NONPARAMETRIC APPROACHES FOR SIMULATION OF STREAMFLOW SEQUENCES

by

Ashish Sharma

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

Approved:

Dr. David G. Tarboton
Major Professor

Dr. Upmanu Lall
Committee Member

Dr. David S. Bowles
Committee Member

Dr. Robert W. Gunderson
Committee Member

Michael C. Minnotte

Dr. Michael Minnotte
Committee Member

James P Shaver

Dr. James P. Shaver
Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

1996

ABSTRACT

Nonparametric Approaches for Simulation of Streamflow Sequences

by

Ashish Sharma, Doctor of Philosophy

Utah State University, 1996

Major Professor: Dr. David G. Tarboton
Department: Civil and Environmental Engineering

This work describes strategies for simulation of synthetic streamflow sequences using nonparametric methods for density estimation. Nearest-neighbor and kernel density estimation methods are used. Nonparametric methods use the observed data to estimate the probability densities required for simulation. Use of these methods enables proper representation of dependence (linear or nonlinear), asymmetry, and multimodality in the probability density of streamflow. These features are not easily modeled by existing parametric streamflow simulation methods. The nonparametric models are applied to synthetic data with known statistical attributes and to actual monthly streamflow data sets. Comparisons with popular parametric models indicate that the nonparametric simulations reproduce not only the linear attributes modeled by parametric stochastic models, but also a broader set of properties based on the additional distributional information contained in the nonparametric probability estimates of streamflow.

(224 pages)

To my parents.

ACKNOWLEDGMENTS

I would like to thank my major advisor, Dr. David G. Tarboton, for the opportunity to work on this project and for guiding me throughout my stay at Utah State University. I thank him for his insightful comments and ideas, for his time and patience in discussing several aspects of the research, and for the financial support he provided during my stay in Logan. I would like to thank Dr. Upmanu Lall for the comments and suggestions he gave at all stages of the research. In addition to his work on my committee, Dr. Lall was a co-principal investigator on the project. His extensive experience with the issues pertinent to my project proved invaluable in the course of the research. I wish to thank Dr. David Bowles, Dr. Mike Minnotte, and Dr. Robert Gunderson for serving on my committee and for the useful suggestions and help they gave all through my research. I feel fortunate to have had such a committee.

I would also like to thank several friends and colleagues for their help and advice at various stages of my work. Many thanks to Balaji, Moon, Aladdin, Unni, Shaleen, Shyamal, Ashutosh, Rajeev, Liu, Sanjay, Seigo, and Tom for their insights and suggestions on several aspects of my research and for making my stay in Logan a pleasant one.

This research was supported in part by the United States Geological Survey, Department of the Interior, under USGS award number 1434-92-G-2265. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. Additional support, office space, computing facilities, and staff help provided by the Utah Water Research Laboratory are gratefully acknowledged.

Ashish Sharma

CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1. GENERAL INTRODUCTION 1

    Problem Statement 1
    Objectives 2
    Outline 3
    References 4

2. A NEAREST-NEIGHBOR BOOTSTRAP FOR RESAMPLING HYDROLOGIC TIME SERIES 6

    Abstract 6
    Introduction 6
    Background 9
    The Nearest-Neighbor Resampling Algorithm 13
    Parameters of the Nearest-Neighbor Resampling Method 15

        Choosing the Weight Function K(v) 15
        Choosing the Number of Neighbors k, and Model Order d 18

    Applications 20

        Example E1 20
        Example E2 21
        Example E3 22

    Summary and Discussion 26
    References 28

3. STREAMFLOW SIMULATION USING NONPARAMETRIC DENSITY ESTIMATION 58

    Abstract 58
    Introduction 58
    Background 61
    Kernel Density Estimation 67
    Nonparametric Order p Markov Streamflow Model, NPp 72
    Testing with Synthetic Data 77

        Tests with AR1 Data 78
        Tests with SETAR Data 79

    Application of NP1 to Simulation of Monthly Streamflow 82
    Discussion and Conclusions 86
    References 88

4. DISAGGREGATION PROCEDURES FOR STOCHASTIC HYDROLOGY BASED ON NONPARAMETRIC DENSITY ESTIMATION 111

    Abstract 111
    Introduction 112
    Background 114
    Kernel Density Estimation 117
    Nonparametric Disaggregation Model, NPD 121
    Model Evaluation 128

        SPIGOT 129
        Tests with Synthetic Data 130
        Tests with Monthly Flow Data 133

    Discussions and Conclusions 138
    References. 140

5. SMALL SAMPLE PERFORMANCE OF FOUR BANDWIDTH ESTIMATORS FOR BIVARIATE KERNEL DENSITY ESTIMATION 162

    Abstract 162
    Introduction 162
    Kernel Density Estimation 163
    Bandwidth Estimation Methods 166

        Gaussian Reference Bandwidth (GREF) 167
        Maximum Likelihood Cross Validation (MLCV) 168
        Unbiased or Least Square Cross Validation (LSCV) 169
        Biased Cross Validation (BCV) 170
        Estimation of Optimal Local Bandwidths 172

Application to Samples from Gaussian Mixture Densities 174

Results Using a Global λ 175
Results Using Local Bandwidths 177

Conclusions and Recommendations 178
References 179

6. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS 195

References 200

APPENDICES 201

APPENDIX A. STATE DEPENDENT CORRELATION COEFFICIENTS 202
APPENDIX B. DERIVATION OF MODEL STATISTICS 205
APPENDIX C. PERMISSION LETTER 210

CURRICULUM VITAE 212

LIST OF TABLES

Table Page

2-1 Statistical Comparison of k-nn and AR1 Model Simulations Applied to an AR1 Sample 32

3-1 State-Dependent Lag 1 Correlations for NP1 Simulations from a SETAR Model 92

3-2 Reservoir Capacities Evaluated for 100 AR1 and NP1 Model Realizations for a Yield of 0.8 Mean Annual Flow 93

4-1 Parameters of the Test Distribution 143

4-2 Reservoir Capacities from 100 Realizations for a Yield of 0.8 Mean Annual Flow 144

5-1 Description of the Bivariate Normal Mixture Densities Studied 181

5-2 Mean ISE for Global Bandwidth-Based Estimators 182

5-3 Sample Bias and Standard Deviation of Estimated Optimal Global Bandwidths 183

5-4 Mean ISE for Local Bandwidth-Based Estimators 184

5-5 Sample Standard Deviations of Estimated Optimal λ_p 185

LIST OF FIGURES

Figure Page

2-1 A time series from the model $x_{t+1}=(1-4(x_t-0.5)^2)$ 33

2-2 A plot of x_{t+1} vs x_t for the time series generated from the model $x_{t+1}=(1-4(x_t-0.5)^2)$ 34

2-3 Illustration of resampling weights, K(j(i)), at selected conditioning points x_i, using k=10 with a sample of size 100 from an exponential distribution with parameter =1 35

2-4 Illustration of weights K(j(i)) versus weights from the uniform kernel applied at three points selected from Figure 2-3 36

2-5 (a) A time series trace from the SETAR model described by Equation 2.11 and (b) a time series trace from a k-nn resample of the original SETAR sample 38

2-6 (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the SETAR sample, the thick curve denoting a LOWESS smooth, and (b) an average of the ASH estimates of the bivariate probability density $f(x_t, x_{t-1})$ across 100 k-nn resamples from the SETAR sample 39

2-7 The Weber river monthly streamflow time series 41

2-8 Monthly and annual statistics - (a) log(mean flow), (b) log(standard deviation), (c) skew, and (d) lag 1 correlation for simulated traces using AR1 and k-nn models for the Weber River data 42

2-9 Marginal probability densities estimated by ASH from AR1 and k-nn samples for the Weber River data 46

2-10 (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the October-November historical flows, the thick curve denoting a LOWESS smooth, (b) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 simulations from an AR1 model, and (c) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 k-nn resamples 49

2-11 (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the April-May historical flows, the thick curve denoting a LOWESS smooth, (b) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 simulations from an AR1 model, and (c) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 k-nn resamples 51

2-12 (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the May-June historical flows, the thick curve denoting a LOWESS smooth, (b) ASH estimate of the probability density of June flows conditional to a May flow of 867 cfs, (c) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 simulations from an AR1 model, and (d) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 k-nn resamples 54

3-1 Histogram and probability density estimates of July monthly streamflow (cfs) in the Beaver River at Beaver, UT (USGS station number 10234500) 94

3-2 Illustration of conditional probability density function 95

3-3 Flowchart of NP1 model 96

3-4 Marginal density estimates for NP1 simulations of an AR1 data set 97

3-5 Bivariate joint density of: (a) SETAR model in (3.5). 98

3-6 Underlying bivariate densities in NP1 model simulations for selected month pairs 101

3-7 Comparison of means of simulated and historical flows 103

3-8 Comparison of standard deviations of simulated and historical flows . . . 104

3-9 Comparison of lag 1 correlations of simulated and historical flows 105

3-10 Comparison of skewness of simulated and historical flows 106

3-11 Marginal density estimates for selected months streamflow 107

3-12 Monthly state-dependent correlations for simulated and historical flows 110

4-1 Illustration of a conditional density estimate $\hat{f}(X_1, X_2|Z)$ with $Z = X_1 + X_2$ as a slice through the joint density function 145

4-2 Bivariate distribution used in the synthetic example to test the disaggregation approach 146

4-3 Marginal distributions for the calibration and disaggregation samples from SPIGOT and NPD as represented by boxplots for components a) X_1 and b) X_2 147

4-4 Comparison of statistics of SPIGOT and NPD for the synthetic example 149

4-5 The average bivariate kernel density estimate of 101 calibration samples from the synthetic test distribution in Figure 4-2 150

4-6 The average bivariate kernel density estimate of 10100 disaggregations from 101 calibration samples 151

4-7 A bivariate kernel density estimate of July-August flows from the Snake River 152

4-8 Simulated and observed seasonal and annual mean streamflow using SPIGOT and NPD 153

4-9 Simulated and observed streamflow standard deviations using SPIGOT and NPD 154

4-10 Simulated and observed streamflow skewness coefficients using SPIGOT and NPD 155

4-11 Simulated and observed cross correlation pairs using SPIGOT and NPD 156

4-12 Simulated and observed marginal density estimates using the univariate kernel density estimator 157

4-13 Simulated and observed state-dependent correlations for sequential month pairs using SPIGOT and NPD 161

5-1 Contour plots of the bivariate test densities listed in Table 5-1 186

5-2 Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density A 187

5-3 Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density B 188

5-4 Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density C 189

5-5 Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density D 190

5-6 Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density A 191

5-7 Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density B 192

5-8 Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density C 193

5-9 Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density D 194

CHAPTER 1

GENERAL INTRODUCTION

Problem Statement

Engineers have always recognized the variability and uncertainty of hydrologic inputs in designing water resources systems. Water supply and management agencies often work in the face of this uncertainty and need carefully synthesized streamflows to properly plan reservoir operations and system expansions to meet future demands. To develop optimal operating rules and control strategies for complex hydrologic systems, one requires synthetic flows that reproduce important statistical properties of the observed streamflow record.

Streamflow simulation has traditionally been performed using linear autoregressive moving average (ARMA) models [*Bras and Rodriguez-Iturbe*, 1985]. These models treat future flows as a function of a finite set of past flows. A joint probability distribution characterizes the interdependence of the flow variables. In the case of ARMA models, this distribution is assumed to be parametric, i.e., defined explicitly by a few parameters (that can be estimated using the method of moments or maximum likelihood from the historical time series). Such a characterization limits the synthetic flows from representing "unusual" features such as nonstationarities (in mean and variance), nonlinearities (in the dependence structure of flow variables), or state dependence (for example, between low flows and succeeding high flows) that may be present in the historical data.

One way of representing such features in the synthetic flow samples is to approximate the true (unknown) probability distribution of flows. This can be done using nonparametric density estimation methods. Nonparametric methods use the historical data to estimate the probability densities needed for simulation. Nonparametric methods are designed to remain true to the data and are consistent, i.e., they converge to a broad class of possible underlying distributions as the sample size is increased. Although extensively studied in the statistics literature, they have seen limited applications in hydrology.

This study used nonparametric methods to develop models for synthetic streamflow generation. Nearest-neighbor and kernel methods were used to develop the models. Approaches for streamflow simulation based on conditioning on past occurrences as well as disaggregation of aggregate (annual) flows to disaggregate (tributary or seasonal) flows were developed. These models are a special class of autoregressive models since flow at the current time step is modeled as a function of prior flows. The dependence between the current and prior flows is characterized by a joint probability distribution, which is estimated nonparametrically, using the available historical record. This marks a deviation from the traditional parametric approaches, which impose an assumed distribution and a rigid form of dependence on the model simulations. The use of nonparametric methods results in simulations that are representative of the historical record. This, in turn, is responsible for an accurate representation of hydrologically relevant features such as storage-yield relations or drought-related characteristics in the simulated time series.

Objectives

This work focused on the application of nonparametric methods to synthetic streamflow simulation. The underlying objective of this study was to develop nonparametric alternatives to parametric stochastic models such as the ARMA, Fractional Gaussian Noise (FGN) [*Mandelbrot and van Ness*, 1968], or Broken Line processes [*Curry and Bras*, 1978], so as to free the modeler from any assumptions as to the form of the probability density function (e.g., Gaussian) or the dependence (e.g., linear or nonlinear) between flows. The specific objectives of this study were:

(1) Use nonparametric methods to develop approaches for simulation of streamflow for a given site, at annual or seasonal time scales. This problem is addressed in a generalized autoregressive framework with flow at time t being modeled as a nonparametric function (estimated based on joint and conditional density estimates) of a finite set of prior flows.

(2) Develop a nonparametric model to disaggregate aggregate (annual or main stream) flows to seasonal or tributary flows. Disaggregation models provide the ability to generate multiseason and multisite streamflow sequences that maintain the proper interdependence between aggregate and disaggregate flows, thus allowing a model to span scales in space and time. Such a dependence is modeled based on nonparametrically estimated joint probability densities of the aggregate and disaggregate flow variables.

(3) Study the statistical attributes and applicability of nonparametric methods to small samples of sizes typically encountered in hydrology.

Outline

This study is presented in a multiple-paper format and describes the investigations of the objectives described in the previous section. This work is divided into six chapters, including the introductory and the conclusion chapters. An outline of the different chapters that follow is given here.

Chapter 2 details a nonparametric model for synthetic streamflow simulation using nearest-neighbor methods. This model is a "bootstrap," i.e., it resamples the data with replacement from the historical streamflow time series. The model preserves linear and nonlinear dependence between flows and admits features such as bimodality in the probability density and state dependence (expressed in this study as the dependence of correlation on flow magnitude, see appendix A) in its simulations. The model is applied to synthetic data with known statistical attributes and to monthly streamflow.

Chapter 3 discusses a nonparametric model for streamflow simulation that uses kernel density estimation techniques. This model uses a kernel estimate of the joint distribution of flows to simulate new realizations. A general multiple order dependence is admitted in the model, though applications are constrained to the simplest (order one) case. Suggestions for dealing with boundary problems (negative realizations) are included. The model is tested and applied to data from linear and nonlinear models and to monthly streamflows.

Chapter 4 describes a nonparametric approach for disaggregating aggregate annual flows to seasonal or tributary components. The kernel density estimation methodology is used to develop the model. The model is applied to synthetic data from a known nonlinear model and also to disaggregate annual flows to monthly flows. The order one nonparametric model described in chapter 3 is used to simulate the annual streamflows that drive this disaggregation model.

Chapter 5 discusses the kernel density estimation methodology used in the models described in chapters 3 and 4. A study of the efficiency of methods to estimate the parameters of a kernel density estimate is performed. Results from this study were used to choose the kernel density estimators in chapters 3 and 4.

The models in chapters 2, 3, and 4 are useful nonparametric alternatives to existing models for streamflow simulation. A brief discussion of each of the models described in chapters 2, 3, and 4 is presented in chapter 6.

References

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, Mass., 1985.

Curry, K., and R. L. Bras, Theory and applications of the multivariate broken line, dissaggregation and monthly autoregressive streamflow generators to the Nile River, *Technology Adaptation Program*, MIT, Cambridge, Mass., 1978.

Mandelbrot, B. B., and J. W. van Ness, Fractional brownian motions, fractional noises and applications, *SIAM Rev.*, 10(4), 422-437, 1968.

CHAPTER 2

A NEAREST-NEIGHBOR BOOTSTRAP FOR RESAMPLING HYDROLOGIC TIME SERIES[1]

Abstract

A nonparametric method for resampling scalar or vector valued time series is introduced. Multivariate nearest-neighbor probability density estimation provides the basis for the resampling scheme developed. The motivation for this work comes from a desire to preserve the dependence structure of the time series while bootstrapping (resampling it with replacement). The method is data driven, and is to be preferred when the investigator is uncomfortable with prior assumptions as to the form (e.g., linear or nonlinear) of dependence and the form of the probability density function (e.g., Gaussian). Such prior assumptions are often made in an ad hoc manner for analyzing hydrologic data. Connections of the nearest-neighbor bootstrap to Markov processes as well as its utility in a general Monte Carlo setting are discussed. Applications to resampling monthly streamflow and some synthetic data are presented. The resampling method is shown to be effective with time series generated by linear and nonlinear Autoregressive models. The utility of the method for resampling monthly streamflow sequences with asymmetric and bimodal marginal probability densities is also demonstrated.

Introduction

Autoregressive Moving Average (ARMA) models for time series analysis are often used by hydrologists to generate synthetic streamflow and weather sequences to aid in the analysis of reservoir and drought management. Hydrologic time series can exhibit the

---

[1] Coauthored by Ashish Sharma and Upmanu Lall, Water Resources Research, 32(3), 679-693, 1996, Copyright by the American Geophysical Union.

following behavior, which can be a problem for linear ARMA models that are commonly used:

(1) asymmetric and/or multimodal conditional and marginal probability distributions;

(2) persistent large amplitude variations at irregular time intervals;

(3) amplitude-frequency dependence (e.g., the amplitude of the oscillations increases as the oscillation period increases);

(4) apparent long memory (this could be related to (b) and/or (c));

(5) nonlinear dependence between x_t vs $x_{t-\tau}$ for some lag τ;

(6) time irreversibility (i.e.,the time series plotted in reverse time is "different" from the time series in forward time). The physics of most geophysical processes is time irreversible. Streamflow hydrographs often rise rapidly, and attenuate slowly, leading to time irreversibility.

Kendall and Dracup [1991] have argued for simple resampling schemes, such as the index sequential method, for streamflow simulation in place of ARMA models, suggesting that the ARMA streamflow sequences usually do not "look" like real streamflow sequences. Some alternatives [*Yakowitz*, 1973; *Yakowitz*, 1979; *Schuster and Yakowitz*, 1979; *Yakowitz*, 1985; *Karlsson and Yakowitz*, 1987a; 1987b; *Smith*, 1991; *Smith et al.*, 1992;*Tarboton et al.*, 1993; *Yakowitz*, 1993] that consider the time series as the outcome of a Markov process and estimate the requisite probability densities using nonparametric methods are available. A resampling technique or bootstrap for scalar or vector valued, stationary, ergodic time series data that recognizes the serial dependence structure of the time series is presented here. The technique is nonparametric, i.e., no prior assumptions as to the distributional form of the underlying stochastic process are made.

The bootstrap [*Efron*, 1979; *Efron and Tibishirani*, 1993] is a technique that prescribes a data resampling strategy using the random mechanism that generated the data. Its applications for estimating confidence intervals and parameter uncertainty are well known [*Tasker*, 1987; *Hardle and Bowman*, 1988; *Woo*, 1989; *Zucchini and Adamson*, 1989]. Usually the bootstrap resamples with replacement from the empirical distribution function of independent, identically distributed data. The contribution of this chapter is the development of a bootstrap for dependent data that preserves the dependence in a probabilistic sense. This method should be useful for the Monte Carlo analysis of a variety of hydrologic design and operation problems where time series data on one or more interacting variables are available.

The underlying concept of the methodology is introduced through Figure 2-1. Consider that the serial dependence is limited to the two previous lags, i.e., x_t depends on the two prior values x_{t-1} and x_{t-2}. Denote this ordered pair or bi-tuple at a time t_i by $\mathbf{D}_i$. Let the corresponding succeeding value be denoted by S. Consider the k nearest neighbors of $\mathbf{D}_i$ as the k bi-tuples in the time series that are closest in terms of Euclidean distance to $\mathbf{D}_i$. The first three nearest neighbors are marked as $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$. The expected value of the forecast S can be estimated as an appropriate weighted average of the successors x_t (marked as 1, 2, and 3, respectively) to these three nearest neighbors. The weights may depend inversely on the distance between $\mathbf{D}_i$ and its k nearest neighbors $\mathbf{D}_1, \mathbf{D}_2, \ldots \mathbf{D}_k$. A conditional probability density $f(x|\mathbf{D}_i)$ may be evaluated empirically using a nearest-neighbor density estimator [*Silverman*, 1986] with the successors $x_1 \ldots x_k$. For simulation, the x_i can be drawn randomly from one of the k successors to the $\mathbf{D}_1, \mathbf{D}_2, \ldots \mathbf{D}_k$ using this estimated conditional density. Here, this operation will be done by resampling the original data with replacement. Hence the procedure developed is termed a nearest-neighbor time series bootstrap. In summary, one finds k patterns in the data that are "similar" to the current pattern, and then operates on their respective successors to define a local regression, conditional density or resampling.

The nearest-neighbor probability density estimator and its use with Markov processes is reviewed in the next section. The resampling algorithm is described next. Applications to synthetic and streamflow data are then presented.

Background

It is natural to pursue nonparametric estimation of probability densities and regression functions through weighted local averages of the target function. This is the foundation for nearest-neighbor methods. The recognition of the nonlinearity of the underlying dynamics of geophysical processes, gains in computational ability, and the availability of large data sets have spurred the growth of the nonparametric literature. The reader is referred to *Silverman* [1986], *Eubank* [1988], *Hardle* [1989, 1990], and *Scott* [1992] for accessible monographs. *Györfi et al.* [1989] provide a theoretical account that is relevant for time series analysis. *Lall* [1995] surveys hydrologic applications. For time series analysis, a moving block bootstrap (MBB) was presented by *Kunsch* [1989]. Here a block of m observations is resampled with replacement as opposed to a single observation in the bootstrap. Serial dependence is preserved within, but not across a block. The block length m determines the order of the serial dependence that can be preserved. Objective procedures for the selection of the block length m are evolving. Strategies for conditioning the MBB on other processes (e.g., runoff on rainfall) are not obvious. Our investigations indicated that the MBB may not be able to reproduce the sample statistics as well as nearest-neighbor bootstrap presented here.

The k nearest-neighbor (k-nn) density estimator is defined as [*Silverman*, 1986]:

$$f_{NN}(x) = \frac{k/n}{V_k(x)} = \frac{k/n}{c_d r_k^d(x)} \qquad (2.1)$$

where k is the number of nearest neighbors considered, d is the dimension of the space, c_d is the volume of a unit sphere in d dimensions ($c_1=2$, $c_2=\pi$, $c_3=4\pi/3$..., $c_d=\frac{\pi^{d/2}}{\Gamma(d/2+1)}$), $r_k(x)$ is the Euclidean distance to the kth nearest data point, and $V_k(x)$ is the volume of a d-dimensional sphere of radius $r_k(x)$.

This estimator is readily understood by observing that for a sample of size n, we expect approximately $\{n\ f(x)\ V_k(x)\}$ observations to lie in the volume $V_k(x)$. Equating this to the number observed, i.e., k, completes the definition.

A generalized nearest-neighbor density estimator [*Silverman*, 1986] defined in (2.2) can improve the tail behavior of the nearest-neighbor density estimator by using a monotonically and possibly rapidly decreasing smooth kernel function.

$$f_{GNN}(x) = \frac{1}{r_k^d(x)n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{r_k(x)}\right) \qquad (2.2)$$

The "smoothing" parameter is the number of neighbors used, k, and the tail behavior is determined by the kernel K(t). The kernel has the role of a weight function (data vectors x_i closer to the point of estimate x are weighted more) and can be chosen to be any valid probability density function. Asymptotically, under optimal Mean Square Error (MSE) arguments, k should be chosen proportional to $n^{4/(d+4)}$ for a probability density that is twice differentiable. However, given a single sample from an unknown density, such a rule is of little practical utility. The sensitivity to the choice of k is somewhat lower as a kernel that is monotonically decreasing with $r_k(\mathbf{x})$, is used. A new kernel function that weights the j^{th} neighbor of $\mathbf{x}_i$ using a kernel that depends on the distance between $\mathbf{x}_i$ and its j^{th} neighbor is developed in the resampling methodology section.

Yakowitz (references cited earlier) developed a theoretical basis for using nearest-neighbor and kernel methods for time series forecasting and applied them in a hydrologic context. In these papers, Yakowitz considers a finite order, continuous parameter Markov Chain as an appropriate model for hydrologic time series. He observes that discretization of the state space can quickly lead to either an unmanageable number of parameters (the curse of dimensionality) or poor approximation of the transition functions, while the ARMA approximations to such a process call for restrictive distributional and structural assumptions. Strategies for the simulation of daily flow sequences, one-step-ahead prediction, and the conditional probability of flooding (flow crossing a threshold) are exemplified with river flows and shown to be superior to ARMA models. Seasonality is accommodated by including the calendar date as one of the predictors. Yakowitz claims that this continuous parameter Markov chain approach is capable of reproducing any possible Hurst coefficient. Classical ARMA models are optimal only under squared error loss, and only for linear operations on the observables. The loss/risk functions associated with hydrologic decisions (e.g., declare a flood warning or not) are usually asymmetric. The nonparametric framework allows attention to be focused directly on calculating these loss functions and evaluating the consequences.

The example of Figure 2-1 is now extended to show how the nearest-neighbor method is used in the Markov framework. One-step Markov transition functions are considered. The relationship between x_{t+1} and x_t is shown in Figure 2-2. The correlation between x_t and x_{t+1} is 0, even though there is clear-cut dependence between the two variables.

Consider an approximation of the model in Figure 2-1 by a multistate, first-order Markov chain, where transitions from say state 1 for x_t (0 to 0.25 in Figure 2-2), to states 1, 2, 3, or 4 for x_{t+1} are of interest. The state i to state j transition probability p_{ij} is evaluated by counting the relative fraction of transitions from state i to state j. The estimated transition probabilities depend on the number of states chosen as well as their actual demarcation (e.g., one may need a nonuniform grid that recognizes variations in data density). For the nonlinear model used in our example, a fine discretization would be needed. Given a finite data set, estimates of the multistate transition probabilities may be unreliable. Clearly, this situation is exacerbated if one considers higher dimensions for the predictor space. Further, a reviewer observed that a discretization of a continuous space Markov process is not necessarily Markov.

Now consider the nearest-neighbor approach. Consider two conditioning points x^*_A and x^*_B. The k nearest neighbors of these points are in the dashed windows A and B, respectively. The neighborhoods are seen to adapt to variations in the sampling density of x_t. Since such neighborhoods represent moving windows (as opposed to fixed windows for the multistate Markov chain) at each point of estimate, we can expect reduced bias in the recovery of the target transition functions. The one-step transition probabilities at x^*_t can be obtained through an application of the nearest-neighbor density estimator to the x_{t+1} values that fall in windows like A and B. A conditional *bootstrap* of the data can be obtained by resampling from this set of x_{t+1} values. Since each transition probability estimate is based on k points, the problem faced in a multistate Markov chain model of sometimes not having an adequate number of events or state transitions to develop an estimate is circumvented.

The Nearest-Neighbor Resampling Algorithm

In this section, a new algorithm for generating synthetic time series samples by bootstrapping (i.e., resampling the original time series with replacement) is presented. Denote the time series by x_t , t=1...n, and assume a known dependence structure, i.e., which and how many lags the future flow will depend on. This conditioning set is termed a "feature vector," and the simulated or forecasted value the "successor". The strategy is to find the historical nearest neighbors of the current feature vector, and resample from their successors. Rather than resampling uniformly from the k successors, a discrete resampling kernel is introduced to weight the resamples to reflect the similarity of the neighbor to the conditioning point. This kernel is monotonically decreasing with distance, adapts to the local sampling density, to the dimension of the feature vector, and to boundaries of the sample space. An attractive probabilistic interpretation of this kernel consistent with the nearest-neighbor density estimator is also offered. The resampling strategy is presented through a flow chart.

(1) Define the composition of the "feature vector" $\mathbf{D}_t$ of dimension d.

e.g., (a) $\mathbf{D}_t : (x_{t-1}, x_{t-2})$; d=2

(b) $\mathbf{D}_t$: $(x_{t-\tau 1},\ x_{t-2\tau 1},\ \ldots.\ x_{t-M_1\tau_1};\ x_{t-\tau_2},\ x_{t-2\tau_2},\ \ldots..\ x_{t-M_2\tau_2})$; d=M_1+M_2

(c) $\mathbf{D}_t$: $(x'_{t-\tau_1},\ \ldots.\ x'_{t-M_1\tau_1};\ x''_t,\ x''_{t-\tau_2},\ \ldots.\ x''_{t-M_2\tau_2})$; d=$M_1$+$M_2$+1

where τ_1 (e.g., 1 month) and τ_2 (e.g., 12 months) are lag intervals, and $M_1, M_2 \geq 0$ are the number of such lags considered in the model.

Case 1 represents dependence on two prior values. Case 2 permits direct dependence on multiple time scales, allowing one to incorporate monthly and interannual dependence. For case 3, x', and x" may refer to rainfall and runoff, or to two different streamflow stations.

(2) Denote the current feature vector as $\mathbf{D}_i$ and determine its k nearest neighbors among the $\mathbf{D}_t$, using the weighted Euclidean distance

$$r_{it} = \left(\sum_{j=1}^{d} w_j (v_{ij} - v_{tj})^2 \right)^{1/2} \tag{2.3}$$

where v_{tj} is the j^{th} component of $\mathbf{D}_t$, w_j are scaling weights (e.g., 1, or $1/s_j$), where s_j is some measure of scale such as the standard deviation or range of v_j.

The weights w_j may be specified a priori, or chosen to provide the best forecast for a particular successor in a least squares sense [*Yakowitz and Karlsson,* 1987].

Denote the ordered set of nearest-neighbor indices by $J_{i,k}$. An element j(i) of this set records the time t associated with the j^{th} closest $\mathbf{D}_t$ to $\mathbf{D}_i$. Denote $x_{j(i)}$ as the successor to $\mathbf{D}_{j(i)}$. If the data are highly quantized, it is possible that a number of observations may be the same distance from the conditioning point. The resampling kernel defined in step 3 is based on the order of elements in $J_{i,k}$. Where a number of observations are the same distance away, the original ordering of the data can impact the ordering in $J_{i,k}$. To avoid such artifacts, the time indices t are copied into a temporary array that is randomly permuted prior to distance calculations and creation of the list $J_{i,k}$.

(3) Define a discrete kernel K(j(i)) for resampling one of the $x_{j(i)}$ as follows:

$$K(j(i)) = \frac{1/j}{\sum_{j=1}^{k} 1/j} \tag{2.4}$$

where K(j(i)) is the probability with which $x_{j(i)}$ is resampled.

This resampling kernel is the same for any i, and can be computed and stored prior to the start of the simulation.

(4) Using the discrete probability mass function (p.m.f.) K(j(i)), resample an $x_{j(i)}$, update the current feature vector and proceed to step 2 if additional simulated values are needed.

A similar strategy for time series forecasting is possible. An m-step-ahead forecast is obtained by using the corresponding generalized nearest-neighbor regression estimator:

$$g_{GNN}(x_{i,m})=\sum_{j=1}^{k} K(j(i))x_{j(i),m} \tag{2.5}$$

where $x_{i,m}$ and $x_{j(i),m}$ denote the m^{th} successor to i, and j(i), respectively.

Parameters of the Nearest-Neighbor Resampling Method

Choosing the Weight Function K(v)

The goals for designing a resampling kernel are to (1) reduce the sensitivity of the procedure to the actual choice of k, (2) keep the estimator local, (3) have k sufficiently large to avoid simulating nearly identical traces, (4) develop a weight function that adapts automatically to boundaries of the domain, and to the dimension d of the feature vector $\mathbf{D}_t$. These criteria suggest a resampling kernel that decreases monotonically as r_{ij} increases.

Consider a d-dimensional Ball of Volume V(r) centered at $\mathbf{D}_i$. The observation $\mathbf{D}_{j(i)}$ falls in this ball when the ball is of volume exactly $V(r_{i,j(i)})$. Assuming that the observations are independent (which they may not be), the likelihood with which the

$j(i)^{th}$ observation should be resampled as representative of $\mathbf{D}_i$ is proportional to $1/V(r_{ij(i)})$.

Now, consider that in a small locale of $\mathbf{D}_i$, the local density can be approximated as a Poisson process, with constant rate λ. Under this assumption, the expected value of $1/V(r_{ij(i)})$ is:

$$E(1/V(r_{ij(i)})) = \lambda/j \tag{2.6}$$

The kernel K(j(i)) is obtained by normalizing these weights over the k nearest neighborhood.

$$K(j(i)) = \frac{c\lambda/j}{\sum_{j=1}^{k} c\lambda/j} = \frac{1/j}{\sum_{j=1}^{k} 1/j} \tag{2.7}$$

where c is a constant of proportionality.

These weights do not explicitly depend on the dimension d of the feature vector $\mathbf{D}_t$. The dependence of the resampling scheme on d is implicit through the behavior of the distance calculations used to find nearest neighbors as d varies. Initially, we avoided making the assumption of a local Poisson distribution, and defined K(j(i)) through a normalization of $1/V(r_{ij(i)})$. This approach gave satisfactory results as well but was computationally more demanding. The results obtained using Equation (2.7) were comparable for a given k.

The behavior of this kernel in the boundary region, the interior, and the tails is seen in Figures 2-3 and 2-4. From Figure 2-3, observe that the nearest-neighbor method allows considerable variation in the "bandwidth" (in terms of a range of values of x) as a function of position, and underlying density. The bandwidth is automatically larger as the density is sparser and flatter. In regions of high data density (left tail or interior), the kernel is nearly symmetric (the slight asymmetry follows the asymmetry in the underlying distribution). Along the sparse right tail, the kernels are quite asymmetric as expected. Some attributes of these kernels relative to a uniform kernel (with the same k), used by the ordinary nearest-neighbor method are shown in Figure 2-4.

For bounded data (e.g., streamflow that is constrained to be greater than 0), simulation of values across the boundary is often a concern. This problem is avoided in the method presented since the resampling weights are defined only for the sample points. A second problem with bounded data is that bias in estimating the target function using local averages increases near the boundaries. This bias can be recognized by observing that the centroid of the data in the window does not lie at the point of estimate. From Figure 2-4 note that while the kernel is biased towards the edges of the data, i.e., the centroid of the kernel does not match the conditioning point, the bias is much smaller than for the uniform kernel. If one insists on kernels that are strictly positive with monotonically decreasing weights with distance, it may not be possible to devise kernels that are substantially better in terms of bias in the boundary region.

The primary advantages of the kernel introduced here are:

(1) It adapts automatically to the dimension of the data.

(2) The resampling weights need to be computed just once for a given k (there is no need to recompute the weights or their normalization to 1).

(3) Bad effects of data quantization or clustering (this could lead to near zero values of $r_{i,j(i)}$ at some points) on the resampling strategy, which arise if one were to resample using a kernel that depends directly on distance (e.g., proportional to $1/V(r_{i,j(i)})$, are avoided. These factors translate into considerable savings in computational time that can

be important for large data sets and high dimensional settings, and to improved stability of the resampling algorithm.

Choosing the Number of Neighbors k, and Model Order d

The order of ARMA models is often picked [*Loucks et al.*, 1981] using the Akaike Information Criteria (AIC). Such criteria estimate the variance of the residual to time series forecasts from a model, appropriately penalized for the effective degrees of freedom in the model. A similar perspective based on cross validation is advocated here. Cross validation involves "fitting" the model by leaving out 1 value at a time from the data, and forecasting it using the remainder. The model that yields the least predictive sum of squares of errors across all such forecasts is picked. One can approximate the average effect of such an exercise on the sum of squares of errors without going through the process.

Here, the forecast is formed (Equation 2.5) as a weighted average of the successors. When using the full sample, define the weight used with the successor to the current point as w_{jj}. This weight recognizes the influence of that point on the estimate at the same location. Hence, the influence of the rest of the points on the fit at that point is $(1-w_{jj})$. This suggests that if estimated full sample forecast error e_j is divided by $(1 - w_{jj})$ a measure of what the error may be if the data point $(\mathbf{D}_j, x_j)$ was not used in developing the estimate is provided. Note that the degrees of freedom (e.g., 0 for a k=1; 1-1/3 for a k=3 using a uniform kernel) of estimate are implicit in this idea. *Craven and Wahba* [1979] present a Generalized Cross Validation (GCV) score function that considers the average influence of excluded observations for estimation at each sample point and approximates the predictive squared error of estimate. The GCV score is given as:

$$\mathrm{GCV} = \frac{\sum_{i=1}^{n} e_i^2/n}{\left(\sum_{j=1}^{n} (1-w_{jj})/n\right)^2} \tag{2.8}$$

The GCV score function can be used to choose both k and d. For the kernel suggested in this chapter, w_{jj} is a constant for a given k, and the GCV can be written as:

$$\mathrm{GCV} = \frac{\sum_{i=1}^{n} e_i^2/n}{\left(1 - 1/\sum_{j=1}^{k} 1/j\right)^2} \tag{2.9}$$

A prescriptive choice of $k=\sqrt{n}$ from experience is also suggested. This is a good choice for $1 \le d \le 6$, and $n \ge 100$. Sensitivity to the choice of k in this neighborhood is small and where computational resources are limited this choice can be recommended. Typically, with a sample size n of 50 to 200, this corresponds to a choice of k ranging from 7 to 14. When using the GCV criteria with the same sample size, it is our experience that varying k within five to 10 units of the optimal selected value does not appreciably change the GCV score.

Criteria such as the GCV and the AIC are known to overfit or over parameterize time series relationships. With the nearest-neighbor resampler, a model with order higher than necessary will have increased variability for a given k. The extra or superfluous coordinates serve to degrade rather than enhance identification of the patterns that describe the system. Likewise, a smaller than optimal choice of d would lead to traces that lack the appropriate memory. Comparison of the attributes of the series generated by models with different values of k and d is consequently desirable. These comparisons can be based on how well attributes of direct interest to the investigator such as run lengths or the frequencies of threshold crossings are reproduced. One can try various combinations of k and d and visually compare resampling attributes with historical sample attributes (sample moments and marginal or joint probability densities).

Applications

Two synthetic examples, one from a linear and one from a nonlinear model, are presented first. These are followed by an application to monthly flows from the Weber River near Oakley, Utah. In all cases a lag one model with k chosen as $\sqrt{n}$ was used.

Comparative performance of the simulations is judged using sample moments and sample probability density functions (p.d.f.'s) estimated using Adaptive Shifted Histograms (ASH) [*Scott*, 1992]. In all applications using the univariate ASH, a bin width of $(x_{max}-x_{min})/9.1$ where x_{max} and x_{min} are the respective maximum and minimum values of the data, and five shifted histograms were used. For bivariate densities, a bin width of $(x_{max}-x_{min})/3.6$ and five shifted histograms in each coordinate direction were used. These are the default settings for the computer code distributed by David Scott. Conditional expectations, $E(x_t \mid x_{t-1})$, are estimated using LOWESS [*Cleveland*, 1979]. LOWESS is a popular robust locally weighted linear regression technique, that allows a flexible curve to be fit between two variables. We used default parameter choices (three iterations for computing the robust estimates based on two third of the data) with the "lowess" function available on Splus [*Chambers and Hastie*, 1992].

Example E1

The first data set considered is a sample of size 500 from a linear autoregressive model of order 1, AR1, defined as:

$$x_t = 0.6\, x_{t-1} + 0.8\, e_t \tag{2.10}$$

where e_t is a mean zero, variance 1, Gaussian random variable.

One hundred realizations of length 500 each were then generated from an AR(1) model fitted to the sample, and from the nearest-neighbor (k-nn) bootstrap. Selected statistics from these simulations are compared in Table 2-1. The sample statistics considered are reproduced by the k-nn method, while only the sample statistics used in fitting the parametric AR1 model are reproduced. The AR1 simulations instead reproduce population or model statistics (e.g., skew =0) for parameters that are not explicitly estimated from the sample. With repeated applications to a number of samples from the same distribution, the k-nn procedure reproduces the population statistics as well. On the other hand, a parametric model only reproduces fitted statistics. The ASH estimated median p.d.f. of x_t, from the k-nn resamples matches the sample p.d.f., and the scatter of the estimated p.d.f.'s across resamples is comparable to the scatter from the AR(1) samples. These results are not reproduced here to save space.

Example E2

A sample of size 200 was generated from a Self-Exciting Threshold Autoregressive (SETAR) model described by *Tong* [1990]. The general structure of such models is similar to that of a linear autoregressive model, with the difference that the parameters of the model change upon crossing one or more thresholds. Such a model may be appropriate for daily streamflow, since crossing a flow threshold (defined on a single past flow or collectively on a set of past flows) with flow increasing may signal runoff response to rainfall or snow-melt, and crossing the threshold with flow decreasing may signal return to base flow or recession behavior. Here a lag 1 SETAR model given below was used:

$$\begin{aligned} x_t &= 0.4 + 0.8\, x_{t-1} + e_t \qquad \text{if } x_t \leq 0.0 \\ &= -1.5 - 0.5\, x_{t-1} + e_t \qquad \text{else} \end{aligned} \tag{2.11}$$

where e_t is a Gaussian random variable with mean 0, and variance 1.

The time series generated from the SETAR model and a time series simulated by the nearest-neighbor method are shown in Figure 2-5. The bivariate probability densities $f(x_t, x_{t-1})$ for the original SETAR sample and for 100 nearest-neighbor samples each of length 200 were computed using ASH. The estimated $f(x_t, x_{t-1})$ from the original sample along with a LOWESS fit of $E(x_t \mid x_{t-1})$ and an average of the $f(x_t, x_{t-1})$ estimates taken across the nearest-neighbor realizations are illustrated in Figure 2-6. We see that the bivariate density of the data is reproduced quite well by the simulations.

Example E3

The 1905-1988 monthly flow record from USGS station 10128500, Weber River near Oakley, Utah, located at 40°44'10"N, and 111°14'45"W, at an elevation of 6600 ft above Mean Sea Level was extracted from the USGS, HCDN CD-ROM [*Slack et al.*, 1992]. This data set is presumed to be free of the effects of regulation, diversion, and similar factors. Weber River at this location is a snow-melt fed, perennial, mountain stream, with a drainage area of 162 square miles. The mean annual flow is 223 cfs. June is the month with the highest flow, subsequent to snow-melt. January is typically the month with the lowest flow. The 1905-1988 monthly time series, and flows for two specific years are presented in Figure 2-7.

A monthly AR1 model was fitted to logarithmically transformed monthly flows, with monthly varying parameters estimated as described in *Loucks et al.* [1981] to preserve moments in real space. This entails sequentially simulating monthly flow moving through the calendar year, using (for example) only the 83 monthly values for January and February over the 83-year record to simulate February flows given January flows. One hundred simulations of 83 years were generated in each case. The k-nn bootstrap is applied in a similar manner using (for example) January flows to find neighbors for a given January flow, and the corresponding February successor. The flow data are not logarithmically transformed for the k-nn bootstrap.

Selected results are presented. A comparison of means, standard deviations, skew, and lag 1 correlations is presented for the 12 months in Figure 2-8. Both the k-nn and the AR1 model seem to be comparable in reproducing the mean monthly flow and its variation across simulations. The standard deviations of monthly flows are somewhat more variable in k-nn simulations than those from the AR1 model. Some months (low flows) have a slight downward bias in the simulated standard deviation, while others (March and July, the months following a minimum and maximum in monthly flow, respectively) show a larger spread in standard deviation than in the AR1 case. While both models seem to do well in reproducing the historical lag 1 correlation, the k-nn statistics appear to be more variable across realizations. A difference between the two simulators is apparent in Figure 2-8(c) where the AR1 model fails to reproduce the monthly and the annual skews as well as the k-nn model. Recall that the ad hoc prescriptive choice of $k=\sqrt{n}$ was used here, with no attempt at fine-tuning the k-nn simulator.

The marginal probability density functions were estimated by ASH for flow in each month. Selected results for simulations from the k-nn and for the AR1 model applied to logarithmically transformed flows for 3 months in Figure 2-9. We see from Figure 2-9 that the k-nn samples are indeed a bootstrap, i.e., the simulated marginal probability densities behave much like the empirical sample probability density. The usual shortcoming of the bootstrap in reproducing only historical sample values is also apparent. We see that while the Lognormal density used by the AR1 model is plausible in a number of months (e.g., October), the Lognormal model seems to be inappropriate in other months, e.g., March, where the skew is too extreme for the AR1, and June, which exhibits a distinct bimodality that may be related to the timing or amount of snow-melt. The latter is interesting, since the 100 simulations from the AR1 model fail to bracket the two prominent modes of the ASH density estimate, lending support to the idea of bimodality under a pseudo-hypothesis test obtained from this Monte Carlo experiment.

The bivariate probability density functions for flows in each consecutive pair of months (e.g., May and June) were also computed by ASH. Results for selected months are presented in Figures 2-10 to 2-12. Other month pairs were found to exhibit features similar to those in Figures 2-10 to 2-12. In each case, we present a scatterplot of the flows (in cfs.) in the 2 months, with a LOWESS [*Cleveland*, 1979] smooth of the conditional expectation $E(x_t \mid x_{t-1})$. An examination of the October-November density in Figure 2-10 reveals that the AR1 model may be quite appropriate for this pair of months. The ASH estimated density from the sample and the averages of the ASH estimated densities from the AR1 and the k-nn samples are all very similar. The LOWESS estimate of the conditional expectation of the November flow given the October flow is very nearly a straight line.

Figures 2-11 and 2-12 refer to the months of April/May and May/June, where runoff from snow-melt becomes important. The timing of the start and of the peak rate of snow-melt vary over this period. Consequently, one can expect some heterogeneity in the sampling distributions of flows in these months. From Figure 2-11 (a), we see that the LOWESS estimate of $E(x_t \mid x_{t-1})$ exhibits some degree of nonlinearity for April/May. The slope of $E(x_t \mid x_{t-1})$ for $x_{t-1}<150$ cfs is quite different from the slope for $x_{t-1}>150$ cfs. This is reminiscent of the SETAR model examined earlier. We could belabor this point through formal tests of significance for difference in slope. Our purpose here is to show that the k-nn approach can adapt to such sample features, while the AR1 model may not. The average bivariate densities of the simulations based on ASH once again reinforce this difference between the k-nn and the AR1 models and their attributes. Note also that the AR1 simulations do not reproduce the sample skews for this period either.

The May/June analysis shown in Figure 2-12 is marked by considerably increased variability in stream flow as the snow-melt runoff develops. Once again, LOWESS shows some degree of nonlinearity in $E(x_t \mid x_{t-1})$, with the slope of the relationship smaller for high flows than for low flows. A comparison of the ASH bivariate density contours in Figures 2-12(a) and (c) reveals that the AR1 density is oriented quite differently from the ASH estimate for the raw sample and is unable to reproduce the degree of heterogeneity in the sample density. Recall that the marginal density of June flows was bimodal, with an antimode around 1000 cfs. The antimode suggests that the data are clustered into two classes of events, those with flow below 1000 cfs (mode at 700 cfs) and those with flow above 1000 cfs (mode at 1300 cfs). The LOWESS fit suggests that the June flows have an expectation close to 1000 cfs for May flows greater than about 700 cfs. It appears that the conditional expectation averages across the two modes for June flows, and that the conditional density (Figure 2-12(b)) of June flows given May flow may be bimodal as seen in the marginal density plot for June flows.

The significance of the findings reported above is that the nearest-neighbor bootstrap provides a rather flexible and adaptive method for reproducing the historical frequency distribution of streamflow. The possibly tenuous issue of choosing between a variety of candidate parametric models, month by month, is avoided. Matching the historical frequency distribution of flows properly is important for properly estimating storage requirements for a reservoir and analyzing reservoir release options. For snow-melt driven streams in arid regions, the timing and amount of melt is important in determining reservoir operation. The bimodality in the probability density of monthly streamflow during the melt months may be connected to structured low frequency (interannual and interdecadal) climatic fluctuations in this area [*Lall and Mann*, 1995]. This would be a significant factor for reservoir operation, since the timing and amount of snow-melt may correspond to a circulation pattern that corresponds to specific flow patterns in subsequent months as well. The nearest-neighbor bootstrap would be an appropriate technique for simulating sequences conditioned on such factors. Work in this direction is in progress.

Summary and Discussion

A nearest-neighbor method for a conditional bootstrap of time series was presented and exemplified. A corresponding forecasting strategy was indicated. Our contributions here lie primarily in the development of a new kernel, suggestion of a parameter selection strategy, application to a conditional bootstrap, and demonstration of the methodology. It was shown that sample attributes are reproduced quite well by this nonparametric method for both synthetic and real data sets. Given the flexibility of these techniques, we consider them to have tremendous practical potential.

The parametric versus nonparametric statistical method debate often veers towards sample size requirements and statistical efficiency arguments. In the context of a resampling strategy, as espoused here, these arguments take a somewhat different complexion. For some processes, such as daily streamflow, identification or even definition of an appropriate parametric model is problematic. In these cases, data are relatively plentiful. For such cases, methods such as those presented here are enticing. For monthly and annual flows, there is progressively less structure and sample sizes are smaller. In these situations, parametric methods may indeed be statistically more efficient provided the correct model is identifiable and parsimonious. In our view, particularly for the snow-fed rivers of the western United States, this may not always be the case. Indeed, for the application presented here at the monthly time scale, it is hard to justify choosing the parametric approach over the nearest-neighbor method. The consideration of parameter uncertainty is justifiably considered a good idea in parametric time series resampling of streamflow [*Grygier and Stedinger*, 1990]. Likewise, it may be useful to think about model uncertainty when developing parametric models. The latter consideration is implicit in the nonparametric approach, since a rather broad class of models is approximated. The impact of varying the "parameters" k, and the model order on specific attributes of the resamples bears further investigation. Our preliminary analyses suggest that the sensitivity of the scheme is limited over a range of k values near the "optimal" with the kernel used here. Formal investigations of this issue are being pursued.

One can devise a strategy that allows nearest-neighbor resampling with perturbation of the historical data in the spirit of traditional autoregressive models, i.e., conditional expectation with an added random innovation. First, one evaluates the conditional expectation using the generalized nearest-neighbor regression estimator for each vector $\mathbf{D}_i$ in the historical record. A residual e_i can be computed as the difference between the successor x_i of $\mathbf{D}_i$ and the nearest-neighbor regression forecast. The simulation proceeds by estimating the nearest-neighbor regression forecast relative to a conditioning vector $\mathbf{D}_i$, and then adding to this one of the e_j corresponding to a data point j, that lie in the k-nearest neighborhood $J_{i,k}$. The innovation e_j is chosen using the resampling kernel K(j(i)). This scheme will perturb the historical data points in the series, with innovations that are representative of the neighborhood, and will thus "fill in" between the historical data values, as well as extrapolating beyond the sample. The computational burden is increased and there is a possibility that the bounds on the variables will be violated during simulation. However, there may be situations where the investigator may wish to adopt this strategy. Further exploration of this strategy is planned.

Issues such as disaggregation of streamflows bear further investigation. One strategy is trivial; resample the flow vector that aggregates to the aggregate flow simulated. A question that arises is whether there is even any need to work with models that disaggregate (especially in time) using these methods. One may wish to work directly with, for example, the daily flows, conditioned on a sequence of past daily flows and weekly or monthly flows.

The real utility of the method presented here may lie in exploiting a dependence structure (e.g., in daily flows) that is difficult to treat by traditional methods, as well as complex relationships between variables, and in estimating confidence limits or risk in problems that have a time series structure. The traditional time series analysis framework directs the researcher's attention towards an efficient estimation of model parameters under some metric (e.g., least squares or maximum likelihood). The performance metric of interest to the hydrologist may not be the one optimal for the estimation of a certain set of parameters and selected model form. There is reason to directly explore other aspects of the problem that may be of direct interest for reservoir operation and flood control, using flexible, adaptive, data exploratory methods. Such investigations using the k-nn bootstrap are in progress.

References

Chambers, J. M., and T. J. Hastie, *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove, Calif., 1992.

Cleveland, W. S., Robust locally weighted regression and smoothing scatterplots, *J. Am. Stat. Assoc.*, 74, 829-836, 1979.

