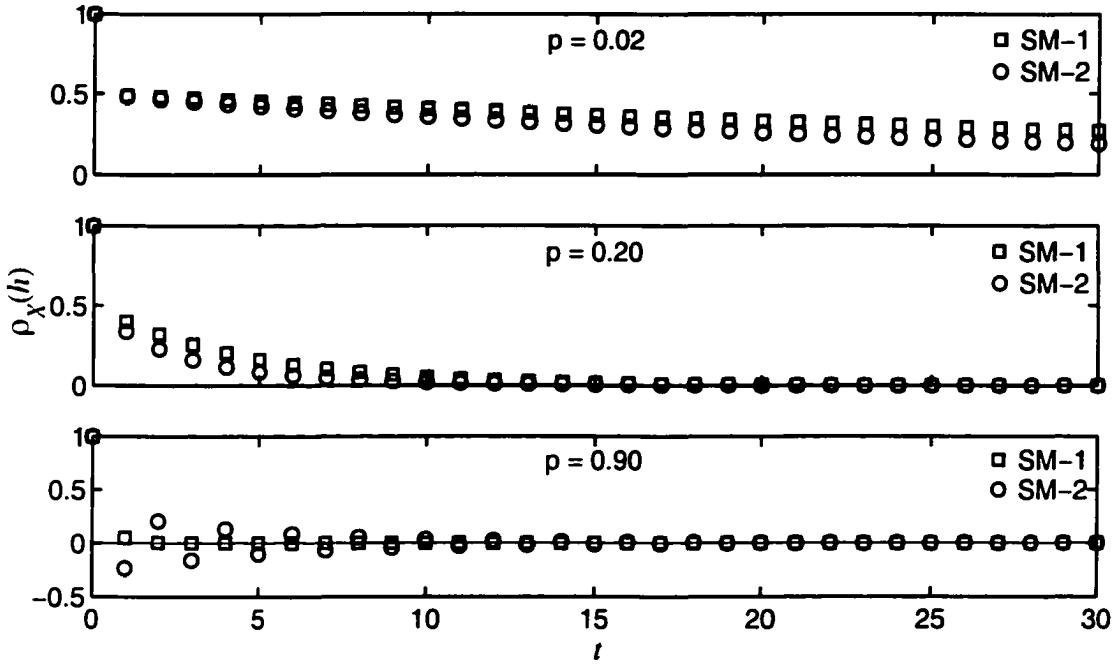

Figure 4.7: Comparison of the autocorrelation functions of SM-1 in Eq (4.6) and of SM-2 in Eq (4.29) for $\sigma_Y^2 = \sigma_M^2$ in Eq (4.6), $\sigma_Y^2 = \beta^2$ in Eq (4.29), and $p \in \{0.02, 0.2, 0.9\}$.

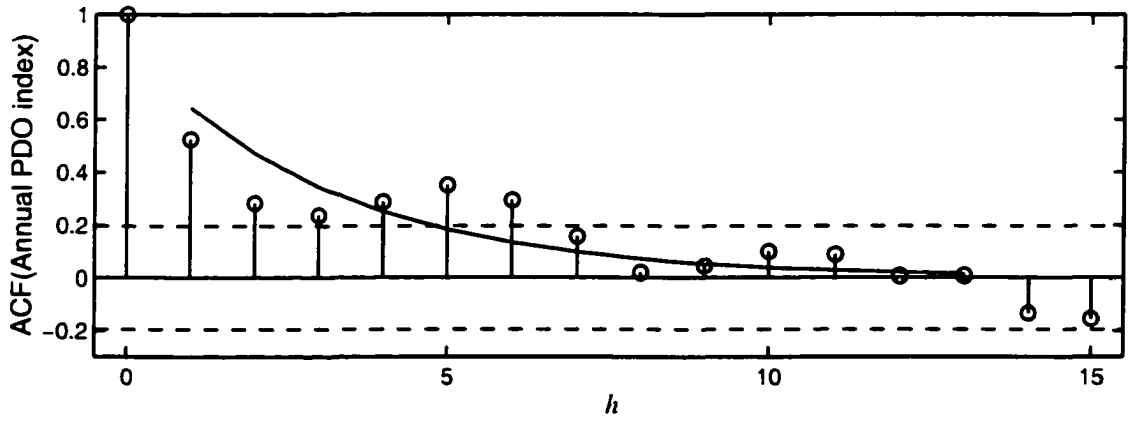

Figure 4.8: Sample autocorrelations up to lag-15 of the Pacific Decadal Oscillation in Fig. 4.1. An exponential decay function is fitted through the ACF at lags 1-13.

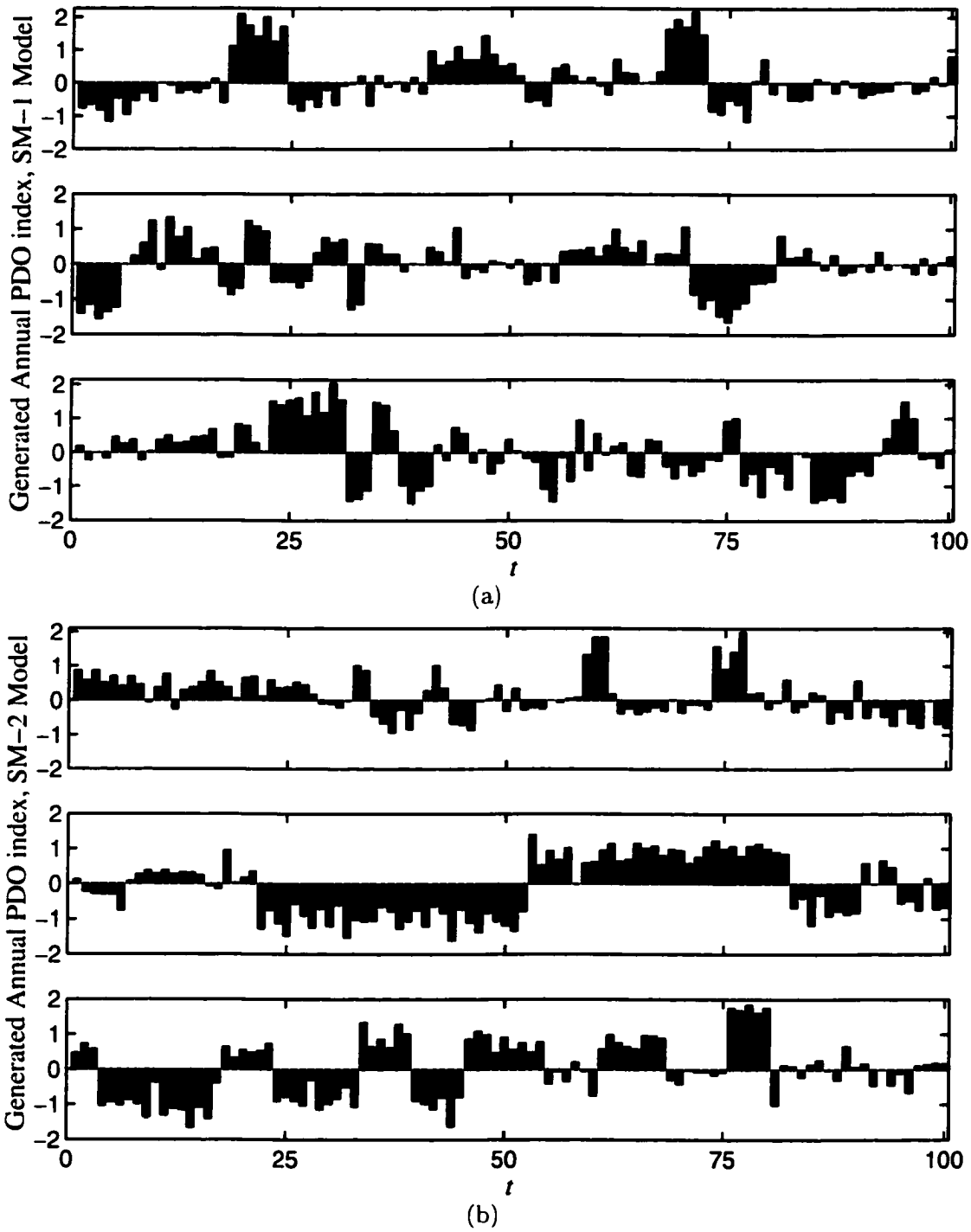

Figure 4.9: Generated sequences of the PDO annual oscillation index using the SM-1 model in (a) and the SM-2 model in (b).

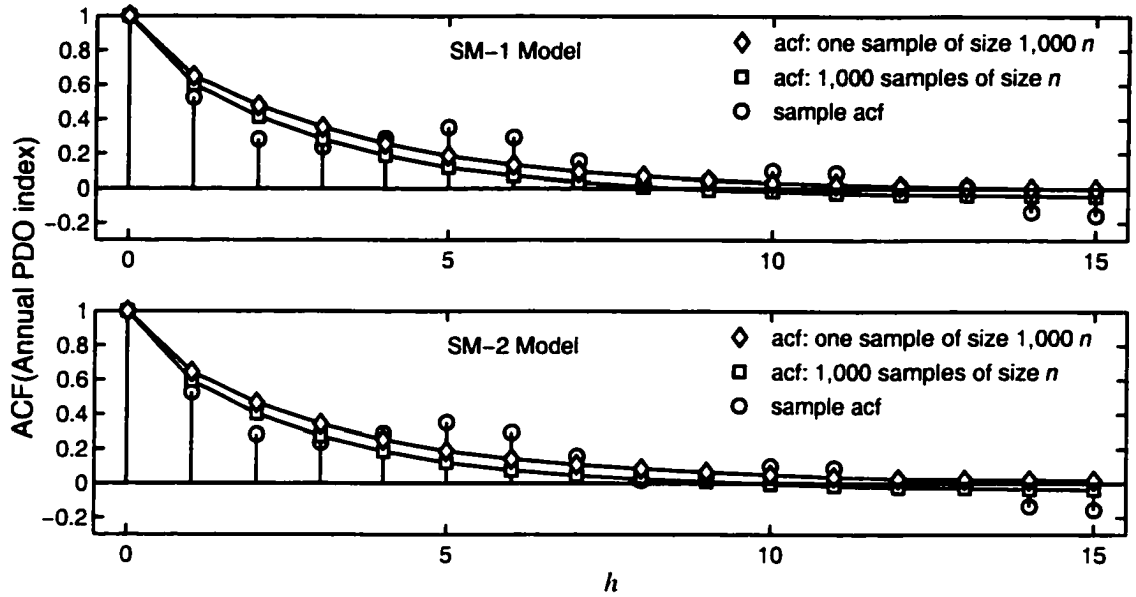

Figure 4.10: Correlograms of generated sequences using the assumed SM-1 and SM-2 models fitted to the PDO data in Fig. 4.1. For each model a correlogram is estimated based on one generated sample of size 1,000 n, and based on averaging the acf's of 1,000 generated samples of the same size as the historical record (n).

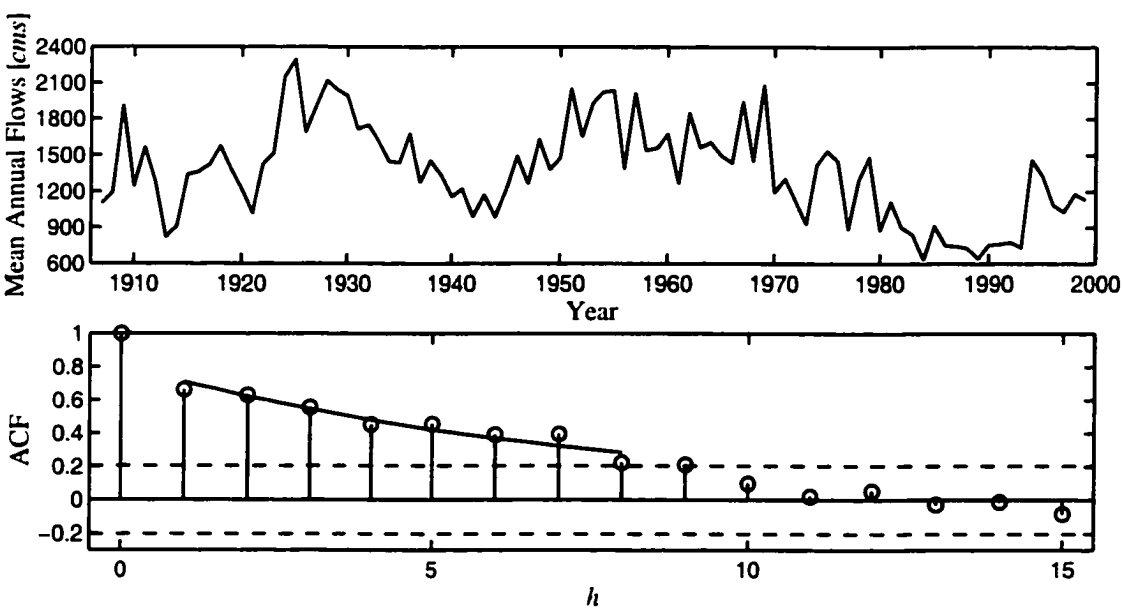

Figure 4.11: 1907–1999 annual mean flows in the Niger River at Koulikoro. The bottom plot shows the correlogram with a fitted exponential decay function at lags 1–8.

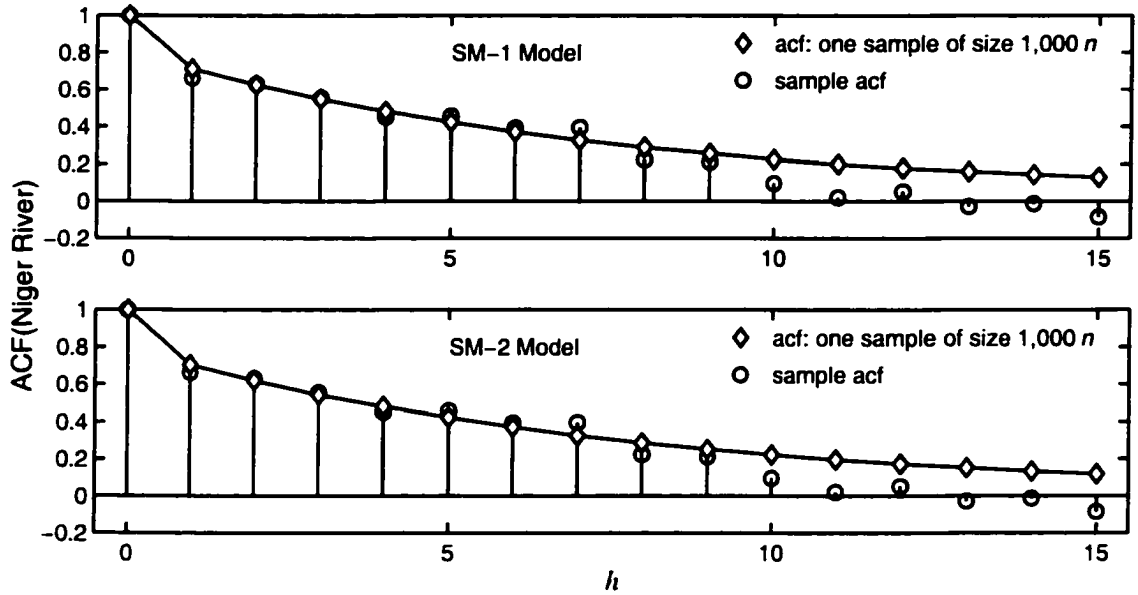

Figure 4.12: Correlograms of generated sequences using the assumed SM-1 and SM-2 models fitted to the Niger River data. For each model the correlogram is estimated based on one generated sample of length 1,000 n, where n is the length of the the historical record.

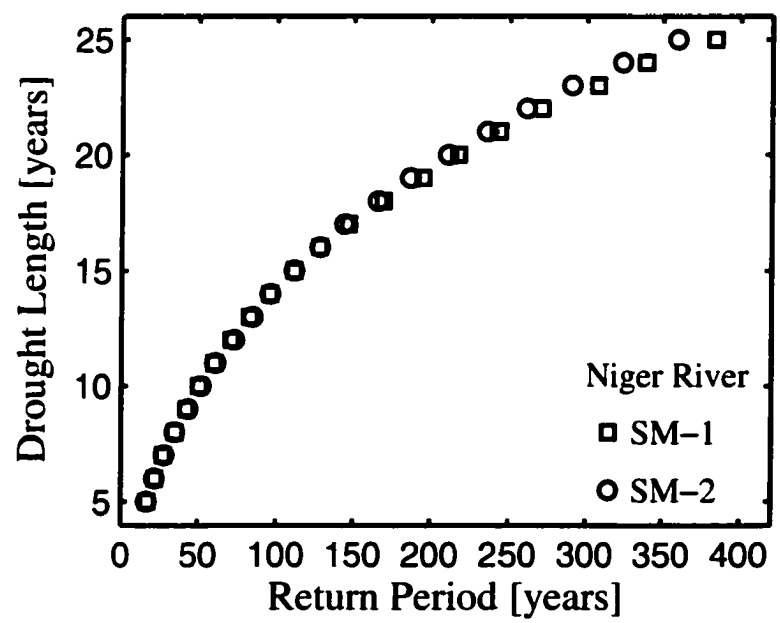

Figure 4.13: Return periods of droughts of various lengths for the Niger River at Koulikoro.

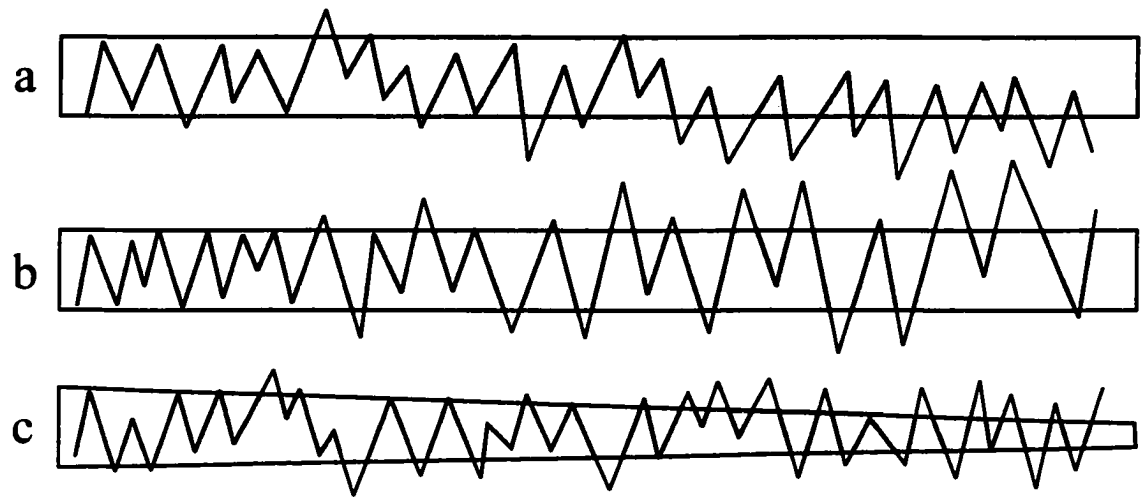

Figure 4.14: A schematic illustration in which risk changes due to variations in the physical system and the socio-economic system. In all the cases risk increases over time (after Smith, 1996; reproduced from Kabat, 2002).

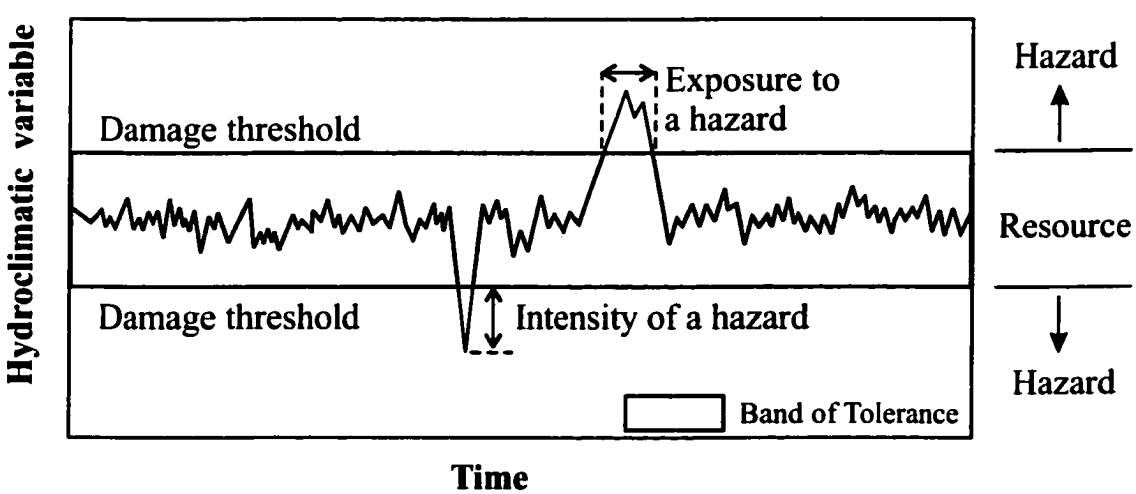

Figure 4.15: The intensity and exposure of environment hazard expressed as a function of the variability of a hydroclimatic variable within the limits of tolerance (after Smith, 1996; reproduced from Kabat, 2002).

Chapter 5

PREDICTION OF EXTREME HYDROLOGIC PROCESSES THAT EXHIBIT SUDDEN SHIFTING PATTERNS

Abstract The objective of this paper is to propose a probabilistic framework for modeling of extreme processes that are autocorrelated, such as annual extreme precipitation and floods, and droughts. The autocorrelation is assumed to arise from a certain type of non-stationarity in the mean of the process under consideration, where the process is assumed to shift abruptly from one "stationary" state to another one around a long term mean. The proposed modeling framework is based upon previously suggested shifting mean models (Sveinsson et al., 2002b), where the marginal distribution covered was the normal distribution and as a result the model skewness was zero. The main objective of this paper is to further extend the referred shifting mean models to incorporate skewed marginal distributions so that the models can be applied for frequency analysis of extreme events, drought analysis, and for generation of synthetic sample records. The proposed models utilizing skewed distributions are successfully applied as illustrated in a number of examples of extreme rainfall and flood data from several sites in North America.

5.1 Introduction

Probabilistic modeling of annual extreme hydrologic processes, such as annual maximum floods, has commonly been based on the assumption of randomness (independent and identically distributed random variables) of the process under consideration. However, in some cases the hydrologic process under consideration may be autocorrelated, not necessar-

ily because of any inherent persistence induced by the effect of storage (as can be the case for example in annual streamflow volumes) but because of sudden shifting patterns induced by large scale low frequency climatic mechanisms. Yet under assumed randomness any such autocorrelation is ignored. Stochastic models that account for autocorrelation do exist (see for example Salas, 1993). While they have been successfully applied to a number of hydrologic processes, such as monthly and annual precipitation and streamflow volumes, they are in general not suitable for modeling extreme events that are autocorrelated and show a type of non-stationary behavior. Autocorrelation in extreme hydrologic data may arise from temporal changes in the mean, variance or other statistical quantities. In this paper the focus is on processes that are characterized by sudden shifts or jumps in the mean. More precisely, the process to be analyzed is assumed to be characterized by multiple stationary states, which only differ from each other by having different means that vary around the long term mean of the process (that is, the process may be considered as being non-stationary in the mean even though its unconditional expected value at any time step is equal to the long term mean of the process). In this paper, these type of processes will be referred to as "shifting mean processes" and their corresponding models will be referred to as "shifting mean models" as in Chapter 4. The main objective here is to build on the results of Chapter 4 for modeling autocorrelated processes that shift abruptly from one stationary state to another around their long term mean. Since our purpose is to predict extreme hydrological events, the focus will be on skewed hydrologic processes. That is, we will extend the non-skewed shifting mean models studied in Chapter 4 so that they can be applicable to extreme events.

Earlier concepts on shifting mean mechanisms for simulating hydrological processes have been suggested by Hurst (1957), Klemes (1974), Potter (1976), and Boes and Salas (1978). In particular, Boes and Salas (1978) mathematically formalized certain shifting level models for the study of the Hurst phenomenon, that have been further developed and studied by Salas and Boes (1980); Ballerini and Boes (1985); Boes (1988), and in Chapter 4. The skewed shifting mean models proposed in this paper are useful for estimation

of event quantiles, and for generation of long hydrologic records. The type of data that can be typically analyzed by these models are: annual maximum floods and precipitation for extreme event studies; and mean annual flows and precipitation for drought and high flow studies. These models can also be used for forecasting, but this subject is no explored in this paper.

Shifts in hydrologic processes may be related to climate variability associated with changes in ocean currents and evaporation from the sea surface. Climatic indices such as the Southern Oscillation (SO), the Pacific Decadal Oscillation (PDO), or the North Atlantic Oscillation (NAO) appear to change quasi-periodically with time or shift from one random stationary state to another. Shifts in hydrologic time series may also be caused by man made changes, such as deforestation, urbanization, or other changes in land use. While these factors may be as important as natural mechanisms, they are not included in this paper.

5.2 The Shifting Mean Model

A general definition of the shifting mean (SM) model is given by

$$X_t = Y_t + Z_t \tag{5.1}$$

where $\{X_t\}$ is a sequence of variables representing the hydrologic process of interest. $\{Y_t\}$ is a sequence of independent and identically distributed (*iid*) variables with mean μ_Y, variance σ_Y^2, and skewness γ_Y. $\{Z_t\}$ is a sequence with mean zero ($\mu_Z = 0$), variance σ_Z^2, and skewness γ_Z. The sequences $\{Y_t\}$ and $\{Z_t\}$ are assumed to be mutually independent of each other. The X_t process is characterized by multiple "stationary" states each of random length N_i, $i = 1, 2, \ldots$. Any two states of the process X_t only differ in location, that is, the process shifts from one state to another, where the Z_ts represent the shifts (in the mean) as compared to the long term mean of the process X_t. The Z_ts are referred to as noise levels since their value remains fixed during each "stationary" state, that is $Z_1 = \cdots = Z_{N_1}$,

$Z_{N_1+1} = \cdots = Z_{N_1+N_2}$, ..., etc.

The process X_t is stationary in the mean (μ_X), the variance (σ_X^2) and the skewness (γ_X) as shown below. The relations between the first three moments of X_t, Y_t, and Z_t processes are

$$\mu_X = E[X_t] = \mu_Y + \mu_Z \tag{5.2}$$

$$\sigma_X^2 = Var(X_t) = \sigma_Y^2 + \sigma_Z^2 \tag{5.3}$$

$$\gamma_X = \frac{E[(X_t - \mu_X)^3]}{\sigma_X^3} = \frac{\gamma_Y \sigma_Y^3 + \gamma_Z \sigma_Z^3}{\sigma_X^3} \tag{5.4}$$

Since the $\{Y_t\}$ sequence is *iid* and independent of $\{Z_t\}$ and $\mu_Z = 0$ the lag h autocovariance function of X_t is given by

$$Cov(X_t, X_{t+h}) = Cov(Z_t, Z_{t+h}) = E[Z_t Z_{t+h}], \qquad h = 1, 2, \ldots \tag{5.5}$$

which is stationary if Z_t is stationary in the covariance A necessary condition for Z_t to be strictly stationary is that $\{N_i\}_{i=1}^{\infty}$ is a discrete, stationary, delayed-renewal sequence (Ballerini and Boes, 1985, Chapter 4). As in previous studies of the SM models the $\{N_i\}_{i=2}^{\infty}$ are assumed to be *iid* positive geometric variables with probability mass function (pmf) given by

$$P(N = n) = p(1-p)^{n-1} I_{\{1,2,\ldots\}}(n) \tag{5.6}$$

where $0 < p < 1$, and $I_{\{\cdot\}}(x)$ is the indicator function equal to one if $x \in \{\cdot\}$ but zero otherwise. As a result of assuming that $\{N_i\}_{i=1}^{\infty}$ is a stationary, delayed-renewal sequence, it follows that N_1 is also positive geometric distributed (Ballerini and Boes, 1985, Chapter 4), with N_1 independent of $\{N_i\}_{i=2}^{\infty}$. The mean and the variance of N are $\mu_N = 1/p$ and $\sigma_N^2 = (1-p)/p^2$, respectively. Furthermore $S_j = N_1 + \cdots + N_j$ is negative binomial distributed with pmf

$$P(S_j = s) = \binom{s-1}{j-1} p^j (1-p)^{s-j} I_{\{j,j+1,\ldots\}}(s) \tag{5.7}$$

Two different types of SM models (SM-1 and SM-2) as in Chapter 4 are considered here. They differ in the treatment of the Z_ts. In the SM-1 model the magnitude of the

noise levels is random, while in the SM-2 model the process X_t shifts from one state to another in a systematic manner so that the magnitudes of any two consecutive noise levels (or shifts) have opposite signs. For ease of reference, the SM-1 and SM-2 models are briefly summarized from Chapter 4. Then some modifications in the treatment of the skewness are introduced so as to make the SM models applicable to the modeling of extreme events.

5.2.1 The SM-1 Model

The SM-1 model is structurally the same as the SM model proposed by Boes and Salas (1978), except that we will consider the underlying processes to be skewed. In the SM-1 model it is assumed that

$$Z_t = \begin{cases} M_1 & \text{if} \quad t \leq N_1 \\ M_2 & \text{if} \quad N_1 < t \leq N_1 + N_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ M_t & \text{if} \quad S_{t-1} < t \leq S_t \end{cases} \tag{5.8}$$

$$= \sum_{i=1}^{t} M_i I_{(S_{i-1}, S_i]}(t)$$

where $S_i = N_1 + N_2 + \ldots + N_i$ with $S_0 = 0$, and $\{M_i\}$ is a sequence of *iid* real valued random variables with mean zero ($\mu_M = 0$), variance σ_M^2, and skewness γ_M. It follows that the mean, the variance, and the skewness of Z_t are

$$\mu_Z = 0 \tag{5.9}$$

$$\sigma_Z^2 = \sigma_M^2 \tag{5.10}$$

$$\gamma_Z = \gamma_M \tag{5.11}$$

Furthermore, for $N_1, N_2, \ldots \overset{iid}{\sim}$ posgeom(p) the autocovariance function of Z_t becomes

$$Cov(Z_t, Z_{t+h}) = \sigma_M^2 (1-p)^h, \qquad h = 1, 2, \ldots \tag{5.12}$$

and hence the autocorrelation function of X_t for the SM-1 model is

$$\rho_X(h) = \frac{\sigma_M^2 (1-p)^h}{\sigma_Y^2 + \sigma_M^2}, \qquad h = 1, 2, \ldots \tag{5.13}$$

There are four parameters in the SM-1 model $\{\mu_Y, \sigma_M, \sigma_Y, p\}$. The parameter estimates in terms of $\hat{\mu}_X, \hat{\sigma}_X$ and $\hat{\rho}_X(h)$ for $h = 1$ and 2 are

$$\hat{\mu}_Y = \hat{\mu}_X \tag{5.14}$$

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 \frac{\hat{\rho}_X^2(1)}{\hat{\rho}_X(2)} \tag{5.15}$$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\sigma}_M^2 \tag{5.16}$$

$$\hat{p} = 1 - \frac{\hat{\rho}_X(2)}{\hat{\rho}_X(1)} \tag{5.17}$$

The parameters estimates are feasible if $\hat{\rho}_X(1) > \hat{\rho}_X(2) > \hat{\rho}_X^2(1)$.

If the observed sample series has zero skewness ($\hat{\gamma}_X = 0$), then one may assume that $Y_1, Y_2, \ldots \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ and $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$ as in Chapter 4. On the other hand if the sample series is skewed ($\hat{\gamma}_X \neq 0$), then the sample skewness can be reproduced by modeling either the Y_ts or the M_is or both by two or three-parameter skewed distributions. In case of two-parameter distribution $\hat{\gamma}_X$ is not needed for parameter estimation since in that case the skewness of the distribution is either fixed (as for the Gumbel) or a function of the mean and the variance (as for the lognormal-2). For three-parameter distribution the skewness of the distribution is estimated using Eq (5.4). In general there are four options available for modeling the skewness of the process under consideration:

(1) Model both $\{M_i\}$ and $\{Y_t\}$ by two-parameter skewed distributions. Hence, $\hat{\gamma}_Y = f(\hat{\mu}_Y, \hat{\sigma}_Y)$ and $\hat{\gamma}_M = g(\hat{\mu}_M, \hat{\sigma}_M)$.

(2) Assume that $\gamma_M = 0$ or model $\{M_i\}$ by a two-parameter skewed distribution (that is $\hat{\gamma}_M = f(\hat{\mu}_M, \hat{\sigma}_M)$). Then use a three-parameter distribution for Y_t, where γ_Y is estimated by

$$\hat{\gamma}_Y = \frac{\hat{\gamma}_X \hat{\sigma}_X^3 - \hat{\gamma}_M \hat{\sigma}_M^3}{\hat{\sigma}_Y^3} \tag{5.18}$$

(3) Assume that $\gamma_Y = 0$ or model $\{Y_t\}$ by a two-parameter skewed distribution (that is $\hat{\gamma}_Y = f(\hat{\mu}_Y, \hat{\sigma}_Y)$). Then use a three-parameter distribution for M_i, where γ_M is

estimated by

$$\hat{\gamma}_M = \frac{\hat{\gamma}_X \hat{\sigma}_X^3 - \hat{\gamma}_Y \hat{\sigma}_Y^3}{\hat{\sigma}_M^3} \tag{5.19}$$

(4) Model both $\{M_i\}$ and $\{Y_t\}$ by three-parameter distributions, and assume that $\gamma_Y = \gamma_M = \gamma$. Then γ is estimated by

$$\hat{\gamma} = \frac{\hat{\gamma}_X \hat{\sigma}_X^3}{\hat{\sigma}_Y^3 + \hat{\sigma}_M^3} \tag{5.20}$$

The first option does not utilize $\hat{\gamma}_X$, while the other three do. In most practical situations, and in most examples in this paper involving the SM-1 model, option 2 with $\gamma_M = 0$ (that is $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$) is used.

Unless the shifting pattern of the historical record suggests otherwise, it may not be appropriate to preserve the sample skewness through the M_is since in a generated series of the same length as the historical record the number of values of M_is used is in most cases only a fraction of the historical record length. As a result the skewness of the generated sequence may be underestimated. In addition, for the special case where two-parameter skewed distributions, such as the Gumbel or the lognormal-2, are used to model the M_is and/or the Y_ts, then information about the sample skewness of X_t ($\hat{\gamma}_X$) is not really needed for estimation of the parameters, since in that case the skewness, γ is either fixed or a function of the other parameters. If $\hat{\gamma}_X$ is to be used for estimation, then the above estimation procedures in Eqs (5.14)-(5.17) need to be reformulated, where for example $\hat{\gamma}_X$ would be used instead of $\hat{\rho}_X(2)$ in the estimation procedure (this special case is discussed for the Gumbel distribution in section 5.2.5). As a result the model may not preserve closely the observed correlogram. This latter approach is not recommended for cases were the lengths of the different stationary states (related to p) are critical for estimating the statistical quantities of interest.

