DISSERTATION

STOCHASTIC SIMULATION OF HYDROLOGIC DATA BASED ON NONPARAMETRIC APPROACHES

Submitted by

Taesam Lee

Department of Civil and Environmental Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, CO.

Fall 2008

COLORADO STATE UNIVERSITY

October 28, 2008

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY TAESAM LEE ENTITLED STOCHASTIC SIMULATION OF HYDROLOGIC DATA BASED ON NONPARAMETRIC APPROACHES BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Jorge A. Ramirez

Jeffrey D. Niemann

Duane C. Boes

Advisor (Jose D. Salas)

Department Head

ABSTRACT OF DISSERTATION

STOCHASTIC SIMULATION OF HYDROLOGIC DATA BASED ON NONPARAMETRIC APPROACHES

Stochastic simulation of hydrologic data has been widely developed for several decades. However, despite the several advances made in literature still a number of limitations and problems remain. The major research topic in this dissertation is to develop stochastic simulation approaches to tackle some of the existing problems such as the preservation of the long-term variability and the joint modeling of intermittent and non-intermittent stations. For this purpose, nonparametric techniques have been applied. For simulating univariate seasonal streamflows, a model is suggested based on k-nearest neighbors resampling (KNNR). Gamma kernel density estimate (KDE) perturbation is employed to generate realistic values of streamflow that are not part of the historical data. Further, aggregate and pilot variables are included in KNNR so as to reproduce the long-term variability. For multivariate streamflows, the moving block bootstrapping procedure is employed considering a random block length, KNNR block selection to avoid the discontinuity between blocks, a Genetic Algorithm mixture, and Gamma KDE perturbation. In addition, the drawbacks of an existing nonparametric disaggregation scheme have been examined and appropriate modifications developed that include accurate adjusting for the disaggregate variable, KNNR, and Genetic Algorithm mixture. The suggested univariate, multivariate, and disaggregation models have been compared

with existing nonparametric models using several cases of streamflow data of the Colorado River System. In all cases, the results showed major improvements. Furthermore, disaggregation from daily to hourly rainfall for a single site has been studied based on three disaggregation models so as to account for the diurnal cycle in hourly data. Those models are (1) Conditional Markov Chain and Simulated Annealing (CMSA), (2) Product Model (GAR(1)-PDAR(1)) with Accurate Adjusting (PGAA), and (3)Stochastic Selection Method with Weighted Storm Distribution (SSMW). Various tests and comparisons have been performed to validate the models and it revealed that PGAA is superior to the others for preserving the diurnal cycle and the key statistics of hourly rainfall.

Taesam Lee
Department of Civil and Environmental Engineering
Colorado State University
Fort Collins, CO 80523
Fall 2008

ACKNOWLEDGMENT

At the beginning, all gratitude and deepest appreciations are due to the living God, Jesus Christ.

Upon completing the graduate work that leads to dissertation, the author would like to state his deep and sincere gratitude to his advisor, Dr. Jose D. Salas, Professor of Civil and Environmental Engineering, for his guidance, and encouragement during the author's work on this research. Thanks and gratitude is expressed to the author's committee members: Dr. Duane C. Boes, Emeritus Professor of Statistics, Dr. Jorge A. Ramirez and Dr. Jeffrey D. Niemann, Professors of Civil and Environmental Engineering. In addition, special thanks to Dr. William L. Lane for his comments and suggestions and to Dr. James R. Prairie for his accessibility and revisions.

The financial support provided by the project "Development of Stochastic Hydrology for the Colorado River System" funded by the U.S. Bureau of Reclamation is gratefully acknowledged. Special thanks are extended to Dr. Don Frevert, and Dr. Terry Fulp for their support.

The author appreciates the support of Korean Friends and Colleagues, especially Daeryong Park and Jonghyun Lee. Thanks to the colleagues in the Hydrology Lab is stated especially to Michael L. Coleman, Mark A. Perry, and Ernesto Trujillo.

The author greatly acknowledges the special help and care to his colleague, Ki-Wook Cha and his wife when the author was in difficulty. Also thanks to great support and encouragement of Sukhyun Kim and his wife.

An expression of thanks and gratitude is given to Hanbit Church members and Moohan Mokzang for their prays and encouragements, especially to the pastor, Jeongsup Choi. Great appreciation is given to Cornerstone church members and pastor and his wife. The author appreciates the wisdom and great care of John Brumbaugh.

My deepest gratitude goes to my dear mother for her endless love and care all over my life. My sincere appreciations to my wife, Misun Kim, for her support and patience to prepare a suitable environment to finish my research are expressed.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Study Motivation and Background

Water allows living creatures to exist on the earth. The amount of water is not proportionally distributed. Complex physical reactions on the surface and in the atmosphere of the earth cause diverse climate regions. If the average annual precipitation is less than 500mm and the average annual potential evapo-transpiration exceeds 800mm in a region, the region is defined as arid (McMahon, 1979). In large river basins one may find a variety of climatic regions such as arid, semi-arid, and moderate. For example, the Colorado River system includes arid zones such as Arizona and Nevada, semi-arid areas such as Colorado and Utah, and the moderate zone of California.

The Colorado River is one of the main sources of water for several states in the western United States. Water management is an important issue in the Colorado River system. Generally, some water is delivered from places where plenty of water exists to the places where water is scarce. Planning the storage, diversion, and delivery of water must consider current and future conditions of the available water resources. Estimating

the future availability of water resources is not be easy task. Stochastic simulation have been suggested to create possible streamflow scenarios that may arise in the future. The simulated data allow water managers checking many possible options. Many models for simulating streamflows at monthly and yearly time scales have been developed and applied in water resources management area.

The time series simulation models that are typical in hydrology and water resources include Autoregressive Moving Average (ARMA), periodic ARMA (PARMA), multisite ARMA (MARMA), contemporaneous ARMA (CARMA), and disaggregation models (Salas, 1980, Loucks et al., 1981). These models are linear and assume normal distribution. Since hydrologic data such as rainfall and streamflow are not normally distributed, data transformation is unavoidable. The data transformation might induce bias on key statistics such as the mean and standard deviation of the original variable in real domain even if there are no biases in the transformed (normal) variable.

In the last two decades, nonparametric simulation techniques such as bootstrapping, k-nearest neighbors resampling, conditional kernel density estimate, nonparametric disaggregation, and more have been developed to provide alternatives and get around some of the shortcomings of parametric models. In this study, the current nonparametric simulation techniques for streamflow data are investigated and drawbacks of the techniques are revealed such as generating only historical values, and the repetition of seasonal and spatial patterns. In order to tackle the drawbacks of the current nonparametric models, a number of modifications are proposed such as Gamma KDE perturbation, the inclusion of aggregate or pilot variable, Genetic Algorithm mixture, and combination of nonparametric disaggregation and accurate adjusting. Furthermore, the

proposed modifications will be useful for simulating intermittent and non-intermittent streamflows jointly at several sites.

In addition, finer time scale data such as hourly precipitation are needed for water quality modeling and dam operation. A disaggregation model from daily to hourly is very useful because often only daily data are available. Even though some disaggregation models have been developed, the diurnal cycle that may be an important feature in some areas has not been considered. Therefore, some disaggregation models from daily rainfall to hourly are proposed that include the effect of the diurnal cycle.

1.2 Objectives of the Study

The general objective of this research is developing nonparametric simulation techniques that are applicable to hydrologic data such as streamflow and rainfall. For streamflow data, nonparametric models are mainly focused considering the long-term variability and the joint modeling of intermittent and non-intermittent data. For rainfall data, existing rainfall disaggregation models are enhanced to account for the diurnal cycle in hourly data. Specific objectives that will be considered are:

(1) To develop a model that is capable of generating seasonal streamflow data at single sites. The model will be nonparametric which will enable one generating data values that are not part of the historical data, it will avoid generating negative values, and it will improve the preservation of long-term variability. This will be accomplished by Gamma kernel density estimate (GKDE) perturbation conditioning on pilot or aggregate variable based on k-nearest neighbors resampling (KNNR).

(2) To develop a model that is capable of simulating streamflows at multiple stations. Nonparametric techniques are applied to generate sequences that will produce feasible mixing (spatial and temporal) and the joint modeling of intermittent and non-intermittent data. This will be done by the multivariate bootstrapping with a random block length, KNNR block selection, and Genetic Algorithm mixture.

(3) To develop a model for disaggregating higher scale variable (e.g. yearly) into a lower scale variable (e.g. monthly). The model must preserve the correlation between the last month of the previous year and the first month of the current year and be capable of spatially disaggregating data for intermittent and non-intermittent sites. This will be accomplished by modifying an existing nonparametric technique with the accurate adjusting and Genetic Algorithm mixture.

(4) To enhance the existing models to disaggregate daily rainfall data to hourly so that the effect of the diurnal cycle are taken into account.

(5) To validate and apply the various models listed in (1)-(4) using actual single site and multisite data, particularly data of the Colorado River system.

1.3 Dissertation Outline

This dissertation includes mainly four different modeling chapters followed by conclusions and recommendations. In Chapter 2, a univariate model for generating seasonal streamflow is proposed. The model uses the nonparametric techniques such as

k-nearest neighbors resampling (KNNR) and Gamma Kernel Density Estimate (GKDE). In Chapter 3, a multivariate model is suggested to simulate multi-station seasonal streamflow with nonparametric techniques such as bootstrapping and KNNR as well as the mixing process with Genetic Algorithm. In Chapter 4, a model is proposed that disaggregates higher-level data (e.g. yearly) into lower-level data (e.g. monthly). The suggested model employs the nonparametric technique KNNR and the accurate adjusting procedure. In Chapter 5, the current existing models that disaggregate the daily rainfall data into hourly are investigated. The disaggregation models are enhanced to account for the diurnal cycle. Chapter 6 presents conclusions and summary of the contributions from this study followed by the recommendations for future study.

CHAPTER II

NON-PARAMETRIC SIMULATION OF SINGLE SITE SEASONAL STREAMFLOWS

2.1 Introduction

Stochastic models of hydrological processes have been developed so as to reproduce key statistical features of the observed hydrological data such as mean, variance-covariance, skewness, seasonality, and long-term dependency (e.g. Salas et al., 1980; Hipel and McLeod, 1994). The synthetic data obtained from these models are used for evaluating alternative designs and operating rules of hydraulic structures, or analyzing the effect of extreme hydrologic events such as droughts (e.g. McMahon, 2006). For this purpose, a number of parametric models have been suggested in literature such as the autoregressive moving average (ARMA) model (e.g. Salas et al., 1980; Loucks et al., 1981), the shifting mean (SM) (Salas and Boes, 1980; Fortin et al., 2004; Sveinsson et al. 2003), and the fractional gaussian noise (e.g. Mandelbrot and Wallis, 1969). However, most parametric models suffer from a number of drawbacks. For example, the marginal distribution of hydrologic data is often non-gaussian which requires transforming the observed data (e.g. using logarithmic or power transformations) prior to fitting the

models. However, modeling in the transformed domain may cause some bias in reproducing the basic historical statistics (e.g. standard deviation) although modifications have been suggested to correct for such bias (e.g. Fiering and Jackson, 1971; Chebaane et al., 1995). Also gamma autoregressive models with gamma marginal distribution have been developed to model AR auto-covariance instead of transforming the data (Fernandez and Salas 1986, 1990). Still, if the historical marginal distribution is bimodal or multimodal, it is difficult to replicate it with parametric models. And if an inappropriate transformation function is applied to transform the skewed data, it may lead to generation of extremely large values which may not be acceptable from the practical standpoint. It is challenging to reproduce skewed data yet avoiding the generation of negative values. Furthermore, non-linear relationships, which may be observed in the historical data can not be captured by the usual parametric time series models.

To overcome some of the mentioned drawbacks (in parametric models,) nonparametric models have been developed such as resampling techniques (bootstrapping.) For example, block bootstrapping has been suggested (Vogel and Shallcross, 1996) but because of discontinuity between blocks (each block is resampled independently) only historical values are generated. Srinivas and Srinivasan (2005 and 2006) devised a hybrid model combining a periodic autoregressive model and block bootstrapping. Their model was used for generating monthly streamflows of the Beaver and Weber rivers in Utah and yield generated values other than the historical ones. However, the model has limited variability especially where multi-season data are generated because it uses yearly block innovations (Lee and Salas, 2008). On the other hand, more elaborate schemes based on k-nearest neighbor resampling (KNNR), has been

developed by Lall and Sharma (1996). This resampling scheme has been further advanced and applied by many researchers in the field such as Rajagopalan and Lall (1999), Yates et al (2003), Prairie et al (2006) and Sharif and Burn (2007). For example, Prairie et al (2006) modified the KNNR technique employing local polynomial regression. Also, the conditional density estimate is used for nonparametric simulation instead of utilizing the resampling scheme [Sharma et al. 1997 and 2002]. Independently, Young (1994) applied a similar approach (called multivariate chain model) for generating weather variables. Young's method employs the discriminant space obtained from multivariate discriminant analysis of the observed data and a set of similar days is chosen from discriminant space using a nearest neighbor algorithm.

One of the major difficulties in simulating seasonal time series has been the reproduction of interannual variability. The traditional approach to tackle this problem has been using temporal disaggregation (e.g. Valencia and Schaake, 1993; Stedinger et al. 1985). In this approach the annual series is modeled and generated first, which are then disaggregated into seasonal data (e.g. monthly). Also a nonparametric disaggregation scheme has been devised by Tarboton et al (1998) employing conditional kernel density estimate. These disaggregation models are not efficient since they employ the entire relation structure in the lower scale time series. On the other hand, Koutsoyiannis and Manetas (1996) suggested a simpler disaggregation scheme where the seasonal and the annual series are generated by two separate models. The lower scale (i.e. seasonal) time series are regenerated until the summation of the seasonal data is close to the higher scale (annual) data, and then the lower scale time series data are readjusted to meet the additivity condition. Still, some bias in the lower scale synthetic data is unavoidable

resulting from the adjustment. Sharma and O'Neil (2002) developed a nonparametric simulation technique that employs the conditional kernel density estimate and the summation of the previous p-months as a condition. Furthermore, the upper scale (e.g. yearly) streamflow data may involve certain unique features such as long memory (O'Connell, 1971) and shifting means (e.g. Salas and Boes, 1980; Sveinsson et al. 2003). These features may be incorporated in the generation of seasonal streamflows indirectly via temporal disaggregation as suggested above. Including these features directly in the generation of seasonal streamflows is not straightforward. Nevertheless, Langousis and Koutsoyiannis (2006) developed a seasonal streamflow generation model that is able to reproduce long memory by using Fractional Gaussian Noise structure and cyclostationarity.

In this study, some representative nonparametric simulation techniques are further reviewed and analyzed. From reviewing their pros and cons new schemes for generating seasonal streamflows are developed utilizing KNNR and a Gamma Kernel density estimate which are geared to reproducing both seasonal and interannual variability. In Section 2, a brief review of key nonparametric simulation techniques is described. The mathematical description of the suggested model is illustrated in Section 3, followed by a description of the data and the procedure for assessing the models in Section 4. Sections 5 and 6 show the results and conclusions, respectively.

2.2 Review of nonparametric simulation models

Let $x_{\nu,\tau}$ define seasonal streamflow at time steps τ =1,..., ω (seasons) and ν =1,...,N (years) where ω and N denote the number of seasons (e.g. months) and years,

respectively. Furthermore, x_t stands for yearly streamflow data where t=1,…,N. The superscripts G or H, e.g. x_t^G or x_t^H, will be used where referring to generated or historical data, respectively. Also random variables for yearly and seasonal time series are denoted by X_t and $X_{v,\tau}$, respectively. Sometimes the lower-scripts are excluded where referring to random variables. Three nonparametric simulation alternatives, are described here such as *k*-nearest neighbor resampling (KNNR) (*Lall and Sharma*, 1996), local polynomial with KNN innovation (LPK) (*Prairie, et al.*, 2006), and nonparametric order *p* simulation with long-term dependence (NPL) (*Sharma and O'Neill*, 2002). Pros and cons of each model are also discussed from the authors' point of view.

2.2.1 KNNR

The KNNR method was developed for the generation of yearly and monthly time series by Lall and Sharma (1996) and applied to streamflow generation of the Weber River in Utah. The mathematical background of this approach relies on *k*-nearest neighbor density estimator that employs the Euclidean distance to the k^{th} nearest data point and its volume containing *k*-data points. KNNR generates a value from the historical data according to the closeness of the distance estimated from the current feature vector and the historical feature vector. Thus the same values of the historical data are obtained but with different combinations and orders. The procedure is summarized below using as example the illustration in Figure 2.1.

Firstly two notations are employed to indicate the yearly scale, namely ν refers to years in the historical data and ν =1,…,*N* while *t* refers to years in the generated data and t=1,…,N^G where N^G is the required length of generation. Then, assume that the initial

value x_0^G is known and set $D_1 = x_0^G$ (x_0^G may be taken randomly from the historical values) and the number of neighbors k, is determined by $k = \sqrt{N}$ (Lall and Sharma, 1996). Then:

(a) We want to generate (resample) x_t^G given the (known) feature vector $D_t = x_{t-1}^G$. For example, $x_{t-1}^G = 2.39$ in Figure 2.1.

(b) The k-nearest neighbors of $x_{t-1}^G = 2.39$ are those values of $x_{\nu-1}^H$ that have the closest Euclidian distances relative to x_{t-1}^G. For example, for k=3 Figure 2.1 shows that from all the values $x_{\nu-1}^H$, ν =2, …, N+1 the set $\{x_{\nu-1}^H\} = \{2.39, 2.36, 2.43\}$ are the three values having distances {0.00, 0.03, 0.04}, respectively relative to the feature value $D_1 = 2.39$ (note that for convenience the distances have been ordered in increasing order of magnitude).

(c) The potential successors of x_{t-1}^G are the values of x_ν^H that correspond to the k-nearest neighbors (as referred to in (b) above). For example, Figure 2.1 shows that for k=3 the potential successors are $\{x_\nu^H\} = \{2.05, 2.55, 2.38\}$, which correspond to the successive values of $\{x_{\nu-1}^H\}$.

(d) From the k potential successors $\{ x_\nu^H \}$ one is selected using the weights

$$w_i = \frac{1/i}{\sum_{j=1}^{k} 1/j}, \qquad i = 1,\ldots,k \tag{2-1}$$

where w_1 is the weight that corresponds to the smallest distance. For example, for k=3, $w_1 = 1/(1/1+1/2+1/3) = 6/11$=0.545, $w_2 = 3/11$ =0.273, and $w_3 = 2/11$=0.182, where $w_1 = 0.545$ corresponds to the first value in the set $\{x_v^H\}$, i.e. x_v^H=2.05. The selection is made at random using the cumulative uniform distribution with values 0.545, 0.818, 1.000. For example, if the uniform random number drawn is 0.625, then the second value 2.55 is selected so that x_t^G=2.55.

(e) The steps (a) to (d) are repeated until the desired generated sample size is obtained.

The good features of the KNNR approach are: (1) preserving the marginal distribution, (2) reproducing linear or nonlinear dependence in the historical data, and (3) easy to extend to higher order dependence and multi-dimension. However, there are a few significant drawbacks: (1) It generates only historical values since it is a resampling technique. This can be a serious drawback because one would expect that synthetic streamflows may exceed the observed maximum and also fall below the observed minimum value. (2) The inter-annual variability will not be preserved unless an appropriate term is included in the feature vector D_t. Up to the present this has not been accomplished and the lack of preservation of interannual variability remains a shortcoming in using the KNNR method. (3) The variability is deflated where using the KNNR technique. The variance deflation has been examined by Buishand and Brandsma (2001) using varying k. They suggested that utilizing an appropriate value of k (around 2

to 5) minimize the deflation. The cause of this variance deflation is further explored below.

Young (1994) argued that data points lying at or near the edges of the cloud of points (refer to Figure 2.1) would not be selected as those located well within the clouds of points. Thus isolated points such as that shown in Figure 2.1 will be undersampled (the clouds of points mean those in the range of high frequency.) KNNR is dissimilar to the bootstrapping method regarding this phenomenon. Each historical data points have equal probability to be selected. However, the resampling data in KNNR procedure have different chances to be selected especially when the points are near the edges of the clouds. In Figure 2.1, two isolated points are illustrated. The isolated points are more likely to be extreme values. If the historical data have the tendency to generate less probability on near extreme values, then the variance of the generated data will be lessened. To investigate further, the histogram of the number of times that each data points are selected is drawn for each month at Figure 2-B.1 from the KNNR simulation of Lees Ferry site at Colorado River. One hundred data sets are simulated with the same data length as historical (98 years). Since the 100 data sets are simulated, each data points should be generated around 100 times. However, some values are generated only 20 to 40 times. The selected times of each data points versus the order of data points are displayed in Figure 2.2(left) for the KNNR simulation. Y-axis presents the increasing order of the data points. The order of the data points which are less than 80 times more likely to be low or high in Figure 2.2. The data points in low order with less than 80 times has 5 points and the ones in high order 10 points. Unlike, bootstrapping simulation, the uniform selection of the historical data, does not show this behavior. In the right side of Figure

2.2, uniform distribution of order according to times sampled is shown for bootstrapping method. It is obvious that the undersampling near extreme values will induce the underestimation or deflation of variance. Furthermore, one might assume that this predisposition will be raised when the data are highly correlated. If the data are serially independent, then the KNNR procedure will not produce this bias.

To examine the relation between the variance deflation and serial correlation, a brief experiment has been performed over the KNNR model using the Beaver River monthly streamflow at Utah as in Lall and Sharma (1996). Here, k=30≈$N^{4/(d+4)}$ is employed as suggested by Fukunaga (1990) and d entails selected lag which is one in this case. From one hundred set of the synthetic data with the same record length, the average value of the relative variance bias (described below) and the centered lag-1 correlation is illustrated in Figure 2.3. The purpose of this experiment is to examine the effect of serial correlation onto the variance of the generated data from KNNR. k=30 might be a little awkward for KNNR. However, this setup will more clearly visualize the effect of the serial correlation that might be hidden. The relative bias of variance at month τ (rv_τ) is denoted as

$$\mathrm{rv}_\tau = \frac{\sigma_\tau^2(X^G) - \sigma_\tau^2(X^H)}{\sigma_\tau^2(X^H)} \qquad (2\text{-}2)$$

where $\sigma_\tau^2(X)$ is the variance of X for month τ. If the value is less than zero, the variance from KNNR model $\sigma_\tau^2(X^G)$ is less than $\sigma_\tau^2(X^H)$ which implies underestimation of the variance with the KNNR model. This statistic is the measurement of the deflation scaled with the historical variance. In Figure 2.3, it reveals that the higher lag-1 correlation, the

lower the relative variance bias. And $^{rv_\tau}$ is negative for all months. It represents that the monthly variance of the synthetic data from the KNNR model is underestimated. And, it is obvious that the bias is negatively related to lag-1 correlation from Figure 2.3, saying that the lag-1 correlation each month leads underestimation of the variance from KNNR model. The higher serial correlation leads the higher deflation of the variance. If k is small (around 2 or 5), the effect of serial correlation is diminished. However, there is still some deflation in case of small k, when the data are serially correlated. This deflation from KNNR is unavoidable since the applied data for simulation are always significantly correlated. If the data are serially independent, simply bootstrapping method can be applied, alternatively. Later, this bias will be compensated through applying the smoothing kernel.

2.2.2 Local Regression with KNN for innovation (LRK)

To improve the KNNR model Prairie et al. (2006) adopted a nonlinear local polynomial regression with the innovation sampled from KNNR and applied it to monthly streamflow generation for the Colorado River at Lees Ferry. The LRK model is given by

$$X_t = g(X_{t-1}) + e_t \tag{2-3}$$

where $g(X_{t-1})$ is a local polynomial and e_t is the residual. For more detail on a local polynomial fit, readers are referred to Simonoff (1996). After fitting the local polynomial the residuals are estimated as $e_t = X_t - g(X_{t-1})$, which then are employed for generation using KNNR.

The good features of LRK includes: (1) values other than historical are simulated, (2) any arbitrary relationship (linear or nonlinear) that is present in the observed data is captured, and (3) heteroscedasticity can be reproduced (Lee and Salas, 2006). On the other hand the shortcomings of the LRK approach are: (1) Negative values may be generated because of the error term, i.e. if $e_t < 0$ and $g(X_{t-1}) < |e_t|$ the generated value will be negative. For highly skewed data, this may occur frequently. (2) The variation generated from this model is limited to a directional pathway as depicted in Figure 2.4. In the figure, the relationship between the generated flows for months 3 and 4 is shown for the Colorado River at Lees Ferry using LRK (the length of the generated data was the same as the length of the historical, i.e. 90 years). This is a natural behavior for any hybrid model such as local regression and KNNR innovation, or PAR(1) and KNNR innovation since the innovation resampling arises from a limited number of data points. (3) It does not preserve the inter-annual variability. The variance of the generated annual series will be degraded and the correlation of the yearly series will not be preserved.

2.2.3 NP and NPL

Utilizing the conditional kernel density estimate, a nonparametric alternative to the lag-p autoregressive (NP) model has been developed by Sharma et al.(1997). The conditional kernel density with normal kernel on the random variables X_t and X_{t-1}, the corresponding values x_t and x_{t-1} is denoted as

$$\hat{f}_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \sum_{i=1}^{n} \frac{1}{(2\pi\lambda^2 S')^{1/2}} w_i \exp\left(-\frac{(x_t - b_i)^2}{2\lambda^2 S'}\right) \tag{2-4}$$

where $x_1,....,x_n$ are observed data, S_{12} = Cov($X_t X_{t-1}$), $S_{11} = S_{22} = \sigma^2(X)$, and $S' = S_{11} - S_{12}^2 S_{22}$. In addition,

$$b_i = x_i + (x_{t-1} - x_{i-1}) S_{11} / S_{12} \quad , \tag{2-5}$$

and

$$w_i = \frac{\exp\left(-(x_{t-1} - x_{i-1})^2 / 2\lambda^2 S_{22}\right)}{\sum_{j=1}^{n} \exp\left(-(x_{t-1} - x_{j-1})^2 / 2\lambda^2 S_{22}\right)} \quad , \tag{2-6}$$

and λ is a smoothing parameter (described below in some detail).

The generation procedure based on Eq.(2-4) is as follows:

(a) Two alternatives for initializing procedure are suggested as

a1. to set X_0 equal to mean and remove warm-up period

a2. to select one of the historical values with equal probability for x_0 and generate from $N(x_0, \lambda^2 S_{11})$

(b) From the given value $X_{t-1} = x_{t-1}$, select one of the observation x_i according to the weight w_i

(c) Simulate X_t from normal distribution $N(b_i, \lambda(S')^{1/2})$.

(d) Repeat the step (b) to (c) until the desired length of data are simulated.

The variance for this model denoted by $\sigma^2(X)'$ is (Sharma et al., 1998)

$$\sigma^2(X)' = S_{11}(1+\lambda^2) = \sigma^2(X) + \lambda^2\sigma^2(X) \tag{2-7}$$

This indicates that the variance of the generated series will be inflated (or overestimated) as much as $\lambda^2\sigma^2(X)$. The estimation of the smoothing parameter λ is crucial to estimating the density accurately. One of the most common approaches for estimating λ is by Least Square Cross Validation (LSCV), to minimize the Integrated Square Error (ISE) and simplified as

$$LSCV(\lambda) = R(\hat{f}(x)) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x) \tag{2-8}$$

where $R(g(x)) = \int g(x)^2 dx$, $\hat{f}(x)$ and $\hat{f}_{-i}(x)$ denote a marginal kernel density estimate of x and the density estimate excluding the i^{th} observed data point, respectively.

Furthermore, Sharma and O'Neil (2002) developed a nonparametric alternative that incorporates the interannual variability (NPL) for monthly simulation. The model denoted as NPL is based on conditioning the variable $X_{\nu,\tau}$ on $X_{\nu,\tau-1}$ and the summation of the previous 12 months, i.e. $\hat{f}_{X_{\nu,\tau}|X_{\nu,\tau-1},Z_{\nu,\tau}}(x_{\nu,\tau} \mid x_{\nu,\tau-1}, z_{\nu,\tau})$ where $Z_{\nu,\tau} = \sum_{j=1}^{12} X_{\nu,\tau-j}$. Note that referring to the subscripts of $X_{\nu,\tau-j}$ in the summation, if $\tau - j \le 0$ then ν must be changed by $\nu - 1$ and $\tau - j$ by $12 - |\tau - j|$. The merits of NPL are: (a) preserves the marginal distribution, (b) reproduces the linear or nonlinear relation embedded in the observed data, and (c) preserves the interannual variability. Nevertheless, there are a few drawbacks such as: (i) The variance of the generated series will be overestimated as

mentioned above. Overestimation of the variance might lead to the exaggeration of the extreme events. This will be corroborated in the simulation results later. (ii) Since it is a normal kernel based model, negative values may be generated unless a modification is conducted. In Sharma et al. (2002), variable kernel is employed to prevent generating negative values. The idea behind is to adjust the smoothing parameter λ such that the probability of generating negative values is not greater than a certain threshold (e.g. 6 percent in the referred paper). However, the variable kernel will lead to larger higher bias on density estimate (Simonoff 1996). This will be elucidated more clearly below in a subsequent section.

2.3 Proposed methods

2.3.1 KNNR with Gamma KDE (KGK)

Since KNNR is a resampling algorithm with discrete conditional density estimate it produces generated values that are identical to the historical values (except in different order). Furthermore, the review of the KNNR model in Section 2.1 above suggests that the variance of the generated data becomes underestimated. To surmount this limitation, a perturbation of the value x_i obtained from KNNR is suggested. As presented in Figure 2.1, the selected historical value (x_i =2.55) from KNNR is treated as the center of a kernel in the kernel density estimate. And a value is generated (perturbed) from a Kernel density according to the smoothing principle of KDE. This is related to the generation from normal distribution with b_i and $\lambda^2 S'$ for mean and variance in NP method (explained in Section 2.3). The main difference between NP and the suggested approach is that NP model generates data with conditional nonparametric distribution while the proposed

approach here uses the KNNR to find x_i (comparable to b_i in NP) and perturbs the value x_i with a selected kernel (e.g. gamma) instead generation from $N(b_i, \lambda(S')^{1/2})$. The perturbation process is performed independently on the previous condition (X_{t-1}). The independent perturbation might weaken the relation between X_{t-1} and X_t. But if KNNR reflects the relation appropriately, the overall bias might not be significant because perturbation is performed with centering the selected value (x_i). The underestimation of the relation is investigated thoroughly in result section. Furthermore, the variance of the generated value is increased from this perturbation procedure since it is more likely to add randomness. The source of the variance comes from two parts such as the selection of x_i with KNNR and the perturbation into the selected value. However, this will not be problematic since we review that the KNNR process underestimates the variance. The underestimation will be compensated with the additional variance from the perturbation. More detail will be discussed about this in the smoothing parameter estimation section later.

The properties of Gamma Kernel Density Estimator

Since the perturbation process is based on the Kernel density estimate, the suggested model requires the selection of a Kernel and the smoothing parameter. Generally, a Gaussian kernel is employed for kernel density estimation in Nonparametric modeling literature (Sharma et al., 1997). The kernel, however, is unbounded. This is shortcoming for generating hydrologic data that are positively skewed and bounded by zero. If the data is highly skewed, the density estimate using a normal kernel is significantly biased and the cumulative probability below zero may be significant. This

indicates that the significant amount of generated data will be negative that is not physically suitable for hydrologic data. To avoid the bias and bound, many different approaches are suggested such as boundary kernels, varying the bandwidth, and transformation-based estimation (Simonoff, 1996). Furthermore, the other types of kernels are developed for the skewed and bounded data such as exponential kernel (Mugdadi, 2004), beta kernel (Chen 1999), and gamma kernel (Chen 2000). The exponential kernel is not smoothed even with high smoothing parameter because of the discontinuity of the exponential distribution nature in the estimate. This leads the unsmoothness in the point of each historical value. The beta kernel is bounded in both sides. Typically, hydrological data such as streamflow is bounded at zero and unbounded for $x > 0$. Therefore, the gamma kernel is most desirable for hydrological data.