Craven, P., and G. Wahba, Smoothing noisy data with spline functions, *Numer. Math.*, 31, 377-403, 1979.

Efron, B., Bootstrap methods: Another look at the jackknife, *Ann. Stat.*, 7, 1-26, 1979.

Efron, B., and R. Tibishirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

Eubank, R. L., *Spline Smoothing and Nonparametric Regression*, Marcel-Dekker, New York, 1988.

Grygier, J. C., and J. R. Stedinger, *Spigot, A Synthetic Streamflow Generation Package, Technical Description, Version 2.5*, School of Civil and Environmental Engineering, Cornell University, Ithaca, N.Y., 1990.

Györfi, L., W. Hardle, P. Sarda, and P. Vieu, *Nonparametric Curve Estimation from Time Series*, vol. 60, Lecture Notes in Statistics, Springer Verlag, New York, 1989.

Hardle, W., *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge University Press, Cambridge, Mass., 1989.

Hardle, W., *Smoothing Techniques With Implementation in S*, Springer Verlag, New York, 1990.

Hardle, W., and A. W. Bowman, Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands, *J. Am. Stat. Assoc.*, 83, 102-110, 1988.

Karlsson, M., and S. Yakowitz, Nearest-neighbor methods for nonparametric rainfall-runoff forecasting, *Water Resour. Res.*, 23, 1300-1308, 1987a.

Karlsson, M., and S. Yakowitz, Rainfall-runoff forecasting methods, old and new, *Stochastic Hydrol. Hydraul.*, 1, 303-318, 1987b.

Kendall, D. R., and J. A. Dracup, A comparison of index-sequential and AR(1) generated hydrologic sequences, *J. Hydrol.*, 122, 335-352, 1991.

Kunsch, H.R., The jackknife and the bootstrap for general stationary observations, *Ann. Stat.*, 17, 1217-1241, 1989.

Lall, U., Nonparametric function estimation: Recent hydrologic applications, U.S. National Report to International Union of Geodesy and Geophysics, 1991-1994, *Rev. Geophys*, 33, 1093, 1995.

Lall, U., and M. Mann, The Great Salt Lake: A barometer of interannual climate variability, *Water Resour. Res.*, 31(10), 2503-2515, 1995.

Loucks, D. P., J. R. Stedinger, and D. A. Haith, *Water Resource Systems Planning and Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1981.

Schuster, E., and S. Yakowitz, Contributions to the theory of nonparametric regression with application to system identification, *Ann. Stat.*, 7, 139-149, 1979.

Scott, D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, 1992.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, England, 1986.

Slack, J. R., J. M. Landwehr, and A. Lumb, A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874-1988, *Rep. 92-129*, U.S. Geol. Surv., Oakley, Utah, 1992.

Smith, J., G. N. Day, and M. D. Kane, Nonparametric framework for long range streamflow forecasting, *J. Water Resour. Plng. Mgmt.*, 118, 82-92, 1992.

Smith, J. A., Long-range streamflow forecasting using nonparametric regression, *Water Resour. Bull.*, 27, 39-46, 1991.

Tarboton, D. G., A. Sharma, and U. Lall, The use of non-parametric probability distributions in streamflow modeling, in *Proceedings of the Sixth South African National Hydrological Symposium*, edited by S. A. Lorentz, S. W. Kienzle, and M. C. Dent, pp. 315-327, University of Natal, Pietermaritzburg, South Africa, 1993.

Tasker, G. D., Comparison of methods for estimating low flow characteristics of streams, *Water Resour. Bull.*, 23, 1077-1083, 1987.

Tong, H., *Nonlinear Time Series Analysis: A Dynamical Systems Perspective*, Academic Press, London, England, 1990.

Woo, M. K., Confidence intervals of optimal risk-based hydraulic design parameters, *Can. Water Resour. J.*, 14, 10-16, 1989.

Yakowitz, S., A stochastic model for daily river flows in an arid region, *Water Resour. Res.*, 9, 1271-1285, 1973.

Yakowitz, S., Nonparametric estimation of Markov transition functions, *Ann. Stat.*, 7, 671-679, 1979.

Yakowitz, S., Nonparametric density estimation, prediction, and regression for Markov sequences, *J. Am. Stat. Assoc.*, 80, 215-221, 1985.

Yakowitz, S., Nearest-neighbor regression estimation for null-recurrent Markov time series, *Stochastic Processes Their Appl.*, 48, 311-318, 1993.

Yakowitz, S., and M. Karlsson, Nearest-neighbor methods with application to rainfall/runoff prediction, in *Stochastic Hydrology*, edited by J. B. Macneil and G. J. Humphries, pp. 149-160, D. Reidel, Hingham, Mass., 1987.

Zucchini, W., and P. T. Adamson, Bootstrap confidence intervals for design storms from exceedance series, *Hydrol. Sci. J.*, 34, 41-48, 1989.

Table 2-1. Statistical Comparison of k-nn and AR1 Model Simulations Applied to an AR1 Sample

	AR1 Sample	Simulations					
		5 % Quantile		Median		95 % Quantile	
		k-nn	AR1	k-nn	AR1	k-nn	AR1
Mean	0.04	-0.14	-0.12	0.02	0.04	0.24	0.20
Standard Deviation	1.11	1.02	1.03	1.10	1.11	1.18	1.20
Skew	-0.17	-0.32	-0.25	-0.18	0.00	-0.03	0.21
Lag 1 Correlation	0.63	0.56	0.57	0.62	0.63	0.68	0.69

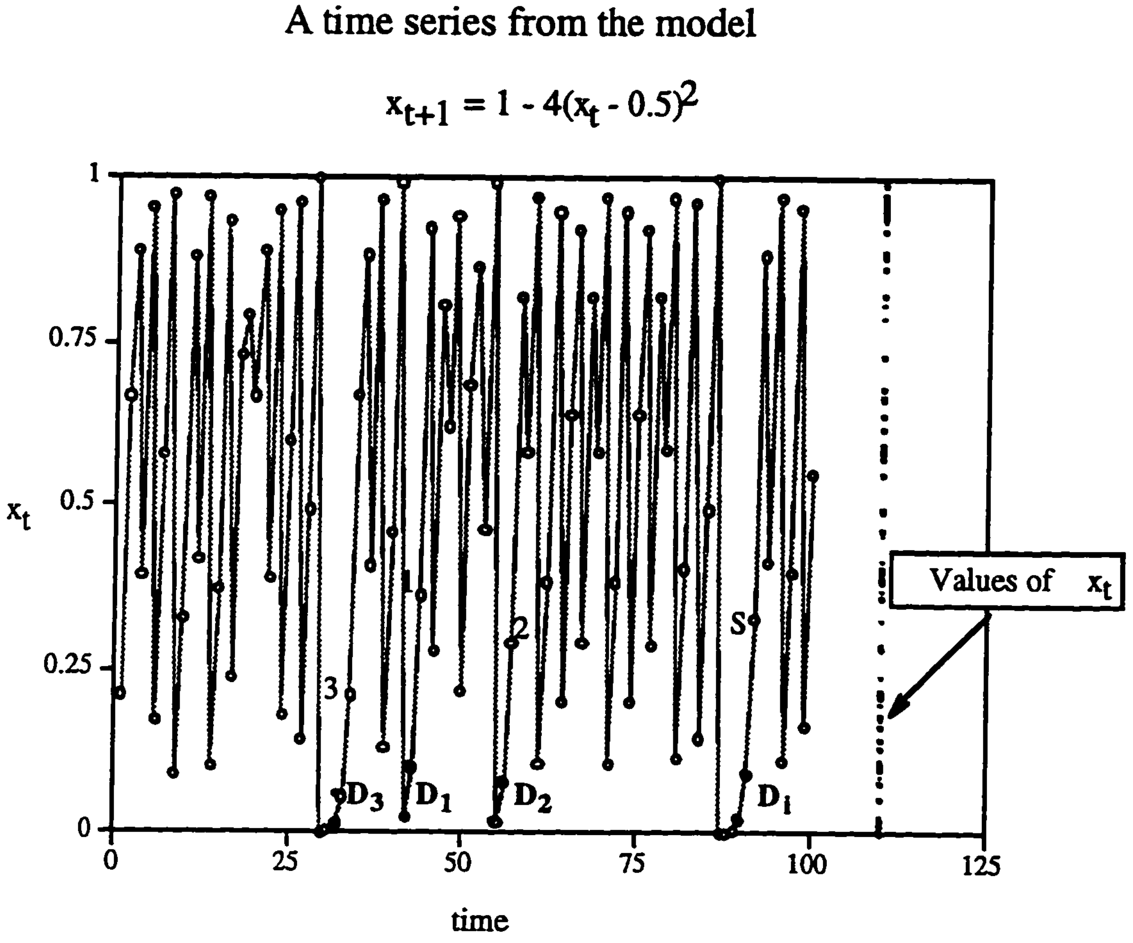


Figure 2-1. A time series from the model $x_{t+1}=(1-4(x_t-0.5)^2)$. This is a deterministic, nonlinear model, with a time series that looks random. A forecast of the successor to the bi-tuple D_i, marked as S in the figure, is of interest. The "patterns" or bituples of interest are the filled circles, near the 3 nearest neighbors, D_1, D_2, and D_3 to the pattern D_i. The successors to these bituples are marked as 1, 2, and 3, respectively. Note how the successor (1) to the closest nearest neighbor (D_1) is closest to the successor (s) of D_i. A sense of the marginal probability distribution of x_t is obtained by looking at the values of x_t shown on the right of the figure. As the sample size n increases, the sample space of x gets filled in between 0 and 1, such that the sample values are arbitrarily close to each other, but no value is ever repeated exactly.

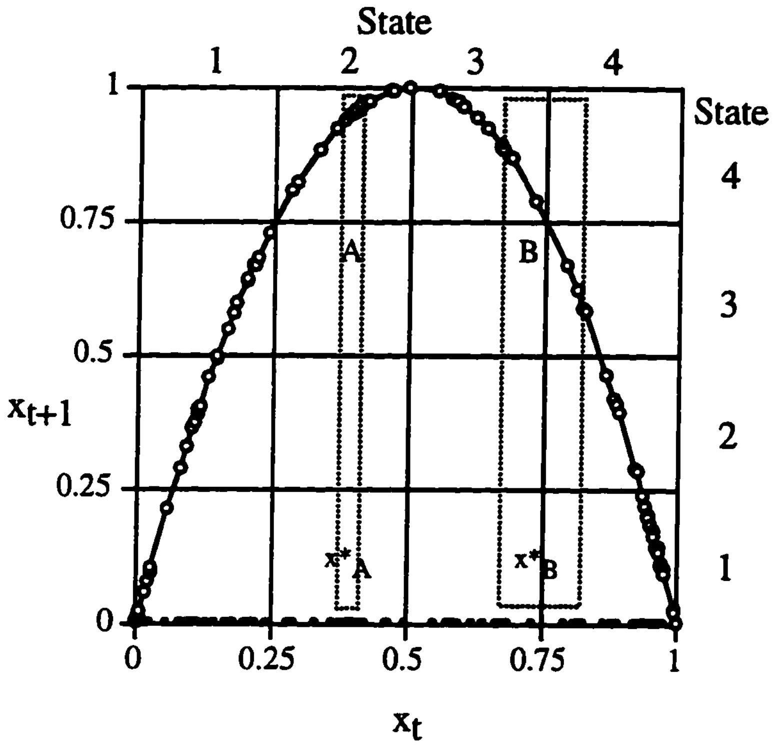


Figure 2-2. A plot of x_{t+1} vs x_t for the time series generated from the model $x_{t+1}=(1-4(x_t-0.5)^2)$. The state space for x is discretized into four states as shown. Also shown are windows A and B with whiskers located over two points x^*_A and x^*_B. These windows represent a k nearest neighborhood of the corresponding x_t. In general, these windows will not be symmetric about the x_t of interest. One can think of state transition probabilities using these windows in much the same way as with the multistate Markov chain. A value of x_{t+1} conditional to point A or B can be bootstrapped by appropriately sampling with replacement one of the values of x_{t+1} that fall in the corresponding window.

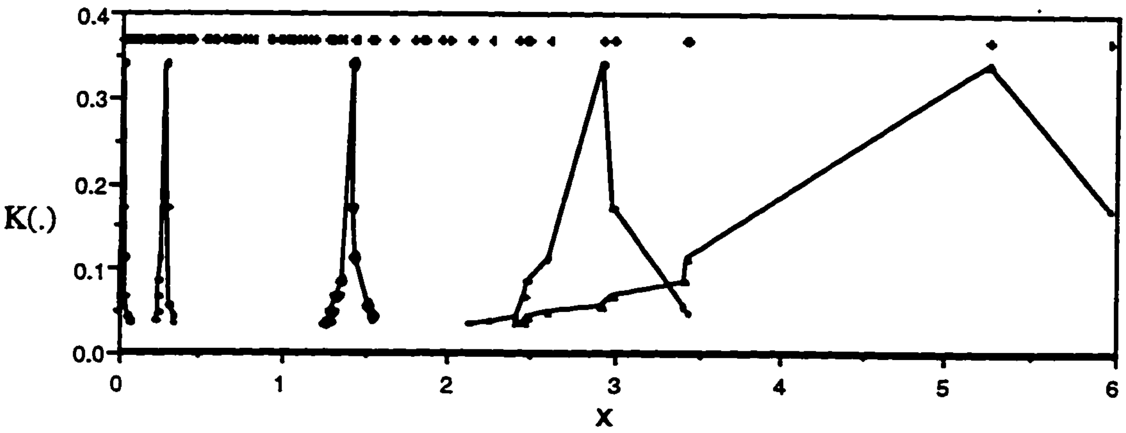
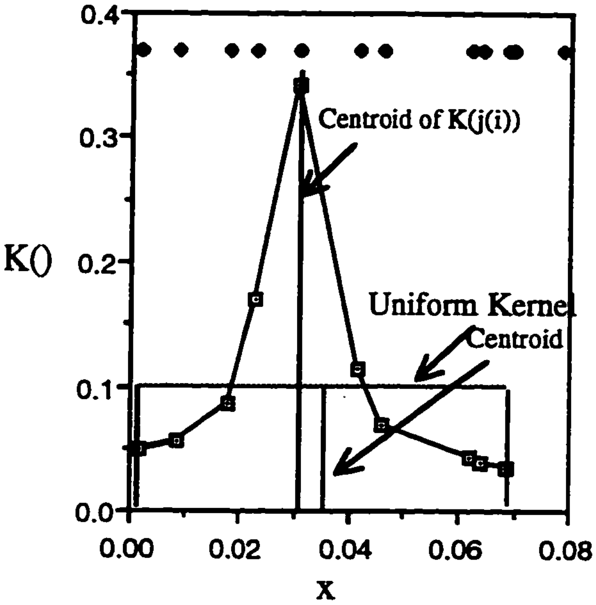
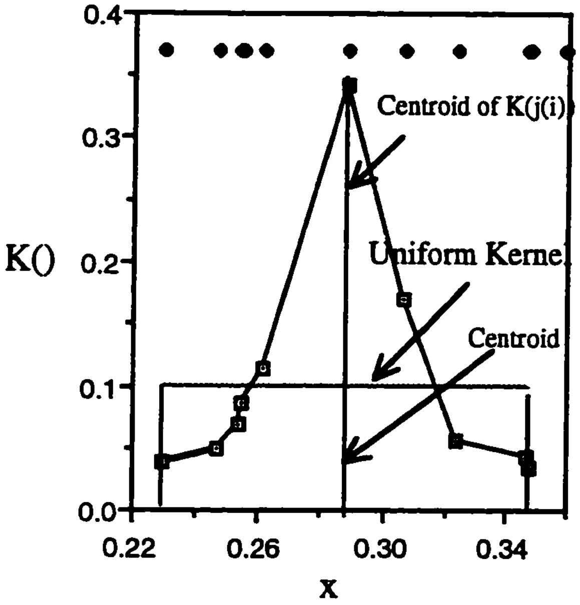


Figure 2-3. Illustration of resampling weights, K(j(i)), at selected conditioning points x_i, using k=10 with a sample of size 100 from an exponential distribution with parameter =1. The original sampled values are shown at the top of the figure. Note how the bandwidth and kernel shape vary with sampling density.

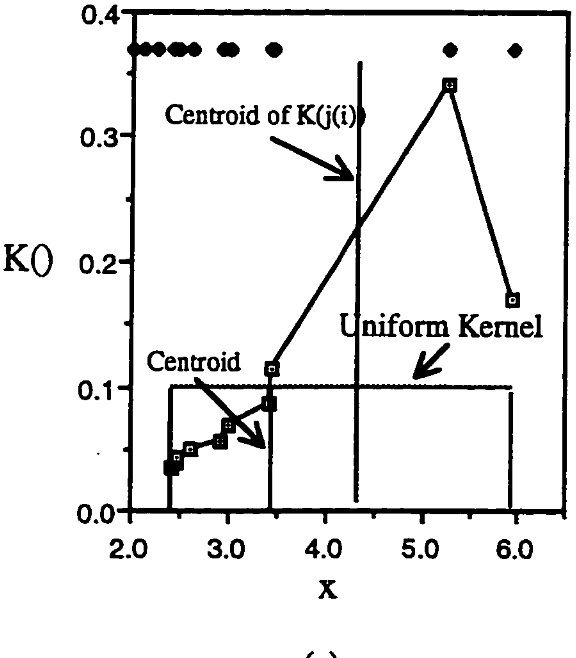


(a)

Figure 2-4. Illustration of weights K(j(i)) versus weights from the uniform kernel applied at three points selected from Figure 2-3. The uniform kernel weights are 1/k for each jεJ(i,k). The effective centroid corresponding to each kernel for each conditioning point i (in each case with the highest value of K(j(i)) is shown. For i in the interior of the data (Figure 2-4b), the centroids of both the uniform kernel and K(j(i)) coincide with i. Towards the edges of the data (Figures 2-4a and 2-4c), the centroid corresponding to K(j(i)) is closer to i, than that for the uniform kernel. The K(j(i)) are thus less biased than the uniform kernel for a given k. The kernel K(j(i)) also has a lower variance than the uniform kernel for a given value of k.



(b)



(c)

Figure 2-4. Continued.

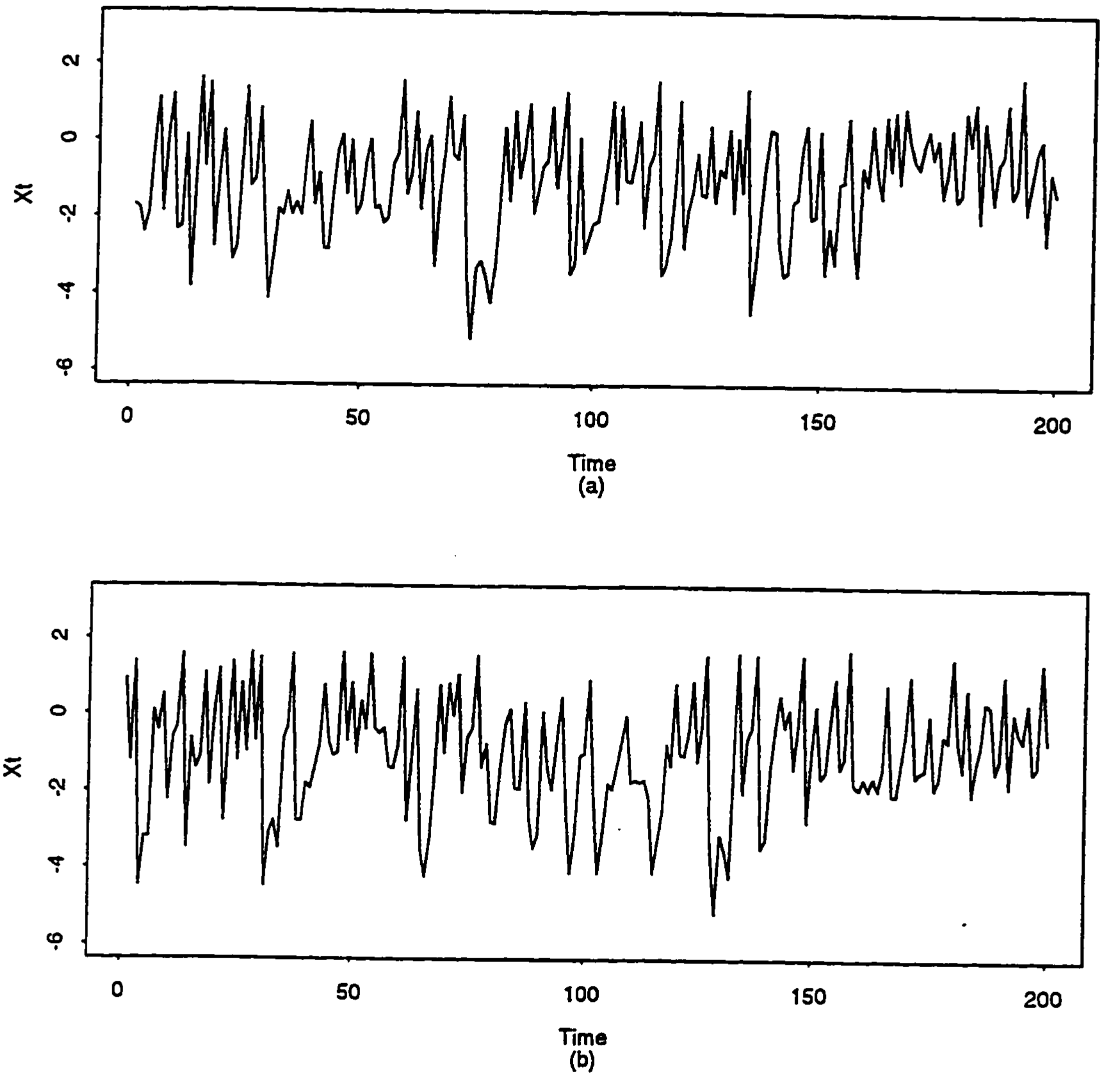
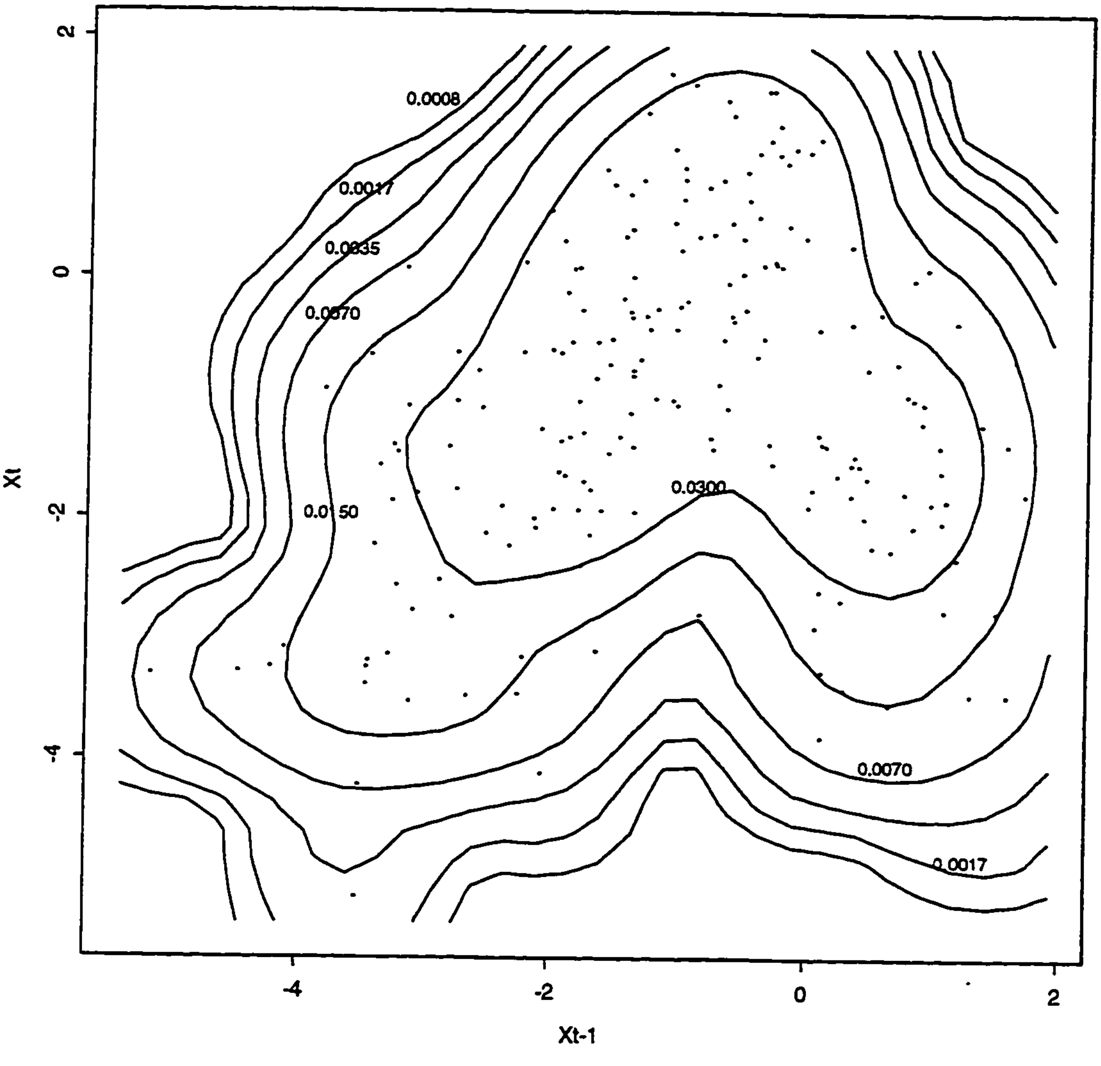
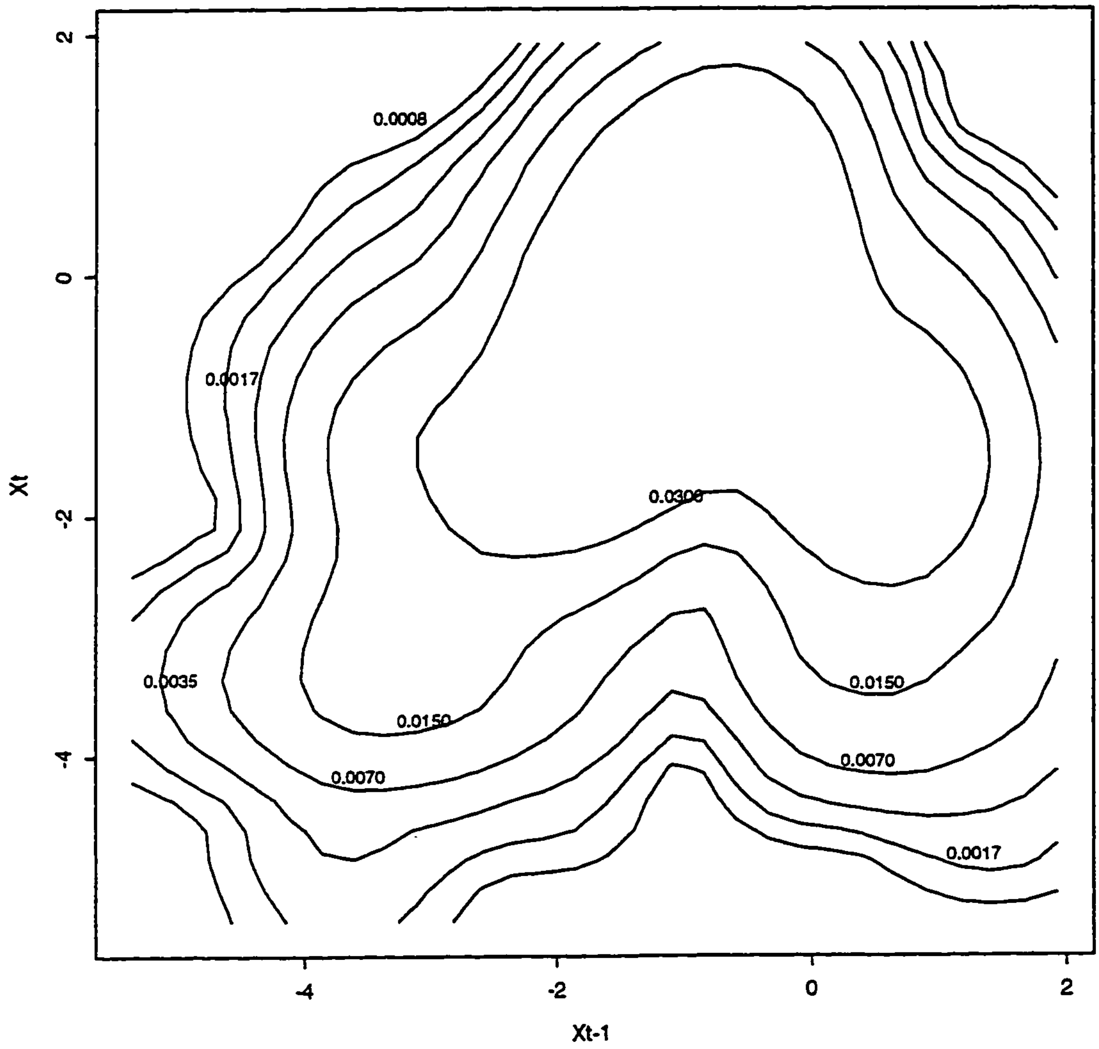


Figure 2-5. (a) A time series trace from the SETAR model described by Equation 2.11 and (b) a time series trace from a k-nn resample of the original SETAR sample.



(a)

Figure 2-6. (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the SETAR sample, the thick curve denoting a LOWESS smooth, and (b) an average of the ASH estimates of the bivariate probability density $f(x_t, x_{t-1})$ across 100 k-nn resamples from the SETAR sample.



(b)

Figure 2-6. Continued.

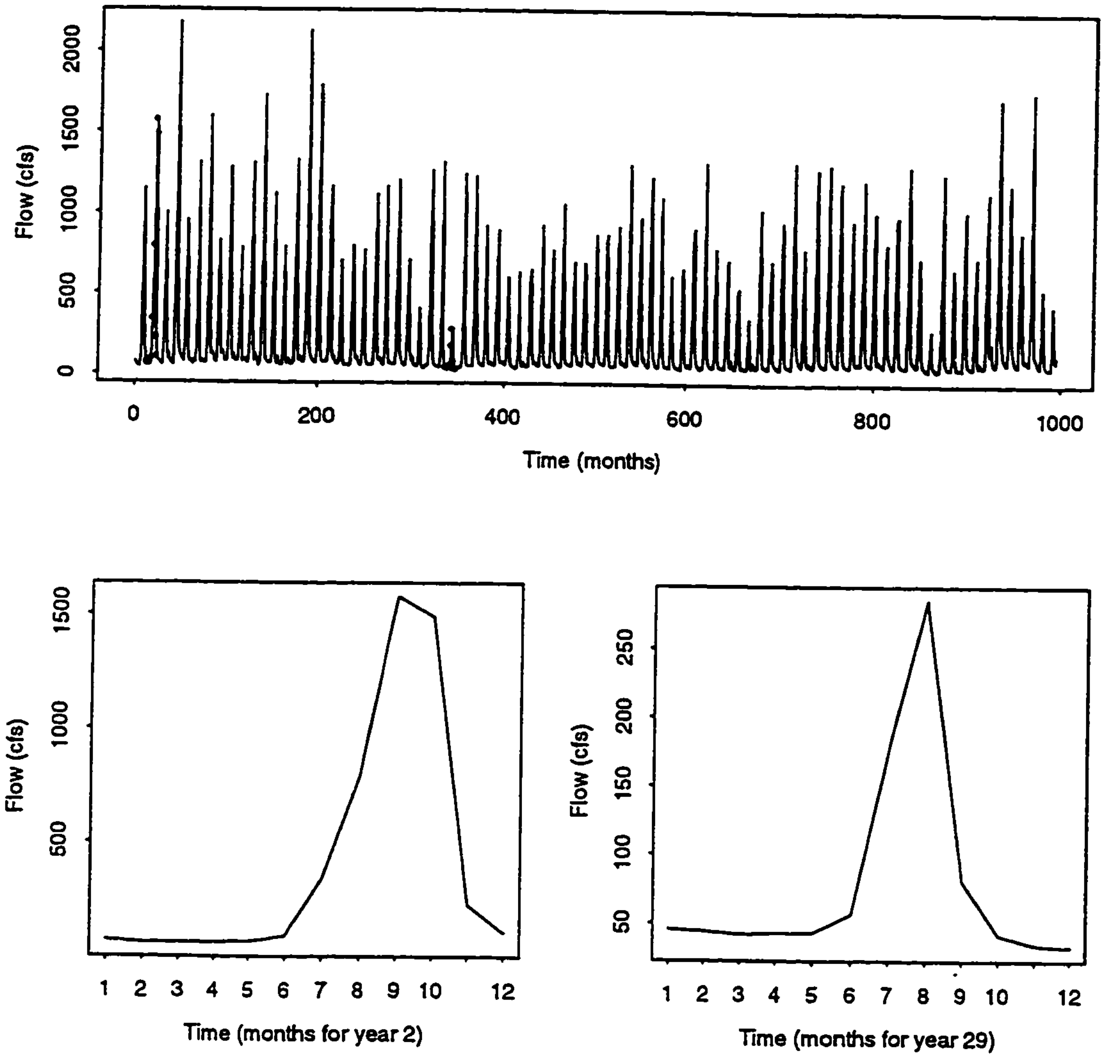
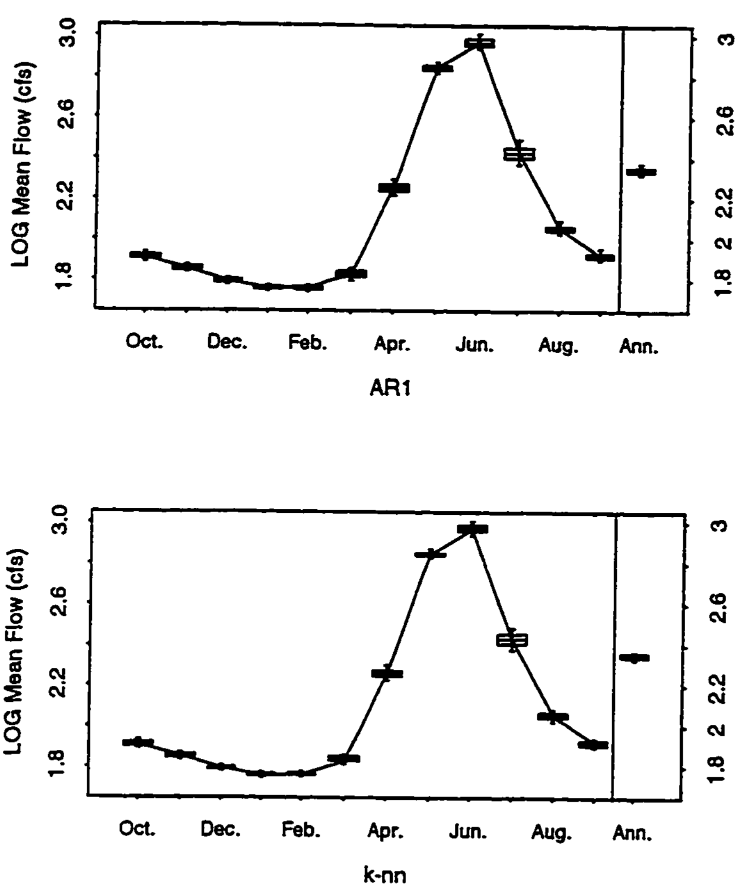
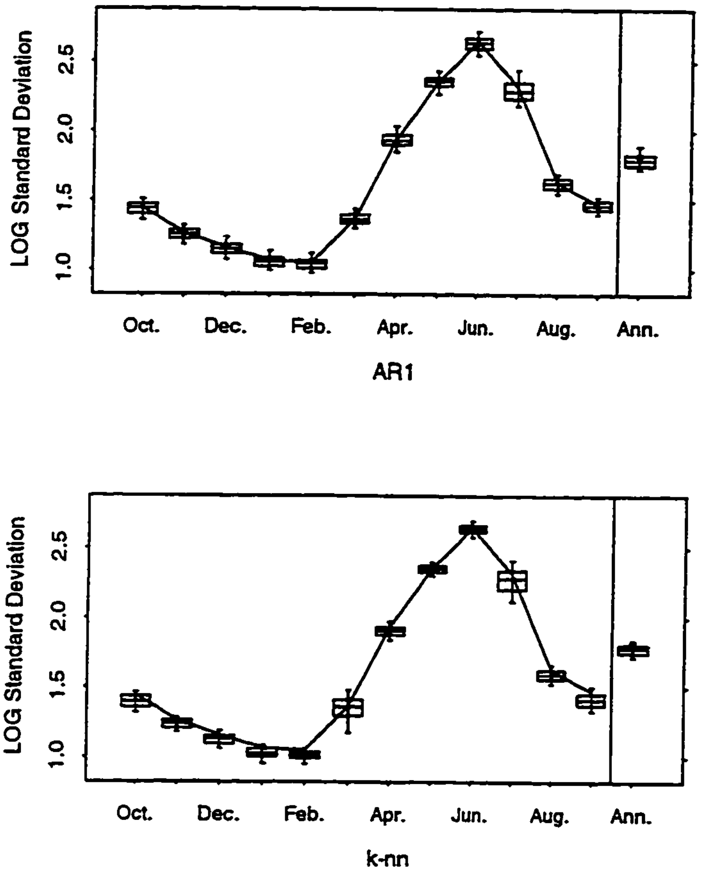


Figure 2-7. The Weber River monthly streamflow time series. Flows for 2 years (1906 and 1933) are shown in the inset. Note the asymmetry about the peak monthly flow in 1906 clearly shows that the time series is irreversible, i.e., its properties will be quite different in reverse time.



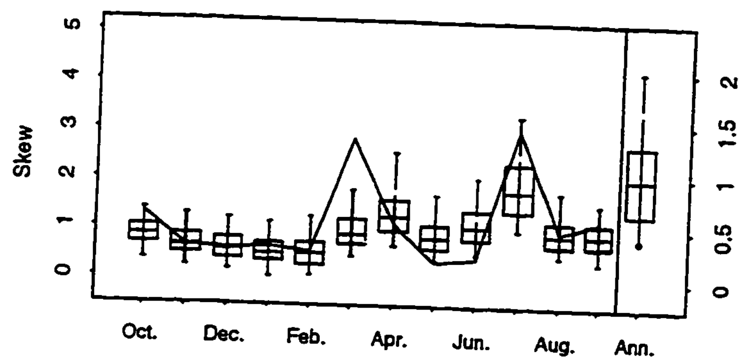
(a)

Figure 2-8. Monthly and annual statistics - (a) log(mean flow), (b) log(standard deviation), (c) skew, and (d) lag 1 correlation for simulated traces using AR1 and k-nn models for the Weber River data. The solid line in each figure represents the statistic for the historical sample. Boxplots compare the statistic over simulations. Boxplots comprise of a box being placed on the interquantile range from the multiple realizations of the statistic being compared, the line in the center of this box being the median. The whiskers extend to the 5% and 95% quantiles of the compared statistic. The dot above Ann. in each figure gives the historical annual statistic. Annual flows are not modeled explicitly by either simulator used.

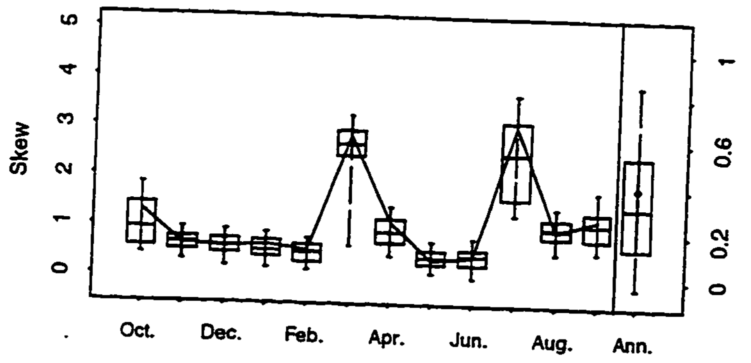


(b)

Figure 2-8. Continued.



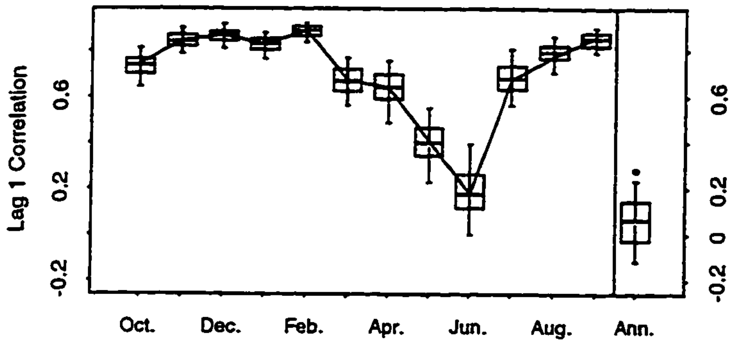
AR1



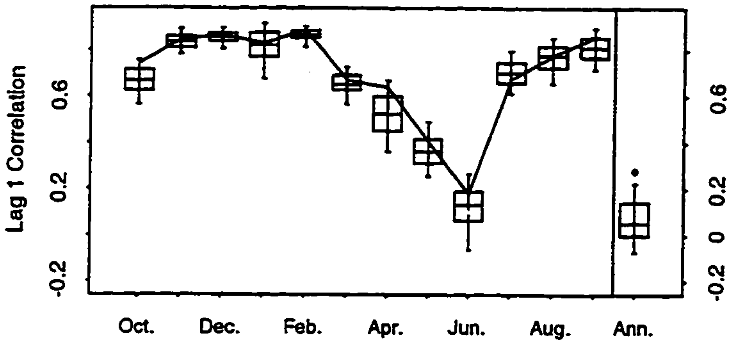
k-nn

(c)

Figure 2-8. Continued.



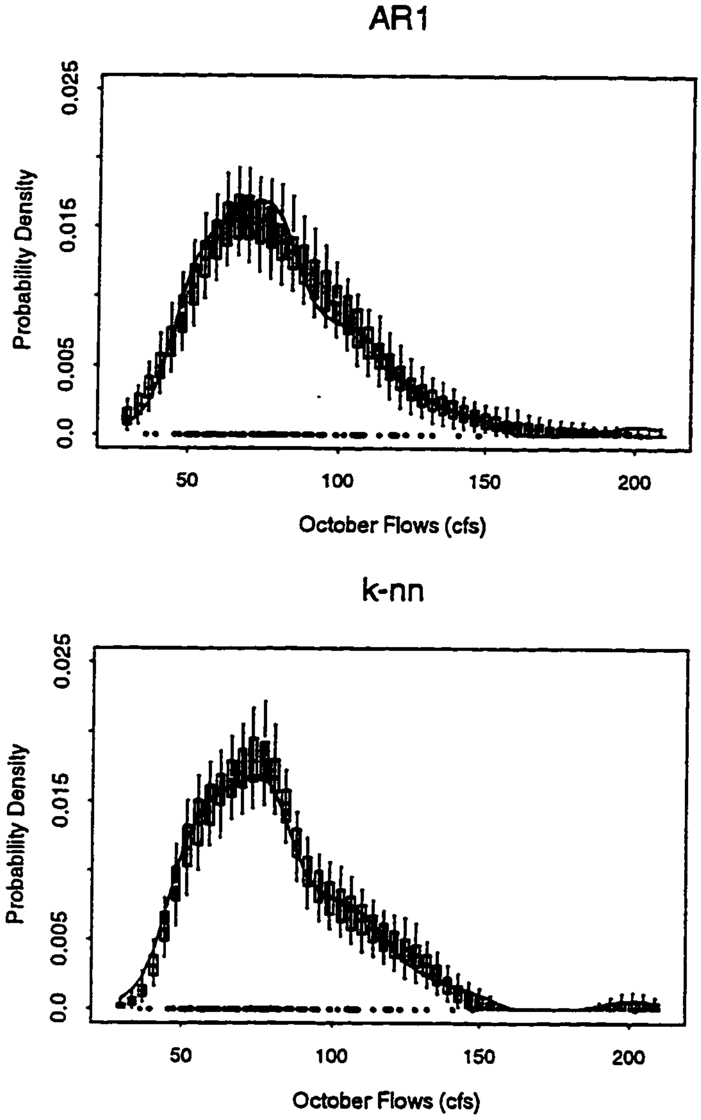
AR1



k-nn

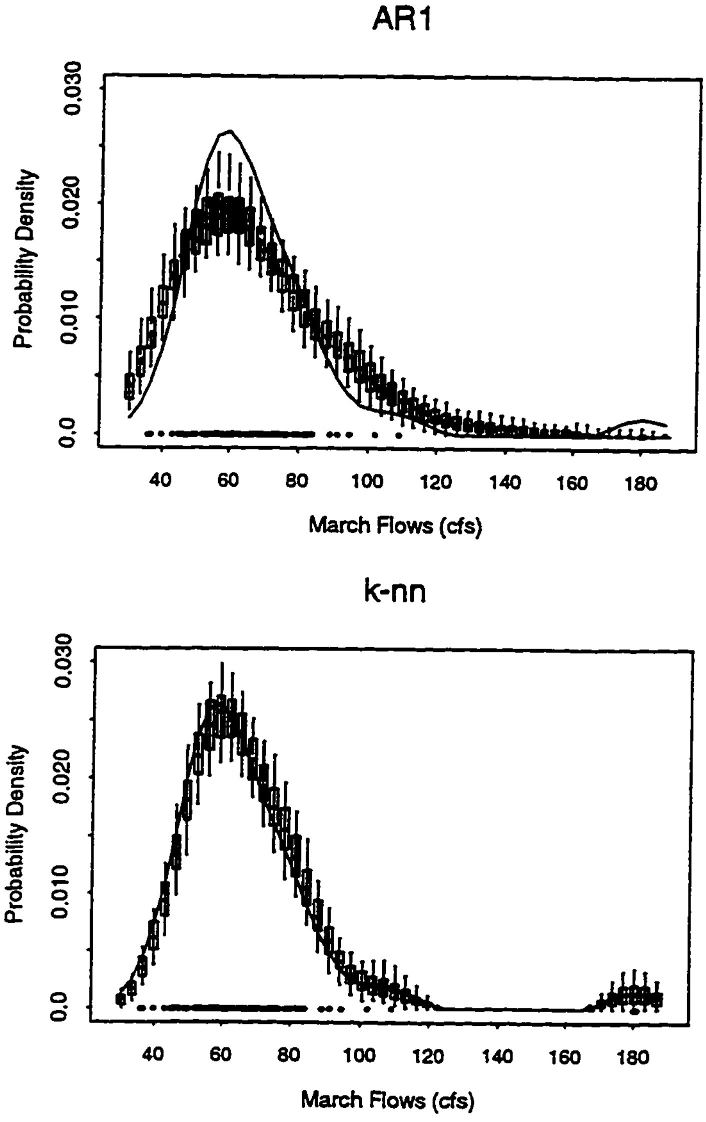
(d)

Figure 2-8. Continued.



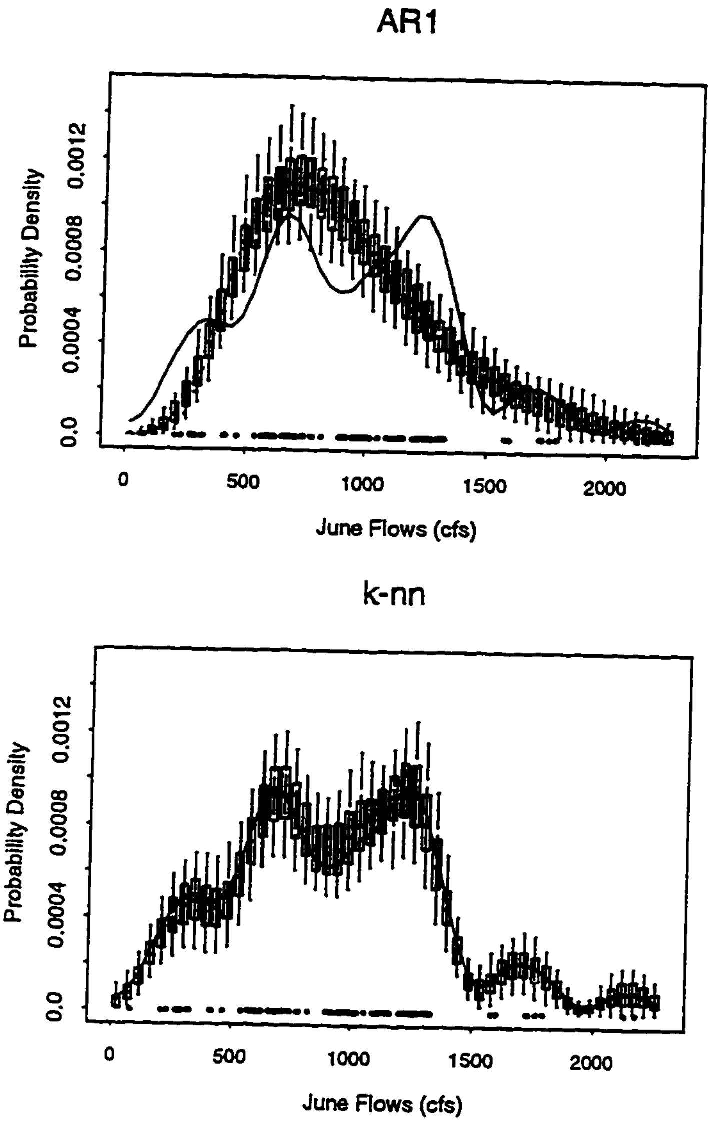
(a)

Figure 2-9. Marginal probability densities estimated by ASH from AR1 and k-nn samples for the Weber River data. In each case, the solid line is the ASH estimate for the historical record; the dashed line in the AR1 figure is the fitted AR1 model; the historical sample points are shown, and the boxplots (see description in Figure 2-8) depict the ASH estimates for all simulations. Results are presented for (a) October, (b) March, and (c) June.



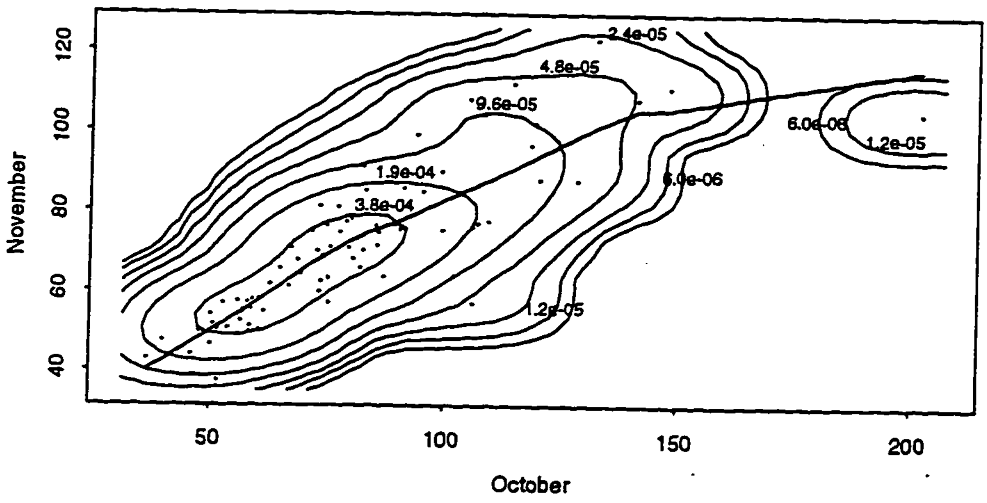
(b)

Figure 2-9. Continued.