5.2.2 The SM-2 Model

The SM-2 model was developed so that the mean oscillates in systematic manner between high and low, and high again pattern (Chapter 4). As developed in Chapter 4, the model assumes that the underlying variables are normally distributed. In this paper we further modify the model to take into account the skewness of the process. In the SM-2 model it is assumed that

$$Z_t = \sum_{i=1}^{t} Q_i M_i I_{(S_{i-1}, S_i]}(t) \tag{5.21}$$

where $S_i = N_1 + N_2 + \ldots + N_i$ with $S_0 = 0$, as in section 5.2.1. The $\{M_i\}$ is a sequence of *iid* positive real valued random variables with mean μ_M, variance σ_M^2, and skewness γ_M. The $\{Q_i\}$ is a simple Markov chain with state space $\{-1, 1\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{5.22}$$

Thus $\mu_Q = 0$, $\sigma_Q^2 = 1$, and $\gamma_Q = 0$. The M_is represent the magnitude of the noise level Z_t while the Q_is represent the sign of the shift as compared to the long term mean of the process X_t. From the state space of $\{Q_i\}$ and Eq (5.22) it follows that two consecutive noise levels $Q_i M_i$ and $Q_{i+1} M_{i+1}$ will always have opposite signs (that is $Q_{i+1} = -Q_i$).

Assuming that $\{N_i\}$, $\{Q_j\}$, and $\{M_k\}$ are mutually independent, it may be shown that the mean, the variance, and the skewness of Z_t are

$$\mu_Z = 0 \tag{5.23}$$

$$\sigma_Z^2 = \sigma_M^2 + \mu_M^2 \tag{5.24}$$

$$\gamma_Z = 0 \tag{5.25}$$

Using Eqs (5.5), (5.6), (5.7) and (5.21) it may be shown that the lag h autocorrelation function of X_t for $h = 1, 2, \ldots$ is

$$\rho_X(h) = \frac{\sigma_M^2 (1-p)^h + \mu_M^2 (1-2p)^h}{\sigma_Y^2 + \sigma_M^2 + \mu_M^2} \tag{5.26}$$

The SM-2 model has five parameters, $\{\mu_Y, \sigma_Y, \mu_M, \sigma_M, p\}$. To reduce the number of parameters one can simplify the model by assuming that the M_is follow a one-parameter distribution. As in Chapter 4 the M_is are assumed to be the absolute zero mean *normal iid* variables. That is, if $W \sim N(\mu = 0, \sigma^2 = \beta^2)$, then $M = |W|$ with probability density function

$$f_M(m) = \sqrt{\frac{2}{\pi}}\beta^{-1}\exp\left(-\frac{m^2}{2\beta^2}\right)I_{[0,\infty)}(m) \tag{5.27}$$

where $E[M] = \sqrt{2/\pi}\,\beta$ and $Var(M) = (1-2/\pi)\beta^2$, respectively. Consequently, the number of parameters in the SM-2 model reduces from five to four. and the ACF in Eq (5.26) changes to

$$\rho_X(h) = \frac{\beta^2}{\pi(\sigma_Y^2+\beta^2)}\left[(\pi-2)(1-p)^h + 2(1-2p)^h\right] \tag{5.28}$$

The following estimation procedure may be used to estimate the parameters $\{\mu_Y, \sigma_Y, \beta, p\}$ in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\rho}_X(1)$, $\hat{\rho}_X(2)$. First the quadratic equation

$$\hat{p}^2\hat{\rho}_X(1)(\pi+6) - \hat{p}(2\hat{\rho}_X(1) - \hat{\rho}_X(2))(\pi+2) + (\hat{\rho}_X(1) - \hat{\rho}_X(2))\pi = 0 \tag{5.29}$$

is solved for $\hat{p}$. and then the estimates of μ_Y, β, and σ_Y^2 are obtained from

$$\hat{\mu}_Y = \hat{\mu}_X \tag{5.30}$$

$$\hat{\beta} = \hat{\sigma}_X\sqrt{\frac{\pi\hat{\rho}_X(1)}{\pi - \hat{p}(\pi+2)}} \tag{5.31}$$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\beta}^2 \tag{5.32}$$

Equation (5.29) may give two feasible estimates of $\hat{p}$, but usually only one of these estimates will yield both $\hat{\beta} > 0$ and $\hat{\sigma}_Y^2 > 0$ in Eqs (5.31) and (5.32). For discussion about the feasibility of parameter estimates refer to Chapter 4.

If the observed sample series has zero skewness ($\hat{\gamma}_X = 0$), then we assume that $Y_1, Y_2, \ldots \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ as in Chapter 4. On the other hand if the observed sample series has non-zero skewness (that is $\hat{\gamma}_X \neq 0$), then the sample skewness can be preserved by modeling the Y_ts by a two-parameter or a three-parameter skewed distribution. Thus the

following two options are available for modeling the Y_t s, where only option 1 is used in this paper:

(1) Model $\{Y_t\}$ by a three-parameter skewed distribution, where γ_Y is estimated by

$$\hat{\gamma}_Y = \frac{\hat{\gamma}_X \hat{\sigma}_X^3}{\hat{\sigma}_Y^3} \tag{5.33}$$

This option is used exclusively for the SM-2 model in this paper.

(2) Model $\{Y_t\}$ by a two-parameter skewed distribution. Hence, $\hat{\gamma}_Y = f(\hat{\mu}_Y, \hat{\sigma}_Y)$. This option is not used in this paper, but shown here as an alternative to option 1.

In some cases when $\sigma_Y^2 \ll \sigma_X^2$ the value of the estimated skewness of the Y_t s in Eq (5.33) can be unrealistically too high, resulting in infeasible or unrealistic parameters estimates of the distribution of the Y_t s. In such cases the SM-2 model is usually not the correct model to use for the process under consideration. As for the SM-1 model, in the special case when the sample skewness is preserved only through the use of two-parameter skewed distributions, then the sample skewness of X_t ($\hat{\gamma}_X$) is not really needed for estimation of parameters (refer to option 2 above). If $\hat{\gamma}_X$ is to be included in the estimation of the parameters, where for example $\hat{\gamma}_X$ is used instead of $\hat{\rho}_X(2)$, then the estimation procedures must be reformulated. Such special case involving the Gumbel distribution is discussed in section 5.2.5.

5.2.3 Limitations of the SM Models and Problems in Parameter Estimation

As a result of modeling $\{N_i\}$ by the positive geometric distribution, the ACF of the SM-1 model in Eq (5.13) can take only positive values and has the same form as the ACF of an ARMA(1, 1) process (Salas and Boes, 1980). The ACF of the SM-2 model (refer to Eq. (5.26) has a similar shape as the ACF of the SM-1 model, but in addition it can take negative values for odd valued lags. As discussed in Chapter 4 the choice of using the positive geometric distribution to model the lengths of the random time spans $\{N_i\}$ will always result in a fitted SM model with an ACF with no periodic components. Despite

that, the SM models may still be a good alternative for modeling observed time series with pseudo-periodic behavior.

The main problem that arises in estimation of parameters for both the SM-1 and SM-2 models is due to relatively tight constraints on the parameter space of the ACF of both models. For example, in the case of the SM-1 model, parameter estimates are feasible if and only if $\hat{\rho}_X(1) > \hat{\rho}_X(2) > \hat{\rho}_X^2(1)$. In Chapter 4 the parameter space for $\rho_X(1)$ in terms of $\rho_X(2)$ is plotted for both models. However, because of sample variability the values of $\hat{\rho}_X(1)$ and $\hat{\rho}_X(2)$ used for parameter estimation may result in infeasible parameters for the SM models. To reduce the effects of sample variability and sample periodic behavior in the estimation of parameters, the sample ACF may be fitted by functions that have exactly the same form as the model ACFs. For the ACF of the SM-1 model in Eq (5.13) such function would be

$$\rho_X(h) = ab^h, \qquad h = 1, 2, \ldots \tag{5.34}$$

where $a = \sigma_M^2/\sigma_X^2$ and $b = 1 - p$, with $0 < a < 1$ and $0 < b < 1$. Least squares estimates for a and b are easily obtained by fitting a straight line to the logs of Eq (5.34). In Chapter 4, Eq (5.34) was also used to fit the ACF of the SM-2 model and was found to give good results. On the other hand, in this paper we will fit the exact form of SM-2 model ACF to the historical ACF. The exact form of the ACF of the SM-2 model in Eq (5.28) is

$$\rho_X(h) = a\left[b^h + \frac{2}{\pi - 2}(2b - 1)^h\right], \qquad h = 1, 2, \ldots \tag{5.35}$$

where $a = \sigma_M^2/\sigma_X^2 = (1 - 2/\pi)\beta^2/\sigma_X^2$ and $b = 1 - p$, with $0 < a < 1$ and $0 < b < 1$. Least squares estimates of a and b using the sample ACF up to lag k are found by minimizing

$$S(a, b; \hat{\rho}_1, \ldots, \hat{\rho}_k) = \sum_{h=1}^{k}\left\{\hat{\rho}_h - a\left[b^h + \frac{2}{\pi - 2}(2b - a)\right]\right\}^2 \tag{5.36}$$

Equation (5.36) is minimized by taking the partial derivative with respect to the parameters

a and b, setting them equal to zero, and solving for the parameters. Hence,

$$\sum_{h=1}^{k}[b^h + c(2b-1)^h]\sum_{h=1}^{k} h\hat{\rho}_h[b^{h-1} + 2c(2b-1)^{h-1}]$$
$$-\sum_{h=1}^{k} h[b^h + c(2b-1)^h]\cdot[b^{h-1} + 2c(2b-1)^{h-1}]\sum_{h=1}^{k}\hat{\rho}_h[b^h + c(2b-1)^h] = 0 \tag{5.37}$$

is solved for b, where $c = 2/(\pi - 2)$, and then a is calculated from

$$a = \frac{\sum_{h=1}^{k}\hat{\rho}_h[b^h + c(2b-1)^h]}{\sum_{h=1}^{k}[b^h + c(2b-1)^h]} \tag{5.38}$$

The fitted correlograms (ACFs) may not always resemble the sample correlogram closely, but in most cases they will result in feasible parameter estimates for both the SM-1 and SM-2 models.

5.2.4 The Skewed Distributions Utilized in this Paper

In Chapter 4 it was assumed that the X_t process had zero skewness and that $Y_1, Y_2, \ldots \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ for both the SM-1 and SM-2 models. Furthermore, it was assumed that the noise levels $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$ for the SM-1 model. On the other hand, if X_t has non-zero skewness (and the modeler wants to preserve it) then skewed distributions must be used to model the Y_t process or/and the M_i process depending on the specific model (refer to sections 5.2.1 and 5.2.2). Three skewed distributions are considered here: the three-parameter generalized extreme value (GEV), the three-parameter Pearson Type III (PE3) , and the two-parameter Gumbel. They are summarized below for ease of reference.

The GEV Distribution

The CDF of the generalized extreme value (GEV) distribution with parameters α (location), β (scale) and $\kappa \neq 0$ (shape) is given by

$$F_Y(y) = \exp\left\{-\left[1 - \frac{\kappa}{\beta}(y-\alpha)\right]^{1/\kappa}\right\} \tag{5.39}$$

where $-\infty < \alpha < \infty$, $\beta > 0$, and $-\infty < \kappa < \infty$. The range of Y is: $\alpha + \beta/\kappa < y < \infty$ for $\kappa < 0$, and $-\infty < y < \alpha + \beta/\kappa$ for $\kappa > 0$. For $\kappa = 0$ the GEV is the Gumbel. The mean,

the variance, and the skewness are respectively

$$\mu_Y = \alpha + \beta[1 - \Gamma(1+\kappa)]/\kappa, \quad \text{for } \kappa > -1 \tag{5.40}$$

$$\sigma_Y^2 = \beta^2[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]/\kappa^2, \quad \text{for } \kappa > \frac{-1}{2} \tag{5.41}$$

$$\gamma_Y = -\text{sign}(\kappa)\frac{\Gamma(1+3\kappa) - 3\Gamma(1+2\kappa)\Gamma(1+\kappa) + 2\Gamma^3(1+\kappa)}{(\Gamma(1+2\kappa) - 2\Gamma^2(1+\kappa))^{3/2}}, \quad \text{for } \kappa > \frac{-1}{3} \tag{5.42}$$

where $\Gamma(\cdot)$ is the gamma function. Moment estimates of α, β, and κ can be obtained from Eqs (5.40)-(5.42). Referring to Eq (5.42) a polynomial approximation is given below for $\hat{\kappa}$ as a function of the sample skewness ($\hat{\gamma}_Y = \hat{\gamma}$) in the range, $-2.000 \leq \hat{\gamma} \leq 13.484$,

$$\hat{\kappa} = \begin{cases} P_1(\hat{\gamma}) & \text{if } 13.484 \geq \hat{\gamma} > 5.605\ , \ -0.3 \leq \kappa < -0.25 \\ P_2(\hat{\gamma}) & \text{if } 5.605 \geq \hat{\gamma} > 1.140\ , \ -0.25 \leq \kappa < 0 \\ P_3(\hat{\gamma}) & \text{if } 1.140 \geq \hat{\gamma} \geq -2.000\ , \ 0 \leq \kappa \leq 1.0 \end{cases} \tag{5.43}$$

with error in $\hat{\kappa}$, $\varepsilon \leq 0.0002$. The above polynomials are given by

$$\begin{aligned} P_1(x) &= -0.0508\,8977 - 0.0657\,3888\,x + 0.0073\,8300\,x^2 \\ &\quad -0.0004\,0487\,x^3 + 0.0000\,0870\,x^4 \\ P_2(x) &= 0.2904\,2499 - 0.3657\,1056\,x + 0.1208\,1829\,x^2 \\ &\quad -0.0233\,1413\,x^3 + 0.0024\,4473\,x^4 - 0.0001\,0736\,x^5 \\ P_3(x) &= 0.2776\,5603 - 0.3219\,6424\,x + 0.0602\,1770\,x^2 \\ &\quad +0.0165\,6822\,x^3 - 0.0058\,5995\,x^4 - 0.0020\,0171\,x^5 \end{aligned}$$

Then estimates of β and α are obtained from

$$\hat{\beta} = \sqrt{\frac{\hat{\sigma}_Y^2\,\hat{\kappa}^2}{\Gamma(1+2\hat{\kappa}) - \Gamma^2(1+\hat{\kappa})}} \tag{5.44}$$

$$\hat{\alpha} = \hat{\mu}_Y - \frac{\hat{\beta}}{\hat{\kappa}}(1 - \Gamma(1+\hat{\kappa})) \tag{5.45}$$

The PE3 Distribution

The pdf of the three-parameter gamma distribution with parameters α (location), β (scale) and κ (shape) is given by

$$f_Y(y) = \frac{1}{\beta\Gamma(\kappa)}\left(\frac{y-\alpha}{\beta}\right)^{\kappa-1} e^{-(y-\alpha)/\beta} I_{(\alpha,\infty)}(y) \tag{5.46}$$

where $-\infty < \alpha < \infty$, $\beta > 0$, and $\kappa > 0$ (the case $\beta < 0$ is not considered here). The mean, variance, and the skewness of Y are

$$\mu_Y = \alpha + \beta\kappa \tag{5.47}$$

$$\sigma_Y^2 = \beta^2\kappa \tag{5.48}$$

$$\gamma_Y = 2/\sqrt{\kappa} \tag{5.49}$$

respectively. Thus the moment estimates are

$$\hat{\kappa} = 4/\hat{\gamma}_Y^2 \tag{5.50}$$

$$\hat{\beta} = \hat{\sigma}_Y\sqrt{\hat{\kappa}} \tag{5.51}$$

$$\hat{\alpha} = \hat{\mu}_Y - \hat{\beta}\hat{\kappa} \tag{5.52}$$

The Gumbel Distribution

The CDF of the Gumbel distribution with parameters α (location) and β (scale) is given by

$$f_Y(y) = \frac{1}{\beta}\exp\left(-\frac{y-\alpha}{\beta} - e^{-(y-\alpha)/\beta}\right) I_{(-\infty,\infty)}(y) \tag{5.53}$$

where $-\infty < \alpha < \infty$ and $\beta > 0$. The mean, variance, and skewness of Y are

$$\mu_Y = \alpha - \Gamma'(1)\beta \tag{5.54}$$

$$\sigma_Y^2 = \beta^2\pi^2/6 \tag{5.55}$$

$$\gamma_Y \approx 1.1396 \tag{5.56}$$

where $\Gamma'(1)$ is the negative Euler's constant. Thus the moment estimates of the parameters are

$$\hat{\beta} = \sqrt{6}\cdot\hat{\sigma}_Y/\pi \tag{5.57}$$

$$\hat{\alpha} = \hat{\mu}_Y + \Gamma'(1)\hat{\beta} \tag{5.58}$$

5.2.5 Special Cases Involving the Gumbel Distribution

As discussed in sections 5.2.1 and 5.2.2, if only two-parameter skewed distributions are used to model the skewed component(s) of the SM models, then information about the skewness of the X_t process is not really needed for estimation of the parameters. As a consequence the fitted SM model may not preserve the observed sample skewness ($\hat{\gamma}_X$) well. If on the other hand, it is of particular interest to the modeler to preserve the observed sample skewness of the X_t process, then information about the skewness of the particular two-parameter distribution of interest along with the estimated sample skewness can be used in the estimation of the parameters of the SM-1 model or the SM-2 model. As a result, the number of sample statistics needed to estimate the parameters of the models is reduced by one, that is in addition to $\hat{\gamma}_X$ only estimates of μ_X, σ_X, and $\rho_X(1)$ are needed. An obvious drawback is that information about the decay rate of the sample correlogram is not used, thus the decay rate of the correlogram of the fitted model may depart significantly from the decay rate of the observed sample correlogram.

In the case of the SM-1 model, both the Y_t and the M_i process can be used in preserving the skewness. In this case the estimated parameters in terms of μ_X, σ_X, $\hat{\gamma}_X$, and $\rho_X(1)$ are found by first estimating μ_Y by

$$\hat{\mu}_Y = \hat{\mu}_X \tag{5.59}$$

Then depending on the modeling choices for $\{Y_t\}$ and $\{M_i\}$ estimates of σ_Y^2 and σ_M^2 are obtained from one of the following options:

(1) $Y_1, Y_2, \ldots \overset{iid}{\sim}$ Gumbel(α, β) and $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 \left(\hat{\gamma}_X / \gamma_Y\right)^{2/3} \tag{5.60}$$

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 - \hat{\sigma}_Y^2 \tag{5.61}$$

(2) $Y_1, Y_2, \ldots \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ and $M_1, M_2, \ldots \overset{iid}{\sim}$ Gumbel(α, β)

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 \left(\hat{\gamma}_X / \gamma_M\right)^{2/3} \tag{5.62}$$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\sigma}_M^2 \tag{5.63}$$

(3) $Y_1, Y_2, \ldots \overset{iid}{\sim}$ Gumbel(α_Y, β_Y) and $M_1, M_2, \ldots \overset{iid}{\sim}$ Gumbel(α_M, β_M). Solve

$$\left(\hat{\sigma}_X^2 - \hat{\sigma}_M^2\right)^3 = \left(\hat{\sigma}_X^3 \frac{\hat{\gamma}_X}{\gamma} - \hat{\sigma}_M^3\right)^2 \tag{5.64}$$

for $\hat{\sigma}_M^2$, where $\gamma_Y = \gamma_M = \gamma$. Usually only the largest feasible solution for $\hat{\sigma}_M^2$ will give feasible estimates of the model parameters. Then σ_Y^2 is estimated as in Eq (5.63).

At last p is estimated by

$$\hat{p} = 1 - \hat{\rho}_X(1)\frac{\hat{\sigma}_X^2}{\hat{\sigma}_M^2} \tag{5.65}$$

In the case of the SM-2 model in which $\{Y_t\}$ are assumed to be *iid* Gumbel variables, the parameters are estimated in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\gamma}_X$, and $\hat{\rho}_X(1)$ as:

$$\hat{\mu}_Y = \hat{\mu}_X \tag{5.66}$$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 \left[\frac{\hat{\gamma}_X}{\gamma_Y}\right]^{2/3} \tag{5.67}$$

$$\hat{\beta} = \sqrt{\hat{\sigma}_X^2 - \hat{\sigma}_Y^2} \tag{5.68}$$

$$\hat{p} = \frac{\pi}{\pi + 2}\left[1 - \frac{\hat{\sigma}_X^2 \hat{\rho}_X(1)}{\hat{\beta}^2}\right] \tag{5.69}$$

where $\gamma_Y \approx 1.1396$. The estimates exist if $0 < \hat{\gamma}_X < \gamma_Y \cdot \min[1, (1 - \hat{\rho}_X(1))^{3/2}, (1 + 0.5\pi\hat{\rho}_X(1))^{3/2}]$. Note that in the above estimation procedure the skewness of the model is preserved through the variance of the Y_t process (see Eq (5.67)). This can result in low estimates for β ($\hat{\beta} \ll \hat{\sigma}_X$) in Eq (5.68), which in turn may cause $\hat{p}$ in Eq (5.69) to become infeasible. As for the SM-1 model, the use of a Gumbel distribution to preserve the skewness of the X_t process, can significantly affect the memory of the model.

5.3 Examples

In this section we illustrate the applicability of the referred SM models using extreme hydrologic data that are autocorrelated and exhibit sudden shifts. Unlike the conventional

frequency analysis in which the underlying variables are assumed to be *iid* and quantiles can be determined in closed form or estimated numerically, the X_ts of the SM models are not identically distributed. Hence frequency analysis of extreme events, such as annual maximum flows or annual maximum precipitation, is conducted based on simulation experiments.

5.3.1 Quarter-Monthly Annual Maximum Outflows from Lake Ontario

Data of quarter-monthly annual maximum flows (AMF) for Lake Ontario are shown in Fig. 5.1 for the period 1900–1989. The data were obtained from Fernández and Salas (1999). The sample correlogram in the lower graph has been fitted by Eqs (5.34) and (5.35) using the observed ACF at lags 1–5. The sample statistics of the 90 year long time series are $\hat{\mu}_X = 7940$ *cms*, $\hat{\sigma}_X = 1001$ *cms*, and $\hat{\gamma}_X = 0.263$. The fitted ACF at lags 1 and 2 has values 0.579 and 0.449 for the SM-1 model, and values 0.568 and 0.444 for the SM-2 model. The parameters for the two SM models estimated in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\gamma}_X$, and the lag 1 and 2 ACF from the fitted correlogram are: for the SM-1 model assuming $\gamma_M = 0$, Eqs (5.14)-(5.18) give $\{\hat{p} = 0.224, \hat{\sigma}_M = 864.6$ *cms*, $\hat{\mu}_Y = 7940$ *cms*, $\hat{\sigma}_Y = 503.8$ *cms*, $\hat{\gamma}_Y = 2.060\}$, and for the SM-2 model Eqs (5.29)-(5.33) give $\{\hat{p} = 0.136, \hat{\beta} = 855.5$ *cms*, $\hat{\mu}_Y = 7940$ *cms*, $\hat{\sigma}_Y = 519.1$ *cms*, $\hat{\gamma}_Y = 1.883\}$.

Note that the sample skewness is $\hat{\gamma}_X = 0.263$. In conventional frequency analysis assuming *iid* series such a small sample skewness may be an indicator that the data is Gaussian. On the other hand, in the SM models, a small value of γ_X does not necessarily mean that γ_Y and γ_M are also small. For example, for the SM-1 model $\hat{\gamma}_Y = 2.059$ (assuming that $\gamma_M = 0$), and for the SM-2 model $\hat{\gamma}_Y = 1.883$. In the following analysis we will assume that $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$ for the SM-1 model. Furthermore, for comparison the Y_ts will be modeled by the normal distribution (zero skewness), the PE3 distribution (nonzero skewness), and the GEV distribution (nonzero skewness). Generated sequences of the same length as the historical record are shown in Figs. 5.2, 5.3, and 5.4 based on the normal, the PE3, and the GEV distributions, respectively. Comparing the figures, the effect of including

skewed distributions is clear. Furthermore, all generated sequences appear to represent the historical time series in Fig. 5.1 reasonably well in terms of variability and shifting pattern.

It is not possible to calculate explicitly return periods of design events from the fitted SM models. Instead simulations have to be used to estimate return periods of specific events. If the random variable τ represents the time between occurrences of the event $X \geq x_T$, then the return period T of the event x_T is defined as the expected value of τ. The fitted SM models were used to estimate the growth curves of the quarter-monthly AMF of Lake Ontario. The simulation experiment was repeated until each quantile had been exceeded 2,000 times. The average return period from the 2,000 occurrences of $X \geq x_T$ is shown in Fig. 5.5. The estimated growth curves for the SM-1 and SM-2 models are clearly different when the distribution of the Y_t s in the models is skewed or not. Furthermore, the estimated growth curves based on the skewed distributions PE3 and GEV are similar to each other and appear to fit the historical data better than the growth curves based on the normal distribution. The reason for the similar growth curves obtained for the PE3 and the GEV models is likely due to the fact that the variance of the Z_t s is dominating (about three times larger) over the variance of the Y_t s.

5.3.2 Annual Maximum Flows of Cache La Poudre River Near Greeley, Colorado

Annual maximum flows of 3-day duration for the Cache La Poudre River are shown in Fig. 5.6 for the period 1925–1998. The lower plot shows the sample correlogram with fitted ACFs for the SM-1 (refer to Eq (5.34)) and SM-2 (refer to Eq (5.35)) models using the observed ACF at lags 1–11. The AMF series was derived from daily raw flow data obtained from USGS. The water year was defined as starting on October 1, and ending on September 30, where the year is defined as the year containing the ending date. The sample statistics of the 74 year long historical sample and the values of the fitted ACFs at lags 1 and 2 are shown in Table 5.1. The estimated parameters of the SM-1 and SM-2 models in terms of

the sample statistics $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\gamma}_X$, the fitted correlogram $\hat{\rho}_X(1)$, and $\hat{\rho}_X(2)$ are: for the SM-1 model with $\gamma_M = 0$, Eqs (5.14)-(5.18) give $\{\hat{p} = 0.166,\ \hat{\sigma}_M = 628.1\ cfs,\ \hat{\mu}_Y = 1200\ cfs,\ \hat{\sigma}_Y = 931.2\ cfs,\ \hat{\gamma}_Y = 2.602\}$; and for the SM-2 model Eqs (5.29)-(5.33) give $\{\hat{p} = 0.107,\ \hat{\beta} = 680.9\ cfs,\ \hat{\mu}_Y = 1200\ cfs,\ \hat{\sigma}_Y = 893.3\ cfs,\ \hat{\gamma}_Y = 2.948\}$.

In order to preserve the sample skewness, $\hat{\gamma}_X$, the Y_ts in both models are assumed to be *iid* GEV or PE3 variables, and the M_is of the SM-2 model are assumed to be *iid* normal variables. Based on the fitted models, 1,000 realizations of the same length (n) as the historical data, and one realization of length $1,000\ n$ of the Cache La Poudre River AMF data were generated. The statistics of the generated realizations are shown in Table 5.1. Focusing on the average statistics from the 1,000 realizations, for all scenarios the mean and the standard deviation are well preserved, while the skewness coefficient and the ACF at lags 1 and 2 are somewhat underestimated. However, the statistics calculated from the $1,000\ n$ long sample appear to be close to the historical ones. Furthermore, the results in Table 5.1 show that in judging the performance of a model for reproducing second and third order moment statistics specially when the underlying data are correlated, it is better to compare the statistics estimated based on a long sample (in our case $1,000\ n$ long) rather than statistics estimated by averaging from several sets of smaller sample sizes (in our case 1,000 sample n years long).

The estimated growth curves for the Lake Ontario data in based on $Y_1, Y_2, \ldots \overset{iid}{\sim}$ GEV and $Y_1, Y_2, \ldots \overset{iid}{\sim}$ PE3 were similar for the range shown in Fig. 5.5. We indicated that a possible explanation for the similarity was that the variability of the noise Z_t dominated over the variability of Y_t. On the other hand, for the AMF series of the Cache La Poudre River $\hat{\sigma}_Y^2 > 2\hat{\sigma}_Z^2$. Thus we expect to see some differences between the estimated growth curves based on $Y_1, Y_2, \ldots \overset{iid}{\sim}$ GEV and $Y_1, Y_2, \ldots \overset{iid}{\sim}$ PE3. The estimated growth curves of the fitted SM-1 and SM-2 models under these scenarios are shown in Fig. 5.7. The growth curves are estimated based on a repeated simulation experiment until each quantile has been exceeded 2,000 times, as in section 5.3.1. For the range shown in Fig. 5.7 the estimated growth curves

based on $Y_1, Y_2, \ldots \overset{iid}{\sim}$ PE3 result in somewhat higher quantile estimates than the curves based on $Y_1, Y_2, \ldots \overset{iid}{\sim}$ GEV.

5.3.3 Maximum Daily Precipitation for Dillon, Colorado

In Fig. 5.8 maximum daily precipitation (1909-1998) for Dillon, Colorado, is plotted along with its correlogram and fitted correlograms (Eqs (5.34) and (5.35)) using the observed ACF up to lag 15. The data were obtained from the National Climatic Data Center. The sample statistics of the 90 year long historical sample and the values of the fitted ACFs at lags 1 and 2 are shown in Table 5.2 and repeated in Table 5.3. In order to preserve the historical skewness an attempt is made to use the skewed models introduced earlier. The estimated parameters for both models in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\gamma}_X$, and the lag 1 and 2 ACF from the fitted correlograms are: for the SM-1 model Eqs (5.14)-(5.19) give $\{\hat{p} = 0.053$, $\hat{\sigma}_M^2 = 0.272\ cm^2$, $\hat{\mu}_Y = 2.618\ cm$, $\hat{\sigma}_Y^2 = 0.887\ cm^2$, $\hat{\gamma}_Y = 1.535$ assuming $\gamma_M = 0$, $\hat{\gamma}_M = 9.050$ assuming $\gamma_Y = 0\}$. and for the SM-2 model Eqs (5.29)-(5.33) give $\{\hat{p} = 0.033$, $\hat{\beta} = 0.521\ cm$, $\hat{\mu}_Y = 2.618\ cm$, $\hat{\sigma}_Y^2 = 0.887\ cm^2$, $\hat{\gamma}_Y = 1.535\}$. As in the previous examples, 1,000 realizations of the Dillon annual maximum precipitation data of the same length as the historical record (n) and one realization of length $1,000\ n$ were generated for the cases: SM-1 model with M_i's $\overset{iid}{\sim} N$ and Y_t's $\overset{iid}{\sim}$ GEV, SM-1 model with M_i's $\overset{iid}{\sim}$ GEV and Y_t's $\overset{iid}{\sim} N$, and SM-2 model with Y_t's $\overset{iid}{\sim}$ GEV. The average statistics of the 1,000 realization and the statistics of the one realization for each case, respectively, are shown in Table 5.2. For Y_t's $\overset{iid}{\sim}$ GEV, the historical sample statistics are reasonably well preserved by both models, except that the average skewness and average ACF at lags 1 and 2 based on 1,000 realizations is underestimated by both models. For the SM-1 model with M_i's $\overset{iid}{\sim}$ GEV, the historical mean and the variance for both types of generated realizations and the ACF based on the one generated sequence of length 1,000 n are well preserved, while the average ACF based on the 1,000 generated sequences is underestimated, and the historical skewness is severely underestimated for both types of generated realizations.

Applying the special case of the Gumbel distribution with parameters estimated based on procedures in section 5.2.5, then the parameter estimates for the SM-2 model are not feasible, but for the SM-1 model the parameters are feasible when $\gamma_Y = 0$ and $\gamma_M = 1.1396$, and for the case when $\gamma_Y = \gamma_M = 1.1396$. In terms of the $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\gamma}_X$ and the lag 1 ACF of the fitted correlogram the estimated parameters for the former case ($\gamma_Y = 0$ and $\gamma_M = 1.1396$) using Eqs (5.59), (5.62)-(5.63), and (5.65) are $\{\hat{p} = 0.762,\ \hat{\sigma}^2_M = 1.082cm^2,\ \hat{\mu}_Y = 2.618\ cm,\ \hat{\sigma}^2_Y = 0.077\ cm^2\}$, and for the latter case ($\gamma_Y = \gamma_M = 1.1396$) Eqs (5.59), (5.63)-(5.65) give $\{\hat{p} = 0.765,\ \hat{\sigma}^2_M = 1.092cm^2,\ \hat{\mu}_Y = 2.618\ cm,\ \hat{\sigma}^2_Y = 0.066\ cm^2\}$. The average statistics of 1,000 generated sequences of length n, and one generated sequence of length 1,000 n for the cases M_i's $\overset{iid}{\sim}$ Gumbel and Y_t's $\overset{iid}{\sim}$ N, and M_i's $\overset{iid}{\sim}$ Gumbel and Y_t's $\overset{iid}{\sim}$ Gumbel are shown in Table 5.3. Note that the ACF decays very quickly resulting in short memory of the fitted model. The reason for this is that only $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\gamma}_X$ and the lag 1 ACF of the fitted correlogram were used in the fitting process of the model. Since only the lag 1 ACF of the correlogram is used in the fitting process, no information about the decay rate of the correlogram is included. In fact the skewness is preserved through the variance of the M_i process (refer to Eq (5.62) or (5.64)). A high value of $\hat{\sigma}^2_M$ can result in high value of p (refer to Eq (5.65)). High value of p results in short memory of the fitted model, while low value of p results in longer memory of the fitted model. Thus, if a particular shape of the correlogram is to be preserved, then it may not be appropriate to use a two-parameter skewed distribution, such as the Gumbel, to preserve the skewness of the historical data.

5.3.4 3-Day Annual Maximum Flows of the Colorado River at Hot Sulphur Springs

In Fig. 5.9 annual maximum 3-day flows (1930-1994) for the Colorado River at Hot Sulphur Springs (Colorado), are plotted along with its correlogram and fitted correlograms (Eqs (5.34) and (5.35)) using the oberved ACF up to lag 15. The raw data were obtained from USGS. The sample statistics of the 65 year long historical sample are $\hat{\mu}_X = 2052\ cfs$,

$\hat{\sigma}_X = 1435$ *cfs*, and $\hat{\gamma}_X = 0.550$. The estimated parameters for the SM-1 and SM-2 models in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$ and the lag 1 and 2 ACF from the fitted correlogram are: for the SM-1 model Eqs (5.14)-(5.17) give $\{\hat{p} = 0.079, \hat{\sigma}_M = 986.4$ *cfs*, $\hat{\mu}_Y = 2052$ *cfs*, $\hat{\sigma}_Y = 1042$ *cfs*, and for the SM-2 model Eqs (5.29)-(5.32) give $\{\hat{p} = 0.050, \hat{\beta} = 1015$ *cfs*. $\hat{\mu}_Y = 2052$ *cfs*, $\hat{\sigma}_Y = 1014$ *cfs*. Furthermore if the intent is to preserve the skewness through the Y_ts of both models, then from Eq (5.18) $\hat{\gamma}_Y = 1.434$ for the SM-1 model, and from Eq-(5.33) $\hat{\gamma}_Y = 1.556$ for the SM-2 model. Assuming that the Y_ts follow a three-parameter PE3 distribution for both models, the growth curves for both the SM-1 and the SM-2 models were estimated using simulations. The results are shown in Fig. 5.10, where the estimated growth curves based on the SM-1 and SM-2 models appear identical over the range of return periods shown in the figure, and both estimated growth curves seem to fit the empirical data quite well. In fact the estimated empirical return period of the largest observation appears to be overestimated.