Chen (2000) proposed the gamma kernel as

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K_{x/h+1,h}(X_i) \tag{2-9}$$

where

$$K_{x/h+1,h}(t) = \frac{t^{x/h} e^{-t/h}}{h^{x/h+1}\Gamma(x/h+1)} \tag{2-10}$$

is a gamma kernel with shape parameter $\alpha = x/h+1$ and scale parameter $\beta = h$, X_i is the random sample of size N, h is the bandwidth of the gamma kernel, and $\hat{f}(x)$ is the density estimator evaluated at x. Chen (2000) showed that the gamma kernel density estimate achieves the optimal rate of convergence for the mean integrated squared error,

and the variance of the gamma kernel estimator gets smaller as x increases. The mean and variance of the gamma kernel above are $x+h$ and $xh+h^2$, respectively, and x is the mode (Chen, 2000). If one uses Chen's gamma kernel for non-parametric data generation as suggested above, it will produce some bias in the mean although it will avoid generating negative values.

In the generation procedure proposed in this paper, a point say $X_i = x$ obtained from the KNNR method will be perturbed with the gamma kernel. If one uses Chen's gamma kernel the mean will be $x+h$, so the mean of the generated data will be overestimated as much as h. Instead, another type of gamma kernel is suggested here to avoid this bias as:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K_{x^2/h^2,h^2/x}(X_i) \tag{2-11}$$

where

$$K_{x^2/h^2,h^2/x}(t) = \frac{t^{x^2/h^2-1}e^{-t/(h^2/x)}}{(h^2/x)^{x^2/h^2}\Gamma(x^2/h^2)} \tag{2-12}$$

where in this case $K_{\alpha,\beta}(t)$ is the gamma kernel with shape parameter $\alpha = x^2/h^2$ and scale parameter $\beta = h^2/x$. The mean and variance of the gamma kernel are $\mu(t) = x$, $\sigma^2(t) = h^2$, respectively. Thus the gamma kernel above is formatted so that the generated data from the suggested approach will preserve the mean. And also the variance for this gamma kernel is independent on x so that the magnitude of the variance from this kernel

is simple to manipulate. Next we verify properties of the gamma kernel density estimator such as the bias and variance.

The expected value of $\hat{f}(x)$ of (11) may be expressed as

$$\begin{aligned} E\{\hat{f}(x)\} &= E\left\{\frac{1}{N}\sum_{i=1}^{N} K_{x^2/h^2,h^2/x}(X_i)\right\} = \frac{1}{N}\sum_{i=1}^{N} E\left\{K_{x^2/h^2,h^2/x}(X_i)\right\} \\ &= E\left\{K_{x^2/h^2,h^2/x}(t)\right\} = \int_0^{\infty} K_{x^2/h^2,h^2/x}(y) f(y)dy = E\{f(Z)\} \end{aligned} \tag{2-13}$$

where Z is $Gamma[x^2/h, h^2/x]$ with mean $\mu(Z) = x$ and variance $\sigma^2(Z) = h^2$.

To find $E\{f(Z)\}$ we will use the Taylor series expansion up to the second order as in Chen (2000), i.e.

$$\begin{aligned} E\{f(Z)\} &\approx f(a) + \left.\frac{\partial f(Z)}{\partial Z}\right|_{Z=a} E[Z-a] + \left.\frac{\partial^2 f(Z)}{\partial Z^2}\right|_{Z=a} E[Z-a]^2 \\ &= f(x) + \frac{1}{2} f''(x)\sigma^2(Z) = f(x) + \frac{1}{2} f''(x) h^2 \end{aligned} \tag{2-14}$$

where $a = \mu(Z) = x$. Therefore,

$$E\{\hat{f}(x)\} \approx f(x) + \frac{1}{2} h^2 f''(x) \tag{2-15}$$

And the bias is

$$Bias[\hat{f}(x)] = E\{\hat{f}(x)\} - f(x) \approx \frac{1}{2} h^2 f''(x) \tag{2-16}$$

In addition, the variance of the density estimator is derived in Appendix A. It gives

$$Var\{\hat{f}(x)\} \approx \begin{cases} \dfrac{1}{2N h\sqrt{\pi}} f(x) & \text{if } x/h \to \infty \\ \dfrac{\Gamma(2\kappa^2 - 1)}{N\Gamma^2(\kappa^2)} (h/\kappa^2)^{-1} 2^{-2\kappa^2+1} f(x) & \text{if } x/h \to \kappa \end{cases} \tag{2-17}$$

For comparison the bias and variance of the density estimator from Chen's Gamma Kernel is (Chen, 2000)

$$Bias[\hat{f}(x)] = E\{\hat{f}(x)\} - f(x) \approx hf'(x) + \frac{1}{2} x f''(x) \tag{2-18}$$

$$Var\{\hat{f}(x)\} \approx \begin{cases} \dfrac{1}{2Nh\sqrt{\pi}} x^{-1/2} f(x) & \text{if } x/h \to \infty \\ \dfrac{\Gamma(2\kappa - 1)}{2^{2\kappa+1} nh\Gamma^2(\kappa+1)} f(x) & \text{if } x/h \to \kappa \end{cases} \tag{2-19}$$

Comparing Eqs.(2-16) and (2-18) one may see that the bias of the kernel density estimator from Chen (2000) has the unpleasant term $f'(x)$. In addition, the second term of the bias in Eq.(2-18) increases with x as opposed to that in Eq.(2-16), which does not depend on x. Thus the suggested Gamma Kernel of Eq.(2-12) leads to smaller bias than the Gamma Kernel by Chen (2000). On the other hand, the variance of the density estimator by Chen (2000) has a better feature when x/h goes to infinity, i.e. the variance decreases, but this does not occur when x/h goes to zero. In conclusion, the results of the bias and variance of the two gamma kernels promulgate that the suggested gamma kernel density estimate is comparable to the kernel of Chen (2000). The applicability of

the suggested gamma kernel density estimate has been verified by data generation. Even though the estimation of the density as such has nothing to do with generation, it is important to assess whether the suggested gamma kernel density estimate is acceptable to apply.

Selection of the bandwidth h for the gamma kernel and the number of neighbors k for KNNR

The variability of the suggested model will come from two sources, KNNR and gamma Kernel density. Therefore, the number k is reduced to $\sqrt{N}/2$ instead of $\sqrt{N}$ suggested by Lall and Sharma (1996), since $20 \le N \le 100$ is in the range between 2 to 5. This also effects the lagged correlation since the smaller number of k results more similar relation to historical data. The lower variability from the smaller k will be compensated through the Gamma kernel perturbation.

The kernel smoothing (perturbation process) yields an extra variance in the simulated data. The suggested generation method KGK is made up of two components: (1) a generated variable say $X' = x'$ obtained from KNNR and (2) a perturbation gamma variable say $X'' = x''$ such that the mean is equal to x' and the standard deviation is equal to the smoothing parameter h. Consequently the variance of the generated variable $X = X' + X''$ is

$$Var(X) = Var(X') + Var(X'') = \sigma^2(X') + h^2 \qquad (2\text{-}20)$$

where $\sigma^2(X')$ refers to the variance obtained from KNNR. ($\sigma^2(KNNR)$) is not explicitly known, it reveals that the variance of the resampled data from KNNR tends to

be underestimated as described in KNNR review section. As indicated in the review section above the variance obtained from KNNR is smaller than the historical variance (i.e. KNNR underestimates the variance). But as suggested in the foregoing analysis such underestimation will be compensated by the variance induced by the gamma random term. Thus, the smoothing parameter h (bandwidth) has two missions: (1) to smooth out the historical values so that the generated data values are placed over the physical range of the hydrologic data and (2) enhance the variance of the generated data.

A possible approach for estimating the bandwidth is the Least Square Cross-Validation (LSCV) as in Chen (2000). It is approximate estimation procedure and requires a fair amount of computation. Instead, an heuristic estimation approach is suggested here as

$$h = \frac{\sigma(X)}{k} = \frac{\sigma(X)}{\sqrt{N}/2} \tag{2-21}$$

Note that as $N \to \infty$, $h \to 0$. This is basic characteristics of the bandwidth. Since the number of data increases as infinity implying that the population of the data is known, the variance burdened into smoothing parameter should be diminished. Also note that the smoothing parameter for normal distribution with normal kernel in the context of minimizing the approximate mean integrated square error is $h = 1.06\sigma(X)N^{-1/5}$ (Silverman, 1986). In this case for a data range $20 \le N \le 100$ the bandwidth h is in the range $0.58\sigma(X)$ and $0.42\sigma(X)$. Since this is for the normal distribution with normal kernel which is one of the most smoothed distributions and season streamflow data tends to be skewed, the smoothing parameter for gamma kernel should be less than this

magnitude. Eq.(2-21) is in the range between $0.45\sigma(X)$ and $0.2\sigma(X)$ for $20 \le N \le 100$ reaching this argument.

2.3.2 Description of the new models

KGK model is to model the dependency structure with KNNR analogous to $f(X_{\nu,\tau} \mid X_{\nu,\tau-1})$ and smoothing with Gamma Kernel perturbation. The KGK based on only the previous month quantity $X_{\nu,\tau-1}$ cannot reproduce satisfactorily the interannual variability. To enhance the model capability to reproduce long-term variability, an additional term should be included as a conditional variable, i.e. $f(x_{\nu,\tau} \mid x_{\nu,\tau-1}, \Psi)$ where Ψ is the addition variable to consider the interannual variability. For this purpose, two schemes are suggested here: (1) employing the aggregate flow variable of the previous p months analogous to the NPL model and (2) utilizing the yearly value generated from separate yearly model to specify the condition of a certain year for monthly time scale generation. The specific description on each model is followed.

Gamma KDE on KNNR with the aggregate variable (KGKA)

The conditional term for interannual variability is the moving aggregate flow variable

$$z_{\nu,\tau} = \sum_{j=1}^{\omega} x_{\nu,\tau-j} \tag{2-22}$$

As noted before in Eq.(2-22) if $\tau - j \le 0$, then $x_{\nu,\tau-j}$ becomes $x_{\nu-1,\omega-|\tau-j|}$. The term $z_{\nu,\tau}$ represents the sum of the previous ω seasons. Since we will find our generated

value $x_{\nu,\tau}^{G}$ by conditioning on $x_{\nu,\tau-1}^{G}$ and $z_{\nu,\tau}$, it is necessary to determine the weighted Euclidean distance between the generated and historical $x's$ of the previous time $\tau-1$ and between the generated and historical sums $z's$ of the previous ω seasons. Thus the weighted distance denoted by $r_{t(\nu,\tau)}$ is given by

$$r_{t(\nu,\tau)} = \left\{w_{\omega}(x^{H})[x_{t-1,\omega}^{G} - x_{\nu-1,\omega}^{H}]^{2} + w_{1}(z^{H})[z_{t,\tau}^{G} - z_{\nu,\tau}^{H}]^{2}\right\}^{1/2} \text{ for } \tau = 1,\ \nu > 1,\ t > 1 \quad \text{(2-23a)}$$

and

$$r_{t(\nu,\tau)} = \left\{w_{\tau-1}(x^{H})[x_{t,\tau-1}^{G} - x_{\nu,\tau-1}^{H}]^{2} + w_{\tau}(z^{H})[z_{t,\tau}^{G} - z_{\nu,\tau}^{H}]^{2}\right\}^{1/2} \quad \text{for } \tau > 1,\ \nu > 1 \qquad \text{(2-23b)}$$

Note that the calculations of r begins at t=2 and $\tau = 1$. The scaling weights $w_{\tau-1}(x^{H})$ and $w_{\tau}(z^{H})$ are given by the inverse of the variance of $x_{\nu,\tau-1}^{H}$ and $z_{\nu,\tau}^{H}$, respectively. Also Mahalanobis distance may be employed as more elaborate work to relation, which is suggested by Yakowitz and Karlsson (1987) for best prediction in least square sense. However, it encumbers on computation and no significant difference is found from the test performed in this study. The benefits from including the term $z_{\nu,\tau}$ are: (a) to take into account the dependency between the current's month's flow and the previous year flow, i.e. the summation of the previous ω seasons, and (b) self-constructed on yearly time scale meaning that it does not require any yearly time series models and values.

The procedure for simulating the synthetic data is:

(1) Estimate the smoothing parameters k and h following the method suggested above.

(2) The initial value $x_{1,1}^{G}$ is randomly selected from the historical data set $x_{\nu,1}^{H}$ where $\nu=1,\ldots,N$. Each historical data has an equal chance to be selected.

(3) General KNN Resampling process in Chapter 3.2 (a)-(d) is employed for data generation of the rest months of the first year. From the second year, the following processes are employed such that $x_{t,\tau}^{G}$ where t=2,…,N and τ=1,…,ω.

(4) At first, the aggregate variable of the generated data are obtained with $z_{t,\tau}^{G}=\sum_{j=1}^{\omega} x_{t,\tau-j}^{G}$. The k-nearest neighbors are obtained from the estimated distances employing Eq.(2-23 a and b). From the same selection procedure in KNNR, the successor of the selected one among k values is taken, say $x_{t,\tau}^{*}$. This step is repeated until the required generation sample is filled.

(5) Perturb with Gamma Kernel Density Estimate meaning that generate from the gamma distribution $x_{t,\tau}^{G}=Gam[x_{t,\tau}^{*\,2}/h^{2},h^{2}/x_{t,\tau}^{*\,2}]$.

Gamma KDE on KNNR with the pilot variable (KGKP)

It is not easy task to generate seasonal streamflow data with long-term (yearly) variability such as Hurst Phenomenon, Shifting-mean, and climate change as well as common serially correlated structure. Here, we suggest a seasonal simulation model that from modeling or obtaining yearly series separately, the data obtained from a simulation

model or other sources is used as the pilot variable. It presents that the known yearly value will be used as conditional variable which adds to the feature vector of the KNNR model. Yearly data with the unique features mentioned above can be generated from yearly model [e.g. FGN (Hurst Phenomenon), Shifting Mean Level (Shifting-Mean Level)], and denoted as x_t'. For example, if x_t' is lower than normal implying drought condition, this will lead the tendency that the whole monthly values of the current year be small. The feature vector of KNNR algorithm is described as

$$r_{t(\nu,\tau)} = \left[w_1 (x^G_{t,\tau-1} - x^H_{\nu,\tau-1})^2 + w_2 (x_t' - x^H_\nu)^2\right]^{1/2} \tag{2-24}$$

The merits of this scheme is that (1) any long-term variability can be adopted into the synthetic seasonal data, (2) no complex unnecessary structure is included, and (3) it is very straightforward to apply the long-term variability into seasonal system structure. This model is not a disaggregation model in that the pre-obtained yearly values are only employed for leading the status of the current year. As an alternative, El Niño/Southern Oscillation (ENSO) index can be employed to define the status of the current year.

The process of KGKP model is followed as:

(1) Estimate the smoothing parameters k and h.

(2) Fit a model for yearly data for the pilot variable x_t'. The same yearly data or exogenous variable might be employed for this variable. And generate annual data for the pilot variable x_t' where t=1,…,N^G and N^G is the required generation length.

(3) The initial value $x_{1,1}^{G}$ is obtained with the same procedure as mentioned in the previous KGKA model (random selection from the historical data of the first month). The other months and years use the following procedure.

(4) The current yearly state x_t' and the previous month state $x_{t,\tau-1}^{G}$ are compared with the historical data and measures the distance with the Eq.(2-24). Obtain the resampled value from the k-nearest neighbors and the weighted probability. The resampled value is assigned into $x_{1,2}^{*}$. With this value and the selected bandwidth from Eq.(2-21), the final generated value will be obtained with Gamma Kernel perturbation. These selection procedures are exactly the same as the KNNR model.

(5) Perturb with Gamma Kernel Density Estimate meaning that generate from the gamma distribution $x_{v,\tau}^{G} = Gam[x_{t,\tau}^{*\,2} / h^2, h^2 / x_{t,\tau}^{*\,2}]$.

The steps (4)-(5) is repeated until the required length of the data generation is met.

2.4 Data Description and Model Assessment

To assess the suggested models, two sets of monthly streamflow data are applied, the Colorado River at Lee Ferry (site 20 of the Colorado River System) and the Niger River at Koulikoro. The historical data of the Colorado River has been naturalized and partially extended for the period 1906-2003 (Lee and Salas, 2006). The historical streamflow data of the Niger River at Koulikoro has been used for the period 1907-1979 to avoid the effect of reservoir regulation after 1979. The tested models for site 20 in

CRS are (1) NPL, (2) Gamma KDE on KNNR with the aggregate variable (KGKA), and (3)Gamma KDE on KNNR with the pilot variable generated from Shifting mean level model (KGKP with SML) and for Niger River flow data at Koulikoro are (1) NPL, (2)Gamma KDE on KNNR with the aggregate variable (KGKA), and (3) Gamma KDE on KNNR with the pilot variable generated from Shifting mean level model (KGKP with SML).

To test the models, one hundred sets of synthetic monthly streamflow data of the same length as the historical data were generated from each model. A number of basic statistics are calculated from each which are displayed using boxplots. For example, Figure 2.5 shows the basic statistics such as the mean, standard deviation, skewness, lag-1 serial correlation, maximum, and minimum constructed from the generated data obtained from a given model. The end line of the box implies the 25 and 75 percent quantiles while the cross lines above the box on the whisker correspond to the 90 percent quantile and the maximum, while the cross lines below the box on the whisker represents the ten percent quantile and the minimum. And the 'x' mark and the segment line connecting the x mark represent the historical statistics. The comparison of boxplot for the referred statistics has been done for both monthly and annual time scales. The kernel density estimate of the generated data is also compared to that of the historical data.

In addition, various drought and surplus statistics as well as storage capacity have been estimated and compared on historical and generated yearly data from the selected models above such as maximum drought length, maximum drought amount, maximum surplus length, maximum drought amount, storage capacity. Maximum drought length is defined as longest length of the deficit which is shortage from the water demand during

the year. The demand is denoted as the mean value multiplied by threshold level [0.6, 0.7, 0.8, 0.9, and 1.0]. Maximum amount length is the maximum amount of the deficit. Surplus is excessive water over the demand. The definition on surplus is the similar to drought. The storage capacity is the extent to which streamflows can be stored for later release. The sequent peak algorithm is employed for estimation of this statistic (Loucks et al., 1981).

2.5 Evaluation and Comparison of the methods

2.5.1 Colorado River Streamflow at Lees Ferry

The time series of yearly streamflow for Colorado River at Lees Ferry is illustrated in Figure 2-B.3. Notice that the of the fist 20 years has higher flows and significant drought period at the last 5 years which is one of the worst droughts in Colorado River. The key monthly statistics of historical and generated data from three selected model is shown in Figure 2.5-Figure 2.7. The behavior of the generated data from KGKA and KGKP are very similar to each other (Figure 2.5 and Figure 2.6). Every key statistics of both models are well preserved except slight underestimation in lag-1 correlation through all months. The underestimation of lag-1 correlation comes from the weakness represented by KNNR algorithm. This behavior is observed also in pure KNNR model (Lall and Sharma, 1996). For the statistics from NPL model, the standard deviation and maximum are overestimated and underestimation for skewness and minimum (Figure 2.7). These deviations from the NPL model imply that the model does not appropriately reproduce the historical distribution. Furthermore, the inflation of the standard deviation by NPL model is expected as Eq.(2-7). In KGKA and KGKP model, however, the

deflation of the variance by KNNR model is compensated by using the Gamma KDE (Figure 2.5 and Figure 2.6). The inflation of the NPL model effects the overestimation of maximum and the underestimation of minimum as shown in Figure 2.7. Furthermore, the underestimation of the skewness (Figure 2.7) is induced from the nature of the Gaussian kernel especially for the highly skewed months such as October, February, and September. The KDE based on symmetric kernel has some difficulty to preserve the high-skewness. From Figure 2-B.4 to Figure 2-B.6, the scatter plots of the generated and historical monthly data to show how well the generated data will reproduce the relation. The generated data of KGKA and KGKP relatively well reproduce the overall local relation of the historical data while NPL eliminates the local relation of the historical data.

In yearly time-scale statistics as illustrated at Figure 2.8, mean and lag-1 correlation are well preserved through all models. Inter-annual variability in KGKA represented as standard deviation is underestimated while KGKP well preserves this statistics. This notices that the aggregate variable is not good enough to deliver the long-term variability combining KGK model. This variable can deliver more sophisticated inter-annual structure into the downscale generation such as shifting mean process. Conversely, the NPL model overestimates the yearly standard deviation propagated from the overestimation of the monthly variability (Figure 2.7). In case of minimum, the NPL is relatively underestimated through whole months while maximum is overestimated. Since NPL model employs normal kernel in generation which is symmetric distribution. Some negative values might be generated. To avoid negative values, Sharma and O'Neill. (2002) suggests employing variable kernel. However this artificial procedure will leads to the bias (Simonoff, 1996). The drought, surplus, and reservoir statistics of the yearly data

with the threshold presented as the historical yearly mean in Figure 2.9 are comparable through whole models. Those statistics with different thresholds behaves similarly. No significant difference between models can be observed. More detailed figures are presented from Figure 2-B.7 to Figure 2-B.21. The storage capacities of the monthly data with different thresholds (multiplying the basic threshold as the historical mean by the threshold levels, 0.3-1.0) are estimated (Figure 2-B.22, Figure 2-B.23, and Figure 2-B.24 for KGKA, KGKP, and NPL respectively). The preservation of the statistics is comparable to all the models. The storage capacities are underestimated through the range of the threshold levels of 0.3-0.6 and overestimated through 0.7-1.0 in KGKA and KGKP models while the statistics are overestimated through all the range in NPL model.

2.5.2 Niger River at Koulikoro

The time series of the yearly data at Niger River station is shown at Figure 2.10 (time series plot with bar at Figure 2-B.25) with the one example of the generated set from KGKP model. The apparent particular pattern of shifting means is revealed from the figure. The basic statistics of the monthly and yearly similarly behaves at the results of Colorado River Site shown at Figure 2-B.26, Figure 2-B.27, Figure 2-B.28, and Figure 2-B.29. Here, Figure 2.11 shows the KDE of the generated time series of NPL (left) and KGKP (right) and for KGKA at Figure 2-B.30 for months 1, 5, and 9. The densities of the main body (near mode) of the generated distribution from the NPL are underestimated while overestimated in the outside of the main body. The distributional behavior illustrated in Figure 2.11 reflects the inflation of the variation through NPL model. Furthermore, to scrutinize the local and overall relations embedded on the historical data, scatter-plots for the month 8 and month 9 are drawn in Figure 2.12 for

KGKP (upper) and for NPL (bottom). The scatter plot of KGKA is no difference to the one of KGKP so that it is not shown in this paper. From these figures, the overall relation between month 8 and month 9 are well preserved through the models. The localized non-linearity is better preserved in KGKP model (Figure 2.12, upper) while the NPL model blurs the local non-linearity (Figure 2.12, bottom). The drought, surplus, and reservoir statistics with the historical yearly mean as the threshold are shown in Figure 2.13. More detailed descriptions are shown in Table 2-1. The KGKP model well preserves the drought, surplus, and storage statistics while KGKA and NPL model underestimate those statistics (Figure 2.13) especially drought and surplus length and drought amount. The drought, surplus, and storage statistics of yearly data with different threshold (multiplying threshold levels,0.6-1.0, by the historical mean) are estimated and shown at the Table 2-1 and Figure 2.14 (only maximum surplus length is shown, the other plots are referred to from Figure 2-B.31 to Figure 2-B.34). In Table 2-1, the results seem to preserve the estimated statistics because of the high standard deviation. But, the boxplot figures of these statistics show differently such that as an example in Figure 2.14, the maximum surplus length is underestimated at NPL model (Figure 2.14, bottom) and KGKA model (Figure 2-B.37) while KGKP model preserves the statistics fairly well through all different threshold levels (Figure 2.14, upper). The same behavior can be seen for maximum surplus amount and maximum drought length. For storage capacity and maximum drought amount, all the applied models fairly preserve those statistics. Overall, the surplus and drought statistics of different thresholds is well reproduced in KGKP model whereas some bias in NPL and KGKA model except storage capacity and maximum drought amount. The ratios of the storage capacity (the statistics of the

monthly generated data divided by the historical one) with different threshold levels are estimated from 0.3 to 1.0 and are illustrated in Figure 2.15 for KGKP and NPL model. The estimated real values are shown at Figure 2-B.44, Figure 2-B.45, and Figure 2-B.46 for KGKA, KGKP, and NPL respectively. The storage capacities of the historical data are well reproduced in the generated data of the KGKP model through all range 0.3-1.0 while the ones of the NPL model are mostly overestimated through all range except the last part 0.9-1.0. The preservation of these statistics is difficult to reproduce. The KGKP model, however, well reproduces these statistics through all different thresholds. Through the test statistics, it can be concluded that the employment of the pilot variable with proper fitting leads to better preservation of the long-term variability. More clear evidence can be observed in Figure 2.10. It is observable that the historical time series shows the sudden shifting mean process. One set of the yearly generated data obtaining from the summation of the generated monthly data into yearly is also illustrated in Figure 2.10. It is shown that the KGKP model reproduces the particular behavior of the historical long-term variability. To model yearly data with shifting mean and to employ it as the pilot variable is very efficient to reproduce the particular long-term process.

2.6 Summary and Conclusions

Synthetic data in hydrology has performed important roles for planning reservoir capacity, drought analysis, etc. Enhancing the model capability from parametric ARMA type, nonparametric models has been employed for its simplicity and for avoiding the bias through transformation procedure. In this study, a generation model employing KNNR algorithm is proposed to overcome the drawbacks of the KNNR such as

generating new feasible values other than observations and reproducing the interannual variability embedded in historical data. For the first part of enhancement (generating new values), Gamma KDE is proposed. Gamma KDE has been proposed by Chen (2000) but it has not been applied or tested in hydrologic fields. Different setup for Gamma kernel parameterization is proposed to preserve the historical mean and standard deviation better. And the second part of development (preserving the annual variability) is achieved through employing the aggregate variable or the pilot variable. Instead of complicated smoothing parameter estimation, heuristic estimation method is proposed employing the monthly variance and k nearest neighbor. The aggregate variable has been suggested by Sharma and O'Neil (2002) applying conditional kernel density with normal kernel while the pilot variable is suggested from this paper to lead the current yearly state. The pilot variable can be either the model of the same station as the monthly data or the exogenous variable. Here, only the yearly model of the same station is tested. KGKA (employing aggregate variable) and KGKP (employing pilot variable) model has been compared with the NPL model (Sharma and O'Neill, 2002) since it has one of the most recently developed nonparametric techniques with the reproduction of the inter-annual variability. Various streamflow data in different rivers are applied and tested. Here, we show only two stations such as Lees Ferry station at Colorado River and Koulikoro station at Niger River.

Various tests are performed with the generated data such as key statistics of yearly and seasonal time scale and drought, surplus, and storage statistics for monthly and annual time scale with different threshold levels. Some conclusions are derived from the results. KGKA and KGKP has superior to preserve the skewness and the variance of

the monthly time scale but a slight underestimation of the month-to-month correlation is unavoidable while NPL model has better performance on the lag-1 correlation with overestimation of the variance and underestimation of the skewness. The gaussian kernel has the limitation to preserve skewed distribution of the historical data, especially highly skewed data. The KGKP preserves the yearly variance while KGKA underestimate this statistics and NPL model overestimate the variance propagated from the overestimation of the monthly data. In drought statistics, KGKP model has a little bit better performance in case of Niger River data. But the results are not consistent in Colorado River case. Furthermore, it is shown that the particular long-term pattern (e.g. shifting means) can be reproduced through employing pilot variable in yearly time scale in KGKP model while the aggregate variable cannot reproduce the pattern.

Overall, the suggested model such as KGKP and KGKA shows the reliable results to generate a univariate seasonal time series. Furthermore, the model procedure is very simple to apply such that the monthly data is obtained from KNNR including the aggregate variable or the pilot variable and then the data is perturbed through the Gamma distribution. Employing exogenous variable with the KGKP model might be a good topic for future research.

Table 2.1 Drought, surplus, and storage statistics of the historical and generated yearly data (mean±stdev) for Niger River at Koulikoro

	T.L	0.6	0.7	0.8	0.9	1
Max Dr. Leng.	Hist	1	2	3	7	11
	KGKA	0.8±0.7	2±0.9	3.5±1.4	5.3±1.7	9.2±3.2
	KGKP	1.1±1	2.6±2.7	4.1±3.7	7.6±5.1	12.1±7.7
	NPL	1.4±0.8	2.3±1.1	3.6±1.5	5.4±1.8	8.2±3
Max Sur. Leng.	Hist	63	28	25	15	11
	KGKA	56.5±14.9	32.3±12.5	20.6±6.5	10.8±4.2	6.3±2.2
	KGKP	51.3±16.8	36.8±15.7	26.4±13.4	16.4±9.7	10.6±6.4
	NPL	44.3±14	29.8±11.5	19.1±7.6	11.3±3.7	8.1±3.1
Max Dr. Amt.	Hist	2.46	11.54	20.99	42.89	91.81
	KGKA	1.7±1.9	7.8±4.4	19.6±8.7	41.5±15.4	86.3±33.9
	KGKP	2.7±2.7	11.7±12.9	29.2±32	67.8±59.9	138.7±113.4
	NPL	6.5±4.9	14.4±7.6	27.5±13.3	51.5±23.4	91.7±38.1
Max Sur. Amt.	Hist	1260.21	481.77	348.8	206.94	136.08
	KGKA	1099.4±325.2	532.2±220.2	283.1±103.8	137.5±54.9	78.2±29.5
	KGKP	1079.2±417.8	680.1±334.9	417.1±253	218.3±154.5	118±92.9
	NPL	940±326.4	547.1±221.1	301.7±123.2	162±60.1	103.1±42.4
Stor. Cap.	Hist	2.46	11.54	20.99	42.89	101.04
	KGKA	1.7±1.9	7.9±4.5	21.2±10	53.8±26.9	146.7±71.9
	KGKP	2.7±2.9	14±21.8	43±55.8	106.1±104.8	229.7±168.2
	NPL	6.5±4.9	15.1±8.7	31±17.6	63.2±33.2	143.9±71.9

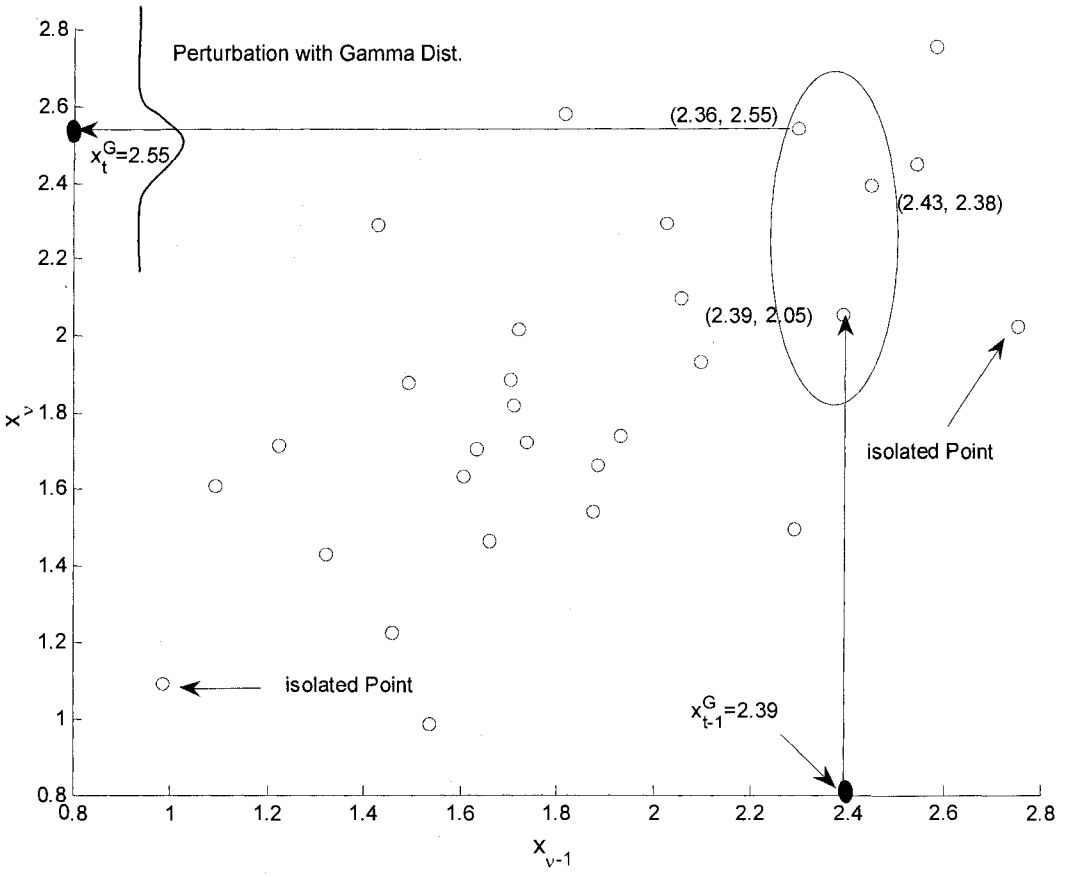


Figure 2.1 Fundamentals of k-nearest neighbor resampling where k=3. The values in parenthesis near the oval refer to the coordinates (x_{v-1}, x_v). Taking x_{t-1}^G=2.39, the 3-nearest historical values (in the oval, x_t^* are 2.55, 2.38, and 2.05, respectively) are selected according to the distance of the feature vector ($D_t = 2.39$) and then selected one value $x_t^G = 2.55$ among three with the probability in Eq.(2-1).