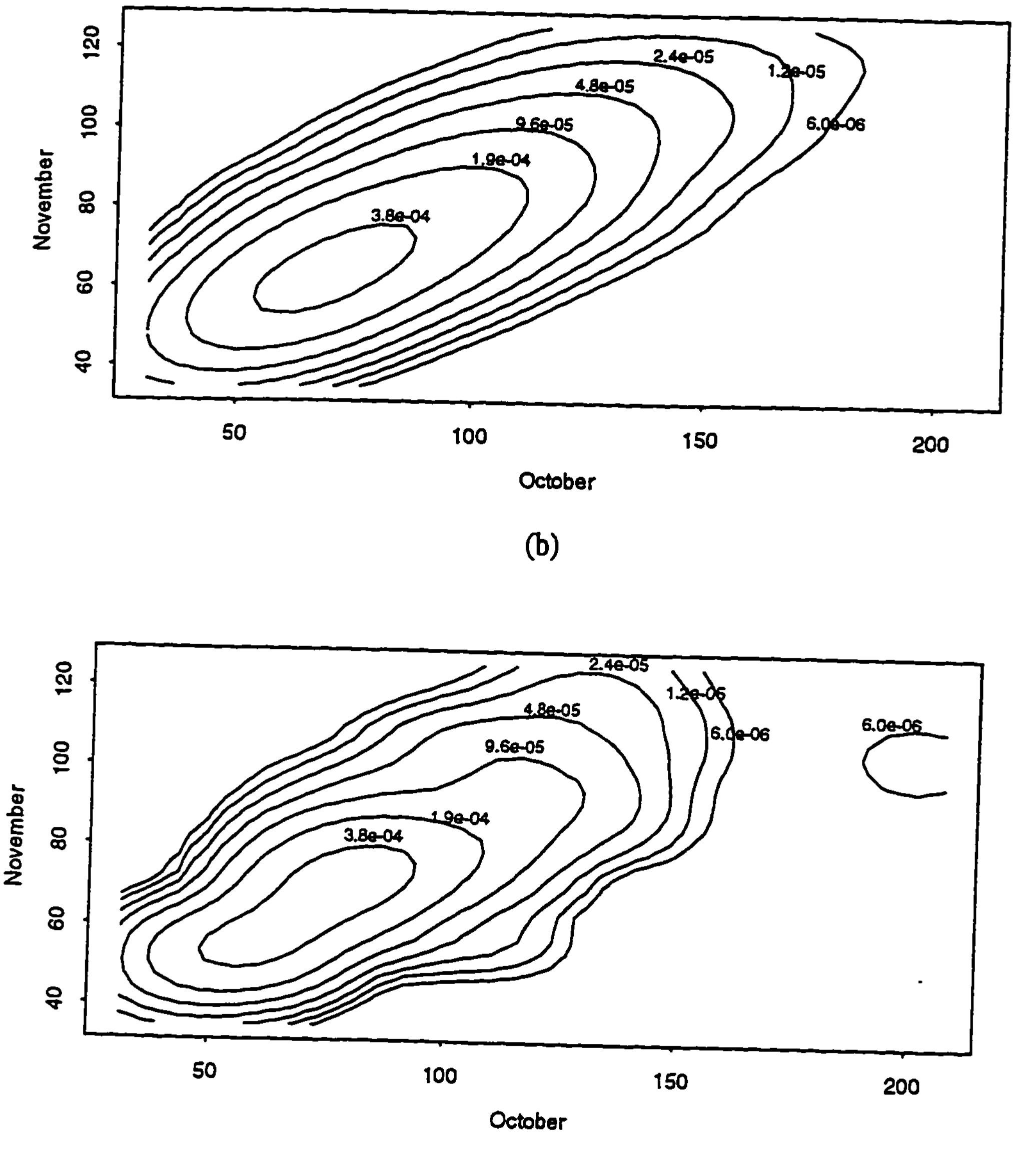
(c)

Figure 2-9. Continued.



(a)

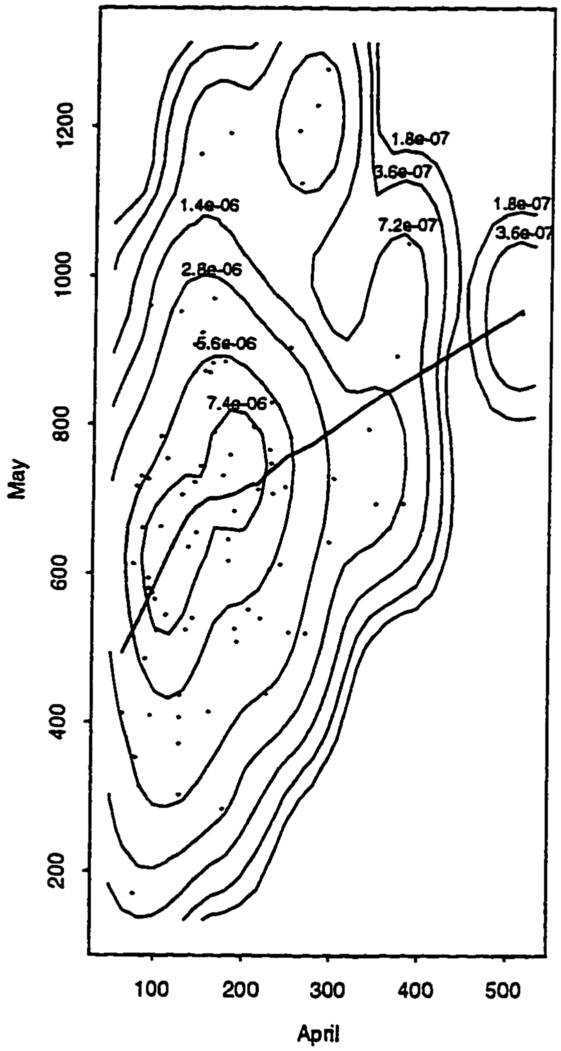
Figure 2-10. (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the October-November historical flows, the thick curve denoting a LOWESS smooth, (b) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 simulations from an AR1 model, and (c) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 k-nn resamples.



(b)

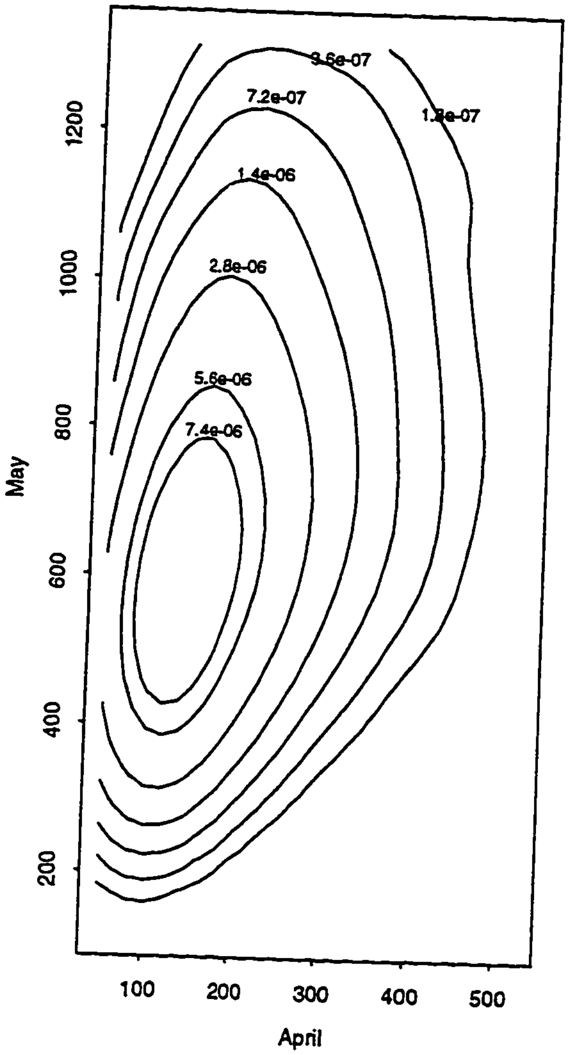
(c)

Figure 2-10. Continued.



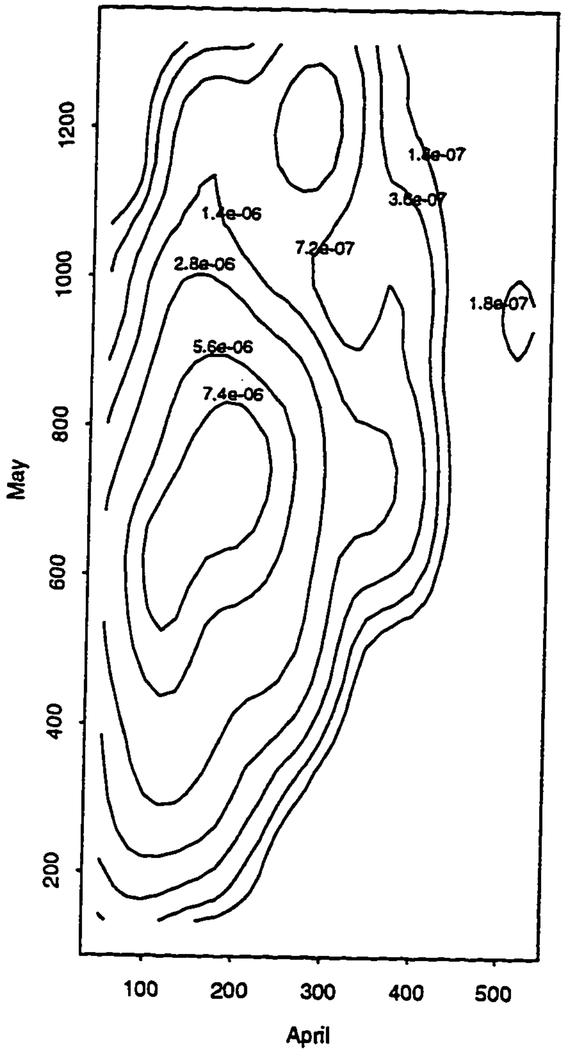
(a)

Figure 2-11. (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the April-May historical flows, the thick curve denoting a LOWESS smooth, (b) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 simulations from an AR1 model, and (c) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 k-nn resamples.



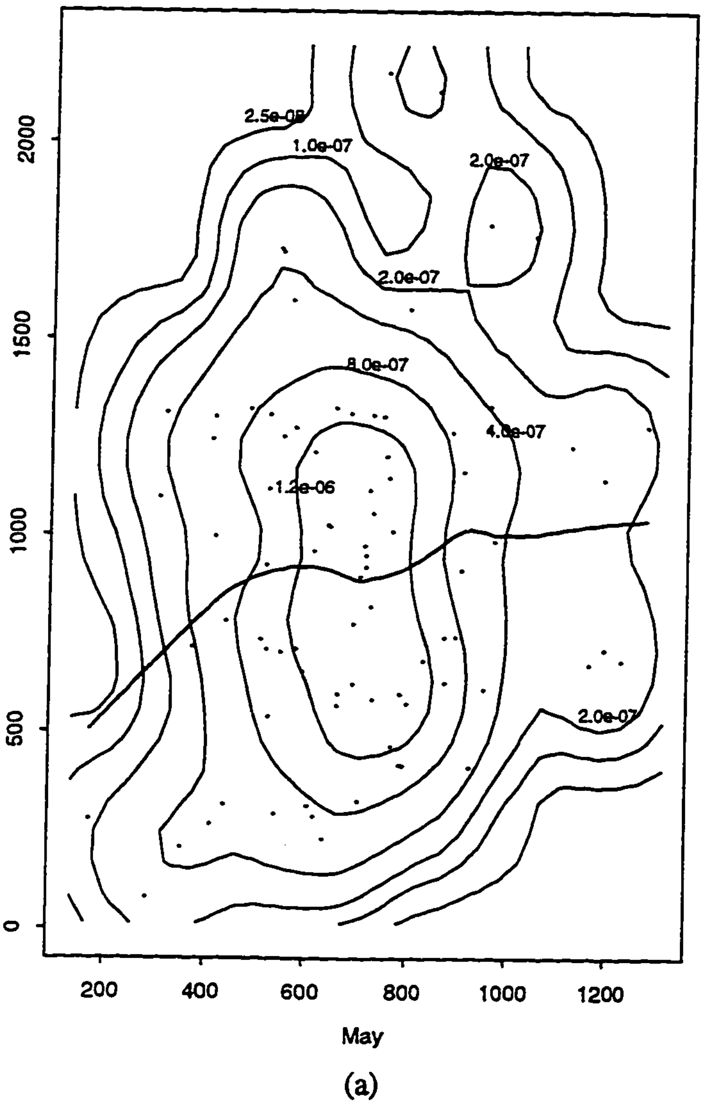
(b)

Figure 2-11. Continued.



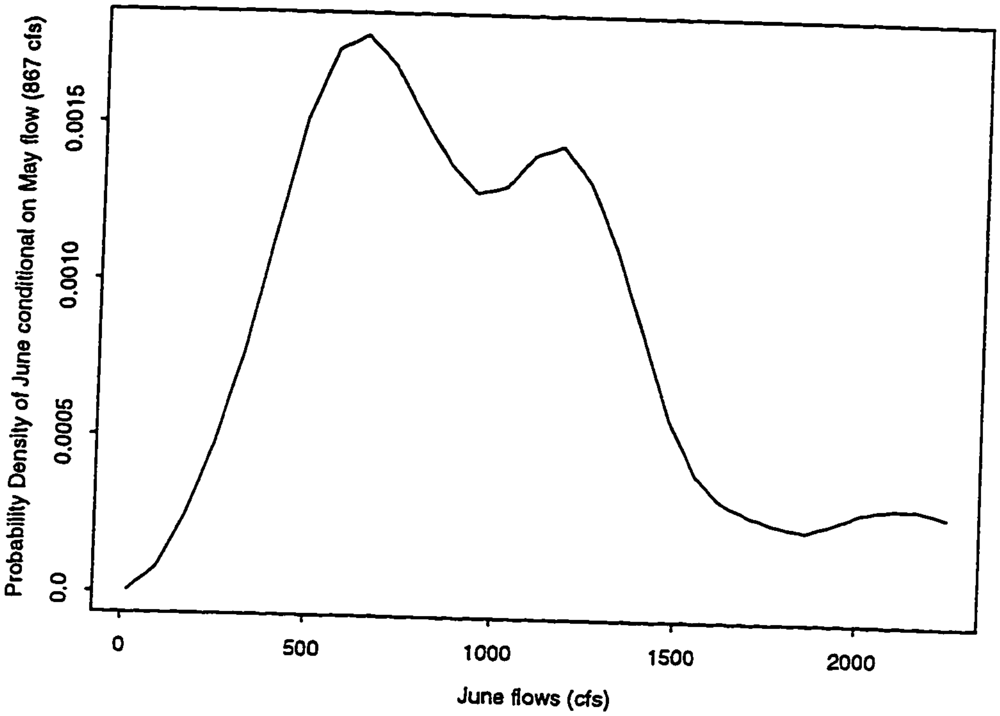
(c)

Figure 2-11. Continued.



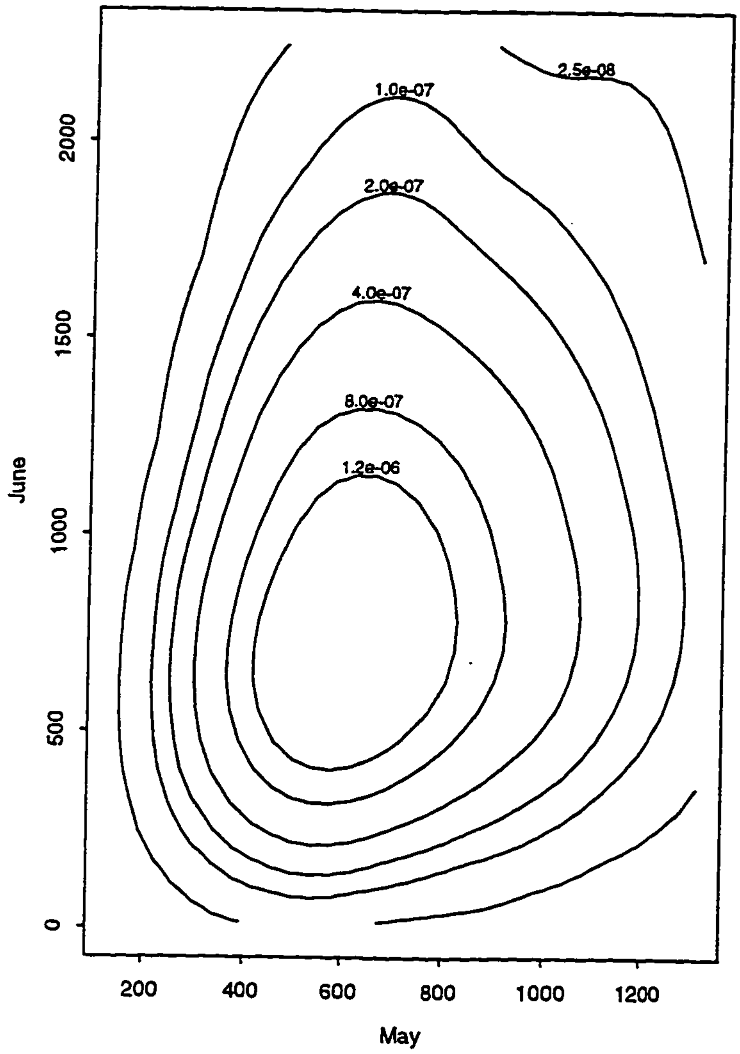
(a)

Figure 2-12. (a) An ASH estimate of the bivariate probability density $f(x_t, x_{t-1})$ for the May-June historical flows, the thick curve denoting a LOWESS smooth, (b) ASH estimate of the probability density of June flows conditional to a May flow of 867 cfs, (c) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 simulations from an AR1 model, and (d) an average of the ASH estimates of $f(x_t, x_{t-1})$ across 100 k-nn resamples.



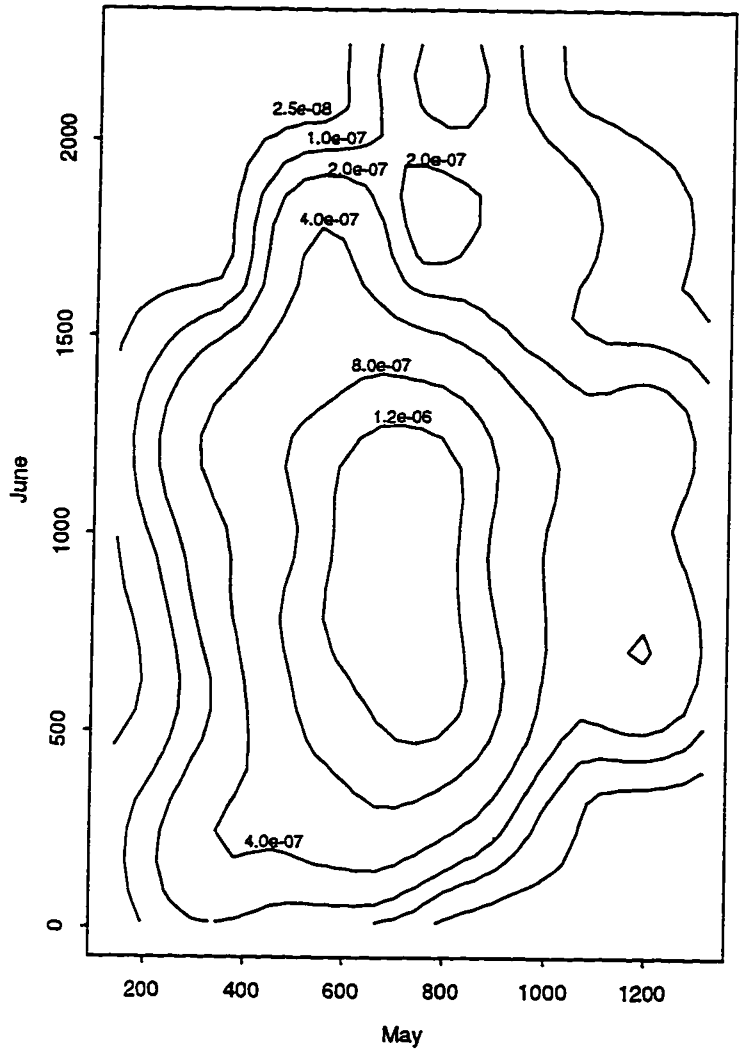
(b)

Figure 2-12. Continued.



(c)

Figure 2-12. Continued.



(d)

Figure 2-12. Continued.

CHAPTER 3

STREAMFLOW SIMULATION USING NONPARAMETRIC DENSITY ESTIMATION[1]

Abstract

In this chapter kernel estimates of the joint and conditional probability density functions are used to generate synthetic streamflow sequences. Streamflow is assumed to be a Markov process with time dependence characterized by a multivariate probability density function. Kernel methods are used to estimate this multivariate density function. Simulation proceeds by sequentially resampling from the conditional density function derived from the kernel estimate of the underlying multivariate probability density function. This is a nonparametric method for the synthesis of streamflow that is data-driven and avoids prior assumptions as to the form of dependence (e.g., linear or nonlinear) and the form of the probability density functions (e.g., Gaussian). We show, using synthetic examples with known underlying models, that the nonparametric method presented is more flexible than the conventional models used in stochastic hydrology and is capable of reproducing both linear and nonlinear dependence. The effectiveness of this model is illustrated through its application to simulation of monthly streamflow at a station in the Snake River Basin.

Introduction

A goal of stochastic hydrology is to generate synthetic streamflow sequences that are statistically similar to observed streamflow sequences. Statistical similarity implies sequences that have statistics and dependence properties similar to those of the historical record. These sequences represent plausible future streamflow scenarios under the

---

[1] Coauthored by Ashish Sharma, David G. Tarboton, and Upmanu Lall.

assumption that the future will be similar to the past. Synthetic streamflow sequences are needed in simulation studies to analyze alternative designs, operation policies, and rules for water resources systems. In this chapter, we present a nonparametric approach for the generation of synthetic streamflow sequences. The utility of this approach relative to conventional parametric methods is demonstrated through applications to monthly streamflow from the Snake River at Weiser, Idaho, and to samples generated from linear and nonlinear models with known statistical attributes.

Consider a time series $\{X_1, X_2, \ldots, X_t, \ldots\}$ where the X_t represent streamflow quantities at time t. In practice, the dependence structure of streamflow sequences is often assumed to be Markovian, i.e., dependent only on a finite set of prior values. With this assumption, *Bras and Rodriguez-Iturbe* [1985] note that stochastic streamflow models are an exercise in conditional probability. An order p model simulates X_t based on the previous values, i.e., $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$. This requires that a $d = p+1$ dimensional joint probability distribution be specified. Simulation can proceed from the conditional density function defined as:

$$f(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p}) = \frac{f(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-p})}{\int f(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-p})\, dX_t} \qquad (3.1)$$

Traditional parametric models specify Equation (3.1) through assumed distributions. Here, it is suggested that streamflow may instead be directly modeled from empirical, data-driven estimates of the joint and conditional density functions given in Equation (3.1). Nonparametric estimates of these density functions are developed directly from the historical data. A method is considered nonparametric if it can reproduce a broad class of possible underlying density functions [*Scott*, 1992].

Nonparametric methods for density estimation strive to approximate the underlying density locally using data from a small neighborhood of the point of estimate [*Lall*, 1995]. They impose only weak assumptions, such as continuity of the target function, rather than a priori specification or choice of a particular parametric probability distribution (e.g., Gaussian, Lognormal, etc.). A perusal of the statistical literature shows that nonparametric statistical estimation, using splines, kernel functions, nearest-neighbor methods, and orthogonal series methods is an active area, with major developments still unfolding. *Silverman* [1986] and *Scott* [1992] are good introductory texts. Applications of these methods in hydrology are reviewed by *Lall* [1995].

Our model is based on a nonparametric kernel density estimate of the p+1 dimensional density function $f(X_t, X_{t-1}, \ldots, X_{t-p})$, which is then used in (3.1) to estimate the conditional density function that forms the basis for generation of synthetic streamflow series. This is called a nonparametric order p or NPp model. It has the following advantages:

(1) Statistical attributes of the data are automatically honored since one works with a smoothed empirical frequency distribution based directly on the historical data. Such attributes include nonlinear dependence and inhomogeneity (i.e., statistical properties that vary by streamflow state).

(2) The somewhat tenuous issue of choosing between different models for the probability distribution is sidestepped.

(3) Considerations related to the above two points lead to a procedure that is more reliable for streamflow simulation and hence for decisions on reservoir operation and design.

We shall first review some of the traditional approaches, noting their shortcomings and motivating the need for the nonparametric approach. Kernel density estimation is reviewed next. We then describe the NPp model and illustrate its use with synthetic data from a linear autoregressive (AR1) model and a self exciting threshold autoregressive (SETAR) model [*Tong*, 1990]. These tests demonstrate the effectiveness of the NPp approach in representing both linear and nonlinear systems, without a prior specification of the model equations. An application of our model to simulate monthly streamflow from the Snake River at Weiser, Idaho, is then presented and results are compared to those from an AR1 model with marginal densities chosen from the best fitting of four commonly used probability density functions.

Background

Annual and monthly streamflows have been modeled extensively using autoregressive moving average (ARMA) type models [*Yevjevich*, 1972; *Hipel et al.*, 1977; *McLeod et al.*, 1977; *Pegram et al.*, 1980; *Salas et al.*, 1980; *Loucks et al.*, 1981; *Stedinger and Vogel*, 1984; *Bras and Rodriguez-Iturbe*, 1985; *Stedinger et al.*, 1985b]. The early Thomas-Fiering model [*Thomas and Fiering*, 1962; *Fiering*, 1967; *Beard*, 1967], an autoregressive lag 1 model with seasonally varying coefficients, is a good example of this approach.

$$(X_{t,j} - m_j) = \rho_j \frac{\sigma_j}{\sigma_{j-1}} (X_{t,j-1} - m_{j-1}) + \sigma_j (1 - \rho_j^2)^{1/2} \, W_{t,j} \qquad (3.2)$$

where $X_{t,j}$ is the seasonal streamflow at year t and season (month) j, ρ_j is the lag-one correlation coefficient between seasons j and j-1, m_j is the mean streamflow in season j, σ_j is the standard deviation of flow in season j and $W_{t,j}$ is an independent random variable with mean 0 and variance 1. By allowing the noise term $W_{t,j}$ to be from a skewed distribution [*Lettenmaier and Burges*, 1977; *Todini*, 1980], streamflow from a skewed distribution can be approximated. Thus this model reproduces the mean, variance, and correlations between monthly streamflows and approximates the skewness. These are the variables traditionally considered as most important by stochastic hydrologists. As written, this model only applies to a single site; however, it is illustrative of a very general class of ARMA models for single sites and in a multivariate context for multiple sites or seasons that have been developed and applied extensively in hydrology over the years and described at length in texts on the subject [*Salas et al.*, 1980; *Loucks et al.*, 1981; *Bras and Rodriguez-Iturbe*, 1985].

Such models can be viewed as special cases of a general multivariate ARMA(p,q) model:

$$X_{t+1} = \sum_{j=0}^{p} A_j X_{t-j} + \sum_{j=0}^{q} B_j W_{t-j} + U \qquad (3.3)$$

where X_t is a vector of the variables of interest, including annual and seasonal flows at all sites, A_j and B_j are coefficient matrices, U is a vector of coefficients, and W_t is a vector of independent random innovations. The first term represents an autoregressive component and the second term a moving average component. In all but the simplest univariate models it is impractical to assume anything but a Gaussian distribution for the W_t. This is equivalent to the assumption of a multivariate Gaussian distribution for the time series dependence structure. The ARMA model is then defined through the estimation of the parameters A_j, B_j, U, and the model order (p,q). To account for the fact that the real streamflows are not Gaussian, the flows are often first transformed to a Gaussian distribution and then the transformed variables are used with (3.3) [*Stedinger*, 1981; *Stedinger and Taylor*, 1982; *Stedinger et al.*, 1985a]. Reproducing moments in the original coordinates may then be difficult.

The general linear model depicted by Equation (3.3) is a special case of the conditional density function of Equation (3.1). This multivariate Gaussian structure with transformed marginal distributions (denoted MGTM here) has with few exceptions [*Yakowitz*, 1985; *Smith*, 1991; *Smith*, 1992; *Lall and Sharma*, 1996] underlain practically all stochastic hydrology to date. The Lall and Sharma work is very similar in spirit to this work, though the approach is that of a nearest-neighbor bootstrap, rather than kernel density estimation. We believe that both are good alternatives that need to be considered for streamflow simulation.

The preceding discussion reveals the basic structure of current time series estimation methods, and hints at their restricted view of the possibilities of variation in hydrological time series. The main reasons for the prevalence of linear ARMA models for hydrologic time series analysis may be:

(1) The framework has been well developed in the statistical literature for stationary processes.

(2) The techniques are well understood and taught.

(3) Software for multivariate analysis has been developed by a number of people, is readily available and does not pose a severe computational burden.

(4) The models have been reasonably successful for the analysis of monthly and annual streamflow records. This is particularly true of shorter records. They likely provide a good first approximation to the underlying time series process.

Some drawbacks of the MGTM approach are:

(1) Only a limited degree of heterogeneity in the statistical dependence structure is admitted through the normalizing transform. The dependence of variance of streamflow

on streamflow magnitude is often noted. There is evidence in some streamflow data that correlations are different depending on whether flows are low or high. We give an example of this in section 6 below using state-dependent correlation statistics defined in Appendix A. In a MGTM model, the correlation structure is fixed regardless of flow magnitude. Among others, *Yevjevich* [1972] has argued for the systematic identification of nonstationarities in the mean of the time series (e.g., jumps, periodicities) and their removal to yield a stationary time series that can be analysed by standard methods. However, such features may be part of the underlying dynamics and important to model behavior (e.g., to a drought regime) that may be related to threshold-dependent processes.

(2) The MGTM models impose a time reversible structure. The joint distributions of $(X_t, X_{t+1}, ...X_{t+m})$ and of $(X_t, X_{t-1}, ...X_{t-m})$ are identical. *Tong* [1990] shows an example of daily streamflow that is not time reversible, and argues that the dynamics of physical processes is time irreversible.

(3) The choice of a distribution for W_t or of an appropriate transform can be problematic. For short series, statistical tests are unable to distinguish between candidate distributions [*Kite*, 1977]. None of the common transformations may be applicable. Figure 3-1 illustrates this problem with July monthly streamflow from the Beaver River at Beaver, Utah. The figure shows the histogram of monthly flow, with three commonly used distributions fitted to the data. The histogram has bimodality that cannot be reproduced by any of the distributions commonly used. This figure also shows a kernel density estimate. Note that this is effectively a smoothing of the histogram. The following Filliben correlation statistics [*Grygier and Stedinger*, 1990] test the goodness of fit for each distribution in Figure 3-1. Kernel density estimate: 0.997; Normal: 0.962; Lognormal: 0.980; Three parameter Lognormal: 0.985 and Gamma: 0.987. Goodness

of fit is measured by how close this statistic is to 1. By this measure the nonparametric density estimate fits better than any of these commonly used parametric choices. A χ^2 test [*Benjamin and Cornell*, 1970] rejected at the 95% level the hypothesis that this histogram was from a Gaussian distribution. The χ^2 test, however, would not reject any of the other distributions, including the nonparametric density estimate, which is typical of the inability to distinguish between candidate distributions.

(4) The synthetic traces generated by MGTM replicate the first few (two or three) moments of the underlying dependence structure. Consequently, the generated series may bear little resemblance to the observed series in terms of persistence and threshold crossings, factors that are of interest to hydrologists. The ARMA models also are incapable of displaying sudden bursts or jumps, a feature that may often be observed during an otherwise prolonged drought.

(5) *Salas and Smith* [1981] and *Salas et al.* [1980] discussed physical justifications for ARMA models and showed that a linear control system representation of basin processes can lead to ARMA models of streamflow. However, other factors emerge, when one considers the relationship of streamflow to some causative factors. For instance, in snow-fed basins, the streamflow response during snow-melt months is a threshold response to temperature. The dynamics of soil moisture is hysteretic and nonlinear. The dynamics of vegetative consumptive use and retention of water is also quite different during wet and dry periods, and as a function of temperature. Runoff generation mechanisms during protracted wet or dry periods will consequently be different. While these comments have more direct bearing on streamflow at time scales shorter than a month or year, they are relevant for the longer time scales in that they influence the variance of the streamflow at these time scales.

(6) Despite the fact that nearly 30 years have elapsed since the classical time series (AR) models were introduced to practicing hydrologists, acceptance and application of these models for drought analysis and reservoir operation by practitioners has been limited. They often prefer to base their decisions on the historical record (or a resampled proxy thereof). *Kendall and Dracup* [1991] note that the index sequential method, which is basically a sequential resampling of the historical record, appears to be the procedure of choice in many water management agencies, including the California Department of Water Resources, U.S. Bureau of Reclamation, Los Angeles Department of Water, and the Metropolitan Water District of Southern California. This is practiced in spite of the recognition that history is unlikely to repeat, and that the record is perhaps woefully short. The parametric ARMA models are regarded with suspicion by many practitioners, in part because they do not seem to replicate aspects of the historical record seen qualitatively by the practitioner, and in part because the "fitting" process undergone in developing such a model may leave the practitioner with a sense of uncertainty equivalent to the uncertainty imparted by just using the record.

In summary, while the MGTM ARMA framework is indeed useful in certain contexts, a more flexible time series analysis method that is capable of reproducing additional features of hydrologic data is needed. The success of linear ARMA models with some hydrologic data sets may be fortuitous, and a consequence of short records. The technical issues are nonlinearity, nonstationarity, and inhomogeneity in the underlying dependence structure. Parametric nonlinear models [*Bendat and Piersol*, 1986; *Tong*, 1990] can be used in place of the linear ARMA models to model nonlinear time series. The use of such models, however, still requires specification of the form of nonlinear dependence, something which may be difficult to do in practice. From a practitioner's perspective the key issue is reproducibility of observed data characteristics, simplicity, and dependability. The nonparametric techniques proposed here avoid the difficult model specification issues associated with parametric linear or nonlinear models. They amount to resampling from the original data, with perturbations, and reproduce directly the characteristics of the original data in a simple and dependable way.

Kernel Density Estimation

Kernel density estimation entails a weighted moving average of the empirical frequency distribution of the data. Most nonparametric density estimators can be expressed as kernel density estimation methods [*Scott*, 1992]. In this chapter we use multivariate kernel density estimators with Gaussian kernels and bandwidth selected using least squares cross validation [*Scott*, 1992]. This bandwidth selection method is one from among the many available methods. Our methodology is intended to be generic and should work with any bandwidth and kernel density estimation method. This section reviews kernel density estimation first in a univariate then in a multivariate setting and gives details of the Least Squares Cross Validation (LSCV) procedure for estimating bandwidth. For a review of hydrologic applications of kernel density and distribution function estimators, readers are referred to *Lall* [1995]. *Silverman* [1986] and *Scott* [1992] are good introductory texts.

A univariate kernel probability density estimator is written:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{n\,h} K\left(\frac{x - x_i}{h}\right) \tag{3.4}$$

where there are n sample data x_i. K(.) is a kernel function that must integrate to 1 and h is a parameter called the bandwidth that defines the locale over which the empirical frequency distribution is averaged. There are many possible kernel functions given in texts such as *Silverman* [1986] and *Scott* [1992]. The Gaussian kernel function, a popular and practical choice, is used here.

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \tag{3.5}$$

The density estimate in (3.4) is formed by summing kernels with bandwidth h centered at each observation x_i. This is similar to the construction of a histogram where individual observations contribute to the density by placing a rectangular box (analogous to the kernel function) in the prespecified bin in which the observation lies. The histogram is discrete and sensitive to the position and size of each bin. By using smooth kernel functions, the kernel density estimate in (3.4) is smooth and continuous.

A multivariate extension of (3.4) and (3.5) for a vector **x** in d dimensions can be written as:

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2} \det(\mathbf{H})^{1/2}} \exp\left(- \frac{(\mathbf{x}-\mathbf{x}_i)^T \mathbf{H}^{-1} (\mathbf{x}-\mathbf{x}_i)}{2} \right) \quad . \tag{3.6}$$

where n is the number of observed vectors $\mathbf{x}_i$ and **H** is a bandwidth matrix that must be from the class of symmetric positive definite d x d matrices [*Wand and Jones*, 1994]. The above density estimate is formed by summing Gaussian kernels with a covariance matrix **H**, centered at each observation $\mathbf{x}_i$. A useful specification of the bandwidth matrix **H** is:

$$\mathbf{H} = \lambda^2 \mathbf{S} \tag{3.7}$$

Here, S is the sample covariance matrix of the data and λ^2 prescribes the bandwidth relative to this estimate of scale. These are parameters of the model that are estimated from the data. The procedure of scaling the bandwidth matrix proportional to the covariance matrix (Equation 3.7) is called "sphering" [*Fukunaga*, 1972] and ensures that all kernels are oriented along the principal components of the covariance matrix.

Silverman [1986] cites results indicating that sufficient conditions for convergence of the kernel density estimate to an underlying density function under broad conditions met by any kernel that is a usable probability density function are that as $n \rightarrow \infty$, $h \rightarrow 0$ and $nh \rightarrow \infty$. This also applies to λ in the multivariate context. However, the rate of convergence depends on how h or λ is chosen. Methods for choosing the bandwidth are based on evaluation of factors such as bias, $E\{f(\mathbf{x})-\hat{f}(\mathbf{x})\}$, variance, $\mathrm{Var}\{\hat{f}(\mathbf{x})\}$, Mean Square Error (MSE), Integrated Square Error (ISE), and Mean Integrated Square Error (MISE) of the estimate.

$$\mathrm{MSE} = E\{[f(\mathbf{x})-\hat{f}(\mathbf{x})]^2\} = \{E[f(\mathbf{x})-\hat{f}(\mathbf{x})]\}^2 + \mathrm{Var}\{\hat{f}(\mathbf{x})\} \tag{3.8}$$

$$\mathrm{ISE} = \int_{\Re^d} \left(\hat{f}(\mathbf{x})-f(\mathbf{x})\right)^2 d\mathbf{x} \tag{3.9}$$

$$\mathrm{MISE} = E \int_{\Re^d} \left(\hat{f}(\mathbf{x})-f(\mathbf{x})\right)^2 d\mathbf{x} \tag{3.10}$$

A small value of the bandwidth (h or λ) can result in a density estimate that appears "rough" and has a high variance. On the other hand, too high an h results in an "over smoothed" density estimate with modes and asymmetries smoothed out. Such an estimate has low variance but is more biased with respect to the underlying density. This bias-variance trade-off [*Silverman*, 1986] plays an important role in choice of h.

Taylor series expansion of the one-dimensional density estimate in Equation (3.4) can be used to show that the Asymptotic Mean Integrated Square Error (AMISE) is [*Silverman*, 1986; *Sain et al.*, 1994]:

$$\mathrm{AMISE}(h) \approx \frac{R(K)}{n\,h} + \frac{1}{4}\sigma_K^4 h^4 R(f'') \qquad (3.11)$$

where $R(g(x)) = \int g(x)^2\,dx$ for any function g(x) (either K(x) or f''(x)), f'' is the second derivative and $\sigma_K^2 = \int u^2 K(u)du$. This can be generalized to higher dimensions.

One choice for the bandwidth is one that directly minimizes (3.11) if the true distribution were known. This value is known as the AMISE optimal bandwidth for that distribution. For a Gaussian distribution with Gaussian kernel functions (estimator defined by Equations (3.6) and (3.7)), *Silverman* [1986] gives this bandwidth as:

$$\lambda = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \qquad (3.12)$$

In the univariate case (d=1) this reduces to $h = 1.06\,\hat{\sigma}\,n^{-1/5}$ where $\hat{\sigma}$ is an estimate of the standard deviation (Silverman advocates a robust estimate) of the data. An upper bound on bandwidth can be obtained by minimizing R(f'') over a class of probability densities. This leads to the optimal bandwidth for the smoothest possible density function. *Scott*

[1992] cites results showing that this upper bound ($1 \le d \le 10$)) is 1.08 to 1.12 times the λ in Equation (3.12).

Data-driven methods have been developed to estimate the bandwidth when the underlying distribution is not known. They minimize estimates of ISE, MISE, or AMISE formed only from the data. Least Squares Cross Validation (LSCV) [*Silverman*, 1986] is one such method based on the fact that the integrated square error (Equation 3.9) can be expanded as:

$$\mathrm{ISE} = R(\hat{f}(\mathbf{x})) - 2 \int \hat{f}(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} + R(f(\mathbf{x})) \qquad (3.13)$$

The first term may be directly evaluated. The second term may be recognized as $E[\hat{f}(\mathbf{X})]$ and estimated using leave-one-out cross validation. The last term, $R(f(\mathbf{x}))$, is independent of the bandwidth and does not need to be considered. The LSCV method in one dimension chooses bandwidth, h, to minimize the following LSCV score, comprising the first two terms in (3.13).

$$\mathrm{LSCV}(h) = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(2)}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{n} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{n h} K\left(\frac{x_i - x_j}{h}\right) \qquad (3.14)$$

Here, $K^{(2)}$ denotes the convolution of the kernel function with itself (for example, if K is the standard Gaussian kernel, then $K^{(2)}$ will be the Gaussian density with variance 2).

Sain et al. [1994] provide an expression for LSCV in any dimension with multivariate Gaussian kernel functions and **H** a diagonal matrix. *Adamowski and Feluch* [1991] provide a similar expression for the bivariate case with Gaussian kernels. Here we generalize these results for use with the multivariate density estimator (3.6) to:

$$\mathrm{LSCV}(\mathbf{H}) = \frac{1+\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\left[\exp(-L_{ij}/4) - 2^{d/2+1}\exp(-L_{ij}/2)\right]}{(2\sqrt{\pi})^{d}\, n \det(\mathbf{H})^{1/2}} \tag{3.15}$$

where

$$L_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T\, \mathbf{H}^{-1}\, (\mathbf{x}_i - \mathbf{x}_j) \tag{3.16}$$

We use numerical minimization of (3.15) over the single parameter λ with bandwidth matrix from Equation (3.7) to estimate all the necessary probability density functions. We recognize that LSCV bandwidth estimation is occasionally degenerate, so based on suggestions in *Silverman* [1986] and the upper bound given by *Scott* [1992] we restrict our search to the range $\lambda/4$ to 1.1λ.

Nonparametric Order p Markov Streamflow Model, NPp

To keep the presentation simple the equations will be presented for a lag 1 (order p=1) model. The formulae presented are readily extended to include higher order lags. Consideration of higher order models raises the issue of determination of the correct order p. This is deferred to future work. Here results are presented for the simplest case (NP1) analogous to the simple AR1 model. In the form presented below, the model can be applied to simulate stationary sequences such as annual flows. We later describe how application of the model to pairs of sequential months is used to simulate seasonally nonstationary (e.g., monthly) streamflow sequences.

The joint distribution of X_t and its prior value X_{t-1} is estimated using Equation (3.6) based on n observed data vectors $\mathbf{x}_i$. For a time series $x_0, x_1, x_2, \ldots x_n$, the data vector $\mathbf{x}_i$ has elements (x_i, x_{i-1}), $1 \le i \le n$. Hence $\mathbf{x}_1 = (x_1, x_0)$, $\mathbf{x}_2 = (x_2, x_1), \ldots \mathbf{x}_n = (x_n, x_{n-1})$. This is a series of ordered pairs. There is one less ordered pair than the length of the time series. The conditional density (Equation 3.1) is written as:

$$f(X_t | X_{t-1}) = \frac{f(X_t, X_{t-1})}{\int f(X_t, X_{t-1})\, dX_t} = \frac{f(X_t, X_{t-1})}{f_m(X_{t-1})} \tag{3.17}$$

where $f_m(X_{t-1})$ is the marginal density of X_{t-1}. Now applying the estimator in (3.6) the joint density estimate is obtained as:

$$\hat{f}(X_t, X_{t-1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi\lambda^2 \det(S)^{1/2}} \exp\left(-\frac{\begin{bmatrix} X_t - x_i \\ X_{t-1} - x_{i-1} \end{bmatrix}^T S^{-1} \begin{bmatrix} X_t - x_i \\ X_{t-1} - x_{i-1} \end{bmatrix}}{2\lambda^2} \right) \tag{3.18}$$

Note that each observation contributes to this density estimate depending on the distance of the observation (x_i, x_{i-1}) to the point (X_t, X_{t-1}), the bandwidth λ and the sample covariance matrix S of (X_t, X_{t-1}). The bandwidth λ is obtained by minimizing the LSCV score function (Equation 3.15).

Denote the terms in the covariance matrix:

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \tag{3.19}$$

Then for a given X_{t-1}, Equation (3.18) substituted in Equation (3.17) reduces to a sum of Gaussian kernels dependent on a single variable X_t:

$$\hat{f}(X_t|X_{t-1}) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda^2 S'}} w_i \exp\left(- \frac{(X_t - b_i)^2}{2\lambda^2 S'} \right) \tag{3.20}$$

where

$$w_i = \exp\left(- \frac{\left(X_{t-1} - x_{i-1}\right)^2}{2\lambda^2 S_{22}} \right) \Bigg/ \sum_{j=1}^{n} \exp\left(- \frac{\left(X_{t-1} - x_{j-1}\right)^2}{2\lambda^2 S_{22}} \right) \tag{3.21a}$$

$$S' = S_{11} - \frac{S_{12}^2}{S_{22}} \tag{3.21b}$$

$$b_i = x_i + \left(X_{t-1} - x_{i-1}\right) \frac{S_{12}}{S_{22}} \tag{3.21c}$$

This is illustrated in Figure 3-2. The conditional density is a slice through the bivariate density function, comprised of a sum of slices through the individual kernels that form the bivariate density estimate. Parameters b_i and $\lambda^2 S'$ give the center and spread of each kernel slice, respectively. The area under each kernel slice is the weight w_i, which controls the contribution of x_{i-1} to the conditional density estimate. Observations that lie close to the conditioning plane (i.e., where $(X_{t-1} - x_{i-1})$ is small) receive greater weight.

A time series realization is simulated by sampling X_t from (3.20), given a current value for X_{t-1}. The simulation then proceeds sequentially through time, updating X_{t-1} as the last sampled value. A flowchart describing the steps needed to simulate a sample of size n_r is provided in Figure 3-3. In practice we provide a "warm-up" period for the simulation scheme such that the length of the sequence simulated (n_r, refer to flowchart in Figure 3-3) is actually greater than the number desired. The first several (10 years in the results below) are discarded to account for the arbitrary initialization used.

Note that in the simulation scheme one does not need to explicitly estimate the conditional density in (3.20). This is avoided by treating it as a mixture of n kernel slices, each slice being selected with probability w_i. Once selected, X_t is simply a random variate from that kernel slice. Each slice is itself a Gaussian p.d.f. with mean b_i and variance $\lambda^2 S'$, so X_t is simulated using

$$X_t = b_i + \lambda\sqrt{S'}\, W_t \tag{3.22}$$

where W_t is N(0,1).

A complication can arise because W_t is unbounded and may result in negative X_t. The Gaussian kernels used in the kernel density estimate have infinite support and assign some (small) probability to regions of the domain where the streamflow is negative (i.e., invalid or out of bounds). This leakage of probability across boundaries is a problem when using kernel density estimates based on kernels with infinite support. It is also present in the parametric context where a Gaussian distribution, or any parametric distribution with support extending to negative values or beyond a lower or upper bound on the process, is used. Here we address the leakage by checking at each step whether the simulated flow values are positive. Whenever a negative X_t is encountered, we generate another sample from the same kernel slice, repeating this process until a positive X_t is obtained. This is achieved by simply generating a new W_t in Equation (3.22). This is equivalent to cutting the portion of each kernel that is out of bounds and renormalizing that kernel to have the appropriate mass. We record how often this is done as frequent boundary normalization is symptomatic of a substantial boundary leakage problem. Although the boundary renormalization procedure results in some bias in the simulated density in the neighborhood of the boundary, this was required for less than 1% of total realizations for the streamflow data sets on which the model was evaluated.

Simulations from this nonparametric approach retain the marginal and joint density structure of the historical time-series, including nonlinearities and state-dependence. One can also analytically calculate the marginal distribution and the values of the NP1 model mean, standard deviation, skewness and lag 1 correlation from the kernel density estimate (Equation 3.18). These are given in Appendix B and compared in the results below to sample statistics from the historical data.

This method has been presented from the perspective of formally estimating the underlying probability density function and then sampling from it. However, when viewed operationally one sees that simulation of each streamflow value effectively amounts to picking a prior data pair (x_i, x_{i-1}) that is nearby, i.e., x_{i-1} near to X_{t-1}. The probability of picking a particular pair falls off with distance according to Equation (3.21a). Then the value x_i is perturbed (Equations 3.21c and 3.22) by an amount related to the density estimate bandwidth. It is this perturbation that is responsible for the inflation of variance given by Equation (B.5). This method can therefore also be viewed as a smoothed bootstrap. The bootstrap [*Efron*, 1979; *Efron and Tibishirani*, 1993] is a statistical method that involves resampling the original data (with replacement) that has applications in estimation of confidence intervals and quantification of parameter uncertainty [*Tasker*, 1987; *Hardle and Bowman*, 1988; *Woo*, 1989; *Zucchini and Adamson*, 1989]. The classical bootstrap assumes data are independent and identically distributed. The nearest-neighbor bootstrap method presented by *Lall and Sharma* [1996] is appropriate for bootstrapping dependent data. It is similar to the approach here in that a prior data pair nearby is picked using a discrete kernel based on a local Poisson approximation to the density function. It differs in that there is no perturbation of the selected point. Consequently, it only reproduces streamflow values that have been observed. The perturbations in our approach effectively fill in the gaps.

Testing with Synthetic Data

In order to evaluate the ability of our model to recover structure from known linear and nonlinear parametric models, we conducted tests using two synthetic models. The purpose of these experiments was to verify the performance of the NP1 model where the true model is known. The first model used was a linear autoregressive order 1 (AR1) model of the type commonly used to model streamflow. The second was a Self-Exciting Threshold Autoregressive (SETAR) model [*Tong*, 1990]. The following procedure was used in both cases:

(1) Generate 100 sample records of length 80 from the true model (AR1 or SETAR).

(2) For each sample record, generate 100 realizations, each of length 80, from the NP1 model.

This provides 10,000 realizations (100 NP1 realizations from each of the 100 sample records) that were used to evaluate how well the NP1 model reproduces statistics of the samples it is based on and the underlying population statistics. Since all realizations are the same length as the record from which they are generated, the 100 realizations from each record provide estimates of the natural sampling variability associated with that record length. Statistics such as the mean, standard deviation, lag 1 correlation, and skewness were estimated, as well as marginal and joint kernel density estimates from the sample records and realizations.

Boxplots are used for the graphical comparisons. These consist of a box that extends over the interquartile range of the quantity being plotted, estimated from the 100 realizations. The line in the center of this box is the median and whiskers extend to the 5% and 95% quantiles of the compared statistic.

Tests with AR1 Data

The AR1 model used was:

$$X_t = 0.5\, X_{t-1} + 0.866\, W_t \tag{3.23}$$

where W_t was a Gaussian random variate with mean zero and standard deviation one. For brevity, comparisons for the standard statistics are not given. The mean, variance, lag 1 correlation, and skewness of each AR1 sample were well reproduced in the simulations based on it. These values were also close to the corresponding model statistic.

Figure 3-4 shows the marginal density estimates from one of the sample records and corresponding 100 NP1 simulations. Shown are the true Gaussian density function, the NP1 model marginal density function (from Equation B.1), and a univariate kernel density estimate based on the sample records, with the boxes giving the univariate kernel density estimates for the 100 NP1 simulations. To ensure that these univariate kernel density estimates are comparable, we used the same bandwidth for each of them, namely the median bandwidth from the set of bandwidths obtained by applying the LSCV procedure to each simulation. Figure 3-4 shows that the marginal density of the data is reproduced quite well by the simulations.

The Integrated Square Error (ISE, see Equation 3.9) of the joint density of each sample record provides a measure of model error. Averaging this across the 100 sample records, we get an estimate of MISE, which was 0.0093. The corresponding MISE from fitting an AR1 model joint density to each sample is 0.0046.

Bivariate density estimates were calculated using the procedure described earlier, for each of the 10,000 NP1 realizations. The ISE for each of these was calculated and the average is 0.0161, which is greater than the 0.0093 given above. This reflects the additional error introduced by reestimating the density function from simulated values. It is representative of the difference between NP1 simulations and the underlying model. This average ISE is still acceptably small. These results show that the NP1 model performs adequately at reproducing the properties of an AR1 process.

Tests with SETAR Data

The SETAR [*Tong*, 1990] model used was:

$$\begin{aligned} X_t &= 0.4 + 0.8\, X_{t-1} + W_t && \text{if } X_{t-1} \le 0.0 \\ X_t &= -1.5 - 0.5\, X_{t-1} + W_t && \text{if } X_{t-1} > 0.0 \end{aligned} \qquad (3.24)$$

where W_t was N(0,1). This is a state-dependent time series model with parameters that depend on the system state as determined by a threshold. This model may be representative of the monthly streamflow time series one could get from threshold-driven hydrologic processes such as snowmelt and evapotranspiration. It is worth noting that in one of the results given in the next section (see Figure 3-6b), the relationship between July and August streamflow is nonlinear in a manner similar to this SETAR model.

As with the AR1 case, mean, variance, lag 1 correlation, and skewness of the each SETAR sample were well reproduced in the simulations based on it. These values were also close to the corresponding model statistic.