5.3.5 Mean Annual Flows of the Colorado River at Hot Sulphur Springs

In Fig. 5.11 mean annual flows (1930-1994) for the Colorado River at Hot Sulphur Springs (Colorado), are plotted along with its correlogram and fitted correlograms (Eqs (5.34) and (5.35)) using the oberved ACF up to lag 15. The raw data were obtained from USGS. The sample statistics of the 65 year long historical sample are $\hat{\mu}_X = 339.1$ *cfs*, $\hat{\sigma}_X = 192.7$ *cfs*, and $\hat{\gamma}_X = 0.649$. The estimated parameters for the SM-1 and SM-2 models in terms of $\hat{\mu}_X$, $\hat{\sigma}_X$ and the lag 1 and 2 ACF from the fitted correlograms are: for the SM-1 model Eqs (5.14)-(5.17) give $\{\hat{p} = 0.078, \hat{\sigma}_M = 141.8$ *cfs*, $\hat{\mu}_Y = 339.1$ *cfs*, $\hat{\sigma}_Y = 130.5$ *cfs*, and for the SM-2 model Eqs (5.29)-(5.32) give $\{\hat{p} = 0.051, \hat{\beta} = 145.1$ *cfs*, $\hat{\mu}_Y = 339.1$ *cfs*, $\hat{\sigma}_Y = 126.9$ *cfs*. Furthermore, in order to preserve the skewness through the Y_ts of both models, then from Eq (5.18) $\hat{\gamma}_Y = 2.093$ for the SM-1 model, and from Eq-(5.33) $\hat{\gamma}_Y = 2.277$ for the SM-2 model. Simulations were used to estimate the return period of drought durations of lengths 5 to 25 years, based on the fitted SM-1 and SM-2 models with the Y_ts

modeled by a three-parameter PE3 distribution for both models. The demand level was assumed equal to be the mean annual flow. Average return periods of 2,000 occurrences of each drought length are shown in Fig. 5.12. The estimated growth curves of drought duration for the SM-1 and SM-2 models in the figure appear similar. In the historical sample the longest drought duration is 8 years with two occurrences. The return period of a 8 year drought is 42.9 years based on the fitted SM-1 model and 41.3 years based on the fitted SM-2 model.

5.4 Final Remarks and Conclusions

The main contribution of this paper is extending the shifting mean models (SM-1 and SM-2) repeated in Chapter 4 for modeling skewed hydrologic processes. The SM models are strictly stationary, but can be considered to exhibit some kind of "non-stationarity" in the mean, in the sense that they are allowed to shift from one stationary state to another around a long term mean. The process of interest (X_t) is assumed to be a sum of two independent random variables Y_t and Z_t, where the Y_ts are *iid* variables and the Z_ts are assumed to represent departure of each "stationary" state from the long term mean of the process. That is, during each "stationary state" the Z_ts remain fixed at a value referred to as a noise level. In the SM-2 model two consecutive stationary states always have noise levels of opposite signs, while in the SM-1 model the noise levels are allowed to fluctuate in random manner. In this paper only the positive geometric distribution was considered for modeling the length the process spent in each stationary state.

Several examples were used to illustrate the applicability of the SM models for fitting hydrologic time series. The skewed distributions considered in the examples, were the GEV, the PE3, and the Gumbel. In general the SM models were capable of preserving the mean, variance, skewness, and the autocorrelation function of the sample series. In addition, the SM models were shown to be useful for generating long sample series for use in frequency analysis of extreme events. The results showed that the SM-1 and SM-2 models usually give

similar results if the corresponding components of each model were modeled by the same type of distribution. Furthermore, the SM models are capable of modeling time series that are correlated and have shifting patterns as shown in the historical samples. On the other hand, the assumed SM models used in this paper do not model trends or time series with changes in the sample variability. In general, the SM models appear to have a wide range of applicability for modeling any type of climatic, hydrologic, and/or geophysical processes.

Table 5.1: Statistics of generated realizations for the Cache La Poudre River AMF data in section 5.3.2. The numbers in parenthesis are statistics from 1 realization of length 1,000 n, while the numbers not in parenthesis are average statistics from 1,000 realizations of the same length as the historical data (n).

Statistic	Historical	$Y_1, Y_2, \ldots \overset{iid}{\sim}$ GEV		$Y_1, Y_2, \ldots \overset{iid}{\sim}$ PE3	
		SM-1	SM-2	SM-1	SM-2
$\hat{\mu}_X$ [cfs]	1200	1194 (1190)	1205 (1186)	1191 (1200)	1213 (1199)
$\hat{\sigma}_X$ [cfs]	1123	1087 (1127)	1102 (1121)	1095 (1121)	1094 (1130)
$\hat{\gamma}_X$	1.483	1.136 (1.426)	1.121 (1.466)	1.331 (1.527)	1.385 (1.484)
$\hat{\rho}_X(1)$ SM-1	0.261	0.201 (0.269)		0.212 (0.256)	
$\hat{\rho}_X(2)$ SM-1	0.218	0.157 (0.223)		0.163 (0.210)	
$\hat{\rho}_X(1)$ SM-2	0.303		0.244 (0.300)		0.239 (0.301)
$\hat{\rho}_X(2)$ SM-2	0.251		0.182 (0.248)		0.184 (0.251)

Table 5.2: Statisticd of generated realizations for the Dillon annual maximum daily precipitation data. The numbers in parenthesis are statistics from 1 realization of length 1,000 n, while the numbers not in parenthesis are average statistics from 1,000 realizations of the same length as the historical data (n).

Statistic	Historical	$Y_1, Y_2, \ldots \overset{iid}{\sim}$ GEV		$M_1, M_2, \ldots \overset{iid}{\sim}$ GEV
		SM-1	SM-2	SM-1
$\hat{\mu}_X$ [cm]	2.618	2.624 (2.608)	2.632 (2.612)	2.613 (2.622)
$\hat{\sigma}^2_X$ [cm^2]	1.159	1.080 (1.165)	1.039 (1.131)	1.081 (1.156)
$\hat{\gamma}_X$	1.028	1.039 (1.036)	1.129 (1.069)	0.060 (0.614)
$\hat{\rho}_X(1)$ SM-1	0.222	0.137 (0.222)		0.088 (0.221)
$\hat{\rho}_X(2)$ SM-1	0.210	0.122 (0.211)		0.076 (0.208)
$\hat{\rho}_X(1)$ SM-2	0.203		0.072 (0.198)	
$\hat{\rho}_X(2)$ SM-2	0.198		0.073 (0.190)	

Table 5.3: Statistics of generated realizations for the Dillon annual maximum daily precipitation data, for the special case of the Gumbel distribution with parameters estimated based on procedures in section 5.2.5. The numbers in parenthesis are statistics from 1 realization of length 1,000 n, while the numbers not in parenthesis are average statistics from 1,000 realizations of the same length as the historical data (n).

Statistic	Historical	$\gamma_M = 1.1396$, $\gamma_Y = 0$	$\gamma_M = 1.1396$, $\gamma_Y = 1.1396$
		SM-1	SM-1
$\hat{\mu}_X$ [cm]	2.618	2.613 (2.627)	2.620 (2.622)
$\hat{\sigma}^2_X$ [cm^2]	1.159	1.144 (1.170)	1.160 (1.166)
$\hat{\gamma}_X$	1.028	0.900 (1.031)	0.938 (1.073)
$\hat{\rho}_X(1)$ SM-1	0.222	0.195 (0.225)	0.193 (0.223)
$\hat{\rho}_X(2)$ SM-1	0.210	0.029 (0.056)	0.026 (0.053)

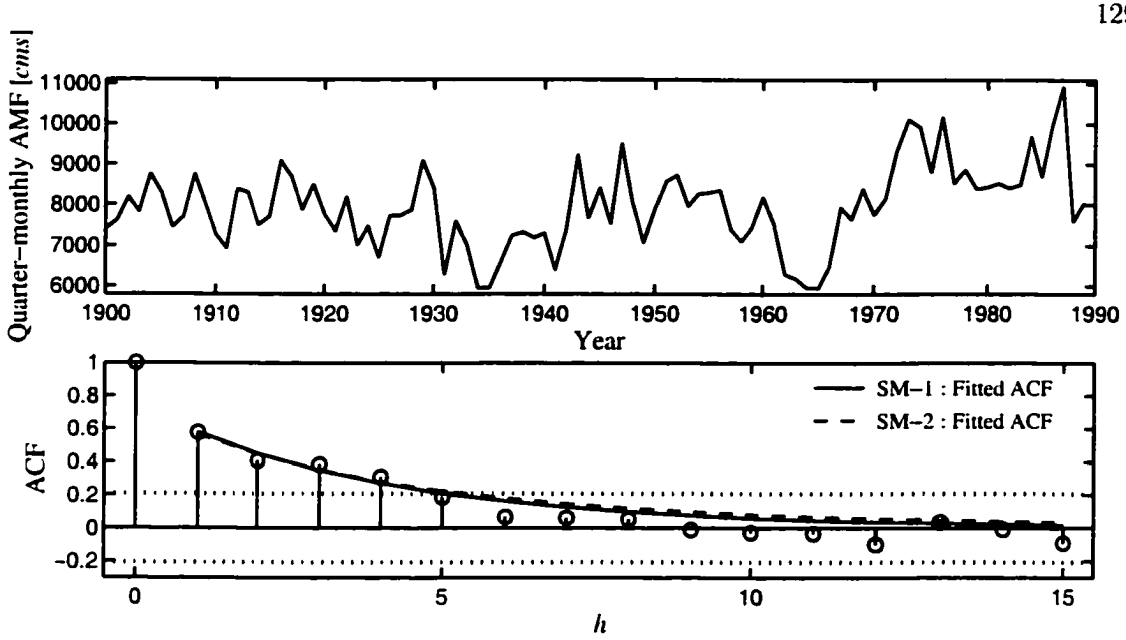

Figure 5.1: Quarter-monthly annual maximum outflow series (1900–1989) for Lake Ontario. The bottom plot shows the sample correlogram and fitted correlograms for the SM-1 and SM-2 models.

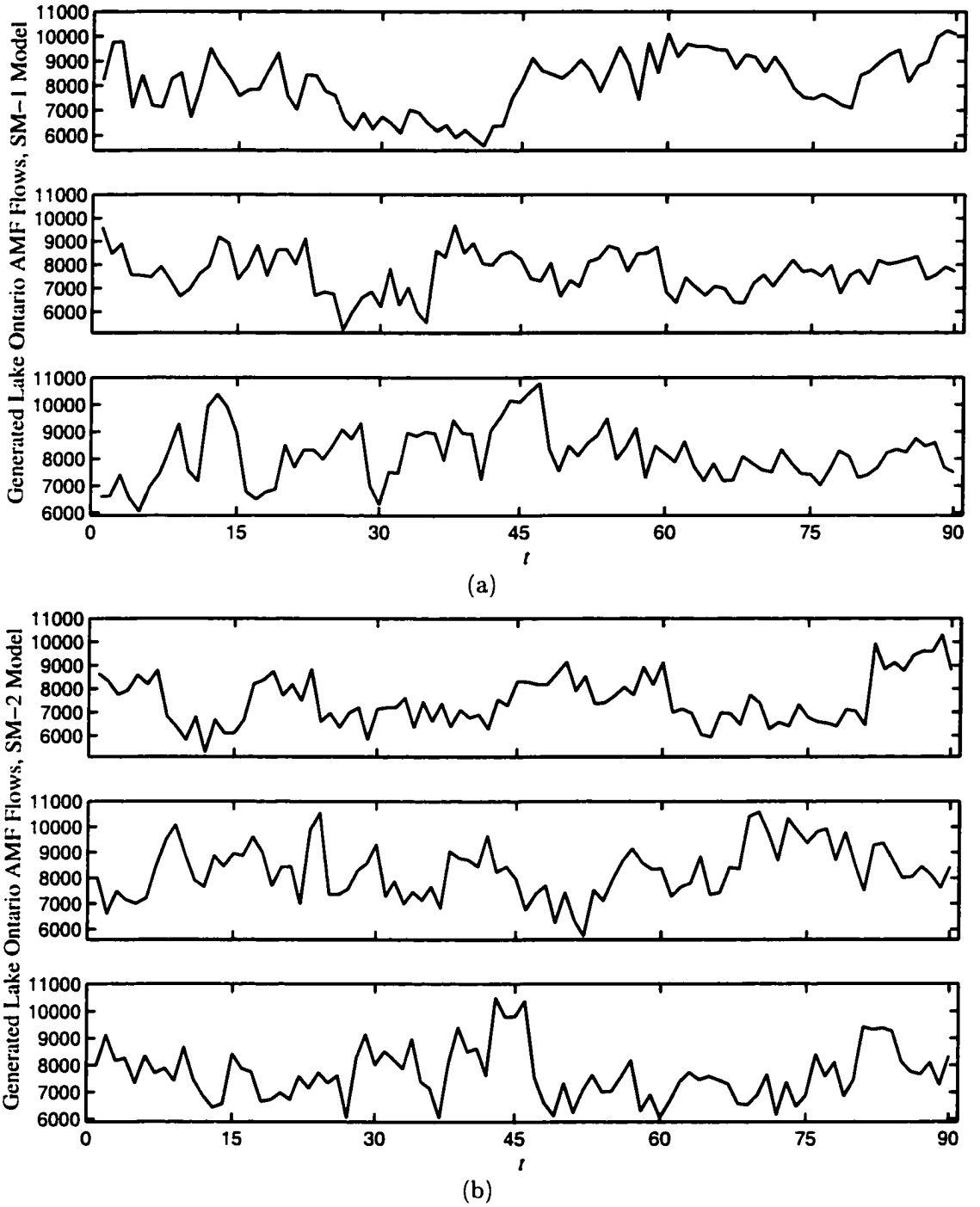

Figure 5.2: Generated sequences of quarter-monthly AMF for Lake Ontario. (a) SM-1 model with $Y_1, Y_2, \ldots \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ and $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$, and (b) SM-2 model with $Y_1, Y_2, \ldots \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$.

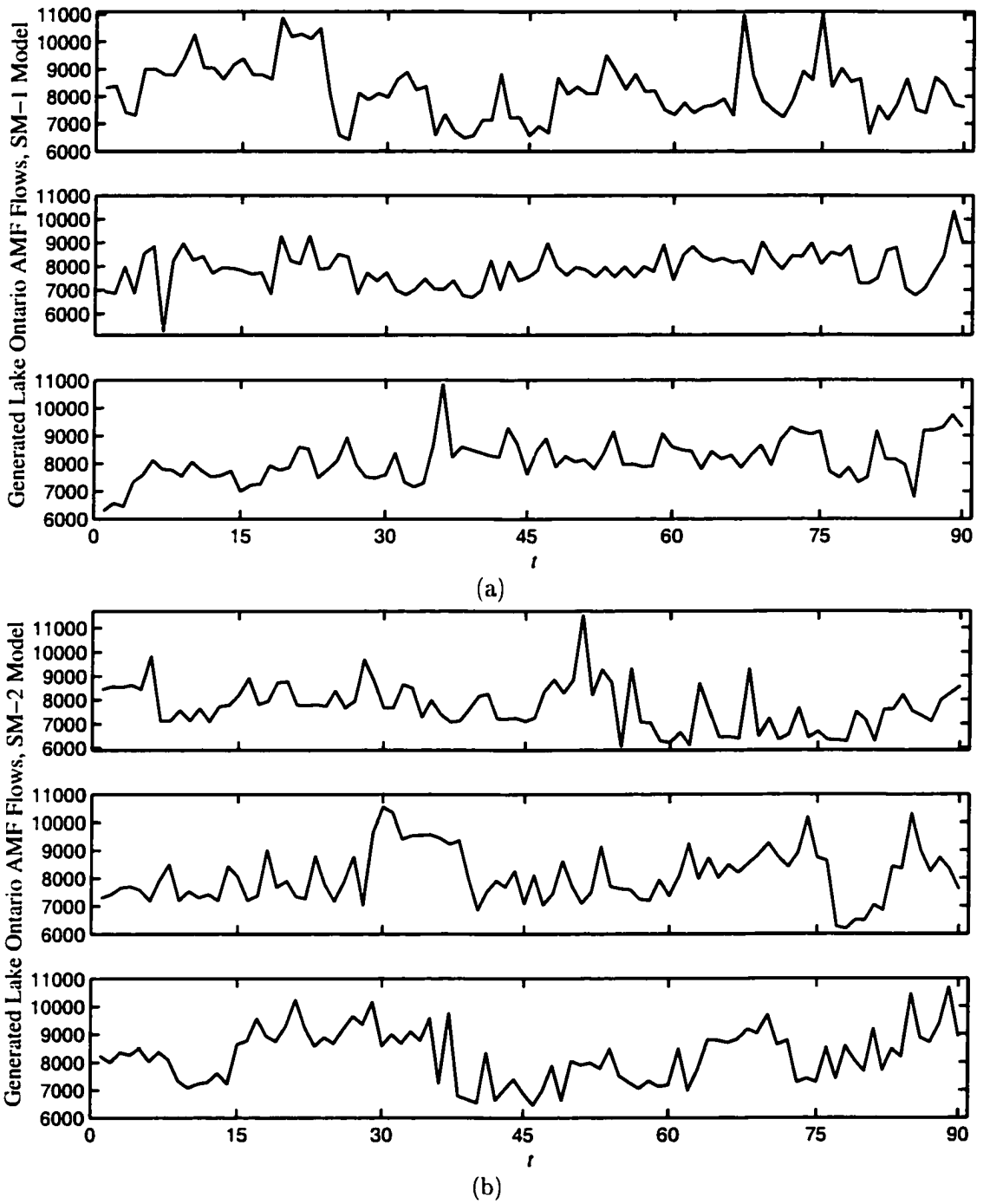

Figure 5.3: Generated sequences of quarter-monthly AMF for Lake Ontario. (a) SM-1 model with $Y_1, Y_2, \ldots \overset{iid}{\sim} \text{PE3}(\alpha, \beta, \kappa)$ and $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$, and (b) SM-2 model with $Y_1, Y_2, \ldots \overset{iid}{\sim} \text{PE3}(\alpha, \beta, \kappa)$.

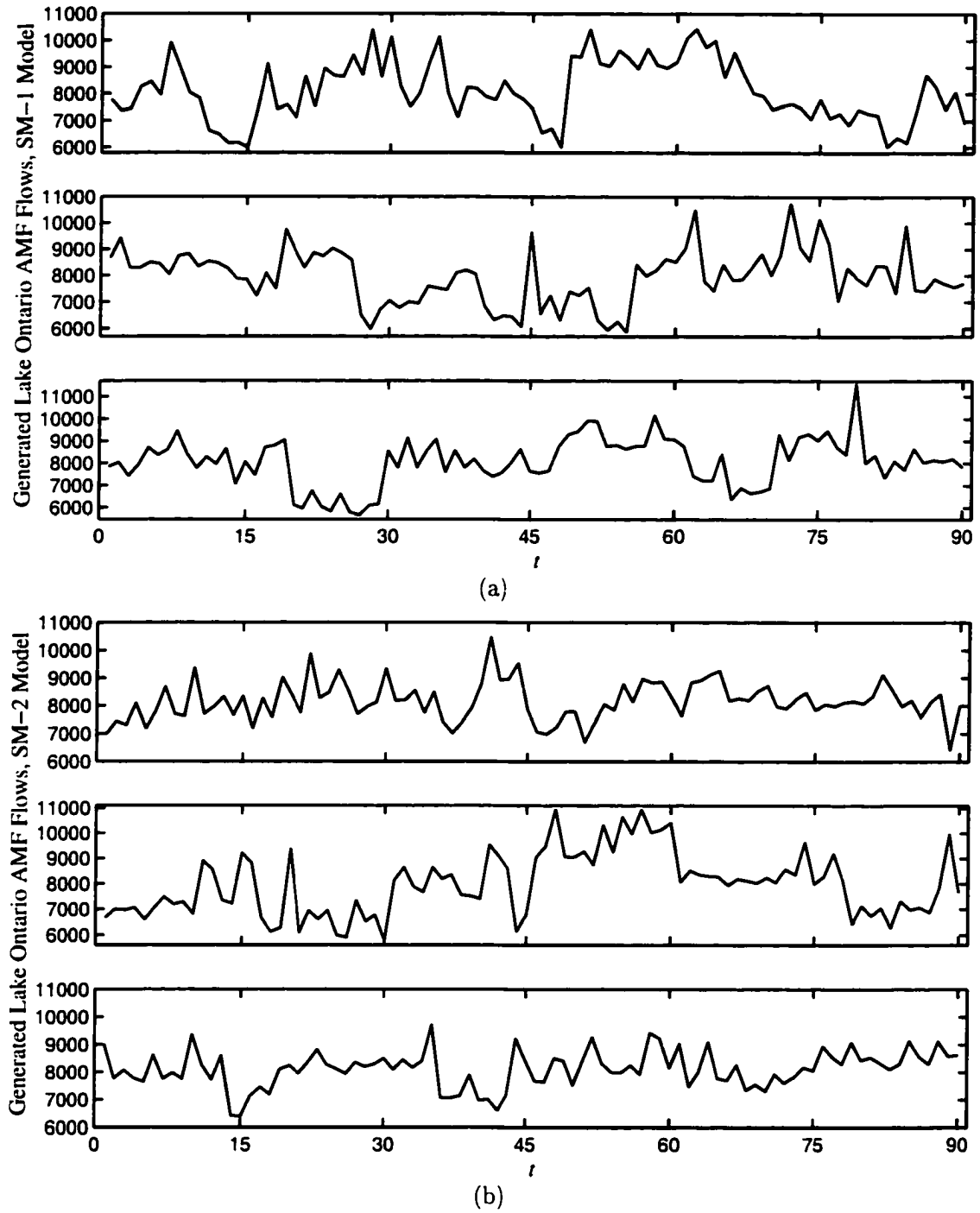

Figure 5.4: Generated sequences of quarter-monthly AMF for Lake Ontario. (a) SM-1 model with $Y_1, Y_2, \ldots \overset{iid}{\sim} \text{GEV}(\alpha, \beta, \kappa)$ and $M_1, M_2, \ldots \overset{iid}{\sim} N(0, \sigma_M^2)$, and (b) SM-2 model with $Y_1, Y_2, \ldots \overset{iid}{\sim} \text{GEV}(\alpha, \beta, \kappa)$.

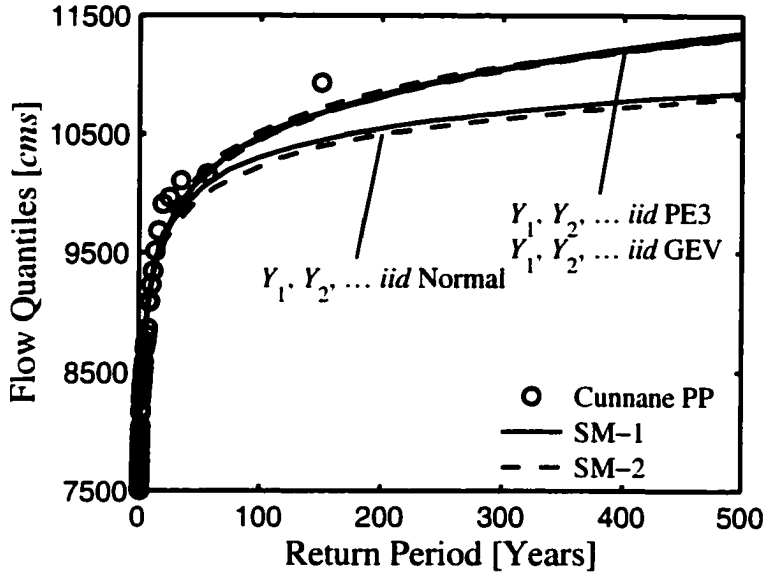

Figure 5.5: Estimated growth curves of quarter-monthly AMF for Lake Ontario based on the fitted SM-1 and SM-2 models. The empirical data points are based on the Cunnane plotting position formula.

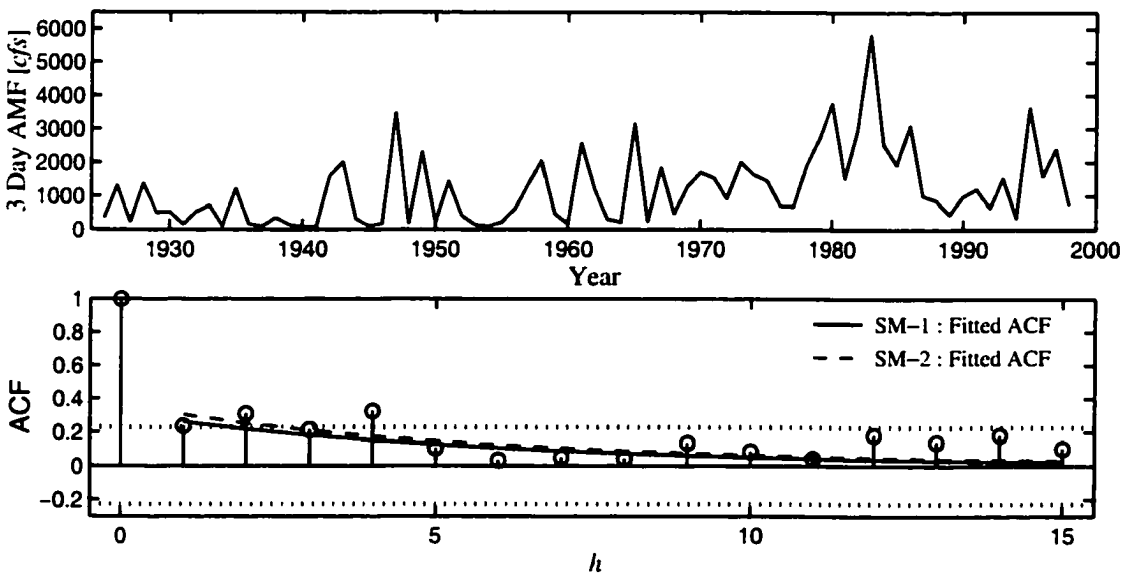

Figure 5.6: AMF series (1925–1998) of 3-day duration of Cache La Poudre River near Greeley, Colorado. The bottom plot shows the sample correlogram and fitted correlograms for the SM-1 and SM-2 models.

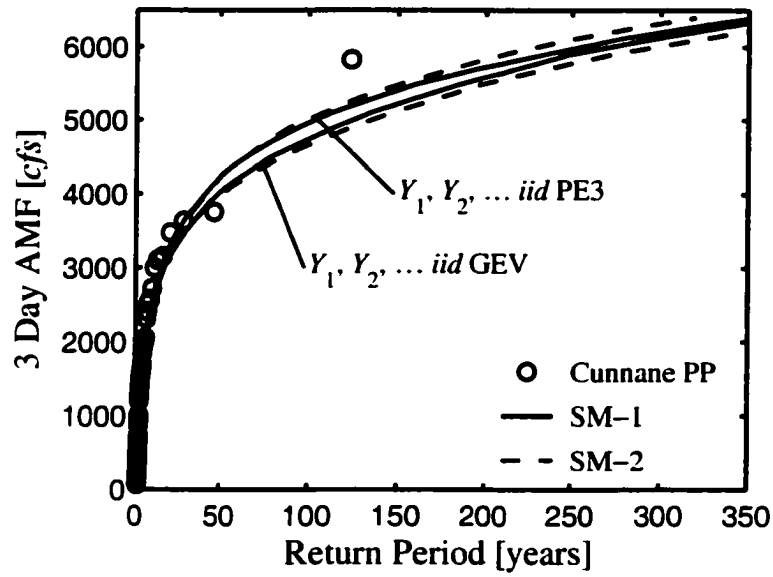

Figure 5.7: Estimated growth curves of 3 day AMF for Cache La Poudre River based on fitted a SM-1 and SM-2 models. The empirical data points are based on the Cunnane plotting position formula.

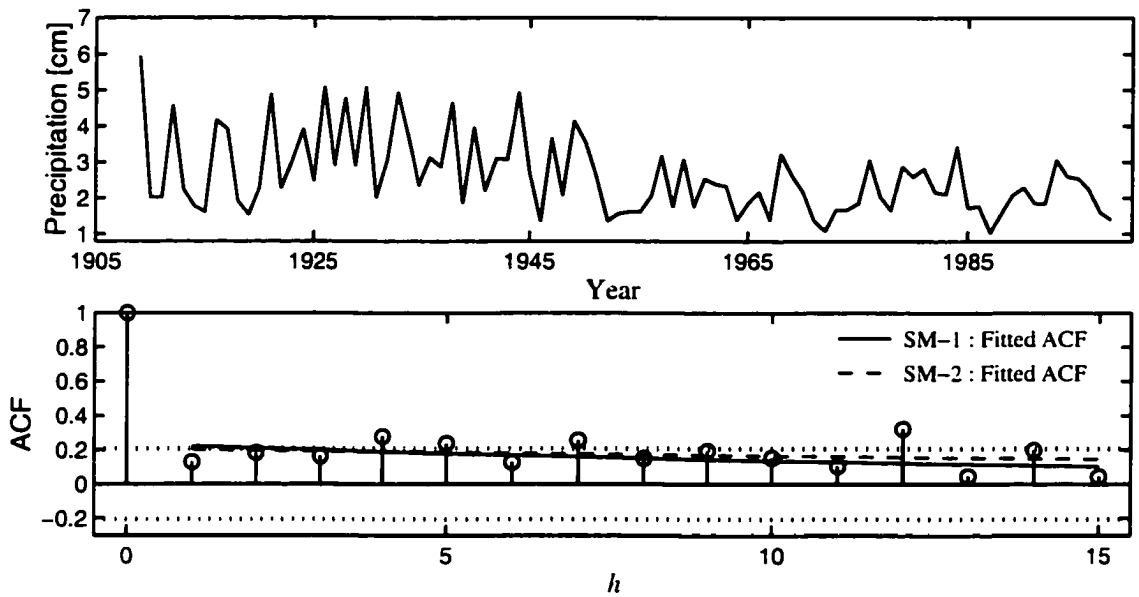

Figure 5.8: 1909–1998 daily maximum precipitation for Dillon, Colorado. The bottom plot shows the the sample correlogram and fitted correlograms for the SM-1 and SM-2 models.

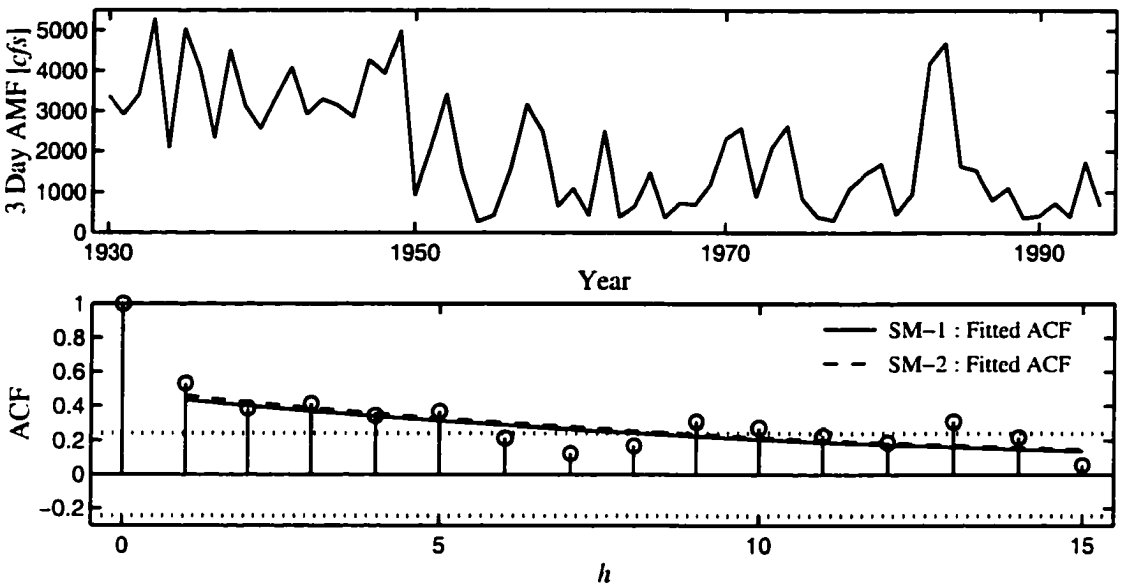

Figure 5.9: 1930–1994 3 day annual maximum flows of the Colorado River at Hot Sulphur Springs, Colorado. The bottom plot shows the sample correlogram and fitted correlograms for the SM-1 and SM-2 models.

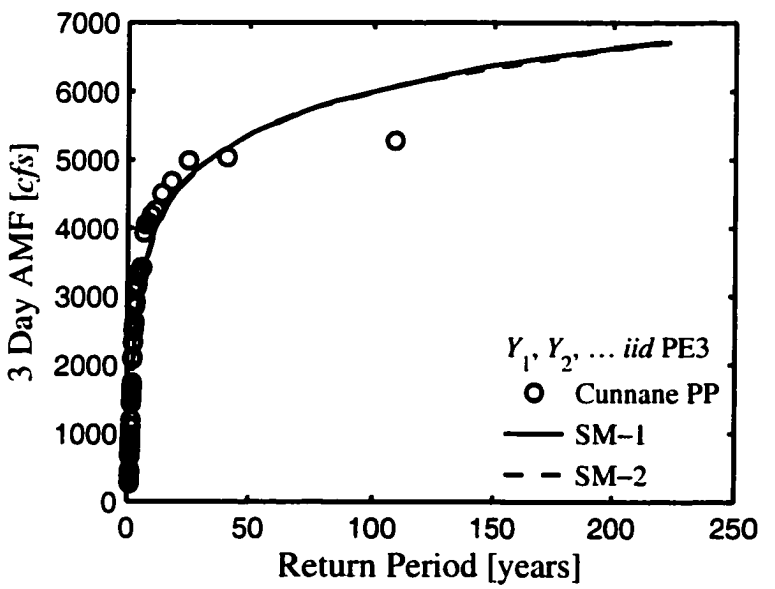

Figure 5.10: Growth curves of 3 day AMF for Colorado River at Hot Sulphur Springs. The estimated growth curves are based on fitting a SM-1 and a SM-2 model to the historical data. The empirical data points are based on the Cunnane plotting position formula.

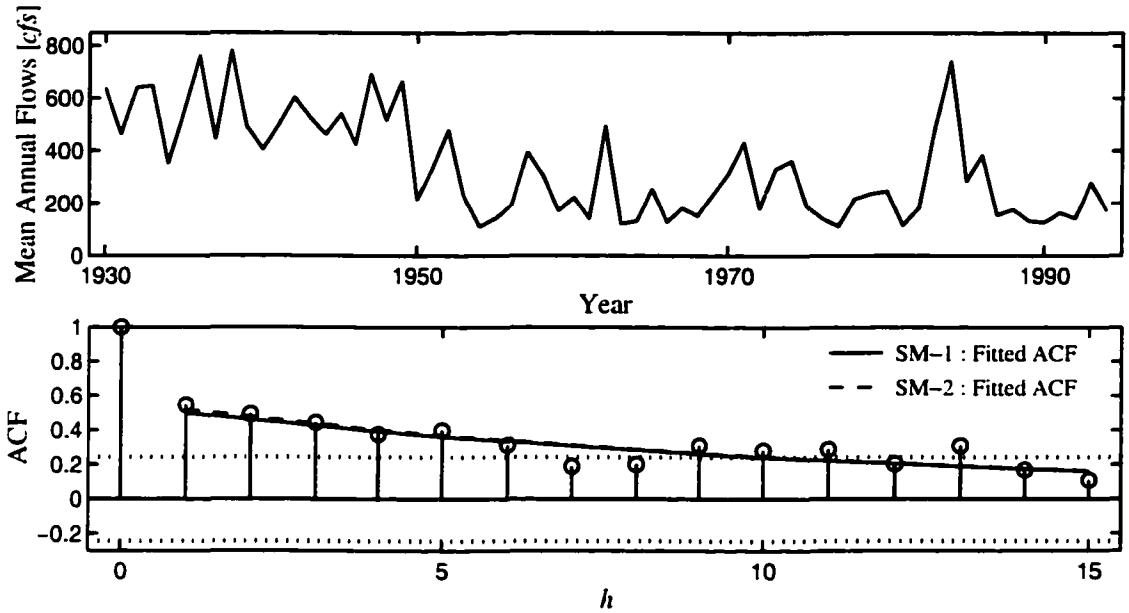

Figure 5.11: 1930–1994 mean annual flows of Colorado River at Hot Sulphur Springs, Colorado. The bottom plot shows the sample correlogram and fitted correlograms for the SM-1 and SM-2 models.