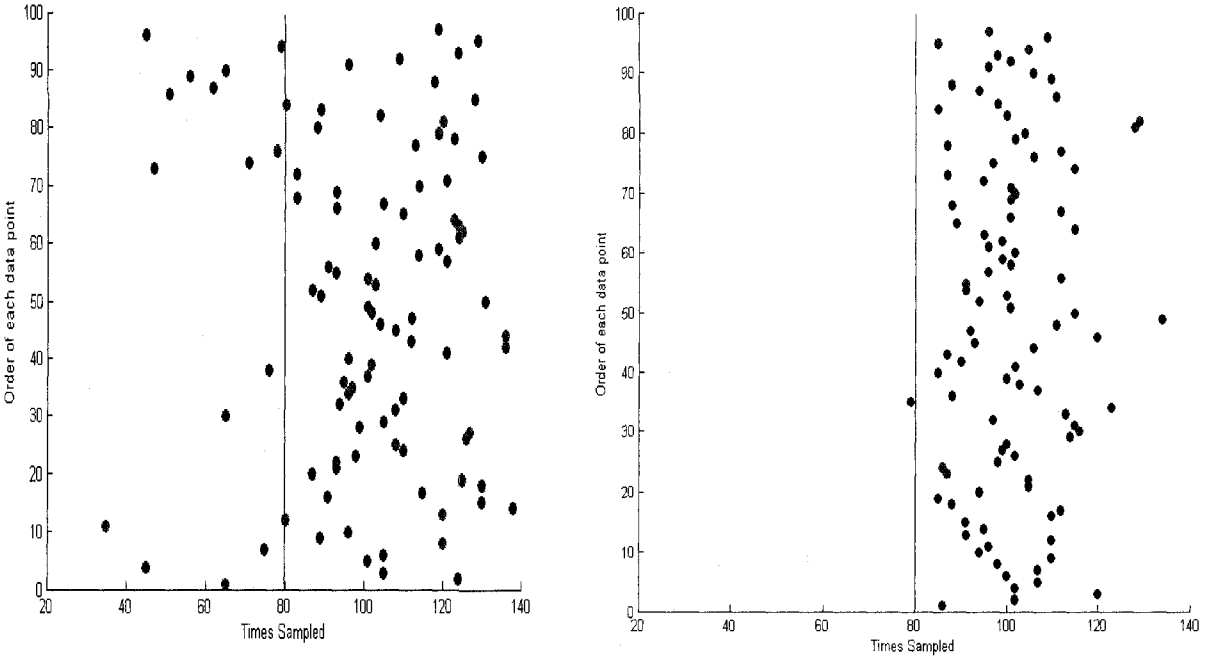


Figure 2.2 Scatterplot of times sampled versus order of the data point from and KNNR simulation (left) and Bootstrapping (right) for month 8 of Site 20 Colorado river; 100 sets are simulated for the length 98 yrs as historical for Colorado River Site 20

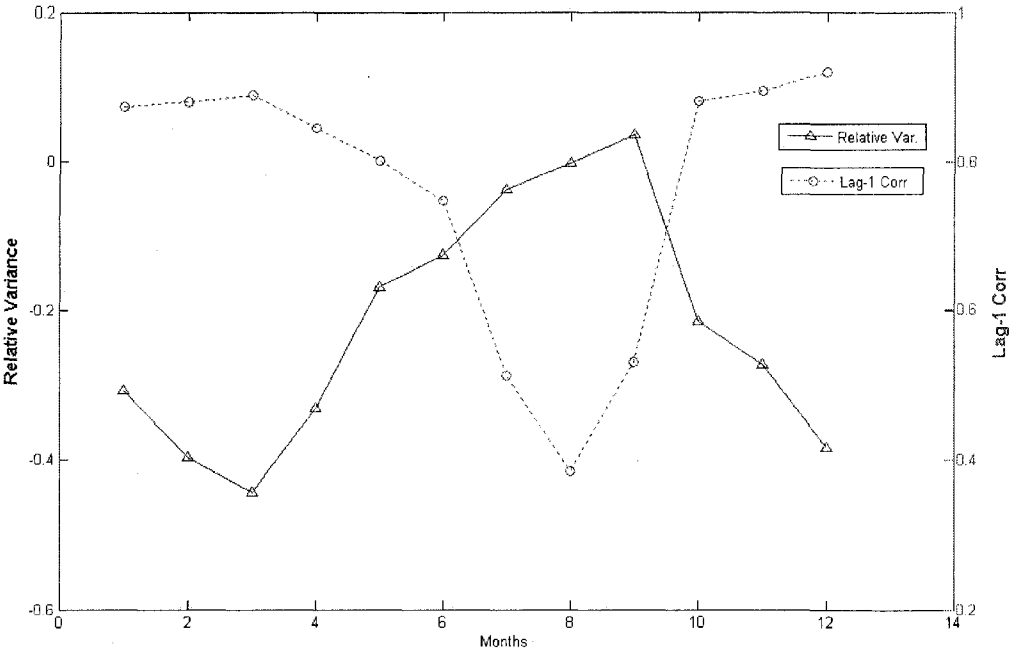


Figure 2.3 Relative bias of variance, Eq.(2-2), (average value of 100 set of the synthetic data from KNNR model) and the lag-1 correlation, for Beaver River monthly streamflow at Utah

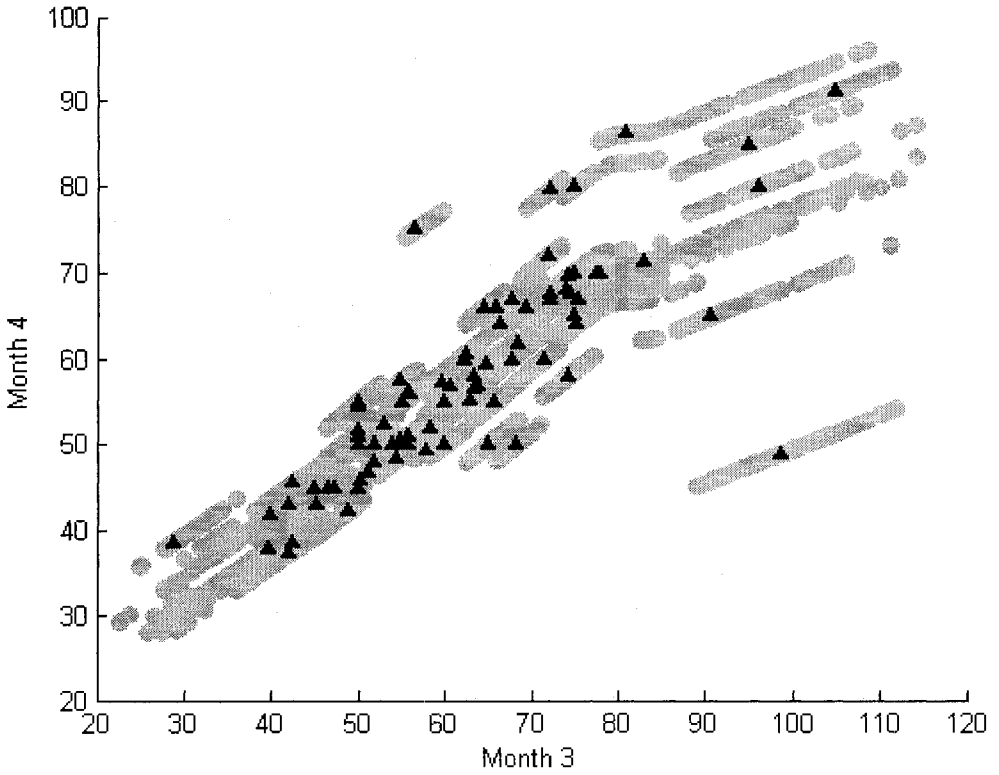


Figure 2.4 Scatter plot of Historical (triangle) and Generated (grey circle) data from local regression with KNNR innovation for Weber River data in m^3/s

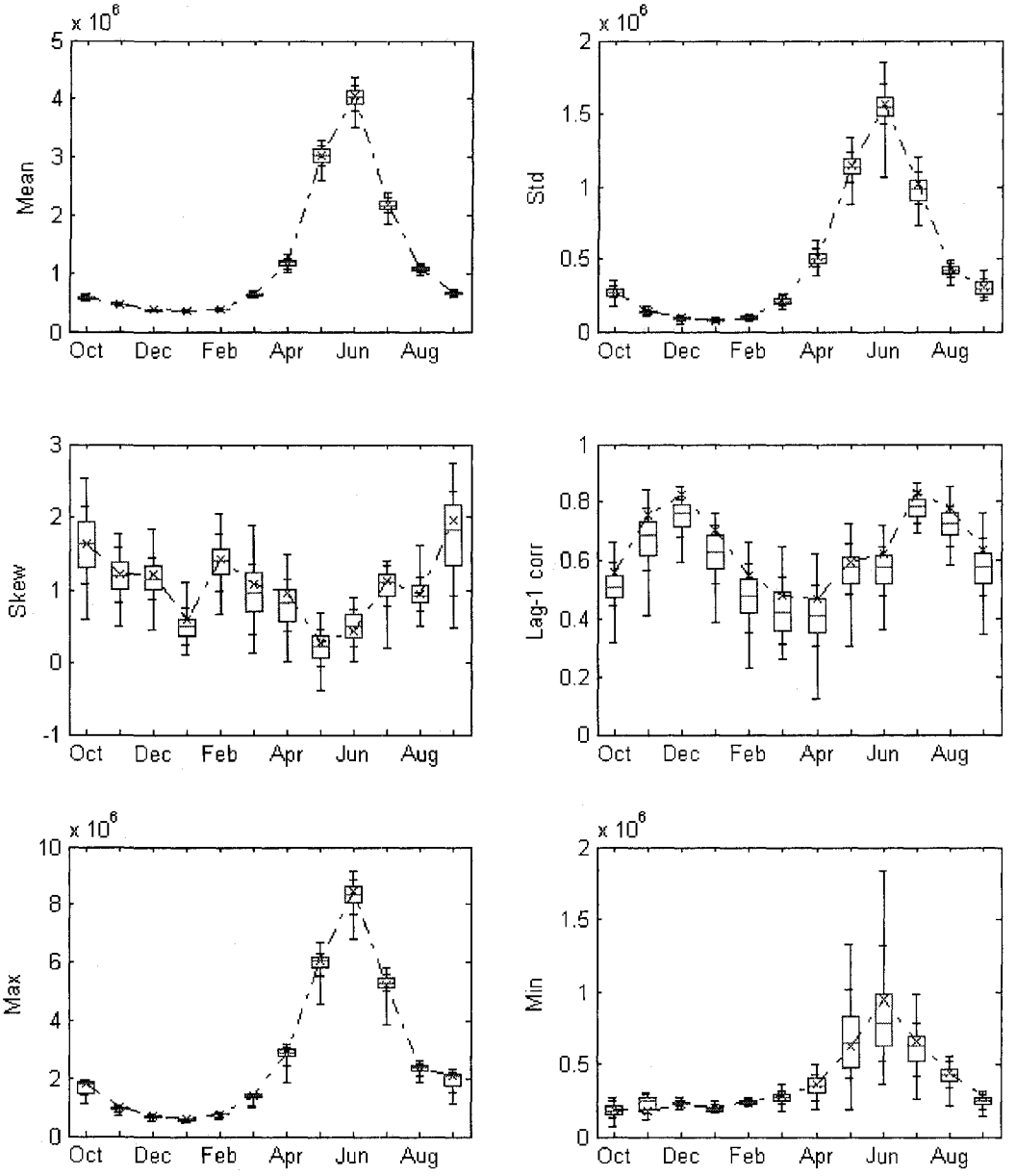


Figure 2.5 Key Statistics of Historical (dot line) and KGKA model simulations (boxplot) of the Colorado River monthly streamflow Unit : Acre-feet

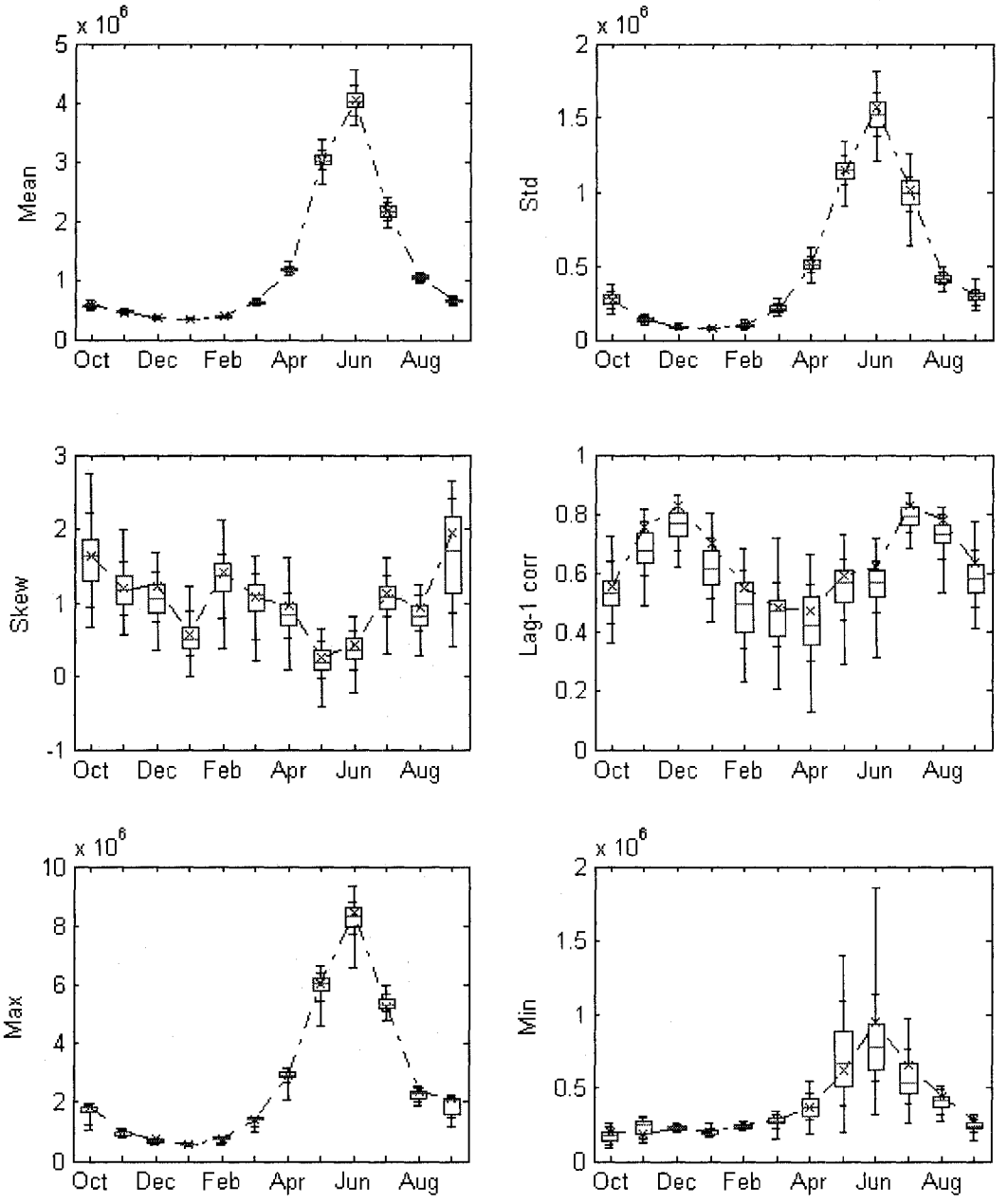


Figure 2.6 Key Statistics of Historical (dot line) and KGKP model simulations (boxplot) of the Colorado River monthly streamflow Unit : Acre-feet

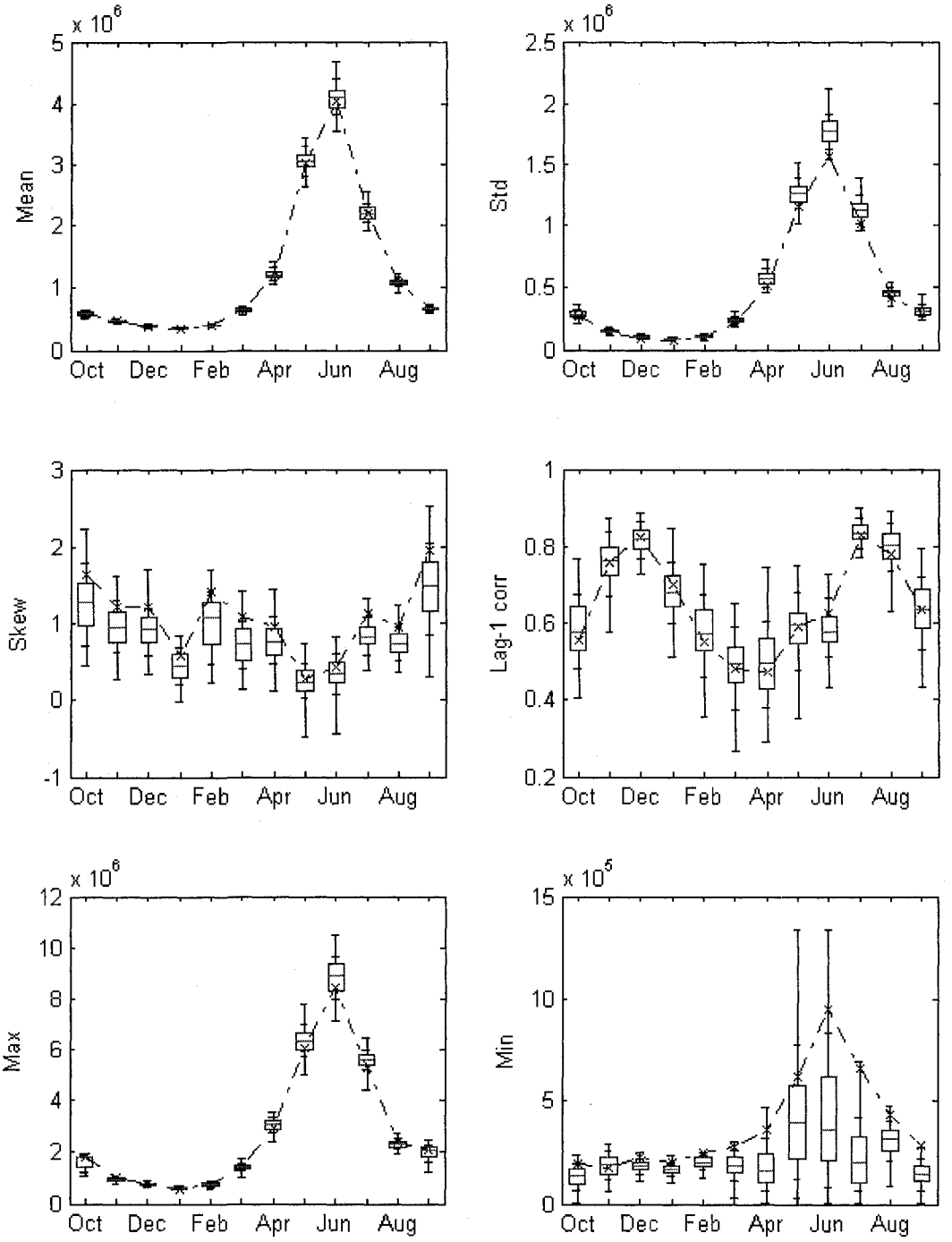


Figure 2.7 Key Statistics of Historical (dot line) and NPL model simulations (boxplot) of the Colorado River monthly streamflow, Unit : Acre-feet

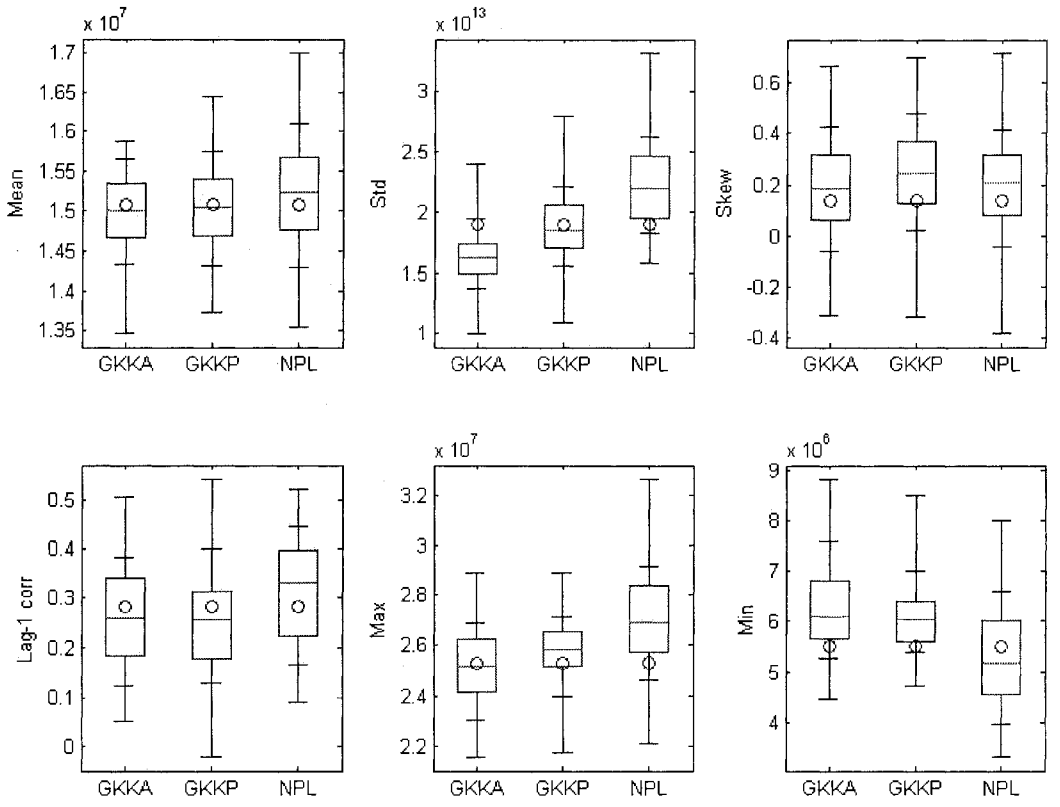


Figure 2.8 Key Statistics of Historical (circle) and KGKA, KGKP, and NPL model simulations (boxplot) of the Colorado River yearly streamflow Unit : Acre-feet

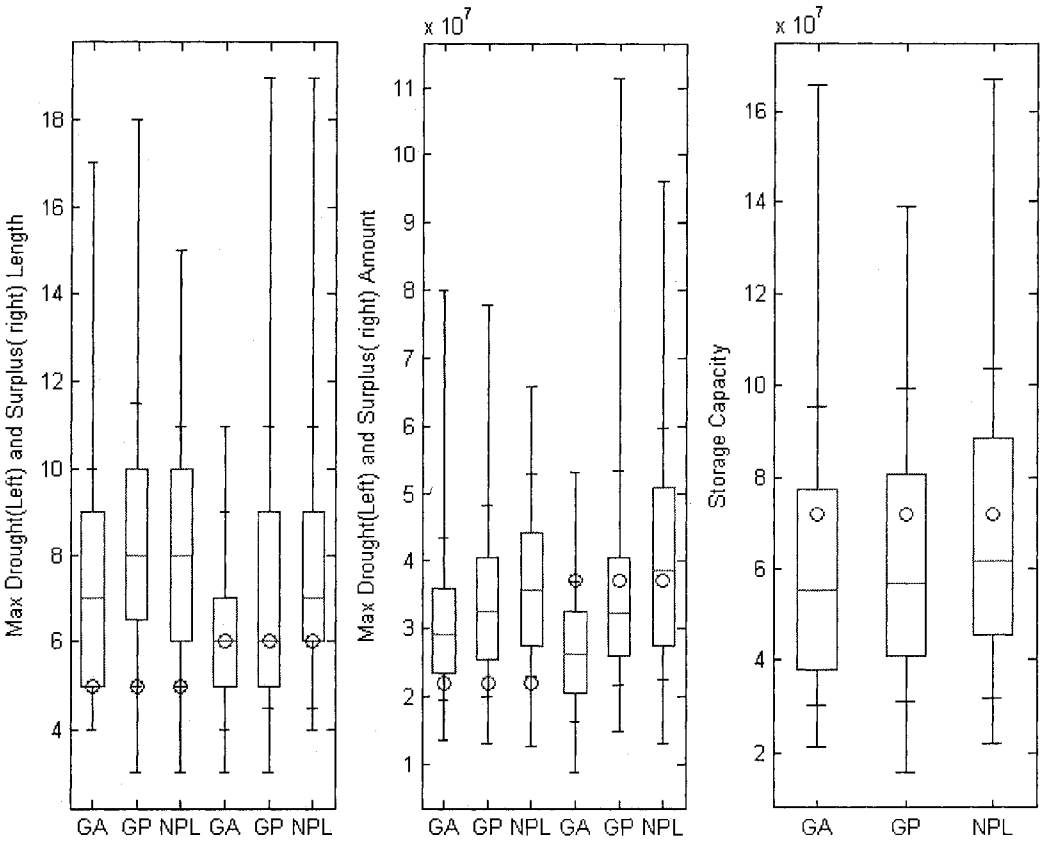


Figure 2.9 Reservoir-related statistics from historical (circle) and generated yearly data from GA (KGKA), GP(KGKP), NPL models (boxplot) for Colorado River at Lees Ferry – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (Unit : Acre-feet)

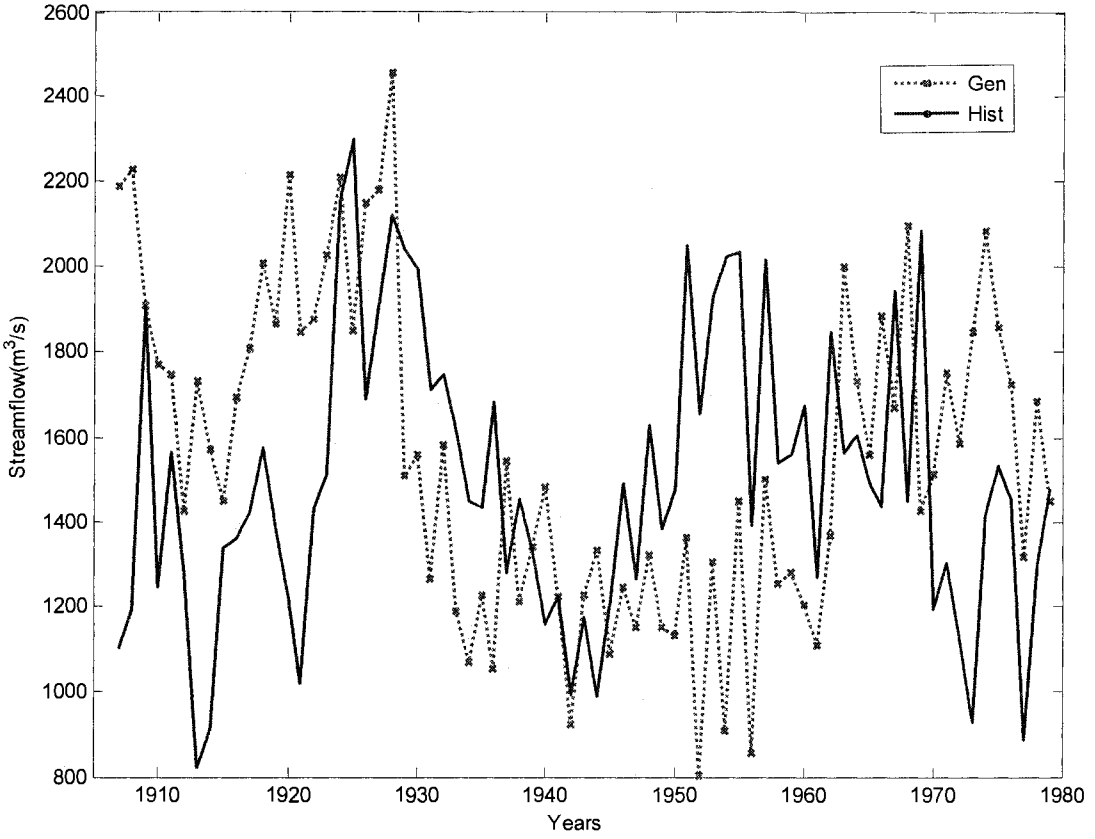


Figure 2.10 Yearly time series of historical data (segment line) and one set (dotted line with 'x') of the summation of generated monthly data (KGKP)

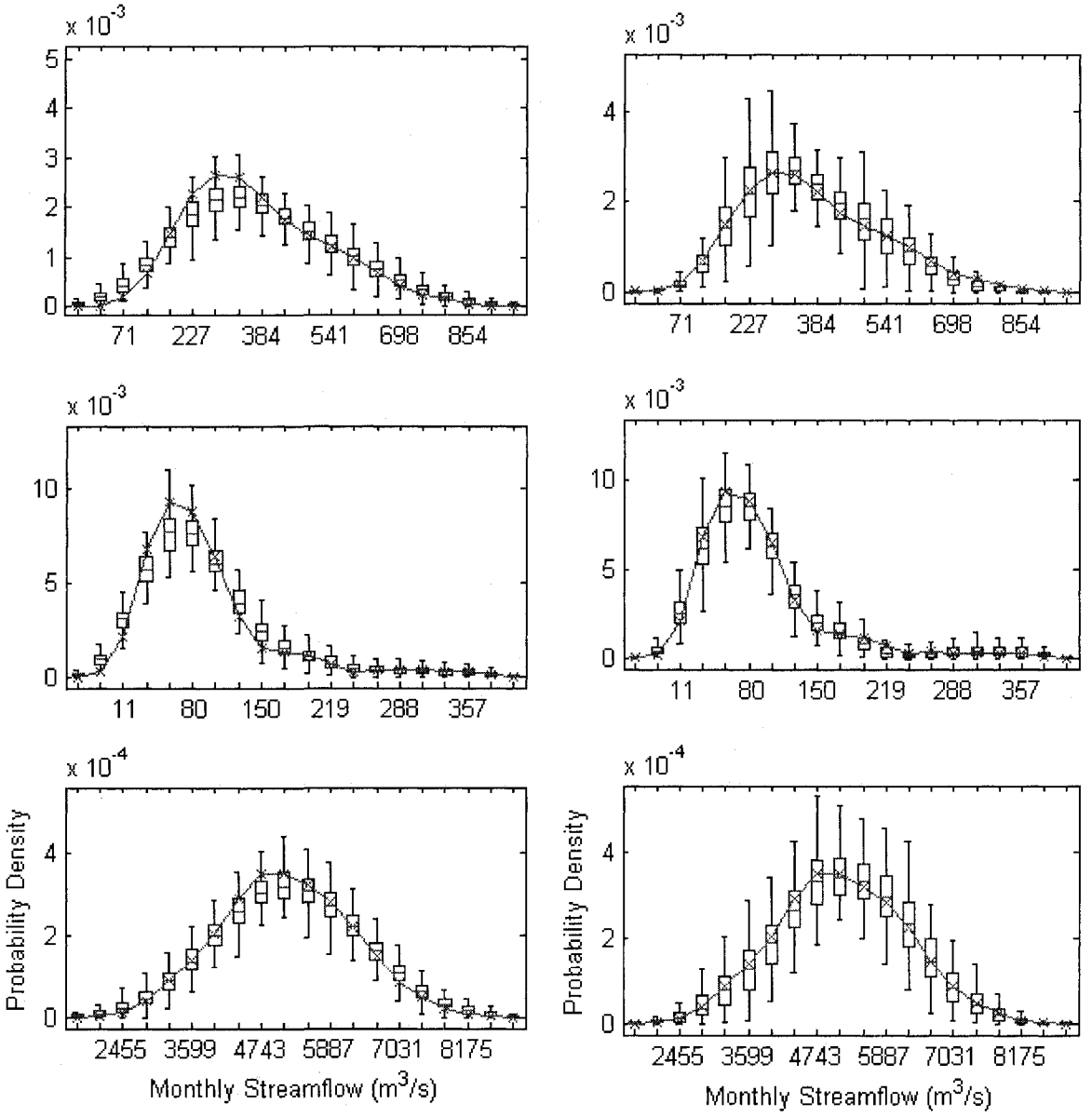


Figure 2.11 Kernel density estimate of historical (segment line) and generated (boxplot) monthly streamflow data for Niger River at Koulikoro from NPL(left) and KGKP(right) model for month 1, 5, and 9 from top to bottom.