Figure 3-5 shows the underlying true joint density $f(X_t, X_{t-1})$ for the SETAR model in (3.24), the bivariate kernel density estimate for one SETAR sample record, and the density estimate of the NP1 simulations averaged over all 10,000 realizations. The line in Figure 3-5a shows the true conditional mean from Equation (3.24) with W_t set to 0. Figure 3-5b shows an estimate of the conditional mean based on the sample record obtained using LOESS [*Cleveland and Devlin*, 1988]. LOESS is a locally weighted regression smoother that calculates a weighted least square fit (assigning weights using a tri-cubic weight function centered at the point of estimation) at each data point based on a fixed number of nearest neighbors. The number of nearest neighbors (expressed as a fraction of the total number of data points, called "span") used to compute the LOESS smooth was chosen as the one that resulted in an optimal value of Mallow's Cp. The function "loess," available in the software package Splus [*Chambers and Hastie*, 1992], was used in our calculations. The LOESS smooth is plotted to show that the sample record and nonparametric density function based on it reproduce the change in conditioning structure (with some smoothing) as the threshold is crossed. The illustrated fit based on an optimal Mallow's Cp has a span of 0.75. The kernel density estimate (Figure 3-5b) has an ISE (see Equation 9) of 0.0084 that was evaluated by integrating the squared differences between Figures 3-5a and b. Averaging across the 100 sample records, we obtain an estimate of the NP1 model fitting MISE as 0.0082. The corresponding MISE from fitting an AR1 joint density to each sample is 0.0131.

As for the AR1 example, bivariate density estimates were calculated using the procedure given earlier in this chapter, for each of the 10,000 realizations. The ISE for each of these was calculated and the average is 0.010, which is greater than the NP1 model MISE (0.0082) due to the additional error added by reestimating the density function from simulated values. It is again representative of the difference between NP1 simulations and the underlying model. Figure 3-5c shows the average density function estimated from the 10,000 realizations. This captures the essential nonlinearity of the SETAR model despite smoothing over the discontinuity. No model from the MGTM class of models is able to reproduce samples that exhibit such nonlinear structure. A bivariate Gaussian distribution with the true mean and covariance of the model in Equation (3.24), i.e., there are no fitting errors, has ISE relative to Figure 3-5a of 0.0119, larger than that obtained from the NP1 model fits with 80 data points. Asymptotically the NP1 kernel density estimate will converge to the underlying SETAR model exactly, i.e., with no fitting errors. This reiterates the point that model misspecification, for example, by selection of the bivariate Gaussian distribution, precludes a model from convergence to whatever the underlying distribution may be.

Table 3-1 shows the state-dependent correlation statistics (described in Appendix A) for NP1 model simulations based on SETAR data. Note how well the NP1 model reproduces the big difference between above median and forward and below median and forward correlations.

It is clear from the synthetic examples presented that the NP1 model is able to: (1) approximate the underlying joint distribution of the data; (2) reproduce the nonlinear structure suggested by the data in model simulations; and (3) approximate both linear and nonlinear dependence between the variables involved. No assumptions about marginal distributions or normalizing transforms are required.

Application of NP1 to Simulation of Monthly Streamflow

This section describes the application of the NP1 model to simulate seasonal streamflow sequences. Assume we have s seasons (or months, s = 12) and n years of data (n x s data values). The model applied to seasonal sequences then consists of s (one for each season) bivariate density functions estimated directly from the historical data. For all seasons except the first, the random vector (X_t, X_{t-1}) is replaced by $(X_{t,j}, X_{t,j-1})$, where subscript t denotes the year and j denotes the season. For the first season the conditioning flow is the flow in the last season of the previous year and the vector is $(X_{t,1}, X_{t-1,s})$. Simulations proceed sequentially from density estimates for one season pair to the next.

Results from an AR1 model representative of current hydrologic practice are also presented for comparison to NP1 simulations. SPIGOT, a synthetic streamflow generation software package developed by *Grygier and Stedinger* [1990], uses four choices for monthly marginal probability densities. These are Gaussian, two parameter Lognormal, three-parameter Lognormal, and approximate three-parameter Gamma distributions. The parameters for each distribution are estimated by matching moments and the best fitting distribution chosen by measuring the correlation of observations to the fitted distribution quantiles (Filliben correlation statistic, *Grygier and Stedinger* [1990]). Here we used the same procedure as SPIGOT to fit a marginal probability distribution and to obtain a normalizing transformation for each month. Then the AR1 model with seasonally varying coefficients given in Equation (3.2) was applied to the transformed monthly flows.

Both the NP1 and AR1 (with normalizing transformations) models were applied to an 83-year (1911 to 1993) record of monthly streamflow in the Snake River at Weiser, Idaho, located at 44° 14' 44" N and 116° 58' 48" W at an elevation of 2086 ft above Mean Sea Level, as recorded by the U.S. Geological Survey (USGS Station Number 13269000). This data set is one of many streamflow data sets with which we have tested the model, all with satisfactory results. We chose to present results from the Snake River because it illustrates well some of the points we want to emphasize.

Figure 3-6 shows the joint density for the March-April and July-August month pairs, using the kernel estimator described previously. The NP1 model simulates streamflow from these density functions. It is notable that the LOESS fit for the July-August joint density is highly nonlinear and resembles the conditional expectation of the SETAR example presented earlier in Figure 3-5b.

One hundred simulations, each with a length of 83 years (initialized with the average flow of the first month, with a warm-up period of 1 year), were made using both the NP1 and AR1 models. Results comparing the simulations from NP1 and AR1 are presented below.

Boxplots of selected monthly statistics are shown in Figures 3-7 to 3-10. The mean flows of the AR1 and NP1 simulations (Figure 3-7) match well those of the streamflow record. The annual means also match well. Figure 3-8 shows standard deviations of flows for the AR1 and NP1 simulations. The standard deviations of the NP1 simulations are slightly inflated with respect to the historical record, as expected from Equation (B.5). The standard deviations of AR1 simulations compare well, although some bias is visible in May simulations. Annual standard deviations, though not modeled directly by either approach, compare well with the historical value. Figure 3-9 shows boxplots of the correlation between sequential month pairs. The NP1 model reproduces lag 1 correlations without any bias as proved in Equation (B.7). The AR1 simulations approximate the lag 1 correlation well, although some bias is present depending on which transformation (or which marginal distribution) is used. Skewness is reproduced well in NP1 simulations (Figure 3-10) although a small downward bias (Equation B.12) is evident. AR1 model simulations sometimes have a higher skewness than observed, another indication of difficulty in fitting the marginal distributions.

The marginal distributions for each month were also compared. Selected marginal distributions are shown in Figure 3-11. In these figures, the model underlying density function (Equation B.1 in the case of NP1 or one of the SPIGOT [*Grygier and Stedinger*, 1990] densities in the AR1 case) is shown as a dashed line. The solid line is a univariate kernel density estimate applied to the original data and the boxes represent the range of univariate kernel density estimates applied to the 100 simulations. For these univariate kernel density estimates, the same bandwidth is used for all, chosen as the median of the set obtained by minimizing LSCV over the 100 simulations. Here the univariate kernel density estimator is being used as a plotting tool to compare observed and simulated data. The dots on the axis represent the historical data. These figures show that for some months the best fitting SPIGOT marginal distribution is inadequate. In particular, the Lognormal density used by the seasonal AR1 model does not compare well with the historical July streamflow data, and therefore simulations based on it do not match.

In addition to the Snake River example, it is instructive to see how the NP1 and AR1 models reproduce the bimodal marginal distributions of streamflow for the Beaver River data discussed earlier (see Figure 3-1). Figure 3-11c shows July marginal distributions for this station. Note how the NP1 model is able to reproduce the bimodality, whereas the fitted three-parameter Gamma distribution does not. Overall we find that the common normalizing transformations are unable to capture a lot of the structure, in particular bimodality, sometimes present in data. This structure is captured by the kernel density estimates.

Recall that the joint density of July-August flows (Figure 3-6b) indicated a highly nonlinear conditional expectation. The associated lag 1 correlation (Figure 3-9) is much lower than for other months. Looking at Figure 3-6b, one sees that the subset of the data with July Streamflow less than 12,000 cfs exhibits a strong positive correlation, whereas above this value the data appears to be weakly negatively correlated. It is possible that such flows are state-dependent. To quantify the dependence of autocorrelation on the magnitude of flow, we split each series into flows above and below the median and then calculated the state-dependent correlation statistic described in Appendix A. These results are illustrated in Figure 3-12. The historical data (solid line) has significant differences (at the 95% level by a hypothesis test for equality of two sample correlations, see Appendix A for details) between forward above and below median correlations for the following five month pairs: Oct-Nov, Dec-Jan, Feb-Mar, Jul-Aug, Aug-Sep. Reproduction of these statistics requires a state-dependent or nonlinear model. Correlations of streamflow above and below median are modeled effectively by the NP1 approach. Simulations from the AR1 approach are unable to accommodate such nonlinear dependence. Particularly notable is the inability of AR1 simulations to model above or below median correlations for July (correlation between above or below median flows in July and succeeding August flows). Obviously the transformations to Gaussian from the set of marginal distributions used in the AR1 model are inadequate in dealing with the state-dependent July-August month pair flows.

The practical use of synthetic streamflow simulations is often the evaluation of the storage capacity of reservoirs required to support a certain yield. For a given streamflow sequence (observed or simulated) the storage required to support a specified yield can be obtained using the sequent peak algorithm [*Loucks et al.*, 1981]. *Vogel and Stedinger* [1988] compared the Root Mean Square Error (RMSE) and bias of this storage statistic computed directly from data and showed the improvements in precision that result from using stochastic streamflow models. Here reservoir storages required to support a firm yield that is 80% of the mean annual flow were estimated for both AR1 and NP1 simulations of Snake River streamflow. Monthly demand fractions given in *Lall and Miller* [1988] were used. Standardized bias and RMSE estimates for both models, relative to the storage required to support a given yield for the historical record, are given in Table 3-2.

$$\text{Bias} / S_h = (S_h - \frac{1}{n_r}\sum_{i=1}^{n_r} S_{s_i})/S_h \tag{3.25}$$

$$\text{RMSE} / S_h = \frac{\sqrt{\frac{1}{n_r}\sum_{i=1}^{n_r}(S_h - S_{s_i})^2}}{S_h} \tag{3.26}$$

where S_h denotes the historical storage, S_{s_i} is the storage estimated from the i'th AR1 or NP1 realization and n_r is the number of realizations. As one can see from Table 3-2, the NP1 bias and RMSE are lower than those for AR1 simulations. The NP1 model is better at providing simulations with storage properties comparable to the historical data.

Discussion and Conclusions

We also computed and checked many other statistical attributes of the NP1 and AR1 simulations, but space limitations prevent presentation of the results. The simulated autocorrelation function (acf) for each month (not shown) showed that both models do not model correlations higher than lag 1 very well. In some months, lag 2 and 3 correlation was preserved though both NP1 and AR1 model correlations decrease to essentially zero by lag 7. Longer range dependence quantified in terms of the annual correlation coefficient or Hurst coefficient [*Hurst*, 1951] was not preserved by either model. Although the bias and errors in the reservoir storages given in Table 3-2 are smaller for the NP1 simulations than the AR1 simulations, they are still relatively large, indicating that the order one dependence assumed in both models is inadequate to model reservoir storage for these data. These all indicate the need for models that capture higher order dependence, such as multivariate or disaggregation models.

The results presented here support the nonparametric approach as a feasible alternative to parametric approaches used to model streamflow. The nonparametric approach presented here is consistent and robust and reproduces not only the linear statistics modeled by the AR1 model but also a broader set of properties based on additional distributional information. The skewness, bimodality, and dependence of correlations on the flow magnitude, when present in the data, can be adequately modeled. One could no doubt find better marginal distributions to use with the AR1 model and improve on some of the AR1 simulations. However, the NP1 approach is effective in sidestepping these difficult model and distribution selection issues that are often somewhat arbitrarily resolved and provides a method that is easy to use, and adapts well to the data.

Although the examples presented here used an order one dependence structure, it is easy to extend the model to higher order dependence. Cross validatory procedures [*Eubank*, 1988] can be applied to evaluate the benefit gleaned from including additional lags in the model dependence structure. These are somewhat analogous to use of Akiake's information criterion in linear models. We intend to evaluate this further in future work. Future work will also apply the nonparametric approach to multivariate problems in stochastic hydrology, specifically to the development of nonparametric analogs to multivariate ARMA and disaggregation models. The purpose here was to introduce this approach in a simple univariate setting with order one dependence and show that results are satisfactory when compared to current hydrologic practice.

We are convinced that nonparametric techniques have an important role to play in improving the synthesis of hydrologic time series for water resources planning and management. They can capture the dependence structure present in the historical data, without imposing arbitrary linearity or distributional assumptions. They have the capability to reproduce nonlinearity, state-dependence, and multimodality while remaining faithful to the historical data and producing synthesized sequences statistically indistinguishable from the historical sequence.

References

Adamowski, K., and W. Feluch, Application of nonparametric regression to groundwater level prediction, *Can. J. Civ. Eng.*, 18, 600-606, 1991.

Beard, L. R., *Monthly Streamflow Simulation*, Hydrologic Engineering Center, U.S. Corps of Engineers, Washington, D.C., 1967.

Bendat, J. S., and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, 2nd ed., John Wiley & Sons, New York, 1986.

Benjamin, J. R., and C. A. Cornell, *Probability, Statistics, and Decision for Civil Engineers*, McGraw-Hill, New York, 1970.

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, Mass., 1985.

Chambers, J. M., and T. J. Hastie, *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove, Calif., 1992.

Cleveland, W. S., and S. J. Devlin, Locally weighted regression : An approach to regression by local fitting, *J. Am. Stat. Assoc.*, 83(403), 596-610, 1988.

Efron, B., Bootstrap methods: Another look at the jackknife, *Ann. Stat.*, 7, 1-26, 1979.

Efron, B., and R. Tibishirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

Eubank, R. L., *Spline Smoothing and Non-Parametric Regression,* Marcel-Dekker, New York, 1988.

Fiering, M. B., *Streamflow Synthesis,* Harvard University Press, Cambridge, Mass., 1967.

Fukunaga, K., *Introduction to Statistical Pattern Recognition,* Academic Press, New York, 1972.

Grygier, J. C., and J. R. Stedinger, *Spigot, A Synthetic Streamflow Generation Package, Technical Description, Version 2.5,* School of Civil and Environmental Engineering, Cornell University, Ithaca, N.Y., 1990.

Hardle, W., and A. W. Bowman, Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands, *J. Am. Stat. Assoc.*, 83(401), 102-110, 1988.

Hipel, K. W., A. I. Mcleod, and W. C. Lennox, Advances in Box-Jenkins modeling, 1: Model construction, *Water Resour. Res.*, 13(3), 567-575, 1977.

Hurst, H. E., Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng.*, 116, 770-799, 1951.

Kendall, D. R., and J. A. Dracup, A comparison of index-sequential and AR(1) generated hydrologic sequences, *J. Hydrol.*, 122, 335-352, 1991.

Kite, G. W., *Frequency and Risk Analysis in Hydrology,* Water Resources Publications, Fort Collins, Colo., 1977.

Lall, U., Nonparametric function estimation: Recent hydrologic applications, U.S. National Report to International Union of Geodesy and Geophysics, 1991-1994, *Rev. Geophys,* 33, 1093, 1995.

Lall, U., and C. W. Miller, An optimization model for screening multipurpose reservoir systems, *Water Resour. Res.*, 24(7), 953-968, 1988.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for time series resampling, *Water Resour. Res.*, 32(3), 679-693, 1996.

Lettenmaier, D. P., and S. J. Burges, An operational approach to preserving skew in hydrologic models of long-term persistence, *Water Resour. Res.*, 13(2), 281-290, 1977.

Loucks, D. P., J. R. Stedinger, and D. A. Haith, *Water Resources Systems Planning and Analysis,* Prentice-Hall, Englewood Cliffs, N.J., 1981.

McLeod, A. I., K. W. Hipel, and W. C. Lennox, Advances in Box-Jenkins modeling, 2: Applications, *Water Resour. Res.*, 13(3), 577-585, 1977.

Pegram, G. G. S., J. D. Salas, D. C. Boes, and V. Yevjevich, *Stochastic Properties of Water Storage*, Colorado State University, Fort Collins, Colo., 1980.

Sain, S. R., K. A. Baggerly, and D. W. Scott, Cross-validation of multivariate densities, *J. Am. Stat. Assoc.*, 89(427), 807-817, 1994.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, Water Resources Publications, Littleton, Colo., 1980.

Salas, J. D., and R. A. Smith, Physical basis of stochastic models of annual flows, *Water Resour. Res.*, 17(2), 428-430, 1981.

Scott, D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, 1992.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, England, 1986.

Smith, J. A., Long-range streamflow forecasting using nonparametric regression, *Water Resour. Bull.*, 27(1), 39-46, 1991.

Smith, L. A., Identification and prediction of low dimensional dynamics, *Phys. D*, 58, 50-76, 1992.

Stedinger, J. R., Estimating correlations in multivariate streamflow models, *Water Resour. Res.*, 17(1), 200-208, 1981.

Stedinger, J. R., D. P. Lettenmaier, and R. M. Vogel, Multisite ARMA(1,1) and disaggregation models for annual streamflow generation, *Water Resour. Res.*, 21(4), 497-509, 1985a.

Stedinger, J. R., D. Pei, and T. A. Cohn, A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res.*, 21(5), 665-675, 1985b.

Stedinger, J. R., and M. R. Taylor, Synthetic streamflow generation, 1: Model verification and validation, *Water Resour. Res.*, 18(4), 909-918, 1982.

Stedinger, J. R., and R. M. Vogel, Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, 20(11), 47-56, 1984.

Tasker, G. D., Comparison of methods for estimating low flow characteristics of streams, *Water Resour. Bull.*, 23(6), 1077-1083, 1987.

Thomas, H. A., and M. B. Fiering, Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, in *Design of Water Resource Systems*, edited by A. Maass, M. M. Hufschmidt, R. Dorfman, H. A. Thomas, S. A. Marglin, and G. M. Fair, pp. 459-493, Harvard University Press, Cambridge, Mass., 1962.

Todini, J., The preservation of skewness in linear disaggregation schemes, *J. Hydrol.*, 47(199), 1980.

Tong, H., *Nonlinear Time Series Analysis: A Dynamical Systems Perspective*, Academic Press, London, England, 1990.

Vogel, R. M., and J. R. Stedinger, The value of stochastic streamflow models in overyear reservoir design applications, *Water Resour. Res.*, 24(9), 1483-1490, 1988.

Wand, M. P., and M. C. Jones, Multivariate plug-in bandwidth selection, *Comp. Stat.*, 9, 97-116, 1994.

Woo, M. K., Confidence intervals of optimal risk-based hydraulic design parameters, *Can. Water Resour. J.*, 14(1), 10-16, 1989.

Yakowitz, S., Nonparametric density estimation, prediction, and regression for Markov sequences, *J. Am. Stat. Assoc.*, 80(389), 215-221, 1985.

Yevjevich, V. M., *Stochastic Processes in Hydrology*, Water Resources Publications, Fort Collins, Colo., 1972.

Zucchini, W., and P. T. Adamson, Bootstrap confidence intervals for design storms from exceedance series, *Hydrol. Sci. J.*, 34(1), 41-48, 1989.

Table 3-1. State-Dependent Lag 1 Correlations for NP1 Simulations from a SETAR Model

Single sample record

Statistic	Record	NP1 Simulation Statistics		
		5 % Quantile	Median	95 % Quantile
r	0.164	-0.094	0.160	0.329
r_{af} (above and forward)	-0.528	-0.595	-0.373	-0.202
r_{bf} (below and forward)	0.504	0.167	0.425	0.620
r_{ab} (above and back)	0.172	-0.056	0.161	0.380
r_{bb} (below and back)	0.310	-0.215	0.239	0.490

Average over 100 sample records and 10,000 realizations

Statistic	100 Records	NP1 Simulation Statistics		
		5 % Quantile	Median	95 % Quantile
r	0.074	-0.105	0.041	0.226
r_{af} (above and forward)	-0.555	-0.557	-0.396	-0.164
r_{bf} (below and forward)	0.494	0.187	0.377	0.558
r_{ab} (above and back)	0.165	0.000	0.131	0.305
r_{bb} (below and back)	-0.020	-0.219	-0.039	0.169

Table 3-2. Reservoir Capacities Evaluated for 100 AR1 and NP1 Model Realizations for a Yield of 0.8 Mean Annual Flow

Model	Bias/S_h	RMSE/S_h
AR1	0.4075	0.4295
NP1	0.2619	0.3282

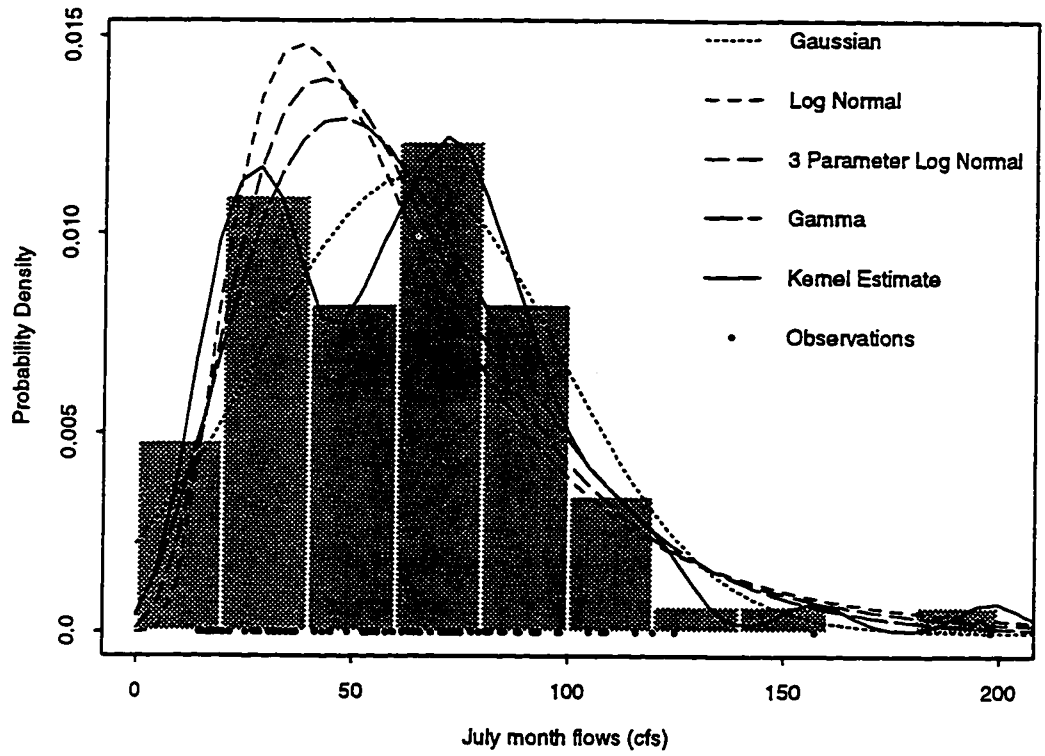


Figure 3-1. Histogram and probability density estimates of July monthly streamflow (cfs) in the Beaver River at Beaver, UT (USGS station number 10234500). The dots on the x-axis denote the individual data points.

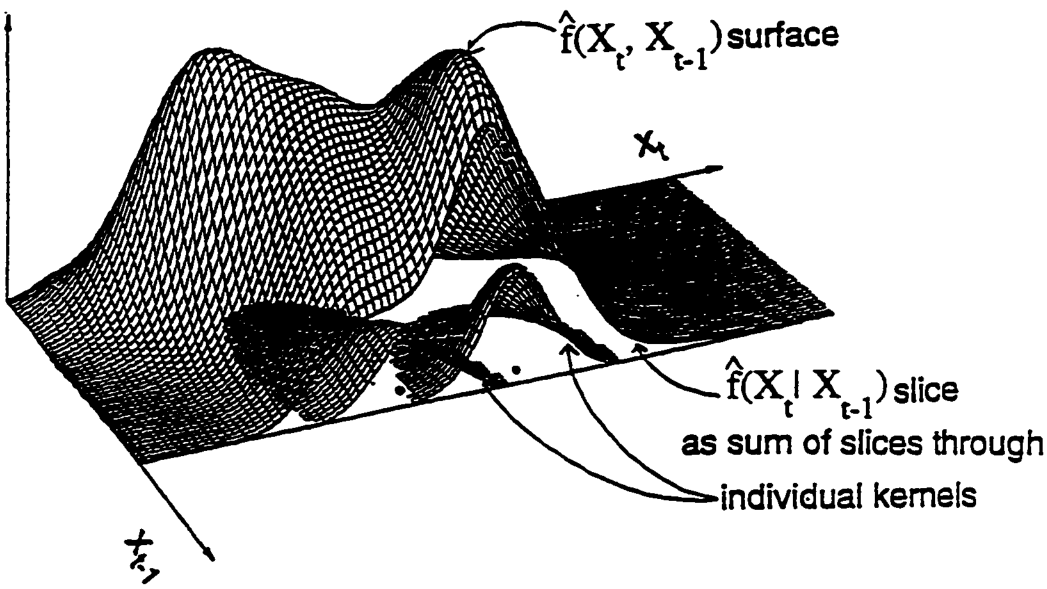


Figure 3-2. Illustration of conditional probability density function.

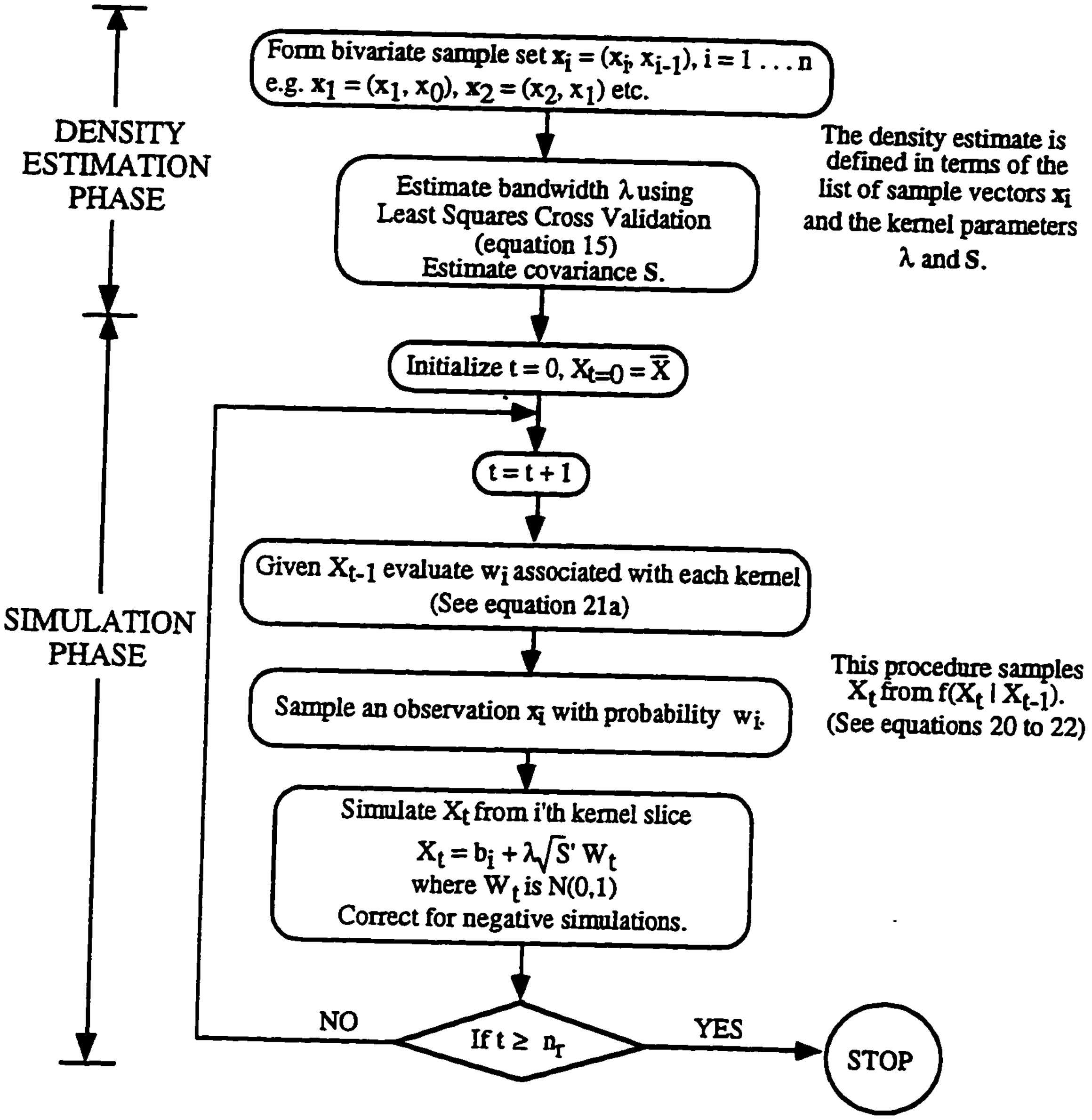

DENSITY ESTIMATION PHASE
SIMULATION PHASE
Form bivariate sample set $\mathbf{x}_i = (x_i, x_{i-1})$, $i = 1 \ldots n$
e.g. $\mathbf{x}_1 = (x_1, x_0)$, $\mathbf{x}_2 = (x_2, x_1)$ etc.
Estimate bandwidth λ using Least Squares Cross Validation (equation 15) Estimate covariance S.
The density estimate is defined in terms of the list of sample vectors $\mathbf{x}_i$ and the kernel parameters λ and S.
Initialize $t = 0$, $X_{t=0} = \overline{X}$
$t = t + 1$
Given X_{t-1} evaluate w_i associated with each kernel (See equation 21a)
Sample an observation $\mathbf{x}_i$ with probability w_i.
This procedure samples X_t from $f(X_t \mid X_{t-1})$. (See equations 20 to 22)
Simulate X_t from i'th kernel slice
$X_t = b_i + \lambda\sqrt{S'}\,W_t$
where W_t is N(0,1)
Correct for negative simulations.
NO
If $t \geq n_T$
YES
STOP

Figure 3-3. Flowchart of NP1 model.

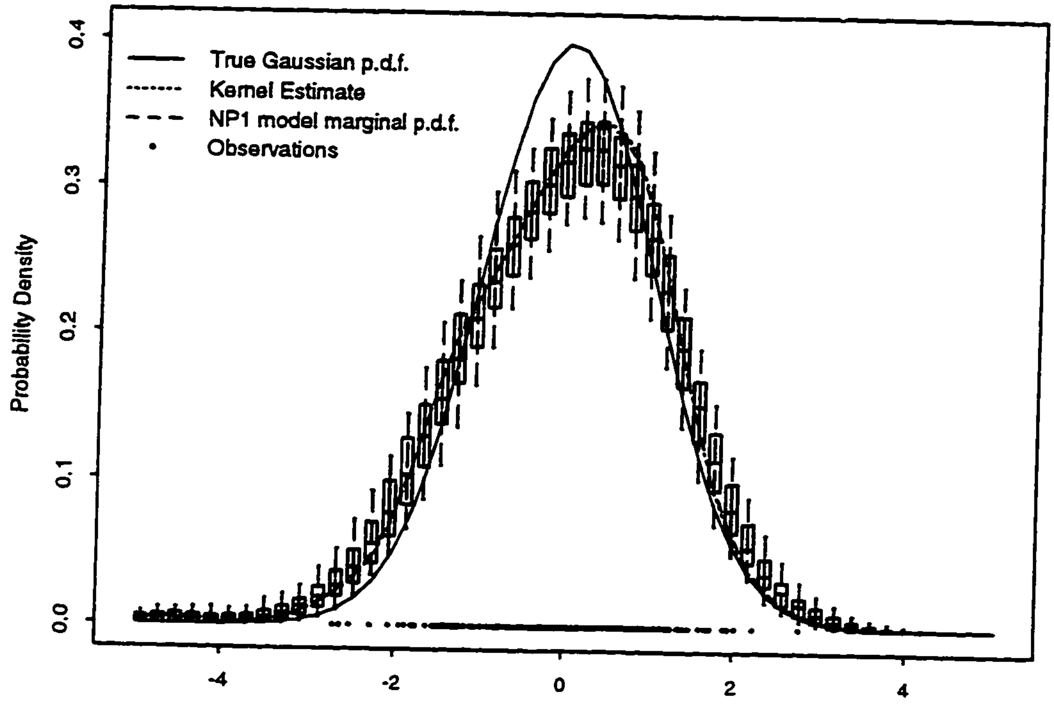
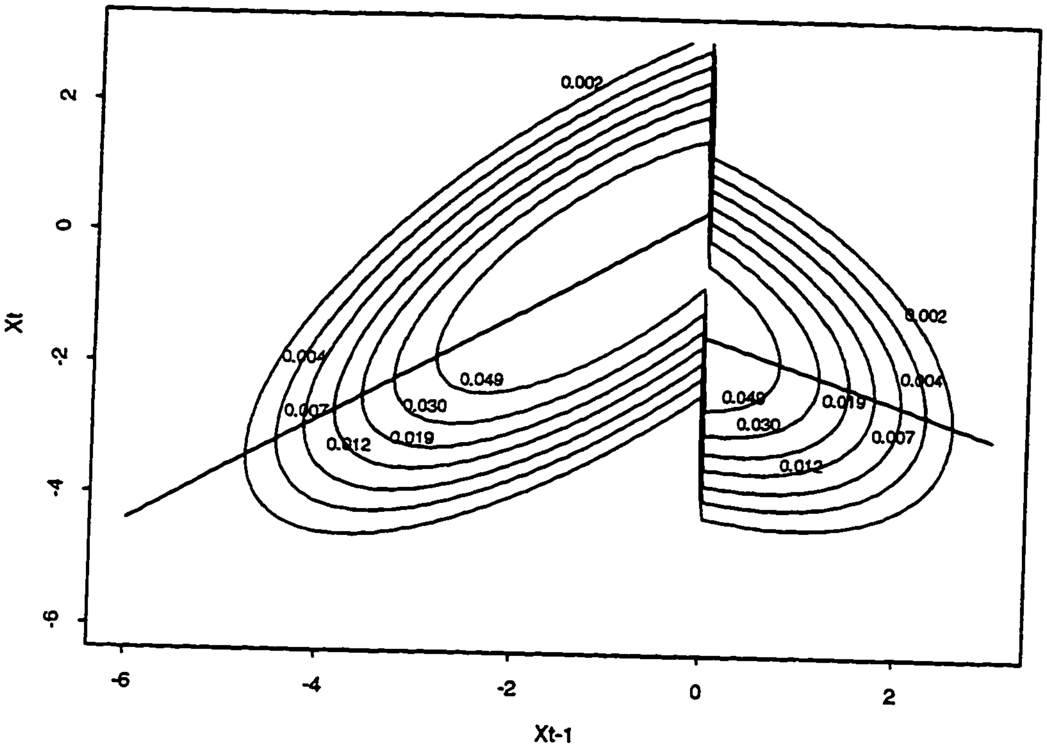
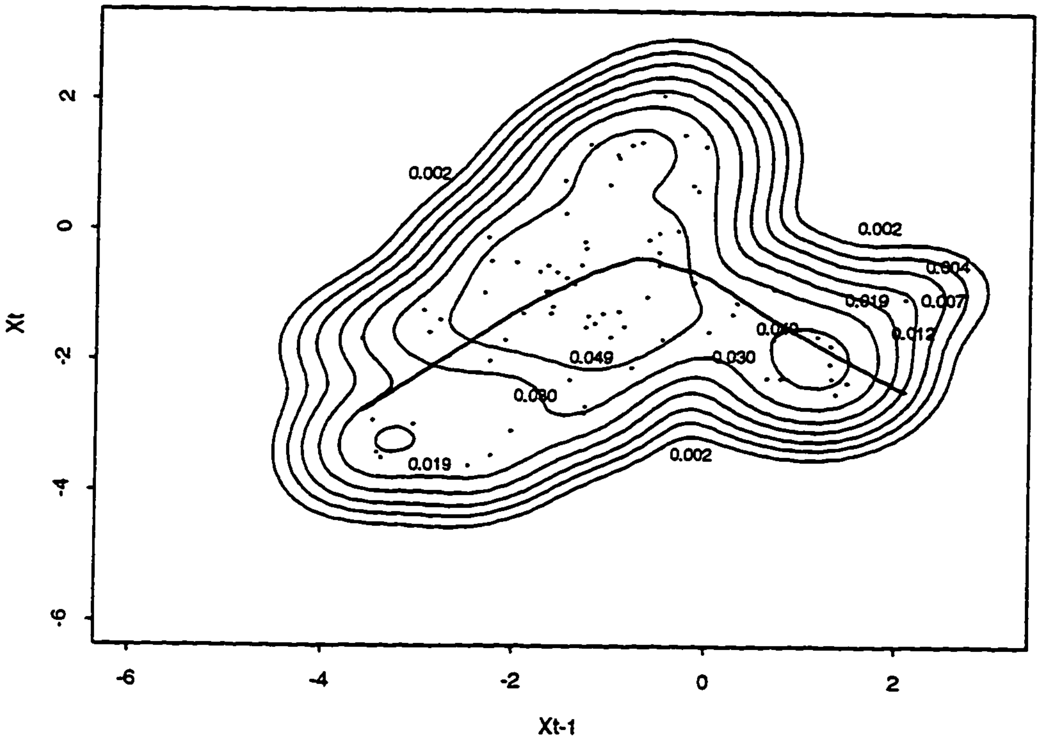


Figure 3-4. Marginal density estimates for NP1 simulations of an AR1 data set. The NP1 underlying density is estimated using (B.1).



(a)

Figure 3-5. Bivariate joint density of: (a) SETAR model in (3.5). The straight lines denote the conditional mean of the model. (b) SETAR model sample. Dots (.) represent individual observations in SETAR sample. The line is a LOESS smooth through the data. (c) NP1 simulations. This is the average of the kernel density estimates from all 10,000 realizations.



(b)

Figure 3-5. Continued.

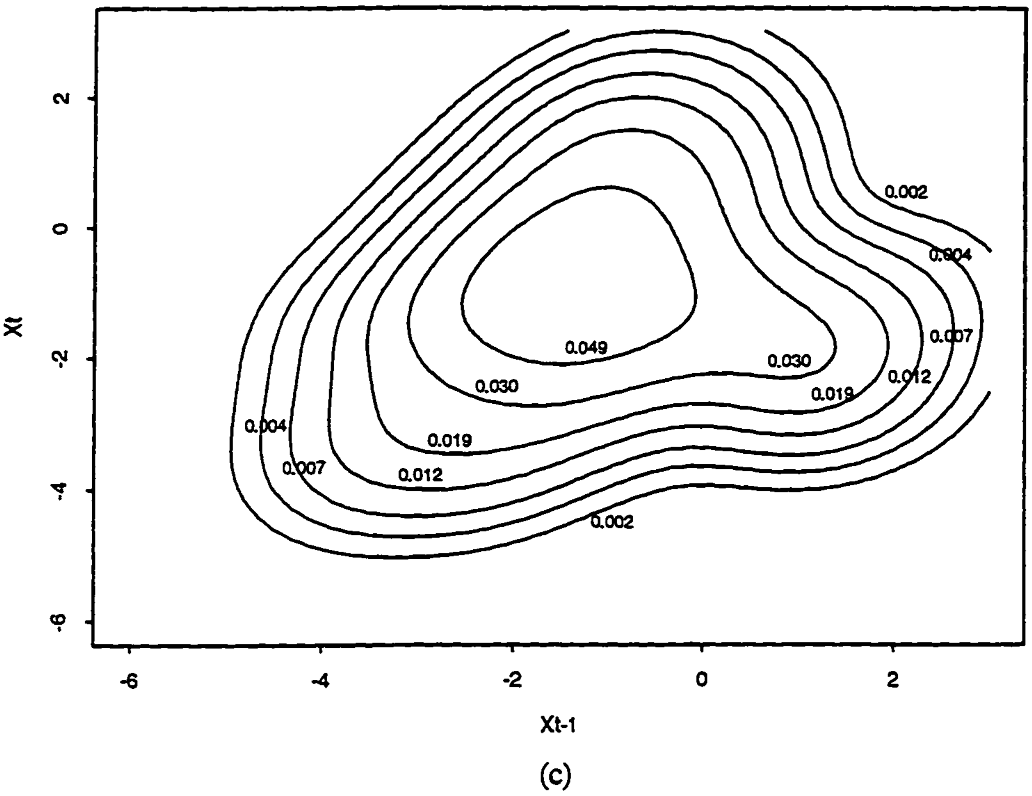
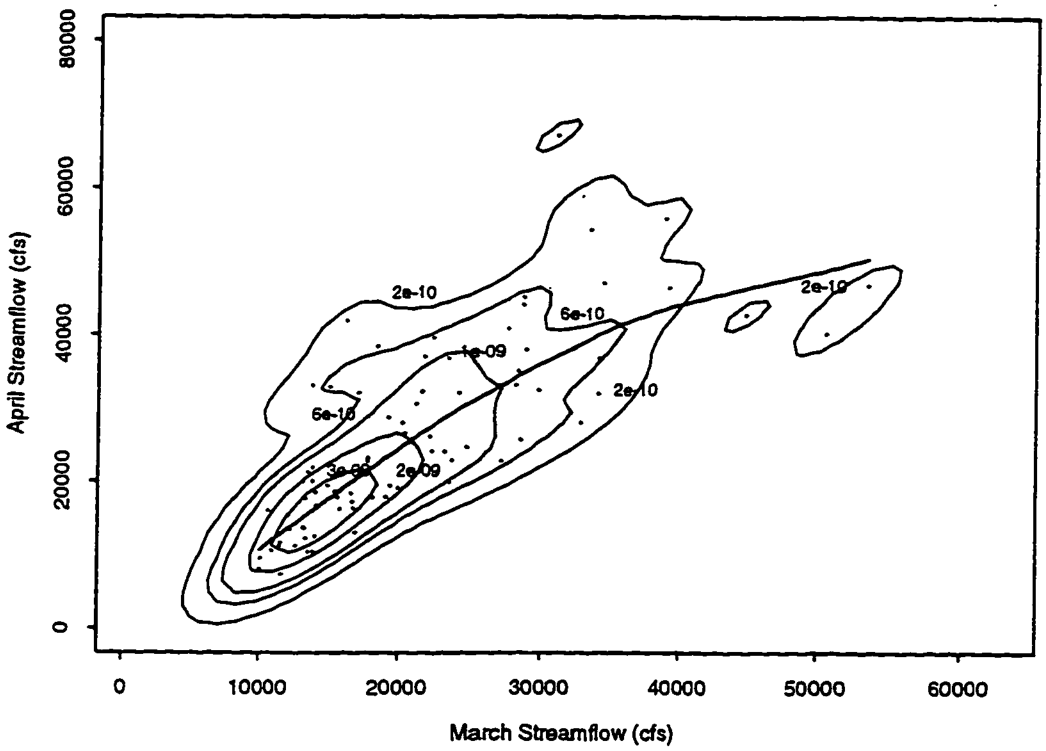
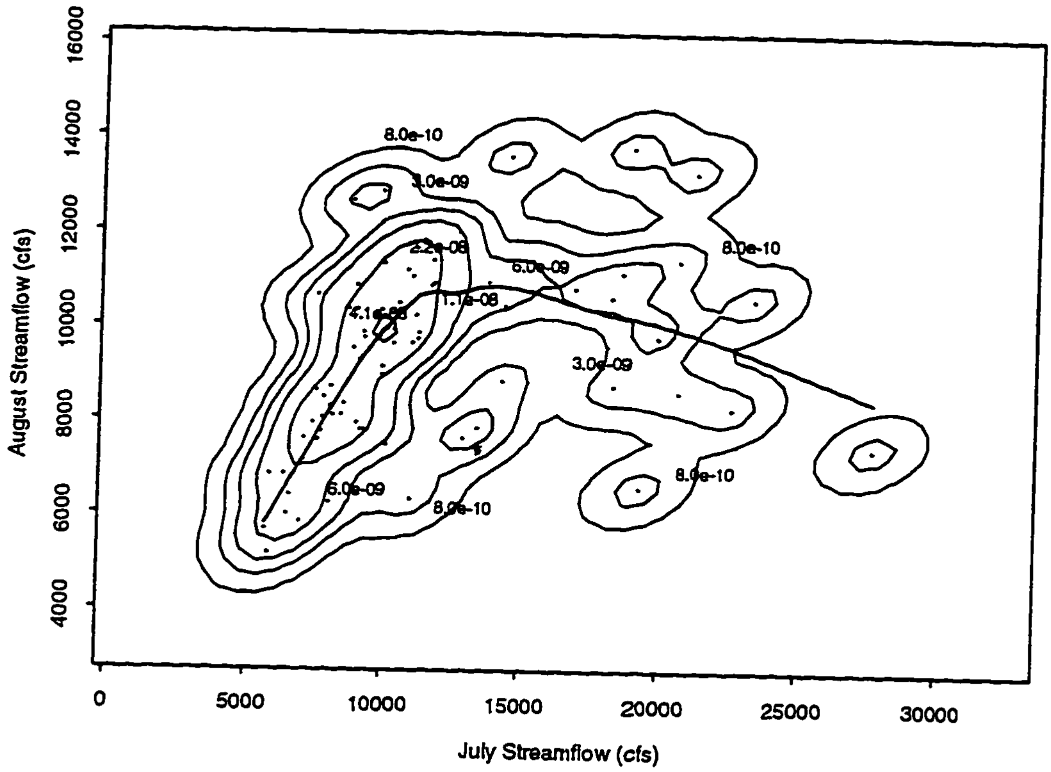


Figure 3-5. Continued.



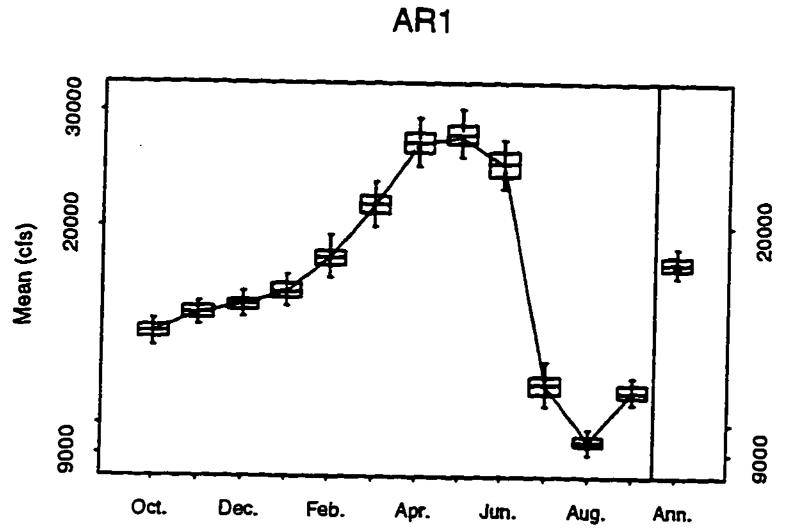
(a)

Figure 3-6. Underlying bivariate densities in NP1 model simulations for selected month pairs. The dots (.) represent observations. (a) March-April flows. The line represents a LOESS smooth with a span of 0.95 corresponding to an optimal Mallow's Cp. (b) July-August flows. The line represents a LOESS smooth with a span of 0.7 corresponding to an optimal Mallow's Cp.



(b)

Figure 3-6. Continued.



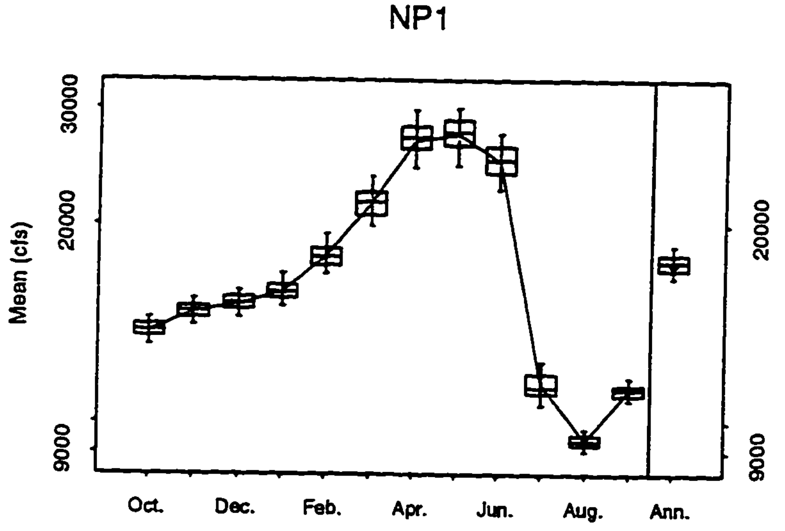


Figure 3-7. Comparison of means of simulated and historical flows. The continuous line represents monthly means of the historical record. The dot in the right panel is the observed annual mean.

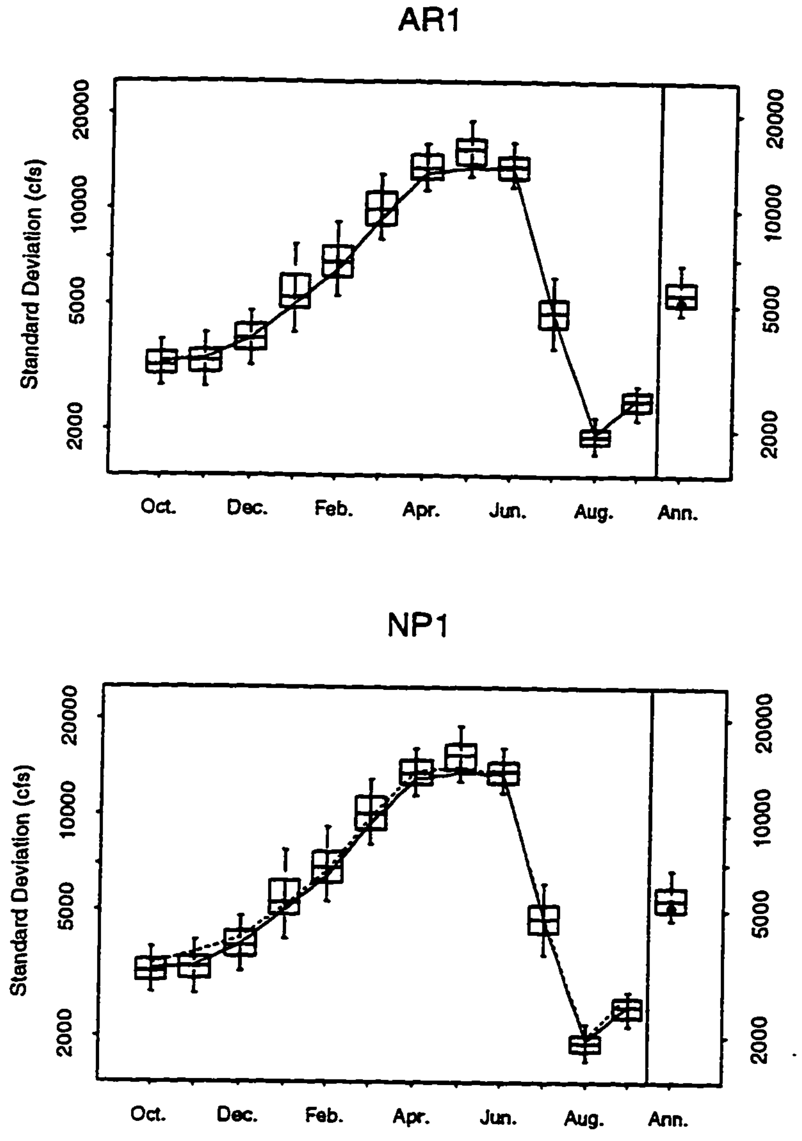


Figure 3-8. Comparison of standard deviations of simulated and historical flows. The continuous line represents monthly standard deviations of the historical record. The dot in the right panel is the observed annual standard deviation. The dashed line in the NP1 figure shows the NP1 model standard deviations (Equation B.5).