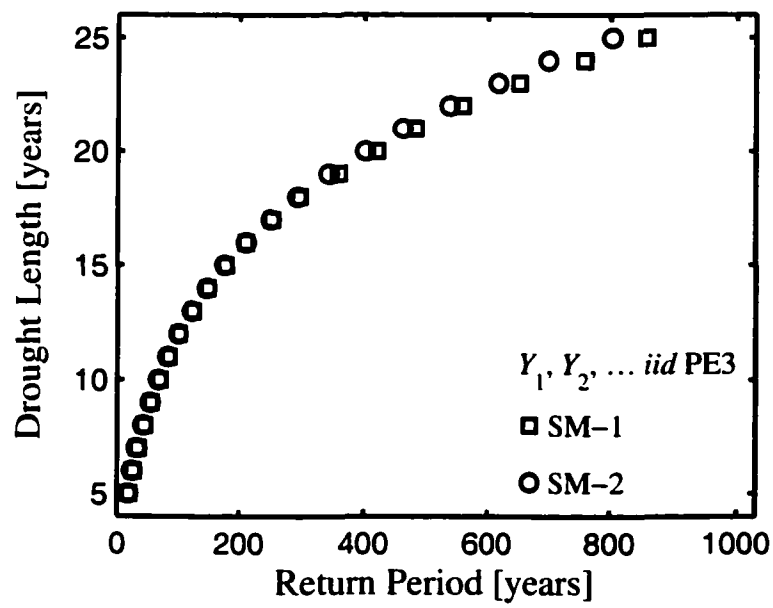

Figure 5.12: Return periods of droughts of various lengths for the Colorado River at Hot Sulphur Springs.

Chapter 6

SHIFTING MEAN PLUS PERSISTENCE MODEL FOR SIMULATING THE GREAT LAKES NET BASIN SUPPLIES

Abstract In current shifting mean models the autocorrelation structure is assumed to arise from the combination of sudden shifts in the mean level of the process under consideration and the time between such shifts. The objective of this study is to add direct persistence feature to the current shifting mean models. This is done by assuming that the underlying process can be represented by a shifting mean AR(1) model. In this study the applicability of the proposed model for simulating the annual net basin supplies (NBS) of the Great Lakes system is analyzed. The NBS of lakes Erie, Ontario, and St. Clair are autocorrelated and show sudden shifting behavior, and thus are successfully modeled by the proposed models. The other lakes, Michigan-Huron and Superior, do not show signs of sudden shifts and they do not appear to be autocorrelated.

6.1 Introduction

Many previous studies have shown that the stochastic characteristics of the Great Lakes Net Basin Supplies (NBS) are quite complex (see for example Yevjevich, 1975; Rassam et al., 1992; Buchberger, 1994) They include spatial and temporal variability with important high and low frequency components and possibly non-stationarity, in addition to the usual seasonality (periodicity) and variance-covariance properties. While determining some of those characteristics can be done by well-known stochastic techniques, the difficulty comes on how to interpret and justify them physically and statistically as not all NBS series show

similar behavior. For example, a feature that is apparent in some but not all the NBS series is the sudden or near sudden shifting pattern. It is well-known that significant changes in some key statistical parameters such as the mean, variability, skewness, and persistence may be the result of human induced interventions such as the construction of diversions dams or the regulation of natural lakes. Also such changes may be the result of climate variability, particularly the effect of low frequency components. The hypothesis that, "sudden changes in climate lasting for time spans of a few decades have naturally occurred in some parts of the globe," has become more plausible during the past years as newer hydroclimatic data such as sea surface temperature, precipitation, and streamflow appear to suggest.

A number of approaches has been suggested in the past decades by several groups in the United States and Canada for analyzing, modeling, and simulating the NBS series of the Great Lakes system. Well known models such as autoregressive (AR), AR with moving average terms (ARMA), and their multisite versions thereof have been utilized. Also alternative modeling schemes including single site, multisite, modeling at monthly/quarter-monthly time scales, and temporal disaggregation models have been used with various degrees of success (see for instance Yevjevich, 1975; Loucks, 1989; Buchberger, 1992; Rassam et al., 1992).

Hydro-Québec, conducted a major modeling and simulation effort in the period 1991-1992, in connection with the Beauharnois-Les Cedres extreme floods study (Rassam et al., 1992). The study involved modeling and simulation of the quarter monthly NBS series: (i) verifying that the simulated samples reproduced certain key NBS historical statistics, and (ii) routing such simulated NBS samples through the Great Lakes System model to obtain simulated lake levels and outflows and verifying that they resemble those corresponding to the routed historical NBS series. Three modeling schemes were compared, namely, a CARMA(1, 1) model with temporal disaggregation, a monthly-annual singular value decomposition model, and a multivariate model in which two sites were fitted by a simple shifting mean model and the rest of the sites fitted by an AR(1) model. The study showed

that simulating annual NBS that incorporates sudden shifts and disaggregating such annual values into monthly or smaller quantities gave quite good results.

The main objective of the study reported herein has been to investigate and develop a modeling scheme so as to incorporate a direct persistence feature in a shifting mean modeling framework. This new feature may offer some advantage in modeling and simulating annual NBS series. In this paper, we focus on a univariate model that will be useful for modeling and simulating individual series (sites) separately. The multivariate model, to enable modeling and simulating NBS series jointly, will be reported elsewhere. Specifically, a single site shifting mean model with AR(1) persistence, which is dubbed here as SMAR(1), is developed and applied to modeling an simulation of the annual NBS series of the Great Lakes system. Section 6.5 provides some illustrations of using the single site shifting mean models.

6.2 SMAR(1) : SM model with persistence in $\{Y_t\}$

A general definition of the shifting mean, SM, model is given by

$$X_t = Y_t + Z_t \tag{6.1}$$

where $\{X_t\}$ is a sequence of variables representing the process of interest. $\{Y_t\}$ is a stationary process with mean μ_Y, variance σ_Y^2 and autocorrelation function, ACF, $\rho_Y(\,)$ (zero at all lags in the regular SM model). $\{Z_t\}$ is a sequence with mean zero and variance σ_Z^2 and ACF $\rho_Z(\,)$. The Z_t's represent noise in the mean of the X_t process, that is characterized by levels.

In the shifting mean autoregressive of order 1, SMAR(1), the $\{Y_t\}$ is an AR(1) process with parameters $-1 < \phi < 1$ and σ_ϵ^2

$$Y_t - \mu_Y = \phi(Y_{t-1} - \mu_Y) + \varepsilon_t \tag{6.2}$$

where $\{\varepsilon_t\} \sim N(0, \sigma_\varepsilon^2)$, with $\sigma_\varepsilon^2 = \sigma_Y^2(1-\phi^2)$. The ACF of Y_t is

$$\rho_Y(h) = \phi^h \qquad h = 0, 1, \ldots \tag{6.3}$$

The noise level process $\{Z_t\}$ can be written as

$$Z_t = \sum_{i=1}^{t} M_i I_{(S_{i-1},S_i]}(t) \tag{6.4}$$

where $\{M_i\}_{i=1}^{\infty} \overset{iid}{\sim} N(0, \sigma_M^2 = \sigma_Z^2)$, $S_i = N_1 + N_2 + \ldots + N_i$ with $S_0 = 0$, and $I_{(a,b)}(t)$ is the indicator function equal to one if $t \in (a, b)$ and zero otherwise. The $\{N_i\}_{i=1}^{\infty}$ is a discrete, stationary, delayed-renewal sequence on the positive integers, such that $N_1, \{N_i\}_{i=2}^{\infty}$ are *iid* positive geometric variables with parameter p (Chapter 4). It follows that the ACF of Z_t is

$$\rho_Z(h) = (1-p)^h \qquad h = 0, 1, \ldots \tag{6.5}$$

Assuming that $\{Y_t\}$, $\{M_i\}$, and $\{N_i\}$ are mutually independent then the mean and the variance of X_t is

$$\mu_X = \mu_Y \tag{6.6}$$

and,

$$\sigma_X^2 = \sigma_Y^2 + \sigma_M^2 \tag{6.7}$$

Furthermore, the ACF of X_t is

$$\rho_h := \rho_X(h) = \frac{\sigma_M^2(1-p)^h + (\sigma_X^2 - \sigma_M^2)\phi^h}{\sigma_X^2} \qquad h = 0, 1, \ldots \tag{6.8}$$

6.2.1 Parameter Estimation : p Known

For modeling of a hydrologic or climatic process in a homogeneous (geographic) region, it may be assumed that p is common for all sites within the region. Given a regional estimate of p, noted as $\hat{p}$, the parameters $\{\mu_Y, \sigma_Y^2, \sigma_M^2, \phi\}$ at each site are estimated in terms of $\{\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\rho}_1, \hat{\rho}_2\}$ from Eqs (6.6)-(6.8). First ϕ is estimated from

$$\hat{\phi} = \frac{\hat{\rho}_1(1-\hat{p}) - \hat{\rho}_2}{1 - \hat{p} - \hat{\rho}_1} \tag{6.9}$$

and then

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 \frac{\hat{\rho}_1 - \hat{\phi}}{1 - \hat{p} - \hat{\phi}} , \tag{6.10}$$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\sigma}_M^2 , \text{ and} \tag{6.11}$$

$$\hat{\mu}_Y = \hat{\mu}_X \tag{6.12}$$

the estimated parameters are feasible only if $\{\hat{p}, \hat{\rho}_1, \hat{\rho}_2\}$ fulfill simultaneously

$$-1 < \frac{\hat{\rho}_1(1-\hat{p}) - \hat{\rho}_2}{1 - \hat{p} - \hat{\rho}_1} < 1 \tag{6.13}$$

and,

$$0 < \frac{\hat{\rho}_2 - \hat{\rho}_1^2}{(1-\hat{p})^2 - 2\hat{\rho}_1(1-\hat{p}) + \hat{\rho}_2} < 1 \tag{6.14}$$

6.2.2 Parameter Estimation : p Unknown

For p unknown the procedure for estimation of the parameters $\{\mu_Y, \sigma_Y^2, \sigma_M^2, p, \phi\}$ in terms of $\{\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3\}$ from Eqs (6.6)-(6.8) is as follows. Solve the quadratic equation

$$(\hat{\rho}_2 - \hat{\rho}_1^2)\hat{\phi}^2 + (\hat{\rho}_1\hat{\rho}_2 - \hat{\rho}_3)\hat{\phi} + \hat{\rho}_1\hat{\rho}_3 - \hat{\rho}_2^2 = 0 \tag{6.15}$$

for $\hat{\phi}$. Then estimates of p and σ_M^2 are obtained from

$$\hat{p} = 1 - \frac{\hat{\rho}_2 - \hat{\phi}\hat{\rho}_1}{\hat{\rho}_1 - \hat{\phi}} \tag{6.16}$$

and,

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 \frac{\hat{\rho}_2 - \hat{\phi}^2}{(1-\hat{p})^2 - \hat{\phi}^2} \tag{6.17}$$

At last σ_Y^2 and μ_Y are estimated from Eqs (6.11) and (6.12). Recall that the parameters estimated in Eqs (6.15)-(6.17) are considered feasible if $-1 < \hat{\phi} < 1$; $0 < \hat{p} < 1$; and $0 < \hat{\sigma}_M^2 < \hat{\sigma}_X^2$. Estimated parameters may be infeasible due to the sample variability of the sample ACF or due to other factors. In general if estimated parameters are feasible, then from the structure of the ACF in Eq (6.8) (see also Eq (6.19)) it follows that only

one set of parameters is feasible if $-1 < \hat{\phi} \leq 0$, but two sets of parameters are feasible if $0 < \hat{\phi} < 1$. In the latter case, if the first set of feasible parameter estimates is denoted by $\{\hat{\mu}_Y(1), \hat{\sigma}^2_Y(1), \hat{\sigma}^2_M(1), \hat{p}(1), \hat{\phi}(1)\}$, then the second set of feasible parameter estimates is given by $\hat{\mu}_Y(2) = \hat{\mu}_Y(1)$; $\hat{\sigma}^2_Y(2) = \hat{\sigma}^2_M(1)$; $\hat{\sigma}^2_M(2) = \hat{\sigma}^2_X - \hat{\sigma}^2_M(1)$; $\hat{p}(2) = 1 - \hat{\phi}(1)$; and $\hat{\phi}(2) = 1 - \hat{p}(1)$. Usually the values of the two parameter set are quite different with only one parameter set with values close to what would be expected of the process under consideration.

6.3 Smoothing the ACF of X_t

The estimation procedures in sections 6.2.1 and 6.2.2 can result in infeasible parameter estimates. The cause can often be related to sample variability of the observed ACF, or other characteristics of the observed ACF such as periodicities. A solution to this problem is to smooth the sample ACF, where the smoothed ACF would be used for parameter estimation in sections 6.2.1 and 6.2.2. The best approach may be to fit the sample ACF with a function that has similar characteristics as the model ACF in Eq (6.8). If the sample ACF is positive at all lags, one may be tempted to fit the sample ACF by the simple function

$$\rho_X(h) = ab^h \qquad h = 1, 2, \ldots \tag{6.18}$$

which represents a straight line in a log-log space for $a > 0$ and $b > 0$. Even though the function in Eq (6.18) may have somewhat similar decaying structure as the model ACF in Eq (6.8), the obvious drawback is that Eq (6.18) models exactly only a part of Eq (6.8). As a result for the general case in section 6.2.2, parameters estimated using a fitted ACF from Eq (6.18) would always result in either $\hat{p} = 1$ or $\hat{\phi} = 0$. To avoid these kind of situations, the sample ACF can be fitted by a function that has exactly the same form as the model ACF. Such function is given by

$$\rho_X(h) = ca^h + (1 - c)b^h \qquad h = 1, 2, \ldots \tag{6.19}$$

where $0 < a < 1$, $-1 < b < 1$, and $0 < c < 1$. If p is unknown then all parameters of Eq (6.19) have to estimated, while if p is known then only b and c in Eq (6.19) have to be estimated since $a = 1 - p$ would be known. In general least squares estimates of all three parameters using the sample ACF up to lag k are obtained by minimizing

$$S(a, b, c; \hat{\rho}_1, \ldots, \hat{\rho}_k) = \sum_{h=1}^{k} \left(\hat{\rho}_h - ca^h - (1-c)b^h\right)^2 \tag{6.20}$$

The sum in Eq (6.20) is minimized by taking the partial derivatives with respect to the unknown parameters, setting them equal to zero, and solving for the parameters. An alternative for estimation of the parameters is to use unconstrained or constrained optimization routines to minimize the sum of the squared errors in Eq (6.20). The first partial derivatives with respect to the parameters a, b, and c are given by

$$\frac{\partial S}{\partial a} = -2c\left[\sum_{h=1}^{k}(\hat{\rho}_h - b^h)ha^{h-1} - c\sum_{h=1}^{k}(a^h - b^h)ha^{h-1}\right] \tag{6.21}$$

$$\frac{\partial S}{\partial b} = -2(1-c)\left[\sum_{h=1}^{k}(\hat{\rho}_h - b^h)hb^{h-1} - c\sum_{h=1}^{k}(a^h - b^h)hb^{h-1}\right] \tag{6.22}$$

$$\frac{\partial S}{\partial c} = -2\left[\sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h) - c\sum_{h=1}^{k}(a^h - b^h)^2\right] \tag{6.23}$$

respectively. Setting Eq (6.23) to zero and solving for c

$$c = \frac{\sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h)}{\sum_{h=1}^{k}(a^h - b^h)^2} \quad , a \neq b \tag{6.24}$$

Similarly setting Eqs (6.21) and (6.22) equal to zero and substituting for c from Eq (6.24) then after simplification

$$0 = \sum_{h=1}^{k}(a^h - b^h)^2 \sum_{h=1}^{k}(\hat{\rho}_h - b^h)ha^{h-1} - \sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h)\sum_{h=1}^{k}(a^h - b^h)ha^{h-1} \tag{6.25}$$

$$0 = \sum_{h=1}^{k}(a^h - b^h)^2 \sum_{h=1}^{k}(\hat{\rho}_h - b^h)hb^{h-1} - \sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h)\sum_{h=1}^{k}(a^h - b^h)hb^{h-1} \tag{6.26}$$

The obvious solution $a = b$ is not feasible, since in that case c in Eq (6.24) is not defined. If p is known then $a = 1 - p$ is known, thus Eq (6.26) is solved numerically for b using for

example the bisection method for $b \in (-1, a) \cup (a, 1)$ to bracket the roots and then Newton-Rhapson to estimate the roots. On the other hand, if a is unknown, then Eqs (6.25) and (6.26) are solved numerically for a and b using for example a Newton-Rhapson iterative procedure with either exact or approximate Jacobian of Eqs (6.25) and (6.26). Finally c is estimated from Eq (6.24). The exact Jacobian of Eqs (6.25) and (6.26) is obtained from

$$\begin{aligned}\frac{\partial \text{Eq (6.25)}}{\partial a} =& \sum_{h=1}^{k}(a^h - b^h)ha^{h-1}\sum_{h=1}^{k}(\hat{\rho}_h - b^h)ha^{h-1} \\ &+ a^{-1}\sum_{h=1}^{k}(a^h - b^h)^2\sum_{h=1}^{k}(\hat{\rho}_h - b^h)h(h-1)a^{h-1} \\ &- a^{-1}\sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h)\sum_{h=1}^{k}ha^{h-1}[(2h-1)a^h - (h-1)b^h]\end{aligned} \tag{6.27}$$

$$\begin{aligned}\frac{\partial \text{Eq (6.25)}}{\partial b} =& \sum_{h=1}^{k}h^2a^{h-1}b^{h-1}\left[\sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h) - \sum_{h=1}^{k}(a^h - b^h)^2\right] \\ &+ \sum_{h=1}^{k}(a^h - b^h)hb^{h-1}\left[\sum_{h=1}^{k}(a^h - b^h)ha^{h-1} - 2\sum_{h=1}^{k}(\hat{\rho}_h - b^h)ha^{h-1}\right] \\ &+ \sum_{h=1}^{k}(\hat{\rho}_h - b^h)hb^{h-1}\sum_{h=1}^{k}(a^h - b^h)ha^{h-1}\end{aligned} \tag{6.28}$$

$$\begin{aligned}\frac{\partial \text{Eq (6.26)}}{\partial a} =& 2\sum_{h=1}^{k}(a^h - b^h)ha^{h-1}\sum_{h=1}^{k}(\hat{\rho}_h - b^h)hb^{h-1} \\ &- \sum_{h=1}^{k}(\hat{\rho}_h - b^h)ha^{h-1}\sum_{h=1}^{k}(a^h - b^h)hb^{h-1} \\ &- \sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h)\sum_{h=1}^{k}h^2a^{h-1}b^{h-1}\end{aligned} \tag{6.29}$$

and

$$\begin{aligned}\frac{\partial \text{Eq (6.26)}}{\partial b} =& \left[\sum_{h=1}^{k}(a^h - b^h)hb^{h-1}\right]^2 - \sum_{h=1}^{k}(a^h - b^h)hb^{h-1}\sum_{h=1}^{k}(\hat{\rho}_h - b^h)hb^{h-1} \\ &+ b^{-1}\sum_{h=1}^{k}(a^h - b^h)^2\sum_{h=1}^{k}hb^{h-1}[(h-1)\hat{\rho}_h - (2h-1)b^h] \\ &- b^{-1}\sum_{h=1}^{k}(\hat{\rho}_h - b^h)(a^h - b^h)\sum_{h=1}^{k}hb^{h-1}[(h-1)a^h - (2h-1)b^h]\end{aligned} \tag{6.30}$$

Initial values of a, b, and c are usually obtained from the sample ACF using the estimation procedures in sections 6.2.1 and 6.2.2. If the initial values are not feasible, then the sample

ACF can be tweaked a little bit until it results in feasible values for a, b, and c. A simple grid search on the parameters a, b, and c that minimizes Eq (6.20) can also be used to come up with initial values.

6.4 The Special Case : $\phi = 0$: The SM-1 Model

In the special case with $\phi = 0$ (no persistence in the Y_t process) the SMAR(1) model reduces to the SM-1 model in Chapter 4, with $\{Y_t\}_{t=1}^{\infty} \overset{iid}{\sim} N(0, \sigma_Y^2)$. Thus, the ACF of X_t in Eq (6.8) becomes

$$\rho_h := \rho_X(h) = \frac{\sigma_M^2(1-p)^h}{\sigma_X^2} \qquad h = 1, 2, \ldots \tag{6.31}$$

The model ACF in Eq (6.31) is simple, and its structural form for fitting purposes is given by

$$\rho_X(h) = ab^h \qquad h = 1, 2, \ldots \tag{6.32}$$

where $a = \sigma_M^2/\sigma_X^2$ and $b = 1 - p$, with $0 < a < 1$ and $0 < b < 1$. In Chapter 4 and 5 estimates of a and b were obtained by fitting a straight line to the logs of Eq (6.32). Often the sample ACF, $\hat{\rho}_h$, may have negative values, and consequently its values in the log-domain do not exist. To prevent such cases, then in this paper the least squares estimates of a and b based on the sample ACF up to lag k are obtained by minimizing

$$S(a, b; \hat{\rho}_1, \ldots, \hat{\rho}_k) = \sum_{h=1}^{k} \left(\hat{\rho}_h - ab^h\right)^2 \tag{6.33}$$

with respect to a and b. Setting the partial derivatives of Eq (6.33) with respect to a and b equal to zero and simplifying, then

$$0 = \sum_{h=1}^{k} b^{2h} \sum_{h=1}^{k} h\hat{\rho}_h b^{h-1} - \sum_{h=1}^{k} hb^{2h-1} \sum_{h=1}^{k} \hat{\rho}_h b^h \tag{6.34}$$

is first solved for b, and then a is calculated from

$$a = \frac{\sum_{h=1}^{k} \hat{\rho}_h b^h}{\sum_{h=1}^{k} b^{2h}} \tag{6.35}$$

6.4.1 Parameter Estimation : p Known

As stated in section 6.2.1 for modeling of a hydrologic or climatic process in a homogeneous (geographic) region, it may be assumed that p is common for all sites within the region. For fitting the ACF using Eq (6.32), $b = 1 - p$ is known, and a is estimated from Eq (6.35). Given a regional estimate $\hat{p}$, the parameters $\{\mu_Y, \sigma_Y^2, \sigma_M^2\}$ at each site are estimated in terms of $\{\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\rho}_1\}$ from Eqs (6.6), (6.7) and (6.31):

$$\hat{\sigma}_M^2 = \hat{\sigma}_X^2 \frac{\hat{\rho}_1}{1-\hat{p}} , \tag{6.36}$$

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \hat{\sigma}_M^2 , \text{ and} \tag{6.37}$$

$$\hat{\mu}_Y = \hat{\mu}_X \tag{6.38}$$

the estimated parameters are feasible if $\hat{\rho}_1 < 1 - \hat{p}$.

6.4.2 Parameter Estimation : p Unknown

For p unknown the parameters $\{\mu_Y, \sigma_Y^2, \sigma_M^2, p\}$ are estimated in terms of $\{\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\rho}_1, \hat{\rho}_2\}$ from Eqs (6.6), (6.7) and (6.31). First p is estimated from

$$\hat{p} = 1 - \frac{\hat{\rho}_2}{\hat{\rho}_1} \tag{6.39}$$

and then σ_M^2, σ_Y^2, and μ_Y are estimated from Eqs (6.36), (6.37) and (6.38), respectively. The parameter estimates are feasible if $\hat{\rho}_1^2 < \hat{\rho}_2 < \hat{\rho}_1$. For fitting the ACF using Eq (6.32), then b and a are estimated from Eqs (6.34) and (6.35), respectively.

6.5 Example : The Great Lakes System

In this section we apply the SMAR(1) and SM-1 models to the annual net basin supplies (NBS) of the lakes in the Great Lakes system. The data were obtained from Hydro-Québec, and span the period 1900–1999 for lakes Erie, Michigan-Huron, Ontario, and Superior, and the period 1900–1989 for Lake St. Clair. Data post 1989 for Lake St. Clair were still preliminary, and hence are not used in this study. The annual NBS time

series of the Great Lakes and their ACFs can be seen in Figs. 6.1–6.5. The data for Lake Superior and Lake Michigan-Huron in Figs. 6.2 and 6.5 do not seem to exhibit any sudden shifts, and in addition the ACFs of the these data do not have shapes that are expected of the SMAR(1) or the SM-1 model. On the other hand, the data for the other lakes in Figs. 6.1, 6.3, and 6.4 appear to be characterized by sudden shifts.

Throughout this paper, for a time series $X_1, X_2, \ldots, X_n$ the sample mean, variance, skewness, and lag h ACF will be estimated by

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6.40}$$

$$\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2 \tag{6.41}$$

$$\hat{\gamma}_X = \frac{n\hat{\sigma}_X^{-3}}{(n-1)(n-2)}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^3 \tag{6.42}$$

$$\hat{\rho}_X(h) = \frac{\sum_{i=1}^{n-h}(X_{i+h} - \hat{\mu}_X)(X_i - \hat{\mu}_X)}{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2} \tag{6.43}$$

respectively. In terms of storage related statistics we will use the Hurst slope K and the storage capacity SC. The Hurst slope K is calculated by forming a series of partial sums $S_i = S_{i-1} + X_i - \hat{\mu}_X$, $i = 1, \ldots, n$, with $S_0 = 0$. Then the range (R_n^*), the rescaled range (R_n^{**}) and the K are calculated from

$$R_n^* = \max_{0\le i\le n}(S_i) - \min_{0\le i\le n}(S_i) \tag{6.44}$$

$$R_n^{**} = R_n^*/\hat{\sigma}_X \tag{6.45}$$

$$K = \ln R_n^{**} / \ln(0.5\,n) \tag{6.46}$$

respectively. In general we will only show calculated values of K, since R_n^* and R_n^{**} can be easily calculated given the additional knowledge of n and $\hat{\sigma}_X$. The storage capacity for a demand level d is calculated from

$$SC = \max_{1\le i\le n} S_i' \tag{6.47}$$

where $S_i' = \max(0, S_{i-1}' + d - X_i)$ with $S_0' = 0$. Furthermore the longest drought length DL corresponding to a demand level d is the longest consecutive period for which $X_i < d$, and

the corresponding magnitude DM of the longest drought is the sum over the deficits $d - X_i$ during the drought.

The sample mean, standard deviation, skewness, Hurst slope, storage capacity, and the longest drought length and the corresponding drought magnitude based on demand level $d = \hat{\mu}_X$ of the Great Lakes data are shown in Table 6.1. The values of the sample ACF for lags 1 to 3, and the values of the fitted ACFs for the SMAR(1) and SM-1 models are also shown in Table 6.1. The variables of the fitted ACFs were estimated using the sample ACF up to lag 15 for all lakes.

Also in Table 6.1 the values of the sample and the fitted ACFs are shown for lags 1 to 3.

6.5.1 Modeling of the Great Lakes System

Table 6.2 shows feasible parameter sets of fitted SMAR(1) and SM-1 models for the Great Lakes system. The parameters are estimated using the sample statistics of the observed data $\hat{\mu}_X$, $\hat{\sigma}_X$, and the values of the fitted ACF $\hat{\rho}_1$, $\hat{\rho}_2$, and $\hat{\rho}_3$ based on Eqs (6.19) and (6.32) for the SMAR(1) and the SM-1 models, respectively. Parameter estimates could not be obtained for Lake Michigan-Huron for the both the SMAR(1) and SM-1 models, and for Lake Superior for the SM-1 model (refer to Table 6.2), the reason being that Eqs (6.20) and/or (6.33) were minimized with one or more of their parameters at the boundaries of the parameters space or outside of the parameter space. For Lake St. Clair in Table 6.2, the two feasible parameter sets of the SMAR(1) model are quite different, where the parameter set in parenthesis appears less realistic. Generated sequences based on the fitted SMAR(1) models in Table 6.2 are plotted in Figs. 6.6–6.9. Notice the differences in the generated sequences for Lake St. Clair between Fig. 6.8 (a) and Fig. 6.8 (b), where Fig. 6.8 (b) is based on what we classify as the less realistic parameter set in Table 6.2.

To investigate how well the fitted models preserve the sample statistics used in the fitting procedure 1,000 realizations of the Great Lakes NBS of the same length as the his-

torical record (n) were generated based on the models with the non-parenthesized values in Table 6.2. The average statistics of the 1,000 realization are shown in Table 6.3. Comparing these statistics with the corresponding sample statistics in Table 6.1, it can be concluded that the mean and the standard deviation are well preserved in all cases. The model skewness is zero for all fitted models as reflected in Table 6.3. The Hurst slope K, the storage capacity SC, and the maximum drought length DL and the drought deficit DM are remarkably well preserved considering that these parameters are not used in the fitting of the models. On the other hand the average ACF of the generated series appears to somewhat underestimate the fitted ACF in Table 6.1. This underestimation of the ACF is often encountered when the ACF is estimated based on the average of a large number of generated samples of relatively small sizes. For further comparison the same statistics based on one generated series of length $1,000$ n are shown in Table 6.4, and the first 2,000 observations of the generated sequences are shown in Fig. 6.10. The estimates of the standard deviation and the ACF are improved compared to Table 6.4. On the other hand it should be expected to get different values for SC, DL, and DM since these statistics are highly dependent on the sample size. Notice that the values of the Hurst slope K in Table 6.4 are considerably lower than the corresponding values in Tables 6.1 and 6.3. A similar behavior was noticed in Boes and Salas (1978), where asymptotically K was shown to approach the value 0.5 for the shifting mean models studied there and for the ARMA(1,1) model. For graphical comparison of the estimated ACFs, correlograms based on 1,000 generated samples of length n and based on one sample of length $1,000$ n are shown in Fig. 6.11 for Lake Ontario and Lake St. Clair. In general the correlograms are well preserved, but as often is the case when average correlograms are estimated based on a number of generated samples of small sizes the average correlogram for St. Clair based on 1,000 generated samples underestimates the model correlogram.

For comparing the two different model, the SMAR(1) model and the SM-1 model (SMAR(1) with $\phi = 0$), then the focus should be on statistics that were not used in fitting

of these models. These statistics are the Hurst slope K, the storage capacity SC, and the maximum drought length DL and the drought deficit DM. Comparing these statistics from Tables 6.3 and 6.4 across the two different models for the lakes Erie, Ontario, and St. Clair, then in general it can be concluded that results of the two models are very similar. One reason for the similarity of the results, might be the that the AR(1) parameters ϕ are close to zero for the fitted SMAR(1) models for the referred lakes in Table 6.2.

6.6 Summary and Final Remarks

In the shifting mean models studied in Chapter 4 and 5 the autocorrelation structure is assumed to arise solely from sudden shifting pattern in the mean level of the process under consideration. That is, the underlying process is assumed to be random if the shifting pattern is removed. The main objective of this paper was to add a persistence feature to existing shifting mean models in such a way that, the underlying process with the shifting pattern removed is an autoregressive AR(1) process. This proposed model, dubbed as SMAR(1), reduces to the SM-1 model in Chapter 4 in the special case when the autoregressive parameter $\phi = 0$. Parameter procedures were developed for the SMAR(1) model for both the cases when the expected length of the mean levels, p, is assumed to be known or unknown. Where, for a geographic region p should be taken as common for all sites, if the shifts at different sites are driven by external processes, such as climate changes, that should affect the whole region.

Historical records of lakes Erie, Ontario, and St. Clair in the Great Lakes system show evidence of sudden shifts in addition to autocorrelation. The time series for these lakes were fitted by the SMAR(1) model, and for comparison also by the SM-1 model. Both models were capable of preserving the key statistics (mean, variance, and autocorrelation) that are used in estimation of the parameters parameter. In addition the models preserved remarkably well other statistics, such as the the Hurst slope K, the storage capacity SC, and the maximum drought length DL and corresponding drought deficit DM, that are not used

in estimation of the parameters, but these statistics are affected by the expected length and the variance of the mean levels. In addition the results based on the two different SMAR(1) and SM-1 model appeared to be very similar for the cases studied, but in all those cases the ϕ parameter of the SMAR(1) model was close to zero.

As a general conclusion regarding to the Great Lakes system, the lakes Erie, Ontario, and St. Clair appear to be well presented by both the SMAR(1) model and the SM-1 model, while the lakes Michigan-Huron and Superior do not seem to exhibit shifting pattern nor to be autocorrelated.

Table 6.1: Sample statistics of the Great Lakes NBS time series from 1900–1999, except for Lake St. Clair where the statistics correspond to the period 1900-1989. Fitted ACFs up to lag 3 for the SMAR(1) model and the SM-1 model are also shown.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
$\hat{\mu}_X$[cms]	574.1	3177	1033	121.7	2043
$\hat{\sigma}_X$[cms]	265.4	737.0	241.6	63.34	478.8
$\hat{\gamma}_X$	0.138	-0.091	0.491	0.311	0.033
K	0.787	0.713	0.786	0.847	0.654
SC[cms]	5506	11978	5083	1529	4755
DL	8	8	11	9	5
DM[cms]	1720	6029	2034	659.3	3850
Sample ACF					
$\hat{\rho}_1$	0.173	0.168	0.250	0.504	0.153
$\hat{\rho}_2$	0.170	0.003	0.212	0.291	0.050
$\hat{\rho}_3$	0.175	-0.088	0.199	0.236	-0.023
Fitted ACF for SMAR(1) Model					
$\hat{\rho}_1$	0.175	*	0.263	0.484	0.149
$\hat{\rho}_2$	0.192	*	0.266	0.362	0.050
$\hat{\rho}_3$	0.173	*	0.199	0.300	-0.003
Fitted ACF for SM-1 Model ($\phi = 0$)					
$\hat{\rho}_1$	0.197	**	0.298	0.456	**
$\hat{\rho}_2$	0.181	**	0.236	0.379	**
$\hat{\rho}_3$	0.166	**	0.187	0.315	**

* Optimal solution, for which Eq (6.20) is minimized, is infeasible.
** Optimal solution, for which Eq (6.33) is minimized, is infeasible.

Table 6.2: Estimated parameters of fitted SMAR(1) and SM-1 qmodels to the Great Lake system using the sample mean and variance, and the fitted ACF.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
Parameters for SMAR(1) Model					
$\hat{\phi}$	-0.0440	*	-0.1225	0.1264 (0.8427)	-0.5528
$\hat{p}$	0.0899	*	0.2249	0.1573 (0.8736)	0.8107
$\hat{\sigma}_M$[cms]	127.3	*	158.3	44.74 (44.84)	465.7
$\hat{\mu}_Y$[cms]	574.1	*	1033	121.7 (121.7)	2043
$\hat{\sigma}_Y$[cms]	232.9	*	182.5	44.84 (44.74)	111.3
Parameters for SM-1 Model ($\phi = 0$)					
$\hat{p}$	0.0827	**	0.2091	0.1693	**
$\hat{\sigma}_M$[cms]	123.7	**	149.1	47.22	**
$\hat{\mu}_Y$[cms]	574.1	**	1033	121.7	**
$\hat{\sigma}_Y$[cms]	236.4	**	191.6	42.75	**

* Optimal solution, for which Eq (6.20) is minimized, is infeasible.
** Optimal solution, for which Eq (6.33) is minimized, is infeasible.