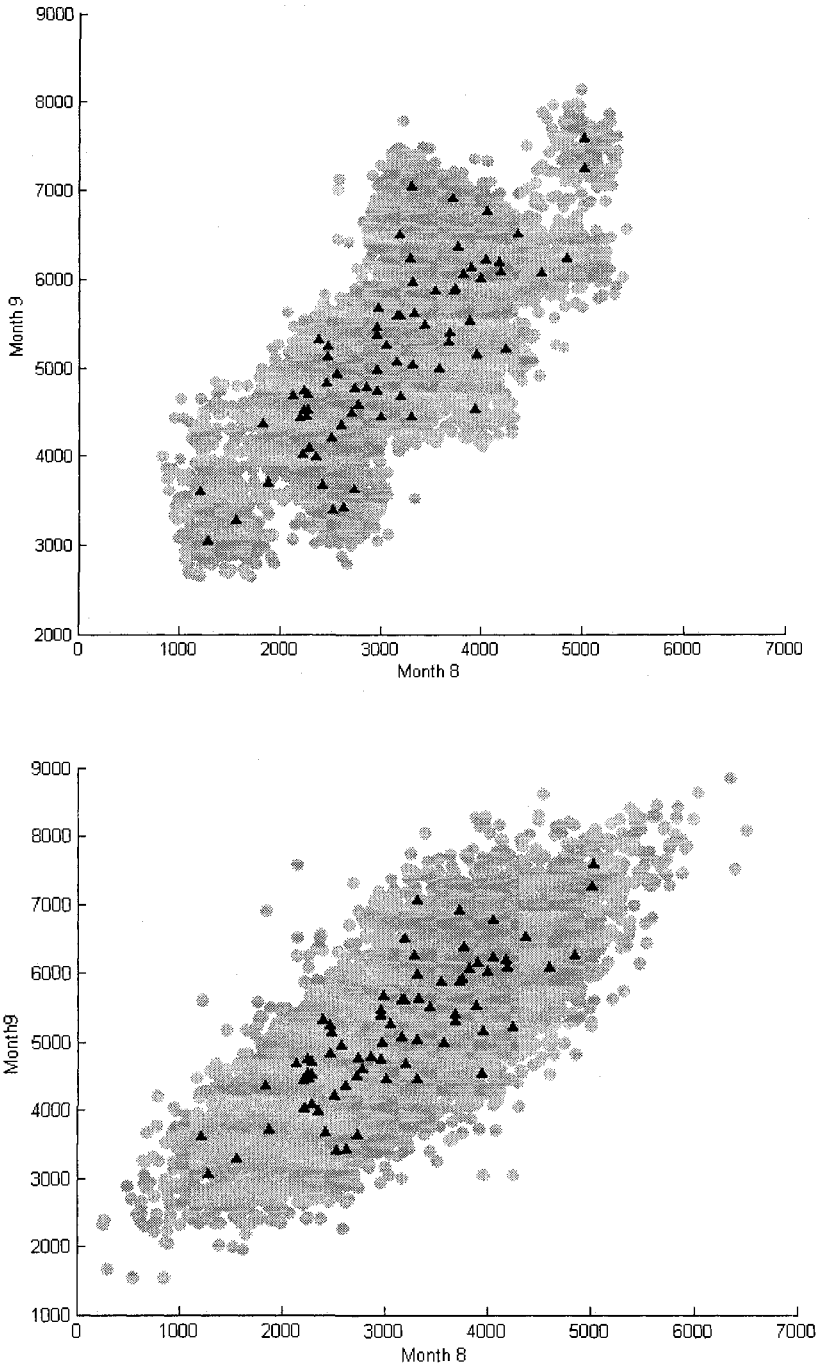


Figure 2.12 Scatter plot of monthly streamflow data with month 8 (x-axis) and month 9 (y-axis) for historical (filled triangle) and 50 sets of the generated data for Niger River at Koulikoro from KGKP(upper) and NPL(bottom) model (grey circle) (unit : m^3/s)

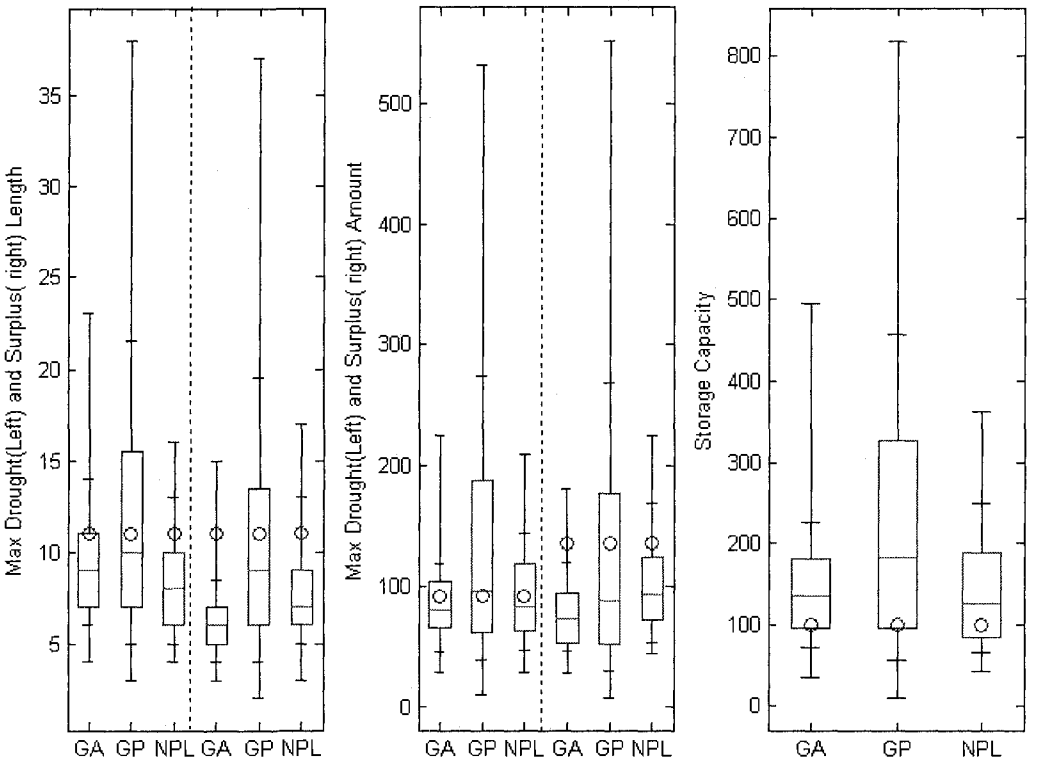


Figure 2.13. Reservoir-related statistics from historical (circle) and generated yearly data from GA (KGKA), GP(KGKP), NPL model (boxplot) for Niger River at Koulikoro – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (unit : 10^9 m^3)

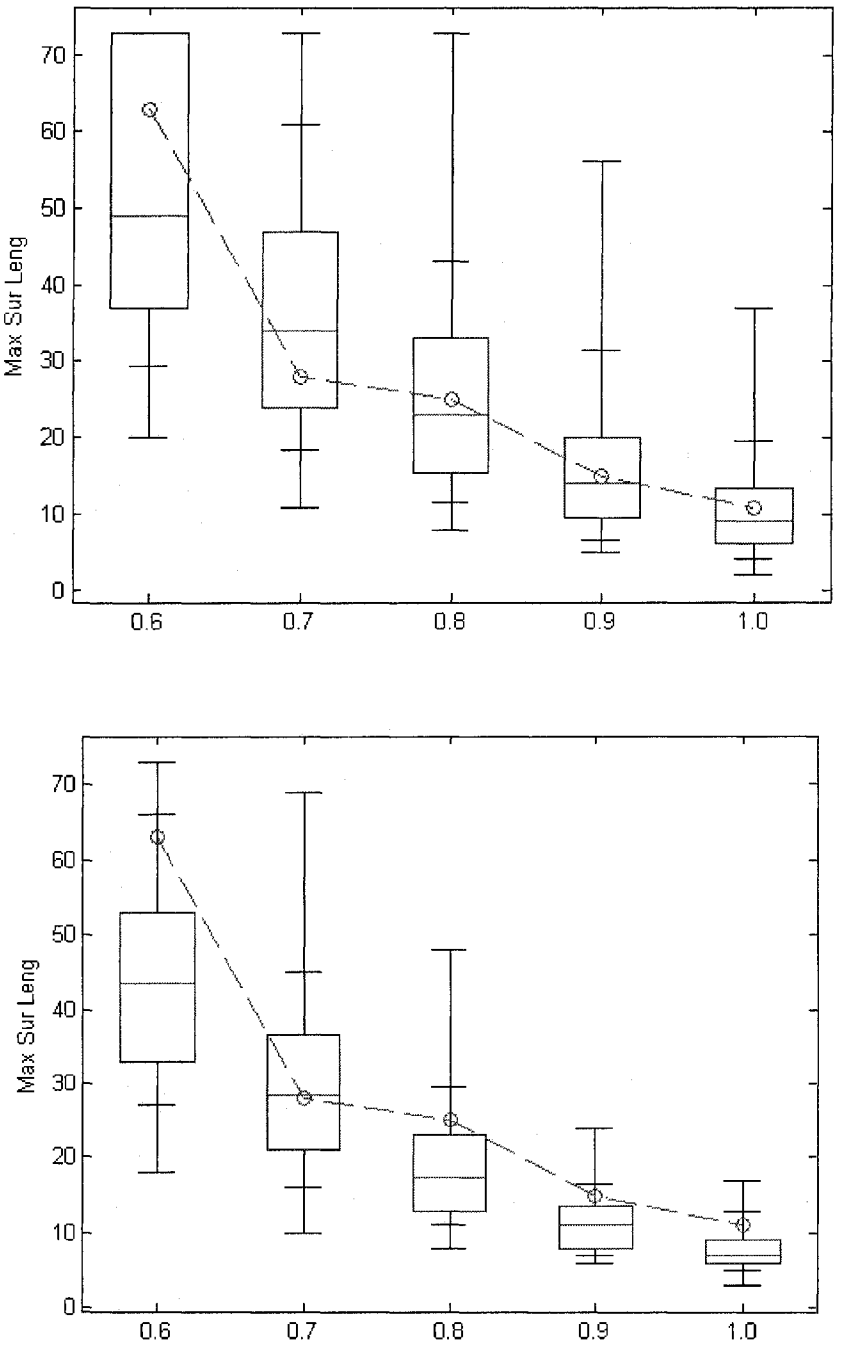


Figure 2.14. Maximum Surplus Length of historical and from historical (circle) and generated yearly data from KGKP (upper) and NPL (bottom) model (boxplot) for Niger River at Koulikoro with different threshold level (0.6~1.0)

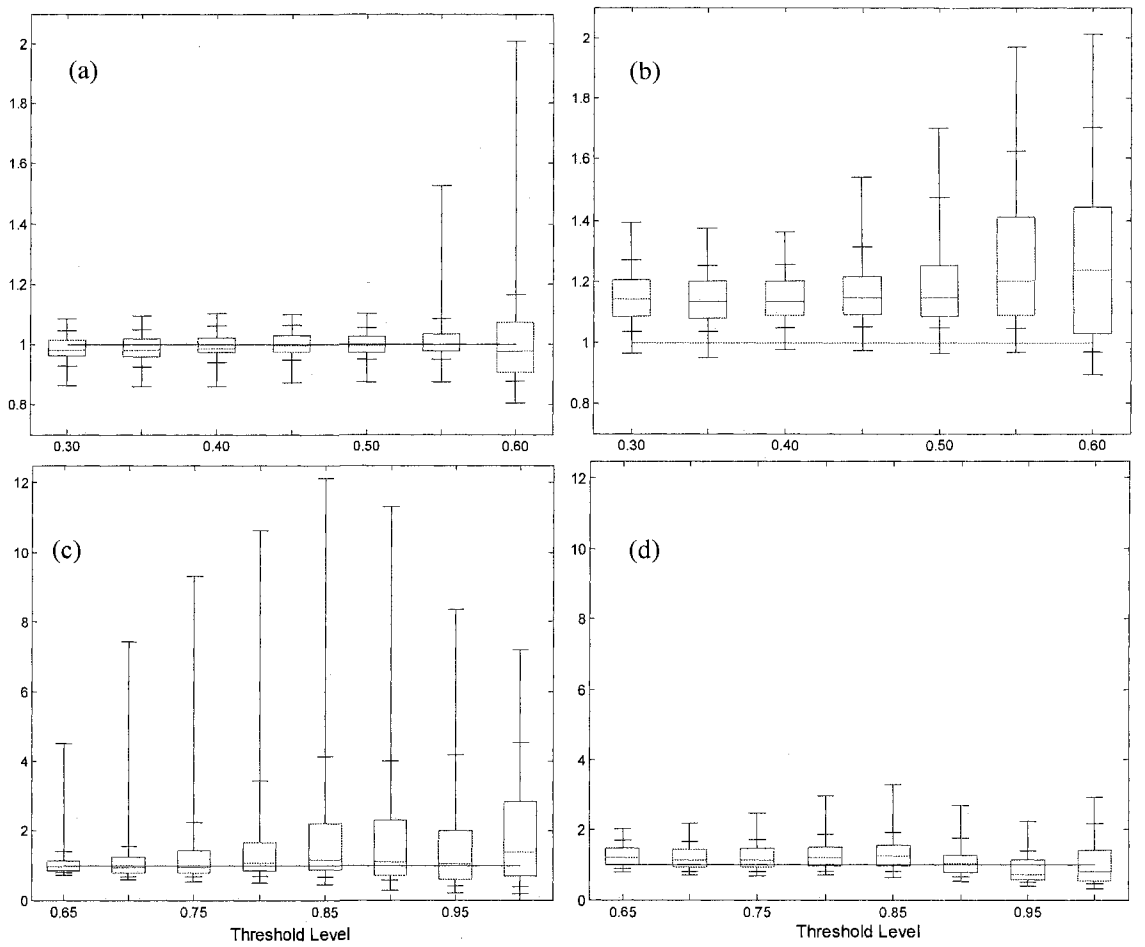


Figure 2.15 The ratios of storage capacity of generated monthly data divided by the one of historical data with different threshold as TL*the overall mean of the historical monthly data for Niger River; the threshold level range 0.3~1.0. The line at 1.0 is presented as a indication mark of perfect match to historical. (a) KGKP with threshold levels 0.3~06; (b) NPL with 0.3~0.6; (c) KGKP with 0.65~1.0; (d) NPL with 0.65~1.0

2.7 References

Buishand, T. A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resources Research, 37*, 2761-2776.

Chebaane, M., Salas JD, and Boes DC (1995), Product Periodic Autoregressive Processes for Modeling Intermittent Monthly Streamflows, *Water Resources Research, 31*, 1513-1518.

Chen, S. X. (1999), Beta kernel estimators for density functions, *Computational Statistics & Data Analysis, 31*, 131-145.

Chen, S. X. (2000), Probability density function estimation using gamma kernels, *Annals of the Institute of Statistical Mathematics, 52*, 471-480.

Fernandez, B., and J. D. Salas (1986), Periodic Gamma-Autoregressive Processes for Operational Hydrology, *Water Resources Research, 22*, 1385-1396.

Fernandez, B., and J. D. Salas (1990), Gamma-Autoregressive Models for Stream-Flow Simulation, *Journal of Hydraulic Engineering-Asce, 116*, 1403-1414.

Fortin, V., Perreault L, and Salas JD (2004), Retrospective analysis and forecasting of streamflows using a shifting level model, *Journal of Hydrology, 296*, 135-163.

Fukunaga, K. (1990), Introduction to Statistical Pattern Recognition, 2 ed., Academic Press.

Goldberg, D. E. (1989), Genetic algorithms in search, optimization, and machine learning, Addison-Wesley Pub. Co.

Hipel K.W., A. I. McLeod. (1996), Time Series modeling of Water Resources and Environmental Systems, Elsevier.

Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resources Research, 32*, 2105-2117.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resources Research, 32*, 679-693.

Langousis, A., and D. Koutsoyiannis (2006), A stochastic methodology for generation of seasonal time series reproducing overyear scaling behaviour, *Journal of Hydrology, 322*, 138-154.

Lee T., J. D. Salas (2008), Periodic Stochastic Model for Simulating Intermittent Monthly Streamflows of the Colorado River System, paper presented at World Environmental & Water Resources Congress 2008, Honolulu, Hawaii.

Lee T, and Salas JD (2006), Record Extension of Monthly Flows for the Colorado River System, USBR Report, in Print

Loucks, D. P., Stedinger J.R., and Haith D.A. (1981), *Water Resources Systems Planning And Analysis*, Prentice-Hall.

Mugdadi, A., and A. Lahrech (2004), The exponential kernel in density estimation, *Far East Journal of Theoretical Statistics*, *14*, 1-14.

Mandelbrot B, and J. R. Wallis (1969), Computer Experiments with Fractional Gaussian Noises .1. Averages and Variances, *Water Resources Research*, *5*, 228-.

McMahon, T. A., et al. (2006), Understanding performance measures of reservoirs, *Journal of Hydrology*, *324*, 359-382.

Prairie, J. R., et al. (2006), Modified K-NN model for stochastic streamflow simulation, *Journal of Hydrologic Engineering*, *11*, 371-378.

Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other weather variables, *Water Resources Research*, *35*, 3089-3101.

Salas, J. D., D.C. Boes (1980), Shifting level modeling of hydrologic series, *Advances in Water Resources*, *3*, 59-63.

Salas, J.D., D.C. Boes, Yevjevich, V, and Pegram, GGS (1979), Hurst Phenomenon as a Pre-Asymptotic Behavior, *Journal of Hydrology*, *44*, 1-15.

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

Sharif, M., and D. H. Burn (2007), Improved K-nearest neighbor weather generating model, *Journal of Hydrologic Engineering*, *12*, 42-51.

Sharma, A., Tarboton DG, and Lall U (1997), Streamflow simulation: A nonparametric approach, *Water Resources Research*, *33*, 291-308.

Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resources Research*, *38*, 5.1-5.10.

Sharma, A., Lall U, and Tarboton DG (1998), Kernel bandwidth selection for a first order nonparametric streamflow simulation model, *Stochastic Hydrology and Hydraulics*, *12*, 33-52.

Silverman, B. W. (1986), Density Estimation for Statistics and Data Analysis : Monographs on Statistics and Applied Probability, Chapman and Hall.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, Springer.

Srinivas, V. V., and K. Srinivasan (2006), Hybrid matched-block bootstrap for stochastic simulation of multiseason streamflows, *Journal of Hydrology*, *329*, 1-15.

Srinivas, V. V., and K. Srinivasan (2005), Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows, *Journal of Hydrology*, *302*, 307-330.

Stedinger, J. R., et al. (1985), A Condensed Disaggregation Model for Incorporating Parameter Uncertainty into Monthly Reservoir Simulations, *Water Resources Research*, *21*, 665-675.

Sveinsson, O. G. B., Salas, J. D., Boes, D. C., and Pielke, R. A. (2003), Modeling the dynamics of long-term variability of hydroclimatic processes, *Journal of Hydrometeorology*, *4*, 489-505.

Tarboton, D. G., et al. (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resources Research*, *34*, 107-119.

Valencia, D., and J. C. Schaake (1973), Disaggregation Processes in Stochastic Hydrology, *Water Resources Research*, *9*, 580-585.

Vogel, R. M., and A. L. Shallcross (1996), The moving blocks bootstrap versus parametric time series models, *Water Resources Research*, *32*, 1875-1882.

Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. (2003), A technique for generating regional climate scenarios using a nearest-neighbor algorithm, *Water Resources Research*, *39*, -.

Yakowitz, S., and M. Karlsson (1987), Nearest-Neighbor Methods with application to rainfall-runoff prediction, 149-160 pp.

Young, K. C. (1994), A Multivariate Chain Model for Simulating Climatic Parameters from Daily Data, *Journal of Applied Meteorology*, *33*, 661-671.

Appendix 2-A : Variance of the Gamma kernel density estimate

In Chen (2000), the variance of the density estimator suggested by Chen(2000) is derived as

$$\begin{aligned} Var\{\hat{f}(x)\} &= \text{var}\{\frac{1}{N}\sum_{i=1}^{N} K_{\alpha,\beta}(X_i)\} = \frac{1}{N^2}\sum_{i=1}^{N}\text{var}\{K_{\alpha,\beta}(X_i)\} \\ &= N^{-1}\,\text{var}(K_{\alpha,\beta}(t)) = N^{-1}\left[E\{K_{\alpha,\beta}(t)\}^2 - (E\{K_{\alpha,\beta}(t)\})^2\right] \\ &= N^{-1}E\{K_{\alpha,\beta}(t)\}^2 + O(N^{-1}) \end{aligned} \tag{2-A.1}$$

In this equation, $E\{K_{\alpha,\beta}(x_i)\}$ is shown in Eq.(2-13) and $[E\{K_{\alpha,\beta}(X_i)\}]^2 \approx f(x)^2$. $N^{-1}f(x)^2$ is negligible by comparing $N^{-1}E\{K_{\alpha,\beta}(t)\}^2$ of Eq.(2-A.1) since $E\{K_{\alpha,\beta}(t)\}^2$ is a function of $f(x)$ and $0 \le f(x) \le 1$. And

$$K_{\alpha,\beta}^2(t) = \frac{t^{2\alpha-2}e^{-2t/\beta}}{\beta^{2\alpha}\Gamma^2(\alpha)} = \frac{t^{(2\alpha-1)-1}e^{-t/(\beta/2)}}{(\beta/2)^{2\alpha-1}\Gamma(2\alpha-1)}\frac{(\beta/2)^{2\alpha-1}\Gamma(2\alpha-1)}{\beta^{2\alpha}\Gamma^2(\alpha)} \tag{2-A.2}$$

Here, $\frac{t^{(2\alpha-1)-1}e^{-t/(\beta/2)}}{(\beta/2)^{2\alpha-1}\Gamma(2\alpha-1)}$ can be interpreted as the gamma distribution function with the shape parameter $2\alpha-1$ and scale parameter $\beta/2$. Therefore, let present this as $g(t)$ where g is the gamma distribution function denoted as $Gama[2\alpha-1, \beta/2]$. From this formation, the gamma function term envisage as:

$$K_{\alpha,\beta}^2(t) = g(t)\frac{(\beta/2)^{2\alpha-1}\Gamma(2\alpha-1)}{\beta^{2\alpha}\Gamma^2(\alpha)} = g(t)\frac{\Gamma(2\alpha-1)}{\Gamma^2(\alpha)}\beta^{-1}2^{-2\alpha+1} \tag{2-A.3}$$

Let, $R(z) = \sqrt{2\pi}e^{-z}z^{z+1/2}\Big/\Gamma(z+1)$ for $z \ge 0$ (Chen, 1998) and

$$\frac{\Gamma(2\alpha-1)}{\Gamma^2(\alpha)} = \frac{R^2(\alpha-1)}{R(2\alpha-2)}\frac{1}{\sqrt{2\pi}}(\alpha-1)^{-1/2}2^{2\alpha-3/2} \tag{2-A.4}$$

Brown and Chen (1998) proved that $R(z)$ is a monotonic increasing function which converges to 1 as $z \to \infty$ and $R(z) < 1$ for any $z > 0$. Thus, $R^2(\alpha-1)/R(2\alpha-2) < 1$ and therefore,

$$E\{K^2_{\alpha,\beta}(t)\} = \frac{R^2(\alpha-1)}{R(2\alpha-2)}\frac{1}{\sqrt{2\pi}}(\alpha-1)^{-1/2}2^{2\alpha-3/2}\beta^{-1}2^{-2\alpha+1}E\{g(t)\}$$

$$= \frac{R^2(\alpha-1)}{R(2\alpha-2)}\frac{1}{2\sqrt{\pi}}(\alpha-1)^{-1/2}\beta^{-1}E\{g(t)\} \tag{2-A.5}$$

Substituting $\alpha = x^2/h^2$ and $\beta = h^2/x$ instead of $\alpha = x/h+1$ and $\beta = h$ by Chen (2000)

$$E\{K_{X^2/h^2,h^2/X}(t)\}^2 = \frac{R^2(x^2/h^2-1)}{R(2x^2/h^2-2)}\frac{1}{2\sqrt{\pi}}(x^2/h^2-1)^{-1/2}(h^2/x)^{-1}E\{g(t)\} \tag{2-A.6}$$

As of Eq.(2-13) with the variance $h^2\left(1/2 - h^2/4x^2\right)$ with *g(t)* distribution, $E\{g(t)\}$ is represented as:

$$E\{g(t)\} = f(x-h^2/2x) + \frac{1}{2}f''(x)\left(\frac{1}{2}h^2 - \frac{h^4}{4x^2}\right) \tag{2-A.7}$$

From the Taylor expansion,

$$f(x-h^2/2x) \approx f(x) - f'(x)\frac{h^2}{2x} \tag{2-A.8}$$

$$\begin{aligned}\operatorname{var}\{\hat{f}(x)\} &= N^{-1}E\{K_{\alpha,\beta}(t)\}^2 + O(N^{-1}) \\ &\approx N^{-1}\frac{R^2(x^2/h^2-1)}{R(2x^2/h^2-2)}\frac{1}{2\sqrt{\pi}}(x^2/h^2-1)^{-1/2}(h^2/x)^{-1}f(x)\end{aligned} \tag{2-A.9}$$

$$\operatorname{var}\{\hat{f}(x)\} \approx \begin{cases} \dfrac{1}{2Nh\sqrt{\pi}}f(x) & \text{if } x/h \to \infty \\ \dfrac{\Gamma(2\kappa^2-1)}{N\Gamma^2(\kappa^2)}(h/\kappa^2)^{-1}2^{-2\kappa^2+1}f(x) & \text{if } x/h \to \kappa \end{cases} \tag{2-A.10}$$

The first term in Eq.(2-A.9) is derived from the $\dfrac{R^2(x^2/h^2-1)}{R(2x^2/h^2-2)} \to 1$ as $x/h \to \infty$ from the theorem in Brown and Chen(1999), and the second term with the case of $x/h \to \kappa$ and replace the term $R^2(x^2/h^2-1)$ and $R(2x^2/h^2-2)$ as follows.

$$R^2(x^2/h^2-1) = \left[\frac{\sqrt{2\pi}e^{-(x^2/h^2-1)}\left(x^2/h^2-1\right)^{x^2/h^2-1+1/2}}{\Gamma(x^2/h^2)}\right]^2$$

$$R(2x^2/h^2-2) = \frac{\sqrt{2\pi}e^{-(2x^2/h^2-2)}\left(2x^2/h^2-2\right)^{2x^2/h^2-2+1/2}}{\Gamma(2x^2/h^2-2+1)}$$

Appendix 2-B: Detailed Figures

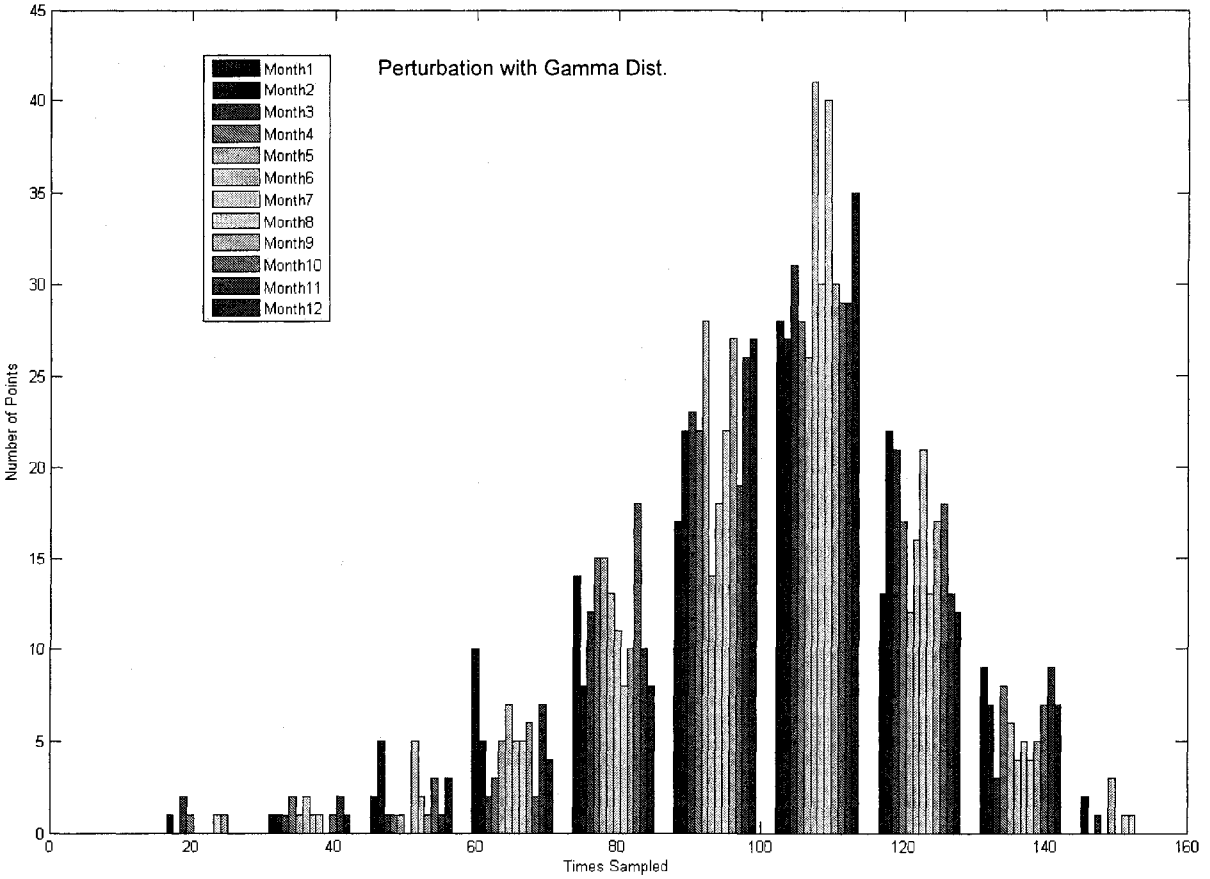


Figure 2-B.1 Histogram of the number of times individual points were selected from KNNR simulation for each month; 100 sets are simulated for the length 98 yrs as historical for Colorado river Site 20