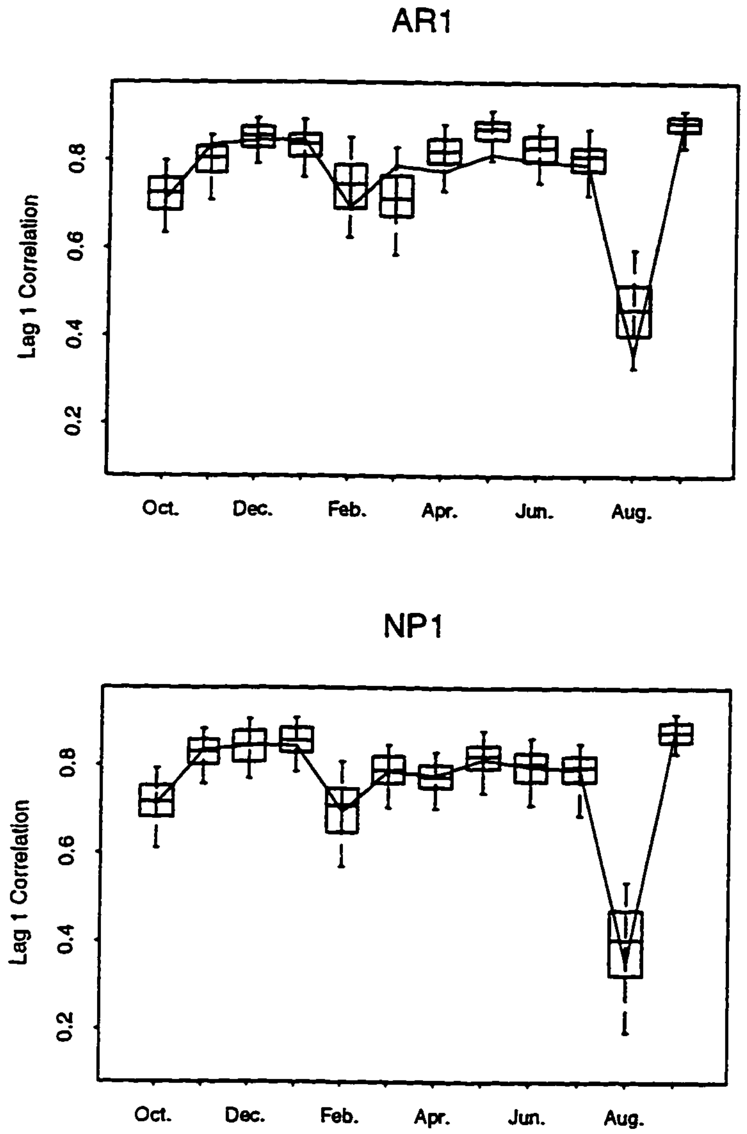


Figure 3-9. Comparison of lag 1 correlations of simulated and historical flows. The continuous line represents lag 1 correlations in the historical record.

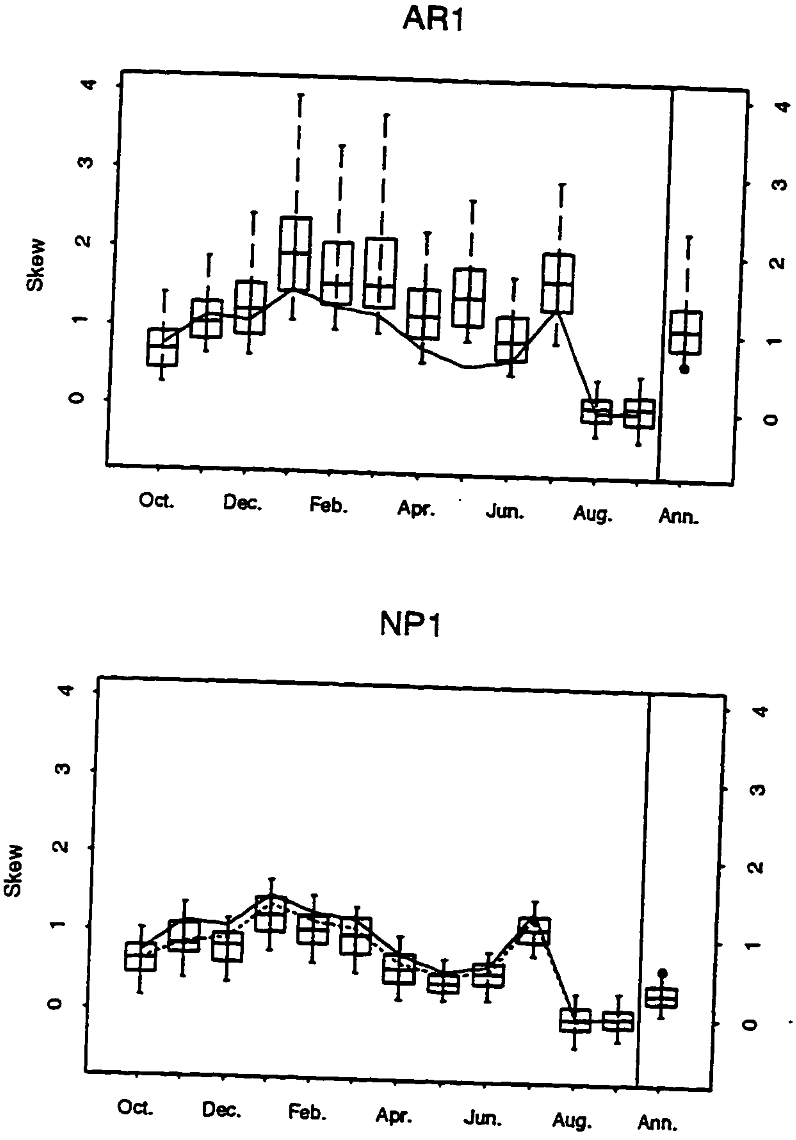
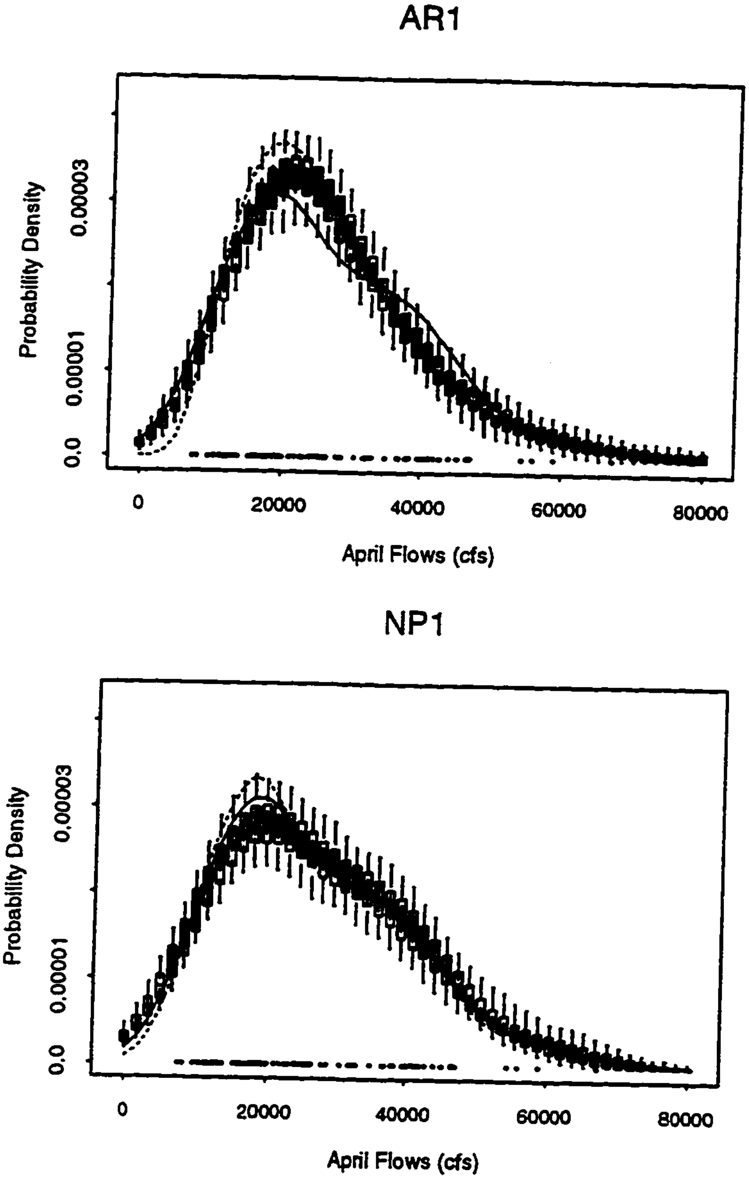
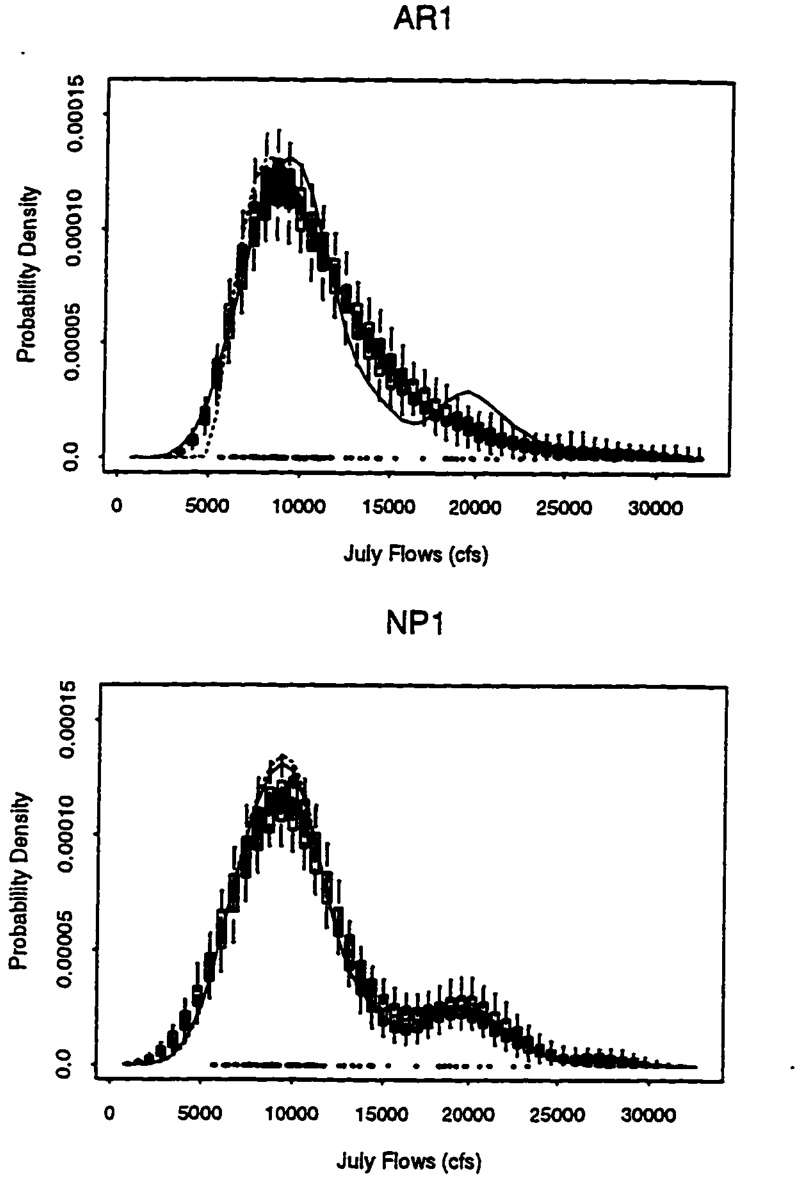


Figure 3-10. Comparison of skewness of simulated and historical flows. The continuous line represents the monthly skewness of the historical record. The dot in the right panel is observed skewness of the annual flows. The dashed line (in NP1 result) shows model skewness (Equation B.12).



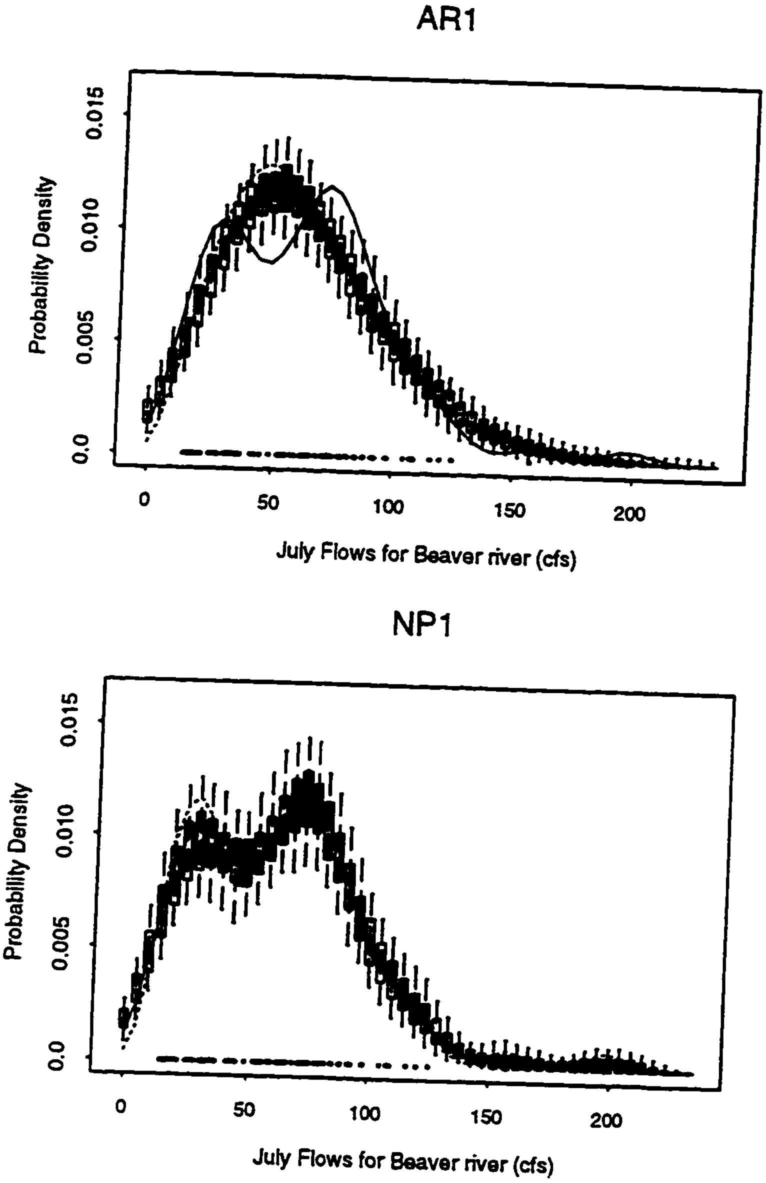
(a)

Figure 3-11. Marginal density estimates for selected months streamflow. The solid line is a univariate kernel density estimate applied to the original data. The model underlying density function (Equation B.1 in the case of NP1 or one of the SPIGOT densities in the AR1 case) is shown as a dashed line. The boxes depict the range of univariate kernel density estimates applied to the 100 simulations. (a) April AR1 and NP1 marginal density estimates. A three-parameter Lognormal distribution is used for the AR1 model fit. (b) July AR1 and NP1 marginal density estimates. A three-parameter Lognormal distribution is used for the AR1 model fit. (c) July AR1 and NP1 marginal density estimates from the Beaver River at Beaver, Utah (USGS Station Number 10234500). A three-parameter Gamma distribution is used in the AR1 model fit.



(b)

Figure 3-11. Continued.



(c)

Figure 3-11. Continued.

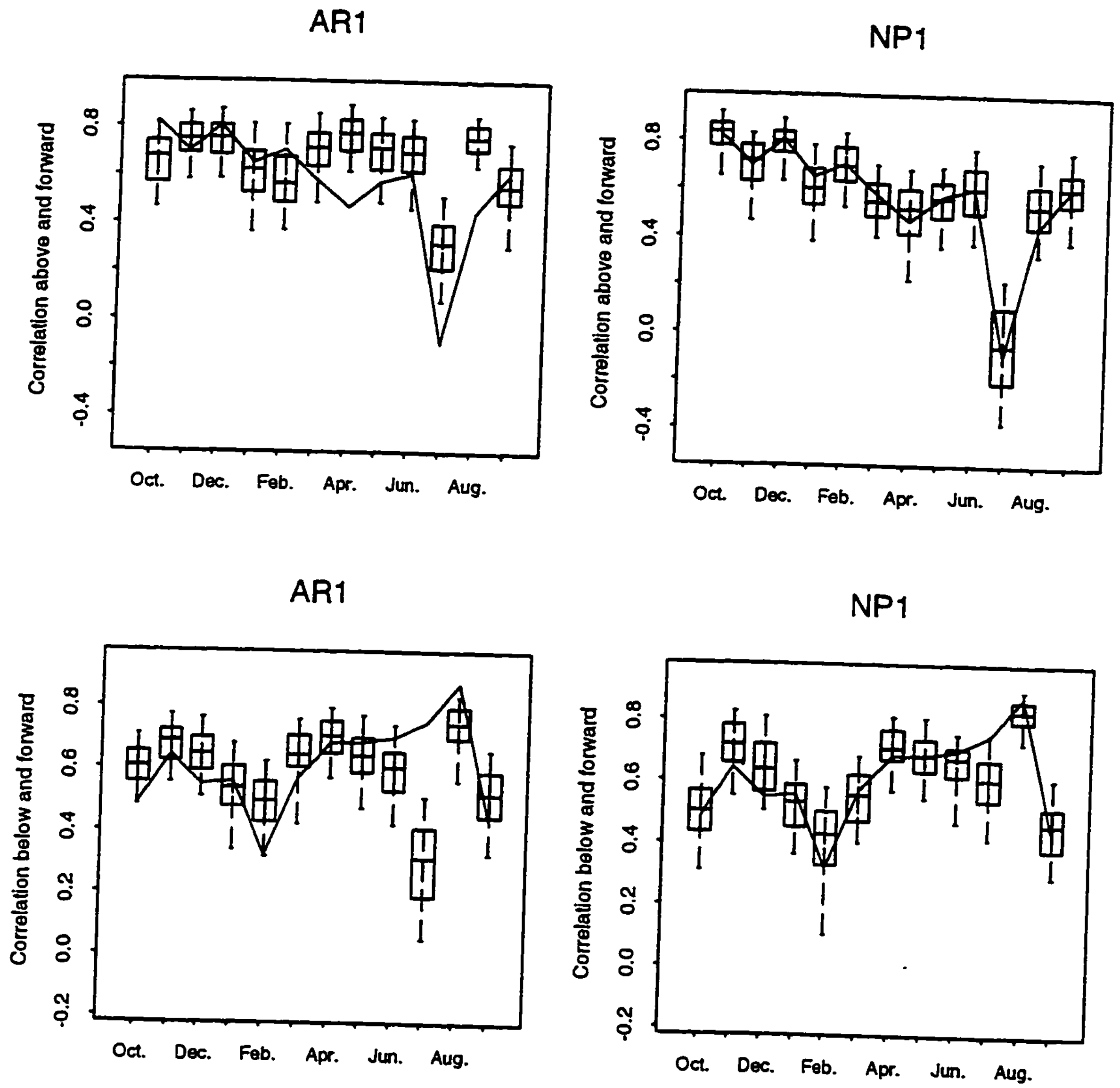


Figure 3-12. Monthly state-dependent correlations for simulated and historical flows. Continuous lines represent respective correlations in the historical record.

CHAPTER 4

DISAGGREGATION PROCEDURES FOR STOCHASTIC HYDROLOGY BASED ON NONPARAMETRIC DENSITY ESTIMATION[1]

Abstract

Synthetic simulation of streamflow sequences is important for the analysis of water supply reliability. Disaggregation models are an important component of the stochastic streamflow generation methodology. They provide the ability to generate multiseason and multisite streamflow sequences that maintain the proper interdependence between the aggregate and disaggregate flows, thus allowing a model to span scales in time or space. In recent papers we have suggested the use of nonparametric methods for streamflow simulation. These methods provide the capability to synthesize time series dependence without a priori assumptions as to the probability distribution of streamflow. They remain faithful to the data and can approximate linear or nonlinear dependence. In this chapter we extend the use of nonparametric methods to disaggregation models. We show how a kernel density estimate of the joint distribution of disaggregate flow variables can form the basis for conditional simulation based on an input aggregate flow variable. This methodology preserves summability of the disaggregate flows to the input aggregate flow. We show through applications to synthetic data and streamflow from the Snake River how this conditional simulation procedure preserves a variety of statistical attributes.

---

[1] Coauthored by Ashish Sharna, David G. Tarboton, and Upmanu Lall.

Introduction

A goal of stochastic hydrology is to generate synthetic streamflow sequences that are statistically similar to observed streamflow records. Such synthetic sequences are needed to analyze alternative designs and policies against a range of sequences that are likely to occur in the future. In studies involving large water resources systems it is often necessary to correctly represent the variation of streamflow across various sites and across the different seasons of the year. This becomes especially important where there is over year storage or sharing and transfers between different sites. A methodology that models the aggregate (annual or main stream) flow and the relation with its tributary or seasonal components becomes important. Disaggregation is one such methodology that preserves the dependence within the disaggregate flow components as well as their relation with the aggregate flows. The motivation behind disaggregation is the desire to parsimoniously represent processes at both the aggregate and disaggregate scales.

Disaggregation was first introduced by *Valencia and Schaake* [1972] and further developed by many others [*Mejia and Rousselle*, 1976; *Curry and Bras*, 1978; *Lane*, 1979; *Salas et al.*, 1980; *Svanidze*, 1980; *Stedinger and Vogel*, 1984; *Bras and Rodriguez-Iturbe*, 1985; *Stedinger et al.*, 1985; *Grygier and Stedinger*, 1988]. An appropriate model is first used to simulate the aggregate level streamflow variable, which is then subdivided into component flows using the disaggregation approach. Since the aggregate flows are simulated using a separate model, representation of the dependence structure at the aggregate level is the task of the aggregate level model and is not part of the disaggregation procedure discussed here.

Historically the disaggregation approach has been used to subdivide annual flows into seasonal and monthly streamflow, and basinwide aggregate streamflow into the streamflow in individual tributaries. In the first case, disparate time scales are involved, and in the latter, disparate spatial scales. Several applications require that the disaggregate variables add up to the aggregate variable. This is called *summability* and is required, for example, for seasonal flows that sum to an annual total. Summability is not a physical requirement for tributary flows with respect to a basin wide aggregate flow due to channel losses, gains and time delays. A "gains" variable representing the difference between aggregate and sum of disaggregate variables may be used in these cases. Combinations of time and space disaggregation have also been used, as well as disaggregation based on multiple (vector) aggregate variables.

In this chapter, we present a nonparametric approach for disaggregation of an aggregate variable to a set of disaggregated variables. Although our presentation considers the disaggregation of a single aggregate variable into a set of disaggregate variables, extensions to disaggregation from vectors is straightforward. The methodology is the same for temporal and spatial disaggregation. Specifically, we consider a d-dimensional vector $\mathbf{X} = (X_1, X_2, \ldots X_d)^T$ with aggregate variable $Z = X_1 + X_2 + \ldots + X_d$. The problem is posed in terms of sampling from the conditional probability density function.

$$f(\mathbf{X}|Z) = f(\mathbf{X}, Z)/\int f(\mathbf{X}, Z)\, d\mathbf{X} \qquad (4.1)$$

In this equation, $f(\mathbf{X}, Z)$ is the joint probability density function of the vector **X** of disaggregate variables (monthly or tributary streamflows) and Z the aggregate variable (annual or main stem streamflow) obtained from an aggregate model at each aggregate time step. The denominator above is the marginal probability density function of the aggregate variable Z derived by integrating the joint distribution over all the components of **X**. We use kernel density estimation techniques to estimate the joint and conditional densities in (4.1). These methods are data adaptive, i.e., they use the historical data (the historical aggregate and component time series) to define the probability densities. Assumptions as to the form of dependence (e.g., linear or nonlinear) or to the probability density function (e.g., Gaussian) are thus avoided.

This chapter is organized as follows. First we review the historic application of disaggregation models noting that they are special cases of the general conditional simulation problem of Equation (4.1). We discuss drawbacks associated with these approaches and use them to motivate the more general approach based on kernel density estimates proposed here. A brief discussion of multivariate kernel density estimation and the kernel estimator used in the disaggregation model is given next. This is followed by a description of our nonparametric disaggregation approach. The performance of the nonparametric disaggregation procedure is then evaluated by applications to synthetic data from a known nonlinear model and to streamflow from the Snake River at Weiser, Idaho (USGS Streamflow Gauging Station Number 13269000). Results from our approach are compared to those from SPIGOT [*Grygier and Stedinger*, 1990], a popular disaggregation software based on linearizing transformations of the historical streamflow time series.

Background

Historically, practically all applications of the disaggregation approach to streamflow synthesis have involved some variant of a linear model of the form

$$\mathbf{X}_t = \mathbf{A}\,\mathbf{Z}_t + \mathbf{B}\,\mathbf{V}_t \qquad (4.2)$$

Here $\mathbf{X}_t$ is the vector of disaggregate variables at time t, $\mathbf{Z}_t$ the aggregate variable and $\mathbf{V}_t$ a vector of independent random innovations, usually drawn from a Gaussian distribution. **A** and **B** are parameter matrices. **A** is chosen or estimated to reproduce the correlation between aggregate and disaggregate flows. **B** is estimated to reproduce the correlation between individual disaggregate components. The many model variants in the literature make different assumptions as to the structure and sparsity of these matrices and which correlations the model should be made to directly reproduce. *Mejia and Rouselle* [1976] suggest adding a term $\mathbf{C}\,\mathbf{X}_{t-1}$ to reproduce correlation between disaggregate flows in the current and prior time steps. *Stedinger and Vogel* [1984] show that this results in an inconsistency because the parameter estimation procedures assume correlation between the current aggregate and prior disaggregate variables that cannot be guaranteed to be represented by the aggregate/disaggregate model combination. They instead suggest approximating the correlation structure between $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$ by building a correlation structure into the $\mathbf{V}_t$ series. It can be shown [*Bras and Rodriguez-Iturbe*, 1985] with a model of the form of Equation (4.1) that summability of the disaggregate variables to the aggregate variables is guaranteed.

Historically the focus of stochastic disaggregation models has been on reproducing the correlation between variables, assuring summability and matching the marginal distributions through appropriate normalizing transformations. The key idea is to recognize that Equation (4.2) provides a mathematical framework where a joint distribution of disaggregate and aggregate variables is specified. Parameters of this model are estimated so as to approximate selected moments of the joint distribution of the observed historical data.

Some of the drawbacks in such an approach are:

(1) Since Equation (4.2) involves linear combinations of random variables, it is mainly compatible with Gaussian distributions. Where the marginal distribution of the streamflow variables involved is not Gaussian (e.g., perhaps with significant skewness), normalizing transformations are required for each streamflow component. Equation (4.2) would then be applied to the normalized flow variables. It is difficult to find a general normalizing transformation and retain statistical properties of the streamflow process in the untransformed multivariable space. This issue is discussed in the context of a first-order Markov model for streamflow by *Tarboton et al.* [1993]. Specifically with respect to disaggregation, normalization transformations destroy the guaranteed summability of disaggregate flows to aggregate flows, so where summability is important, various empirical adjustment techniques are used [*Grygier and Stedinger*, 1988]. However, these are rather ad hoc and can introduce bias in the model statistics.

(2) The linear nature of Equation (4.2) limits it from representing any nonlinearity in the dependence structure between variables, except through the normalizing transformation used. Given the current recognition of the importance of nonlinearity in many physical processes [*Tong*, 1990; *Schertzer et al.*, 1991], we prefer at the outset not to preclude or limit the representation of nonlinearity.

In the spirit of our recent work [*Tarboton et al.*, 1993; *Lall and Sharma*, 1996], the purpose of this chapter is to show how nonparametric techniques can be used with the disaggregation methodology. Much like the traditional approaches, our approach attempts to approximate the joint probability distribution of flow variables. However, instead of using the linear dependence structure of Equation (4.2), with a priori assumptions as to marginal distributions, or marginal distributions from a narrow set of common distributions, we estimate the necessary joint probability density functions directly from the historic data using kernel density estimates. These methods (described in next section) sidestep the difficulties associated with skewness and normalizing transformations and also retain summability. The nonparametric techniques also admit nonlinear dependence. The methods are data driven and relatively automatic so nonlinear dependence will be incorporated to the extent suggested by the data. Difficult subjective choices as to appropriate marginal distributions and normalizing transformations are avoided. The disadvantages outlined above are thus overcome by the methods we present.

Kernel Density Estimation

Kernel density estimation entails a weighted moving average of the empirical frequency distribution of the data. Most nonparametric density estimators can be expressed as kernel density estimation methods [*Scott*, 1992]. In this chapter, we use multivariate kernel density estimators with Gaussian kernels and bandwidth selected using least squares cross validation [*Scott*, 1992]. This bandwidth selection method is one of many available methods. Our choice of the bandwidth estimator is based on a simulation study (A. Sharma, unpublished report, 1996) that compared various cross validation estimators for samples of sizes typically encountered in hydrology. Our methodology is intended to be generic and should work with any bandwidth and kernel density estimation method. This section reviews kernel density estimation first in a univariate then in a multivariate setting and gives details of the least squares cross validation (LSCV) procedure for estimating bandwidth. For a review of hydrologic applications of kernel density and distribution function estimators, readers are referred to *Lall* [1995]. *Silverman* [1986] and *Scott* [1992] are good introductory texts.

A univariate kernel probability density estimator is written as:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{n\,h} K\left(\frac{x - x_i}{h}\right) \tag{4.3}$$

where x_i, $i=1, \ldots, n$, are the observed data; K(.) is a kernel function that must integrate to 1; and h is the bandwidth that defines the locale over which the empirical frequency distribution is averaged. Many possible kernel functions are given in texts such as *Silverman* [1986] and *Scott* [1992]. *Silverman* [1986] notes that the choice of the kernel function does not result in appreciable differences in the mean integrated square errors of the density estimates. This choice is more often based on considerations such as the computational effort or the degree of differentiability desired in the resulting density. The Gaussian kernel function, a popular and practical choice, is used here.

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \tag{4.4}$$

A multivariate extension of (4.3) with a multivariate Gaussian kernel for a vector **x** in d dimensions can be written as:

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\, \det(\mathbf{H})^{1/2}} \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_i)^T \mathbf{H}^{-1} (\mathbf{x}-\mathbf{x}_i)}{2}\right) \tag{4.5}$$

where det(.) denotes determinant, n is the number of observed vectors $\mathbf{x}_i$, and **H** is a symmetric positive definite d x d bandwidth matrix [*Wand and Jones*, 1994]. The density estimate is thus formed by summing multivariate Gaussian kernels with a covariance matrix **H**, centered at each observation $\mathbf{x}_i$.

A useful specification of the bandwidth matrix **H** is:

$$\mathbf{H} = \lambda^2 \mathbf{S} \tag{4.6}$$

Here, S is the sample covariance matrix of the data and λ^2 prescribes the bandwidth relative to this estimate of scale. These are parameters of the model that are estimated from the data. The procedure of scaling the bandwidth matrix proportional to the covariance matrix (Equation 4.6) is called "sphering" [*Fukunaga*, 1972] and ensures that all kernels are oriented along the estimated principal components of the covariance matrix.

The choice of the bandwidth (h or λ) is an important issue in kernel density estimation. A small value of the bandwidth (h or λ) can result in a density estimate that appears "rough" and has a high variance. On the other hand, too high a bandwidth results in an "oversmoothed" density estimate with modes and asymmetries smoothed out. Such an estimate has low variance but is more biased with respect to the underlying density. This bias-variance trade-off [*Silverman*, 1986] plays an important role in choice of h.

Several methods have been proposed to estimate the "optimal" bandwidth for a given data set. These methods are based on evaluation of factors such as bias, $E\{f(\mathbf{x})-\hat{f}(\mathbf{x})\}$, variance, $\mathrm{Var}\{\hat{f}(\mathbf{x})\}$, Mean Square Error (MSE), $E\{[f(\mathbf{x})-\hat{f}(\mathbf{x})]^2\}$, Integrated Square Error (ISE), and Mean Integrated Square Error (MISE) of the estimate.

$$\mathrm{ISE} = \int_{\Re^d} \left(\hat{f}(\mathbf{x})-f(\mathbf{x})\right)^2 d\mathbf{x} \tag{4.7}$$

$$\mathrm{MISE} = \mathrm{E} \int_{\Re^d} \left(\hat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x} \tag{4.8}$$

One choice for the bandwidth is one that directly minimizes a first-order Taylor series approximation of the MISE (4.8) if the true distribution were known. For a Gaussian distribution with Gaussian kernel functions (estimator defined by Equations (4.5) and (4.6)), *Silverman* [1986] gives the Gaussian reference bandwidth (denoted λ_{ref}) as:

$$\lambda_{ref} = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \tag{4.9}$$

In the univariate case (d=1), this translates to $h = 1.06\,\hat{\sigma}\,n^{-1/5}$ where $\hat{\sigma}$ is an estimate of the standard deviation (Silverman advocates a robust estimate) of the data. Such a bandwidth is optimal only for data arising from a known Gaussian distribution.

Data-driven methods are used in cases where the underlying distribution is not known. They minimize estimates of the ISE or MISE formed only from the data. Least squares cross validation (LSCV) [*Silverman*, 1986] is one such method that is based on minimizing an estimate of the ISE of the kernel density estimate.

Sain et al. [1994] provide an expression for the LSCV score in any dimension with multivariate Gaussian kernel functions and **H** a diagonal matrix. *Adamowski and Feluch* [1991] provide a similar expression for the bivariate case with Gaussian kernels. Here we generalize these results for use with the multivariate density estimator (4.5) to:

$$\mathrm{LSCV}(\mathbf{H}) = \frac{1+\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{\exp(-L_{ij}/4)}{n} - \frac{2^{d/2+1}\exp(-L_{ij}/2)}{n-1}\right]}{(2\sqrt{\pi})^{d}\det(\mathbf{H})^{1/2}} \tag{4.10}$$

where

$$L_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{4.11}$$

We use numerical minimization of (4.10) over the single parameter λ with bandwidth matrix from Equation (4.6) to estimate all the necessary probability density functions. We recognize that LSCV bandwidth estimation is occasionally degenerate, so based on suggestions in *Silverman* [1986] and the upper bound given by *Scott* [1992], we restrict our search to the range $\lambda_{\mathrm{ref}}/4$ to $1.1\lambda_{\mathrm{ref}}$.

Nonparametric Disaggregation Model, NPD

In this section a d-dimensional disaggregation model (denoted NPD) is developed. The model can be used to simulate d-dimensional disaggregate vectors $\mathbf{X}_t$ based on an input aggregate series Z_t. Z_t can be obtained from any suitable model for the aggregate streamflow series; however, we recommend a nonparametric model such as those described in *Tarboton et al.* [1993] or *Lall and Sharma* [1996]. The subscript t above was only shown to reinforce the similarity to historic models like Equation (4.2). Since the same procedure is applied for each time step, from here on the subscript t on $\mathbf{X}_t$ is dropped to save notation.

Disaggregation is posed here in terms of resampling from the conditional density function of Equation (4.1). We need a model that, given Z, provides realizations of **X**. This model is to be based on (calibrated from) n observations of X and Z, denoted $\mathbf{x}_i$ and z_i. The components of $\mathbf{x}_i$ are the historical disaggregate components, such as monthly, seasonal, or tributary flows that comprise the historical aggregate z_i. To use Equation (4.1) an estimate of the d+1 dimensional joint density function $f(X_1, X_2, \dots X_d, Z)$ is required. However, because of summability this has all its mass on the d-dimensional hyperplane defined by:

$$X_1 + X_2 + \dots + X_d = Z \tag{4.12}$$

This probability density can then be represented as:

$$f(X_1, X_2, \dots X_d, Z) = f(X_1, X_2, \dots X_d)\, \delta(Z - X_1 - X_2 \dots - X_d) \tag{4.13}$$

where $\delta(.)$ is the dirac delta function. Kernel density estimation is used to estimate $f(X_1, X_2, \dots X_d)$ based on the data. The conditional density function is then:

$$f(X_1, X_2, \dots X_d | Z) = \frac{\delta(Z - X_1 - X_2 - \dots - X_d)\, f(X_1, X_2, \dots X_d)}{\int_{\text{over plane } X_1 + X_2 + \dots + X_d = Z} f(X_1, X_2, \dots X_d)\, dA} \tag{4.14}$$

For a particular Z, this conditional density function can be visualized geometrically as the probability density on a d-1 dimensional hyperplane slice through the d-dimensional density $f(X_1, X_2, \dots X_d)$, the hyperplane being defined by $X_1 + X_2 + \dots + X_d = Z$. This is illustrated in Figure 4-1, for d = 2. There are really only d-1 degrees of freedom in the conditional simulation. The conditional p.d.f. in (4.14) can then be specified through a coordinate rotation of the vector $\mathbf{X} = (X_1, X_2, \ldots X_d)^T$ into a new vector $\mathbf{Y} = (Y_1, Y_2, \ldots Y_d)$ whose last coordinate is aligned perpendicular to the hyperplane defined by (4.12). Gram Schmidt orthonormalization [*Lang*, 1970] is used to determine this rotation.

Once this rotation has been done, simulation amounts to straightforward conditional resampling from a multivariate density function as has been demonstrated before [*Tarboton et al.*, 1993]. The approach used does not require full evaluation of the conditional density function and basically amounts to perturbed resampling of the data.

Gram Schmidt orthonormalization is a procedure for determining an orthonormal set of basis vectors for a vector space from any suitable basis. The standard basis (basis vectors aligned with the coordinate axes) is orthonormal, but does not have a basis vector perpendicular to the conditioning plane defined by (4.12). We therefore drop one of the standard basis vectors and replace it by a vector perpendicular to the conditioning plane. The basis set is now not orthonormal. We then apply the Gram Schmidt procedure to obtain an orthonormal basis vector set that includes a vector perpendicular to the conditioning plane. The result is a rotation matrix **R** such that

$$\mathbf{Y} = \mathbf{R}\,\mathbf{X} \tag{4.15}$$

where **R** has rows that consist of the basis vectors for the rotated coordinate space.

$$\mathbf{R} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{e}_d \end{pmatrix} \tag{4.16}$$

Denote the standard basis as

$$\begin{aligned} \mathbf{i}_1 &= (1, 0, 0, \dots 0)^T \\ \mathbf{i}_2 &= (0, 1, 0, \dots 0)^T \\ &\dots \\ \mathbf{i}_d &= (0, 0, \dots 0, 1)^T \end{aligned} \tag{4.17}$$

Define

$$\mathbf{e}_d = (1/\sqrt{d}, 1/\sqrt{d}, \dots 1/\sqrt{d})^T = 1/\sqrt{d}\,\mathbf{i}_1 + 1/\sqrt{d}\,\mathbf{i}_2 + \dots + 1/\sqrt{d}\,\mathbf{i}_d \tag{4.18}$$

This is a unit vector perpendicular to the conditioning plane. Now apply Gram Schmidt orthonormalization to obtain an orthonormal basis including $\mathbf{e}_d$. For j decreasing from d-1 to 1:

$$\begin{aligned} \mathbf{e}'_j &= \mathbf{i}_j - \sum_{k=j+1}^{d} (\mathbf{e}_k \bullet \mathbf{i}_j)\, \mathbf{e}_k \\ \mathbf{e}_j &= \mathbf{e}'_j / |\mathbf{e}'_j| \end{aligned} \tag{4.19}$$

The first step above obtains a vector orthogonal to the basis vectors e_k, $k = j+1 \ldots d$, obtained thus far and the second step normalizes it to unit length. Since **R** is defined with unit orthogonal basis vectors, $\mathbf{R}\,\mathbf{R}^T = \mathbf{I}$ so $\mathbf{R}^{-1} = \mathbf{R}^T$.

With this rotation (Equation 4.16), the last coordinate of **Y**, Y_d, is in fact a rescaling of Z:

$$Y_d = Z/\sqrt{d} \tag{4.20}$$

It is convenient to define a subset of **Y** , $\mathbf{U} = (Y_1, Y_2, \ldots Y_{d-1})^T$ that is the first d-1 components of **Y** and reflects the true d-1 degrees of freedom in the conditional simulation. We then denote $\mathbf{Y} = (\mathbf{U}^T, Z')^T$ where $Z' = Z/\sqrt{d}$. We actually resample from $f(\mathbf{U}|Z') = f(Y_1, Y_2, \ldots Y_{d-1}|Z')$ and recover the disaggregate components of **X** by backrotation. The kernel density estimate in rotated coordinates is obtained by substituting $\mathbf{X} = \mathbf{R}^T\mathbf{Y}$ into (4.5) with bandwidth matrix **H** from Equation (4.6):

$$\hat{f}(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\,\lambda^d\,\det(\mathbf{S})^{1/2}}\exp\left(-\frac{(\mathbf{Y}-\mathbf{y}_i)^T\mathbf{R}\,\mathbf{S}^{-1}\,\mathbf{R}^T\,(\mathbf{Y}-\mathbf{y}_i)}{2\,\lambda^2}\right). \tag{4.21}$$

One should recognize that $\mathbf{R}\,\mathbf{S}^{-1}\,\mathbf{R}^T = (\mathbf{R}\,\mathbf{S}\,\mathbf{R}^T)^{-1} = \mathbf{S}_y^{-1}$ represents a rotation of the covariance matrix **S** into $\mathbf{S}_y$. Also $\det(\mathbf{S}_y) = \det(\mathbf{S})$. Therefore, this is equivalent to applying (4.5) to the rotated data. Now with $Z' = Z/\sqrt{d}$ given in $\mathbf{Y} = (\mathbf{U}^T, Z')^T = (Y_1, Y_2, \ldots Y_{d-1}, Z')^T$, the conditional density function we resample from is:

$$\hat{f}(\mathbf{U}|Z') = \hat{f}(Y_1, Y_2, \ldots Y_{d-1}|Z') = \frac{\hat{f}(\mathbf{U}, Z')}{\int\hat{f}(\mathbf{U}, Z')\,d\mathbf{U}} \tag{4.22}$$

where $\hat{f}(\mathbf{U}, Z') = \hat{f}(Y)$ is obtained from (4.21) recalling that **U** denotes $(Y_1, Y_2, \ldots, Y_{d-1})^T$, the vector **Y** without the last component. The covariance matrix $\mathbf{S_y}$ is partitioned as follows:

$$S_y = \begin{bmatrix} S_u & S_{uz} \\ S_{uz}^T & S_z \end{bmatrix} \tag{4.23}$$

S_u is the d-1 x d-1 covariance matrix of **U**. S_z is the 1 x 1 variance of Z' and S_{uz} a vector of cross covariance between each component of **U** and Z'. Substituting (4.21) in (4.22), we obtain:

$$\hat{f}(\mathbf{U}|Z') = \frac{1}{(2\pi\lambda^2)^{(d-1)/2} \det(S')^{-1/2}} \sum_{i=1}^{n} w_i \exp\left(\frac{(\mathbf{U} - \mathbf{b}_i)^T S'^{-1} (\mathbf{U} - \mathbf{b}_i)}{2\lambda^2} \right) \tag{4.24a}$$

where

$$w_i = \exp\left(- \frac{(Z' - z'_i)^2}{2\lambda^2 S_z} \right) \Bigg/ \sum_{j=1}^{n} \exp\left(- \frac{(Z' - z'_j)^2}{2\lambda^2 S_z} \right) \tag{4.24b}$$

$$S' = S_u - S_{uz} S_z^{-1} S_{uz}^T \tag{4.24c}$$

$$\mathbf{b}_i = \mathbf{u}_i + \mathbf{S}_{uz}\, S_z^{-1}(Z' - z'_i) \tag{4.24d}$$

This is illustrated in Figure 4-1 for d = 2 and shows the conditional density function $\hat{f}(U|Z')$ as a weighted sum of n Gaussian density functions each with mean $\mathbf{b}_i$ and covariance λ^2 **S'**. The area (volume for d > 2) under each kernel slice is the weight w_i that controls the contribution of point i to the conditional density estimate. Equation (4.24b) shows that this weight depends on the distance of z'_i from the conditioning value Z'. Observations that lie closer to the conditioning value (i.e., where $(Z' - z'_i)$ is small) receive greater weight. The weights are normalized to add to unity.

Resampling from Equation (4.24) proceeds as follows:

Preprocessing:

(1) Compute the sample covariance matrix **S** from the data $\mathbf{x}_i$.

(2) Solve for λ by numerically minimizing (4.10) (using 0.25 and 1.1 times the solution of (4.9) to bracket the search) with **H** from (4.6).

(3) Compute **R**, $\mathbf{S}_z$ and **S'** from Equations (4.16), (4.23) and (4.24c).

(4) Use singular value decomposition to obtain **B** such that $\mathbf{B}\,\mathbf{B}^T \doteq \mathbf{S}'$.

At each time step:

(5) Given Z from the aggregate model at each time step first calculate the weight w_i associated with each observation, using Equation (4.24b).

(6) Pick a point i with probability w_i.

(7) Generate a d-1 dimensional unit Gaussian vector **V**. Each component in **V** is independent N(0,1).

(8) Obtain the simulated U from $\mathbf{U} = \mathbf{b}_i + \lambda\, \mathbf{B}\, \mathbf{V}$. $\mathbf{Y} = (\mathbf{U}^T, Z)^T$.

(9) Rotate back to the original coordinate space. $\mathbf{X} = \mathbf{R}^T\, \mathbf{Y}$.

Steps 5 through 9 are repeated for each aggregate time step. A complication can arise because the Gaussian kernels used in the kernel density estimate have infinite support. Thus they assign some (hopefully small) probability to regions of the domain where streamflow is negative (i.e., invalid or out of bounds). This leakage of probability across boundaries is a problem associated with kernel density estimates based on kernels with infinite support. Kernel density estimates also suffer from problems of bias near the boundaries. Here we address the leakage by checking the flows for validity (positiveness) and if they are invalid, repeat steps 7 to 9 for a given time step. That is, we regenerate a new vector **V** and try again. This amounts to, in the kernel density estimate, removing (cutting) the portion of each kernel that is out of bounds and renormalizing that kernel to have the appropriate mass over the within bounds domain, the so-called cut-and-normalize approach applied to each kernel. We record how often this is done as frequent boundary normalization is symptomatic of substantial boundary leakage. Alternative approaches that use special boundary kernels [*Hall and Wehrly*, 1991; *Wand et al.*, 1991; *Djojosugito and Speckman*, 1992; *Jones*, 1993] or work with log transformed data could be used in cases where this method for handling the boundaries is found to be unsatisfactory.

Model Evaluation

This section explores the use and effectiveness of the NPD approach. It is first applied to data from a specified bimodal distribution. This tests the model's ability to maintain distributional characteristics such as nonlinearity and bimodality. It is then applied to simulate monthly streamflow in the Snake River.

To provide a point of reference, we also generate results using SPIGOT [*Grygier and Stedinger*, 1990]. SPIGOT is a parametric disaggregation model representative of current hydrologic practice. The first subsection below describes SPIGOT. The next subsection then describes the tests against the specified distribution. This is followed by the Snake River application.

SPIGOT

SPIGOT [*Grygier and Stedinger*, 1988; *Grygier and Stedinger*, 1990] is a parametric synthetic streamflow generation package that includes an annual streamflow generation module, an annual to monthly disaggregation module, and a spatial disaggregation module. We used the first two modules in our applications. An Autoregressive model of order 1 (AR1) or order 0 (AR0) is used to generate the annual streamflow. The annual to monthly disaggregation model in SPIGOT is the condensed model described by *Grygier and Stedinger* [1988, 1990] as:

$$X_i = \alpha_i + \beta_i Y + \gamma_i X_{i-1} + \delta_i \sum_{j=1}^{i-1} w_j X_j + \varepsilon_i V \tag{4.25}$$

where Y is the normalized annual flow, X_i the normalized flow in month i, w_i a set of weights depending on the marginal distribution used, chosen to maximize the likelihood of the untransformed monthly flows adding up to the untransformed annual flow, and V is a random innovation taken as a Gaussian random variate with mean 0 and variance 1. α_i, β_i, γ_i, δ_i, and ε_i are parameters estimated by regression for each month. For the first month, γ_i is constrained to be zero and for the first two months δ_i is constrained to be zero to ensure self-consistency in the sense described by *Stedinger and Vogel* [1984].

SPIGOT first transforms the historical annual and monthly (or seasonal) flows to Gaussian using four choices for the marginal probability densities. These are:

(1) Gaussian

(2) Two-parameter Lognormal

(3) Three-parameter Lognormal

(4) Three-parameter Gamma using the Wilson-Hilferty transformation [*Loucks et al.*, 1981].

The parameters for each distribution are estimated by matching moments and the best fitting distribution chosen by measuring the correlation of observations to the fitted distribution quantiles (Filliben correlation statistic, *Grygier and Stedinger* [1990]).

Tests with Synthetic Data

Here we describe a Monte Carlo investigation to test the ability of the NPD approach to approximate a specified underlying distribution. Our test distribution, illustrated in Figure 4-2, is based on distribution J in *Wand and Jones* [1993]. It consists of a mixture of three bivariate Gaussians having different weights α_i, stated as:

$$f = \sum_{i=1}^{3} \alpha_i \, N(\mu_i, \Sigma_i) \tag{4.26}$$

where $N(\mu_i, \Sigma_i)$ denotes a Gaussian distribution with mean μ_i and a covariance matrix Σ_i. Individual weights, means, and covariances are shown in Table 4-1. Simulation from it is achieved by picking one of the three Gaussian distributions with probability α_i, then simulating a value from that distribution.

We simulated 101 bivariate samples each consisting of 80 data pairs from this distribution. The first sample is designated as the "calibration" sample and is used to calibrate the NPD and SPIGOT models. In the case of NPD, this involves estimating the sample covariance and bandwidth parameter λ (based on minimizing the LSCV score as described in previous section). Calibration of SPIGOT involves selection of the best marginal density transformation based on Filliben correlation statistic, and estimation of the coefficients in the condensed disaggregation model in Equation (4.25). The remaining 100 samples are used to form 100 aggregate test realizations by adding the components $Z = X_1 + X_2$. These 100 aggregate test realizations are input to both NPD and SPIGOT to generate 100 disaggregate realizations from both models. These disaggregate series are designated "test" samples and serve as a basis to test how closely the model reproduces statistics of the specified true distribution and of the calibration sample.

SPIGOT was modified to accept the same aggregate flows as the NPD model. Boundary corrections (discussed in the previous section for the NPD approach; and specified as an option in the SPIGOT software) were not imposed on either model.

To evaluate the reproduction of marginal distributions by each model, we applied a univariate kernel density estimate to both components (X_1 and X_2) of each disaggregated sample. Figures 4-3a and 4-3b illustrate marginal densities of the calibration and disaggregated samples for variables X_1 and X_2. Disaggregated sample p.d.f.'s are represented using boxplots, which consist of a box that extends over the interquartile range of the quantity (in this case the p.d.f.) being plotted. The line in the center of this box is the median and whiskers extend to the 5% and 95% quantiles of the compared statistic. The marginal densities of both the calibration sample and the SPIGOT or NPD disaggregations were estimated using a single common bandwidth rather than optimal bandwidths from the minimization of (4.10) for each simulated or historical sample. This was done to avoid differences due to different bandwidths for individual samples. The bandwidth we used was estimated as the median amongst the set of optimal bandwidths for the historical sample and the NPD and SPIGOT realizations. The univariate KDE curve (see NPD and aggregate flow results in Figure 4-3) is also estimated using this median bandwidth. The curve marked "calibrated" represents the marginal density that is theoretically reproduced in simulations from either approach. This is estimated from the joint density of the calibration sample, and is a univariate parametric p.d.f. (depending on the transformation used) in case of SPIGOT results, and a numerically evaluated integral of the joint density of X_1 and X_2 (with respect to X_2 for marginal density of X_1, and X_1 for marginal density of X_2) in case of NPD results. One must note that while disaggregation model marginal densities *will always* be similar to the calibrated marginals, they are *supposed* to resemble the true or univariate KDE curves instead. In the case of NPD results in Figures 4-3a and 4-3b, the true, calibrated, and univariate KDE curves all show the same structure as the disaggregations. This is in contrast to SPIGOT disaggregations in Figure 4-3a, where imposition of a three-parameter Lognormal distribution on variable X_1 results in realizations that bear little resemblance to the sample density estimate or underlying true p.d.f. Figure 4-3b, however, shows good performance by SPIGOT, because the underlying distribution of variable X_2 can be well represented by the Gamma marginal density transformation.

Figure 4-4 illustrates some common statistics for realizations from both approaches. Both models reproduce these statistics well. The poor performance of SPIGOT on the marginal density of variable X_1 (Figure 4-3a) does not show up in this comparison of sample statistics.