Table 6.3: Average sample statistics of 1,000 generated NBS time series of the Great Lakes. Each generated time series is of the same length as the corresponding historical record. The model parameters are the non-parenthesized values in Table 6.2.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
SMAR(1) Model					
$\hat{\mu}_X$[cms]	574.4	*	1033	121.8	2046
$\hat{\sigma}_X$[cms]	257.8	*	237.9	61.37	473.9
$\hat{\gamma}_X$	-0.011	*	-0.008	0.000	0.000
K	0.726	*	0.733	0.781	0.653
SC[cms]	4955	*	4385	1293	6081
DL	7.855	*	9.175	10.713	6.815
DM[cms]	2085	*	2322	735.3	3129
$\hat{\rho}_1$	0.120	*	0.221	0.417	0.132
$\hat{\rho}_2$	0.137	*	0.224	0.287	0.038
$\hat{\rho}_3$	0.115	*	0.149	0.220	-0.017
SM-1 Model ($\phi = 0$)					
$\hat{\mu}_X$[cms]	575.7	**	1034	121.6	**
$\hat{\sigma}_X$[cms]	260.0	**	237.8	61.97	**
$\hat{\gamma}_X$	0.001	**	-0.009	0.002	**
K	0.725	**	0.732	0.778	**
SC[cms]	4966	**	4387	1309	**
DL	8.028	**	8.921	10.677	**
DM[cms]	2133	**	2241	738.6	**
$\hat{\rho}_1$	0.142	**	0.254	0.387	**
$\hat{\rho}_2$	0.128	**	0.189	0.303	**
$\hat{\rho}_3$	0.105	**	0.138	0.236	**

* Optimal solution, for which Eq (6.20) is minimized, is infeasible.

** Optimal solution, for which Eq (6.33) is minimized, is infeasible.

Table 6.4: Sample statistics of 1 generated NBS time series of the Great Lakes of lengths $1,000\ n$, where n is the length of the historical records. The model parameters are the non-parenthesized values in Table 6.2.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
SMAR(1) Model					
$\hat{\mu}_X$[cms]	576.7	*	1036	121.6	2042
$\hat{\sigma}_X$[cms]	264.6	*	242.5	63.05	479.4
$\hat{\gamma}_X$	0.003	*	-0.001	-0.003	-0.001
K	0.620	*	0.631	0.594	0.589
SC[cms]	143098	*	83385	35694	337047
DL	37	*	32	67	25
DM[cms]	16113	*	10065	5481	12757
$\hat{\rho}_1$	0.169	*	0.260	0.479	0.148
$\hat{\rho}_2$	0.190	*	0.261	0.360	0.057
$\hat{\rho}_3$	0.166	*	0.194	0.300	-0.001
SM-1 Model ($\phi = 0$)					
$\hat{\mu}_X$[cms]	579.1	**	1033	122.3	**
$\hat{\sigma}_X$[cms]	265.9	**	241.2	63.26	**
$\hat{\gamma}_X$	0.006	**	0.012	0.012	**
K	0.649	**	0.593	0.627	**
SC[cms]	148914	**	123846	30623	**
DL	31	**	35	55	**
DM[cms]	11177	**	14670	5089	**
$\hat{\rho}_1$	0.203	**	0.297	0.454	**
$\hat{\rho}_2$	0.184	**	0.234	0.379	**
$\hat{\rho}_3$	0.170	**	0.186	0.312	**

* Optimal solution, for which Eq (6.20) is minimized, is infeasible.
** Optimal solution, for which Eq (6.33) is minimized, is infeasible.

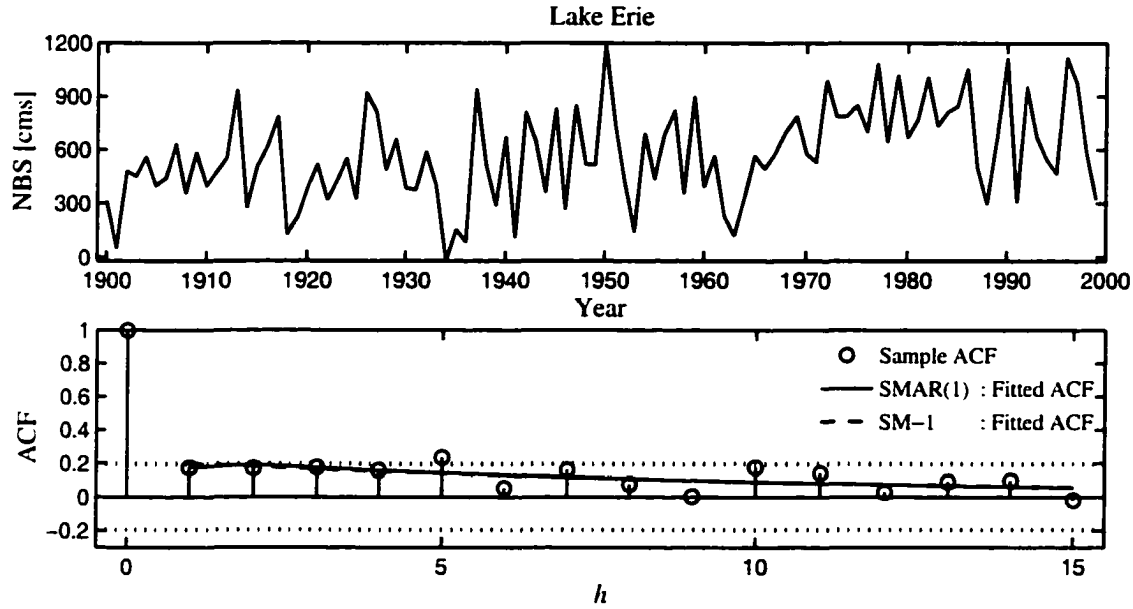

Figure 6.1: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Erie.

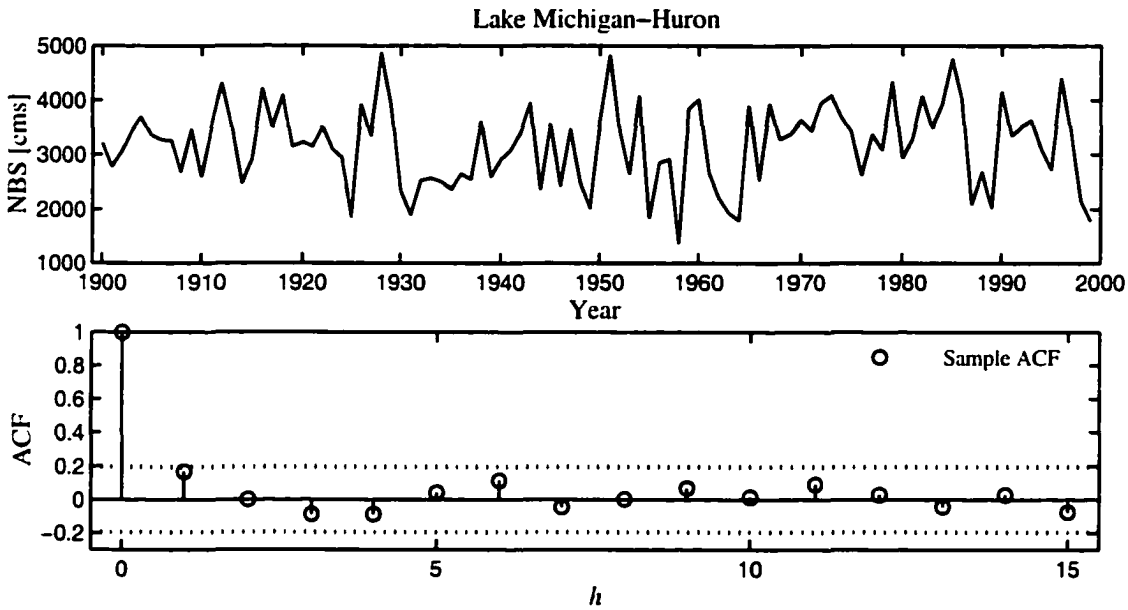

Figure 6.2: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Michigan-Huron.

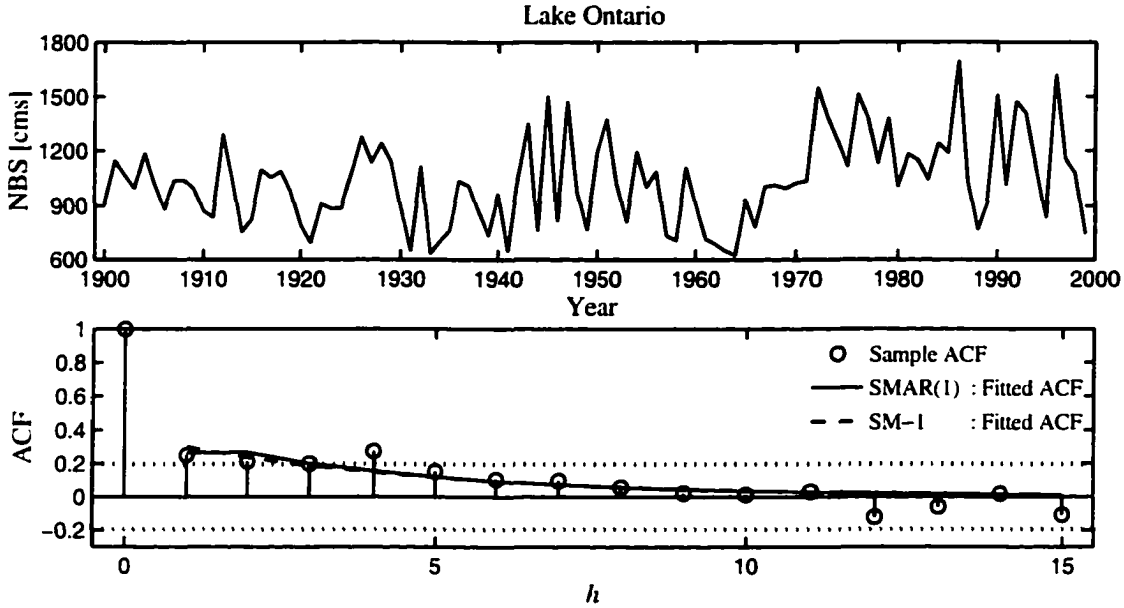

Figure 6.3: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Ontario.

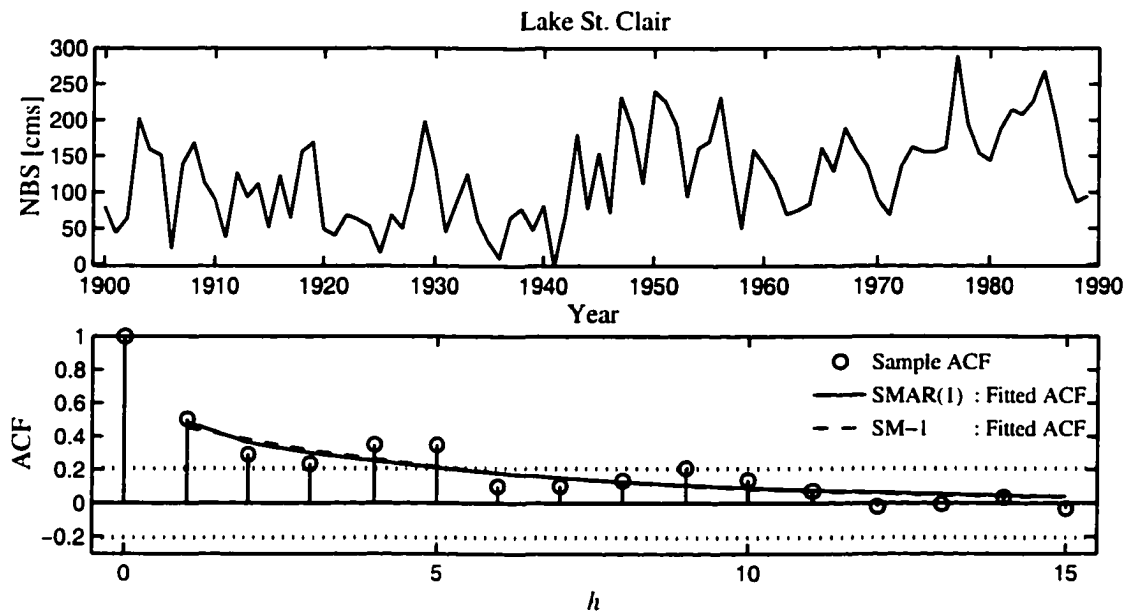

Figure 6.4: Net Basin Supply series (1900–1989) and the autocorrelation function for Lake St. Clair.

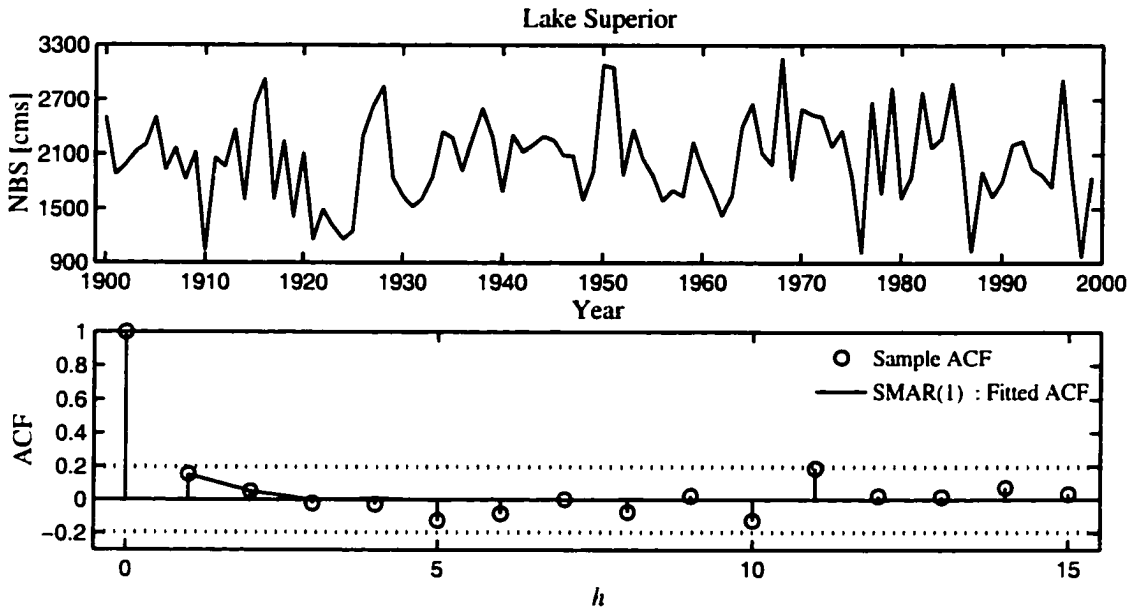

Figure 6.5: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Superior.

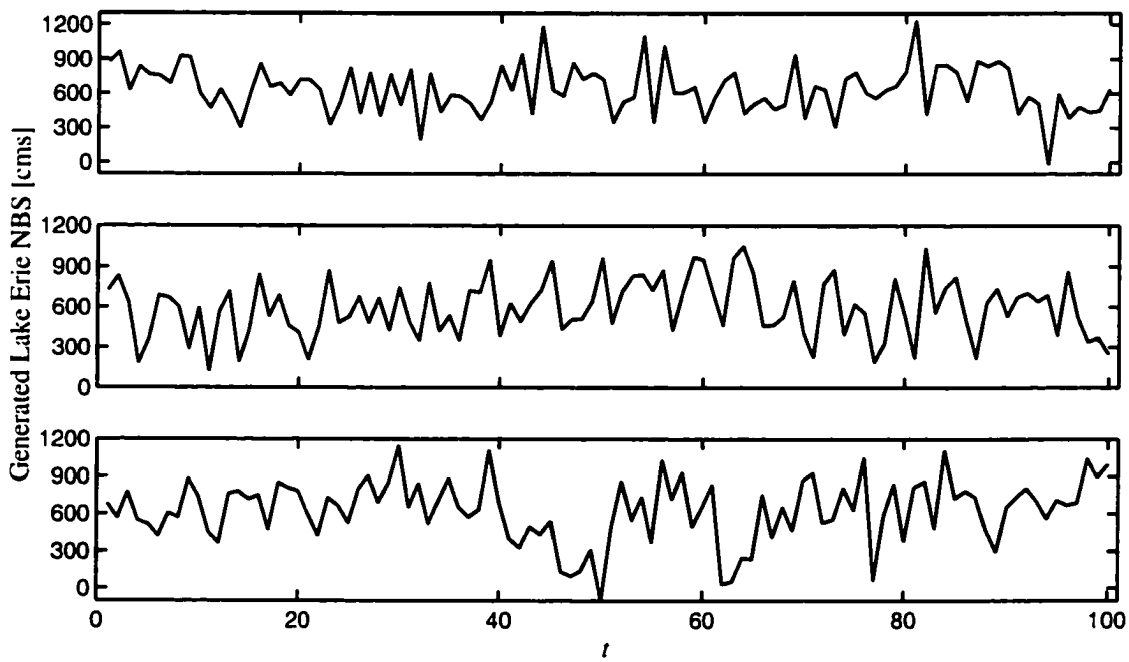

Figure 6.6: Generated sequences of annual NBS for Lake Erie based on the fitted SMAR(1) model in Table 6.2.

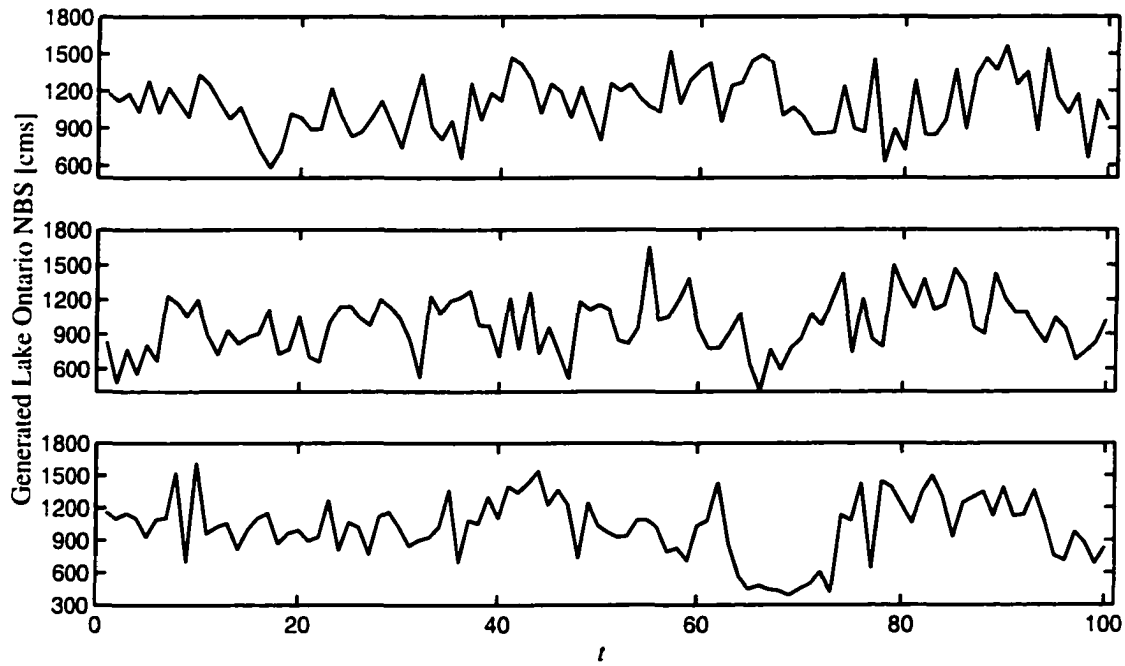

Figure 6.7: Generated sequences of annual NBS for Lake Ontario based on the fitted SMAR(1) model in Table 6.2.

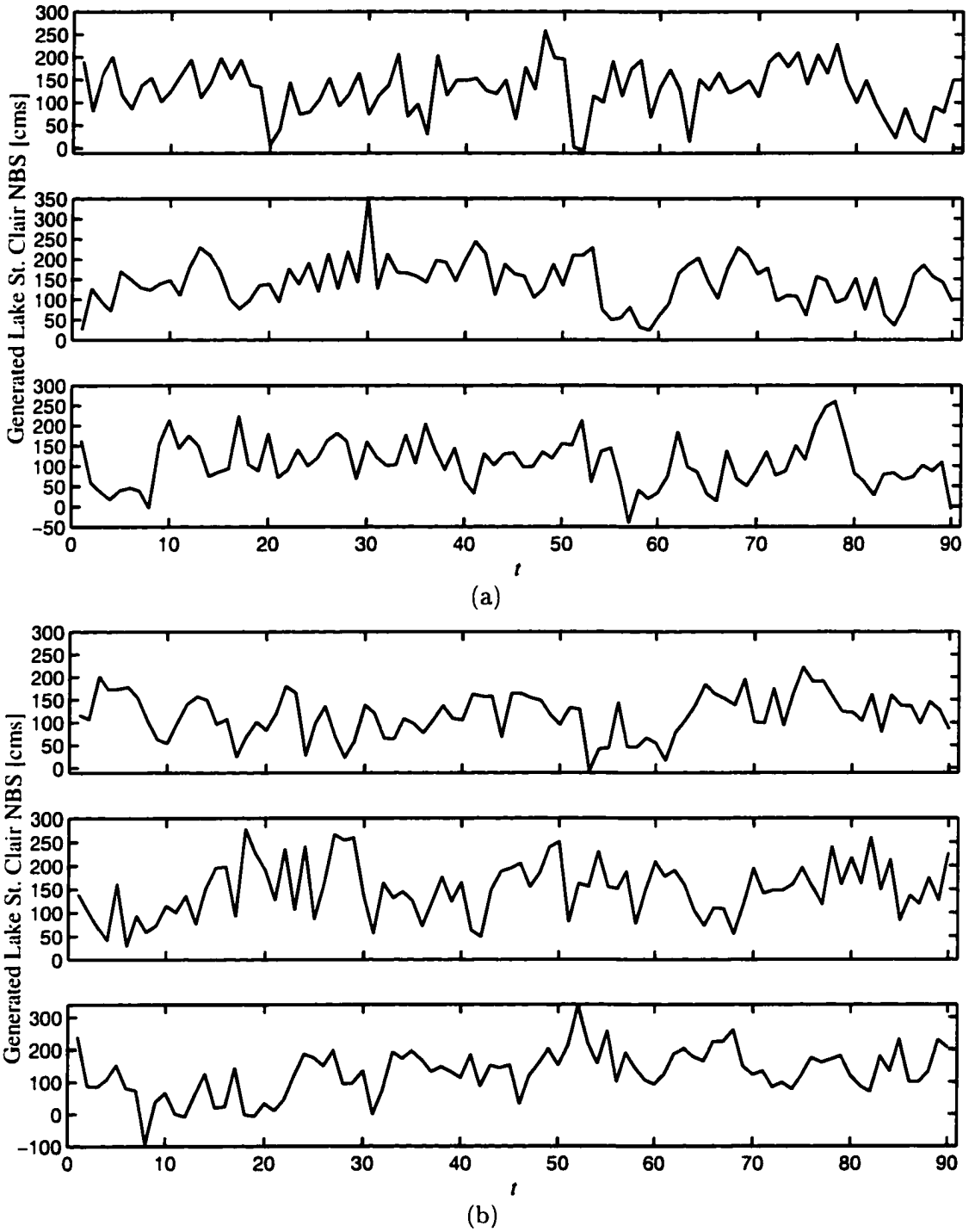

Figure 6.8: Generated sequences of annual NBS for Lake St. Clair based on the fitted SMAR(1) model in Table 6.2. (a) parameter values not in parenthesis in Table 6.2, and (b) parameter values in parenthesis in Table 6.2.

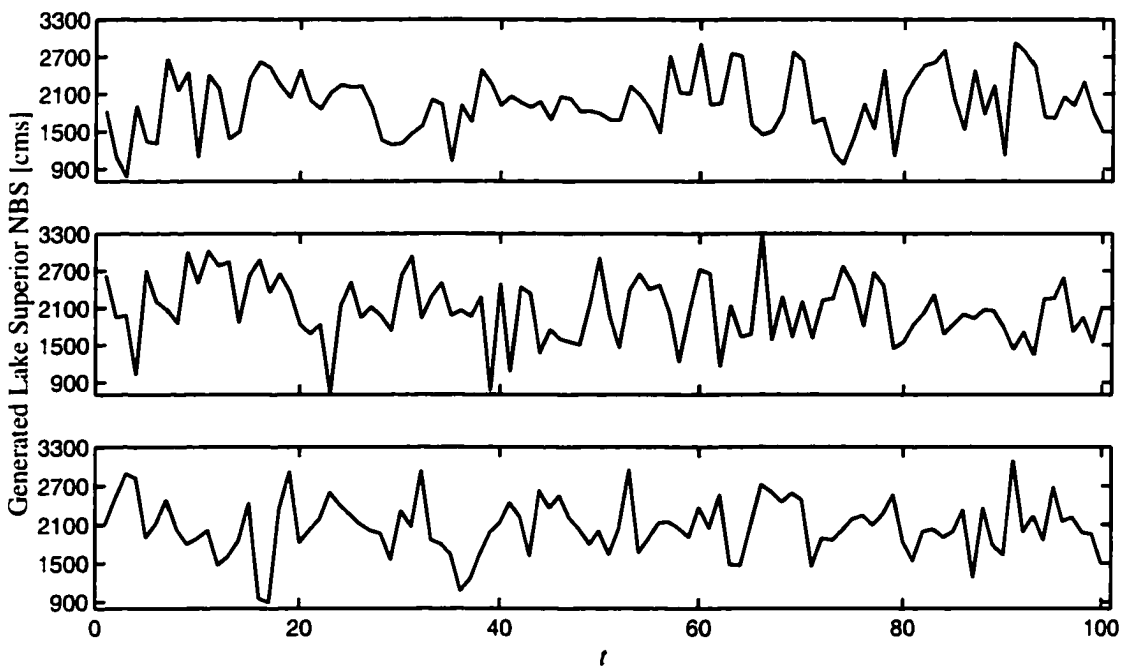

Figure 6.9: Generated sequences of annual NBS for Lake Superior based on the fitted SMAR(1) model in Table 6.2.

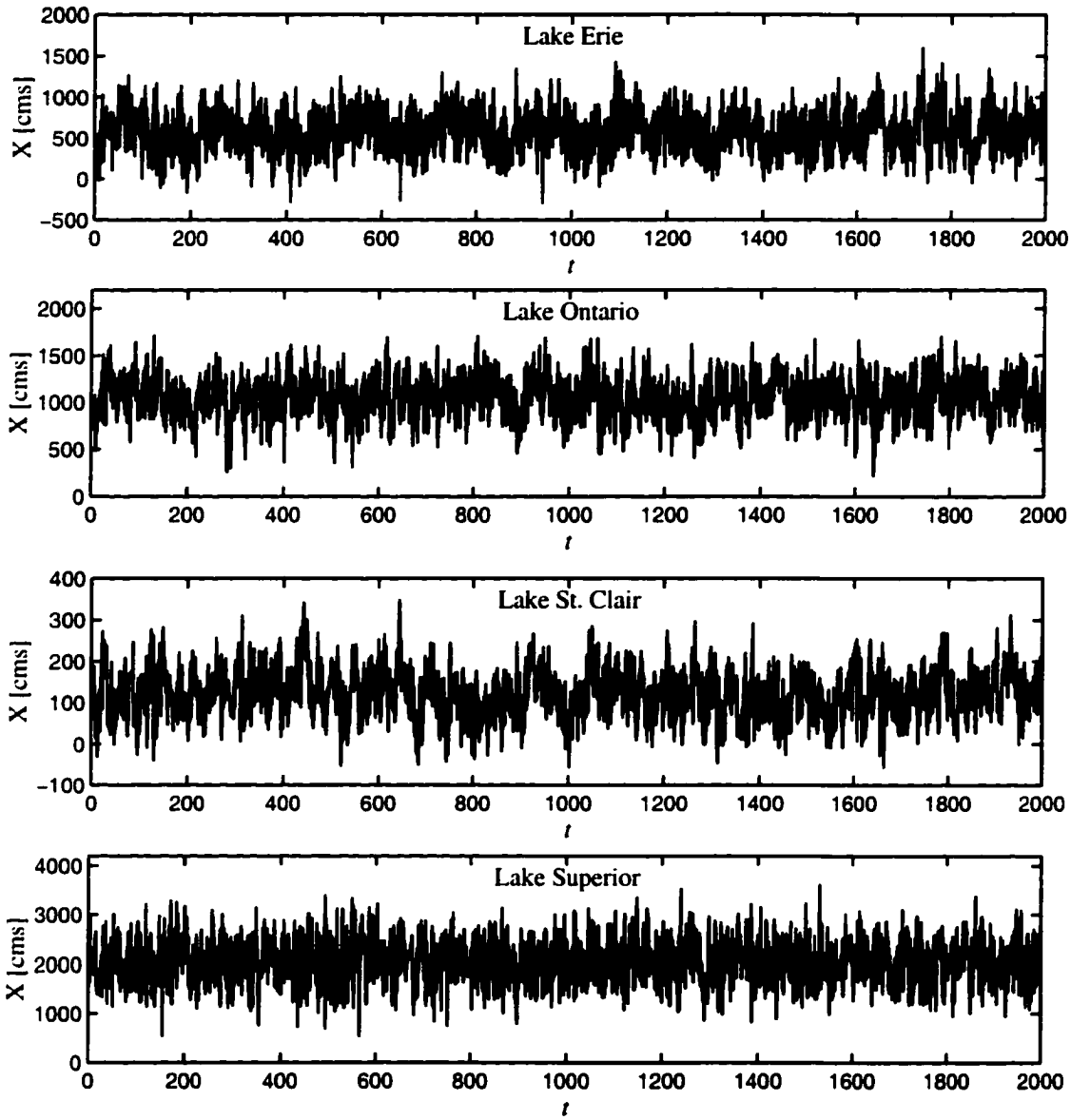

Figure 6.10: First 2,000 observations of generated sequences of length 1,000 n based on the fitted SMAR(1) models with non-parenthesized parameters values in Table 6.2.

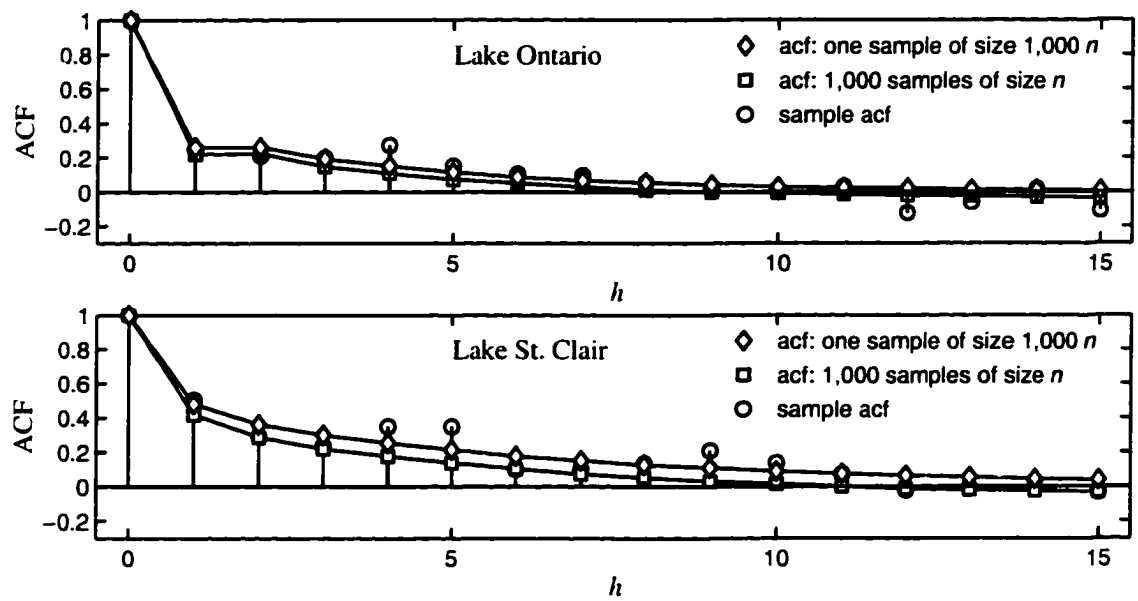

Figure 6.11: Correlograms of generated sequences for Lake Ontario and Lake St. Clair based on the fitted SMAR(1) models with non-parenthesized parameters values in Table 6.2. For each model a correlogram is estimated based on one generated sample of size 1,000 n, and based on averaging the acf's of 1,000 generated samples of the same size as the historical record (n).

Chapter 7

MULTIVARIATE SHIFTING MEAN PLUS PERSISTENCE MODEL FOR SIMULATING THE GREAT LAKES NET BASIN SUPPLIES

Abstract The focus of this paper is to develop a multivariate model to model the net basin supplies (NBS) of the Great Lakes. Not all NBS series show similar behavior. For example, a feature that is apparent in some but not all NBS series is a sudden shifting pattern. In this paper we expand previous studies of univariate shifting mean models to develop contemporaneous shifting mean models. These multivariate models are further mixed with CARMA models in such a way, that the lag zero correlation in space is conserved between the underlying processes of the different models. The full contemporaneous shifting mean CARMA models are successfully applied for modeling jointly the whole Great Lakes system, preserving the spatial correlation at lag zero between different lakes, and preserving other important statistical characteristics of the individual lakes.

7.1 Introduction

The Great Lakes System is one of the major lake systems in the world. It involves a series of five interconnected lakes (Superior, Michigan-Huron, St. Clair, Erie, and Ontario) that are subject to inter-basin flows and net basin supplies (NBS). Lake St. Clair is small compared to the other four lakes but being the middle lake, it is strategically located. Lakes Superior and Ontario have been regulated for the past several decades while the intermediate lakes are not regulated, although modifications in the connecting channels have caused some effect on the lake outflows (Quinn, 1985). Regulation of the two lakes

depends on the expected NBS. In addition, the regulation of Lake Ontario, being the furthest downstream lake of the system, depends on the characteristics of the entire system, such as the expected NBS for all the lakes, the corresponding lake levels, and outflows. Thus the analysis, modeling and simulation of the NBS series for the various lakes have been of interest not only for testing alternative regulation plans but for re-evaluating the capacity of existing waterworks, re-examining the performance of existing water systems, and assessing the capacity of new water resources systems.

Several studies have been made in the past for analyzing and modeling the NBS series of the entire Great Lakes system based on stochastic techniques. The NBS time series show complex patterns that are reflected in some of the statistical characteristics such as the mean, variance, persistence, high flow and low-flow statistics, short and long memory, and shifting level behavior (Rassam et al., 1992). In addition, monthly and quarter-monthly data show periodic basic statistics such as mean, variance, skewness, and correlations (Yevjevich, 1975). Some studies have been made attempting to understand and model some of the stochastic features of the NBS series. For example, Buchberger (1994) used a conceptual analysis based on water balance of the lakes to derive covariance properties of the annual NBS series.

Direct and indirect modeling schemes have been proposed and applied for modeling monthly and quarter monthly NBS series (Yevjevich, 1975; Loucks, 1989; Rassam et al., 1992; Buchberger, 1992). Direct modeling schemes imply using (for instance) monthly data and building a model to simulate monthly data directly at this time scale. For example, Yevjevich (1975) used a multivariate autoregressive (AR) model after seasonally standardizing the NBS series. The drawback with this type of modeling scheme is that while the monthly statistics are generally well preserved, statistics at higher time scales (for example years) are generally underestimated. Likewise, other statistics related to low frequency components such as random apparent shifts in the series are not represented. On the other hand, indirect modeling schemes imply modeling and generating monthly NBS in two or more

steps (stages), that is firstly the time series is modeled at a higher time scale such as years so as to reproduce key annual statistics, subsequently annual NBS series generated from such a model are then disaggregated into smaller time scales such as months in such a way as to reproduce monthly statistics. For example, Rassam et al. (1992) employed two indirect modeling and generation schemes by using the so-called SPIGOT computer package (Grygier and Stedinger, 1990).