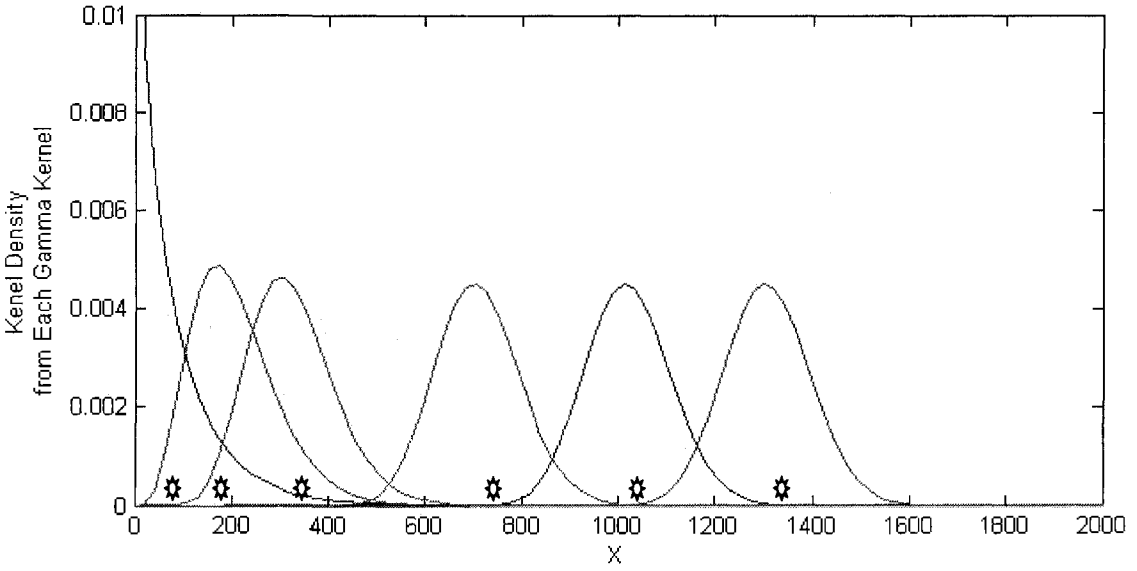


Figure 2-B.2 Example of Gamma Kernel Density Estimate with different location of

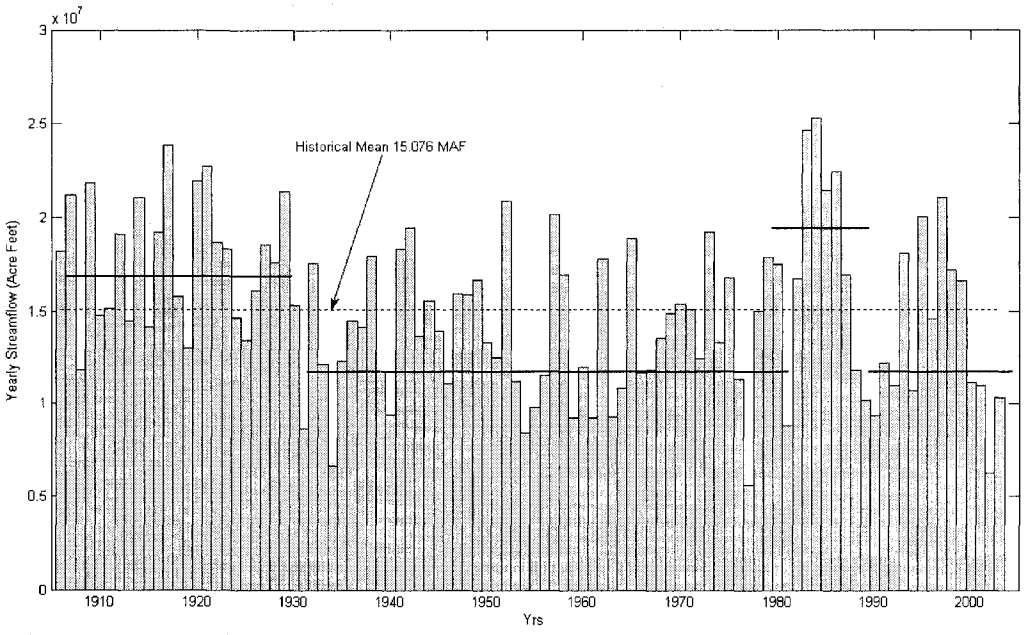


Figure 2-B.3 Time Series of Yearly streamflow for Colorado River at Lees Ferry with the annual mean (15.076 million acre feet)

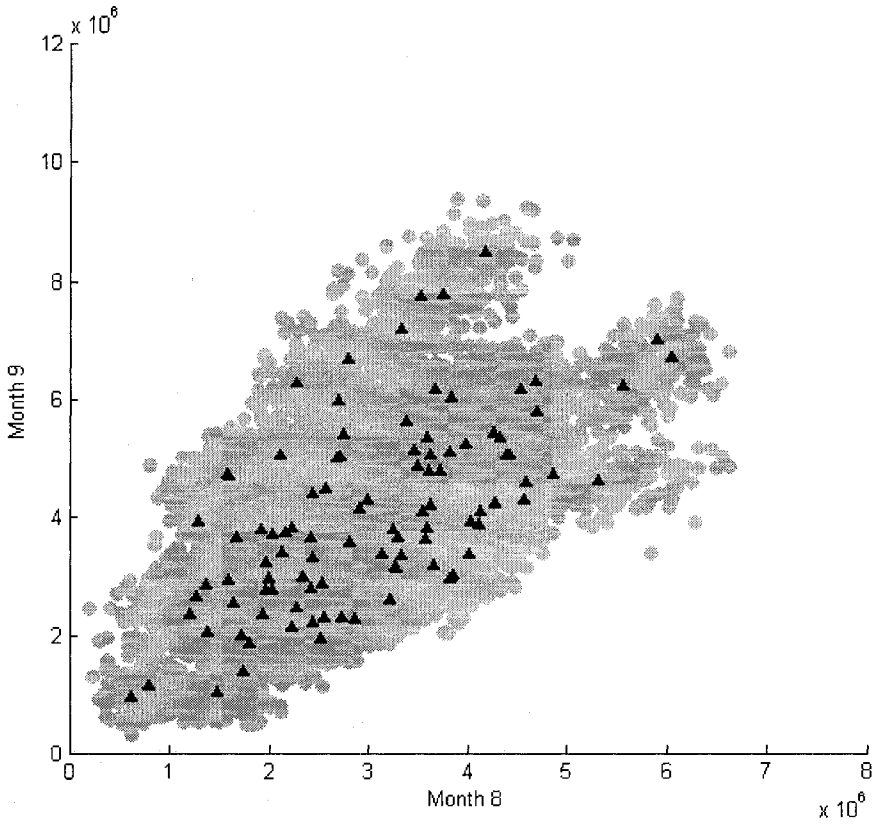


Figure 2-B.4 Scatter plot of monthly streamflow data with month 8 (x-axis) and month 9 (y-axis) for historical (filled triangle) and 50 sets of the generated data from KGKA model (grey circle) for Colorado River at Lees Ferry Unit : Acre-feet

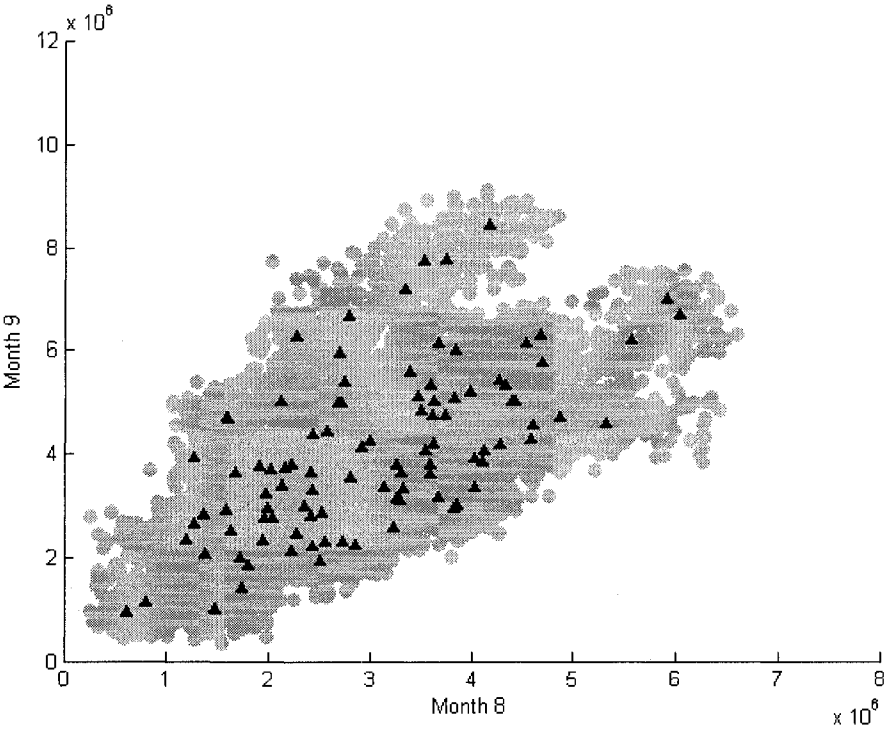


Figure 2-B.5 Scatter plot of monthly streamflow data with month 8 (x-axis) and month 9 (y-axis) for historical (filled triangle) and 50 sets of the generated data for Colorado River at Lees Ferry from KGKP model (grey circle) Unit : Acre-feet

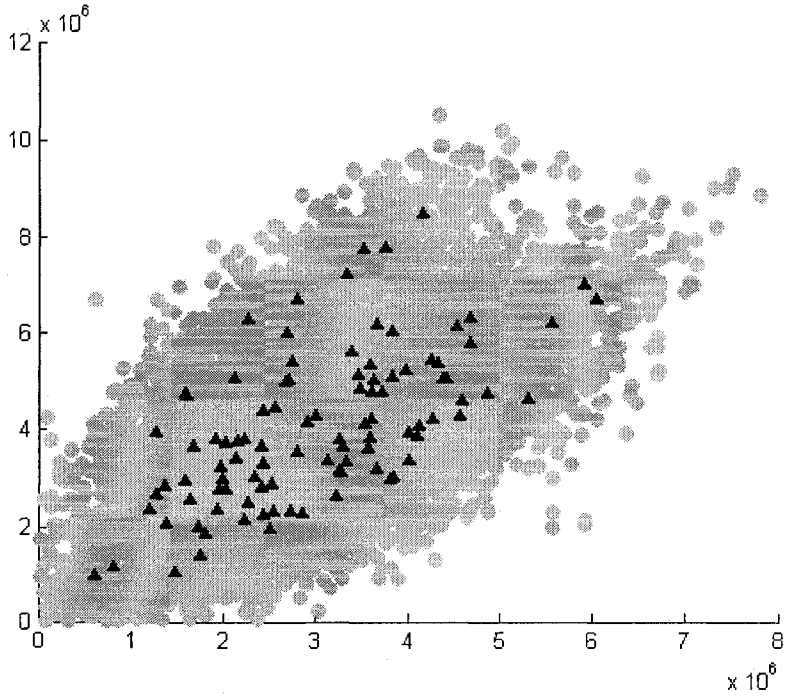


Figure 2-B.6 Scatter plot of monthly streamflow data with month 8 (x-axis) and month 9 (y-axis) for historical (filled triangle) and 50 sets of the generated data for Colorado River at Lees Ferry from NPL model (grey circle) Unit : Acre-feet

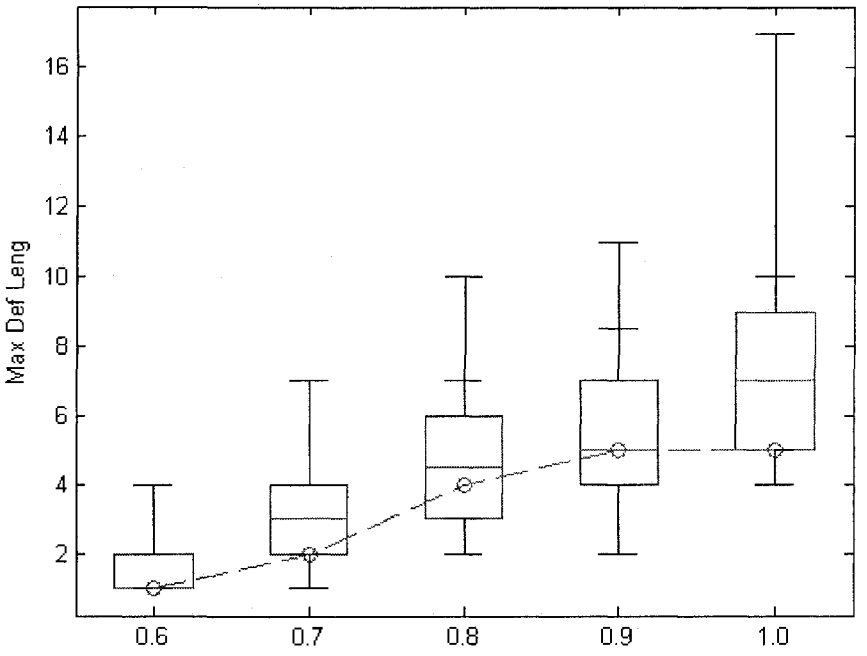


Figure 2-B.7 Maximum Deficit Length of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from KGKA model (boxplot) with different threshold level

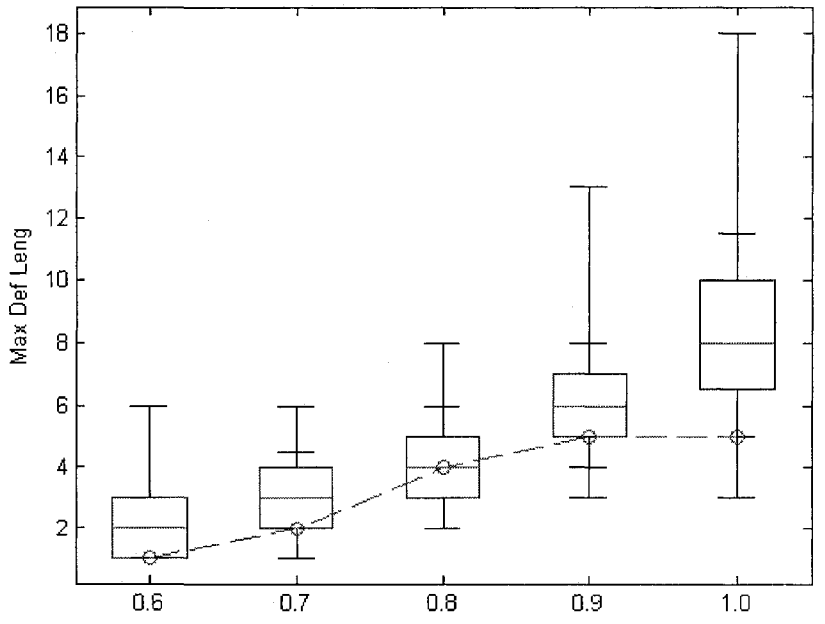


Figure 2-B.8 Maximum Deficit Length of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Colorado River at Lees Ferry with different threshold level

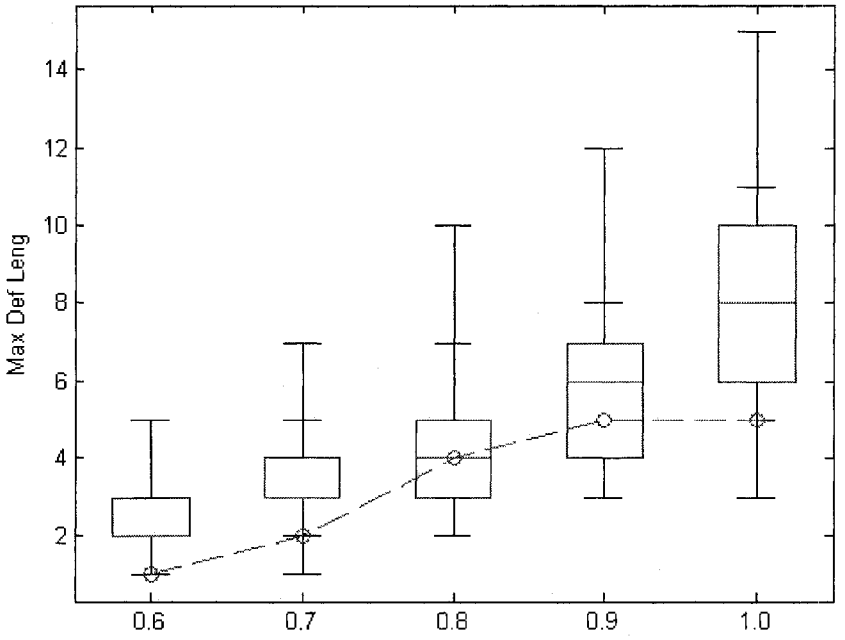


Figure 2-B.9 Maximum Deficit Length of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from NPL model (boxplot) with different threshold level

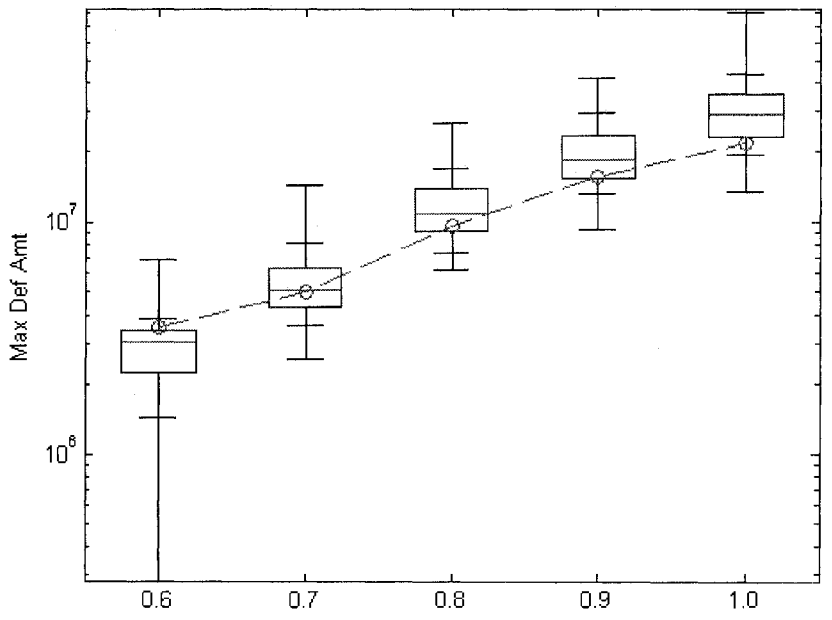


Figure 2-B.10 Maximum Deficit Amount (AF) of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from KGKA model (boxplot) with different threshold level

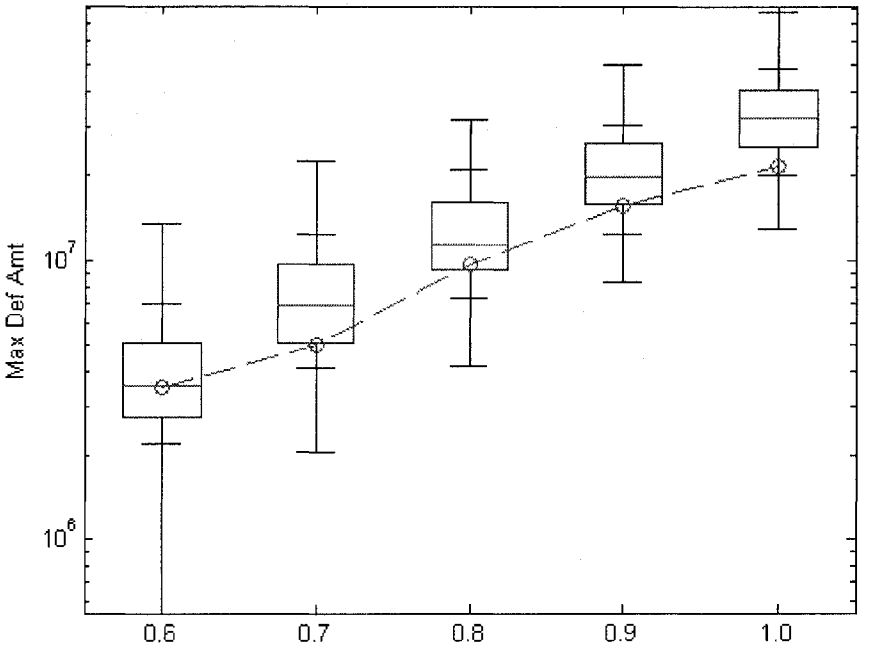


Figure 2-B.11 Maximum Deficit Amount (AF) of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Colorado River at Lees Ferry with different threshold level

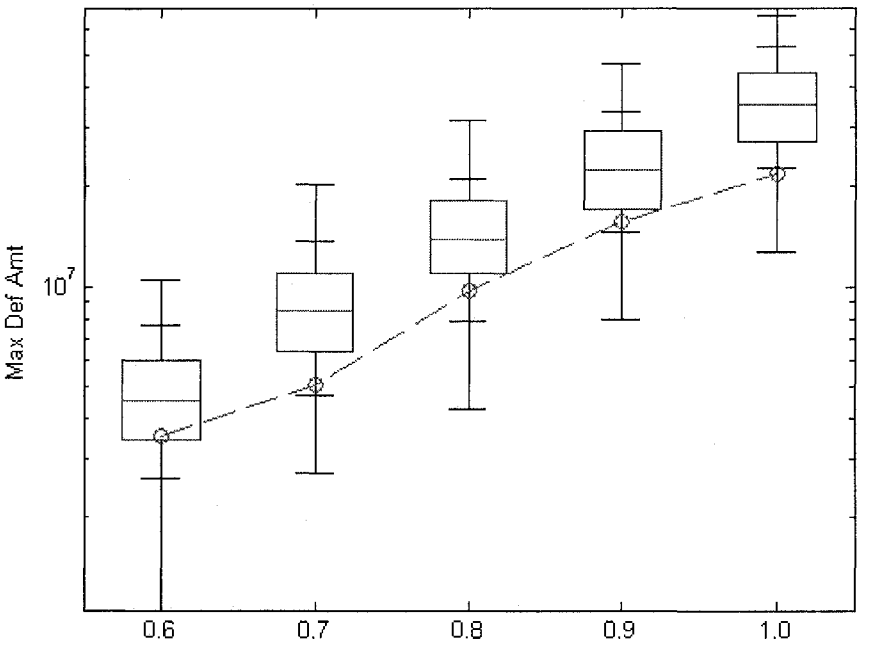


Figure 2-B.12 Maximum Deficit Amount (AF) of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from NPL model (boxplot) with different threshold level

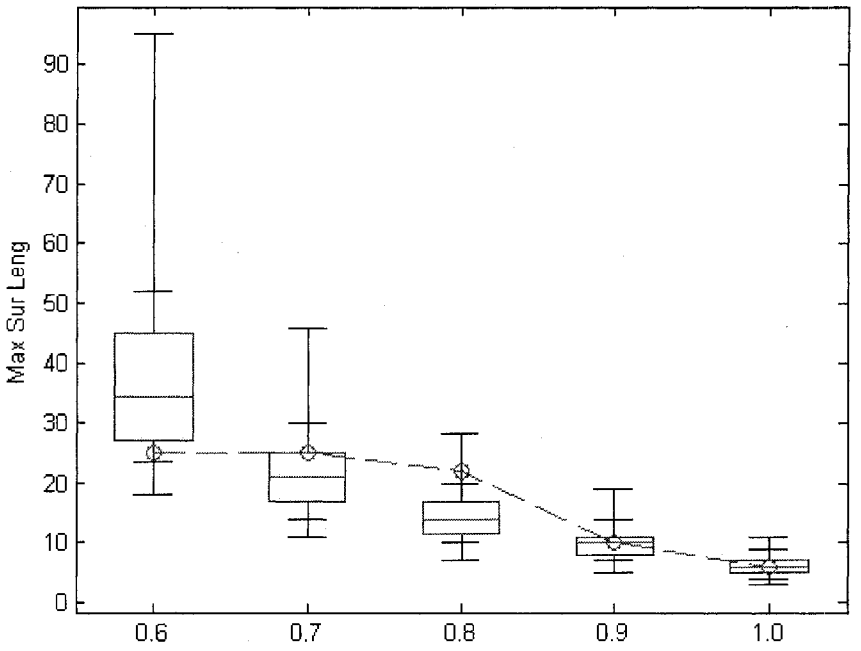


Figure 2-B.13 Maximum Surplus Length of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from KGKA model (boxplot) with different threshold level

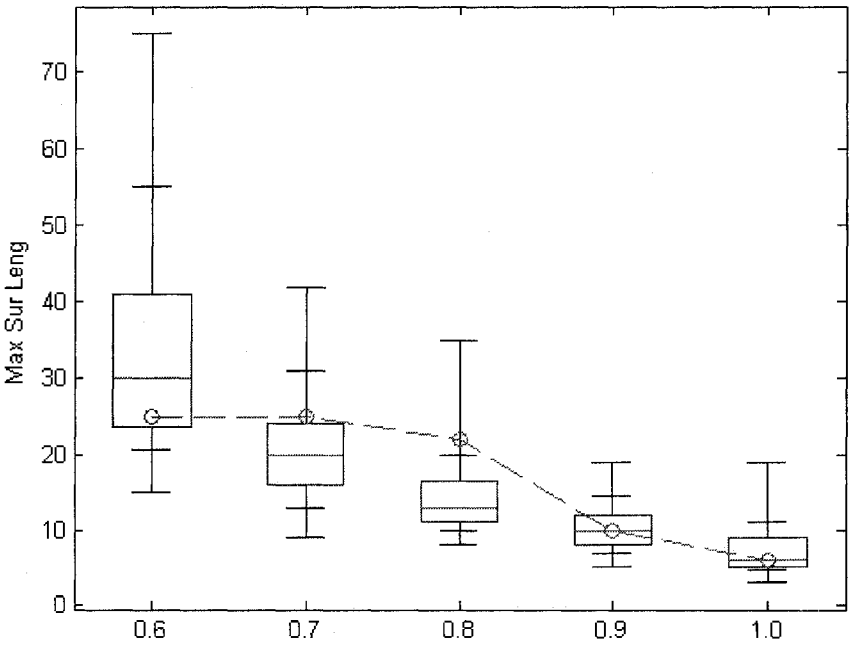


Figure 2-B.14 Maximum Surplus Length of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Colorado River at Lees Ferry with different threshold level

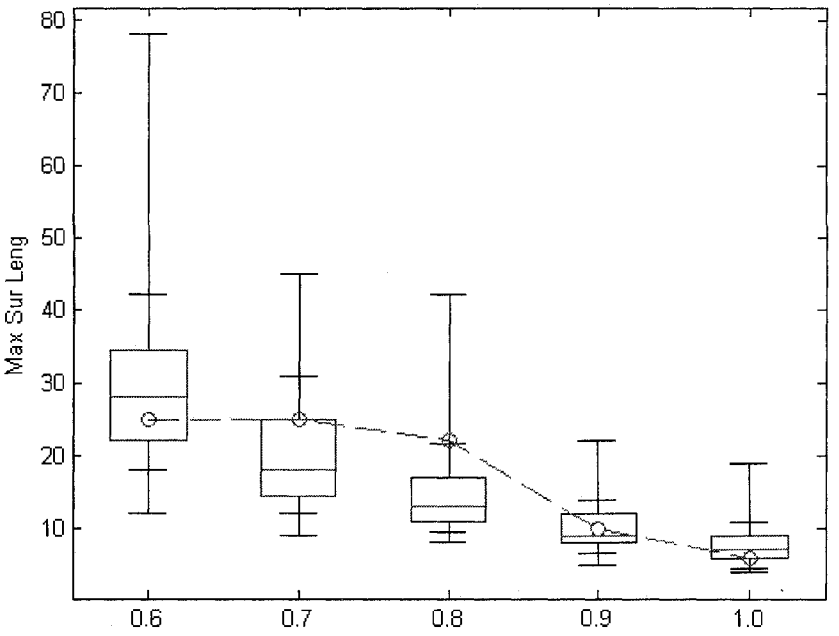


Figure 2-B.15 Maximum Surplus Length of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from NPL model (boxplot) with different threshold level

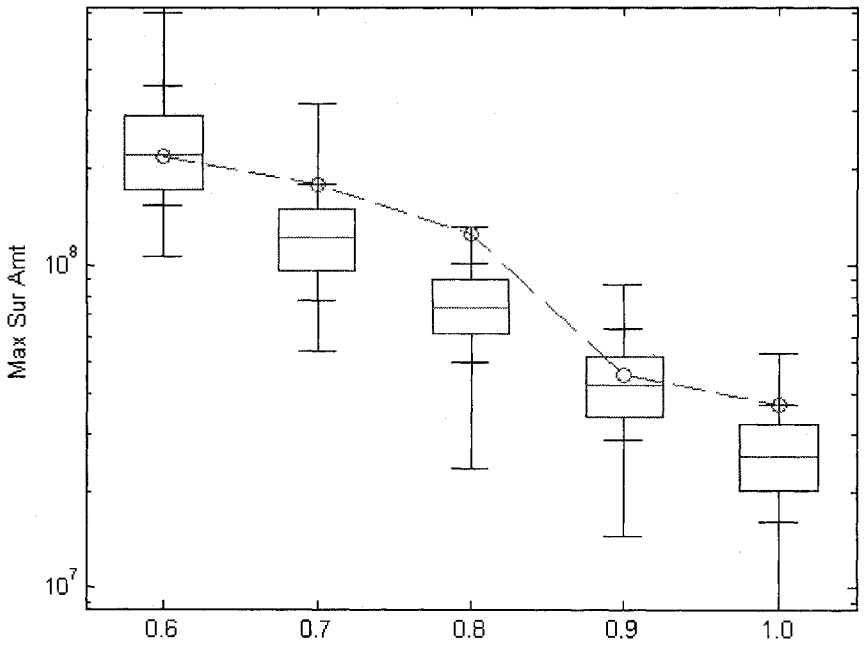


Figure 2-B.16 Maximum Surplus Amount (AF) of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from KGKA model (boxplot) with different threshold level

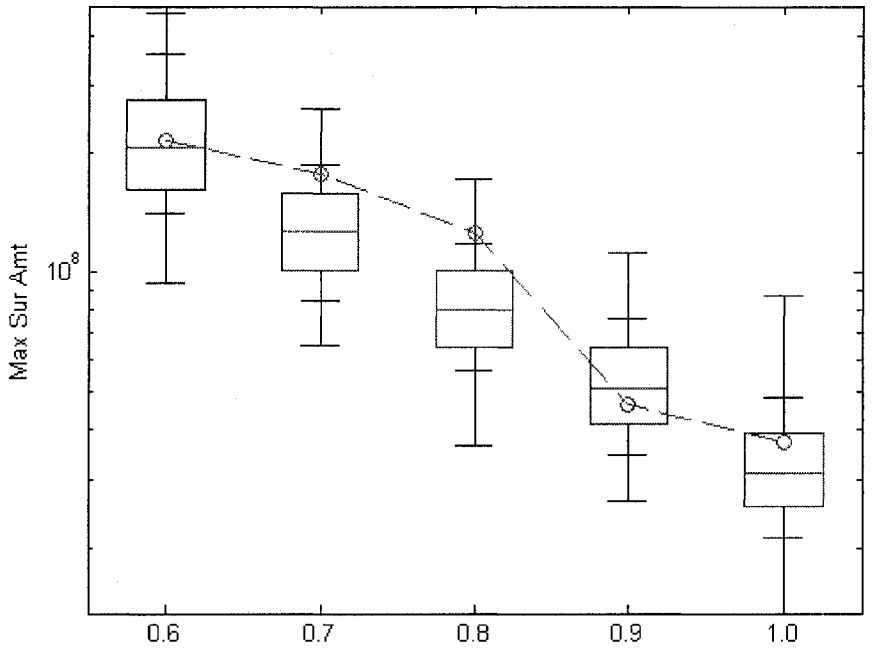


Figure 2-B.17 Maximum Surplus Amount (AF) of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Colorado River at Lees Ferry with different threshold level