The above test shows that the nonparametric approach is able to model the joint distribution of X_1 and X_2 estimated based on a single sample. We also tested the ability of the nonparametric approach to reproduce the underlying distribution in Figure 4-2. We rotated the samples in the above test such that each of the 101 samples was used once for calibration with the remaining 100 being used for disaggregation. The joint density of each calibration sample and its associated disaggregations were computed and stored using the kernel estimate of (4.5) with bandwidth estimated using (4.10). The average joint density of 101 calibration samples is illustrated in Figure 4-5. The average joint density of 10100 (101x100) NPD disaggregations is shown in Figure 4-6. The mean of the 101 ISE (see Equation 4.7) estimates for bivariate densities of calibration samples is 0.0207. The mean of the 10100 NPD disaggregation ISE's is 0.0339. In comparing these figures to Figure 4-2, it is apparent that there is some smoothing in the kernel estimates of both calibration and disaggregation samples. Smoothing in the calibration p.d.f.'s (Figure 4-5) is proportional to the second derivative of the true density [*Scott*, 1992] and could be reduced by use of a "local" bandwidth (one that varies with location) instead of the global bandwidth in (4.5). The average p.d.f. formed by the NPD disaggregations (Figure 4-6) is more smoothed than the calibration p.d.f. of Figure 4-5. This extra smoothing is because of addition of noise (the unit Gaussian vector, **V**, in step 7 of the simulation algorithm in previous section) to the resampled data points (vector $\mathbf{b}_i$ in step 8 of same algorithm). Another outcome of this extra smoothing is a slight inflation in the variance and deflation in the skewness of NPD realizations. Expressions for these and other statistics (mean, variance, correlation, and skewness) are derived for the nonparametric order one streamflow simulation model (denoted NP1) in Appendix B and can be extended (with modest modifications) to our nonparametric disaggregation approach.

Tests with Monthly Flow Data

Here we describe the application of the nonparametric disaggregation (NPD) model to 83 years (1911 to 1993) of monthly streamflow in the Snake River at Weiser, Idaho, located at 44° 14' 44" N and 116° 58' 48" W at an elevation of 2086 ft above Mean Sea Level, as recorded by the U. S. Geological Survey (USGS Station Number 13269000). This data set was one amongst many streamflow data sets on which we tested our model, all with satisfactory results. This site was chosen because in other work (Chapter 3) we identified it as a river that had a marked nonlinear relationship between July and August month flows, as is illustrated in the joint density for the July-August month pair in Figure 4-7, estimated using the kernel estimator in (4.5). Such nonlinearity cannot be easily represented with commonly used parametric density functions; therefore, these streamflow data were a promising candidate for the use of nonparametric methods.

We compare results from application of the NPD model with those from SPIGOT. Aggregate flows for the NPD application were simulated using the NP1 model (Chapter 3). The NP1 model is a nonparametric model constructed to preserve first-order Markov dependence in a time series. It consists of a bivariate nonparametric density estimate $\hat{f}(Z_t, Z_{t-1})$ formed by applying Equation (4.5) to each sequential data pair. Flow values are then obtained by sequentially resampling from the conditional density estimate $\hat{f}(Z_t|Z_{t-1})$. Aggregate flows for the SPIGOT application were simulated using an autoregressive lag 1 (AR1) model. A marginal density transform based on the best Filliben correlation statistic [*Grygier and Stedinger*, 1990] was used to transform historical annual flows to Gaussian. One hundred realizations, each of length 83 years, were generated from both approaches. The length of 83 years was chosen to be the same as the length of available historic data so that the variability of sample statistics across these realizations is representative of the sampling variability of the historic data. Negative realizations from NPD (amounting to about 0.1% of the total number of realizations) were resimulated using the boundary correction procedure described in the previous section. Both models were tested for their ability to reproduce the following statistics of the historic data.

(1) Mean

(2) Standard deviation

(3) Coefficient of skewness

(4) Cross correlation between seasonal streamflows and between seasonal and annual streamflow

(5) Kernel estimates of marginal distributions

(6) 'State-dependent correlations'; correlations between different month pairs as a function flow magnitude.

Figure 4-8 shows the monthly and annual aggregate mean streamflow from both SPIGOT and NPD models, comparing historical and simulated flows using boxplots. The fact that the historical means fall within the range of the boxes indicates that both models reproduce mean flows.

Figures 4-9 and 4-10 use boxplots to compare standard deviation and skewness of simulated and historical streamflows, respectively. Again both models reproduce these statistics well, though some extra smoothing is visible in the nonparametric simulations. This leads to inflation in the standard deviations and deflation in the skewness of disaggregate flows from the nonparametric model. On the other hand, SPIGOT tends to inflate the skewness in the months where the marginal density transform is inadequate (for example, the Lognormal transformation performed on the May flows theoretically leads to a coefficient of skewness of 1.6 as compared to the actual skewness of 0.54).

Figure 4-11 compares the cross correlations of the monthly and aggregate flows from both models. The nonparametric model reproduces this statistic without bias while SPIGOT is unable to model the dependence between certain month pairs (for example, the simulated correlation between flows of month pairs 4-10 and 4-11 is lower than the observed). This could be due to either some bias because of the marginal density transform or the use of a condensed model (4.25) instead of a comprehensive model such as in (4.2).

In Figure 4-12 the probability density estimates of the observed and simulated flows are compared. As in Figure 4-3, we used a common bandwidth (chosen as the median of a set estimated by minimizing LSCV for historical and simulated samples) to compute these univariate density estimates. The aggregate annual flows from AR1 and NP1 models that drive the SPIGOT and NPD models as well as monthly flows from April, June, and July are compared. The dotted line in case of SPIGOT flows represents the modeled p.d.f. as suggested by the Filliben correlation statistic. Annual flows are modeled well by both approaches. April flows are well modeled by both approaches, though a cluster of observations around 40,000 cfs manifests itself better in NPD realizations. June flows have a peculiar plateau-shaped p.d.f., which is not well represented by the SPIGOT simulations. This is because none of the four marginal density transformations available as options in SPIGOT can represent this particular density shape. July flows have a bimodal p.d.f. that SPIGOT is unable to represent. Although the NPD model does reproduce the bimodality, it attenuates the main mode peak, again due to the extra smoothing imparted by the nonparametric model (see discussion in the synthetic example in previous subsection).

It is also worth emphasizing that the NPD monthly flows here, by construction, add up to the simulated aggregate flow used as input. There is therefore no need for adjustments to fix this such as is necessary in SPIGOT [*Grygier and Stedinger*, 1988].

Recall that July-August flows (Figure 4-7) suggested dependence of correlation on the flow magnitude (see difference in slopes of conditional mean for flows less than or greater than 12,000 cfs in Figure 4-7) that is not easily modeled by parametric models such as SPIGOT. In earlier work [*Tarboton et al.*, 1993], we used a statistic that quantified the dependence of correlation on the magnitude of flow. This statistic (denoted the state-dependent correlation statistic [*Tarboton et al.*, 1993]) measures the correlation between flows above or below the median in month i with succeeding flows in month j. For example, the "Correlation above and forward" for the July-August month pair would be the correlation between July flows that are above the median July flow and their succeeding August flows. Differences between above median and below median correlations are indicative of nonlinear state-dependence in the correlation structure.

Figure 4-13 shows the "Above and Forward" and "Below and Forward" state-dependent correlations for flows in adjacent months from both models. The last month pair (September-October) is not well represented because over-year dependence is not modeled by either approach. What is notable though, is that while SPIGOT flows are unable to reproduce the July and August state-dependent correlations, the nonparametric approach approximates the "above and forward" correlations well, though showing a bias in the "below and forward" correlations for the July-August month pair. This bias is due to smoothing on the calibration p.d.f. in the NPD realizations (recall the differences between Figures 4-5 and 4-6 in the synthetic flow example in previous subsection). On the whole, the nonparametric approach shows less bias than SPIGOT.

The statistics compared thus far do not give the full picture relevant for hydrology. Stochastic streamflow sequences are frequently used to evaluate storage and water resources issues. Therefore, it is necessary to ensure that simulated sequences are representative of the historic data with respect to these. Table 4-2 presents the bias and Root Mean Square Error (RMSE) of the reservoir storage capacity corresponding a fixed yield of 80% of the mean annual streamflow. These storages were estimated based on the sequent peak algorithm [*Loucks et al.*, 1981] using equal monthly demands (1/12 of the fixed yield fraction), and the bias and RMSE evaluated as fractions of the storage estimated from the historical record:

$$\text{Bias} / S_h = (S_h - \frac{1}{n_r}\sum_{i=1}^{n_r} S_{s_i})/S_h \qquad (4.27)$$

$$\mathrm{RMSE} / S_h = \frac{\sqrt{\frac{1}{n_r}\sum_{i=1}^{n_r}(S_h - S_{s_i})^2}}{S_h} \tag{4.28}$$

where S_h denotes the historical storage, S_{s_i} is the storage from the i'th realization, and n_r is the total number of realizations. The nonparametric model has a smaller bias and RMSE than SPIGOT. Some other measures of hydrologic relevance were also compared but space limitations prevent presentation of these results. Long-range dependence quantified in terms of the Hurst Coefficient [*Hurst*, 1951] was preserved by both models. The minimum average streamflow associated with different averaging durations was computed to test the ability of each model to reproduce short- and long-term droughts. Both models again performed well.

Discussions and Conclusions

Disaggregation models are useful for generating cross correlated sequences of multisite annual and seasonal flows. We have shown how they can be formulated and used in a nonparametric density estimation framework. This avoids the difficulty and arbitrariness associated with distribution fitting and normalization transformations in parametric approaches. It also directly preserved summability, avoiding the necessity for adjustments to ensure this.

We used two test cases to illustrate that the nonparametric method performs as well as parametric disaggregation approaches. The first test case demonstrated the ability of the NPD approach to model statistical attributes for samples from a bimodal distribution. We also showed that the NPD approach represents the marginal and joint density functions (which for this case were known beforehand) effectively, something that parametric approaches are not able to do in general. The second test application (on observed monthly flows in the Snake River) presented insights into actual performance of the NPD model on a water resources system. It was encouraging to note that the NPD approach was able to portray both conventional and hydrologically relevant statistics as well as conventional methods, while providing a better representation of the sample joint distribution in its realizations.

While this was a small application, chosen to illustrate and present this method, applications to larger situations should be straightforward extensions of the theory described here. One should, however, take note of problems that can arise if too few data points are used to estimate the joint probability density in a high dimensional space. It is advisable to use staging procedures [*Loucks et al.*, 1981] that can help reduce the dimensionality of the problem, and hence provide greater workability with the limited size samples that are typical in hydrologic applications. Future work will focus on such issues and also on devising disaggregation alternatives using the nearest-neighbor resampling philosophy espoused in *Lall and Sharma* [1996].

Least squares cross validation was used to estimate the bandwidths necessary for kernel density estimation in this study. However, the disaggregation method does not depend on this and is applicable with any kernel density estimation procedure and any bandwidth. Procedures for bandwidth and kernel selection are an area of active research in the nonparametric statistics community, and as better methods become available they can be easily incorporated into our model.

We are convinced that nonparametric techniques have an important role to play in improving the synthesis of hydrologic time series for water resources planning and management. They can capture the dependence structure present in the historic data, without imposing arbitrary linearity or distributional assumptions. They have the capability to reproduce nonlinearity, state dependence, and multimodality while remaining faithful to the historic data and producing synthesized sequences statistically indistinguishable from the historic sequence.

References

Adamowski, K., and W. Feluch, Application of nonparametric regression to groundwater level prediction, *Can. J. Civ. Eng.*, 18, 600-606, 1991.

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, Mass., 1985.

Cleveland, W. S., and S. J. Devlin, Locally weighted regression : An approach to regression by local fitting, *J. Am. Stat. Assoc.*, 83(403), 596-610, 1988.

Curry, K., and R. L. Bras, Theory and applications of the multivariate broken line, dissaggregation and monthly autoregressive streamflow generators to the Nile River, *Technology Adaptation Program*, MIT, Cambridge, Mass., 1978.

Djojosugito, R. A., and P. L. Speckman, Boundary bias correction in nonparametric density estimation, *Comm. Stat.-Theory and Methods*, 21(1), 69-88, 1992.

Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.

Grygier, J. C., and J. R. Stedinger, Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, 24(10), 1574-1584, 1988.

Grygier, J. C., and J. R. Stedinger, *Spigot, A Synthetic Streamflow Generation Package, Technical Description, Version 2.5*, School of Civil and Environmental Engineering, Cornell University, Ithaca, N.Y., 1990.

Hall, P., and T. E. Wehrly, A geometrical method for removing edge effects from kernel-type nonparametric regression estimators, *J. Am. Stat. Assoc.*, 86(415), 665-672, 1991.

Hurst, H. E., Long-term storage capacity of reservoirs, *Trans.Am.Soc.Civ. Eng.*, 116, 770-799, 1951.

Jones, M. C., Simple boundary correction for kernel density estimation, *Stat. Comp.*, 3, 135-146, 1993.

Lall, U., Nonparametric function estimation: Recent hydrologic applications, U.S. National Report to International Union of Geodesy and Geophysics, 1991-1994, *Rev. Geophys*, 33, 1093, 1995.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for time series resampling, *Water Resour. Res.*, 32(3), 679-693, 1996.

Lane, W. L., *Applied Stochastic Techniques, Users Manual,* Bureau of Reclamation, Engineering and Research Center, Denver, Colo., 1979.

Lang, S., *Linear Algebra,* 2nd ed., Addison-Wesley, Reading, Mass., 1970.

Loucks, D. P., J. R. Stedinger, and D. A. Haith, *Water Resource Systems Planning and Analysis,* Prentice-Hall, Englewood Cliffs, N.J., 1981.

Mejia, J. M., and J. Rousselle, Disaggregation models in hydrology revisited, *Water Resour. Res.*, 12(2), 185-186, 1976.

Sain, S. R., K. A. Baggerly, and D. W. Scott, Cross-validation of multivariate densities, *J. Am. Stat. Assoc.*, 89(427), 807-817, 1994.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series,* Water Resources Publications, Littleton, Colo., 1980.

Schertzer, D., S. Lovejoy, F. Schmitt, and D. Lavallee, Multifractal analysis and simulations of nonlinear geophysical signals and images, *Treizieme Colloque Gretsi,* 16(20), 1313-1325, 1991.

Scott, D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization,* John Wiley & Sons, New York, 1992.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis,* Chapman and Hall, London, England, 1986.

Stedinger, J. R., D. Pei, and T. A. Cohn, A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res.*, 21(5), 665-675, 1985.

Stedinger, J. R., and R. M. Vogel, Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, 20(11), 47-56, 1984.

Svanidze, G. G., *Mathematical Modeling of Hydrologic Series,* Water Resources Publications, Fort Collins, Colo., 1980.

Tarboton, D. G., A. Sharma, and U. Lall, The use of non-parametric probability distributions in streamflow modeling, in *Proceedings of the Sixth South African National Hydrological Symposium,* edited by S. A. Lorentz, S. W. Kienzle, and M. C. Dent, pp. 315-327, University of Natal, Pietermaritzburg, South Africa, 1993.

Tong, H., *Nonlinear Time Series Analysis: A Dynamical Systems Perspective,* Academic Press, London, England, 1990.

Valencia, D. R., and J. L. Schaake, A dissaggregation model for time series analysis and synthesis, *Ralph M. Parsons Laboratory, M.I.T.*, 1972.

Wand, M. P., and M. C. Jones, Comparison of smoothing parameterizations in bivariate kernel density estimation, *J. Am. Stat. Assoc.*, 88(422), 520-528, 1993.

Wand, M. P., and M. C. Jones, Multivariate plug-in bandwidth selection, *Comp. Stat.*, 9, 97-116, 1994.

Wand, M. P., J. S. Marron, and D. Ruppert, Transformations in density estimation, *J. Am. Stat. Assoc.*, 86(414), 343-361, 1991.

Table 4-1. Parameters of the Test Distribution

Gaussian density	α_i	μ_i	Σ_i
1	0.4	(1.3, 2.5)	$\begin{pmatrix} 0.36 & 0.252 \\ 0.252 & 0.36 \end{pmatrix}$
2	0.4	(3.7, 2.5)	$\begin{pmatrix} 0.36 & 0.252 \\ 0.252 & 0.36 \end{pmatrix}$
3	0.2	(2.5, 2.5)	$\begin{pmatrix} 0.36 & -0.252 \\ -0.252 & 0.36 \end{pmatrix}$

Table 4-2. Reservoir Capacities from 100 Realizations for a Yield of 0.8 Mean Annual Flow

Model	Bias/S_h	RMSE/S_h
SPIGOT	0.2366	0.4319
NP	0.0316	0.3983

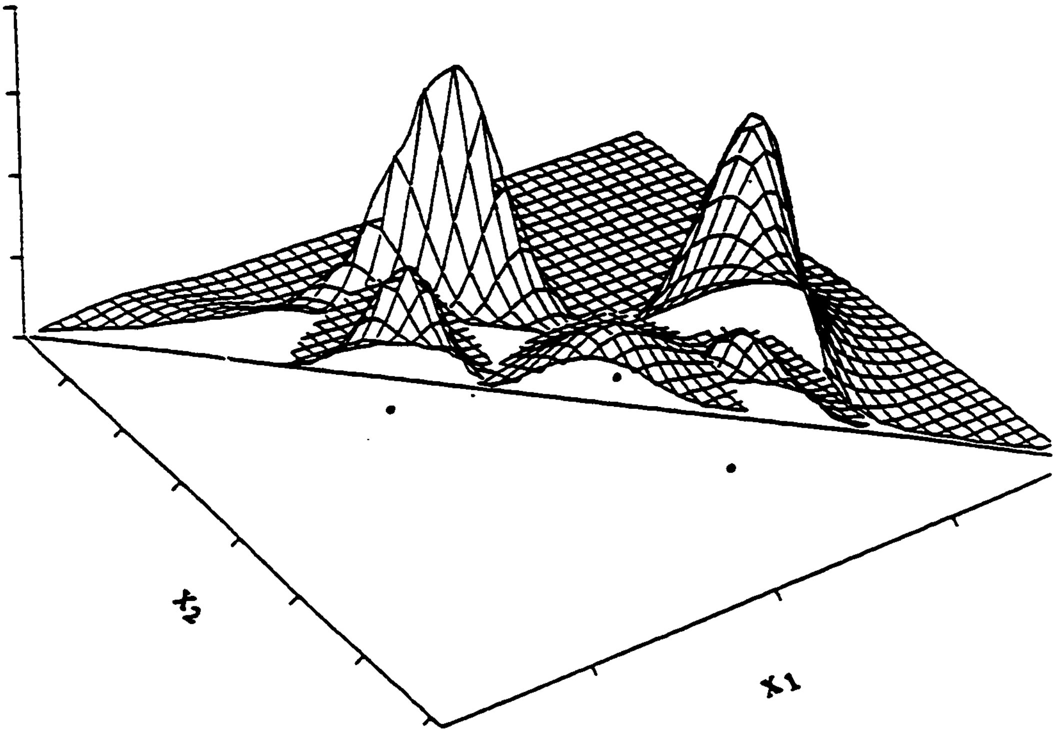


Figure 4-1. Illustration of a conditional density estimate $\hat{f}(X_1, X_2|Z)$ with $Z = X_1 + X_2$ as a slice through the joint density function. Since the joint density estimate is formed by adding bivariate kernels, the conditional density is estimated as a sum of kernel slices.

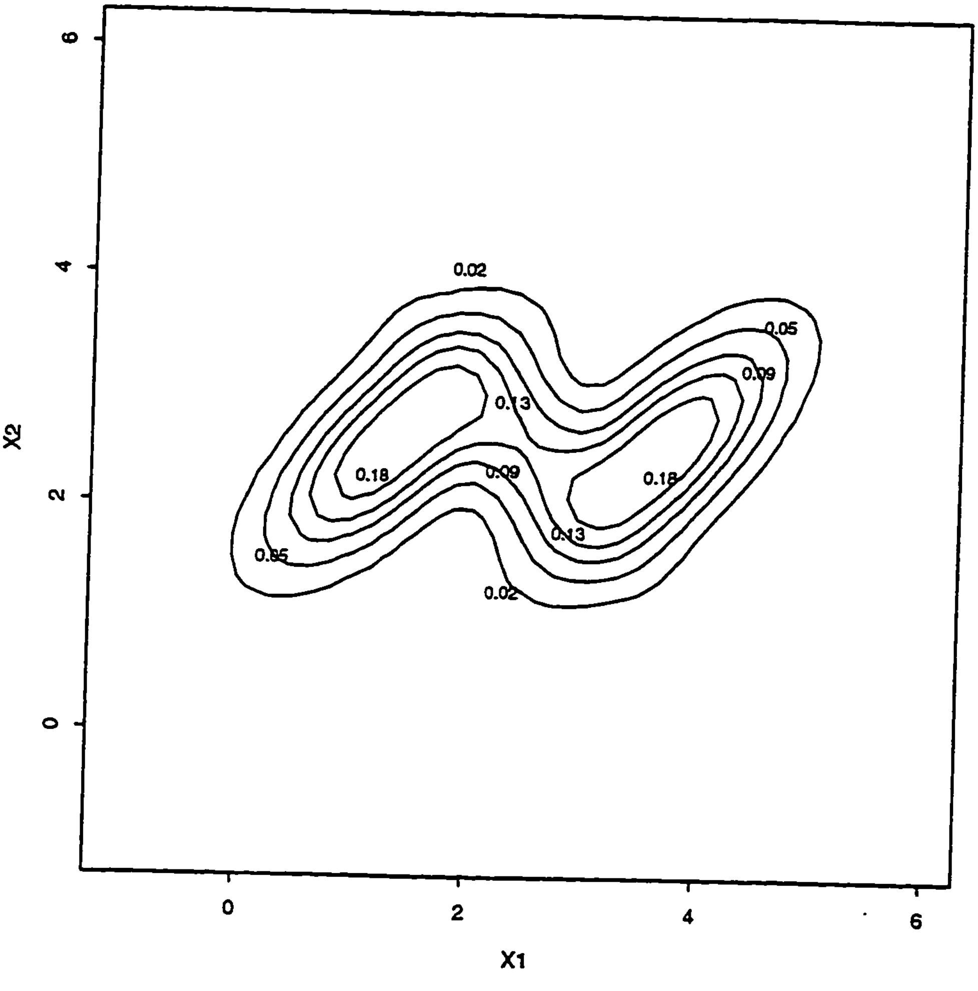
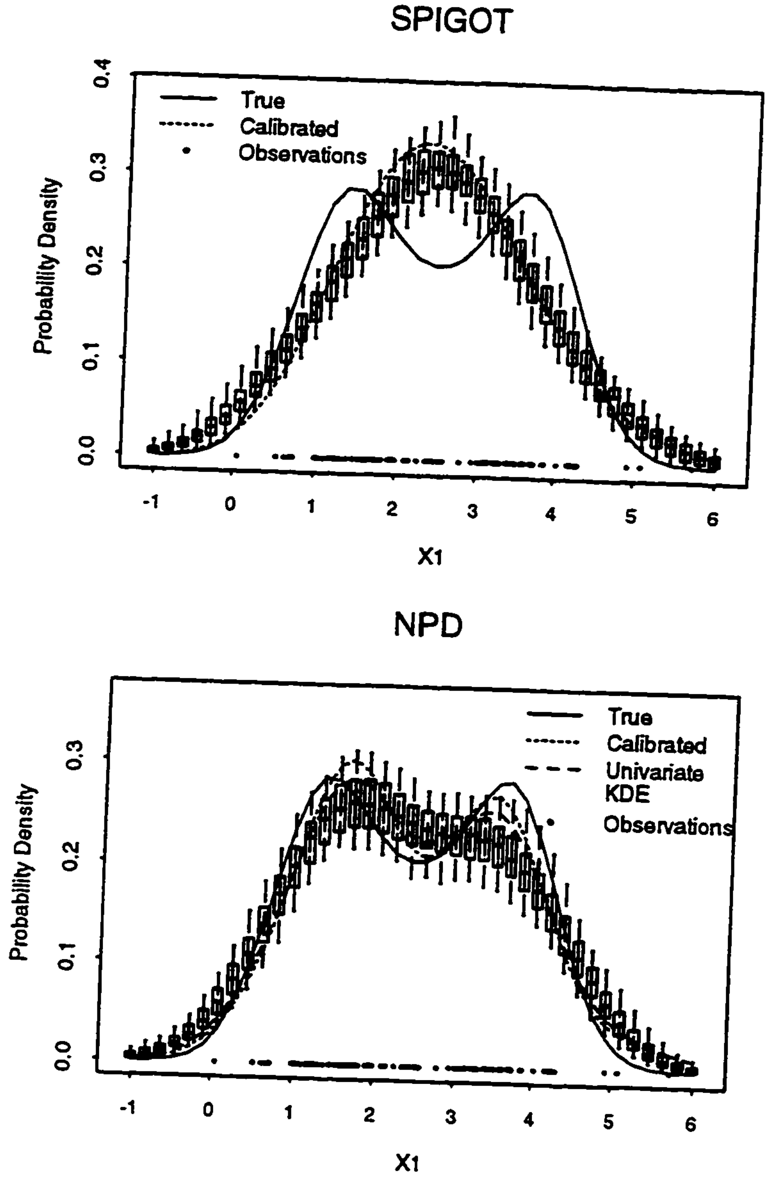
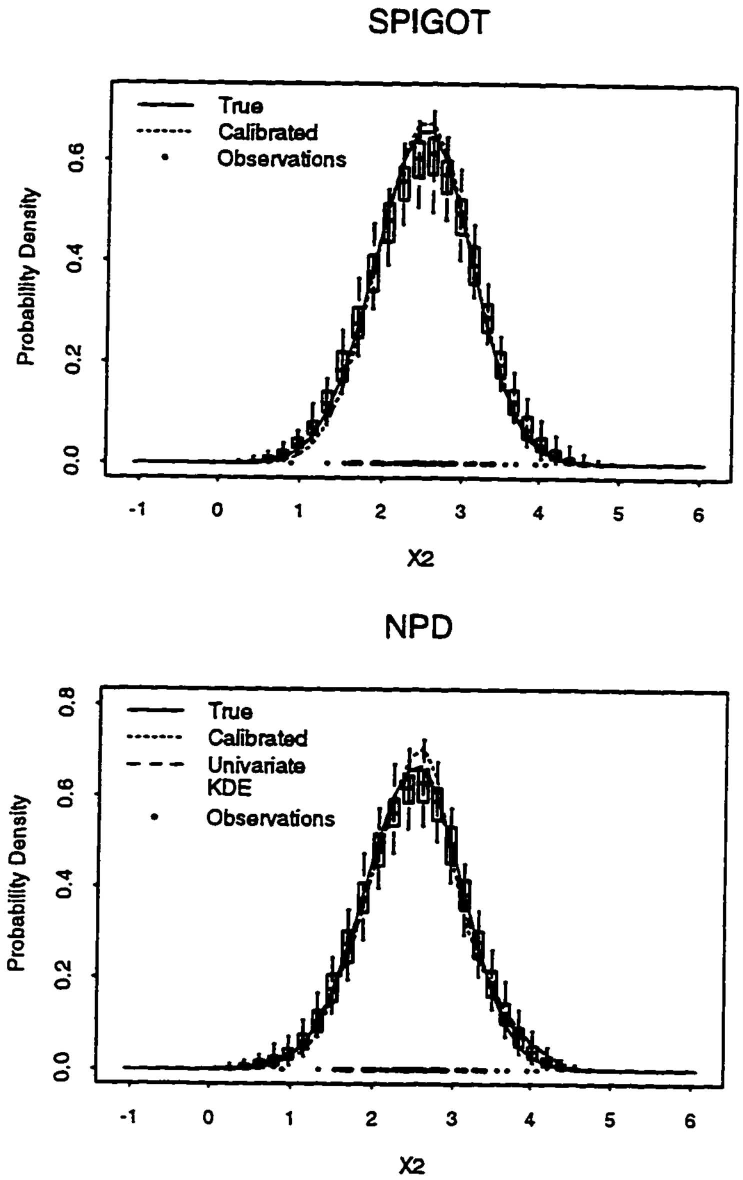


Figure 4-2. Bivariate distribution used in the synthetic example to test the disaggregation approach. This is a mixture of the three bivariate Gaussian density functions described in Table 4-1.



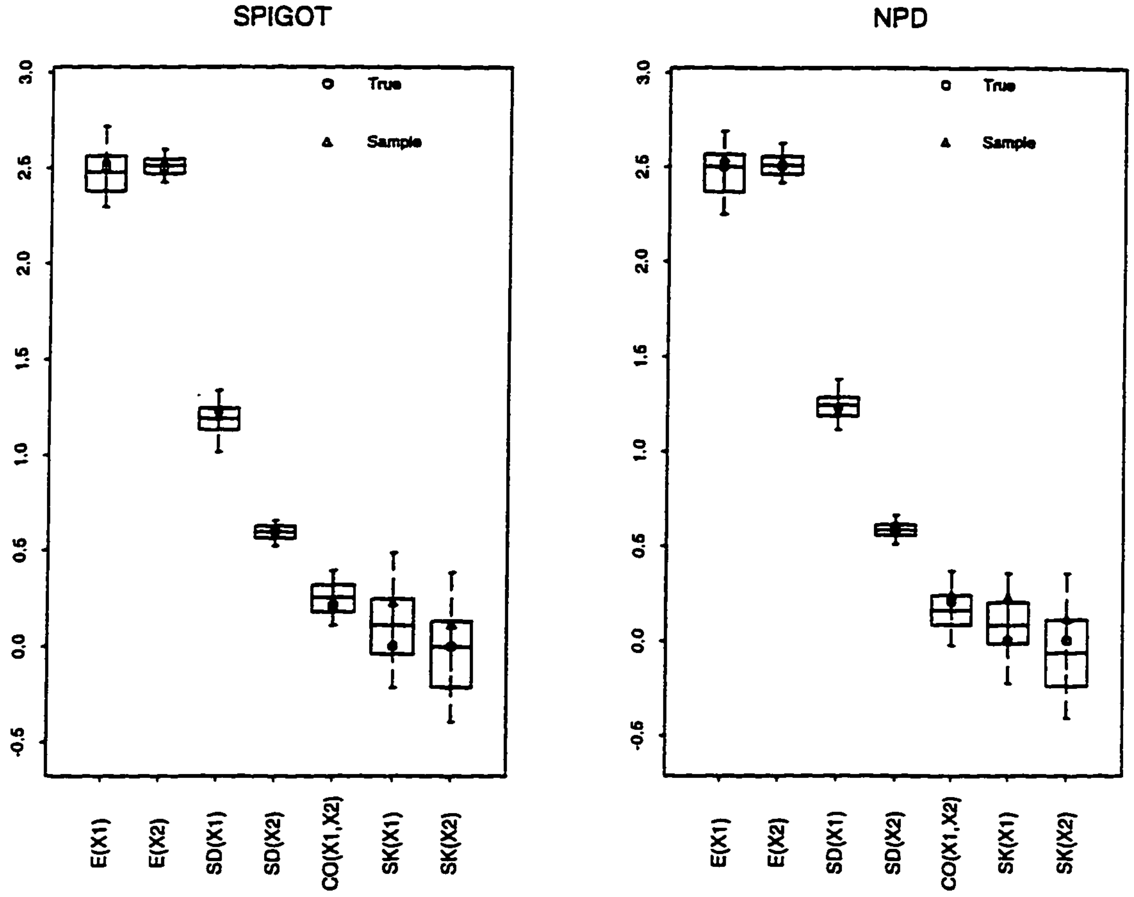
(a)

Figure 4-3. Marginal distributions for the calibration and disaggregation samples from SPIGOT and NPD as represented by boxplots for components (a) X_1 and (b) X_2. The true marginal density is obtained by integrating the p.d.f. in Figure 4-2. The calibration p.d.f. is estimated by integrating the sample joint density of variables X_1 and X_2 (a parametric distribution in case of SPIGOT, and a kernel density estimate in case of NPD). The boxes show the ranges of the univariate kernel density estimates applied to the 100 disaggregation samples with a common bandwidth chosen as the median amongst set of optimal LSCV bandwidths for each sample. The dots above the x-axis represent the calibration sample data points.



(b)

Figure 4-3. Continued.



SPIGOT

NPD

Figure 4-4. Comparison of statistics of SPIGOT and NPD for the synthetic example. The boxes show the ranges from the regenerated samples. Also shown are the true statistic (based on Figure 4-2) and the calibration sample statistic.

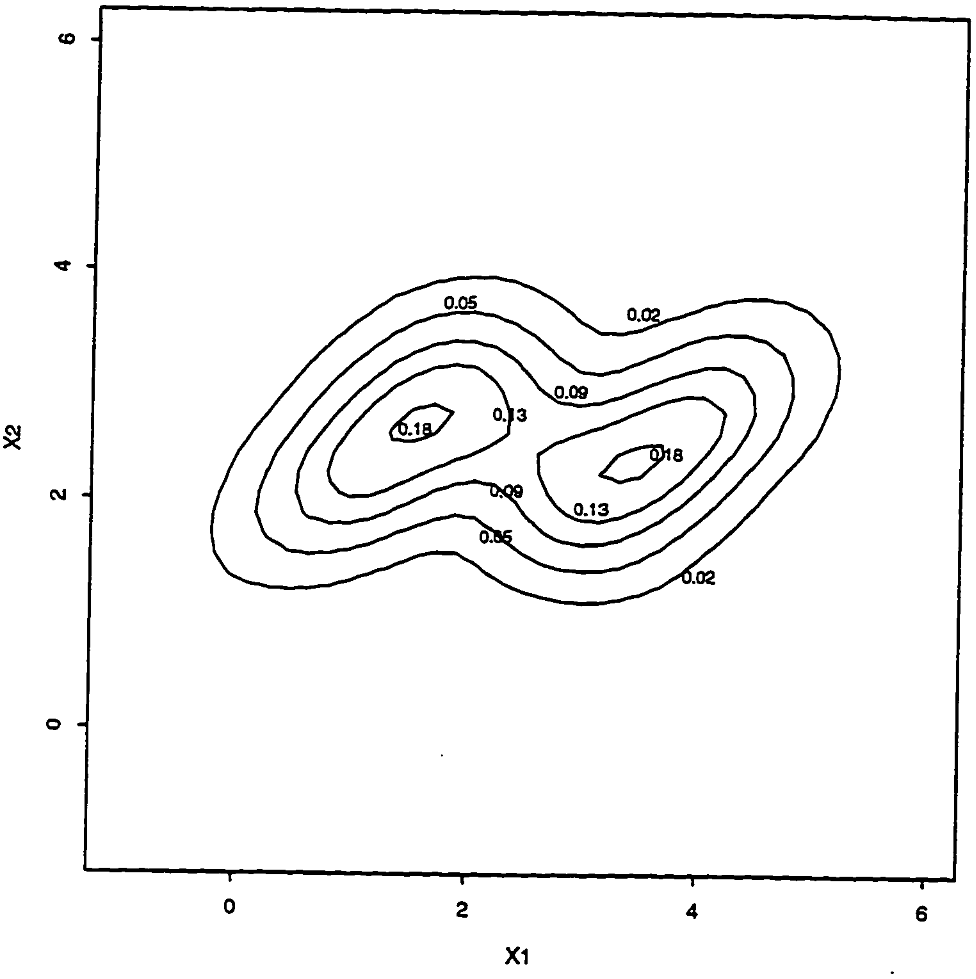


Figure 4-5. The average bivariate kernel density estimate of 101 calibration samples from the synthetic test distribution in Figure 4-2. The average ISE of these calibration joint density estimates is 0.0207.

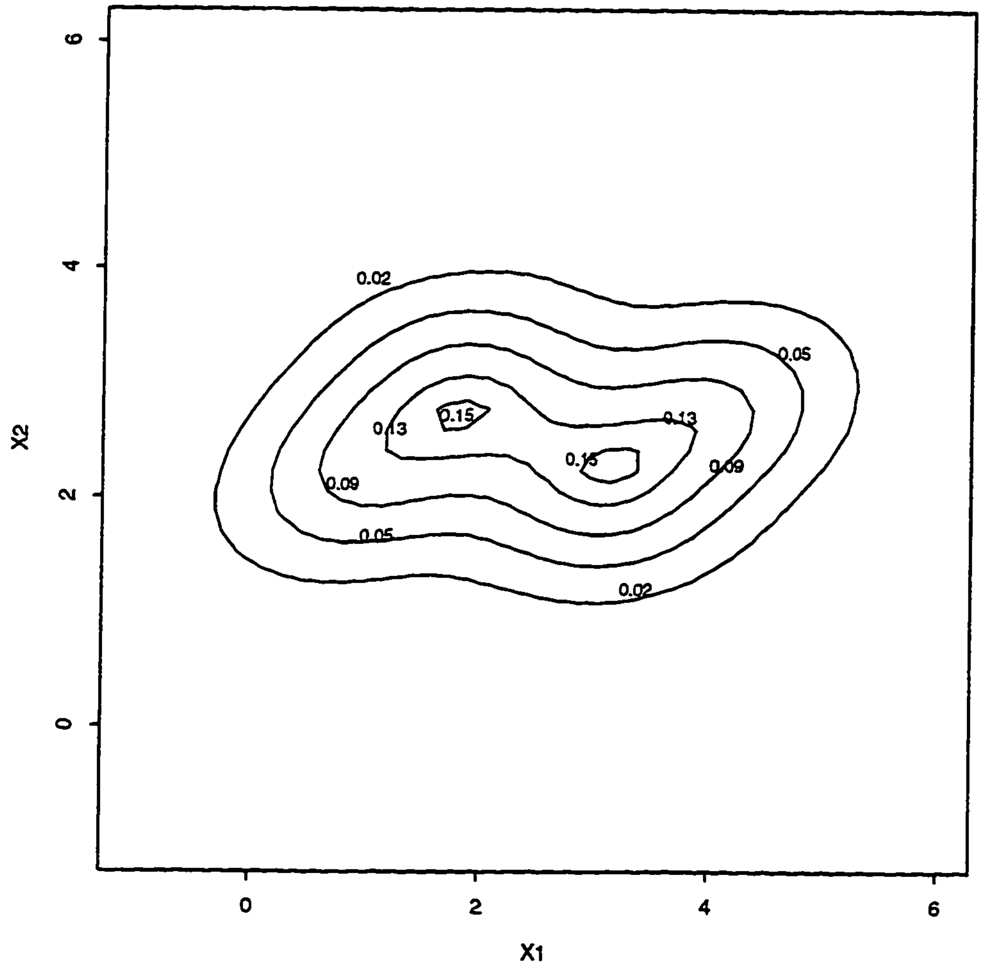


Figure 4-6. The average bivariate kernel density estimate of 10100 disaggregations from 101 calibration samples. The average ISE of these joint density estimates is 0.0339.

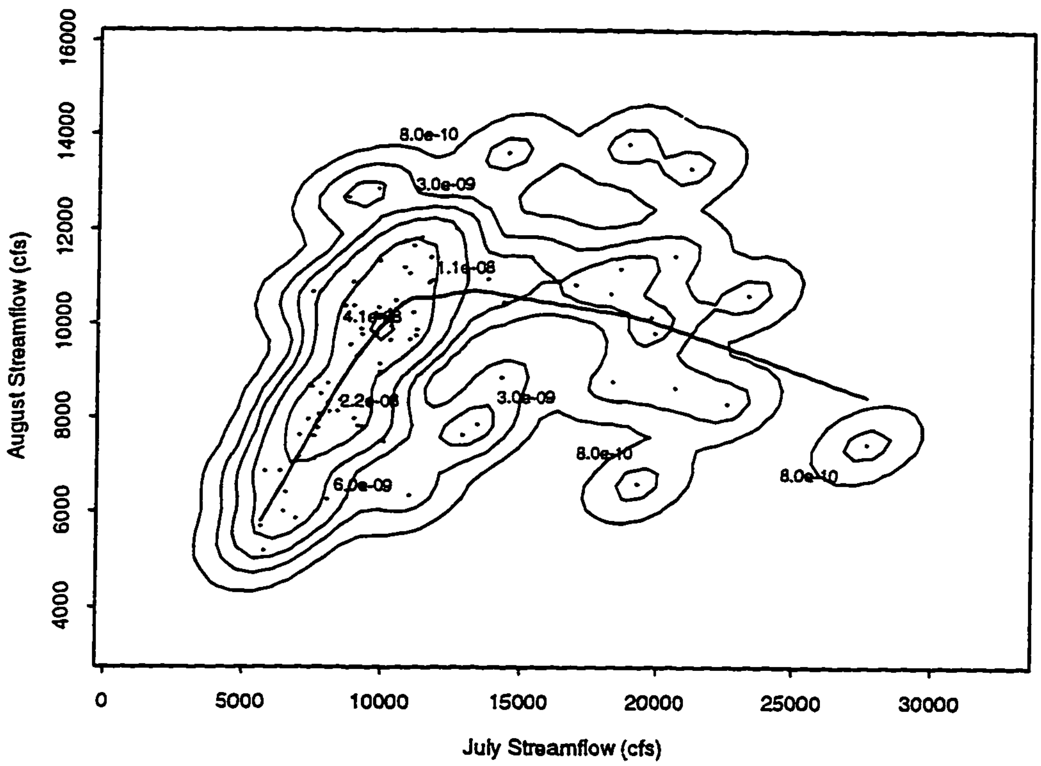


Figure 4-7. A bivariate kernel density estimate of July-August flows from the Snake River. The thick line denotes the conditional mean of August flows conditional to July flows. Lowess, a locally weighted regression smoother [*Cleveland and Devlin*, 1988], was used in our computations. Default parameter choices (number of iterations = 3, fraction of data used for smoothing at each point = 2/3) were used in the "lowess" code in the statistical package S-plus.

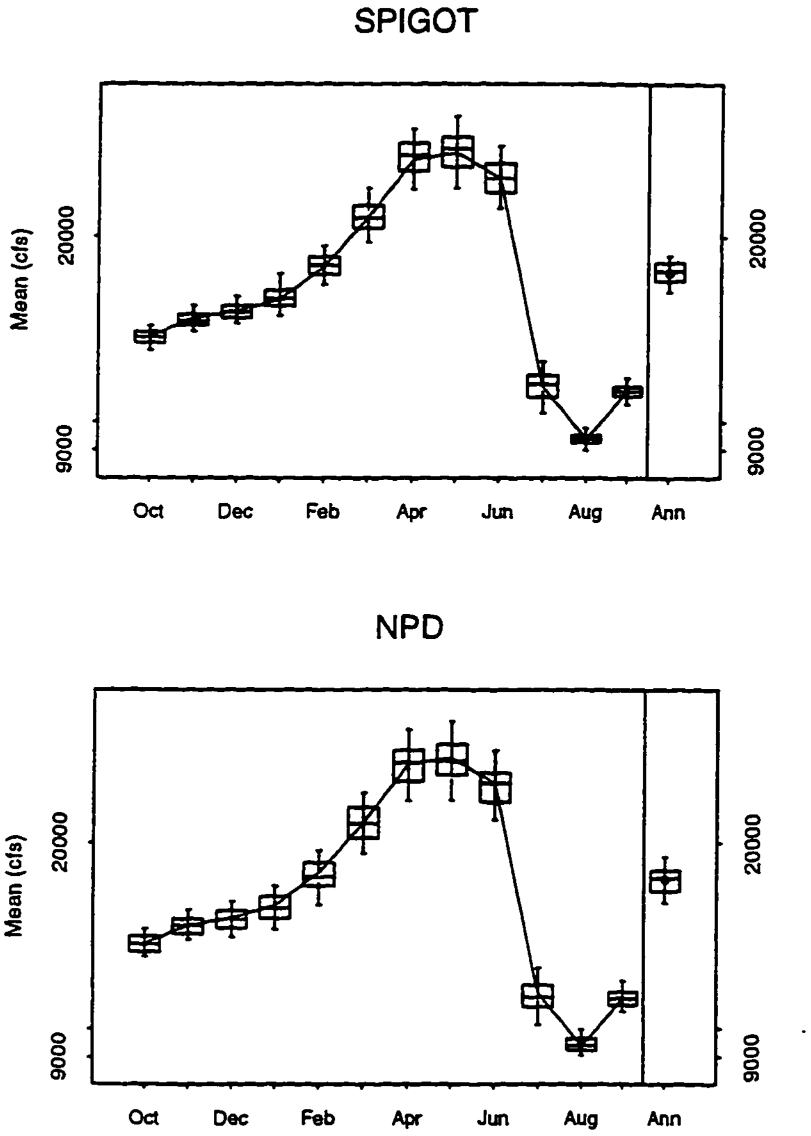
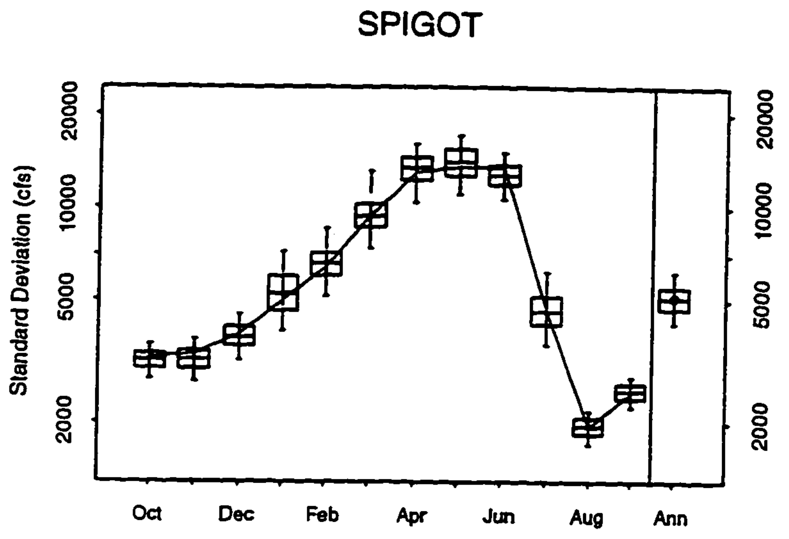


Figure 4-8. Simulated and observed seasonal and annual mean streamflow using SPIGOT and NPD. The line represents the monthly means of the historical record. The dot in the right panel is the observed annual mean.



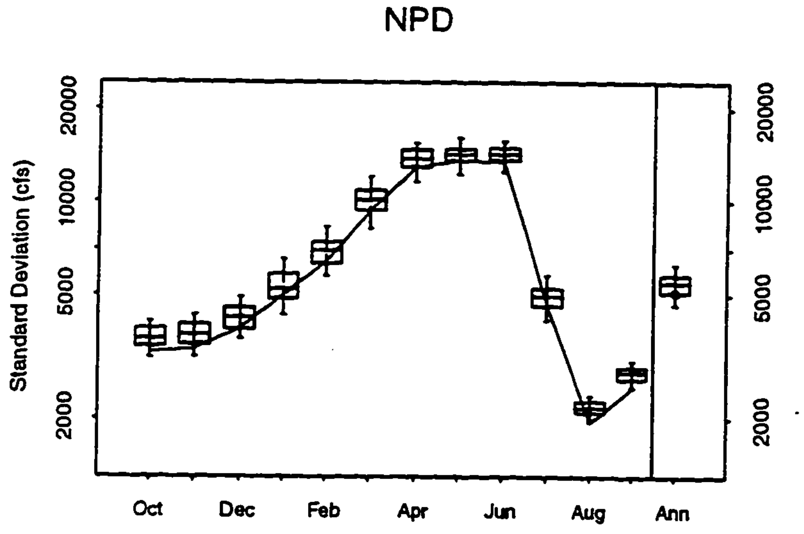


Figure 4-9. Simulated and observed streamflow standard deviations using SPIGOT and NPD. The line denotes the observed monthly standard deviations. The dot above "Ann" represents the standard deviation for the observed annual flows.

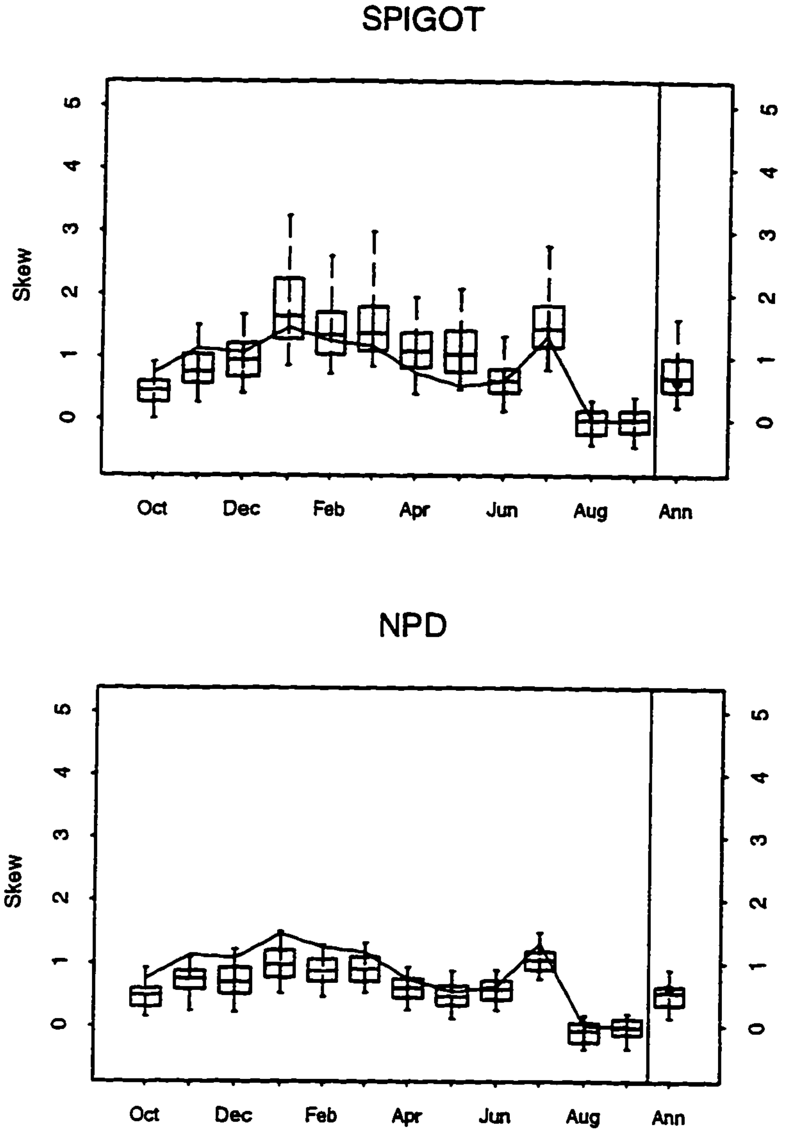
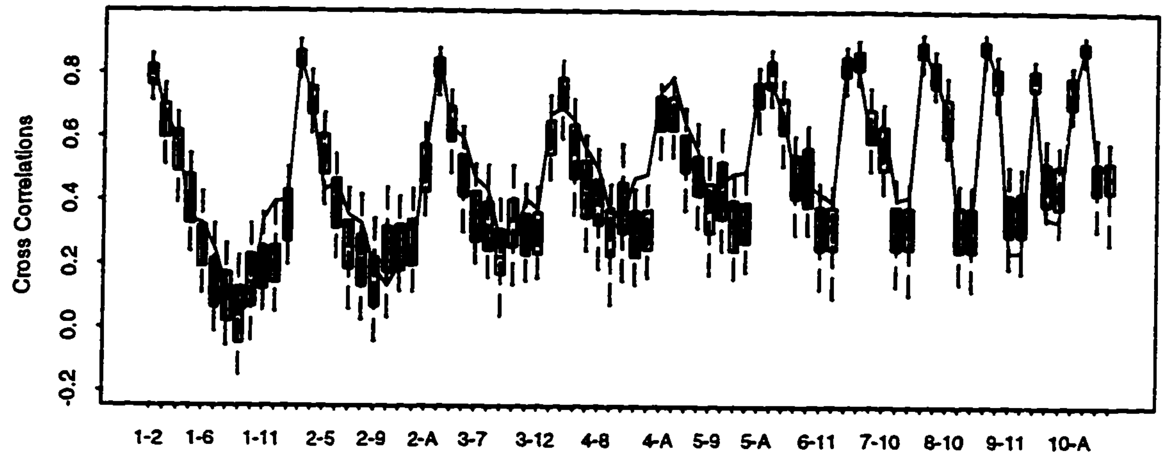


Figure 4-10. Simulated and observed streamflow skewness coefficients using SPIGOT and NPD. The line denotes the observed monthly skews. The dot above "Ann" represents the skew in the observed annual flows.