Rassam et al. (1992) compared three multivariate modeling schemes two of which fall in the category of indirect approach as described above. The two indirect approaches included, a CARMA(1, 1) model with temporal disaggregation, and a mixture of multivariate AR(1) and shifting mean model with temporal disaggregation. The shifting mean model was included for generating the annual NBS series of Lakes Erie and Ontario because it was capable of reproducing the relevant statistics related to lake levels and outflows better than the other alternatives. This study also suggested the need of further developing multisite shifting mean models.

In this paper we develop a multivariate modeling framework using shifting mean models. Two contemporaneous models are developed, namely: the contemporaneous shifting mean model plus AR(1) persistence dubbed as CSMAR(1) and a mixture of contemporaneous shifting mean and a contemporaneous ARMA, dubbed CSMAR(1)-CARMA. The CSMAR(1) model is based on the single site plus persistence model, SMAR(1), suggested in Chapter 6. In addition, simpler versions of the models assuming no direct AR(1) persistence are included. The various models are illustrated and compared using the NBS data of the Great Lakes system.

7.2 Contemporaneous SMAR(1) : CSMAR(1)

The CSMAR(1) model is a contemporaneous SMAR(1) model that can be used to model multiple time series that are correlated in space. For detailed description of the SMAR(1) model refer to Chapter 6. If $\mathbf{X}_t = [X_t^{(1)}\ X_t^{(2)}\ \ldots\ X_t^{(n)}]^{\mathrm{T}}$ is a column vector of

observations at time t for n different sites, where each site is assumed to follow a SMAR(1) process, then the CSMAR(1) process can be expressed as

$$\mathbf{X}_t = \mathbf{Y}_t + \mathbf{Z}_t \tag{7.1}$$

where $\mathbf{Y}_t$ and $\mathbf{Z}_t$ are column vectors defined in the same way as $\mathbf{X}_t$. For a single site the noise level process $\{Z_t\}$ can be written as

$$Z_t = \sum_{i=1}^{t} M_i I_{(S_{i-1},S_i]}(t) \tag{7.2}$$

where $\{M_i\}_{i=1}^{\infty} \overset{iid}{\sim} N(0, \sigma_M^2 = \sigma_Z^2)$, $S_i = N_1 + N_2 + \ldots + N_i$ with $S_0 = 0$, and $I_{(a,b)}(t)$ is the indicator function equal to one if $t \in (a,b)$ and zero otherwise. The $\{N_i\}_{i=1}^{\infty}$ is a discrete, stationary, delayed-renewal sequence on the positive integers, such that $N_1, \{N_i\}_{i=2}^{\infty}$ are *iid* positive geometric variables with parameter p (Chapter 6). The cross covariance function (CCVF) of $\{\mathbf{X}_t\}$ at lag h is denoted by

$$\mathbf{C}_{\mathbf{X}}(h) = E[(\mathbf{X}_{t+h} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}_t - \boldsymbol{\mu}_{\mathbf{X}})^{\mathrm{T}}] = \begin{bmatrix} c_{\mathbf{X}}^{11}(h) & \ldots & c_{\mathbf{X}}^{1n}(h) \\ \vdots & \ddots & \vdots \\ c_{\mathbf{X}}^{n1}(h) & \ldots & c_{\mathbf{X}}^{nn}(h) \end{bmatrix} \tag{7.3}$$

where $c_{\mathbf{X}}^{ij}(h) = E[(X_{t+h}^{(i)} - \mu_X^{(i)})(X_t^{(j)} - \mu_X^{(j)})]$ is the CCVF at lag h between site i and site j, and $\boldsymbol{\mu}_{\mathbf{X}}$ is the mean vector of $\mathbf{X}_t$. In the CSMAR(1) model the following assumptions are made about the independent sequences $\{\mathbf{Y}_t\}$ and $\{\mathbf{Z}_t\}$:

(1) The sequences $\{Y_t^{(1)}\}$, $\{Y_t^{(2)}\}$, ..., $\{Y_t^{(n)}\}$ are modeled by a contemporaneous AR(1), CAR(1), process given by

$$\mathbf{Y}_t - \boldsymbol{\mu}_{\mathbf{Y}} = \Phi(\mathbf{Y}_{t-1} - \boldsymbol{\mu}_{\mathbf{Y}}) + \boldsymbol{\varepsilon}_t \tag{7.4}$$

where Φ is a diagonal $n \times n$ matrix, and $\{\boldsymbol{\varepsilon}_t\} \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varepsilon}}(0))$. Multiplying Eq (7.4) on the right by $(\mathbf{Y}_t - \boldsymbol{\mu}_{\mathbf{Y}})^{\mathrm{T}}$ and taking expectations gives

$$\mathbf{C}_{\boldsymbol{\varepsilon}}(0) = \mathbf{C}_{\mathbf{Y}}(0) - \Phi\mathbf{C}_{\mathbf{Y}}^{\mathrm{T}}(1) \tag{7.5}$$

and repeating for $(\mathbf{Y}_{t-h} - \boldsymbol{\mu}_{\mathbf{Y}})^{\mathrm{T}}$ gives

$$\mathbf{C}_{\mathbf{Y}}(h) = \Phi^h \mathbf{C}_{\mathbf{Y}}(0) \qquad h = 0, 1, \ldots \tag{7.6}$$

Thus $\mathbf{C}_{\mathbf{Y}}(h)$ has the same decaying structure, with respect to h, in space as in time. That is, $c^{ij}_{\mathbf{Y}}(h) = (\phi^{ii})^h c^{ij}_{\mathbf{Y}}(0)$ for $i, j \in \{1, 2, \ldots, n\}$ and $h = 0, 1, \ldots$, where ϕ^{ii} is the ith row and ith column element of Φ.

(2) The sequences $\{M_i^{(1)}\}$, $\{M_i^{(2)}\}$, ..., $\{M_i^{(n)}\}$ are correlated in space only at lag zero. That is, $\{\mathbf{M}_i\} \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \mathbf{C}_{\mathbf{M}}(0))$. It can be shown (see Appendix A) that a necessary and sufficient condition for $\{\mathbf{Z}_t\}$ to be stationary in the covariance is that $N_1, N_2, \ldots \overset{iid}{\sim} \mathrm{posgeom}(p)$ is a common sequence for all sites. In that case the covariance function of $\mathbf{Z}_t$ at lag h is (refer to Appendix A for derivation)

$$\mathbf{C}_{\mathbf{Z}}(h) = (1-p)^h \mathbf{C}_{\mathbf{M}}(0) \qquad h = 0, 1, \ldots \tag{7.7}$$

The condition that $\{N_i\}_{i=1}^{\infty}$ is a common sequence for all sites may also be supported in practice, if the shifts in the means are thought of being caused by changes in natural processes, such as changes in climate. In such cases it should be expected that time series of the same hydrologic variable within a geographic region would all exhibit shifts at the same times. Thus, in general the CSMAR(1) model should not be applied for multivariate analysis of time series if it is clear that shifts in different time series do not coincide in time. Such cases can come up if a a shift in a time series is caused by a construction of a dam or other man made constructions, where the construction does not affect the other time series being analyzed. Note that if $\mathbf{M}_t$ is assumed uncorrelated in space then the condition for stationarity that $\{N_i\}_{i=1}^{\infty}$ is a common sequence for all sites is not necessary any more.

7.2.1 Parameter Estimation for the CSMAR(1) model

The parameter estimation procedure is relatively simple for the CSMAR(1) model. First the CSMAR(1) model is uncoupled into univariate SMAR(1) models. If the common

p is not known, then $p^{(i)}$ is first estimated at each site i using the procedures in Chapter 6. The common p can then be estimated as a weighted average of the $\hat{p}^{(i)}$s

$$\hat{p} = \frac{1}{n^{(1)} + n^{(2)} + \cdots + n^{(n)}} \sum_{i=1}^{n} n^{(i)} \hat{p}^{(i)} \tag{7.8}$$

Given $\hat{p}$ the parameters of each univariate model are reestimated using the estimation procedures in Chapter 6.

After estimating the parameters of the univariate SMAR(1) models, what remains is estimating the non-diagonal elements of $\mathbf{C}_\varepsilon(0)$ and $\mathbf{C}_\mathbf{M}(0)$. The following procedure can be used to estimate $\mathbf{C}_\varepsilon(0)$ and $\mathbf{C}_\mathbf{M}(0)$ in general. Using Eqs (7.1) and (7.6)-(7.7), and the independence of $\{\mathbf{Y}_t\}$ and $\{\mathbf{Z}_t\}$ it follows that

$$\mathbf{C}_\mathbf{X}(h) = \Phi^h \mathbf{C}_\mathbf{Y}(0) + (1-p)^h \mathbf{C}_\mathbf{M}(0) \qquad h = 0, 1, \ldots \tag{7.9}$$

Estimates of $\mathbf{C}_\mathbf{M}(0)$ and $\mathbf{C}_\mathbf{Y}(0)$ are obtained by solving Eq (7.9) with $h = 0$ and $h = 1$ for $\mathbf{C}_\mathbf{M}(0)$ and $\mathbf{C}_\mathbf{Y}(0)$. It follows that

$$\hat{\mathbf{C}}_\mathbf{M}(0) = [\hat{\Phi} - (1-\hat{p})\mathbf{I}]^{-1} (\hat{\Phi}\hat{\mathbf{C}}_\mathbf{X}(0) - \hat{\mathbf{C}}_\mathbf{X}(1)) \tag{7.10}$$

and

$$\hat{\mathbf{C}}_\mathbf{Y}(0) = \hat{\mathbf{C}}_\mathbf{X}(0) - \hat{\mathbf{C}}_\mathbf{M}(0) \tag{7.11}$$

Finally using Eqs (7.5) and (7.6), $\mathbf{C}_\varepsilon(0)$ is estimated from

$$\hat{\mathbf{C}}_\varepsilon(0) = \hat{\mathbf{C}}_\mathbf{Y}(0) - \hat{\Phi}\hat{\mathbf{C}}_\mathbf{Y}^\mathrm{T}(0)\hat{\Phi}^\mathrm{T} \tag{7.12}$$

Note that if either the $\mathbf{Y}_t$ or the $\mathbf{M}_t$ process is assumed to be independent in space (that is either $\mathbf{C}_\mathbf{Y}(0)$ or $\mathbf{C}_\mathbf{M}(0)$ is diagonal), then only $\hat{\mathbf{C}}_\mathbf{X}(0)$ is needed to estimate $\mathbf{C}_\mathbf{Y}(0)$ and $\mathbf{C}_\mathbf{M}(0)$.

Furthermore, using that Φ is a diagonal matrix the ith row and jth column element of $\hat{\mathbf{C}}_\mathbf{M}(0)$ in Eq (7.10) can also be estimated from

$$\hat{c}_\mathbf{M}^{ij}(0) = \frac{\hat{\phi}^{ii}\hat{c}_\mathbf{X}^{ij}(0) - \hat{c}_\mathbf{X}^{ij}(1)}{\hat{\phi}^{ii} - (1-\hat{p})} \tag{7.13}$$

and $\hat{c}_{\epsilon}^{ij}(0)$ can be estimated from

$$\hat{c}_{\epsilon}^{ij}(0) = \hat{c}_{\mathbf{X}}^{ij}(0) - \hat{c}_{\mathbf{M}}^{ij}(0) - \hat{\phi}^{ii}\hat{\phi}^{jj}(\hat{c}_{\mathbf{X}}^{ji}(0) - \hat{c}_{\mathbf{M}}^{ji}(0)) \tag{7.14}$$

7.2.2 Problems Arising in Parameter Estimation

It is required that $\hat{\mathbf{C}}_{\mathbf{M}}(0)$ and $\hat{\mathbf{C}}_{\boldsymbol{\epsilon}}(0)$ are symmetric matrixes. In order for $\hat{\mathbf{C}}_{\mathbf{M}}(0)$ in Eq (7.10) to be symmetric the following relationship is needed between $\hat{c}_{\mathbf{X}}^{ij}(1)$ and $\hat{c}_{\mathbf{X}}^{ji}(1)$,

$$\hat{c}_{\mathbf{X}}^{ij}(1) = [\hat{\phi}^{jj} - 1 + p]^{-1}[(\hat{\phi}^{jj} - \hat{\phi}^{ii})(1-p)\hat{c}_{\mathbf{X}}^{ij}(0) + (\hat{\phi}^{ii} - 1 + p)\hat{c}_{\mathbf{X}}^{ji}(1)] \tag{7.15}$$

where it has been used that $\hat{c}_{\mathbf{X}}^{ij}(0) = \hat{c}_{\mathbf{X}}^{ji}(0)$. Since it is very unlikely that $\hat{c}_{\mathbf{X}}^{ij}(1)$ and $\hat{c}_{\mathbf{X}}^{ji}(1)$ calculated from the data will follow the relationship in Eq (7.15) an adjustment is needed to make $\hat{\mathbf{C}}_{\mathbf{M}}(0)$ symmetric. The simplest such adjustment is to replace all $\hat{c}_{\mathbf{M}}^{ij}(0)$ and $\hat{c}_{\mathbf{M}}^{ji}(0)$ with their respective averages. If $\hat{\mathbf{C}}_{\mathbf{M}}(0)$ is symmetric then no further adjustment is needed in the estimation of $\mathbf{C}_{\boldsymbol{\epsilon}}(0)$. Furthermore, if fitted ACFs were used instead of sample ACFs in the estimation of the univariate SMAR(1) models, then for consistency the fitted ACFs should be used in estimation of the diagonal elements of $\hat{\mathbf{C}}_{\mathbf{X}}(1)$ in Eq (7.10).

7.3 CSMAR(1)-CARMA : Mixture of CSMAR(1) and CARMA(p,q)

Analyzes of multiple time series of different hydrologic variables may require mixing of models. For example shifts in time series of one hydrologic variable may not be present in a time series of another hydrologic variable. Or, if different geographic locations are used for analysis of a single hydrologic variable, then characteristics of the corresponding times series may be dependent on their geographic location. In such cases mixing of multiple CSMAR(1) models and other time series models, such as CARMA(p,q), may be desirable. In this section we will formulate a mixture of one CSMAR(1) model with one CARMA(p,q) model, where the lag zero cross correlation function (CCF) in space is preserved between the

CARMA(p,q) model and the CAR(1) component of the CSMAR(1) model. The analysis can though easily be extended to incorporate more than one CSMAR(1) model, where for multiple CSMAR(1) the lag zero CCF is preserved between the different CAR(1) components of the models but not between the different level shift components.

Lets assume that there are total of n sites, of which n_1 sites follow a CSMAR(1) model and n_2 sites follow a CARMA(p,q) model, where $n_1+n_2=n$. The model of the n sites can be presented by Eq (7.1), where the first n_1 elements of $\mathbf{X}_t$ represent the CSMAR(1) model and the remaining n_2 elements of $\mathbf{X}_t$ represent the CARMA(p,q) model

$$\begin{bmatrix} X_t^{(1)} \\ \vdots \\ X_t^{(n_1)} \\ X_t^{(n_1+1)} \\ \vdots \\ X_t^{(n)} \end{bmatrix} = \begin{bmatrix} Y_t^{(1)} \\ \vdots \\ Y_t^{(n_1)} \\ Y_t^{(n_1+1)} \\ \vdots \\ Y_t^{(n)} \end{bmatrix} + \begin{bmatrix} Z_t^{(1)} \\ \vdots \\ Z_t^{(n_1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{7.16}$$

In general the whole vector $\mathbf{Y}_t$ can be looked at as being modeled by a CARMA(p,q) model

$$\mathbf{Y}_t - \boldsymbol{\mu}_{\mathbf{Y}} = \sum_{j=1}^{p} \Phi_j(\mathbf{Y}_{t-j} - \boldsymbol{\mu}_{\mathbf{Y}}) + \boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \Theta_j \boldsymbol{\varepsilon}_{t-j} \tag{7.17}$$

where $\{\boldsymbol{\varepsilon}_t\} \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varepsilon}}(0))$, and the parameters $\Phi_1, \Phi_2, \ldots, \Phi_p$, $\Theta_1, \Theta_2, \ldots, \Theta_q$ are diagonal $n \times n$ matrixes. Each of the first n_1 elements of $\mathbf{Y}_t$ is an AR(1) process, and each of the remaining n_2 elements of $\mathbf{Y}_t$ follows some ARMA(p,q) process. That is, $Y_t^{(i)}$ is an ARMA(p_i,q_i) process, $i=1,2,\ldots,n$, where the p_is can be different and the q_is can be different. The p and the q of the CARMA(p,q) model are $p=\max(p_1,p_2,\ldots,p_n)$ and $q=\max(q_1,q_2,\ldots,q_n)$.

The parameter matrixes of the CARMA(p,q) are diagonal, thus estimation of parameters of the CSMAR(1)-CARMA model can be done in a similar way as for the CSMAR(1) model, where Eq (7.16) is uncoupled into univariate model. For the CSMAR(1) portion of Eq (7.16), parameters are estimated using procedures in section 7.2.1. For estimation of each

of the univariate ARMA(p_i, q_i), $i = n_1 + 1, n_1 + 2, \ldots, n$, models refer to Salas (1993); Hipel and McLeod (1994); and Brockwell and Davis (1996). Hipel and McLeod (1994) also give a joint multivariate estimation algorithm for estimation of the parameters of the CARMA(p, q) model. The algorithm to estimate $\mathbf{C}_{\boldsymbol{\epsilon}}(0)$ is simple, but a necessary condition is that the CARMA(p, q) is causal. This is equivalent to requiring each of the estimated univariate ARMA(p, q) models to be causal (often a common requirement in estimation procedures for ARMA models). Causality implies that $\mathbf{Y}_t$ can be written out as an infinite moving average model

$$\mathbf{Y}_t - \boldsymbol{\mu}_{\mathbf{Y}} = \sum_{j=0}^{\infty} \Psi_j \boldsymbol{\epsilon}_{t-j} \tag{7.18}$$

where $\{\Psi_j\}$ are matrixes with absolutely summable elements. Substituting from Eq (7.18) to Eq (7.17) for $\mathbf{Y}_t$ implies that

$$(I - \Phi_1 z - \cdots - \Phi_p z^p)(\Psi_0 + \Psi_1 z + \cdots) = (I - \Theta_1 z - \cdots - \Theta_q z^q) \tag{7.19}$$

where I is the identity matrix. From Eq (7.19) $\{\Psi_j\}$ can be derived by equating the coefficients of z^j

$$\Psi_0 = I \tag{7.20}$$

$$\Psi_j = -\Theta_j + \sum_{k=1}^{p} \Phi_k \Psi_{j-k} \qquad j = 1, 2, \ldots \tag{7.21}$$

where $\Psi_j = \mathbf{0}$ for $j < 0$ and $\Theta_j = \mathbf{0}$ for $j > q$. For the special case when $p = 1$ and $q = 0$ then $\Psi_j = \Phi_1^j$, for $j = 1, 2, \ldots$. Multiplying Eq (7.18) on the right by $(\mathbf{Y}_t - \boldsymbol{\mu}_{\mathbf{Y}})^{\mathrm{T}}$ and taking expectations gives

$$\mathbf{C}_{\mathbf{Y}}(0) = \sum_{j=0}^{\infty} \Psi_j \mathbf{C}_{\boldsymbol{\epsilon}}(0) \Psi_j^{\mathrm{T}} \tag{7.22}$$

Since Ψ_j, $j = 0, 1, \ldots$, are diagonal matrixes an estimate of the ith row and jth column element of $\mathbf{C}_{\boldsymbol{\epsilon}}(0)$ is

$$\hat{c}_{\boldsymbol{\epsilon}}^{ij}(0) = \frac{\hat{c}_{\mathbf{Y}}^{ij}(0)}{\sum_{k=0}^{\infty} \hat{\psi}_k^{ii} \hat{\psi}_k^{jj}} \tag{7.23}$$

The $\hat{\psi}_k^{ii}$ decay rather quickly, thus the sum in Eq (7.23) can usually be truncated at a fairly low value of k.

7.4 The Special Case : $\phi = 0$: The CSM-1-CARMA Model

In the special case with $\phi = 0$ (no persistence in the Y_t process) the CSMAR(1) model in section 7.2 reduces to a contemporaneous SM-1 model, dubbed here as CSM-1. Thus, the sequences $\{Y_t^{(1)}\}$, $\{Y_t^{(2)}\}$, ..., $\{Y_t^{(n)}\}$ are correlated in space at lag 0 only, and independent in time, with $\{\mathbf{Y}_t\} \overset{iid}{\sim} \text{MVN}(\boldsymbol{\mu}_\mathbf{Y}, \mathbf{C}_\mathbf{Y}(0))$. The properties of the $\{M_i^{(1)}\}$, $\{M_i^{(2)}\}$, ..., $\{M_i^{(n)}\}$ do not change. In this case the covariance function of $\mathbf{X}_t$ in Eq (7.9)) becomes

$$\mathbf{C}_\mathbf{X}(h) = \begin{cases} \mathbf{C}_\mathbf{Y}(0) + \mathbf{C}_\mathbf{M}(0) & \text{if } h = 0 \\ (1-p)^h \mathbf{C}_\mathbf{M}(0) & \text{for } h = 1, 2, \ldots \end{cases} \tag{7.24}$$

7.4.1 Parameter Estimation for the CSM-1 model

The parameter estimation procedure for the CSM-1 model follows the same steps as the parameter estimation procedure for the CSMAR(1) model in section 7.2.1. That is, first the CSM-1 is coupled into univariate SM-1 models and the parameters are estimated at each site using procedures in Chapter 6. Then the common p for all sites is estimated as a weighted average of the estimated $p^{(i)}$s of the univariate SM-1 models (refer to Eq (7.8)). Given $\hat{p}$ the parameters of the univariate SM-1 models are reestimated using procedures in Chapter 6 for the case p known. What remains is estimating the non-diagonal elements of $\mathbf{C}_\mathbf{Y}(0)$ and $\mathbf{C}_\mathbf{M}(0)$. Using Eq (7.24), then $\mathbf{C}_\mathbf{M}(0)$ is estimated from

$$\hat{\mathbf{C}}_\mathbf{M}(0) = (1-\hat{p})\hat{\mathbf{C}}_\mathbf{X}(1)) \tag{7.25}$$

where if necessary $\hat{\mathbf{C}}_\mathbf{M}(0)$ is made symmetric by replacing $\hat{c}_\mathbf{M}^{ij}(0)$ and $\hat{c}_\mathbf{M}^{ji}(0)$ with their respective averages. Then $\mathbf{C}_\mathbf{Y}(0)$ is estimated from

$$\hat{\mathbf{C}}_\mathbf{Y}(0) = \hat{\mathbf{C}}_\mathbf{X}(0) - \hat{\mathbf{C}}_\mathbf{M}(0) \tag{7.26}$$

7.4.2 Parameter Estimation for the CSM-1-CARMA model

The CSM-1-CARMA follows the same concept as the CSMAR(1)-CARMA model in section 7.3. Given the CSM-1 model then parameters of the CSM-1-CARMA model are estimated using the procedures for estimation of the CSMAR(1)-CARMA parameters in section 7.3, where each of the elements of $\{\mathbf{Y}_t\}$ corresponding to the CSM-1 process is looked at as being modeled by an ARMA$(0,0)$ process.

7.5 The Great Lakes System

The intent here is to fit a multivariate model to the annual net basin supplies (NBS) of the lakes in the Great Lakes system using the procedures presented in this paper. The data were obtained from Hydro-Quebec, and span the period 1900–1999 for lakes Erie, Michigan-Huron, Ontario, and Superior, and the period 1900–1989 for Lake St. Clair. Data post 1989 for Lake St. Clair were still preliminary, and hence are not used in this study. The annual NBS time series of the Great Lakes and their ACFs can be seen in Figs. 7.1–7.5. The data for Lake Superior and Lake Michigan-Huron in Figs. 7.2 and 7.5 do not seem to exhibit any sudden shifts, and in addition the ACFs of the data do not have shapes that are expected of the SMAR(1) model. On the other hand, the data for the other lakes in Figs. 7.1 and 7.3–7.4 appear to be characterized by sudden shifts. Furthermore, the cross correlations of the Great Lakes data are plotted in Fig. 7.6. In all cases in Fig. 7.6, the lag zero cross correlation coefficient is significant. Thus, contemporaneous models could be used to preserve the lag zero cross correlation coefficient between different lakes. The Lake St. Clair is the smallest of the five lakes considered here. It is upstream from Lake Erie, which is upstream from Lake Ontario. Looking at Fig. 7.6, the annual NBS of Lake St. Clair is significantly correlated with past annual NBS of Lake Erie and Lake Ontario for up to at least four lags as shown in the figure.

Throughout this paper, for a time series $X_1, X_2, \ldots, X_n$ the sample mean, variance,

skewness, and lag h ACF will be estimated as in Chapter 6.5. In terms of storage related statistics we will use the Hurst slope K, the storage capacity SC, and the longest drought length DL and the corresponding magnitude DM with respect to a demand level d. For definitions of these storage related statistics refer to Chapter 6.5.

The sample mean, standard deviation, skewness, Hurst slope, storage capacity, and the longest drought length and the corresponding drought magnitude based on demand level $d = \hat{\mu}_X$ of the Great Lakes data are shown in Table 7.1. Also in Table 7.1 the values of the sample and the fitted model ACFs, of the CSMAR(1) and the CSM-1 models, are shown for lags 1 to 3.

7.5.1 Fitting a Multivariate Contemporaneous Model to the Great Lakes System

We will attempt to fit a mixture of CSMAR(1) and a CARMA(p,q) model to the data, where the lakes Erie, Ontario, and St. Clair will be modeled by a CSMAR(1) model, and the lakes Michigan-Huron and Superior will be modeled by CARMA(p,q) model. Note that the lag one cross correlation coefficient between the lakes Erie, Ontario, and St. Clair appears significant in Fig. 7.6, but this cross correlation coefficient is needed for fitting of the CSMAR(1) model. The ACF and the partial ACF (not shown) of lakes Michigan-Huron and Superior in Figs. 7.2 and 7.5 suggests a CARMA$(0,0)$ model (or a bivariate normal model). The CSMAR(1) and the CSMAR(1)-CARMA can be fitted to sample series of different lengths spanning different time-spans, to cover all possibilities the following general approach is used to estimate the sample lag h cross covariance function (CCVF) in Eq (7.3) between two sample series $X^{(1)}_{t_1}, X^{(1)}_{t_1+1}, \ldots, X^{(1)}_{t_1+n^{(1)}-1}$ and $X^{(2)}_{t_2}, X^{(2)}_{t_2+1}, \ldots, X^{(2)}_{t_2+n^{(2)}-1}$

$$\hat{c}^{12}_{\mathbf{X}}(h) = \frac{1}{k_2 - k_1 + 1} \sum_{i=k_1}^{k_2} (X^{(1)}_{i+h} - \overline{X}^{(1)})(X^{(2)}_{i} - \overline{X}^{(2)}) \tag{7.27}$$

similarly the sample cross correlation function (CCF) can be estimated from

$$\hat{\rho}_{\mathbf{X}}^{12}(h) = \frac{\sum_{i=k_1}^{k_2}(X_{i+h}^{(1)} - \overline{X}^{(1)})(X_i^{(2)} - \overline{X}^{(2)})}{\left[\sum_{i=k_1}^{k_2}(X_{i+h}^{(1)} - \overline{X}^{(1)})^2 \sum_{i=k_1}^{k_2}(X_i^{(2)} - \overline{X}^{(2)})\right]^{1/2}} \tag{7.28}$$

where $k_1 = \max(t_1 - h, t_2)$ and $k_2 = \min(t_1 + n^{(1)} - h - 1, t_2 + n^{(2)} - 1)$, and $\overline{X}^{(1)} = (\sum_{i=k_1}^{k_2} X_{i+h}^{(1)})/(k_2 - k_1 + 1)$ and $\overline{X}^{(2)} = (\sum_{i=k_1}^{k_2} X_i^{(2)})/(k_2 - k_1 + 1)$ respectively. Note that the diagonal elements of covariance matrixes are the autocovariance estimated as $\hat{c}_X(h) = \hat{\sigma}_X^2 \hat{\rho}_X(h)$.

To estimate the parameters of the CSMAR(1) model for lakes Erie, Ontario, and St. Clair first p is estimated from Eq (7.8) and then the procedures in section 7.2.1 are followed for estimation of the other parameters. The estimated parameters are shown in Table 7.2, where non-parenthesized values are estimated using the sample ACFs and parenthesized values are estimated using re-fitted ACFs based on assuming that $\hat{p}$ is known and using the sample ACFs up to lag 15 in the ACF fitting process (refer to Chapter 6). The model ACFs for these two different cases are compared with the sample ACFs in Fig. 7.7. From the figure it is evident that for Lake St. Clair the two model ACFs are quite different, with the model ACF based on the fitted ACF representing the sample ACF better. For Lake Erie and Lake Ontario the model ACFs are more similar. Thus there are several possibilities for selecting the CSMAR(1) model, where one has to choose whether to use only the models fitted based on the sample ACFs, or the models based on the fitted ACFs, or a mixture of both. Another possibility could be to experiment with fitted ACFs using the sample ACFs up to different lags in the fitting process. For the purpose of this study we will use a CSMAR(1) model with parameters estimated based on the fitted ACFs in Table 7.2 for lakes Erie and Lake Ontario, and St. Clair. Other possible choices such as using the sample ACFs for all three lakes, or using the sample ACFs for Lake Erie and Lake Ontario, and the fitted ACF for Lake St. Clair, resulted in $\hat{\mathbf{C}}_{\mathbf{M}}(0)$ that was not positive definite (i.e. had at least one negative eigenvalue). Furthermore, a CARMA(0,0) model is used for Lake Michigan-Huron and Lake Superior. The parameters of the full CSMAR(1)-CARMA model are shown in

Table 7.3, where for comparison under same assumptions the model parameters for the full CSM-1-CARMA model with parameters estimated based on procedures in section 7.4. Note that for the CSM-1-CARMA model only $\hat{\mathbf{C}}_{\mathbf{Y}}(0)$ is shown in Table 7.3, the reason being that there are no autoregressive components nor moving average components in the fitted CSM-1-CARMA model, thus $\hat{\mathbf{C}}_{\mathbf{Y}} = \hat{\mathbf{C}}_{\boldsymbol{\varepsilon}}$.

To analyze how capable the fitted models are in preserving the sample statistics used in the fitting procedures, 1,000 realizations of the same lengths as the historical records were generated for the full models in Table 7.3, and one realization of length 1,000 × the length of the historical records was generated. All the lakes have historical records of the same length, $n = 100$, except Lake St. Clair, which has a record length $n = 90$. Thus the generated records for Lake St. Clair were truncated to match the length of the historical record. The average sample statistics of the 1,000 generated realizations are shown in Table 7.4, and the sample statistics of the one generated realization of length $1,000\ n$ are shown in Table 7.5. Comparing with the historical sample statistics in Table 7.1, the mean and the standard deviation are well preserved in all cases. The model ACFs of lakes Michigan-Huron and Superior is zero, as is reflected in Tables 7.4 and 7.5. The ACFs of the generated sequences for lakes Erie, Ontario, and St. Clair should be compared with the fitted ACFs in Table 7.2. As commonly is observed the average ACFs based on 1,000 realization of the same length as the historical records underestimate the model ACF, while the ACFs of the one realization of length 1,000 n preserve the model ACF well. Comparing the storage related statistics K, SC, DL, and DM in Table 7.4 with the corresponding historical statistics in Table 7.1, it can be said that they are in general relatively well preserved. Recall that the storage related statistics depend on the sample size, and are thus not expected to be preserved in Table 7.5. In comparing the results among the two different models, then the focus should be on statistics not used in the fitting procedures of the models, since in general statistics used in the fitting procedures are expected to be preserved. Thus comparing the storage related statistics of the two fitted models, then in general the two different models, the CSMAR(1)-

CARMA and the CSM-1-CARMA, give very similar results. A reason for the similarity may be that the ϕ parameters are close to zero in the CSMAR(1) part of the CSMAR(1)-CARMA model. Recall that the CSM-1 model is a special case of the CSMAR(1) model with $\phi = 0$.

In Tables 7.6 and 7.7 the lag 0 and lag 1 historical CCF matrixes are shown along with the corresponding CCF matrixes based on the 1,000 realizations of length n and based on the one realization of length $1,000$ n, for the CSMAR(1)-CARMA model and the CSM-1-CARMA model, respectively. Comparing the CCF matrixes based on the generated sequences with the historical CCF matrixes, then as expected the lag 0 CCF is very well preserved between all stations for both cases. The lag 1 historical CCF was used in estimation of $\mathbf{C_M}(0)$ and $\mathbf{C_Y}(0)$ in the CSMAR(1) part of the model (refer to section 7.2). As a result $\hat{\mathbf{C}}_{\mathbf{M}}(0)$ and $\hat{\mathbf{C}}_{\mathbf{Y}}(0)$ were not necessarily symmetric and an adjustment was made to make these matrixes symmetric. Thus the lag 1 CCF for the CSMAR(1) part of the model (the upper-left 3×3 sub-matrix of $\hat{\boldsymbol{\rho}}_{\mathbf{X}}(1)$ in Table 7.6) may not be exactly preserved, but in general the off-diagonal averages of $\hat{\rho}_{\mathbf{X}}^{ij}(1)$ and $\hat{\rho}_{\mathbf{X}}^{ji}(1)$ should be relatively well preserved. The values of the lag 1 CCF in Table 7.6 support this. Note that any lag 1 CCF including Lake Michigan-Huron or Lake Superior is not expected to be preserved. Comparing the results among the two different models, then again both models give similar results.

For further comparison the first three realizations of the 1,000 generated sequences are shown in Figs. 7.8–7.10 based on the CSMAR(1)-CARMA model. The lag zero dependence between different lakes can be observed in the figures. Furthermore, the graphs for each lake are drawn on the same scale in all three figures to make comparison easier.

7.6 Summary and Final Remarks

In this paper a multivariate shifting mean modeling framework was developed. More precisely, a contemporaneous version of the univariate shifting mean autoregressive AR(1) model, SMAR(1), in Chapter 6, was developed and dubbed as CSMAR(1). In addition, a general contemporaneous model mixing CSMAR(1) and CARMA models was developed for

modeling of systems, where some of the sites exhibit sudden shifting patterns while others do not. This model was dubbed as CSMAR(1)-CARMA. The special cases of the above models assuming no direct AR(1) persistence in the CSMAR(1) model were also developed. The special cases were, dubbed as CSM-1 and CSM-1-CARMA. A necessary condition for stationarity of the CSMAR(1) is that the sequence of the mean level lengths is common for all sites, that is that shifts at different sites coincide in time. The above models are capable of preserving the lag zero cross correlation in space between different sites. In addition, for sites modeled by the CSMAR(1) or the CSM-1 models, some characteristics related to the lag one cross correlation in space are also preserved.

Historical records of some of the lakes in the Great Lakes system show evidence of sudden shifts in addition to autocorrelation, while records for other lakes do not indicate such behavior. The proposed models, where applied for modeling jointly the Great Lakes system as a whole, with lakes Erie, Ontario, and St. Clair modeled by contemporaneous shifting mean models, and lakes Michigan-Huron and Superior modeled by a CARMA(0,0) model. The models were capable of preserving the lag zero spatial correlation between different lakes, in addition to preserving other important statistical characteristics of the individual lakes.

As a general conclusion, the proposed mixture models mixing contemporaneous shifting mean models and contemporaneous ARMA models appear to be robust and seem to have a wide range of applicability for modeling of hydroclimatic and geophysical systems.