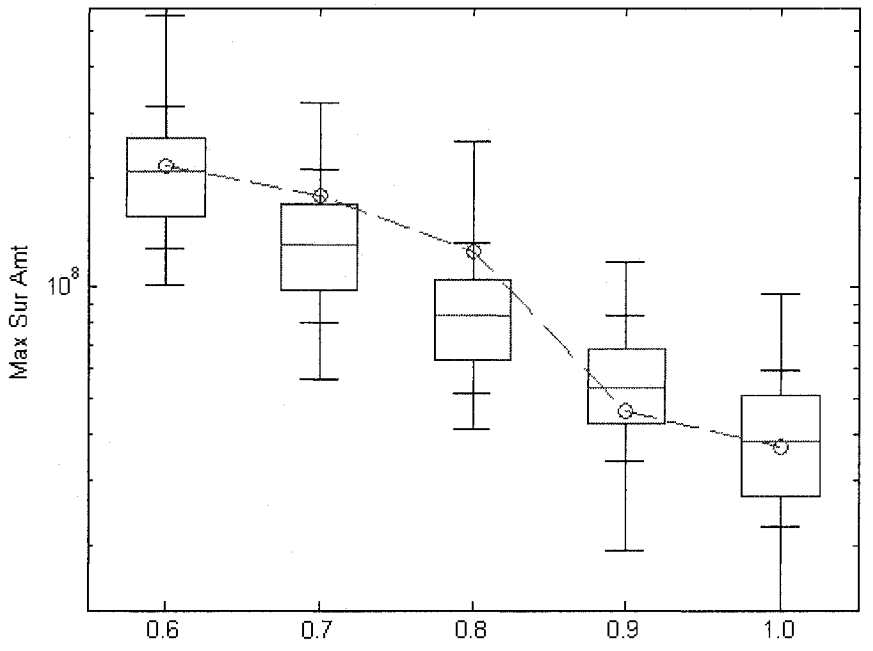


Figure 2-B.18 Maximum Surplus Amount (AF) of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from NPL model (boxplot) with different threshold level

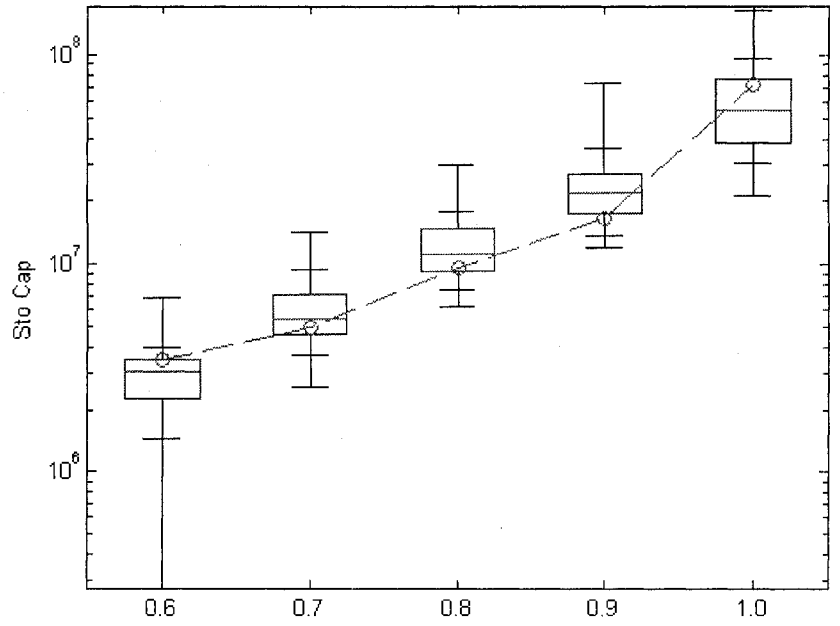


Figure 2-B.19 Storage Capacity (AF) of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from KGKA model (boxplot) with different threshold level

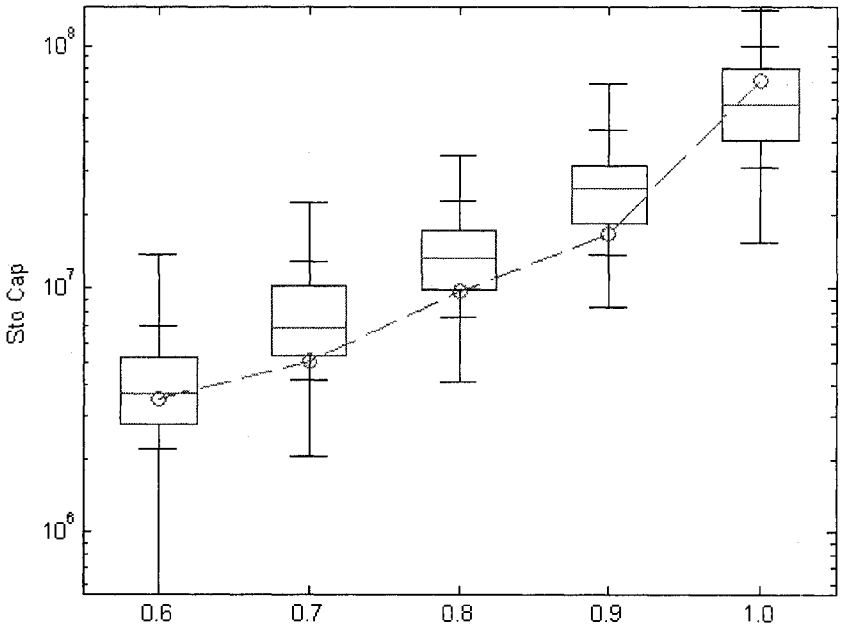


Figure 2-B.20 Storage Capacity (AF) of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Colorado River at Lees Ferry with different threshold level

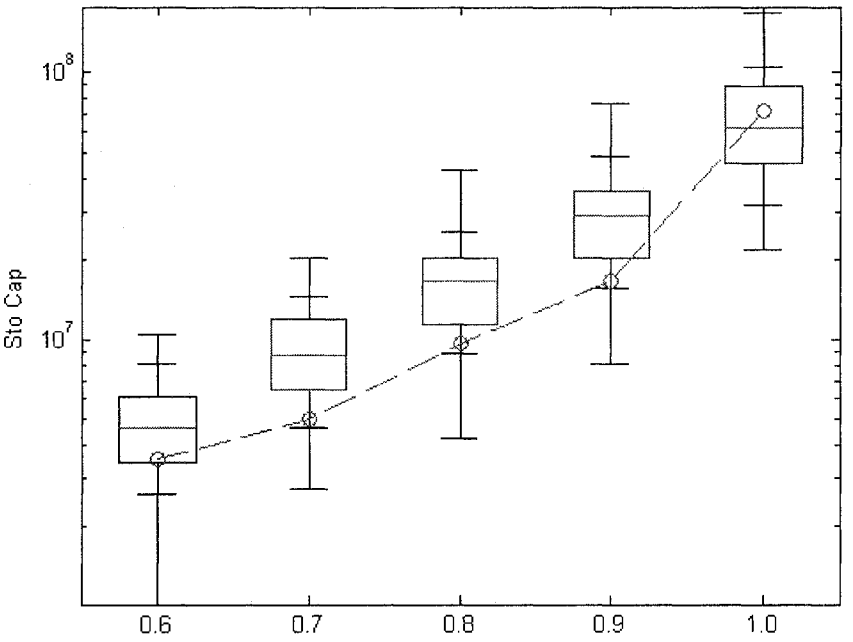


Figure 2-B.21 Storage Capacity (AF) of historical and from historical (circle) and generated yearly data for Colorado River at Lees Ferry from NPL model (boxplot) with different threshold level

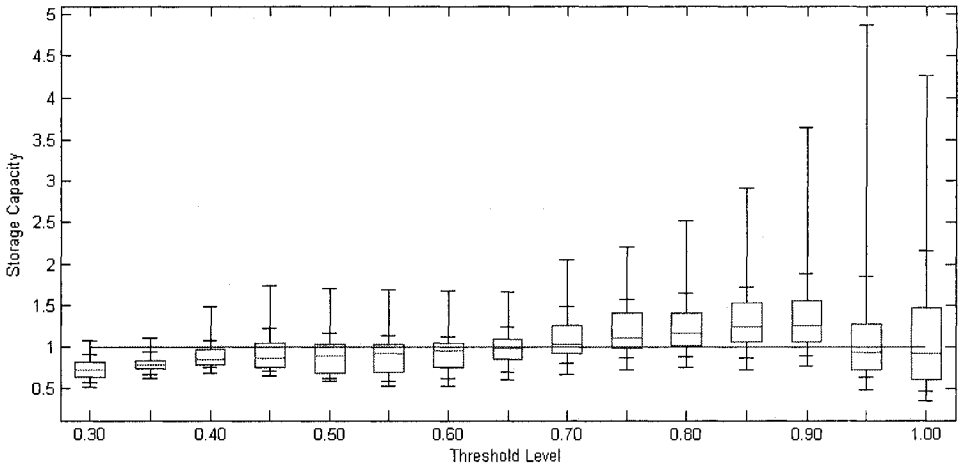


Figure 2-B.22 Storage Capacity of historical and generated (KGKA) monthly data with different threshold as TL*the overall mean of the historical monthly data for site 20 CRS, (unit : Acre-feet)

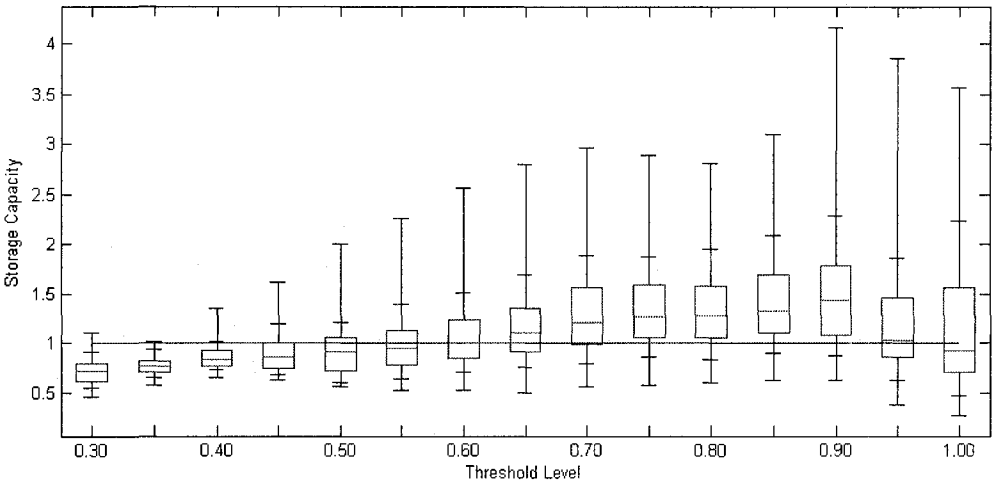


Figure 2-B.23 Storage Capacity of historical and generated (KGKP) monthly data with different threshold as TL*the overall mean of the historical monthly data for site 20 CRS , (unit : Acre-feet)

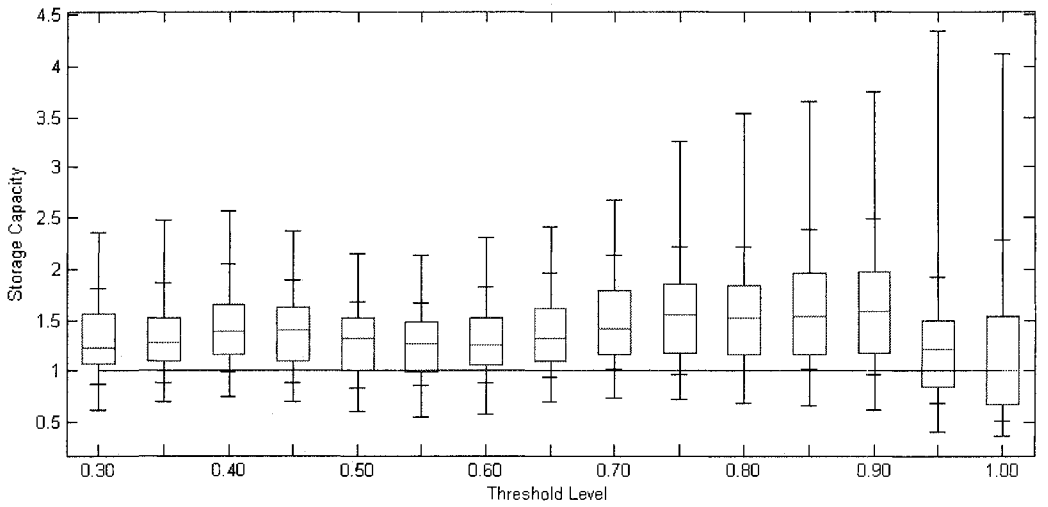


Figure 2-B.24 Storage Capacity of historical and generated (NPL) monthly data with different threshold as TL*the overall mean of the historical monthly data for site 20 CRS , (unit : Acre-feet)

Niger River

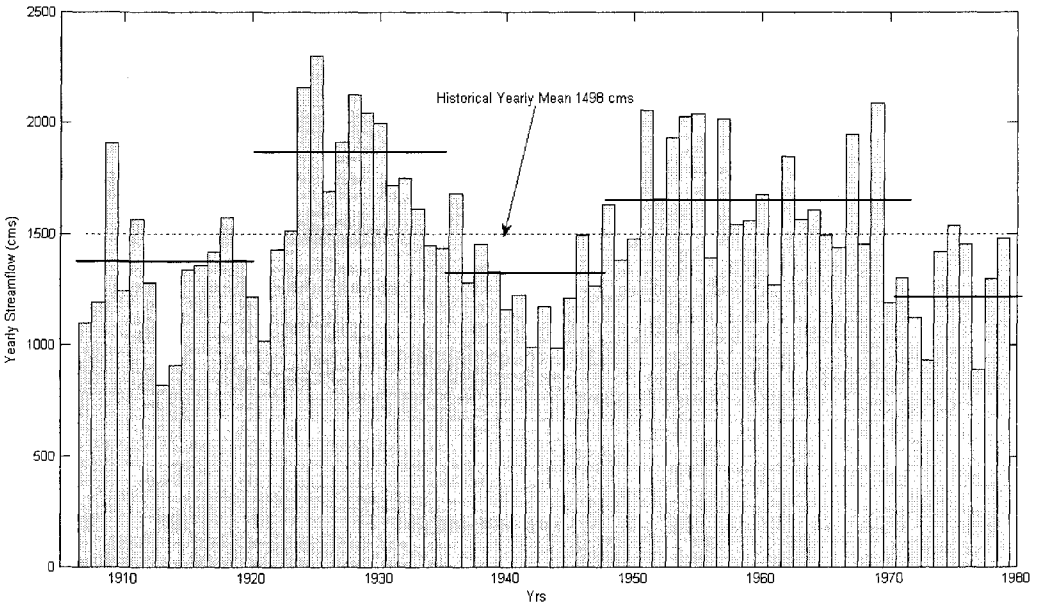


Figure 2-B.25 Time Series of Yearly streamflow for Niger River at Koulikoro with yearly mean (1498 m^3/s)

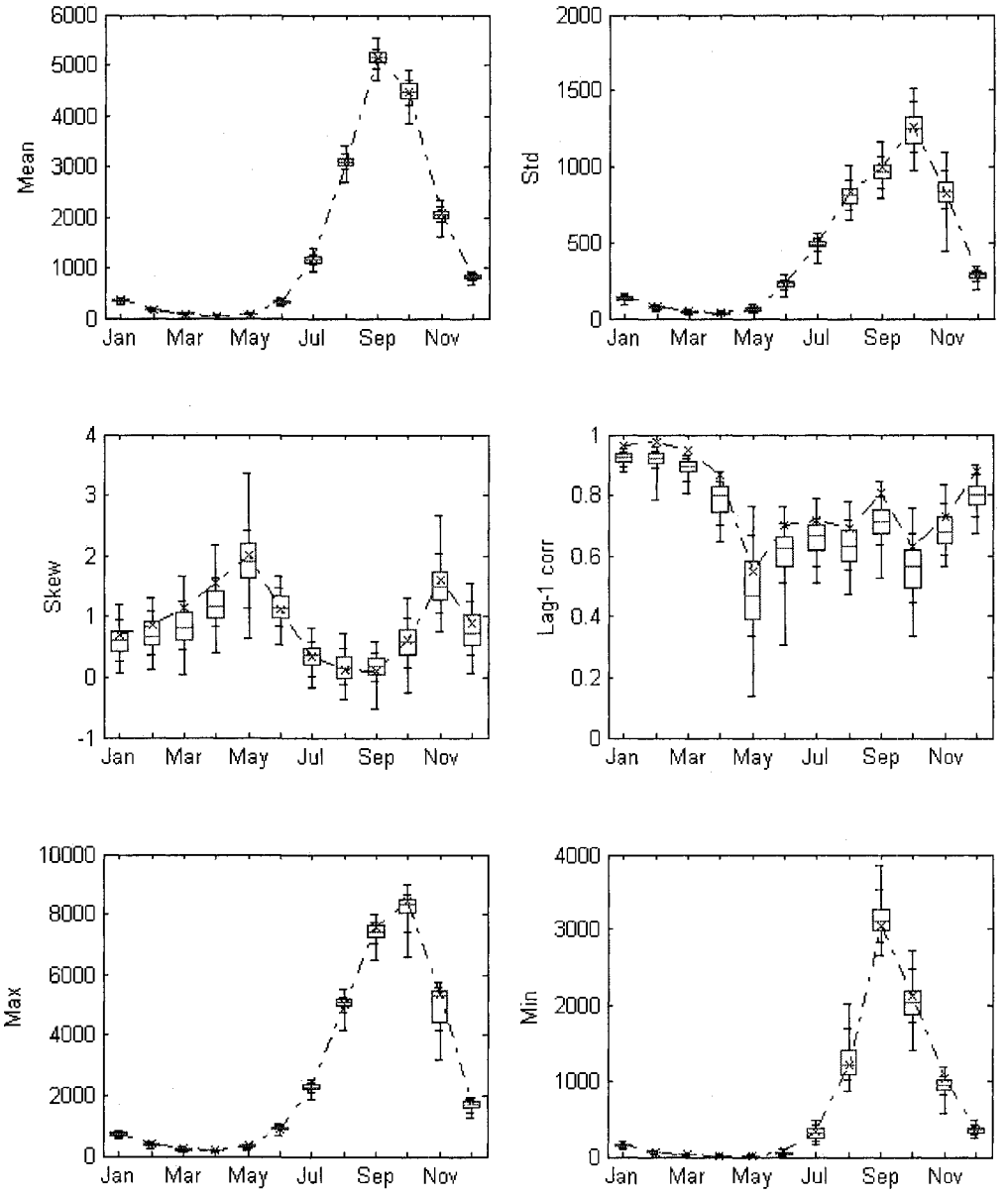


Figure 2-B.26 Key Statistics of Historical (dot line) and KGKA model simulations (boxplot) of the Niger River monthly streamflow (unit : m^3/s)

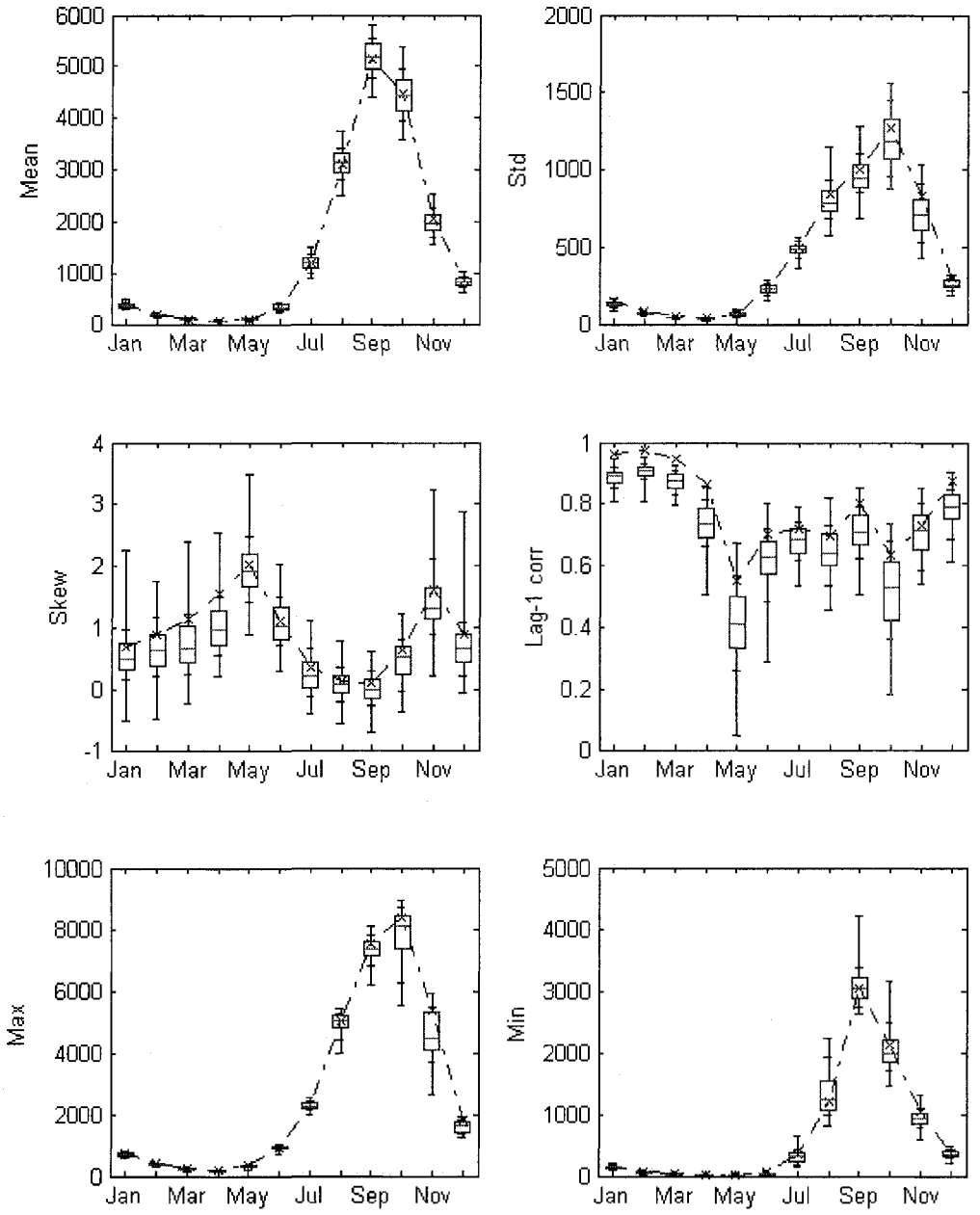


Figure 2-B.27 Key Statistics of Historical (dot line) and KGKP model simulations (boxplot) of the Niger River monthly streamflow (unit : m^3/s)

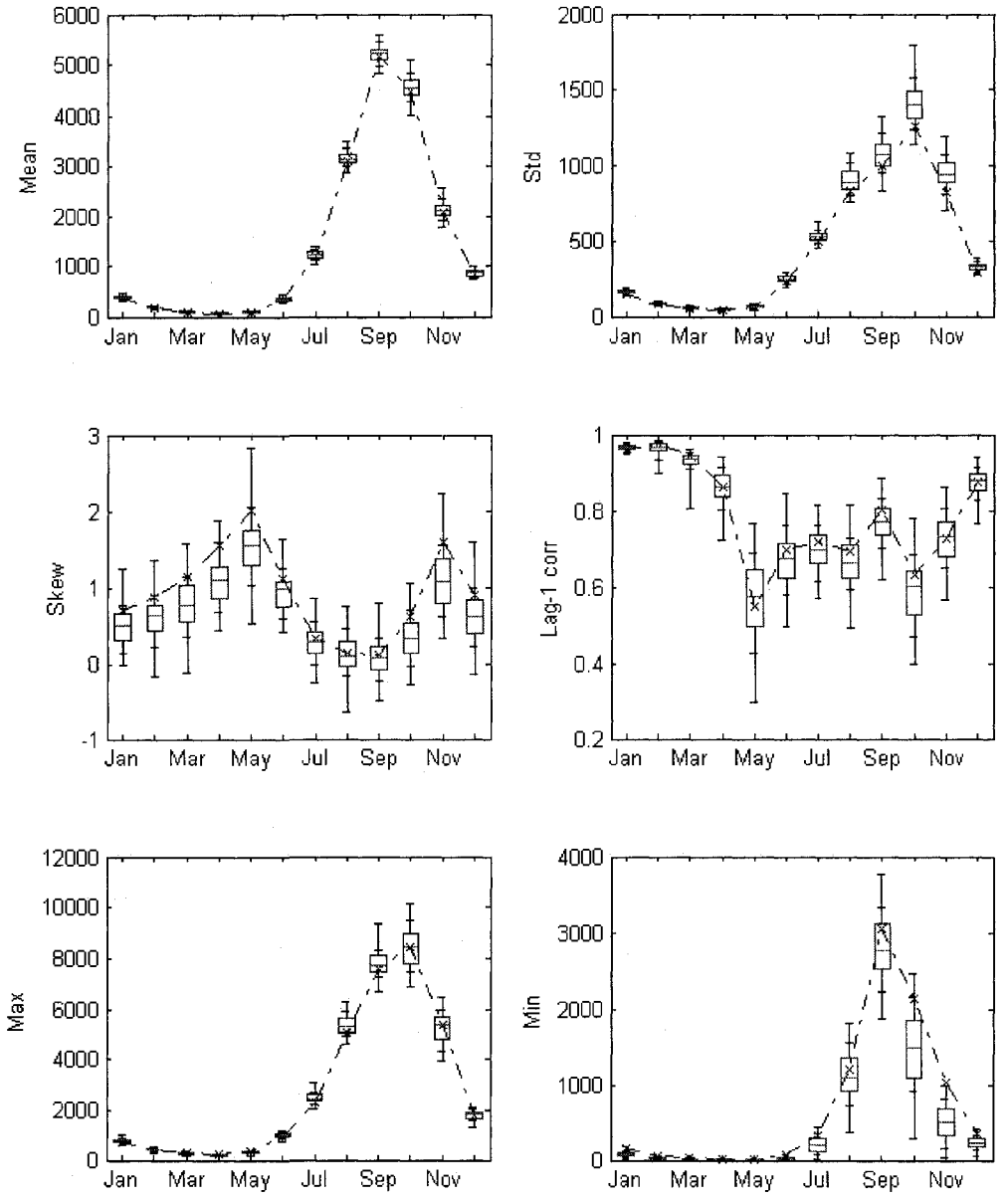


Figure 2-B.28 Key Statistics of Historical (dot line) and NPL model simulations (boxplot) of the Niger River monthly streamflow (unit : m^3/s)

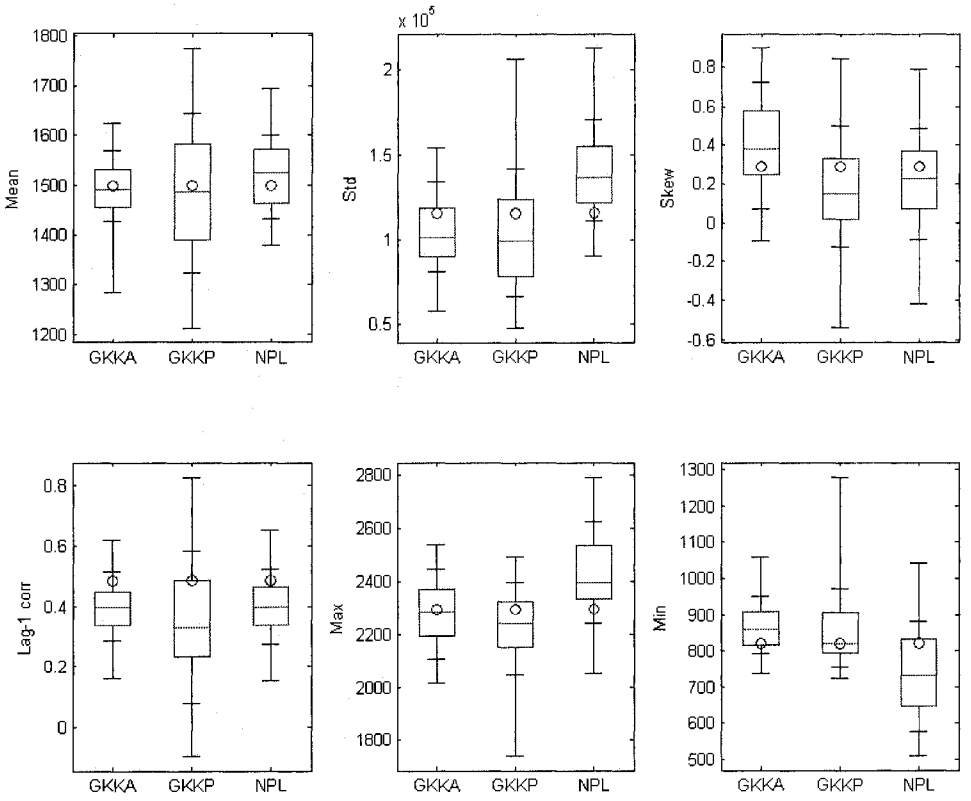


Figure 2-B.29 Key Statistics of Historical (circle) and simulated from KGKA, KGKP, and NPL model (boxplot) of the Niger River yearly streamflow (unit : m^3/s)

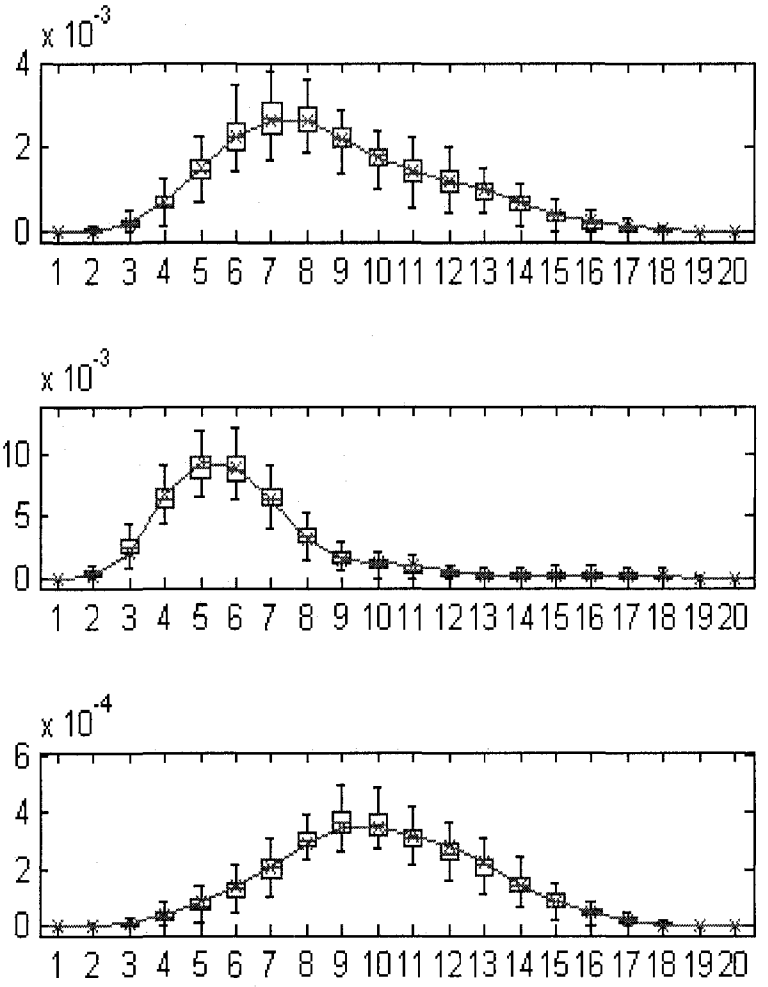


Figure 2-B.30 Kernel density estimate of historical (segment line) and generated (boxplot) monthly streamflow data for Niger River at Koulikoro from KGKA model for month 1, 5, and 9

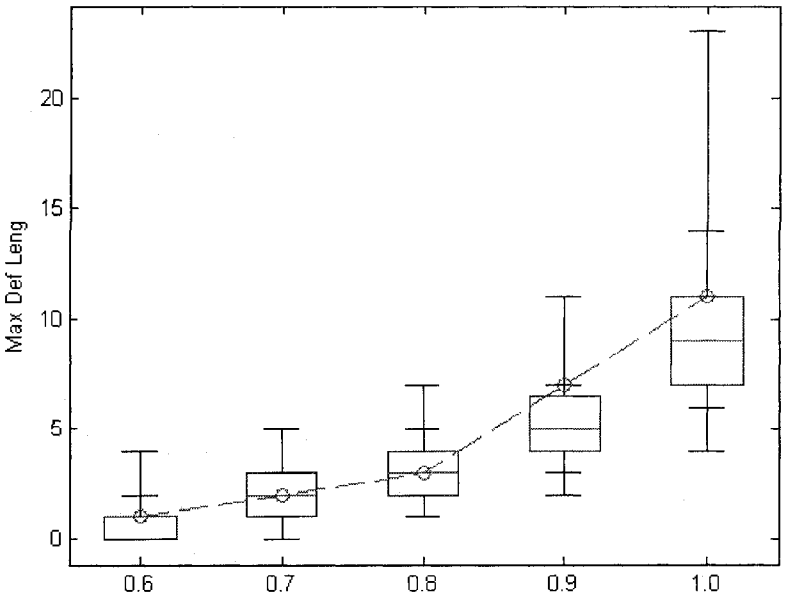


Figure 2-B.31 Maximum Deficit Length of historical and from historical (circle) and generated yearly data from KGKA model (boxplot) for Niger River at Koulikoro with different threshold level