SPIGOT



NPD

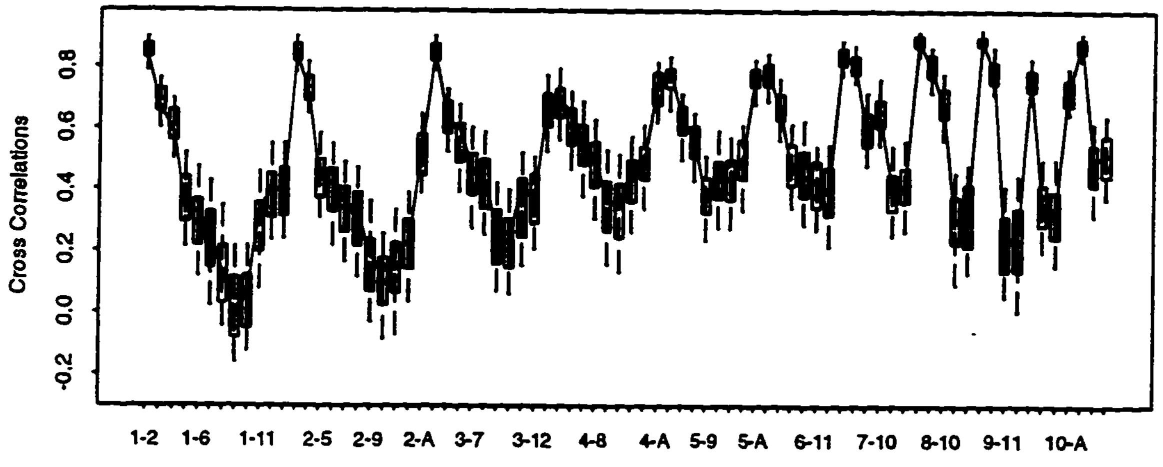
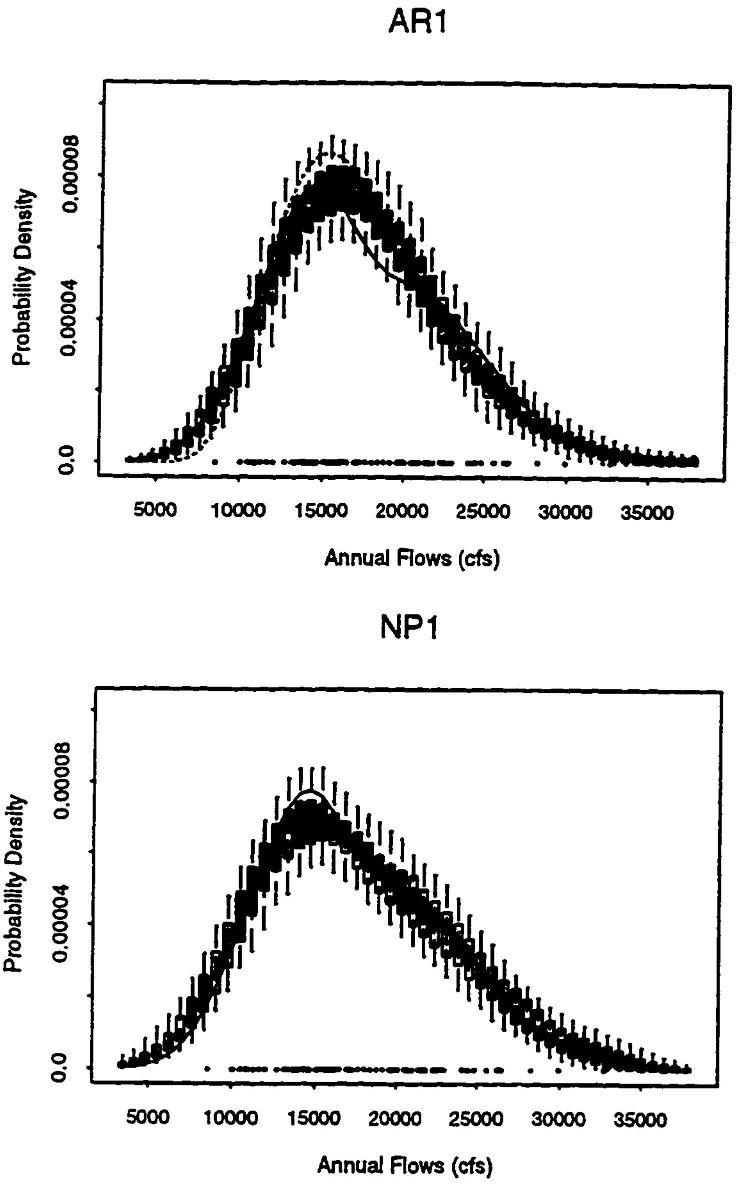
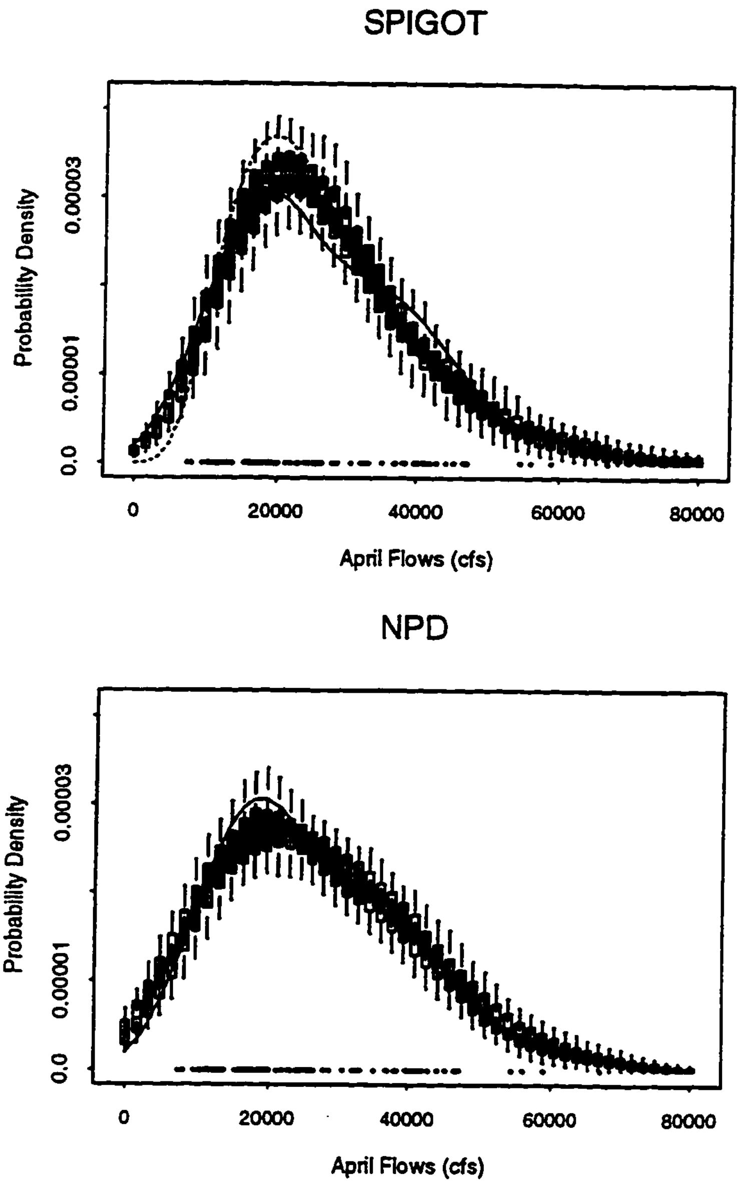


Figure 4-11. Simulated and observed cross correlation pairs using SPIGOT and NPD. The sequence along the x-axis is 1-2, 1-3,. . . , 1-12, 1-A, 2-3, 2-4 . . , 2-12, 2-A, 3-4 . . and so on. (1,2) indicates cross correlation between months 1 and 2, (1,A) indicates cross correlation between month 1 and annual aggregate.



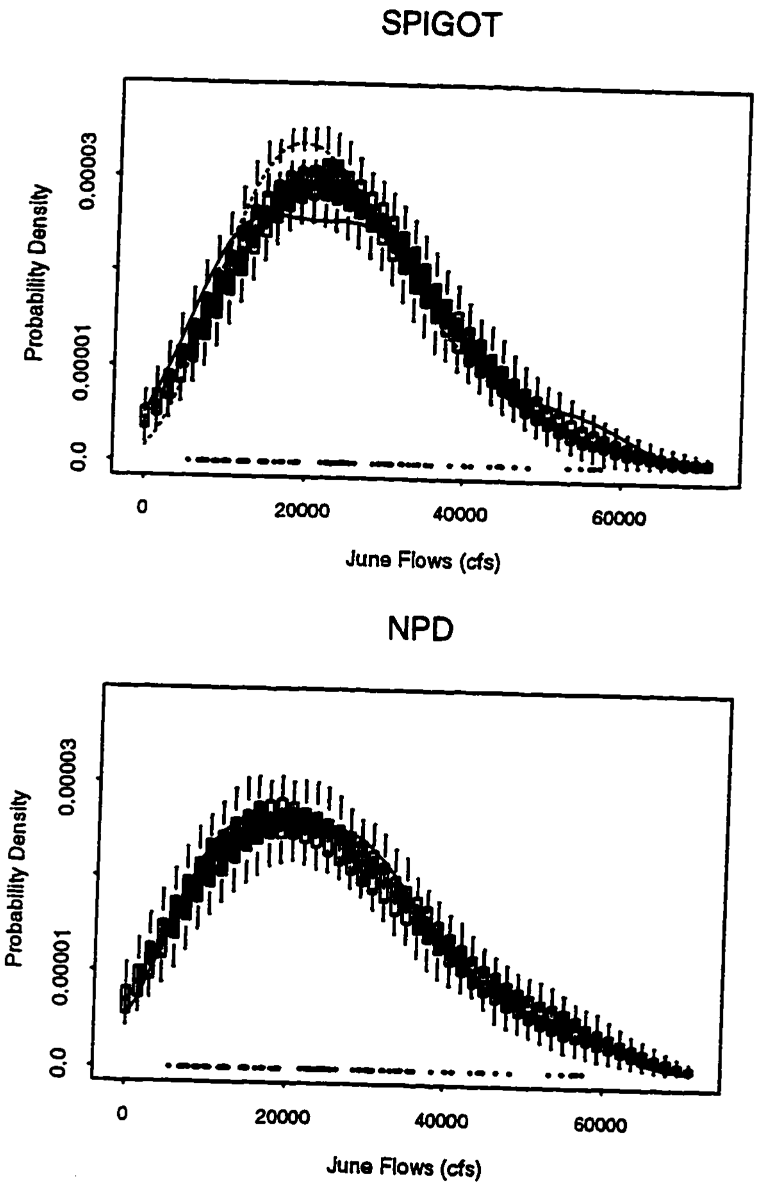
(a)

Figure 4-12. Simulated and observed marginal density estimates using the univariate kernel density estimator. The dotted line in the SPIGOT marginal densities represents the underlying density based on the Filliben correlation statistic. (a) Annual AR1 and NP1 marginal density estimates. A three-parameter Lognormal distribution is used in the AR1 model fit. (b) April SPIGOT and NPD marginal density estimates. A three-parameter Lognormal distribution is used in the SPIGOT fit for this month. (c) June SPIGOT and NPD marginal density estimates. A three-parameter Lognormal distribution is used in the SPIGOT fit for this month. (d) July SPIGOT and NPD marginal density estimates. A three-parameter Lognormal distribution is used in the SPIGOT fit for this month.



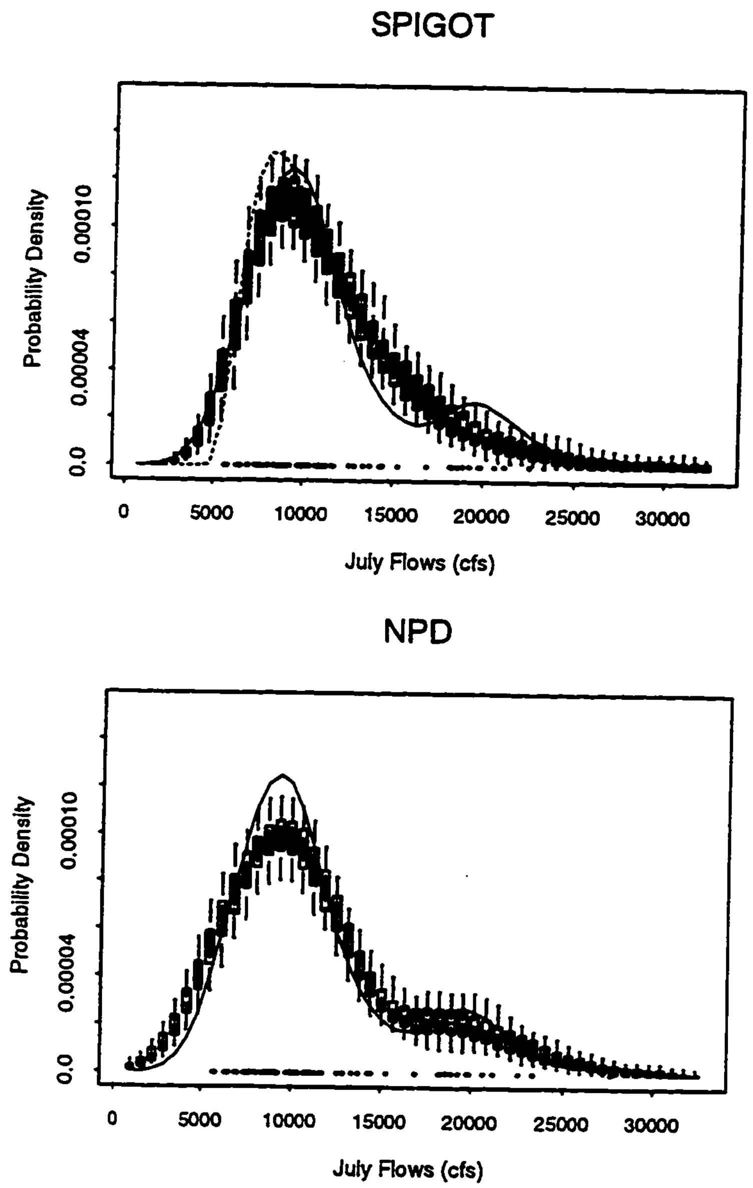
(b)

Figure 4-12. Continued.



(c)

Figure 4-12. Continued.



(d)

Figure 4-12. Continued.

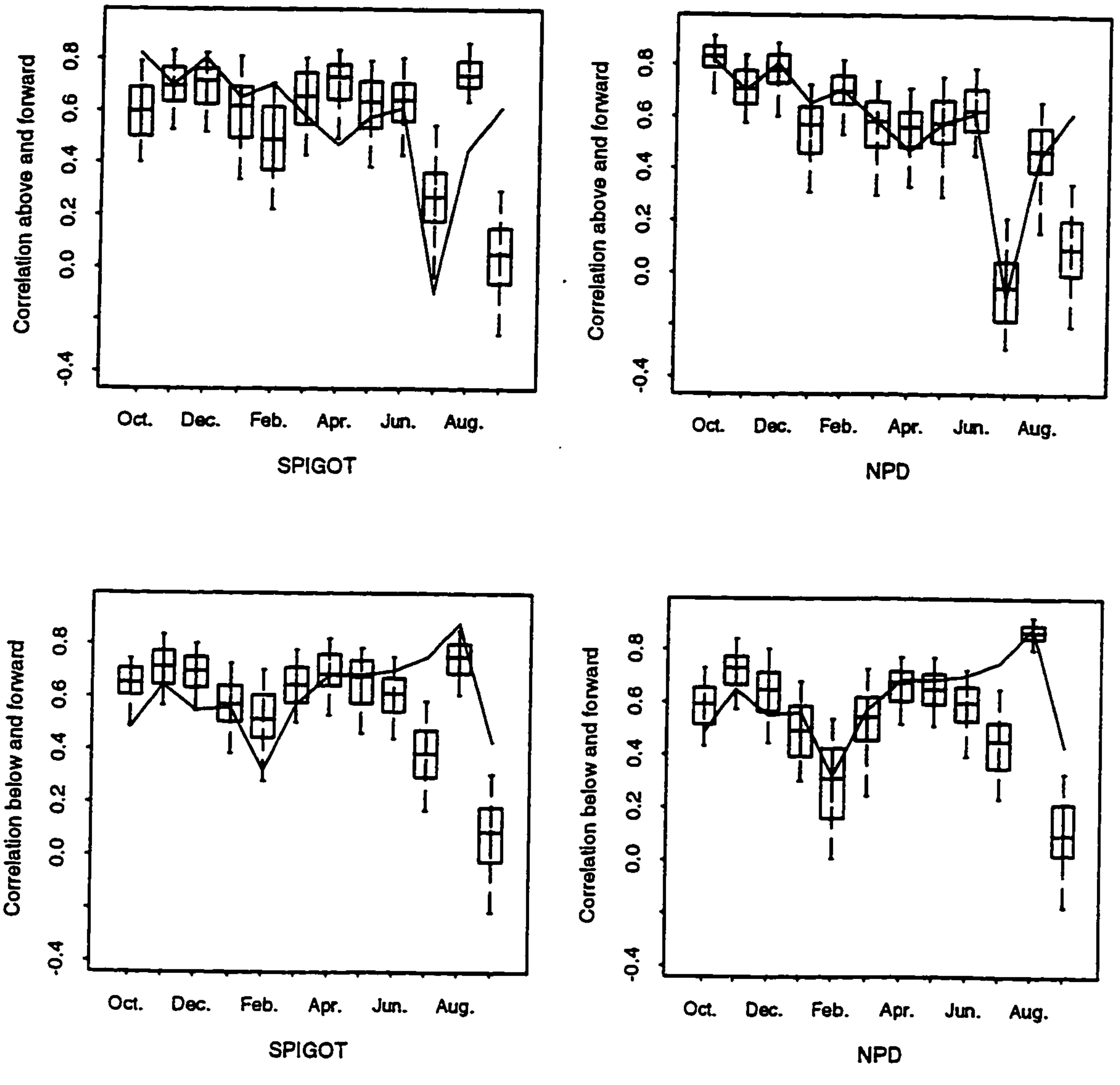


Figure 4-13. Simulated and observed state-dependent correlations for sequential month pairs using SPIGOT and NPD.

CHAPTER 5

SMALL SAMPLE PERFORMANCE OF FOUR BANDWIDTH ESTIMATORS FOR BIVARIATE KERNEL DENSITY ESTIMATION[1]

Abstract

Issues related to selection of optimal smoothing parameters for kernel density estimation with small samples (200 or fewer data points) are examined. Both reference to a Gaussian density and data-based specifications are applied to estimate bandwidths for samples from bivariate normal mixture densities. The three data-based methods studied are Maximum Likelihood Cross Validation (MLCV), Least Square Cross Validation (LSCV), and Biased Cross Validation (BCV2). Modifications for estimating optimal local bandwidths using MLCV and LSCV are also examined. The use of local bandwidths does not necessarily improve the density estimate with small samples. Of the global bandwidth estimators compared, MLCV and LSCV show lower variability and higher accuracy, while BCV2 suffers from multiple optimal bandwidths for samples from strongly bimodal densities.

Introduction

An important issue concerning kernel density estimation is the selection of optimal smoothing parameters. Bandwidth estimation for kernel density estimation and regression has been widely studied in the last decade. While several asymptotically optimal methods exist for bandwidth estimation, limited investigations of their small sample performance are available.

---

[1] Coauthored by Ashish Sharma, Upmanu Lall, and David G. Tarboton.

The aim of this chapter is primarily to share some of our experience with bandwidth selection procedures for small samples (samples less than 200 data points) in a bivariate setting. An overview of three bandwidth selection procedures is presented from a practitioner's perspective. Their performance is evaluated with samples drawn from bivariate normal mixture densities. The efficiency of local bandwidth-based estimators with estimators using a single or global smoothing parameter is also compared. The Mean Integrated Square Error (MISE) between the estimated and parent densities is used as a criterion to evaluate the performance of the bandwidth estimation method.

Kernel density estimation in a multivariate setting including a method for estimation of local bandwidths is described first. The next section describes the bandwidth estimation procedures compared in this chapter. Next, a description of the numerical experiments used to evaluate the performance of the alternative bandwidth selection procedures, is given. Results for the global bandwidth based estimators are followed by results from those that use local bandwidths.

Kernel Density Estimation

The Gaussian kernel density estimate of a d-dimensional probability density function f(x) is written as:

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\,\det(\mathbf{H})^{1/2}}\exp\left(-\frac{(\mathbf{x}-\mathbf{x}_i)^T\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)}{2}\right) \tag{5.1}$$

where n is the number of observed vectors $\mathbf{x}_i$ and **H** is a bandwidth matrix that must be from the class of symmetric, positive-definite d x d matrices [*Wand and Jones*, 1994]. The above density estimate is formed by summing Gaussian kernels with a covariance matrix **H**, centered at each observation $\mathbf{x}_i$. The bandwidth matrix **H** here is analogous to the covariance matrix for a multivariate Gaussian density. It can be specified in several ways. *Wand and Jones* [1993] suggest three ways to parameterize **H** for the bivariate case (d = 2). These are:

$$\mathbf{H} = h^2\,\mathbf{I}: h > 0 \tag{5.2a}$$

$$\mathbf{H} = \mathrm{diag}\,(h_1^2, h_2^2): h_1, h_2 > 0 \tag{5.2b}$$

$$\mathbf{H} = \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix} : h_1, h_2 > 0,\ |h_{12}| < h_1 h_2 \tag{5.2c}$$

where **I** represents the identity matrix and h, h_1, etc. are elements of the bandwidth matrix **H**. The first case (Equation 5.2a) represents a spherical kernel (with circular contours). The addition of extra smoothing parameters in the second case (Equation 5.2b) makes the kernel elliptical, though aligned parallel to the coordinate axes. The last case (Equation 5.2c) represents an elliptical kernel aligned in a direction dictated by the cross diagonal term (h_{12}). The number of parameters required in (5.2c) can be reduced by considering the following parameterization, first proposed by *Fukunaga* [1972].

$$\mathbf{H} = \lambda^2\,\mathbf{S} \tag{5.3}$$

Here, **S** is the sample covariance matrix of the data and λ^2 prescribes the bandwidth relative to this estimate of scale. Use of this parameterization amounts to transforming the data so that the sample covariance matrix is the identity. This procedure is called "sphering" [*Fukunaga*, 1972] and ensures that all kernels are oriented along the principal components of the covariance matrix. Such a parameterization has significant advantages when the variables are strongly correlated. *Wand and Jones* [1993] demonstrate the utility of this method for bivariate Gaussian data but note that densities having multiple modes in one of the coordinate directions may be poorly estimated. We have used the estimator in (5.1) with the bandwidth matrix described in (5.3) in all our results in sections 3 and 4. This estimator was chosen primarily because it requires the optimization of only one parameter (λ) while still providing an elliptical kernel oriented as dictated by the sample correlation. For ease of reference, his factor (λ) is called the bandwidth in the rest of the chapter.

Optimization of the bandwidth is subject to an appropriate criterion. While criteria such as the Integrated Square Error (ISE) and the Mean Integrated Square Error (MISE) give a estimate of the overall or global goodness of fit, the Mean Square Error (MSE) gives a pointwise measure of the error in the kernel density estimate. Minimization of the MSE can be used to estimate optimal pointwise or local bandwidths. Assigning individual bandwidths to each data point (a bandwidth λ_i corresponding to data point $\mathbf{x}_i$) is an effective specification for local bandwidths. Density estimation based on this specification can proceed using $\mathbf{H}_i = \lambda_i^2\, \mathbf{S}$ instead of $\mathbf{H}$ in Equation (5.1). Such additional flexibility in the bandwidth may be used to restrict the smoothing imposed by the kernel function in regions of high density or high curvature, where there are a lot of data points, or where the smoothing introduces bias. Conversely, in regions of low density the bandwidth can be increased to average over more points.

Abramson [1982] proposed a method for estimating local bandwidths by considering the Taylor series expansion of the squared bias of $\hat{f}(\mathbf{x})$. The second-order term in the Taylor series can be eliminated if the local bandwidth is inversely proportional to the square root of true density. The implementation of Abramson's method by *Silverman* [1986] involves perturbing an appropriate global or fixed bandwidth λ_p (denoted as the pilot global bandwidth to distinguish from λ) into a sequence of local bandwidths λ_i at each observation $\mathbf{x}_i$. This implementation (which we shall call the Abramson-Silverman method) can be stated as:

$$\lambda_i = \lambda_p \, (\hat{f}_p(\mathbf{x}_i) / g)^{-1/2} \tag{5.4}$$

where $\hat{f}_p(.)$ is a pilot density estimate and g is the geometric mean of $\hat{f}_p(\mathbf{x}_i)$ at data points $\mathbf{x}_i$. The pilot density specifies the amount of perturbation the local bandwidths receive and may be estimated using any acceptable scheme. In results presented later, a kernel density estimate using the global bandwidth λ_p as the pilot density in (5.4) was used. Note that the inverse relationship of a local bandwidth with the estimated density in effect provides a higher bandwidth in regions of low density and a lower bandwidth where the density is high. Several authors have noted that local bandwidths need to be "clipped" or restricted to lie within certain upper and lower bounds. We chose not to clip the λ_i due to the subjectivity introduced by a prescriptive choice for these upper and lower bounds. Use of local bandwidths for estimating the density, though computationally intensive, can result in certain gains as demonstrated in *Scott* [1992].

Bandwidth Estimation Methods

This section describes some methods for estimation of the optimal bandwidth. These are:

(1) Reference to a standard distribution

(2) Maximum Likelihood Cross Validation (MLCV)

(3) Unbiased or Least Square Cross Validation (LSCV)

(4) Biased Cross Validation (BCV2).

Each of these methods and their associated advantages / disadvantages are discussed, followed by modifications needed to estimate optimal local bandwidths using the Abramson-Silverman method in Equation (5.4).

Gaussian Reference Bandwidth (GREF)

The simplest automated choice for the bandwidth λ is the reference bandwidth. A reference bandwidth is optimal for an assumed (reference) distribution using an appropriate criterion. A Taylor series expansion of the MISE is used to develop expressions for the optimal reference bandwidth. *Scott* [1992] gives an expression for the first-order Taylor series approximation of the univariate MISE. This expression, using a Gaussian kernel, can be stated as:

$$\mathrm{AMISE}(h) = \frac{1}{2\sqrt{\pi}\, nh} + \frac{h^4}{4} R(f'') \qquad (5.5)$$

where AMISE stands for the Asymptotic Mean Integrated Square Error, $R(g(x)) = \int g(x)^2 dx$ (in this case $g(x) = f''(x)$) and $f''(x)$ is the second derivative of the true density. Minimization of the multivariate version of (5.5) with the true density assumed to be Gaussian results in the following expression for the AMISE optimal Gaussian reference bandwidth [*Scott*, 1992]:

$$\lambda_{\mathrm{GREF}} = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \qquad (5.6)$$

Here d and n denote the dimension and sample size, respectively. Although simple, this choice suffers from the obvious disadvantage of being optimal only for a Gaussian density. The following methods avoid assuming an underlying density and instead choose bandwidth by optimizing data-based estimates of likelihood or square error.

Maximum Likelihood Cross Validation (MLCV)

This method is a natural development of the idea of using likelihood to judge the goodness of fit of any statistical model. The use of MLCV to choose the bandwidth for kernel density estimation was proposed by *Habbema et al.* [1974] and *Duin* [1976]. The rationale behind this method is to estimate the log-likelihood of the density at observation $\mathbf{x}_i$ based on all observations except $\mathbf{x}_i$. Averaging this log-likelihood over all observations results in the following MLCV score:

$$\mathrm{MLCV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_{-i}(\mathbf{x}_i) \tag{5.7}$$

where $\hat{f}_{-i}(\mathbf{x}_i)$ denotes the density estimated from all the data points except $\mathbf{x}_i$. Using the estimator in (5.1), the MLCV score can be stated as:

$$\mathrm{MLCV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\sum_{j\neq i}^{n} \exp(-L_{ij}/2)}{(2\pi)^{d/2}\,(n-1)\,\det(\mathbf{H})^{1/2}} \right) \tag{5.8}$$

where

$$L_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T\, \mathbf{H}^{-1}\, (\mathbf{x}_i - \mathbf{x}_j) \tag{5.9}$$

and **H** is the bandwidth matrix specified in (5.3). Maximizing the score in (5.8) results in an MLCV optimal bandwidth λ_{MLCV}.

For long-tailed densities, the MLCV criterion can lead to degenerate bandwidth choices and inconsistent density estimates where a global bandwidth is used [*Silverman*, 1986]. *Schuster* [1985] recommended optimizing the MLCV function over $\mathbf{x} \varepsilon \mathbf{X}$ where **X** is an appropriate subset of the sample space that excludes the tails.

Unbiased or Least Square Cross Validation (LSCV)

This method was first proposed by *Bowman* [1984] and *Rudemo* [1982]. LSCV is based on the direct minimization of the ISE. The ISE for a multivariate density f can be expanded as:

$$\mathrm{ISE} = \int(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} = R(\hat{f}(\mathbf{x})) - 2\int \hat{f}(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} + R(f(\mathbf{x})) \tag{5.10}$$

The first term in (5.10) depends solely on the data, bandwidth, and kernel used. The last term, $R(f(\mathbf{x}))$, is independent of the bandwidth and does not need to be considered. The middle term involving the product of the true and estimated densities may be recognized as $E[\hat{f}(\mathbf{X})]$ and estimated using leave-one-out cross validation. The LSCV criterion can then be stated as:

$$\mathrm{LSCV}(\mathbf{H}) = R(\hat{f}(\mathbf{x})) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{x}_i) \tag{5.11}$$

Sain et al. [1994] provide an expression for LSCV in any dimension with multivariate Gaussian product kernels (Gaussian kernels with a diagonal **H** matrix). The LSCV score for a generalized **H** matrix (one with cross diagonal terms) can be stated as:

$$\mathrm{LSCV}(\mathbf{H}) = \frac{1 + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \left[\frac{\exp(-L_{ij}/4)}{n} - \frac{2^{d/2+1} \exp(-L_{ij}/2)}{n-1} \right]}{(2\sqrt{\pi})^{d} \det(\mathbf{H})^{1/2}} \tag{5.12}$$

where L_{ij} follows the definition in (5.9). Minimizing the LSCV score in (5.12) results in an LSCV optimal bandwidth λ_{LSCV}.

Though unbiased, bandwidth estimated using the LSCV score function has been reported to suffer from the disadvantages of a tendency to undersmooth [*Chiu*, 1990, 1991] and a high variance as compared to the BCV estimator of *Scott and Terrell* [1987]. The higher variance corresponds to a tendency for the LSCV score function to have multiple local minima and hence a tendency to undersmooth [*Chiu*, 1990]. Solutions to this problem are suggested among others by *Chiu* [1990, 1991]. These solutions, however, were not implemented in our analysis. Compared to the BCV2 score function that is described next, *Sain et al.* [1994] report that LSCV has a marginally smaller variance when applied to estimate a univariate Gaussian density.

Biased Cross Validation (BCV2)

Proposed by *Sain et al.* [1994], biased cross validation provides a data-based estimate of the AMISE in (5.5). We shall refer to this as BCV2, the notation used by *Sain et al.* in their paper. The term R(f") in Equation (5.5) is estimated using leave-one-out cross validation. For a vector **x**, this term can be stated as:

$$R(f'') = \int f''(\mathbf{x})^2\, d\mathbf{x} = \int f^{iv}(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} = E[f^{iv}(\mathbf{x})] \tag{5.13}$$

or,

$$R(f'') \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}^{iv}_{-i}(\mathbf{x}_i) \tag{5.14}$$

where $\hat{f}^{iv}_{-i}(\mathbf{x}_i)$ is the fourth derivative of the kernel density estimate at $\mathbf{x}_i$ formed by leaving out data point $\mathbf{x}_i$. *Sain et al.* [1994] provide an expression for BCV2 using Gaussian product kernels. Extending their result to the case of a general bandwidth matrix **H**, we obtain:

$$\mathrm{BCV2}(\mathbf{H}) = \frac{1}{(2\sqrt{\pi})^{d}\, n \det(\mathbf{H})^{1/2}} + \frac{\sum_{i=1}^{n} \sum_{j\neq i}\left[L_{ij} - (2d+4)L_{ij} + d^{2} + 2\,d \right]}{4n(n-1)\det(\mathbf{H})^{1/2}\,(\sqrt{2\pi})^{d}} \exp(-L_{ij}/2) \tag{5.15}$$

with L_{ij} as defined in (5.9). A global bandwidth λ_{BCV2} can be optimized by minimizing the score in (5.15).

BCV2 suffers from the problem of being a biased estimator of the optimal bandwidth. An earlier version of biased cross validation (denoted BCV by *Scott and Terrell* [1987] and BCV1 by *Sain et al.* [1994]) had the advantage of having smaller variance though being more heavily biased than BCV2. Apart from the reduction in bias, another justification for BCV2 over BCV is the relative ease with which it can be implemented in multivariate settings. *Sain et al.* [1994] conducted a simulation study to compare the performance of BCV2 with LSCV. They found that LSCV tends to have higher variance than BCV2 for a standard Gaussian density, although their derivations for the asymptotic standard deviations [*Sain et al.*, 1994] indicate otherwise. *Sain et al.* support their results by noting that LSCV tends to undersmooth, which may increase the variance unnecessarily.

Estimation of Optimal Local Bandwidths

Results in the next section use the Gaussian reference, MLCV, and LSCV, as criteria for estimation of local bandwidths. The Abramson-Silverman method (Equation 5.4) is used. Details of our implementation of the MLCV and LSCV score functions for local bandwidths are given here.

As mentioned in the previous section, local bandwidths may be estimated based on a pilot density estimate (see Equation 5.4). We have used a kernel density estimate based on a global bandwidth λ_p as the pilot density. Local bandwidths can then be estimated by perturbing λ_p as given in Equation (5.4). These steps amount to the following algorithm:

(1) Estimate the pilot density using the pilot bandwidth λ_p. Call this pilot density $\hat{f}_p(.)$.

(2) Estimate the corresponding local bandwidths λ_i using the Abramson-Silverman method in Equation (5.4).

(3) Calculate the criterion function (MLCV or LSCV) for local bandwidths λ_i.

The only specification used here is that of the global bandwidth λ_p. Silverman's suggestion can be taken to optimize the global bandwidth λ_p and then simply perturb it to a vector of local bandwidths. In this study, the target score or optimization function was computed using the density estimates based on the local bandwidths λ_i obtained upon perturbing the global λ_p. The optimal λ_p is then selected as the optimizer of the score function computed using the local bandwidths. This procedure was followed in all cases except when the Gaussian reference bandwidth is used, where that bandwidth was used directly to determine corresponding local bandwidths. The global Gaussian reference bandwidth (GREF) is denoted as $\lambda_{p_{GREF}}$.

The MLCV criterion using local bandwidths is a natural extension of Equation (5.8):

$$\mathrm{MLCV} = \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{j \neq i}^{n} \frac{\exp(-U_1/2)}{(2\pi)^{d/2}\,(n-1)\,\det(\mathbf{H}_j)^{1/2}} \right) \tag{5.16}$$

where:

$$U_1 = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}_j^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{5.17}$$

$\mathbf{H}_j$ equals λ_j^2 S in this specification. Optimization proceeds by maximizing the score in (5.16) with respect to the pilot bandwidth λ_p. The MLCV optimal pilot bandwidth is denoted $\lambda_{p_{MLCV}}$.

An equivalent of the LSCV in (5.12) for local bandwidths is:

$$\mathrm{LSCV} = \frac{1}{(2\sqrt{\pi})^d\, n} \sum_{i=1}^{n} \left\{ \frac{1}{\det(\mathbf{H}_i)^{1/2}} + 2^{d/2} \sum_{j \neq i} \left[\frac{\exp(-U_2/2)}{n \det(\mathbf{H}_i + \mathbf{H}_j)^{1/2}} - \frac{2 \exp(-U_1/2)}{(n-1)\det(\mathbf{H}_j)^{1/2}} \right] \right\} \tag{5.18}$$

where U_1 follows the definition in (5.17) and U_2 equals:

$$U_2 = (\mathbf{x}_i-\mathbf{x}_j)^T (\mathbf{H}_i+\mathbf{H}_j)^{-1}(\mathbf{x}_i-\mathbf{x}_j) \tag{5.19}$$

As before, $\mathbf{H}_i$ equals λ_i^2 S for this specification. Note that Equation (5.18) reduces to Equation (5.12) when $\mathbf{H}_i = \mathbf{H}_j = \mathbf{H}$. Optimization proceeds by minimizing Equation (5.18) with respect to the pilot bandwidth λ_p. The LSCV optimal pilot bandwidth is referred to as $\lambda_{p_{LSCV}}$.

Four global bandwidth selectors GREF, MLCV, LSCV, and BCV2, were discussed in this section. Extensions for local bandwidths were developed for MLCV and LSCV. Results using these procedures applied to samples from selected Gaussian mixture densities are provided in the next section.

Application to Samples from Gaussian Mixture Densities

Here we evaluate the performance of bandwidth estimation methods described in the previous section. These methods were tested with samples drawn from mixtures of bivariate Gaussian densities. Small samples (less than 200 data points for each bivariate sample) were considered. We chose the class of Gaussian mixture densities since exact results for the MISE and AMISE given a bandwidth matrix (or matrices if local bandwidths are used) are known [*Wand and Jones*, 1995]. We used the Numerical Recipes function BRENT [*Press et al.*, 1989] to optimize the bandwidth (λ or λ_p) in the range $\lambda_{GREF}/10$ to $2\lambda_{GREF}$. This optimization function is based on inverse parabolic interpolation that ensures convergence given an initial range in which the minimum can be found. Several spot checks indicated that the minimum found by this routine was indeed a global minimum. The performance of each estimator was measured by recording the ISE between the true and estimated densities for each sample. The ISE was then averaged over the samples to get a measure of the MISE. The fact that the integral of the product of two Gaussians can be represented as a single Gaussian, as reported in *Wand and Jones* [1993], reduced the ISE calculations considerably. It should be noted that since our samples are drawn from Gaussian mixture densities, each term in the ISE in (5.10) can be represented as a sum of Gaussian p.d.f.'s.

Results for three sample sizes (50, 100, and 200) from the four distributions listed in Table 5-1 are evaluated. These distributions were chosen based on a list of bivariate distributions in *Wand and Jones* [1993]. Contour plots of these p.d.f.'s are illustrated in Figure 5-1. Our testing procedure involved drawing 50 samples of sizes 50, 100, and 200 from each test distribution. The ISE was used to evaluate the performance of various methods and was estimated by integrating the square of the difference in the true and estimated densities. Also presented is the bandwidth that minimizes the exact MISE for each of the test densities. Estimates of this exact MISE are based on relations in *Wand and Jones* [1993]. Results for global bandwidth estimators are followed by those using local bandwidths as described in section 3.5.

Results Using a Global λ

ISE's were computed as a function of the optimal global bandwidth for each sample. Table 5-2 shows the average of these ISE scores for all test densities and sample sizes. Histograms of the optimal bandwidths for each of the test densities are illustrated in Figures 5-2 (density A), 5-3 (density B), 5-4 (density C), and 5-5 (density D). Also shown are the MISE optimal bandwidth and the Gaussian reference bandwidth for each density and sample size. The bias and standard deviation of the optimal bandwidths (from

MLCV, LSCV and BCV2) are presented in Table 5-3.

As expected, the reference Gaussian bandwidth results in the minimum MISE (see Table 5-2) for density A (the standard normal p.d.f.) for all sample sizes, with MLCV coming a close second. Higher MISE's for BCV2 and LSCV are partly due to the greater variability in optimal bandwidths from both methods (see Figure 5-2 and Table 5-3). It is notable that MLCV (a method that is not based on minimizing an L2 measure of error) gives a lower variance and a smaller MISE than other cross validation-based methods. A low bias in all methods is apparent from Table 5-3.

Density B is bimodal with the modes aligned along one coordinate axis. Note how close the reference bandwidth lies to the MISE optimal bandwidth for samples of size 50 from this density. This reflects in the ISE scores too, with the Gaussian reference bandwidth resulting in the best MISE amongst all methods. Poor performance for BCV2 is apparent and is due to the high variability in its optimal bandwidths (see Figure 5-3 and Table 5-3). MLCV again gives the best results amongst all cross validation methods.

Density C is more distinctly bimodal than density B. Gaussian reference is a poor choice for the bandwidth for this density (see Figure 5-4). Both MLCV and LSCV prove comparable and perform better than other methods. Both present comparable standard deviations and ISE scores, though LSCV shows a lower bias than MLCV. BCV2 proves disappointing for this density. The histogram showing BCV2 optimal bandwidths in Figure 5-4 indicates the presence of two distinct modes for BCV2 optimal bandwidths. Considering the fact that our bandwidth search procedure is not permitted to extend beyond the limit of the plot, the standard deviation for this method would actually be higher than what is reported in Table 5-3. It is notable, though, that the BCV2 results for sample size 200 show a distinct improvement (and a distinct mode in the histogram) over other sample sizes.

Figure 5-5 shows a large difference in the Gaussian reference and the MISE optimal bandwidths for density D. Much like the earlier case, the Gaussian reference choice compares poorly with the other methods. Multiple modes are again evident for BCV2 (for n = 50 and 100), hence the higher standard deviation (see Table 5-3). Variability in the optimal bandwidth and the associated ISE is lowest for MLCV, followed closely by LSCV.

On the whole, BCV2 does not appear to be a stable estimator of the optimal bandwidth. Amongst other choices, while Gaussian reference is a fairly good guess for Gaussian or near Gaussian data, MLCV and LSCV are the two most consistent bandwidth estimators amongst the ones studied.

Results Using Local Bandwidths

Results from using the optimal local bandwidth estimators developed earlier based on the Abramson-Silverman method (Equation 5.4) are given here. Table 5-4 shows the mean ISE for all test densities and sample sizes using the Gaussian reference, MLCV, and LSCV criteria for selecting local bandwidths. Although there are n local bandwidths λ_i, they are all keyed to a single global pilot bandwidth λ_p through Equation (5.4). We present results for these pilot bandwidths in our comparisons. Histograms of optimal pilot bandwidths for each of the four test densities are illustrated in Figures 5-6 (density A), 5-7 (density B), 5-8 (density C), and 5-9 (density D). The Gaussian reference bandwidth (Equation 5.8) is also shown. Standard deviations of the optimal pilot bandwidths (using MLCV and LSCV) are presented in Table 5-5.

The ISE scores in Table 5-4 indicate no improvement over global bandwidth based estimators. Density A shows that LSCV performs marginally better than MLCV for sample size 100 and 200, though both are much worse than the global bandwidth results presented in Table 5-2. This could be due in part to the higher variability of the optimal pilot density (see Figure 5-6 and Table 5-5).

GREF is picked over MLCV and LSCV for Density B, though no improvement over global bandwidth estimators is visible. A higher variability in the pilot bandwidths (see Figure 5-7) could again be the cause for this performance.

Results from density C are slightly reassuring. The Gaussian reference bandwidth when used as the pilot bandwidth shows better performance than its global counterpart. This is, however, countered by the fact that GREF is the worst choice (except BCV2) for density C in Table 5-2 and is significantly higher than the optimal bandwidths picked by MLCV or LSCV (see Figure 5-8). MLCV and LSCV pick the best local bandwidths though there are no improvements over global MLCV or LSCV choices.

Comparing to Table 5-2, improvements in the ISE with GREF are evident for density D. It, however, heavily oversmooths (see Figure 5-9) as compared to the MLCV or LSCV optimal bandwidths, which are again inferior to their global counterparts.

On the whole, using local bandwidths did not result in any improvements over global bandwidth estimators. This was contrary to our expectations.

Conclusions and Recommendations

Four methods for bandwidth estimation were evaluated. Results using both global and local bandwidths were compared. Although the list of test distributions over which these methods were compared is limited, we feel that these results are representative of plausible practical cases where limited data are available, and do provide guidance and insight for the practitioner.

The conclusions can be summarized as follows:

(1) Local bandwidth estimation methods did not improve MISE performance relative to methods using a global smoothing parameter for the sample sizes and test distributions used.

(2) Gaussian reference can be a good choice when the densities are weakly Gaussian (as in density A and B in Table 5-1). Use of the Gaussian reference in other cases can lead to significant oversmoothing of the density estimate.

(3) Biased Cross Validation (BCV2) does a poor job in estimating the optimal bandwidth, more so for sample size 50 or 100 than for size 200. One disturbing aspect of this method is the high variability of BCV2 optimal bandwidths. Two distinct peaks were visible in the histograms for BCV2 optimal bandwidths for densities C and D. This small sample performance of BCV2 is disappointing and contrary to the results in *Sain et al.* [1994]. By choosing an estimate that is biased towards oversmoothing, one hopes that the variance in bandwidth selection can be reduced. An increase in the variance of LSCV has been demonstrated in the past due to occasional extreme undersmoothing. While BCV2 appears to avoid such extreme undersmoothing, its choice of λ for $n < 200$ appears to be more diffuse than that from other cross validation criteria. The high variance and bias lead to a poor performance overall in the simulations presented here.

(4) LSCV and MLCV both perform well. However, given the computational simplicity of MLCV, we recommend MLCV over LSCV for use on small samples. Cautionary notes on the consistency of MLCV with fixed bandwidths and long-tailed densities do, however, apply.

References

Abramson, I. S., On bandwidth variation in kernel estimates-a square root law, *Ann. Stat.*, 10(4), 1217-1223, 1982.

Bowman, A. W., An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, 71(2), 353-360, 1984.

Chiu, S. T., Why bandwidth selectors tend to choose smaller bandwidths, and a remedy, *Biometrika*, 77(1), 222-6, 1990.

Chiu, S. T., Bandwidth selection for kernel density estimation, *Ann. Stat.*, 19(4), 1883-1905, 1991.

Duin, R. P. W., On the choice of smoothing parameters for parzen estimators of probability density functions, *IEEE Transactions on Computers*, 1175-1179, 1976.

Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.

Habbema, J. D. F., J. Hermans, and K. Vandenbroek, A stepwise discriminant analysis program using density estimation, in *COMPSTAT 1974*, edited by G. Bruckmann, pp. 101-110, Physica Verlag, Vienna, Austria, 1974.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing (Fortran Version)*, Cambridge University Press, New York, 1989.

Rudemo, M., Empirical choice of histograms and kernel density estimators, *Scand. J. Stat.*, 9, 65-78, 1982.

Sain, S. R., K. A. Baggerly, and D. W. Scott, Cross-validation of multivariate densities, *J. Am. Stat. Assoc.*, 89(427), 807-817, 1994.

Schuster, E. F., Incorporating support constraints into nonparametric estimators of densities, *Commun. Statist.-Theor. Meth.*, 14(5), 1123-1136, 1985.

Scott, D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, 1992.

Scott, D. W., and G. R. Terrell, Biased and unbiased cross-validation in density estimation, *J. Am. Stat. Assoc.*, 82(400), 1131-1146, 1987.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, England, 1986.

Wand, M. P., and M. C. Jones, Comparison of smoothing parameterizations in bivariate kernel density estimation, *J. Am. Stat. Assoc.*, 88(422), 520-528, 1993.

Wand, M. P., and M. C. Jones, Multivariate plug-in bandwidth selection, *Comp. Stat.*, 9, 97-116, 1994.

Wand, M. P., and M. C. Jones, *Kernel Smoothing*, Chapman and Hall, London, England, 1995.

Table 5-1. Description of the Bivariate Normal Mixture Densities Studied

Density	$w_1 N(\mu_1, \mu_2, \sigma_{11}^2, \sigma_{12}^2, \rho_1) + \ldots + w_k N(\mu_k, \mu_k, \sigma_{k1}^2, \sigma_{k2}^2, \rho_k)$
(A) Standard	N(0, 0, 1, 1, 0)
(B) Bimodal	1/2 N(-1, 0, $(2/3)^2$, $(2/3)^2$, 0) + 1/2 N(1, 0, $(2/3)^2$, $(2/3)^2$, 0)
(C) Bimodal	1/2 N(1, -1, $(2/3)^2$, $(2/3)^2$, 7/10) + 1/2 N(-1, 1, $(2/3)^2$, $(2/3)^2$, 0)
(D) Bimodal	0.4 N(-1.2, 0, $(3/5)^2$, $(3/5)^2$, 7/10) + 0.4 N(1.2, 0, $(3/5)^2$, $(3/5)^2$, 7/10) + 0.2 N(0, 0, $(3/5)^2$, $(3/5)^2$, -7/10)

In the notation used, N(.) refers to a bivariate Gaussian distribution with mean (μ_{j1}, μ_{j2}), variance (σ_{j1}^2, σ_{j2}^2) and correlation ρ_j where j ranges from 1 to k, k being the number of mixtures used. $w_1, \ldots, w_k$ denote the weights for individual Gaussian p.d.f.'s.

Table 5-2. Mean ISE for Global Bandwidth-Based Estimators

Density	Method	n = 50	n = 100	n = 200
(A)	GREF	0.0075	0.0048	0.0031
	MLCV	0.0077	0.0050	0.0032
	LSCV	0.0089	0.0054	0.0038
	BCV2	0.0083	0.0057	0.0036
(B)	GREF	0.0109	0.0077	0.0049
	MLCV	0.0119	0.0082	0.0050
	LSCV	0.0134	0.0089	0.0053
	BCV2	0.0145	0.0095	0.0054
(C)	GREF	0.0331	0.0263	0.0208
	MLCV	0.0259	0.0175	0.0109
	LSCV	0.0261	0.0173	0.0107
	BCV2	0.0467	0.0293	0.0107
(D)	GREF	0.0274	0.0215	0.0158
	MLCV	0.0253	0.0180	0.0116
	LSCV	0.0286	0.0186	0.0119
	BCV2	0.0369	0.0227	0.0122

Table 5-3. Sample Bias and Standard Deviation of Estimated Optimal Global Bandwidths

		n = 50		n = 100		n = 200	
Density	Method	$E(\lambda-\lambda_{MISE})$	$\sigma(\lambda)$	$E(\lambda-\lambda_{MISE})$	$\sigma(\lambda)$	$E(\lambda-\lambda_{MISE})$	$\sigma(\lambda)$
(A)	MLCV	0.0493	0.0681	0.0192	0.0613	0.0273	0.0404
	LSCV	0.0107	0.1151	0.0081	0.0771	-0.0216	0.0774
	BCV2	0.0510	0.1163	0.0023	0.1115	0.0057	0.0909
(B)	MLCV	0.0362	0.0793	0.0319	0.0596	0.0364	0.0415
	LSCV	0.0435	0.1205	0.0271	0.0836	0.0326	0.0619
	BCV2	0.0979	0.2034	0.0489	0.1370	0.0214	0.0741
(C)	MLCV	0.0664	0.0586	0.0472	0.0359	0.0315	0.0277
	LSCV	0.0125	0.0687	-0.0012	0.0411	0.0036	0.0259
	BCV2	0.5116	0.3318	0.2488	0.3153	0.0113	0.0329
(D)	MLCV	0.0273	0.0652	0.0224	0.0407	0.0298	0.0335
	LSCV	0.0186	0.0959	0.0067	0.0483	0.0016	0.0362
	BCV2	0.2604	0.3175	0.0619	0.1866	0.0002	0.0501

Table 5-4. Mean ISE for Local Bandwidth-Based Estimators

Density	Method	n = 50	n = 100	n = 200
(A)	GREF	0.0089	0.0054	0.0039
	MLCV	0.0090	0.0067	0.0062
	LSCV	0.0150	0.0058	0.0060
(B)	GREF	0.0129	0.0086	0.0051
	MLCV	0.0163	0.0117	0.0072
	LSCV	0.0181	0.0115	0.0059
(C)	GREF	0.0303	0.0223	0.0158
	MLCV	0.0268	0.0190	0.0123
	LSCV	0.0327	0.0219	0.0117
(D)	GREF	0.0277	0.0203	0.0136
	MLCV	0.0319	0.0231	0.0152
	LSCV	0.0372	0.0233	0.0147

Table 5-5. Sample Standard Deviations of Estimated Optimal λ_p

Density	Method	n = 50	n = 100	n = 200
(A)	MLCV	0.1086	0.0926	0.0759
	LSCV	0.1892	0.1324	0.1572
(B)	MLCV	0.0843	0.0667	0.0471
	LSCV	0.1648	0.1117	0.0785
(C)	MLCV	0.0470	0.0295	0.0231
	LSCV	0.0771	0.0606	0.0389
(D)	MLCV	0.0583	0.0327	0.0268
	LSCV	0.1082	0.0635	0.0501

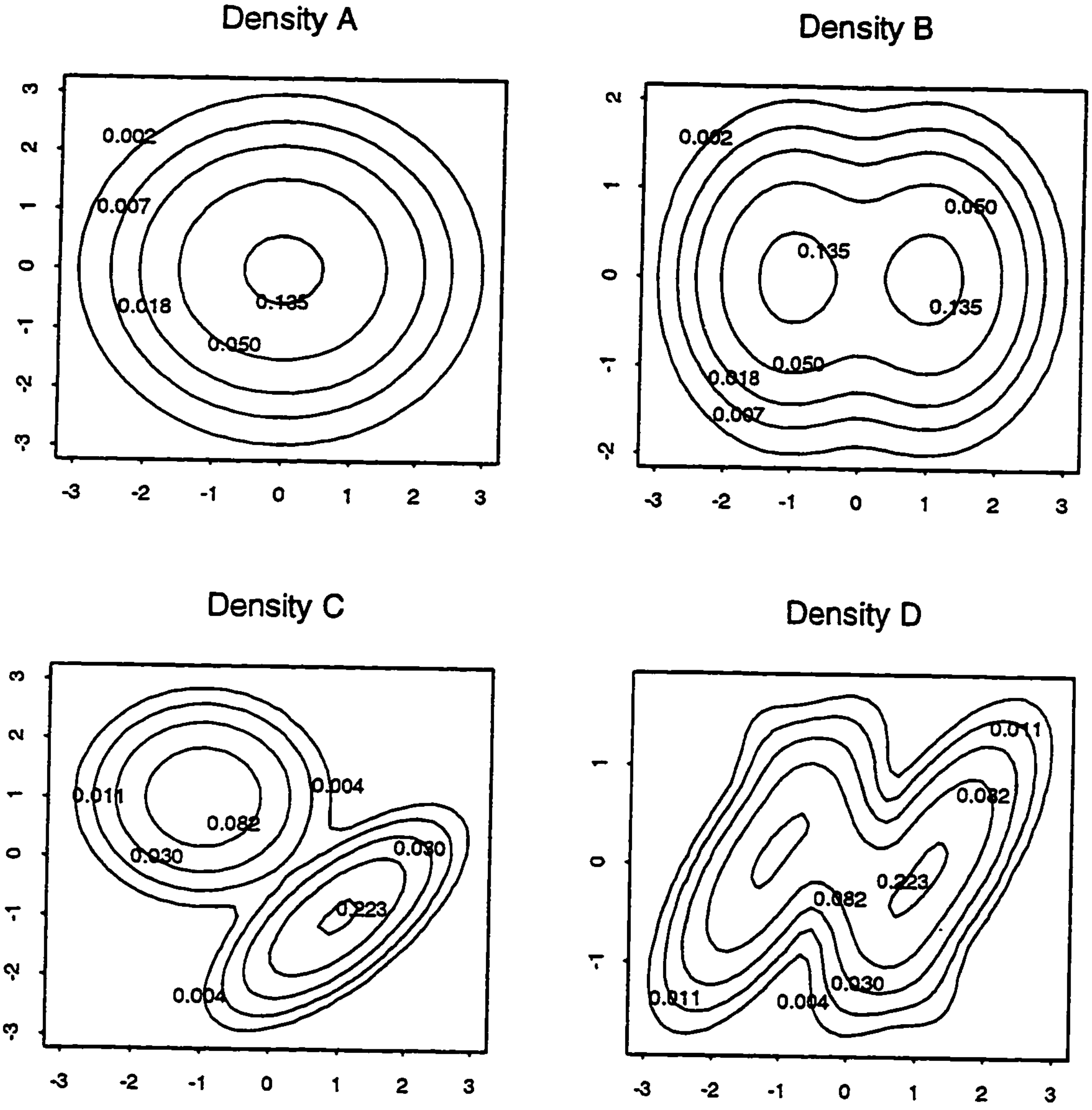


Figure 5-1. Contour plots of the bivariate test densities listed in Table 5-1.