Table 7.1: Sample statistics of the Great Lakes NBS time series from 1900–1999, except for Lake St. Clair where the statistics correspond to the period 1900-1989. Fitted ACFs up to lag 3 for the CSMAR(1) model and the CSM-1 model are also shown.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
$\hat{\mu}_X$[cms]	574.1	3177	1033	121.7	2043
$\hat{\sigma}_X$[cms]	265.4	737.0	241.6	63.34	478.8
$\hat{\gamma}_X$	0.138	-0.091	0.491	0.311	0.033
K	0.787	0.713	0.786	0.847	0.654
SC[cms]	5506	11978	5083	1529	4755
DL	8	8	11	9	5
DM[cms]	1720	6029	2034	659.3	3850
Sample ACF					
$\hat{\rho}_1$	0.173	0.168	0.250	0.504	0.153
$\hat{\rho}_2$	0.170	0.003	0.212	0.291	0.050
$\hat{\rho}_3$	0.175	-0.088	0.199	0.236	-0.023
Fitted ACF for CSMAR(1) Model					
$\hat{\rho}_1$	0.175		0.263	0.484	
$\hat{\rho}_2$	0.192		0.266	0.362	
$\hat{\rho}_3$	0.173		0.199	0.300	
Fitted ACF for CSM-1 Model ($\phi = 0$)					
$\hat{\rho}_1$	0.245		0.254	0.436	
$\hat{\rho}_2$	0.208		0.216	0.369	
$\hat{\rho}_3$	0.176		0.183	0.313	

Table 7.2: Estimated parameters of the CSMAR(1) model for lakes Erie, Ontario, and St. Clair. The non-parenthesized values are estimated using the sample ACF, while the parenthesized values are estimated using the fitted ACF.

Parameter	Lake		
	Erie	Ontario	St. Clair
$\hat{\phi}$	-0.0361 (-0.1170)	-0.0018 (-0.0162)	0.3932 (0.1262)
$\hat{p}$	0.1574	0.1574	0.1574
$\hat{\mu}_Y$ [cms]	574.1	1033	121.7
$\hat{\mathbf{C}}_\mathbf{M}(0)$ [cms^2]	16796 (22401)	16586 (18016)	4820 (5519)
	16586 (18016)	17415 (18121)	3334 (3973)
	4820 (5519)	3334 (3973)	988 (2002)
$\hat{\mathbf{C}}_\mathbf{Y}(0)$ [cms^2]	53663 (48058)	28083 (26654)	4144 (3445)
	28083 (26654)	40944 (40238)	4962 (4323)
	4144 (3445)	4962 (4323)	3025 (2010)
$\hat{\mathbf{C}}_\boldsymbol{\epsilon}(0)$ [cms^2]	53593 (47400)	28081 (26603)	4203 (3496)
	28081 (26603)	40944 (40228)	4965 (4331)
	4203 (3496)	4965 (4331)	2557 (1978)

Table 7.3: Estimated parameters of the full CSMAR(1)-CARMA and CSM-1-CARMA models for the Great Lakes. The lakes Erie, Ontario, and St. Clair are fitted by a CSMAR(1) and CSM-1 models dependign on the case, and the lakes Michigan-Huron and Superior are fitted by a CARMA(0,0) model (or a bivariate normal model).

Parameter	Lake				
	Erie	Ontario	St. Clair	Michigan-Huron	Superior
Parameters for CSMAR(1)-CARMA Model					
$\hat{\phi}$	-0.1170	-0.0162	0.1262		
$\hat{p}$	0.1574	0.1574	0.1574		
$\hat{\mu}_Y$ [cms]	574.1	1033	121.7	3177	2043
$\hat{\mathbf{C}}_\mathbf{M}(0)$ [cms^2]	22401	18016	5519		
	18016	18121	3973		
	5519	3973	2002		
$\hat{\mathbf{C}}_\mathbf{Y}(0)$ [cms^2]	48058	26654	3445	103113	37944
	26654	40238	4323	114183	31060
	3445	4323	2010	20964	7431
	103113	114183	20964	543228	195286
	37944	31060	7431	195286	229280
$\hat{\mathbf{C}}_\boldsymbol{\epsilon}(0)$ [cms^2]	47400	26603	3496	103113	37944
	26603	40228	4331	114183	31060
	3496	4331	1978	20964	7431
	103113	114183	20964	543228	195286
	37944	31060	7431	195286	229280
Parameters for CSM-1-CARMA Model ($\phi = 0$)					
$\hat{p}$	0.1532	0.1532	0.1532		
$\hat{\mu}_Y$ [cms]	574.1	1033	121.7	3177	2043
$\hat{\mathbf{C}}_\mathbf{M}(0)$ [cms^2]	20387	15869	5234		
	15869	17538	4035		
	5234	4035	2066		
$\hat{\mathbf{C}}_\mathbf{Y}(0)$ [cms^2]	50072	28800	3730	103113	37944
	28800	40821	4260	114183	31060
	3730	4260	1946	20964	7431
	103113	114183	20964	543228	195286
	37944	31060	7431	195286	229280

Table 7.4: Average sample statistics of 1,000 generated NBS time series of the Great Lakes of the same lengths as the historical records for the full CSMAR(1)-CARMA and CSM-1-CARMA models in Table 7.3.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
CSMAR(1)-CARMA Model					
$\hat{\mu}_X$[cms]	573.3	3175	1033	121.9	2042
$\hat{\sigma}_X$[cms]	261.2	732.2	237.1	61.33	474.9
$\hat{\gamma}_X$	-0.014	-0.010	0.006	-0.011	0.004
K	0.727	0.616	0.730	0.779	0.617
SC[cms]	4877	8434	4484	1290	5475
DL	8.561	5.903	8.702	10.46	5.957
DM[cms]	2344	4026	2180	722.5	2609
$\hat{\rho}_1$	0.142	-0.009	0.204	0.417	-0.004
$\hat{\rho}_2$	0.194	-0.007	0.170	0.285	-0.004
$\hat{\rho}_3$	0.142	-0.008	0.138	0.217	-0.010
CSM-1-CARMA Model ($\phi = 0$)					
$\hat{\mu}_X$[cms]	573.3	3175	1033	121.9	2041
$\hat{\sigma}_X$[cms]	261.3	732.1	237.0	61.31	474.9
$\hat{\gamma}_X$	-0.013	-0.010	0.008	-0.006	0.004
K	0.728	0.616	0.729	0.779	0.617
SC[cms]	4890	8433	4490	1300	5476
DL	8.522	5.898	8.619	10.40	5.962
DM[cms]	2345	4014	2163	715.5	2610
$\hat{\rho}_1$	0.203	-0.009	0.209	0.365	-0.004
$\hat{\rho}_2$	0.170	-0.007	0.166	0.292	-0.004
$\hat{\rho}_3$	0.130	-0.008	0.133	0.228	-0.010

Table 7.5: Sample statistics of 1 generated NBS time series of the Great Lakes of lengths 1,000 × the lengths of the historical records for the full CSMAR(1)-CARMA and CARMACSM-1-CARMA models in Table 7.3.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
CSMAR(1)-CARMA Model					
$\hat{\mu}_X$[cms]	572.9	3178	1032	121.8	2042
$\hat{\sigma}_X$[cms]	265.9	738.5	242.0	63.15	480.6
$\hat{\gamma}_X$	-0.006	0.012	0.010	0.005	0.004
K	0.639	0.564	0.643	0.652	0.558
SC[cms]	325061	236522	318396	54542	170831
DL	43	15	47	51	17
DM[cms]	13120	10468	14724	5104	6901
$\hat{\rho}_1$	0.188	-0.001	0.249	0.482	-0.004
$\hat{\rho}_2$	0.232	-0.002	0.219	0.361	0.001
$\hat{\rho}_3$	0.190	0.006	0.185	0.298	0.007
CSM-1-CARMA Model ($\phi = 0$)					
$\hat{\mu}_X$[cms]	572.9	3178	1032	121.8	2042
$\hat{\sigma}_X$[cms]	266.2	738.4	242.4	63.44	480.6
$\hat{\gamma}_X$	0.007	0.012	0.018	0.011	0.005
K	0.642	0.563	0.642	0.642	0.558
SC[cms]	330920	237104	312468	53003	170903
DL	36	15	36	60	17
DM[cms]	13611	10403	10302	6031	6930
$\hat{\rho}_1$	0.245	-0.001	0.255	0.438	-0.004
$\hat{\rho}_2$	0.206	-0.002	0.217	0.369	0.001
$\hat{\rho}_3$	0.179	0.006	0.186	0.314	0.007

Table 7.6: Historical and generated cross correlation function (CCF) matrixes of the Great Lakes NBS time series for the CSMAR(1)-CARMA model in Table 7.3.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
Historical CCF Matrixes					
$\hat{\rho}_{\mathbf{X}}(0)$	1	0.697	0.549	0.527	0.299
	0.697	1	0.569	0.641	0.269
	0.549	0.569	1	0.452	0.245
	0.527	0.641	0.452	1	0.553
	0.299	0.269	0.245	0.553	1
$\hat{\rho}_{\mathbf{X}}(1)$	0.173	0.198	0.144	0.030	0.156
	0.220	0.250	0.084	0.151	0.255
	0.393	0.380	0.504	0.196	0.240
	0.181	0.129	0.027	0.168	0.322
	0.006	-0.014	0.036	-0.066	0.153
Average CCF Matrixes from 1,000 generated series of size n					
$\hat{\rho}_{\mathbf{X}}(0)$	1	0.692	0.521	0.536	0.301
	0.692	1	0.537	0.652	0.271
	0.521	0.537	1	0.461	0.251
	0.536	0.652	0.461	1	0.552
	0.301	0.271	0.251	0.552	1
$\hat{\rho}_{\mathbf{X}}(1)$	0.142	0.164	0.226	-0.065	-0.036
	0.204	0.204	0.189	-0.018	-0.005
	0.275	0.232	0.417	0.052	0.029
	-0.002	-0.004	-0.002	-0.009	0.000
	-0.001	0.003	-0.001	-0.003	-0.004
CCF Matrixes from 1 generated series of size 1,000 n					
$\hat{\rho}_{\mathbf{X}}(0)$	1	0.695	0.528	0.527	0.303
	0.695	1	0.540	0.644	0.274
	0.528	0.540	1	0.448	0.249
	0.527	0.644	0.448	1	0.555
	0.303	0.274	0.249	0.555	1
$\hat{\rho}_{\mathbf{X}}(1)$	0.188	0.186	0.248	-0.063	-0.037
	0.228	0.249	0.208	-0.012	-0.007
	0.301	0.252	0.482	0.055	0.029
	-0.002	0.001	-0.003	-0.001	-0.006
	-0.002	0.001	0.004	-0.003	-0.004

Table 7.7: Historical and generated cross correlation function (CCF) matrixes of the Great Lakes NBS time series for the CSM-1-CARMA model in Table 7.3.

Statistic	Lake				
	Erie	Michigan-Huron	Ontario	St. Clair	Superior
Historical CCF Matrixes					
$\hat{\rho}_{\mathbf{X}}(0)$	1	0.697	0.549	0.527	0.299
	0.697	1	0.569	0.641	0.269
	0.549	0.569	1	0.452	0.245
	0.527	0.641	0.452	1	0.553
	0.299	0.269	0.245	0.553	1
$\hat{\rho}_{\mathbf{X}}(1)$	0.173	0.198	0.144	0.030	0.156
	0.220	0.250	0.084	0.151	0.255
	0.393	0.380	0.504	0.196	0.240
	0.181	0.129	0.027	0.168	0.322
	0.006	-0.014	0.036	-0.066	0.153
Average CCF Matrixes from 1,000 generated series of size n					
$\hat{\rho}_{\mathbf{X}}(0)$	1	0.694	0.522	0.536	0.301
	0.694	1	0.536	0.652	0.271
	0.522	0.536	1	0.463	0.252
	0.536	0.652	0.463	1	0.552
	0.301	0.271	0.252	0.552	1
$\hat{\rho}_{\mathbf{X}}(1)$	0.203	0.189	0.239	-0.002	-0.001
	0.186	0.209	0.198	-0.007	-0.001
	0.237	0.202	0.365	-0.004	0.000
	-0.002	-0.004	0.000	-0.009	0.000
	-0.001	0.003	0.001	-0.003	-0.004
CCF Matrixes from 1 generated series of size 1,000 n					
$\hat{\rho}_{\mathbf{X}}(0)$	1	0.696	0.532	0.528	0.302
	0.696	1	0.543	0.644	0.272
	0.532	0.543	1	0.448	0.247
	0.528	0.644	0.448	1	0.554
	0.302	0.272	0.247	0.554	1
$\hat{\rho}_{\mathbf{X}}(1)$	0.245	0.209	0.265	-0.001	-0.002
	0.209	0.255	0.224	-0.002	-0.005
	0.265	0.222	0.438	0.000	-0.001
	-0.001	0.003	0.000	-0.001	-0.006
	-0.002	0.000	0.002	-0.003	-0.004

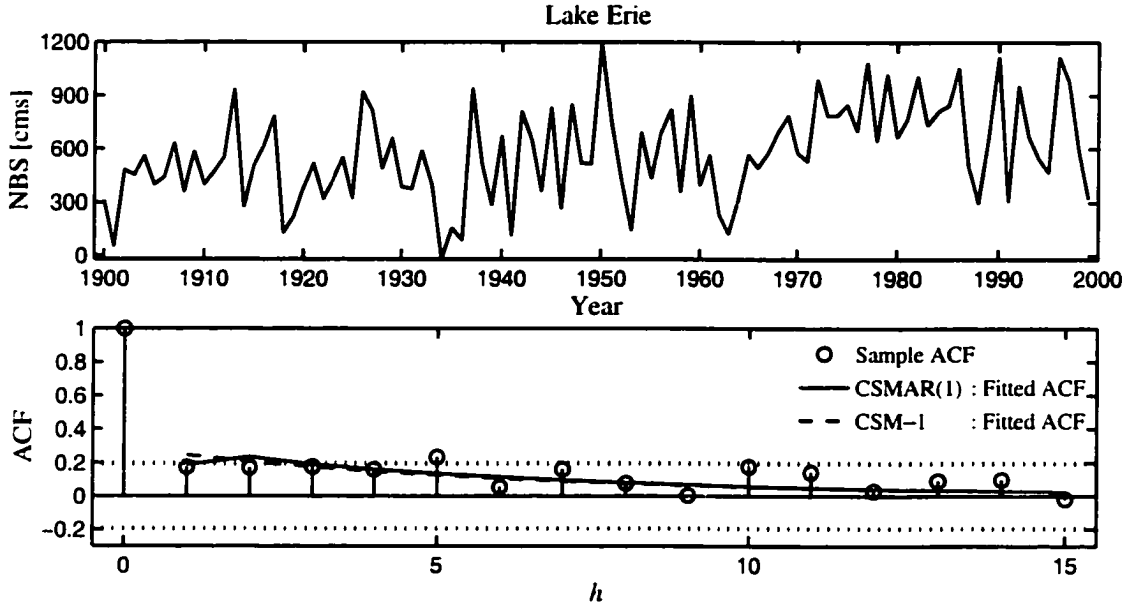

Figure 7.1: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Erie.

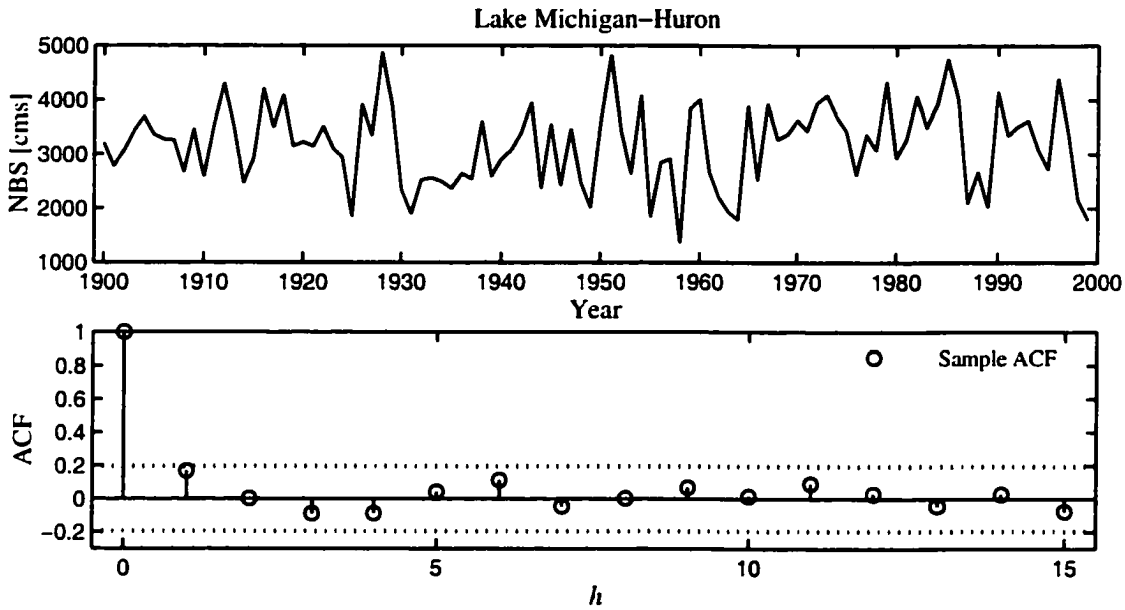

Figure 7.2: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Michigan-Huron.

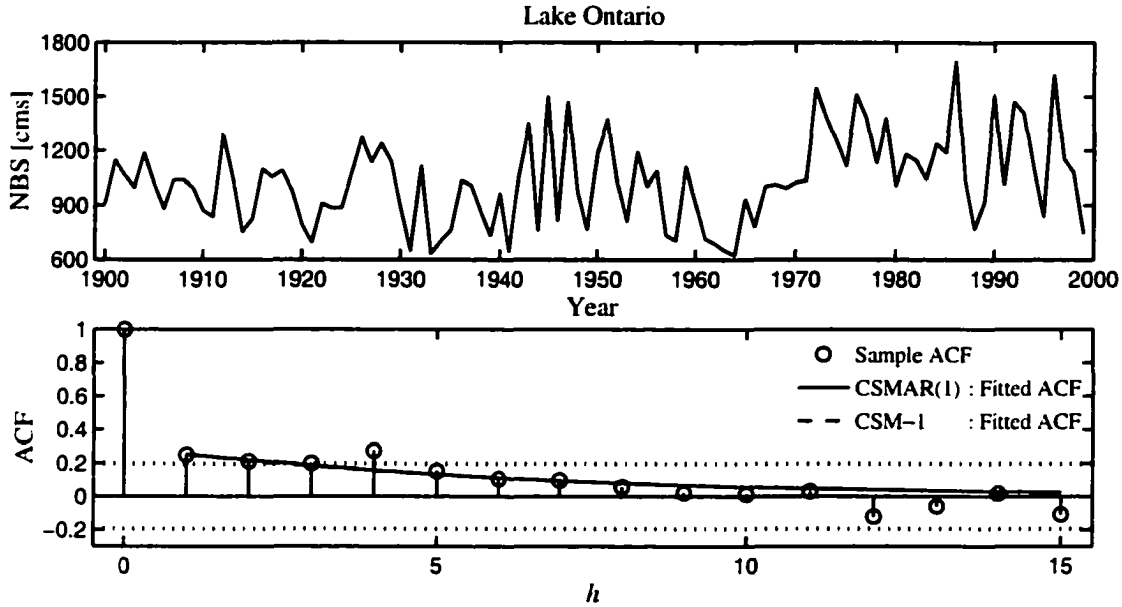

Figure 7.3: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Ontario.

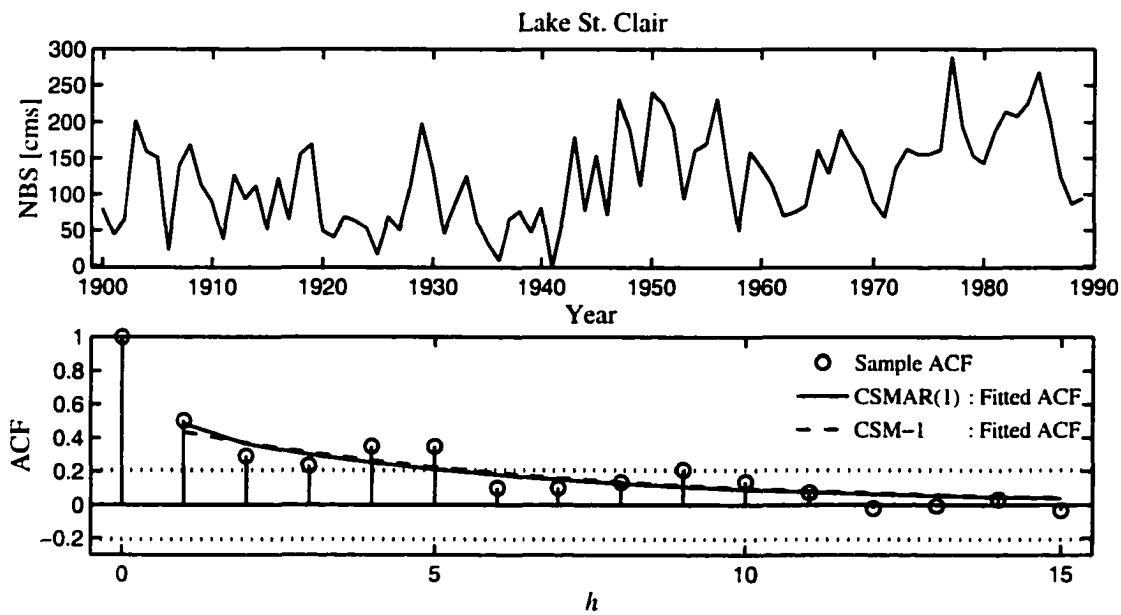

Figure 7.4: Net Basin Supply series (1900–1989) and the autocorrelation function for Lake St. Clair.

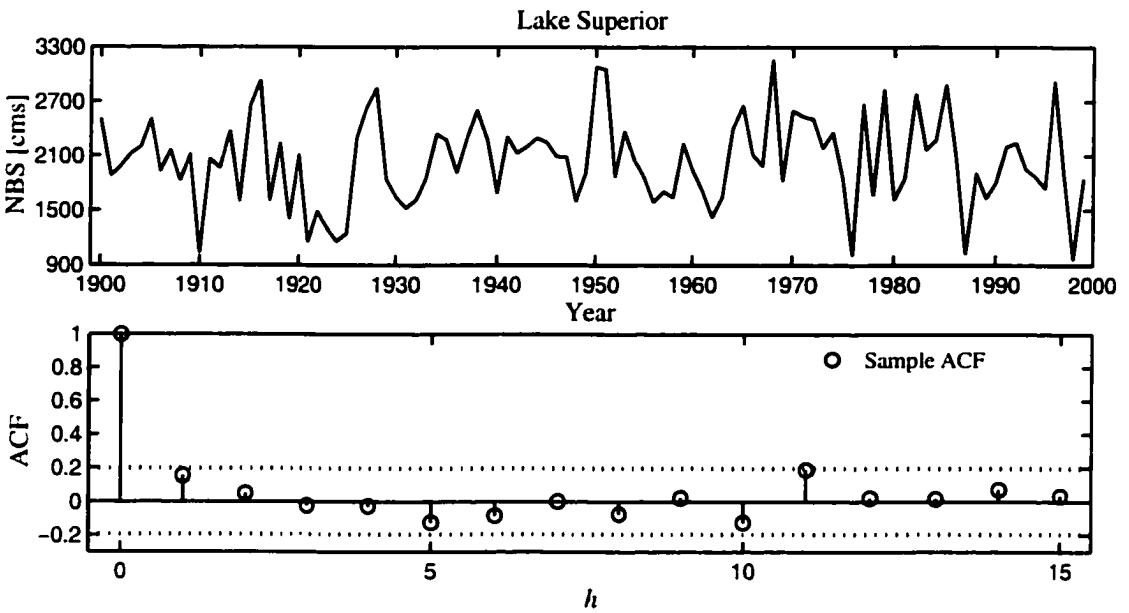

Figure 7.5: Net Basin Supply series (1900–1999) and the autocorrelation function for Lake Superior.

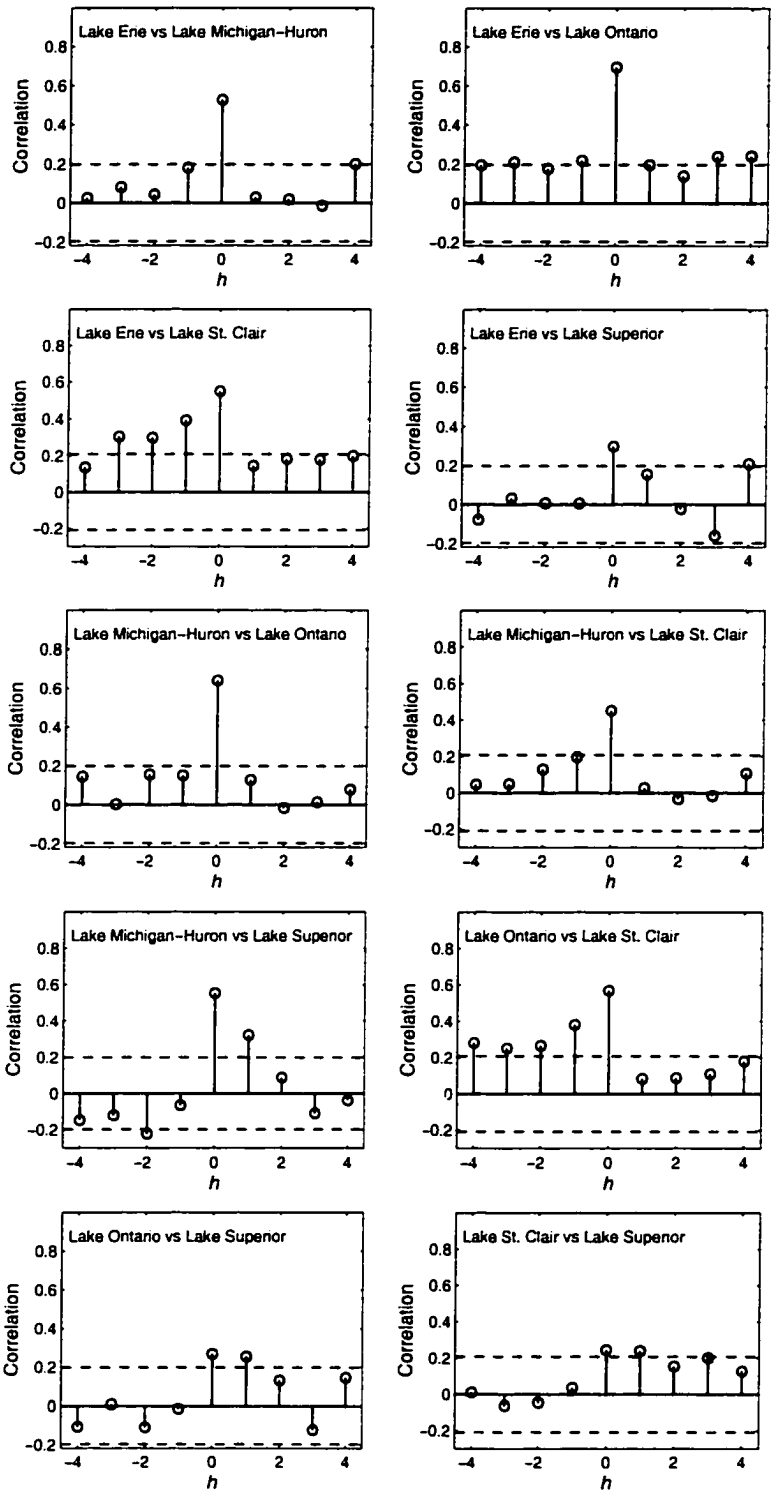

Figure 7.6: Cross correlation between the lakes in the Great Lakes system.

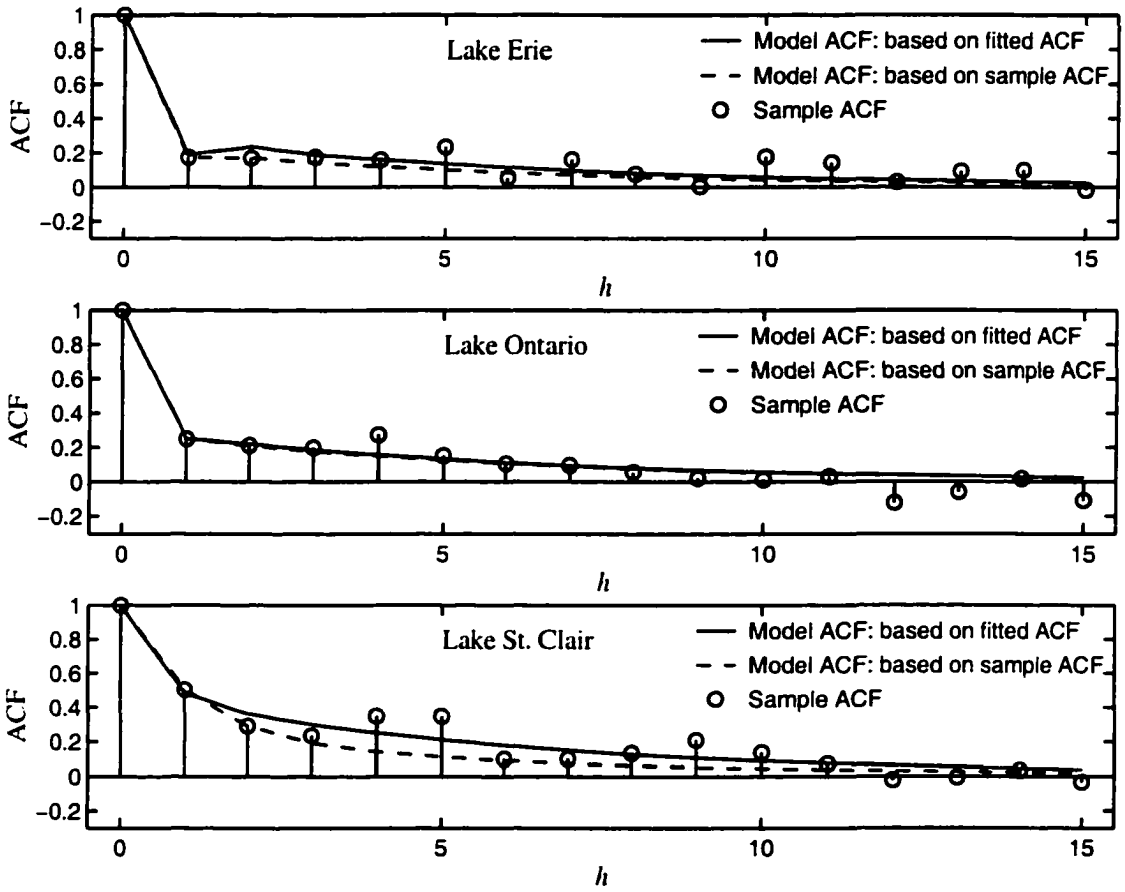

Figure 7.7: The sample ACFs of Lake Erie, Lake Ontario, and Lake St. Clair.

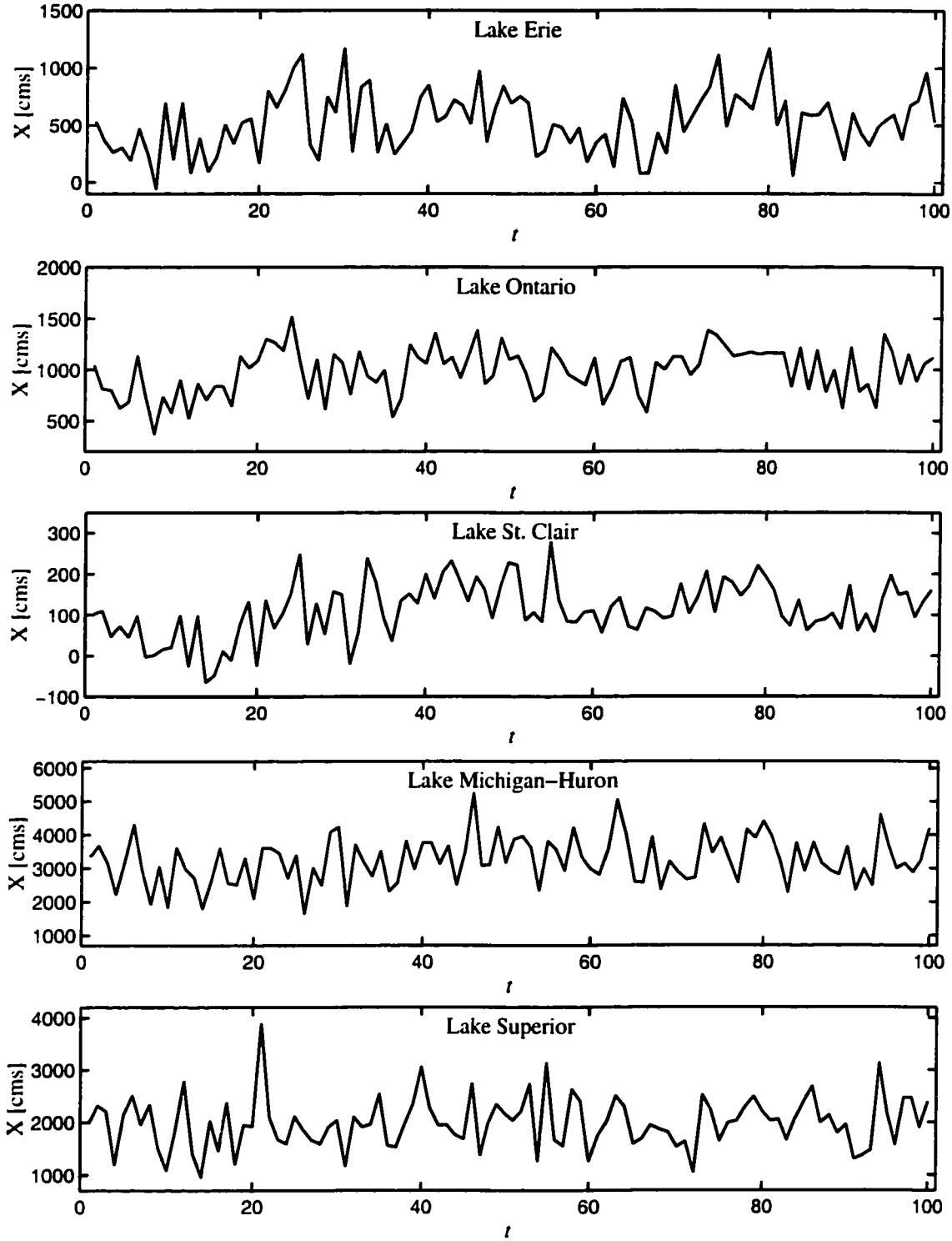

Figure 7.8: Generated sequences of length 100 based on the fitted CARMA-CSMAR(1) model in Table 7.3.