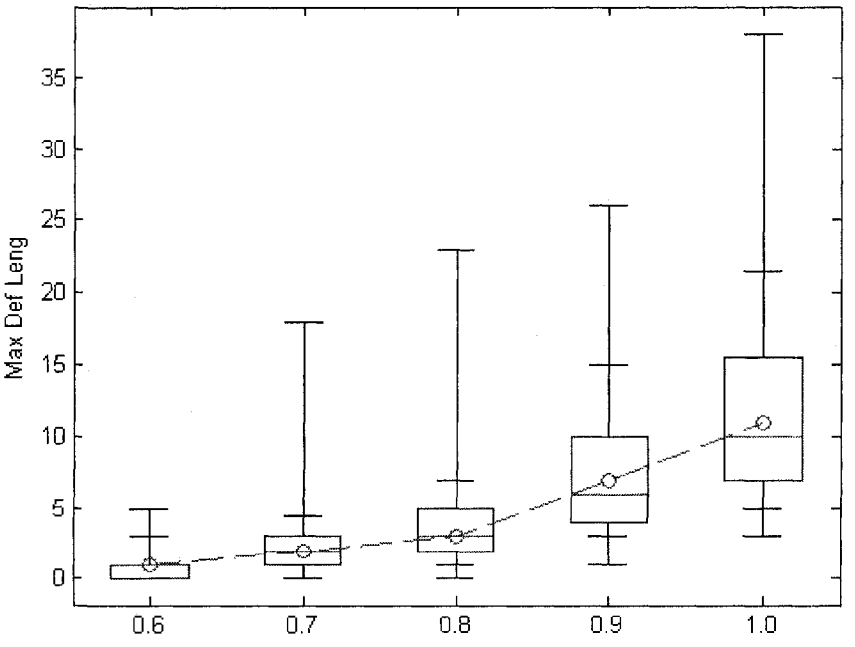


Figure 2-B.32 Maximum Deficit Length of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Niger River at Koulikoro with different threshold level

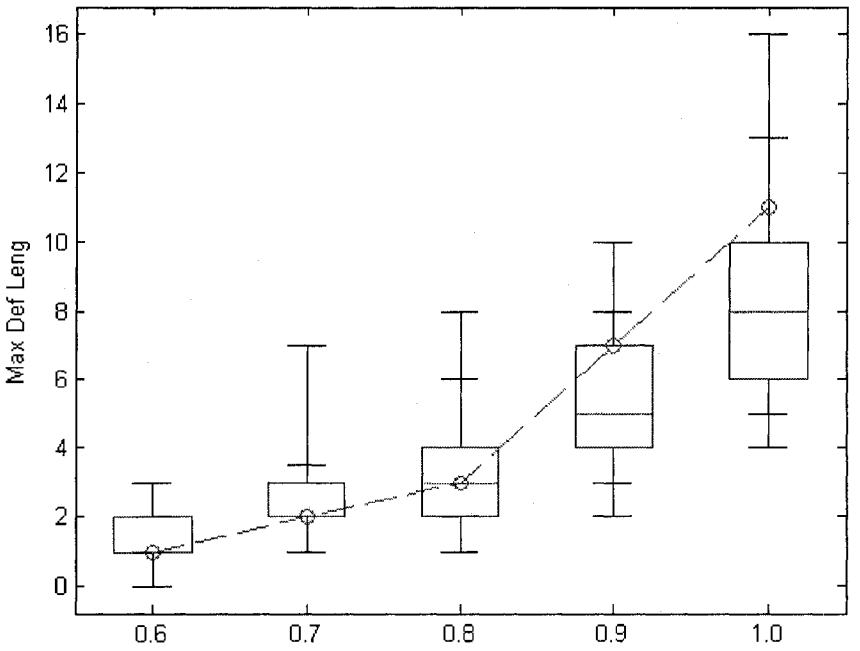


Figure 2-B.33 Maximum Deficit Length of historical and from historical (circle) and generated yearly data from NPL model (boxplot) for Niger River at Koulikoro with different threshold level

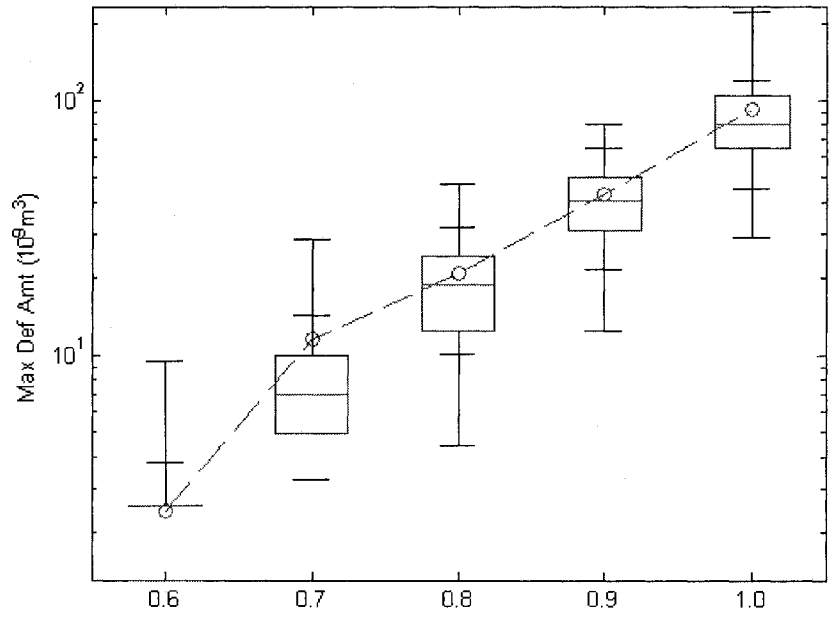


Figure 2-B.34 Maximum Deficit Amount (10^9 m^3) of historical and from historical (circle) and generated yearly data for Niger River at Koulikoro from KGKA model (boxplot) with different threshold level

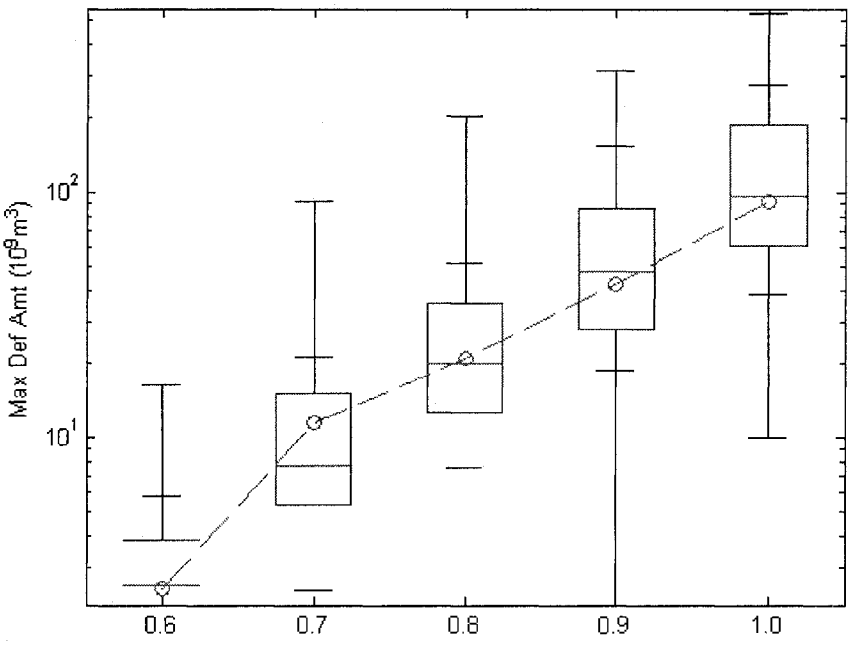


Figure 2-B.35 Maximum Deficit Amount (10^9 m^3) of historical and from historical (circle) and generated yearly data for Niger River at Koulikoro from KGKP model (boxplot) with different threshold level

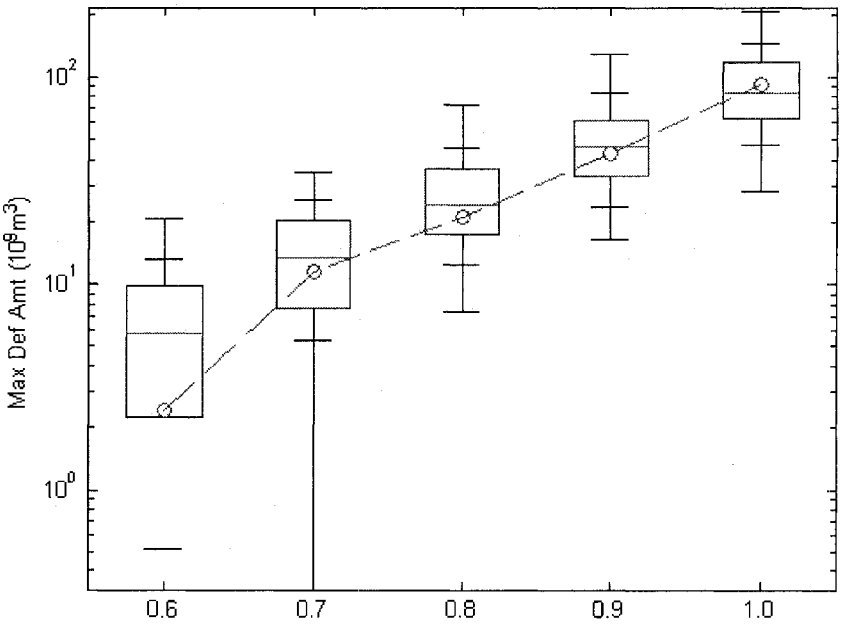


Figure 2-B.36 Maximum Deficit Amount (10^9 m^3) of historical and from historical (circle) and generated yearly data for Niger River at Koulikoro from NPL model (boxplot) with different threshold level

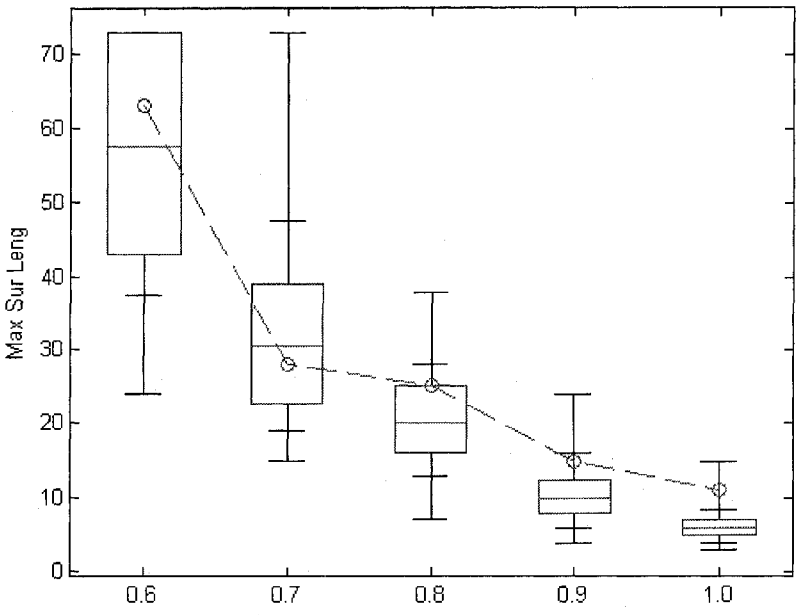


Figure 2-B.37 Maximum Surplus Length of historical and from historical (circle) and generated yearly data from KGKA model (boxplot) for Niger River at Koulikoro with different threshold level

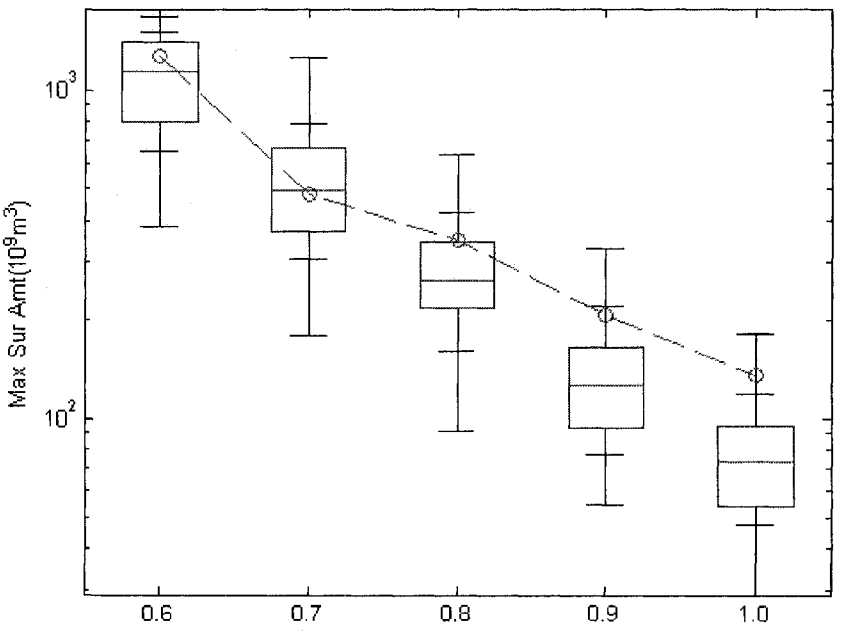


Figure 2-B.38 Maximum Surplus Amount (10^9 m^3) of historical and from historical (circle) and generated yearly data from KGKA model (boxplot) for Niger River at Koulikoro with different threshold level

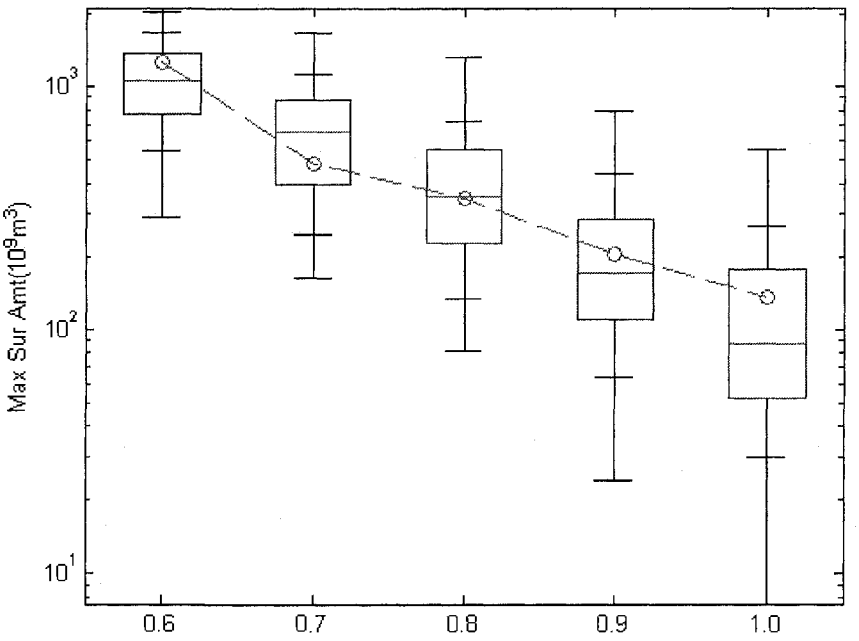


Figure 2-B.39 Maximum Surplus Amount (10^9 m^3) of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Niger River at Koulikoro with different threshold level

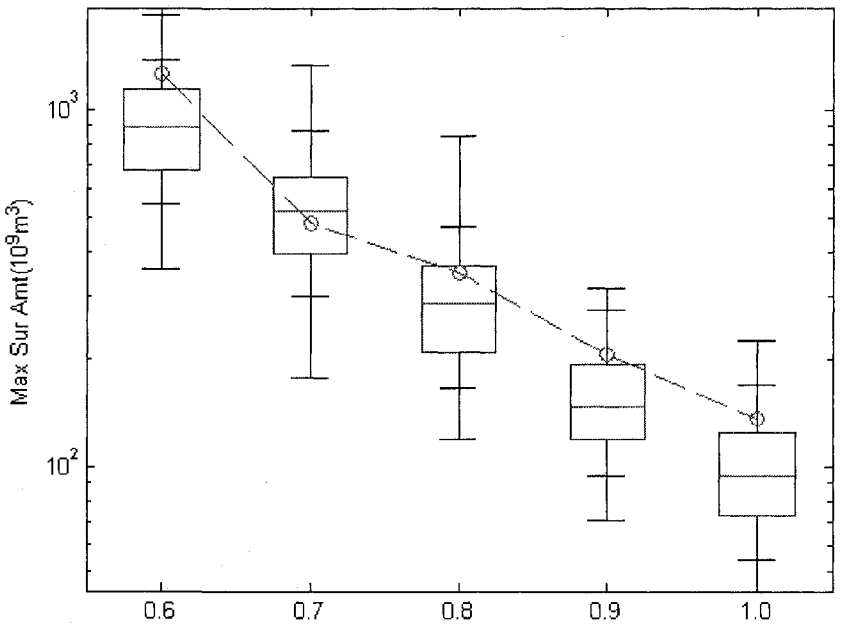


Figure 2-B.40 Maximum Surplus Amount (10^9 m^3) of historical and from historical (circle) and generated yearly data from NPL model (boxplot) for Niger River at Koulikoro with different threshold level

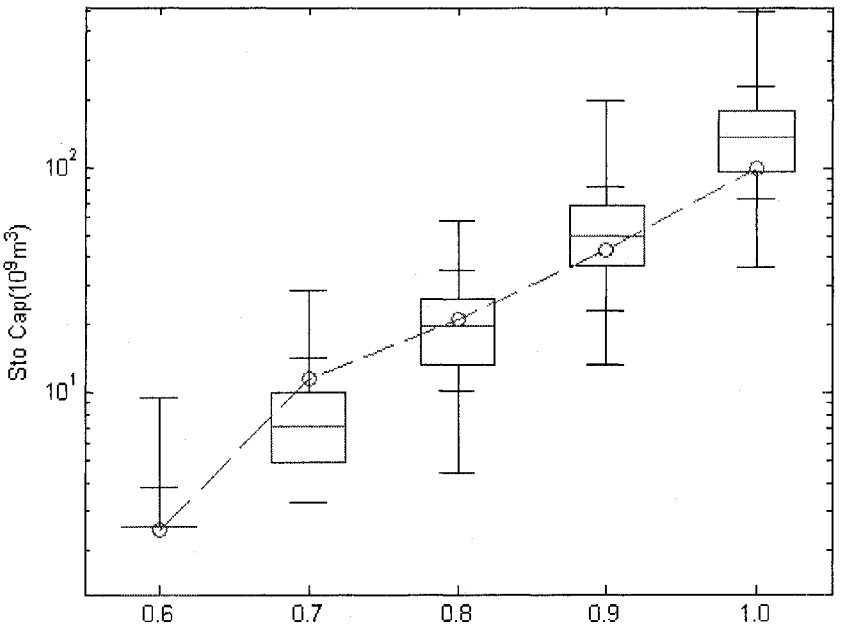


Figure 2-B.41 Storage Capacity (10^9 m^3) of historical and from historical (circle) and generated yearly data from KGKA model (boxplot) for Niger River at Koulikoro with different threshold level

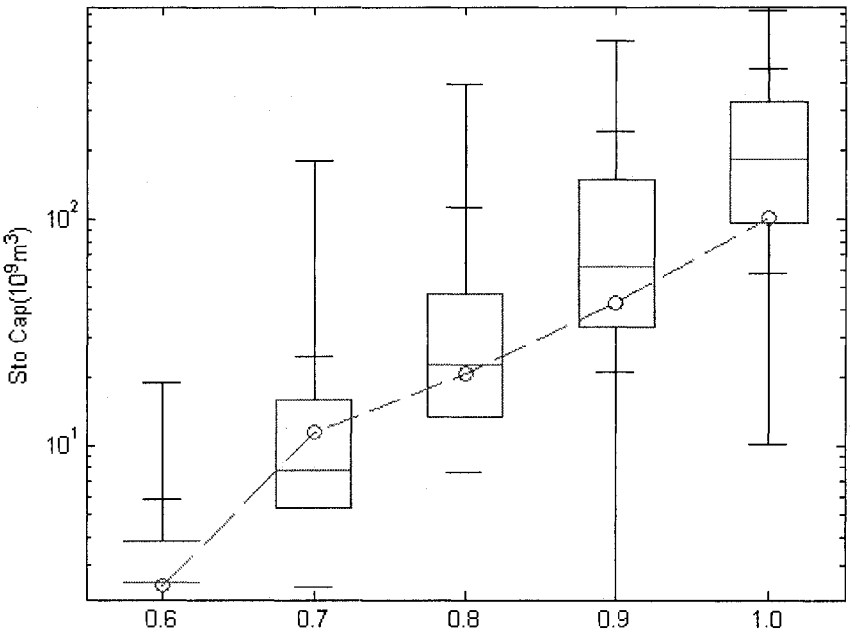


Figure 2-B.42 Storage Capacity (10^9 m^3) of historical and from historical (circle) and generated yearly data from KGKP model (boxplot) for Niger River at Koulikoro with different threshold level

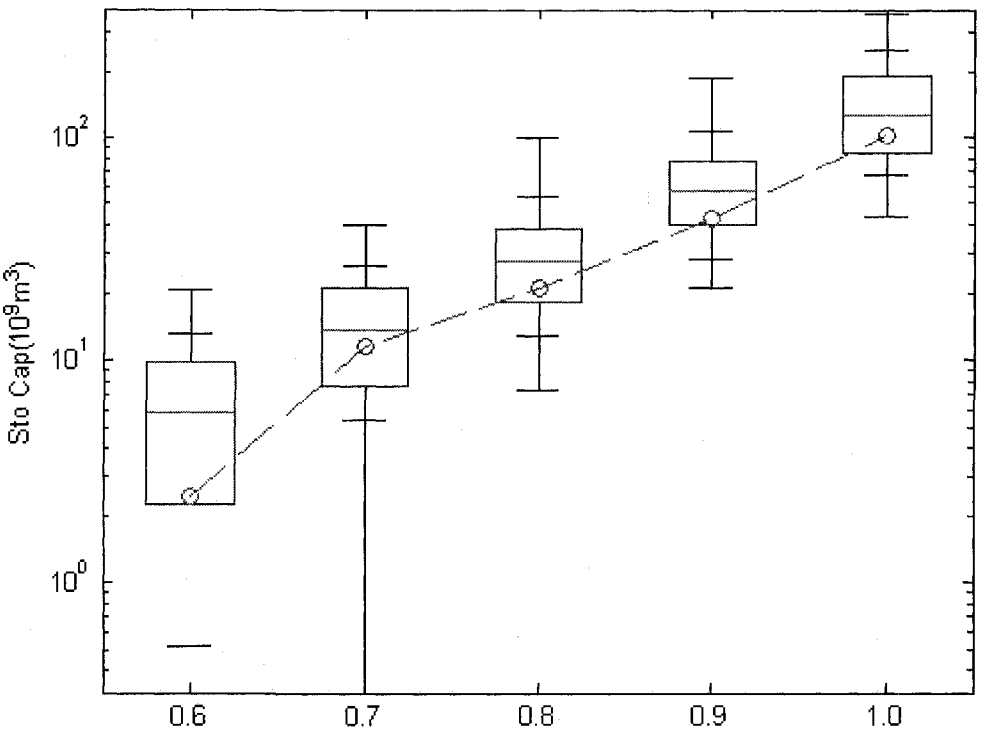


Figure 2-B.43 Storage Capacity (10^9 m^3) of historical and from historical (circle) and generated yearly data from NPL model (boxplot) for Niger River at Koulikoro with different threshold level

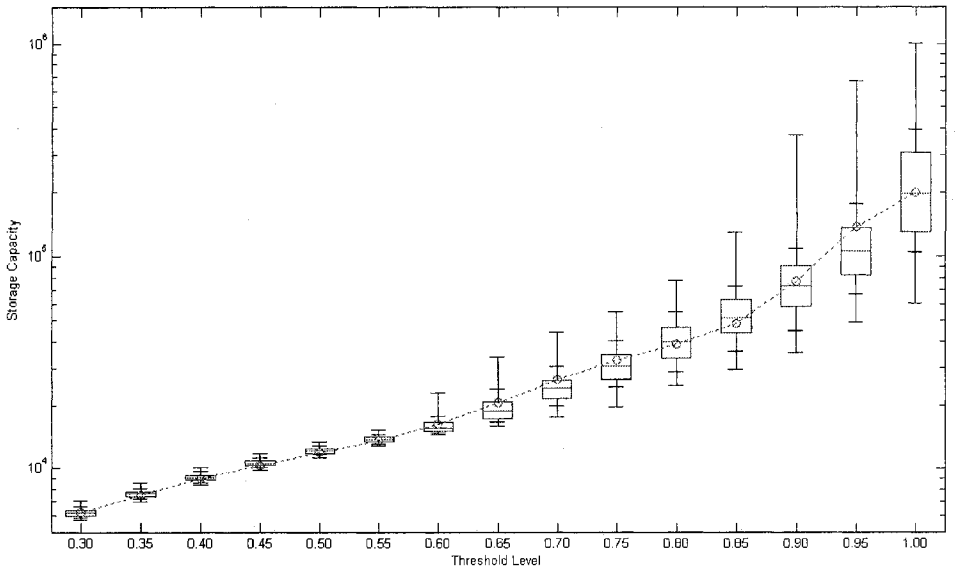


Figure 2-B.44 Storage Capacity of historical data and generated (KGKA) data with different threshold as (TL*the overall mean of the historical monthly data), (unit : m^3)

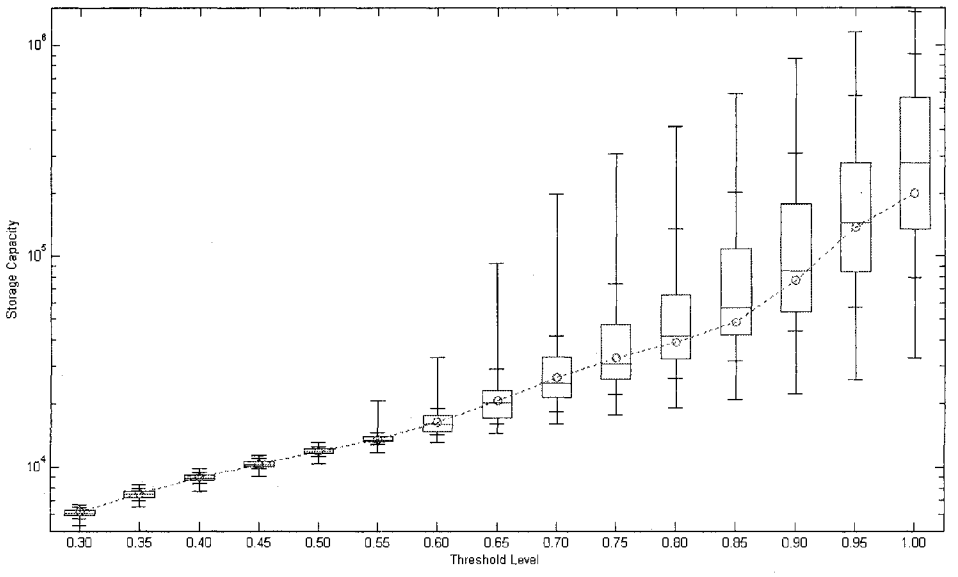


Figure 2-B.45 Storage Capacity of historical and generated (KGKP) monthly data with different threshold as (TL*the overall mean of the historical monthly data) (unit : m^3)

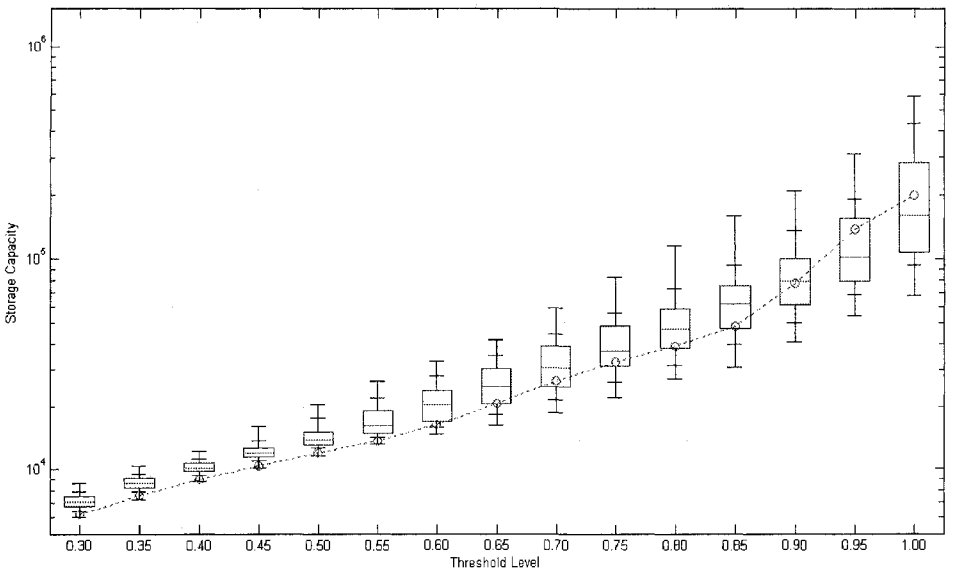


Figure 2-B.46 Storage Capacity of historical and generated (NPL) monthly data with different threshold as (TL*the overall mean of the historical monthly data) (unit : m^3)

CHAPTER III

NON-PARAMETRIC MULTIVARIATE SIMULATION OF INTERMITTENT AND NON-INTERMITTENT MONTHLY STREAMFLOWS

3.1 Introduction

Stochastic simulation models have been broadly employed in water resources to analyze the statistical characteristics of drought or flood, control plans on water resources systems. Multivariate seasonal time scale generation data are generally used for the drought analysis and planning of the water resources in a river network system. Seasonal streamflow data, however, has peculiar characteristics that hinder proper modeling such as high skewness, bimodality, intermittency, long-term persistency and non-linear relations, as well as seasonality and serial and cross correlation.

The main objective of this paper is to develop a simulation model for multivariate seasonal streamflow data with the combination of the intermittent sites and the non-intermittent sites. For this purpose, multivariate nonparametric techniques are employed based on the matched block bootstrapping technique. To simulate variable data between the generated sequences and to produce the values not in historical data, a perturbation

process will be utilized after selection of the historical data point with Gamma Kernel Density estimate (Lee and Salas, 2008a). The suggested model scheme will preserve the interannual variability also. Furthermore, the conceptual Genetic Algorithm process is applied to blend the generated sequences so that the generated data can be mixed spatially.

In Section 2, how those problems have been handled in simulation modeling literature is described. And the suggested modeling procedure is explained, including the techniques to preserve the annual variability and shun the repetition of the historical seasonal and spatial patterns in Section 3. The data and statistics for model verification are described in Section 4. The application and results are shown in Section 5, with the following conclusion in Section 6.

3.2 Brief Review of Literature

First of all, seasonal streamflow data are generally positively skewed while annual streamflows are close to be symmetric although, in some cases, annual may be significantly skewed. The majority of the existing Multivariate Autoregressive Moving Average (MARMA) models, as well as disaggregation models, assume that the data are normally distributed. For skewed streamflow data, various data transformation techniques for the current parametric normal based models (e.g. Autoregressive Moving Average, ARMA models) have been developed such as log, power, Box-Cox, and gamma (Salas, 2006). Still, the generated data in the original domain yields some bias during the back-transformation without bias on the transformation domain.

Secondly, some researchers believe that the marginal distribution of the monthly or higher time scale streamflow data has bimodality or multimodality (e.g. Lall and Sharma, 1996; Prairie et al., 2006; Sharma and O'Neill, 2002). However, this is still controversial since two or more different systems (e.g. snow-melt and precipitation) should affect the streamflow in a certain scale to have a bi- or multi-modal marginal distribution. Otherwise, it might be just a matter of degree of smoothing or inducing from record shortage. For example, a short record can easily produce the bi or multi modal distribution because of random sampling. To prove what causes the multimodality distribution, meticulous work for each river streamflow should be performed. For example, the regional consistency of the multimodality in marginal distributions may buttress that those multimodality are real.