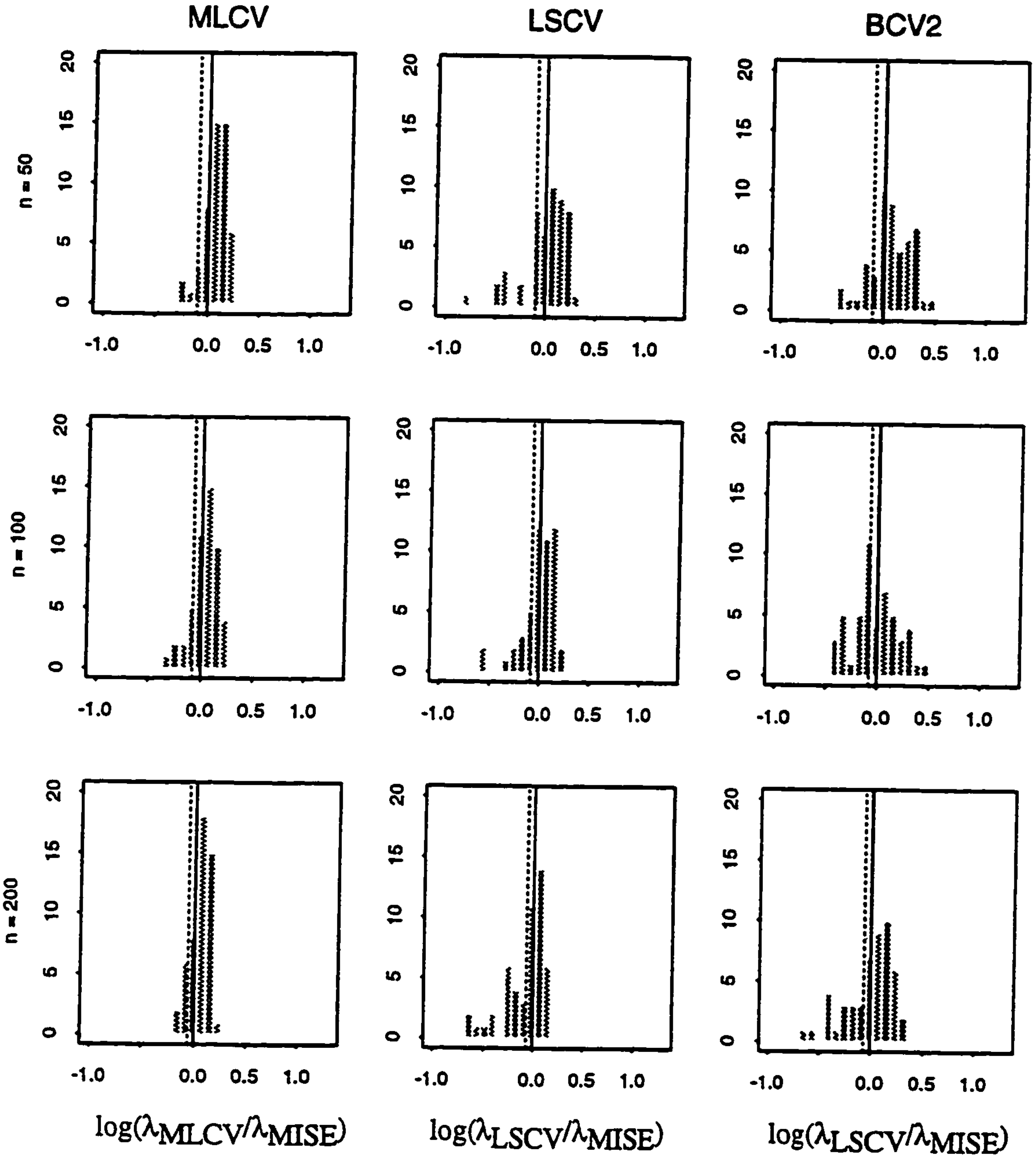


Figure 5-2. Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density A. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (5.6), respectively.

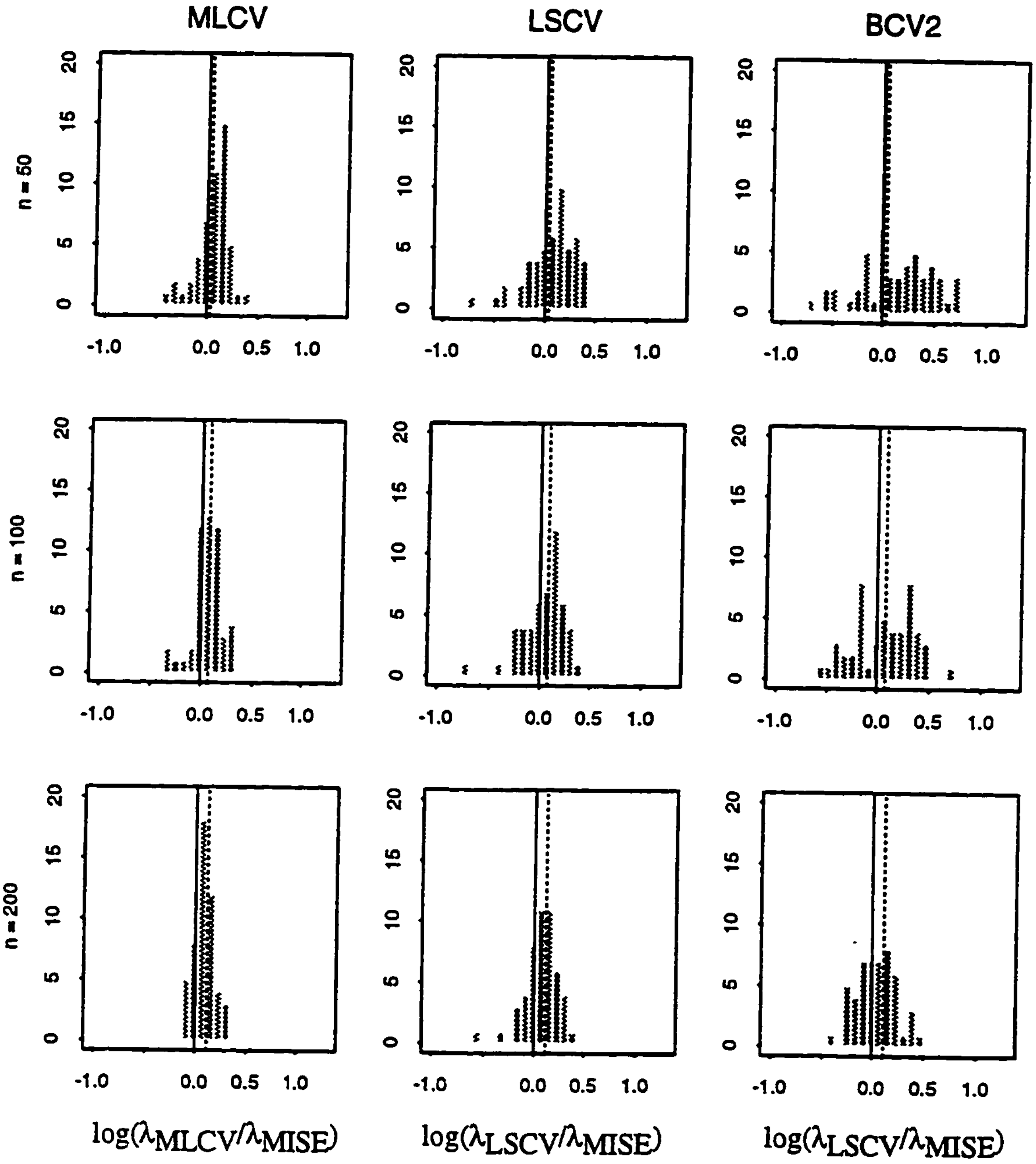


Figure 5-3. Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density B. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (5.6), respectively.

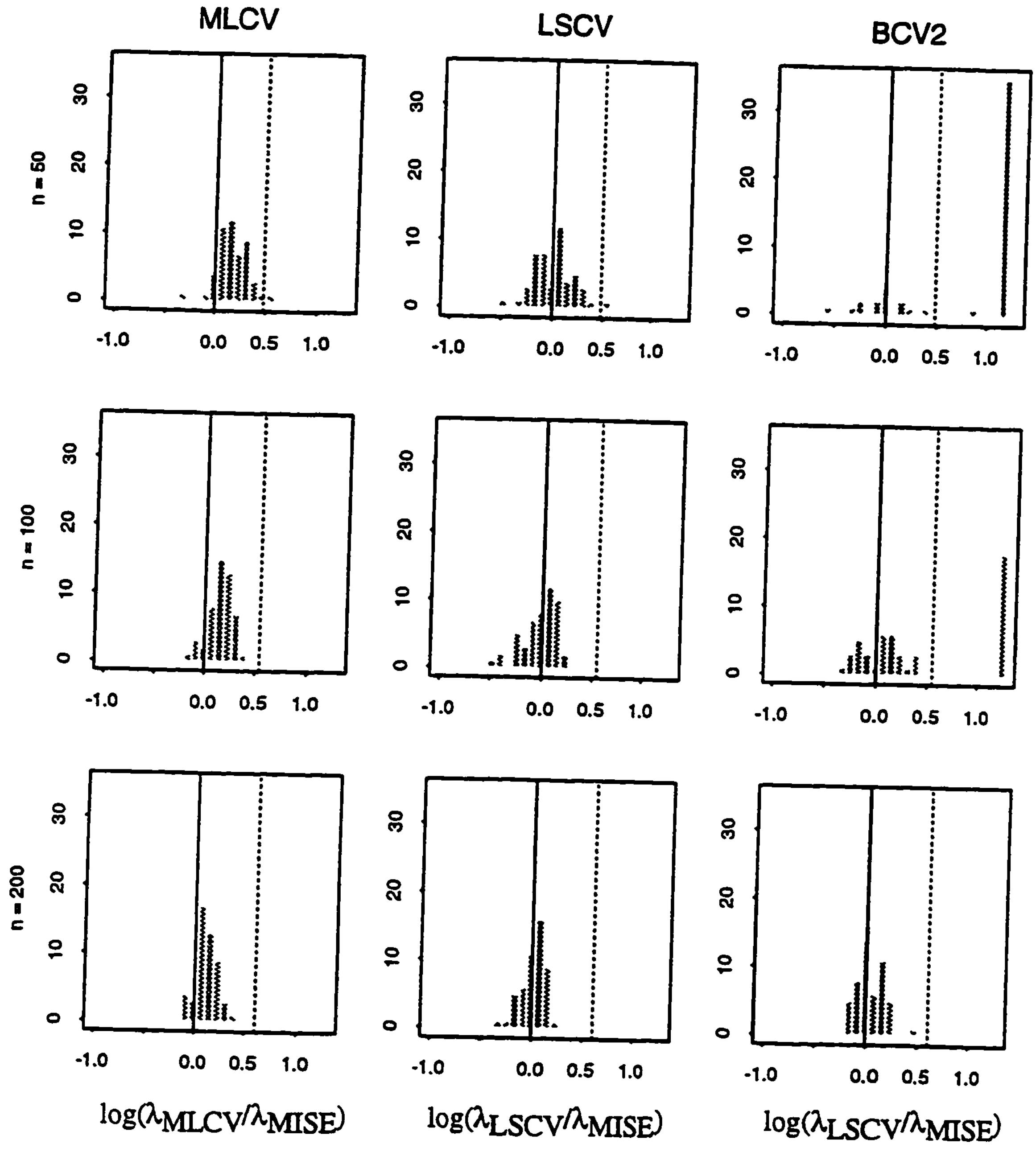


Figure 5-4. Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density C. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (5.6), respectively.

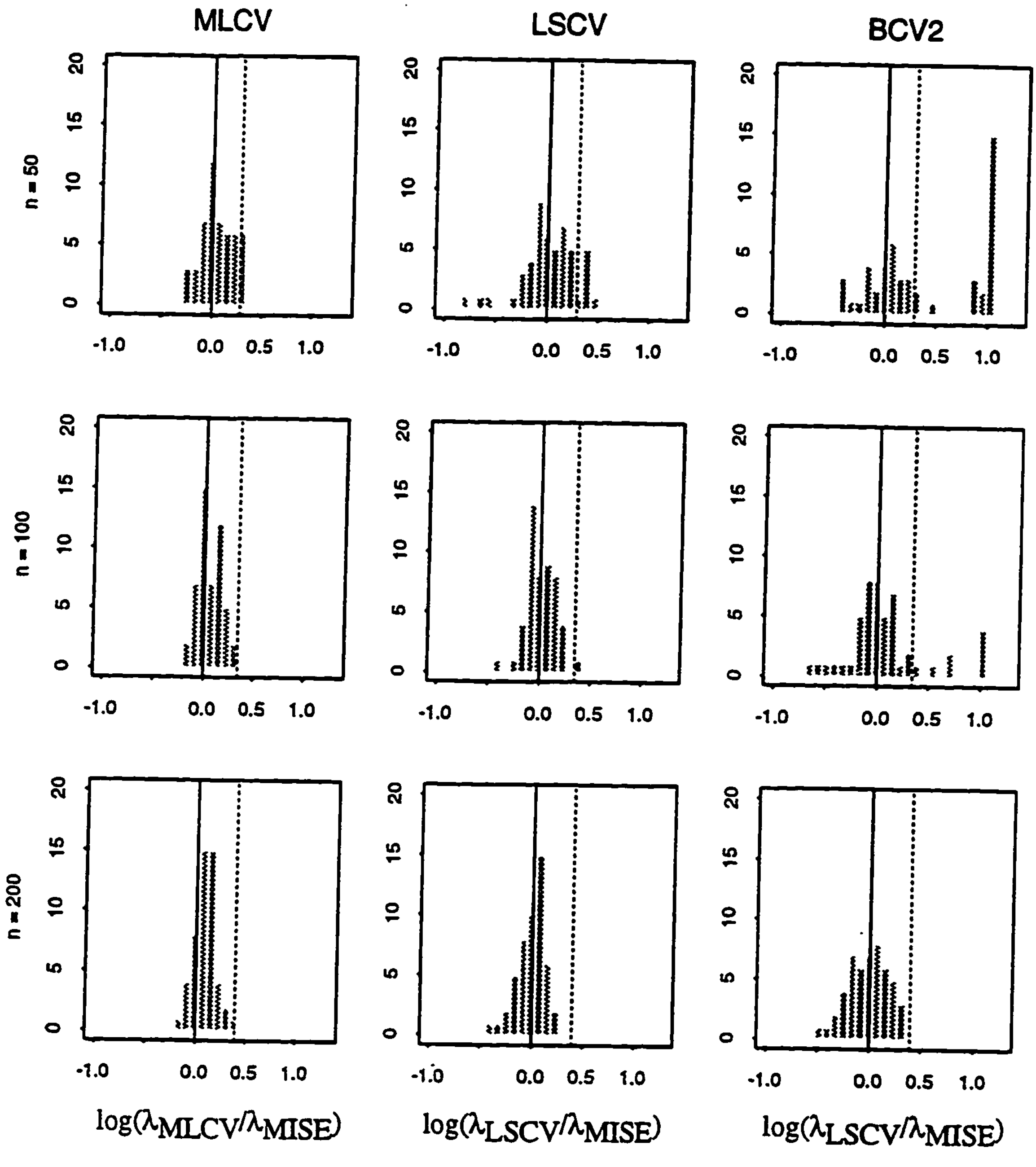


Figure 5-5. Histograms of optimal global bandwidths from MLCV, BCV2, and LSCV for the test density D. The bandwidths are divided by the MISE optimal bandwidth and plotted on a log scale. The continuous and dotted lines denote the MISE optimal bandwidth and the Gaussian reference bandwidth in (5.6), respectively.

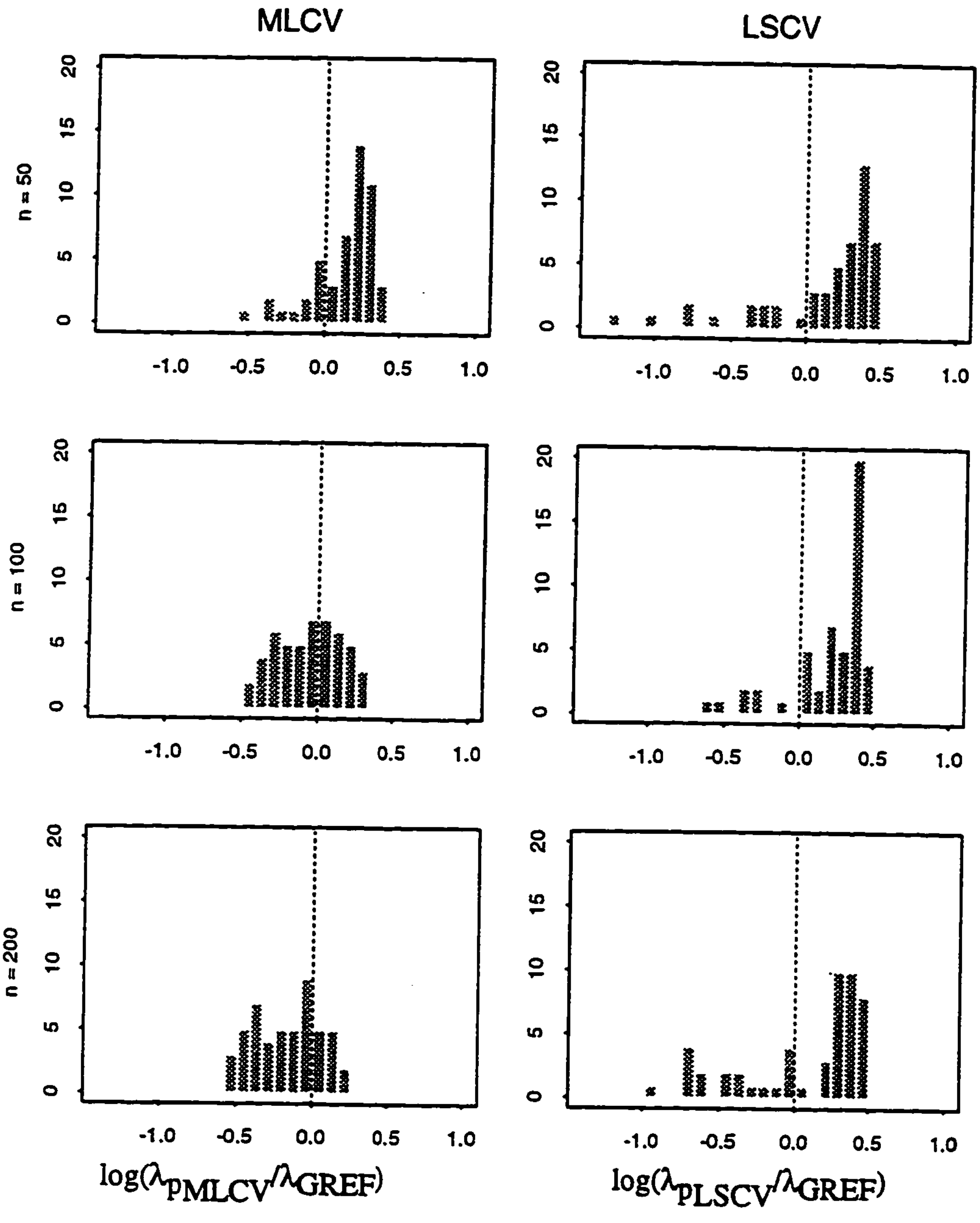


Figure 5-6. Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density A. The bandwidths are scaled by the Gaussian reference bandwidth (shown dotted in figure) in (5.6) and plotted on a log scale. The scaling is done merely to provide a common scale for the different sample sizes for which the results are portrayed.

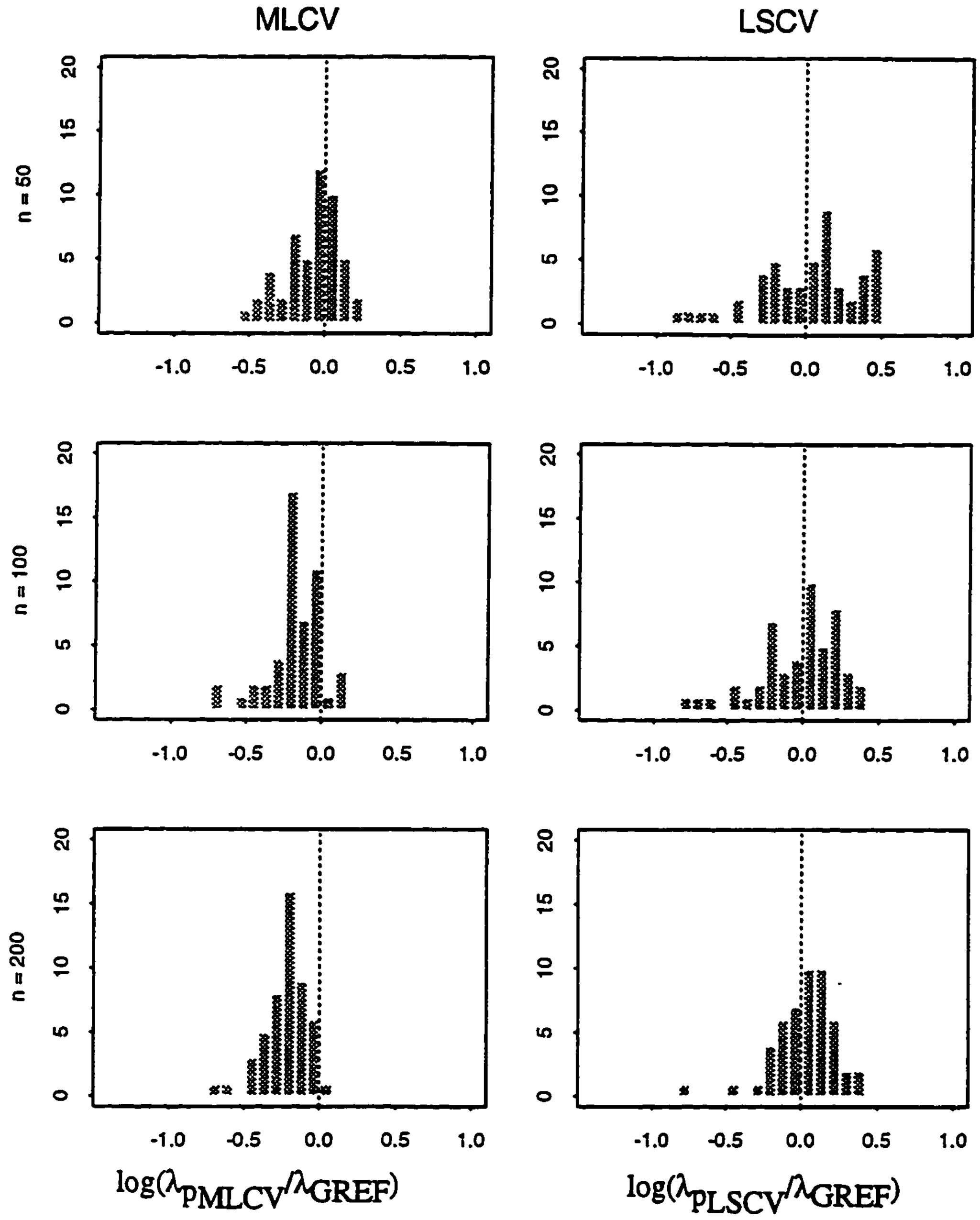


Figure 5-7. Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density B. The bandwidths are divided by the Gaussian reference bandwidth (shown dotted in figure) in (5.6) and plotted on a log scale.

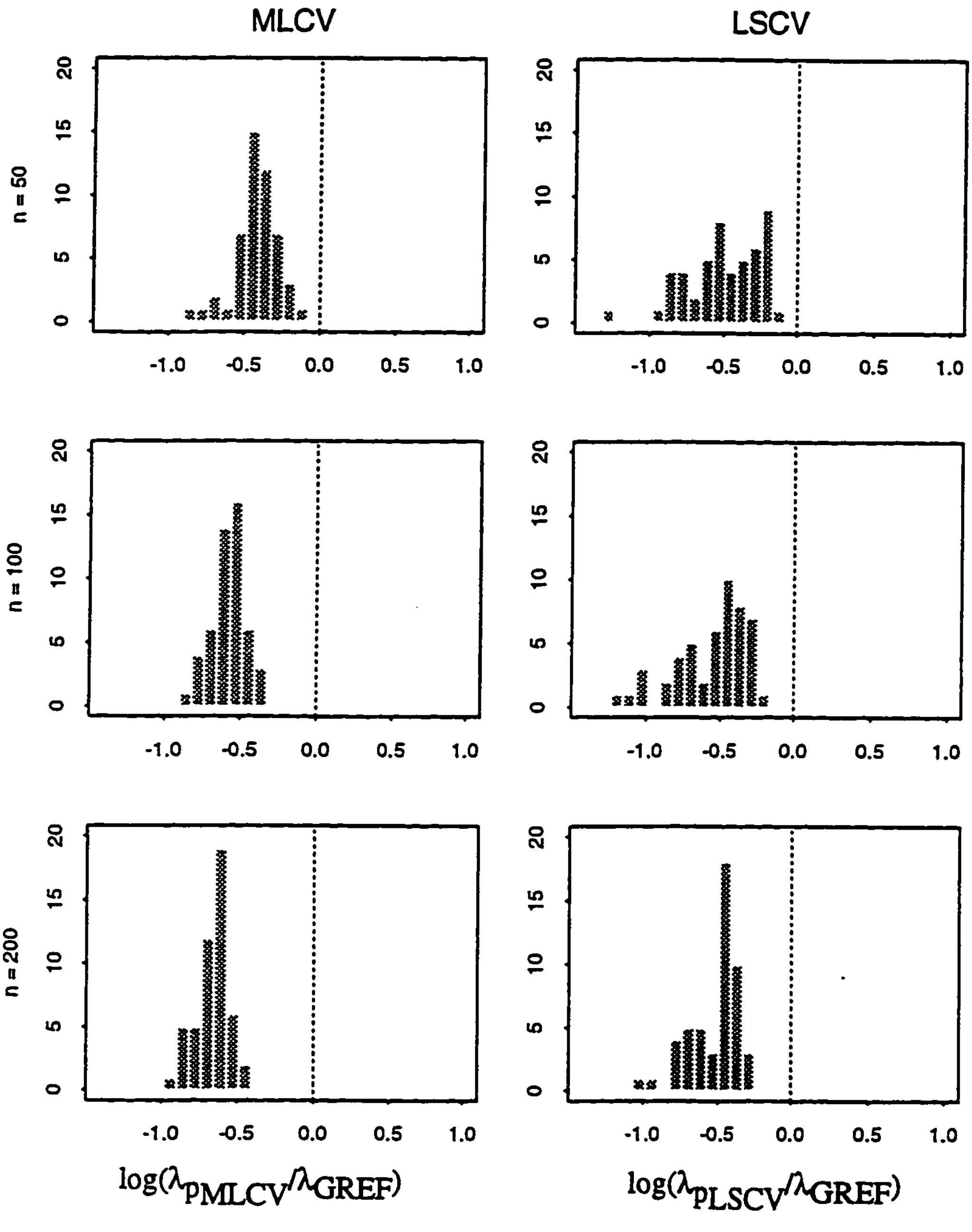


Figure 5-8. Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density C. The bandwidths are divided by the Gaussian reference bandwidth (shown dotted in figure) in (5.6) and plotted on a log scale.

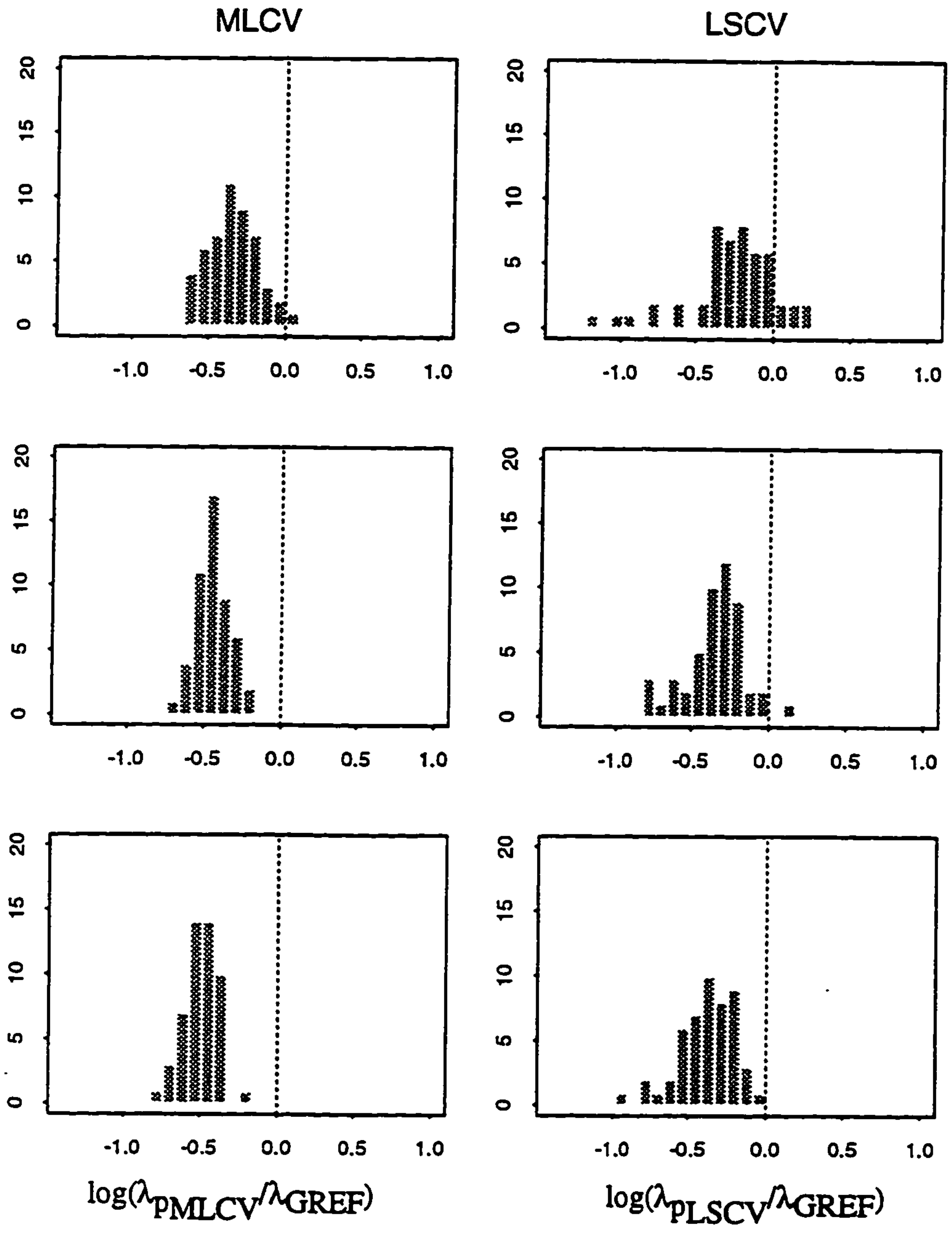


Figure 5-9. Histograms of the optimal pilot bandwidths from MLCV and LSCV for the test density D. The bandwidths are divided by the Gaussian reference bandwidth (shown dotted in figure) in (5.6) and plotted on a log scale.

CHAPTER 6

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Representation of dependence structure has been an important problem in stochastic hydrology. Prior work with a few exceptions [*Adamowski*, 1989; *Adamowski and Feluch*, 1990; *Karlsson and Yakowitz*, 1987; *Yakowitz*, 1987] has addressed this from a parametric viewpoint. A parametric approach usually involves choosing and fitting probability distributions that are completely described by a set of parameters, usually based on the first two or three moment characteristics of the observed streamflow data. One popular class of parametric models are the Autoregressive Moving Average (ARMA) models. Some of the drawbacks of stochastic streamflow modeling using ARMA models are:

(1) ARMA models assume that the dependence between flows is linear, or, that the flows can be adequately described as originating from a Gaussian joint probability distribution. In cases where this is not true, the data are required to be transformed to Gaussian by an appropriate transformation. Choosing a transformation may not be a simple issue, particularly so if the underlying process is one that cannot be characterized by one of the classical probability distributions such as Gamma or Lognormal.

(2) The normalizing transform admits only a limited degree of heterogeneity in the statistical dependence structure of the simulated sequences. This can be a drawback for data sets showing dependence of variance or correlation on the magnitude of streamflow.

(3) In spite of the fact that ARMA models have been around for the last 30 years or so, practitioners often tend to base their decisions on the historical record (or a resampled proxy thereof). This may reflect a consensus among practitioners that ARMA models are deficient at representing the relevant features of real streamflow data.

These points indicate the need for better approaches. The nonparametric methods introduced in this work are hopefully a step towards fulfilling these needs. Nonparametric methods avoid prior assumptions as to the form of dependence or of the nature of probability distribution of streamflow variables. The models based on nonparametric approaches are thus representative of the historical data rather than a set of statistics derived from the data. Although this study dealt with streamflow simulation, the models were kept general enough to be possibly applied to other areas of science and engineering. The specific contributions include:

(1) Nearest-Neighbor Bootstrap--This resampling method was described in chapter 2. Some advantages of the nearest-neighbor bootstrap algorithm are its ability to model linear and nonlinear dependence and multimodality in the probability density, and the computational ease with which the procedure can be used.

(2) Nonparametric order p simulation model--This simulation method, described in chapter 3, was based on kernel density estimation techniques. Unlike the nearest-neighbor approach, simulations here are not required to be a part of the observed flows. In addition, much like the nearest-neighbor approach, the dependence and distributional features that can be inferred by the data are modeled effectively.

(3) Nonparametric disaggregation model--This simulation method, described in chapter 4, was developed to generate a vector of component or disaggregate flows (seasonal or tributary flows), conditional to a given aggregate flow (annual or main stream flow). Kernel density estimation methods were used. In addition to the dependence between the disaggregate flow variables, the dependence between the aggregate and the disaggregate components was modeled. The sum of the disaggregate flows was required to be equal to the aggregate flow.

All of these models were analogous to the parametric methods of representing dependence between different flow variables. However, use of the nonparametric framework resulted in the simulations preserving a broader range of attributes of the observed streamflow time series. Features such as asymmetry and multimodality in the probability density, and nonlinearity in the dependence structure were modeled effectively and automatically in the nonparametric model simulations. It is important that not only were linear attributes modeled, but also a broader set of properties based on additional distributional information translated automatically to the nonparametrically generated sequences.

In developing the simulation models in chapters 3 and 4, a study of several bandwidth estimation methods for kernel density estimation was conducted on samples of sizes typically encountered in hydrology. Of the methods tested, results indicated that maximum likelihood cross validation (MLCV) was computationally simple and gave good estimates of the kernel bandwidth. Results from this study were presented in chapter 5.

Historically, stochastic methods have been used in hydrology to study droughts, floods, and issues relating to reservoir storage and reliability. By definition, events such as droughts or floods are infrequent and result in uncertain estimates because of the small samples usually available. The design decisions based on the parametric methodology have often involved extrapolations beyond the region where there are any real data. The precision of parametric extrapolations in quantifying design variables may have lent an unjustified sense of security. Nonparametric approaches may sometimes be incapable of providing extrapolations beyond the range of the data (for example, in case of bootstrap based approaches). However, this needs to be weighed against the fact that extrapolations based on parametric approaches are inherently flawed in cases where the parametric joint density is incorrectly specified. This is likely to be the case when normalizing transformations based on a limited set of options are applied on the data. Extrapolations in the nonparametric framework are more accurate because a nonparametric approach is a consistent estimator of the distributional characteristics of the data or the dependence information thereof. Also, use of only a local neighborhood reduces the possibility of the extrapolation being unduly biased by outliers. In the case of the bootstrap (which, by construction, is unable to extrapolate beyond the range of the historical data), modifications to the existing methodology (see chapter 2 for more information) can be used to provide reasonable extrapolations. In short, in spite of the inability of some nonparametric methods to extrapolate substantially beyond the range of the data, their use is recommended because they offer consistent estimates of the underlying distribution and thus have a lesser bias in estimates of the conditional mean, as compared to wrongly specified parametric methods.

Although the various simulation approaches espoused in this work are complete in themselves, several improvements can be suggested. The basic kernel density estimator or choice of the number of nearest neighbors can be chosen in a more rigorous manner. A substantial improvement in the kernel estimator would be to have kernels that are local, i.e., they have locally defined orientations and spread parameters. Use of such kernels would, amongst other advantages, result in lesser bias in regions where there are few data points. Extrapolations based on such a kernel density estimate would therefore be more accurate.

The disaggregation approach (chapter 4) can be improved by involving strategies that reduce the dimensionality of the application. Recall that the application in chapter 4 was disaggregation of annual flows to monthly, which needs specification of a 12-dimensional joint density estimate constructed from a small sample set. A more parsimonious representation could be to consider dependence on only the adjacent months, which would reduce the dimensionality considerably. Approaches that model year-to-year correlations could also be incorporated to provide better representation of the persistence that is observed in streamflow sequences.

An important issue concerning the models described in this work is the identification of model order. This issue was sidestepped by assuming the order for most applications. It is well known that consideration of fewer than the correct number of lags in streamflow simulation can result in biased estimates for reservoir storage. Estimation of the model order poses many challenges since consideration of additional lags increases the dimensionality of the density estimates, which is undesirable given the small samples that are available in practice. An approach that identifies the important dependencies in the data, based on their effect on the final application the streamflow sequences are put to, should be considered. Such an approach should be flexible enough to accommodate non-sequential lags (for example, lags 1, 3, and 5 instead of lags 1 through 5) that are able to reproduce the dependence structure present in the data.

Other possible directions of future work include strategies for modeling flows at smaller (daily or hourly) time scales; consideration of other hydrologic variables (such as precipitation) in the modeling exercise; consideration of higher order noise terms (much like the MA terms in ARMA models) that offer a better representation of the persistence that may be present in the data; direct nonparametric modeling of reservoir storages or severe droughts (approaches that sidestep the issue of simulating flow sequences); application in alternative spaces using Fourier or Wavelet bases so as to capture scale-dependent properties; approaches for modeling extreme events (floods and droughts) that use nonparametric methods developed specially for the tails of the data; and extensions of the simulation models to a forecasting scenario.

The above-mentioned applications are only a few of the possible directions to which nonparametric methods could be applied. Several other areas where quantification of uncertainty is an important issue should also be examined.

References

Adamowski, K., A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies, *J. Hydrol.*, 108, 295-308, 1989.

Adamowski, K., and W. Feluch, Nonparametric flood-frequency analysis with historical information, *J. Hydr. Eng.*, 116(8), 1035-1047, 1990.

Karlsson, M., and S. Yakowitz, Nearest-neighbor methods for nonparametric rainfall-runoff forecasting, *Water Resour. Res.*, 23(7), 1300-1308, 1987.

Yakowitz, S., Neares-neighbor methods for time series analysis, *J. Time Series Anal.*, 8(2), 235-247, 1987.

APPENDICES

APPENDIX A. STATE DEPENDENT CORRELATION COEFFICIENTS

This appendix describes the measures we used to quantify nonlinear dependence in data. The usual estimator of lag 1 correlation is:

$$r = \frac{1}{(n-1)\, s_x^2} \sum_{t=1}^{n-1} \left(x_t - \bar{x}\right)\left(x_{t+1} - \bar{x}\right) \tag{A.1}$$

where $\bar{x}$ and s_x^2 are the mean and variance of x_t, $t = 1 \ldots n$.

r_{af}: Forward, above median correlation, is defined as the correlation between above median flows and flows in the subsequent time step. This is calculated by replacing the sum over all t in the expression above by the sum over those t for which x_t is greater than the median flow x_{median}, replacing the s_x^2 by the product of the standard deviations of the above median flows and flows one time step ahead of above median flows, replacing $\bar{x}$ by the mean of the above median, and one time step ahead of above median flows and adjusting n accordingly.

r_{bf}: Forward, below median correlation, is the correlation between all below median flows and the subsequent time steps flow, calculated in a similar manner with the sum over those t for which

$x_t < x_{median}$.

r_{ab}: Backward, above median correlation, is the correlation between above median flows and the preceding time step's flow, calculated in a similar manner with the sum over those t for which $x_{t+1} > x_{median}$.

r_{bb}: Backward, below median correlation, is the correlation between below median flows and the preceding time step's flow, calculated in a similar manner with the sum over those t for which $x_{t+1} < x_{median}$.

For a linear Gaussian process the above and below pair of correlations in either the forward or backward direction should be the same. Differences indicate nonlinearity, or state dependence in the dependence structure. To test the significance of differences between sample correlation coefficients r_1 and r_2 the following test from *Kendall and Stuart* [1979] was used. The test is based on the transformation of the correlation coefficient r as:

$$z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) \tag{A.2}$$

The quantity $z_1 - z_2$ is closely normally distributed with zero mean and variance $1/(n_1-3)+1/(n_2-3)$, where n_1 and n_2 are the sample sizes, under the null hypothesis that z_1 and z_2 are calculated from sample correlation coefficients from populations with the same correlation coefficient. Therefore the significance test compares $(z_1-z_2)/\sqrt{1/(n_1-3)+1/(n_2-3)}$ to the standard normal distribution. This test is approximate unless the samples are from independent bivariate normal populations. In section 6 we used this test to investigate the significance of the difference between r_{af} ($= r_1$) and r_{bf} ($= r_2$). Sets of above and below median flows are effectively censored samples, inconsistent with the independence assumptions. Nevertheless, an approximate measure of whether these quantities are significantly different can be obtained by use of this test.

References

Kendall, M., and A. Stuart, *The Advanced Theory of Statistics, Volume 2, Inference and Relationship*, 4th ed., MacMillan Publishing Co., New York, 1979.

APPENDIX B. DERIVATION OF MODEL STATISTICS

We derive here the expected values of selected statistics of the NP1 model. These depend on the observed data x_i, kernel parameters λ and $\mathbf{S}$ and the Gaussian kernel function.

Marginal Distribution of X_t

The marginal density of X_t (denoted $\hat{f}_m(X_t)$) is estimated as

$$\hat{f}_m(X_t) = \int \hat{f}(X_t, X_{t-1})\, dX_{t-1} = \frac{1}{n}\sum_{i=1}^{n} f_G(X_t - x_i, \lambda^2 S_{11}) \tag{B.1}$$

where

$$f_G(X_t - x_i, \lambda^2 S_{11}) = \frac{1}{\sqrt{2\pi\lambda^2 S_{11}}} \exp\left(- \frac{(X_t - x_i)^2}{2\lambda^2 S_{11}} \right) \tag{B.2}$$

denotes a Gaussian density function with mean x_i and variance $\lambda^2 S_{11}$. This follows from Equation (3.6) with **H** from Equation (3.7) and S expressed as Equation (3.19). Equation (3.6) is the sum of n multivariate Gaussians each of which when integrated over X_{t-1} results in the univariate Gaussian given above. This marginal distribution is used to calculate model mean, covariance and skewness.

Mean μ' of X_t

This can be evaluated using the marginal distribution in (B.1). Since each kernel is symmetric and centered at a data point, the NP1 model mean (μ') is the sample mean.

$$\mu' = E[X_t] = \int X_t \hat{f}_m(X_t)\, dX_t = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{B.3}$$

Standard Deviation of X_t

The variance under the NP1 model can be written as

$$\sigma'^2 = E\left[(X_t-\mu')^2\right] = \int (X_t-\mu')^2 \hat{f}_m(X_t)\, dt \tag{B.4}$$

where the expectation is over the marginal distribution, Equation (B.1). Since $\hat{f}_m(X_t)$ from Equation (B.1) is a sum of Gaussian p.d.f.'s with individual means x_i and variances $\lambda^2 S_{11}$ and the x_i have sample variance S_{11}, we get

$$\sigma'^2 = S_{11}(1+\lambda^2) \tag{B.5}$$

Note the inflation in the underlying variance by the factor $(1+\lambda^2)$.

Lag 1 Correlation

The lag 1 correlation (ρ_1') under the NP1 model is expressed as the ratio

$$\rho'_1 = \frac{E\left[(X_t-\mu')(X_{t-1}-\mu')\right]}{\sigma'^2} \tag{B.6}$$

where expectation is over the joint density estimate in (3.6). This expression simplifies to

$$\rho'_1 = \frac{(1+\lambda^2)\, S_{12}}{(1+\lambda^2)\sqrt{S_{11}S_{22}}} = r \tag{B.7}$$

where r denotes the sample lag 1 correlation:

$$r = \frac{S_{12}}{\sqrt{S_{11} S_{22}}} \tag{B.8}$$

Skewness

The coefficient of skewness (γ') under the NP1 model is defined as the ratio

$$\gamma' = \frac{E\left[(X_t-\mu')^3 \right]}{\sigma'^3} = \frac{\int (X_t-\mu')^3\, \hat{f}_m(X_t)\, dt}{\sigma'^3} \tag{B.9}$$

where the expectation is over the marginal distribution in (B.1). By integrating over the marginal distribution the numerator can be evaluated as

$$E\left[(X_t-\mu')^3 \right] = \frac{1}{n}\sum_{i=1}^{n} x_i^3 + 3\lambda^2 S_{11}\frac{1}{n}\sum_{i=1}^{n} x_i - 3\mu'\lambda^2 S_{11} - 3\mu'\frac{1}{n}\sum_{i=1}^{n} x_i^2 + 3\mu'^2\frac{1}{n}\sum_{i=1}^{n} x_i - \mu'^3 \tag{B.10}$$

Now recognizing (B.3) the second and third terms cancel and this is equivalent to

$$E\left[(X_t-\mu')^3 \right] = \frac{1}{n}\sum_{i=1}^{n} (x_i-\mu')^3 \tag{B.11}$$

The expression for γ' then becomes

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu')^3}{\sigma'^3} = \frac{g}{\left(1+\lambda^2\right)^{3/2}} \tag{B.12}$$

where g is the skewness estimator

$$g = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu')^3}{S_{11}^{3/2}} \tag{B.13}$$

and S_{11} the sample variance. The decrease in skewness is due to the inflation of variance by (B.5). The results derived here do not account for the cut and normalize boundary corrections applied.

APPENDIX C. PERMISSION LETTER

AGU

American Geophysical Union
2000 Florida Avenue, N.W.
Washington, D.C. 20009
Phone (202) 462-6900
TWX 710-822-9300
FAX 202-328-0566

April 17, 1996

Ashish Sharma
Utah Water Research Laboratory
Utah State University
Logan, Utah 84322-8200

Dear Mr. Sharma:

We are pleased to grant permission for the use of the material requested for inclusion in your thesis, including microfilm editions thereof. Permission is restricted to the use stipulated. The original publication must be appropriately cited. The credit line should read: "authors, journal or book title, volume number, page numbers, year," and the phrase "Copyright by the American Geophysical Union." Substitute the last phrase with "*Published* by the American Geophysical Union" if the paper is not subject to U.S. copyright – see the copyright line on the first page of the published paper for such classification.

Please feel free to contact me again if you need further assistance. Thank you.

Sincerely,

Julie A. Hedlund

Julie A. Hedlund
Manager
Publications Administration

The American Geophysical Union encompasses the Earth and space sciences:
Geodesy, Seismology, Atmospheric Sciences, Geomagnetism and Paleomagnetism,
Ocean Sciences, Hydrology, Volcanology, Geochemistry, and Petrology,
Tectonophysics, Planetology, Solar-Planetary Relationships

CURRICULUM VITAE

Ashish Sharma
(July, 1996)

EDUCATION:

Ph.D. in Civil and Environmental Engineering (July, 1996),
Utah State University, Logan Utah.
Master of Technology in Water Resources Engineering (March, 1991),
Indian Institute of Technology, New Delhi, India.
Bachelor of Engineering in Civil Engineering (May, 1989),
University of Roorkee, Roorkee, India.

EXPERIENCE:

Research Assistant at Utah Water Research Laboratory, Logan, Utah (1992 - 1996).
Research Assistant at U. S. D. A. Forest Sciences Laboratory, Logan, Utah (1994-1995).
Worked at Water and Power Consultancy Services (WAPCOS), New Delhi, India (1991-1992).

PUBLICATIONS AND PRESENTATIONS:

Sidle, R. C. and A. Sharma, Stream Channel Changes associated with Mining and Grazing in the Great Basin, in press, *Journal of Environmental Quality*, 1996.

Sharma, A., U. Lall and D. G. Tarboton,Streamflow Synthesis using Nonparametric Methods, *Eos Transactions AGU*, 74(43): Fall Meeting Supplement, 141, 1993.

Sharma, A., D. G. Tarboton and U. Lall, The use of Nonparametric Probability Distributions in Streamflow Modeling, Presentation at the *International Conference on Hydrology and Water Resources*, New Delhi, India, December 22-24, 1993.

Sharma, A., D. G. Tarboton and U. Lall, A Disaggregation Streamflow Simulation Model Using Nonparametric Density Estimation Techniques, *Eos Transactions AGU*, 75(16): Spring Meeting Supplement, 160, 1994

Sharma, A., and D. G. Tarboton, Conceptual Models of Snowmelt Processes, *Eos Transactions AGU*, 76(46): Fall Meeting Supplement, 1995.