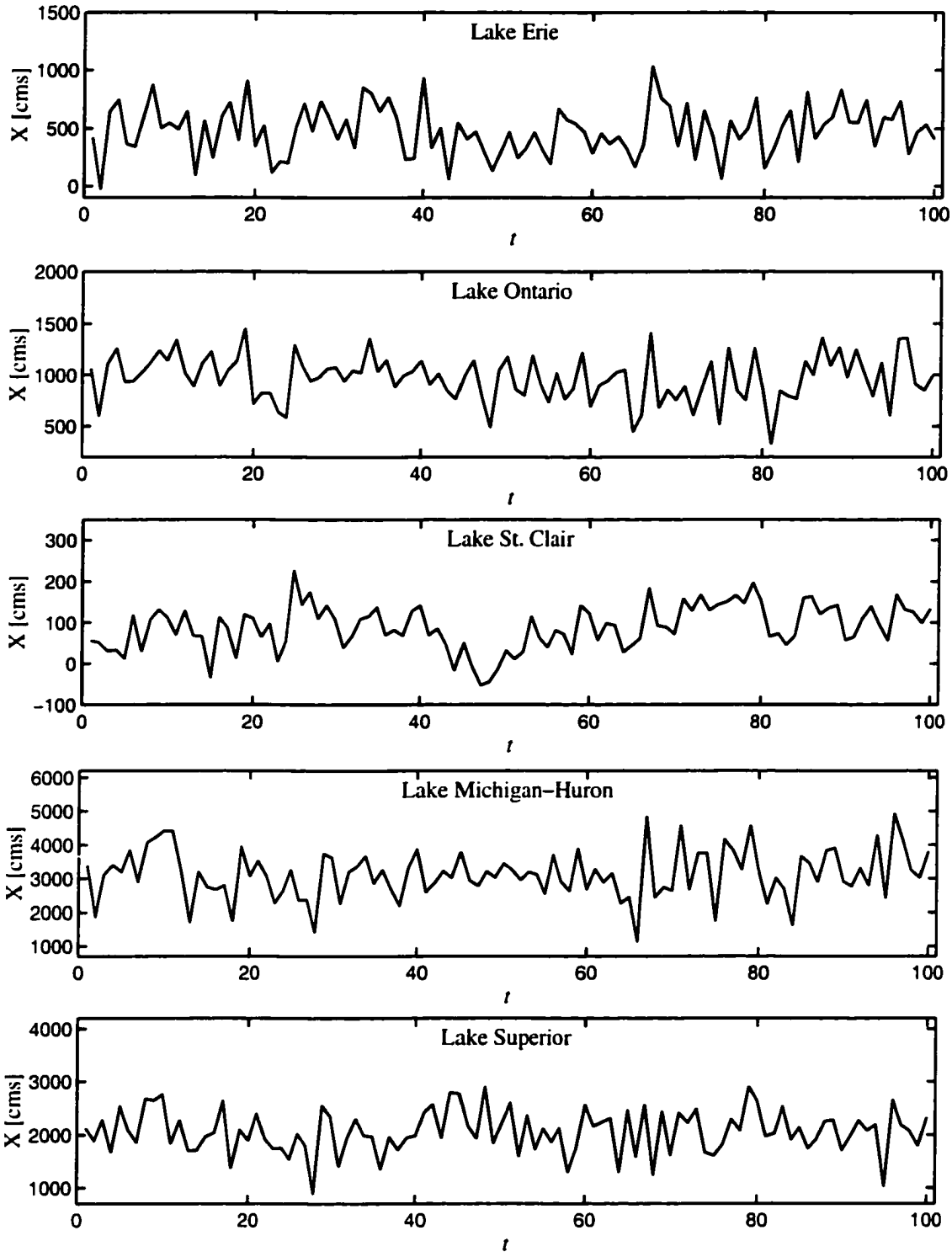

Figure 7.9: Generated sequences of length 100 based on the fitted CARMA-CSMAR(1) model in Table 7.3.

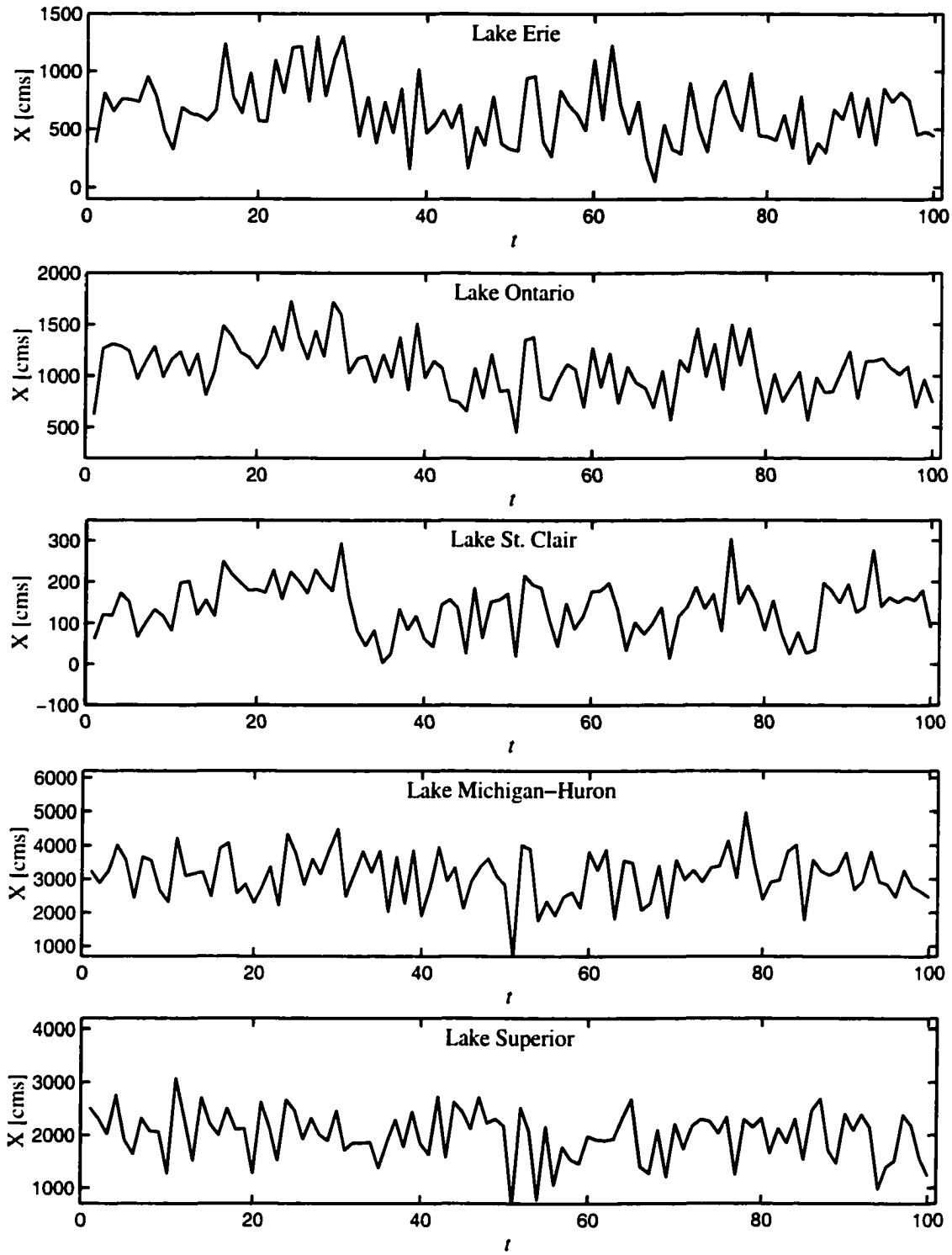

Figure 7.10: Generated sequences of length 100 based on the fitted CARMA-CSMAR(1) model in Table 7.3.

Chapter 8

Further Remarks and Recommendations

The chapters in this dissertation were written as stand alone. For comprehensive final remarks and conclusions about the different methods and procedures studied in this dissertation, refer to the corresponding chapters.

In Chapters 1 and 2 the population index flood (PIF) was introduced as an alternative to traditional index flood methods for regional frequency analysis. The PIF is purely analytic and its name arises from using population quantities in estimation of the index flood. As a result, regional homogeneity is embedded in the structure of the parameter space of the underlying regional distribution. In Chapter 1 PIF models were developed for commonly used two- and three-parameter distributions in hydrology. Due to the analytical structure of the PIF models, maximum likelihood (ML) can be used for parameter estimation in addition to other moment based methods. While methods based on moments, such as probability weighted moments (PWM) or traditional moments, are usually simpler to use for parameter estimation than maximum likelihood, it is not always clear how their regional counterparts should be estimated. This is demonstrated in Chapter 1, where PIF with the index flood estimated by the at-site population mean is compared with the traditional index flood method (*HW* scheme) for the case when the GEV is assumed as the underlying regional distribution. Parameters are estimated based on ML and PWM in the PIF-GEV model, and based on PWM in the *HW* scheme. Simulation experiments show, that for the PWM based methods bias of at-site GEV quantile estimators is significantly reduced when regional

PWM-ratios (or L-moment-ratios) are estimated as in Hosking et al. (1985a) as opposed to Hosking and Wallis (1997). In addition, the simulation experiments indicate that PIF with ML estimation performs better in terms of both bias and rmse of quantile estimators than PIF with PWM estimation. The simulation experiments also indicate that the approximate (non-analytical) *HW* scheme is quite robust and in some cases performs as well or perhaps better as the PIF method. In Chapter 2 it is demonstrated how the analytical structure of the PIF models can be used for deriving explicit equations based on Fischer's information for estimating the asymptotic variance of at-site quantile estimators, assuming the GEV as the underlying regional distribution model. Simulation experiments indicate that the difference in using the observed and expected information matrix is minor for estimating the asymptotic variance of at-site quantile estimators, and that both methods agree well with the observed mean-squared-error (based on simulation experiments). In addition newly suggested procedures (De Michele and Rosso, 2001) for estimating the standard error of at-site GEV quantile estimators in the *HW* scheme are shown to be quite erratic in some cases, although for the shape $\kappa > -0.2$ they can be quite useful. The PIF methods in Chapters 1 and 2 focused on the GEV distribution, where PIF models for other distributions were derived but not applied. Thus, there is still room to extend or repeat the studies in Chapters 1 and 2 for other distributions than the GEV.

In Chapter 3 procedures were developed for estimation of extreme upper quantiles when the observed sample appears to come from a mixed population. A Pareto model was used in fitting the upper tail of the empirical growth curves, where only the largest sample order statistics were used for parameter estimation based on maximum likelihood. The explicit formulas for the parameter estimators made it possible to derive exact equations for the mean-squared-errors of quantile estimators. The procedures were extended for use in a regional context utilizing the PIF. An obvious extension of the Pareto model in Chapter 3 would be to include historical information into the modeling framework. The cost of using historical information is that explicit equations for parameter estimators based on ML can

not be derived, hence, parameters have to be estimated numerically. As a result, exact formulas for the mean-squared-error of quantile estimators can not be derived, but instead procedures for estimating the asymptotic variance of quantile estimators can be developed in similar manner as is done in Chapter 2.

Chapters 4–7 focused on the so-called shifting mean models (SM models), which are capable of modeling autocorrelated processes that show a type of non-stationarity in the mean represented by sudden shifts. The causes for these shifts can be for example climatic fluctuations or other hydroclimatic or geophysical changes that can directly affect the process under consideration. In Chapters 4 and 5 the autocorrelation structure of the process under consideration was assumed to arise solely from the sudden shifting pattern, while in Chapters 6 and 7 the autocorrelation structure was assumed to arise also from an AR(1) persistence in the natural variability of the process itself. In Chapter 4 the general shifting mean model was introduced and its characteristics were investigated. Characteristics such as, storage related statistics, quantiles and characteristics other than moments and autocorrelation can not be calculated explicitly from fitted SM models. Given a fitted SM model, these characteristics have to be estimated using simulation experiments. Chapter 5 extended the studies of Chapter 4 for modeling of skewed SM processes. In Chapter 6 the SMAR(1) model (SM model + AR(1) persistence) was developed, and in Chapter 7 multivariate versions of the previously introduced SM models were developed capable of preserving the lag zero cross correlation in space between different sites. For example the multivariate version of the SMAR(1) model was dubbed as the contemporaneous SMAR(1) model or CSMAR(1). Furthermore, for systems where only some of the sites exhibit sudden shifting patterns while others do not. The sites not exhibiting sudden shifts were assumed to follow a CARMA(p, q) model, and mixture models such as the CSMAR(1)-CARMA were developed capable of preserving the lag zero cross correlation in space between all sites. The applicability of the SM models was demonstrated using numerous examples such as the Great Lakes system. In general the SM models were shown to be capable of preserving key

statistical characteristics in addition to the lag zero spatial correlation. The applicability of the different shifting mean models in Chapters 4–7 can be further investigated by fitting them to various hydroclimatic and geophysical systems and comparing the results with other traditional models. In addition, the shifting mean models can be extended for modeling of quasi-periodic processes by using different distributions than the geometric for modeling of the lengths of the random time spans between shifts. Such distributions that are able to reproduce periodic behavior in the autocorrelation function are for example the Poisson and the binomial.

REFERENCES

Angel, J. R. and Huff, F. A. (1997). Changes in heavy rainfall in midwestern United States. *Journal of Water Resources Planning and Management*, 123(4):246–249.

Ballerini, R. and Boes, D. C. (1985). Hurst behavior of shifting level processes. *Water Resources Research*, 21(11):1642–1648.

Boes, D. C. (1988). Schemes exhibiting hurst behavior. In Srivastava, J. N., editor, *Essays in Honor of Franklin A. Graybill*, Probability and Statistics, pages 21–42. Elsevier Science.

Boes, D. C., Heo, J., and Salas, J. D. (1989). Regional flood quantile estimation for a Weibull model. *Water Resources Research*, 25(5):979–990.

Boes, D. C. and Salas, J. D. (1978). Nonstationarity of the mean and the Hurst phenomenon. *Water Resources Research*, 14(1):135–143.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer, first edition.

Buchberger, S. (1992). Modeling and forecasting Great Lakes monthly net basin supplies. Technical report, U.S. Army Corps of Engineers, Detroit, MI. 48pp.

Buchberger, S. (1994). Modeling and forecasting of Great Lakes annual net basin supplies. *Water Resources Research*, 30(10):2725–2735.

Chiew, F. H. S. and McMahon, T. A. (1993). Detection of trend or change in annual flow of Australian rivers. *International Journal of Climatology*, 13:643–653.

Dalrymple, T. (1960). Flood frequency analysis. Water Supply Paper 1543-A, U.S. Geological Survey, Washington, D.C.

De Michele, C. and Rosso, R. (2001). Uncertainty assessment of regionalized flood frequency estimates. *Journal of Hydrologic Engineering*, 6(6):453–459.

Elíasson, J. (1997). A statistical model for extreme precipitation. *Water Resources Research*, 33(3):449–455.

Eltahir, E. A. B. (1996). El Niño and the natural variability in the flow of the Nile River. *Water Resources Research*, 32(1):131–137.

Fernández. B. and Salas, J. D. (1999). Return period and risk of hydrologic events. II: Applications. *Journal of Hydrologic Engineering*, 4(4):308–316.

Gray, W. M., Landsea, C. W., Mielke, P. W., Berry, K. J., and Blake, E. (2000). Extended range forecast of atlantic seasonal hurricane activity and US landfall strike probability for 2001. *http://typhon.atmos.colostate.edu/forecasts/2001/fcst2001/index.html*, pages 1–22.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5):1049–1054.

Grygier, J. C. and Stedinger, J. R. (1990). *SPIGOT a Synthetic Streamflow Generation Software Package, Technical Description, Version 2.6.* Cornell University, Ithaca, New York.

Hamlet, A. F. and Lettenmaier, D. P. (1999). Columbia river streamflow forecasting based on Enso and Pdo climate signals. *Journal of Water Resources Planning and Management*, 125(6):333–341.

Hipel, K. W. and McLeod, A. I. (1994). *Time Series Modelling of Water Resources and*

Environmental Systems. Number 45 in Developments in Water Science. Elsevier, first edition.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B*, 52(1):105–124.

Hosking, J. R. M. and Wallis, J. R. (1997). *Regional Frequency Analysis: An Approach Based on L-moments.* Cambridge University Press, first edition.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985a). An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. *Hydrological Sciences Journal*, 30(1):85–109.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985b). Estimation of the Generalized Extreme Value Distribution by the method of probability weighted moments. *Technometrics*, 27(3):251–261.

Hurst, H. E. (1957). A suggested statistical model of some time series which occur in nature. *Nature.* 180:494.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continous Univariate Distributions, Volume 1.* New York: Wiley, second edition.

Kabat, P., editor (2002). *Vegetation, water, humans and the climate: A new perspective on an interactive system.* A Synthesis of the IGBP Core Project, Biosphere Aspects of the Hydrological Cycle, in preparation.

Kite, G. (1989). Use of time series analysis to detect climate change. *Journal of Hydrology*, 111:259–279.

Klemes, V. (1974). The Hurst phenomenon: a puzzle? *Water Resources Research*, 10(4):675–688.

Klemes, V. (1987). Hydrological and engineering relevance of flood frequency analysis. In Singh, V. P., editor, *Hydrologic Frequency Modeling*, pages 1–18. Reidel Publishing Company.

Leiva, R. A. (1983). *Properties, Convergence and Range Behavior of Shifting Level Processes.* PhD thesis, Colorado State University.

Linsley, R. K., Kohler, M. A., and Paulhaus, J. L. H. (1982). *Hydrology for Engineers.* McGraw-Hill Book Company, third edition.

Loucks, E. (1989). *Modeling the Great Lakes Hydrologic–Hydraulic System.* PhD thesis, University of Wisconsin.

Lu, L.-H. and Stedinger, J. R. (1992a). Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test. *Journal of Hydrology*, 138(1/2):223–245.

Lu, L.-H. and Stedinger, J. R. (1992b). Variance of two and three-parameter GEV/PWM quantile estimators: Formulae, confidence intervals, and a comparison. *Journal of Hydrology*, 138(1/2):247–267.

Mantua, N., Hare, S., Zhang, Y., Wallace, J. M., and Francis, R. (1997). A Pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, 78(6):1069–1079.

Matalas, N. C. (1997). Stochastic hydrology in the context of climate change. *Climatic Change*, 37:89–101.

NERC (1975). *Flood Studies Report Vols I-V.* Natural Environment Research Council, London.

Niebauer, H. J. (1998). Variability in Bering Sea ice cover as affected by a regime shift in the North Pacific in the period 1947–1996. *Journal of Geophysical Research*, 103(C12):27,717–27,737.

Obeysekera, J. T. B. (1981). *Run and Range Analysis of Shifting Level Models.* PhD thesis, Colorado State University.

Potter, K. W. (1976). *A stochastic model of the Hurst phenomenon: Nonstationarity in hydrologic processes.* PhD thesis, John Hopkins University.

Prescott, P. and Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, 67(3):723–724.

Prescott, P. and Walden, A. T. (1983). Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. *Journal of Statistical Computation and Simulation*, 16:241–250.

Quinn, F. H. (1985). Temporal effects of St. Clair river dredging on lakes St. Clair and Erie water levels and connecting channel flow. *Journal of Great Lakes Research*, 11(3):400–403.

Rassam, J.-C., Faherazzi, L. D., Bobée, B., Mathier, L., Roy, R., and Carballada, L. (1992). Beauharnois-Les Cedres spillway: Design flood study with stochastic approach. Final report to the experts committee, Hydro-Québec, Montreal, Quebec, Canada. 105pp.

Saada, N. (1998). *Modeling the Uncertainty of Hydrologic Processes Exhibiting Changes.* PhD thesis, Colorado State University.

Salas, J. D. (1993). *Analysis and Modeling of Hydrologic Time Series*, chapter 19. Handbook of Hydrology. McGraw-Hill.

Salas, J. D. and Boes, D. C. (1980). Shifting level modelling of hydrologic series. *Advances in Water Resources*, 3(2):59–63.

Smith, K. (1996). *Environmental Hazards: Assessing Risk and Reducing Disaster.* Routledge, London.

Stedinger, J. (1983). Estimating a regional flood frequency distribution. *Water Resources Research*, 19(2):503–510.

Sveinsson, O. G. B. and Salas, J. D. (2001). Comparison of the population index flood and the index flood methods using extreme precipitation data in colorado. In Ramirez, J. A., editor, *Proceedings of the Twenty First Annual A.G.U. Hydrology Days*, pages 1–12.

Sveinsson, O. G. B., Salas, J. D., and Boes, D. C. (2002a). Regional frequency analysis of extreme precipitation in northeastern Colorado and the Fort Collins flood of 1997. *Journal of Hydrologic Engineering*, 7(1):49–63.

Sveinsson, O. G. B., Salas, J. D., Boes, D. C., and Pielke, R. A. (2002b). Modeling the dynamics of long term variability hydroclimatic processes. *Journal of Hydroclimatology*, submitted.

Taylor, K. (1999). Rapid climate change. *American Scientist*, 87:320–327.

Viessman, W. and Lewis, G. L. (1996). *Introduction to Hydrology*. HarperCollins College Publishers, fourth edition.

Waylen, P. R. and Caviedes, C. N. (1986). El Niño and annual floods on the north Peruvian littoral. *Journal of Hydrology*, 89(1/2):141–156.

Yevjevich, V. (1975). Generation of hydrologic samples-case study of the Great Lakes. Hydrology Paper 72, Colorado State University, Fort Collins, CO. 39pp.

Appendix A

DERIVATIVES IN THE PIF 1 and PIF 2 METHODS

A.1 PIF 1 Derivatives

The partial derivatives of the log-likelihood for the PIF 1 model in Eq (2.11) with respect to the parameters are given by

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial \theta_j} &= \frac{n_j}{\theta_j} - \gamma \sum_{i=1}^{n_j} x_{ji} \Delta_{ji} \quad , j = 1, \ldots, m \\
\frac{\partial \ln \mathcal{L}}{\partial \gamma} &= \sum_{j=1}^{m} \left\{ \frac{n_j}{\gamma} - \sum_{i=1}^{n_j} (\theta_j x_{ji} - 1) \Delta_{ji} \right\} \\
\frac{\partial \ln \mathcal{L}}{\partial \kappa} &= \frac{1}{\kappa} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left\{ \frac{1}{\kappa} \Lambda_{ji} - \gamma (\theta_j x_{ji} - 1) \Delta_{ji} \right\}
\end{aligned} \tag{A.1}$$

where $\Delta_{ji} = \zeta_{ji}^{-1}(1 - \kappa - \zeta_{ji}^{1/\kappa})$ and $\Lambda_{ji} = (\zeta_{ji}^{1/\kappa} - 1) \ln \zeta_{ji}$, with $\zeta_{ji} = 1 - \gamma \kappa (\theta_j x_{ji} - 1)$. So the elements of the Fisher sample information matrix, *SI*, of the ML-estimators are given by

$$\begin{aligned}
-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_j^2} &= \frac{n_j}{\theta_j^2} + \gamma \sum_{i=1}^{n_j} x_{ji} \frac{\partial \Delta_{ji}}{\partial \theta_j} \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} &= 0 \quad , i \neq j \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \gamma \partial \theta_j} &= \sum_{i=1}^{n_j} x_{ji} \left(\Delta_{ji} + \gamma \frac{\partial \Delta_{ji}}{\partial \gamma} \right) \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa \partial \theta_j} &= \gamma \sum_{i=1}^{n_j} x_{ji} \frac{\partial \Delta_{ji}}{\partial \kappa} \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \gamma^2} &= \sum_{j=1}^{m} \left\{ \frac{n_j}{\gamma^2} + \sum_{i=1}^{n_j} (\theta_j x_{ji} - 1) \frac{\partial \Delta_{ji}}{\partial \gamma} \right\}
\end{aligned} \tag{A.2}$$

$$-\frac{\partial^2 \ln \mathcal{L}}{\partial\kappa\partial\gamma} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_j x_{ji} - 1)\frac{\partial\Delta_{ji}}{\partial\kappa}$$

$$-\frac{\partial^2 \ln \mathcal{L}}{\partial\kappa^2} = \frac{1}{\kappa}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left\{\frac{1}{\kappa}\left(\frac{2}{\kappa}\Lambda_{ji} - \frac{\partial\Lambda_{ji}}{\partial\kappa}\right) - \gamma(\theta_j x_{ji} - 1)\left(\frac{1}{\kappa}\Delta_{ji} - \frac{\partial\Delta_{ji}}{\partial\kappa}\right)\right\}$$

where

$$\frac{\partial\Delta_{ji}}{\partial\theta_j} = \gamma(1-\kappa)x_{ji}\zeta_{ji}^{-2}(\kappa + \zeta_{ji}^{1/\kappa}) \tag{A.3}$$

$$\frac{\partial\Delta_{ji}}{\partial\gamma} = (\theta_j x_{ji} - 1)(1-\kappa)\zeta_{ji}^{-2}(\kappa + \zeta_{ji}^{1/\kappa}) \tag{A.4}$$

$$\frac{\partial\Delta_{ji}}{\partial\kappa} = \zeta_{ji}^{-1}\left(\frac{1}{\kappa^2}\zeta_{ji}^{1/\kappa}\ln(\zeta_{ji}) - 1\right) + \frac{\gamma}{\kappa}\frac{\partial\Delta_{ji}}{\partial\gamma} \tag{A.5}$$

and

$$\frac{\partial\Lambda_{ji}}{\partial\kappa} = \gamma(\theta_j x_{ji} - 1)\zeta_{ji}^{-1}\left[1 - \zeta_{ji}^{1/\kappa}\left(1 + \frac{1}{\kappa}\ln\zeta_{ji}\right)\right] - \frac{1}{\kappa^2}\zeta_{ji}^{1/\kappa}\ln^2\zeta_{ji} \tag{A.6}$$

In addition, under this parameterization the elements of the gradient of the GEV *qth* quantile in Eq (2.6) for site j, $j = 1, \ldots, m$, are

$$\begin{aligned}
\frac{\partial\xi_j(q)}{\partial\theta_j} &= -\frac{1}{\theta_j^2\gamma\kappa}\left[\gamma\kappa + 1 - (-\ln q)^k\right] \\
\frac{\partial\xi_j(q)}{\partial\gamma} &= -\frac{1}{\theta_j\gamma^2\kappa}\left[1 - (-\ln q)^k\right] \\
\frac{\partial\xi_j(q)}{\partial\kappa} &= -\frac{1}{\theta_j\gamma\kappa^2}\left[1 - (-\ln q)^k + \kappa\ln(-\ln q)(-\ln q)^k\right]
\end{aligned} \tag{A.7}$$

A.2 PIF 2 Derivatives

In the same way as for the PIF 1 model in appendix A.1, the first partial derivatives of the log-likelihood of the PIF 2 model in Eq (2.13) are given by

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial\alpha_j} &= \theta_j\sum_{i=1}^{n_j}\Delta_{ji} \quad , j = 1, \ldots, m \\
\frac{\partial \ln \mathcal{L}}{\partial\theta_j} &= \frac{n_j}{\theta_j} - \sum_{i=1}^{n_j}(x_{ji} - \alpha_j)\Delta_{ji} \quad , j = 1, \ldots, m \\
\frac{\partial \ln \mathcal{L}}{\partial\kappa} &= \frac{1}{\kappa}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left\{\frac{1}{\kappa}\Lambda_{ji} - \theta_j(x_{ji} - \alpha_j)\Delta_{ji}\right\}
\end{aligned} \tag{A.8}$$

where $\Delta_{ji} = \zeta_{ji}^{-1}(1-\kappa-\zeta_{ji}^{1/\kappa})$ and $\Lambda_{ji} = (\zeta_{ji}^{1/\kappa}-1)\ln\zeta_{ji}$ as before, but with $\zeta_{ji} = 1 - \theta_j\kappa(x_{ji}-\alpha_j)$. The elements of the Fisher sample information matrix, SI, of $\boldsymbol{\theta}$ are then given by

$$
\begin{aligned}
-\frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_j^2} &= -\theta_j \sum_{i=1}^{n_j} \frac{\partial \Delta_{ji}}{\partial \alpha_j} \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_i \partial \alpha_j} &= 0 \quad , i \neq j \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_j \partial \alpha_j} &= -\sum_{i=1}^{n_j} \left(\Delta_{ji} + \theta_j \frac{\partial \Delta_{ji}}{\partial \theta_j} \right) \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa \partial \theta_j} &= -\theta_j \sum_{i=1}^{n_j} \frac{\partial \Delta_{ji}}{\partial \kappa} \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_j^2} &= \frac{n_j}{\theta_j^2} + \sum_{i=1}^{n_j} (x_{ji}-\alpha_j) \frac{\partial \Delta_{ji}}{\partial \theta_j} \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} &= 0 \quad , i \neq j \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa \partial \theta_j} &= \sum_{i=1}^{n_j} (x_{ji}-\alpha_j) \frac{\partial \Delta_{ji}}{\partial \kappa} \\
-\frac{\partial^2 \ln \mathcal{L}}{\partial \kappa^2} &= \frac{1}{\kappa} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left\{ \frac{1}{\kappa} \left(\frac{2}{\kappa} \Lambda_{ji} - \frac{\partial \Lambda_{ji}}{\partial \kappa} \right) - \theta_j (x_{ji}-\alpha_j) \left(\frac{1}{\kappa} \Delta_{ji} - \frac{\partial \Delta_{ji}}{\partial \kappa} \right) \right\}
\end{aligned}
\tag{A.9}
$$

where the partial derivatives of Δ_{ji} and Λ_{ji} in Eq (A.9) are given by

$$\frac{\partial \Delta_{ji}}{\partial \alpha_j} = -\theta_j(1-\kappa)\zeta_{ji}^{-2}(\kappa + \zeta_{ji}^{1/\kappa}) \tag{A.10}$$

$$\frac{\partial \Delta_{ji}}{\partial \theta_j} = (x_{ji}-\alpha_j)(1-\kappa)\zeta_{ji}^{-2}(\kappa + \zeta_{ji}^{1/\kappa}) \tag{A.11}$$

$$\frac{\partial \Delta_{ji}}{\partial \kappa} = \zeta_{ji}^{-1} \left(\frac{1}{\kappa^2} \zeta_{ji}^{1/\kappa} \ln(\zeta_{ji}) - 1 \right) + \frac{\theta_j}{\kappa} \frac{\partial \Delta_{ji}}{\partial \theta_j} \tag{A.12}$$

and

$$\frac{\partial \Lambda_{ji}}{\partial \kappa} = \theta_j (x_{ji}-\alpha_j)\zeta_{ji}^{-1} \left[1 - \zeta_{ji}^{1/\kappa} \left(1 + \frac{1}{\kappa} \ln \zeta_{ji} \right) \right] - \frac{1}{\kappa^2} \zeta_{ji}^{1/\kappa} \ln^2 \zeta_{ji} \tag{A.13}$$

Under this parameterization the gradient of the GEV qth quantile in Eq (2.6) has the following elements for site j , $j = 1, \ldots, m$

$$\frac{\partial \xi_j(q)}{\partial \alpha_j} = 1$$

$$\frac{\partial \xi(q)}{\partial \theta_j} = -\frac{1}{\theta_j^2 \kappa}\left[1 - (-\ln q)^{\kappa}\right] \tag{A.14}$$

$$\frac{\partial \xi(q)}{\partial \kappa} = -\frac{1}{\theta_j \kappa^2}\left[1 - (-\ln q)^{\kappa} + \kappa \ln(-\ln q)(-\ln q)^{\kappa}\right]$$

Appendix B

STATIONARITY OF $\mathbf{Z}_t$ IN CHAPTER 7.2

Let us first look at the general case in which $\{\mathbf{Z}_t\}$ in section 7.2 is non-stationary in the covariance. In this case $N_1^{(i)}, N_2^{(i)}, \ldots \overset{iid}{\sim} \text{posgem}(p^{(i)})$ for $i = 1, 2, \ldots, n$, and $S_r^{(i)} = N_1^{(i)} + N_2^{(i)} + \ldots + N_r^{(i)}$ with $S_0^{(i)} = 0$. It follows $S_r^{(i)}$ has the negative binomial distribution with parameters r and $p^{(i)}$. For this general case the ACVF of $\mathbf{Z}_t$ at lag $h \geq 0$ between two sites i and j is

$$\begin{aligned} c_{\mathbf{Z}}^{ij}(h) &= E[Z_{t+h}^{(i)} Z_{t+h}^{(j)}] \\ &= E\left[\sum_{k=1}^{t+h} M_k^{(i)} I_{(S_{k-1}^{(i)}, S_k^{(i)}]}(t+h) \sum_{i=r}^{t} M_r^{(j)} I_{(S_{r-1}^{(j)}, S_r^{(j)}]}(t)\right] \end{aligned} \tag{B.1}$$

where we have substituted for $Z_t^{(i)}$ and $Z_t^{(j)}$ from Eq (7.2). Since $M_k^{(i)}$ and $M_r^{(j)}$ are only correlated at lag zero, and $E[\mathbf{M}] = \mathbf{0}$, then

$$\begin{aligned} c_{\mathbf{Z}}^{ij}(h) &= \sum_{k=1}^{t} E\left[M_k^{(i)} I_{(S_{k-1}^{(i)}, S_k^{(i)}]}(t+h) M_k^{(j)} I_{(S_{k-1}^{(j)}, S_k^{(j)}]}(t)\right] \\ &= \sum_{k=1}^{t} E\left\{E\left[M_k^{(i)} M_k^{(j)} I_{(S_{k-1}^{(i)}, S_k^{(i)}]}(t+h) I_{(S_{k-1}^{(j)}, S_k^{(j)}]}(t) \mid \{S_t^{(i)}\}, \{S_t^{(j)}\}\right]\right\} \\ &= c_{\mathbf{M}}^{ij}(0) \sum_{k=1}^{t} P(S_{k-1}^{(i)} < t, S_k^{(i)} \geq t+h) \cdot P(S_{k-1}^{(j)} < t, S_k^{(j)} \geq t) \end{aligned} \tag{B.2}$$

For a generic $N_1, N_2, \ldots \overset{iid}{\sim} \text{posgeom}(p)$ and $S_r = N_1 + N_2 + \ldots + N_r$ with $S_0 = 0$, it is easily shown that

$$\begin{aligned} P(S_{k-1} < t, S_k \geq t+h) &= \sum_{r=k-1}^{t-1} P(N > t+h-r) \cdot P(S_{k-1} = r) \\ &= \binom{t-1}{k-1} p^{k-1} (1-p)^{t+h-k} \end{aligned} \tag{B.3}$$

Using these results into Eq (B.2) it follows that

$$\begin{aligned} c_{\mathbf{Z}}^{ij}(h) &= c_{\mathbf{M}}^{ij}(0) \sum_{k=1}^{t} \binom{t-1}{k-1} (p^{(i)})^{k-1} (1-p^{(i)})^{t+h-k} \cdot \binom{t-1}{k-1} (p^{(j)})^{k-1} (1-p^{(j)})^{t-k} \\ &= c_{\mathbf{M}}^{ij}(0)(1-p^{(i)})^{h} \sum_{k=0}^{t-1} \binom{t-1}{k}^{2} (p^{(i)}p^{(j)})^{k} [(1-p^{(i)})(1-p^{(j)})]^{t-k-1} \end{aligned} \tag{B.4}$$

which depends on t for all $0 < (p^{(i)}, p^{(j)}) < 1$. That is for the above general case we have shown that $\mathbf{Z}_t$ is non-stationary in the covariance.

If it is now assumed that $N_1, N_2, \ldots \overset{iid}{\sim} \text{posgeom}(p)$ is a common sequence for all the n sites, then

$$\begin{aligned} c_{\mathbf{Z}}^{ij}(h) &= E\left[\sum_{k=1}^{t+h} M_k^{(i)} I_{(S_{k-1}, S_k]}(t+h) \sum_{i=r}^{t} M_r^{(j)} I_{(S_{r-1}, S_r]}(t)\right] \\ &= c_{\mathbf{M}}^{ij}(0) \sum_{k=1}^{t} P(S_{k-1} < t, S_k \geq t+h) \\ &= c_{\mathbf{M}}^{ij}(0)(1-p)^{h} \qquad h = 0, 1, \ldots \end{aligned} \tag{B.5}$$

which is equivalent to Eq (7.7), and stationary with respect to t.