Thirdly, preservation of long-term (e.g. annual) persistency from a lower time scale (seasonal) model simulation is a difficult task in modeling. A monthly model preserving the long-term persistency should include a longer number of seasons in parameterized terms (Vecchia et al, 1983). Disaggregation models have been used with (1) generating the yearly model first, then (2) disaggregating it into seasonal data (Valencia Schaake, 1973; Stedinger, 1985). Disaggregation models generally request a tremendous number of parameters. A nonparametric disaggregation approach has been devised by Tarborton et al.(1998), and Prairie (2007) to avoid the oversized parameterization. However, those cannot preserve the serial correlation between the last month of the previous year and the first month of the current year.

Fourthly, parametric time series models generally use the linear relations (serial and cross between variables). After transforming data into normal domain, the ARMA

type parametric models employs the linear relations. However, there are still many cases in which the relations are not linear and also modeling in transformed domain induce some biases on a back-transformed original domain.

Lastly, intermittency, non-event occurrence between events, in coarser than seasonal time scale streamflow data may occur in arid or semi-arid regions, as well as weekly or daily. The intermittency hinders to apply the existing simulation models such as Periodic ARMA (PARMA) model since it yields a lot of negative simulation values combining with high skewness and brings discontinuity. Beard (1973) and Srikanthan (1979) developed the truncated type model in that if the generated value is negative, assign it as zero. The model, however, yields significant bias on the basic statistics of the generated data. Alternatively, a product model has been suggested combining a binary occurrence process with the amount process (Chebaane et al., 1995). For the binary process, a periodic discrete ARMA (PDARMA) has been fitted and a PARMA or Periodic Gamma Autoregressive (1) processes (PGAR(1)) have been applied for the amount process. Since the PARMA model is restricted to a normal distribution based, some bias on the generated data produce some bias on the key statistics. The PGAR(1) model might be a good alternative instead of the PARMA model. However, the parameter region is so restricted that some data might not be applicable or require further analysis such as Fourier transformation.

Because of the intricate modeling procedures and drawbacks of the intermittent seasonal time series, a simple nonparametric approach has been developed by Svanidze (1978) and Srikanthan and McMahon (1980), named the Method of Fragments (MF). Here, the fragments refer to the ratios of the seasonal values divided by the yearly so that

the sum of the ratios of all seasons at a certain year is unity. The drawback of this approach is that the serial correlation between the last month of the previous year and the first month of the present year is not preserved. Furthermore, Potter and Pink (1991) argued that the drawback of MF is the noticeable duplication of seasonal patterns occurring especially on generating a time series longer than the historical or large number of sets. One of the purposes of data generation is to analyze risks from the unprecedented event of high flows or low flows and to employ the risk analysis into the water resources system design. The repetitive seasonal patterns of the generated data from MF might lead to significant mistakes on decision making.

A large river basin, such as the Colorado River, may include intermittent and non-intermittent flows together. The simulation modeling techniques for multisite data combining intermittent and non-intermittent seasonal streamflow have not been fully developed in literature. Some plausible techniques that surmount part of the difficulties described above are overviewed as follows. One of the simplest approaches is block bootstrapping, resampling the data block from the historical data for synthesized data. This method was developed originally by Kunsch (1989) and applied for hydrologic simulation by Vogel and Shallcross (1996). This approach, however, encounters some drawbacks such as: (1) discontinuity between blocks from block-by-block sampling; (2) repetition of the same sequences of the blocks; and (3) generation of historical values only. In the Srinivas & Srinivasan (2005) article, they proposed a hybrid model with low order Periodic Autoregressive (PAR) and the block bootstrapping of the innovation terms to overcome those drawbacks. Firstly, the seasonal streamflow data are standardized and fitted using PAR(1) model, where the innovation term for each site and each season are

stored. In simulation, (a) The innovation terms are block bootstrapped where a block implies the multisite and multi-season block, and the block length should be a multiple of the number of seasons (e.g. 24 months are used in the paper); (b) After setting the initial value to zero for all sites, synthetic data are simulated recursively with the innovation terms and the fitted PAR(1) model; (c) The data are inverse standardized to transform the generated data back to the original domain. However, this hybrid model has some significant drawbacks too. At first, it is not feasible for intermittent time series modeling. Even if the original block bootstrapping can reproduce the intermittent process, the fitting procedure with PAR(1) cannot be applicable for the intermittent data. Second, it may generate negative values especially where the seasonal streamflow data are highly skewed (e.g. exponential type marginal distribution). And third, the seasonal generated data of later seasons will be almost the same as historical. The first part of the seasons (e.g. month 1, 2, and 3) generates different values from the historical. However, the other parts of the seasons (month 10, 11, and 12) generate almost the same values of the historical. The more elaborate explanation will be followed in the application section.

Moreover, Markovian Matched Block Bootstrapping (MBB) method was developed by Carlstein et al.(1998) to surmount the block discontinuity. The principle of this method is to compare the last element of the historical blocks and the last value of the generated data. From assigning a probability for each block according to the distance estimated from the comparison, choose the next block. Srinivas and Srinivasan (2006) applied this procedure for the resampling of a seasonal hydrologic time series. They used uniform distribution for the assigning probability with a certain number of blocks as a

range. But, the MBB method employs a somewhat intricate procedure. Here, a simple k-nearest neighbor resampling technique replaces this model.

Furthermore, other approaches have been applied for at modeling daily weather variables. The weather simulators require combining the intermittent variable, precipitation, with the non-intermittent variables such as maximum temperature, minimum temperature, and wind velocity. This might be useful for modeling the river network combining intermittent and non-intermittent stations which have not been applied for streamflow data. Rajagopalan and Lall (1999) extended the k-nearest neighbor resampling approach (Lall and Sharma, 1996) to simulate multivariate weather variables. Also, many researchers have improved this technique (Yates et al., 2003; Buishand and Brandsma, 2001). However, both the MBB and the extended K-Nearest Neighbors Resampling (KNNR) models have the drawback that there is no variability in cross relation. In other words, there is no chance to be mixed between variables. The resampled multivariate sequences are mixed on not spatially but temporally. This is the same drawback as the repetition of the same seasonal pattern.

Even though the seasonal streamflow (combined with intermittent and non-intermittent) and daily weather variables have many of similar characteristics for simulation modeling, the seasonal streamflow data has stronger seasonality than daily weather variables. Therefore, a periodic model should be fitted into the seasonal streamflow data, while a stationary model or ranged stationary model is employed for a daily weather variable with a certain period, such as a month.

3.3 Mathematical Description of Model Components

The model scheme suggested here employs matched block bootstrapping targeting on generating a unique data set with the preservation at key statistics of seasonal and yearly time scales. The description below starts from the exhibition of the fundamental block bootstrapping followed by the extension to MBB and the modifications thereof. A distinctive blending process adopting the Genetic Algorithm is applied in order to obtain the sequences with different spatial combinations. Afterward, further improvements are stated, such as perturbing the resampled data with Gamma KDE to attain the new unprecedented values other than historical data and manipulating.

The elementary notations employed in this paper are expressed here. A set notation is employed as $a \in [b,c]$, implying that the integer index, a, is ranged from b to c. For instance, $\{X_a\}_{a\in[b,c]} = \{X_b, X_{b+1},..., X_c\}$, where $a \le c$. Suppose that seasonal streamflow data is available with ω number of seasons and N number of years, and x_ν is employed to express yearly observed data at year ν, while $x_{\nu,\tau}$ is for monthly data at year ν and month τ. For multisite data, $x^s_{\nu,\tau}$ is used for the monthly streamflow for year ν, month τ at site s with $s \in [1,S]$, where S is the number of sites. The symbol is capitalized to illustrate a variable or generated data corresponding to the observed data. For instance, $X_{\nu,\tau}$ expresses a random variable or generated data in contrast to $x_{\nu,\tau}$. In addition, a vector notation is utilized to represent a set of data, e.g. $\mathbf{x}_{\nu,\tau} = \{x^s_{\nu,\tau}\}_{s\in[1,S]}$. The same vector notation is also applied for yearly data. $\mu^s_\tau(X)$ and $\sigma^s_\tau(X)$ represent the mean and

standard deviation of x for month τ at site s, respectively while $\hat{\mu}_\tau^s(x)$ and $\hat{\sigma}_\tau^s(x)$ represent corresponding estimates from the observed data. Also, $\mu(X \mid \Theta)$ is the mean of the variable X satisfying the condition Θ . Additional notations are specified with necessity along with description. Likewise, t is used to differentiate the generated data time step from the historical data time step, e.g. $X_{t,\tau}$ and t=1,…,T instead of $X_{\nu,\tau}$ when T is the required years of simulation.

3.3.1 Matched block bootstrapping and different block length

The simple bootstrapping scheme for a stationary (e.g. yearly) time series is sketched in Figure 3.1. Define $x_{B(i,j)}$ as the value of the j^{th} element of the i^{th} block. Each block consists of l elements, such that $\mathbf{x}_{B(i)} = \{x_{B(i,j)}\}_{j\in[1,l]}$ for i^{th} block. The historical data set with the record length N contain (N-l + 1) blocks since the blocks overlap (refer to Figure 3.1) the historical data. For a generation, choose a block randomly among the overlapped (N-l+1) blocks, each having the same probability 1/(N-l + 1), and the selected block is the length l generated data. For example, in Figure 3.1, it schematically illustrates the simple block bootstrapping. With nine years of record length and block length as three, seven historical blocks are structured. The second block of the historical blocks is selected for the first block of the resampled data, such that $X_1 = x_{B(2,1)}, X_2 = x_{B(2,2)}$, and $X_3 = x_{B(2,3)}$. This generated block is presented as $\mathbf{X}_{B(1)}$.The next resampled block $\mathbf{X}_{B(2)}$ is the seventh block of the historical data such that $\mathbf{X}_{B(2)} = \{X_4 = x_{B(7,1)} = x_7,\ \ X_5 = x_{B(7,2)} = x_8,\ \ X_5 = x_{B(7,3)} = x_9\}$ presented in Figure 3.1.

Notice that while the historical blocks do overlap the historical data, the generated blocks do not.

For a seasonal time series, the block length l should be a multiple of the total number of seasons, i.e. $l = \omega \times m$, where m is the number of years considered to preserve inter-annual variability (Srinivas and Srinivasan, 2001). The reason for this format is to reproduce seasonality and annual key statistics. Furthermore, m should be greater than one to take into account interannual variability (i.e. yearly serial correlation). Subsequently, a block of the seasonal time series is

$$\mathbf{x}_{B(i)} = \{x_{B(i,1)},\ldots, x_{B(i,l)}\} = \{x_{\nu,1}, x_{\nu,2},\ldots, x_{\nu,\omega}, x_{\nu+1,1},\ldots, x_{\nu+m-1,\omega}\}$$

where $i \in [1, N-m+1]$. A block is overlapped in the yearly scale such that $\mathbf{x}_{B(1)} = \{x_{1,1}, x_{1,2},\ldots, x_{1,\omega}, x_{2,1},\ldots, x_{m,\omega}\}$, $\mathbf{x}_{B(2)} = \{x_{2,1}, x_{2,2},\ldots, x_{2,\omega}, x_{3,1},\ldots, x_{m+1,\omega}\}$, and so on.

If yearly data are not serially correlated, which might often be the case for seasonal intermittent data due to the discontinuity of seasonal streamflow induced from zero values, then one is good enough for m. The major drawback of this bootstrapping method is the same as the one for the method of the fragments discussed by Poter and Pink (1991) and Lee and Salas (2008). The obvious seasonal patterns will occur repeatedly in the simulated data. To circumvent this shortcoming, Srinivas and Srinivasan (2006) suggested subdividing a year, in other words, splitting the monthly data of one year into non-overlapping within year blocks. The subdivided blocks of historical data are not overlapped. However, this subdivision will underestimate the yearly variance because the correlation of the seasonal data is attributed to the yearly variance and it does

not depend on only the lag-1 correlation. But it is broken from subdividing the yearly data. And, the seasonal pattern will be incurred recurrently but in subdivided time scale. The first month is the division point in the general block bootstrapping technique. And, seasonal division points (Srinivas and Srinivasan, 2006) are always the same through the generation. This yields that the yearly data (for general block bootstrapping) or the seasonal data are always the same as the historical.

Alternatively, we suggest assuming the block length as a random variable. Instead of the same division point for blocks, the division points are randomly changed since the block length is alternated at each block resampling. Suppose the block length is a discrete random variable with any feasible discrete distribution such as geometric or Poisson distribution. The Geometric and Poisson distributions were tested for a random variable of the block length and the results were no difference. From now on, the method description is based on the Poisson distribution. If the block length l is Poisson random variable, then the length can be generated from the Poisson distribution as:

$$p(l') = \frac{e^{-\lambda}}{\lambda^{l'}(l')!} \tag{3-1}$$

where l'=0, 1,..... and $l=l'+1$. l' is employed instead of l to abstain from generating zero values. An example procedure of the block length variable is as follows:

(1) A block length is generated from the Poisson distribution Eq.(3-1), say l=4.

(2) Choose a block from the historical data. The historical blocks should start with the same month of the division point. If the previous generated blocks end at τ, then the division point for all the historical blocks are τ+1. For

example, if one start with $\mathbf{X}_{B(1)} = \{X_{1,j}\}_{j\in[1,4]}$, then the possible historical blocks consist of $\mathbf{x}_{B(1)} = \{x_{B_{1,1}},..,x_{B_{1,4}}\} = \{x_{1,1},..,x_{1,4}\}$, $\mathbf{x}_{B(2)} = \{x_{2,1},..,x_{2,4}\}$, …, $\mathbf{x}_{B(N)} = \{x_{N,1},..,x_{N,4}\}$. One among *N* blocks is selected for $\mathbf{X}_{B(1)}$.

(3) Suppose *l*=3 for the second generate block (suppose it is generated from the Poisson distribution) with conducting Step(1) above so that $\mathbf{X}_{B(2)} = \{X_{1,5}, X_{1,6}, X_{1,7}\}$ and the historical possible blocks are $\mathbf{x}_{B(1)} = \{x_{B_{1,1}},..,x_{B_{1,3}}\} = \{x_{1,5}, x_{1,6}, x_{1,7}\}$, …, $\mathbf{x}_{B(N)} = \{x_{N,5}, x_{N,6}, x_{N,7}\}$. One among those N historical blocks is selected as $\mathbf{X}_{B(2)}$. Repeat this process for as many years of required simulations as are necessary.

Notice that this process does not produce any discontinuity between years since a block can crossover two years, when $\mathbf{X}_{B(i)}$ contains the elements with two different years. For example, if the previous generated block stops at year ν and month $\omega - 2$ and the generated block length *l*=3, then $\mathbf{X}_{B(i)} = \{X_{\nu,\omega-1}, X_{\nu,\omega}, X_{\nu+1,1}\}$. The parameter λ of Poisson distribution in Eq.(3-1) is directly related to the mean value of *l* as $E[l] = E[l'] + 1 = \lambda + 1$. λ is estimated from E[*l*]. The higher the serial correlation of monthly and/or yearly data is, the larger this mean will be. Srinivas and Srinivasan (2005) suggest that the block length as four is the suitable number to preserve the serial correlation and cross-yearly serial correlation with matching the following block algorithm explained later. By the same token, the mean value of the block length can take this amount. From an experiment on the different mean values in this study (not shown), three to six is appropriate if there is no strong cross-year correlation, otherwise six to twelve might be used. The random

variation of the record length brings different combinations of the seasonal data so that the repetition of the same seasonal pattern will not occur. Speculatively, the block length variable simply eliminates the repetition issue. Furthermore, it allows producing the generated data set with exploring different combinations of the historical seasonal data.

Yet, the discontinuity between blocks still cannot be resolved. The way to connect a block with neighbor blocks should be proposed. Carlstein et al. (1998) attains the requirement using matched block bootstrapping. A following block is selected from the probability assigned corresponding to the distances from the last elements of the recently generated block to the previous condition of the historical candidate blocks. Srinivas and Srinivasan (2005) applied this procedure for resampling a univariate seasonal hydrologic time series. From the subdivided within a year block of the historical data, the last elements from each block are ordered at first. The nearest neighbors are obtained according to the order of the last element of the recent generated block. Here, one among a certain number of neighbors is selected randomly with the same probability. The following block of the selected neighbor is taken as the next generated block. The different numbers of neighbors were tested to find an appropriate number, and the result revealed that five neighbors reproduced the monthly serial correlation and inter-year serial correlation well. Alternatively, instead of using the uniform distribution and ordering to choose the matched block, we offer to utilize the k-NN resampling algorithm (KNNR; Lall and Sharma, 1996). A subsequent block is selected from the condition of the last element of the preceding generated block and the previous condition of the candidate historical blocks. The selection is attained with KNNR. After being suggested by Lall and Sharma (1996), KNNR has been flourished in hydrologic literature because

of the simple and effective way to model serial relations (Yates et al. 2003, Buishand and Brandsma, 2001). The original algorithm of KNNR in Lall and Sharma (1996) for stationary time series generation is summarized below since it is employed many other places of this paper as well.

(1) Define the current and historical feature vectors D_t and D_v respectively, and the number k. D_t and D_v are vectors whose components are the conditional variables to resample. Here X_{t-1} is used for the feature vector such as $D_t = \{X_{t-1}\}$ so as x_{v-1} for D_v. And the number of neighbors (k) is estimated from the heuristic method, $\sqrt{N}$, suggested by Lall and Sharma (1996). Assuming that we know the initial value X_0, the next key steps are followed.

(2) Estimate the distance between the feature vector of the historical and the current state as

$$r_v = \left[\sum_{j=1}^{J} w_i (d_{tj} - d_{vj})^2 \right]^{1/2} \tag{3-2}$$

where w_i is the scaling factor of each J component where J is the number of the conditional variables. This factor is employed for which each conditional variable equally attributes to the distance. Since only one variable is utilized as suggested in Step (1), w_i is not necessary. Therefore, it is expressed as $r_v = |X_{t-1} - x_{v-1}|$. d_{tj} stands for j^{th} component of the current feature vector and d_{vj} is the j^{th} component of the i^{th} year historical feature vector.

(3) Among the smallest k neighbors, one of them is selected from the weighted distribution as:

$$p_{is} = \frac{1/is}{\sum_{ls=1}^{k} 1/ls} \tag{3-3}$$

where is=1,…,k. This probability shows that the neighbor with the closer distance weighs high probability to be selected and vice versa. The selection from the discrete weighted probability within a certain range (1,..,k) can be also done by Roulette wheel selection in the Genetic Algorithm literature (Goldberg 1989).

(4) The subsequent value of the selected neighbor is obtained as X_t. This procedure continues until obtaining the T length of the generated data as is supposed.

The application of KNNR to find a matched block is facile. Schematically, the first element of the following generate block is found using the KNNR approach with the same step above. Here, the last element of the recently generated block, is assigned as the feature vector $D_t = X_{t-1}$ and the last elements of the plausible historical blocks for the historical feature vector such as $D_v = \{x_{B(i-1,l)}\}$,and i=2,…, N-l+1. Afterward, the subsequent l-1 data values of the first point are chosen to complete the synthetic data block.

The simplified algorithm is (a) to choose a value from one to k, say k* and (b) to find the neighbor of the k*th smallest distance in Eq.(3-2). This scheme is less time

consuming in a generation since ordering all distances have to be performed at each generation in general. The searching algorithm to find k*th smallest distance is well described in Press (2002, in Chapter 8.5).

3.3.3 Blending process with Genetic Algorithm

The objective of this paper is to develop a methodology for the multisite seasonal streamflow data in which part or all sites are intermittent. The proposed approach for this objective here is the matched block bootstrapping with the variable block length and the KNNR algorithm to find the matched following block. To manipulate the cross-correlation of multisite data, a summary statistics, suggested by Buishand and Brandsma (2001), is employed to abridge the multivariate dimension problem into the univariate one, explained later. The handling of the multisite data in this way, however, resulted in the fact that the generated data are not mixed between different sites. The generated multisite data set of the certain generate year and month always originates from the same historical year. For example, suppose that we need to generate the S site data starting from year t=10 and month τ =5; the generated block length is one, and the historical year of the bootstrapped data is eight. And the selected data elements from the historical data is $\mathbf{X}_{10,5} = \{x_{8,5}^1,...,x_{8,5}^S\}$ and $\mathbf{X}_{10,6} = \{x_{8,6}^1,...,x_{8,6}^S\}$. As you can observe, S number of multisite data values are derived from the same historical year. The repetition of the same multisite pattern will occur in the generated data set from this summary statistic formation. Instead, a procedure to blend the bootstrapped multisite data might be preferred with preserving the cross-correlation between the sites. Here, we propose a

scheme to blend the multisite data with the conceptual Genetic algorithm (GA). The fundamental and the process of GA are expressed briefly below.

Genetic algorithm is a search technique based on a biological metaphor, such as natural selection and mutation (Goldberg, 1989). GA explores the whole range of the feasible region and evolves toward a better solution with a probability manner. A better solution implies the maximization or minimization of the specified fitness function. This technique is an efficient and robust search process, since it produces a near-optimal solution through the traveling around all possible regions. The GA application needs to encode each parameter or target variable as an array of bits (binary code), called strings. The initialization performs establishing the starting searching points with a certain number of populations. A fitness function is required to evaluate the preference of each population. For a simple example, suppose that the feasible region of the target variable, denoted as Z, is from zero to sixty three with only integer values, and we try to find a value that maximize the fitness function $f(Z) = Z^2$. The example of GA for this problem is expressed in Table 3.1 and Figure 3.2 with four population sizes. The steps are as follows:

(1) The string length should be six because six strings of binary can be decoded as the range of zero to sixty three, such as from 000000 decoded as 0 and 111111 decoded as 63; ($\sum_{i=1}^{6} 2^{6-i} \times b_i$, where b_i is the binary values at each bit i). The strings of the initialized four populations are displayed in the second column of Table 3.1. The probability to be reproduced, the fifth column of Table 3.1, is estimated from

dividing the fitness value of each population (the fourth column) by the summation of the total fitness value (sixth row of the fourth column).

(2) Among the current populations, two populations are independently selected from the estimated probability. Notice that the selected two populations can possibly be the same.

(3) The two selected populations (population 2 and population 4 in Figure 3.2, left) mate and generate two new populations with the crossover of some elements (Figure 3.2, left) or without the crossover corresponding to the crossover probability, p_c. If a uniform random number, $u_c \sim Unif[0,1]$, is less than p_c, two selected populations crossover their values from the first elements to the cross point. On the left side of Figure 3.2, the crossover point is three, three values are crossovered, and generate two new populations. Otherwise, explicitly $u \geq p_c$, two selected populations, itself, become new populations. The cross point is assigned randomly from one to five in this example. Goldberg (1989) suggested the appropriate probability as $p_c = 0.6$. Further process can be applied at this stage, called tournament selection. Instead selecting both cross-over populations, only one of them is chosen in the favor of a certain criterion (i.e. fitness function).

(4) Each string of the new population mutates randomly as shown in Figure 3.2 with a mutation probability (p_m). The second element is mutated. Goldberg (1989) suggested $p_m = 1/30$.

Furthermore, the real-coded GA has been used such that real numbers are used for bits instead of binary (zero or one) code (Dasgupta and Michalewicz, 1997). And tournament selection in which two individuals compete for selection, only one remaining has been employed with better performance. The capability of GA to explore the whole feasible region of a variable with crossover and mutation is profitable for the MBB method with summary statistics since the method suffers from the repetition of the same pattern between sites as mentioned previously. The GA algorithm, however, should be modified to embed in this method because the purpose of GA here is to yield different generated data combinations between sites without loosing the cross relationship between sites. The embedded GA with modification is explained.

The reproduction process of GA is analogous to bootstrapping in that the data are resampled from the existing data $\mathbf{X}_{t,\tau} = \{X^s_{t,\tau}\}_{s\in[1,S]}$. Let the resampled data from MBB be the reproduced data from GA. The next procedure of GA is the crossover. For this process, it entails one more multisite data set denoted $\mathbf{X}^*_{t,\tau}$, and this additional multisite data should be similar to $\mathbf{X}_{t,\tau}$, avoiding the decrease of the cross-relation of the generated data. For this purpose, KNNR is employed to find the neighbors of $\mathbf{X}_{t,\tau}$. Subsequently, the portion of the original resampled data set $\mathbf{X}_{t,\tau}$ is substituted randomly with the values of another data set of $\mathbf{X}^*_{t,\tau}$ shown in Figure 3-A.1 with the crossover probability p_c. This probability is rather downsized as 0.3333 in application because of the cross-relation preservation. If the multisite data set $\mathbf{X}_{t,\tau}$ is probabilistically rejected to perform the crossover, each element (each site value) is replaced with another historical value among

$\{x^s_{v,\tau}\}_{v\in[1,N]}$ randomly with the mutation probability p_m illustrated in Figure 3-A.2. The replacement might be constrained to k- nearest neighbors of $X^s_{t,\tau}$ in order to preserve the cross-correlation.

The differences between the rudimentary GA and the modified are: (1) only one multisite-data set is obtained instead of two data sets (i.e. another data set $\mathbf{X}^*_{t,\tau}$ is discarded after the crossover); (2) instead of cross-point exchange, the exchange data points are selected randomly with probability 0.5; (3) assign the probability p_c as one-third rather smaller than what Goldberg (1989) suggested as 0.6; (4) a mutation process is performed only on the multisite generated data set for which the crossover is probabilistically refused. The difference between (3) and (4) is applied because higher probability for crossover and higher chance of mutation might incur lessening the cross-relation between sites. The exact procedure is described in the next section, with a simple example as shown in Appendix A.

The serious shortcoming of the simulated data from the suggested process above is that it generates only historical value. Short term water resources planning will be significantly affected from this deficiency. This might lead to failure of the analysis for the most significant drought with short duration. Lee and Salas (2008) suggested generating the unprecedented value utilizing the Gamma kernel density estimate.

3.3.3 Perturbation process with Gamma Kernel

One of the major drawbacks for bootstrapping is that the generated sequences are historical values. In literature, a few of the methods have strived to solve this problem on

nonparametric generation models (Prairie et al., 2006; Srinivas and Srinivasan, 2006; Sharif and Burn, 2007). The hybrid method, devised by Srinivas and Srinivasan (2006), with the low-order periodic autoregressive and moving block bootstrapping on the innovation term, is not plausible in the case of the mixture with non-intermittent site and intermittent site. The normal kernel density, derived by the bandwidth from the approach of Sharma et al. (1997), with KNN on weather variable can be a good candidate (Sharif and Burn 2007). However it employs the normal variable kernel which yields some bias on the marginal distribution in case that the observed data is significantly skewed and bounded. This might be a plausible approach when the record range of the variable includes the negative part such as the intervening flow (Lee and Salas, 2006).

Meanwhile, Lee and Salas (2008) suggested that the Gamma kernel density estimate with KNNR and Gamma kernel does not produce any bias in case of positive bounded data range. It can be applied into the MBB without any hassles. The vital point is described succinctly. The Gamma kernel suggested in the paper is:

$$K_{x^2/h^2,h^2/x}(t) = \frac{t^{x^2/h^2-1}e^{-t/(h^2/x)}}{(h^2/x)^{x^2/h^2}\Gamma(x^2/h^2)} \tag{3-4}$$

where $K_{\alpha,\beta}(t)$ is the gamma kernel function with shape parameter α and scale parameter β . The mean and variance from the gamma kernel are $\mu(t) = x$, $\sigma^2(t) = h^2$ respectively. The heuristic parameter estimation of the bandwidth suggested in Lee and Salas (2008) is employed here such that:

$$h = \frac{\sigma_x}{\sqrt{N/2}} \tag{3-5}$$

The application of Gamma KDE is that the resampled data (x) from the procedure described in the previous section is perturbed with Gamma kernel by replacing the data with a generated value from the gamma distribution such as $Gam(x^2/h^2, h/x^2)$ only if $x > 0$. Notice that Gamma kernel perturbation does not vary the resampled zero values.

The perturbation with the Gamma kernel with mean x and variance h^2 might not be appropriate in case of the highly skewed data since the variance of the kernel is fixed. The coefficient of variance (CV) of the kernel (h/x) is too large in case of low value x. The high variance on low value will yield frequent extreme low values. Moreover, the lofty value x case has relatively low CV. This often occurs in highly skewed data. In general, the intermittent streamflow are significantly skewed. Here, additional Gamma kernel only with different parameter formulation is proposed as:

$$K_{h,x/h}(t) = \frac{t^{h-1} e^{-t/(x/h)}}{(x/h)^h \Gamma(h)} \tag{3-6}$$

where $\mu(t) = x$ and $\sigma^2(t) = x^2/h$. Notice that the variance of Gamma kernel is varied along with x. And the expected value of the variance of the Gamma kernel is

$$E[\sigma^2(t)] = E[x^2/h] = \frac{\sigma_x^2 + \mu_x^2}{h} \tag{3-7}$$

For estimation of the bandwidth, a heuristic approach is suggested here with the similar quantity as Eq.(3-5), whereby the variance of the gamma kernel in Eq.(3-4) is the same as the mean variance of the gamma kernel as $E[\sigma^2(t)] = h^2 = \frac{\sigma_x^2}{N/4}$. Then,

$$E[\sigma^2(t)] = \frac{\sigma_x^2 + \mu_x^2}{h} = \frac{\sigma_x{}^2}{N/4} \tag{3-8}$$

The smoothing parameter h is

$$h = \frac{N}{4} \cdot \frac{\sigma_x^2 + \mu_x^2}{\sigma_x{}^2} \tag{3-9}$$

If the coefficient of variance (σ / μ) is close to one, then Eq.(3-7) is simplified as:

$$h = \frac{N}{2} \tag{3-10}$$

3.4 Applied Model Procedure

In Section 2, the employed model components are described. With the rudimentary matched block bootstrapping, the Genetic Algorithm and KNNR matched block process with the block length variable are included in order to attain the plausible diverse combination of the seasonal data without lessening the serial and cross relations. And the Gamma KDE is utilized to synthesize the unprecedented values from the MBB generated values. In addition, the way to manipulate the interannual variability is suggested by means of the pilot variable. The model components are formulated into one procedure described below.

Description of the applied Model Procedures

To implement the MBB with the suggested modification into multisite data, some preprocessing work is required. Instead of dealing with multisite data, Buishand and Brandsma (2001) proposed employing summary statistics over the different variables.

Those have been commonly employed in multivariate nonparametric modeling literature (Yates et al., 2003; Buishand and Brandsma, 2001). Here, the summary statistics are utilized to reduce the multivariate problem into the univariate one. From the outset, the multi-site seasonal streamflow data should be scaled as follows so that each of the data sites contributes equally on the summarize statistics:

$$y_{\nu,\tau}^{s} = \frac{x_{\nu,\tau}^{s}}{\hat{\mu}_{\tau}^{s}(x \mid x > 0)} \tag{3-11}$$

for the multi-site seasonal streamflow data on which part or all of sites are intermittent and $\hat{\mu}_{\tau}^{s}(x \mid x > 0)$ is the mean estimate from the observed data that is greater than zero. Eq.(3-11) is formatted to prevent scaling over zero values, or if the data consist of only non-intermittent data:

$$y_{\nu,\tau}^{s} = \frac{[x_{\nu,\tau}^{s} - \hat{\mu}_{\tau}^{s}(x)]}{\hat{\sigma}_{\tau}^{s}(x)} \tag{3-12}$$

This scaling, however, is avoidable if the historical multisite data are not significantly different from each site. In this case, $y_{\nu,\tau}^{s} = x_{\nu,\tau}^{s}$. After scaling the data with Eq.(3-11) or Eq.(3-12), the summary statistics are attained for each year and month such that:

$$\tilde{y}_{\nu,\tau} = \frac{1}{S}\sum_{s=1}^{S} y_{\nu,\tau}^{s} \tag{3-13}$$

With the summary statistics, the proposed model procedures follow.

A. Set the block length parameter of Poisson distribution (λ) allowing the mean block length ($E[l]$) around three to six in proportion to the magnitude of the temporal relation of monthly and yearly time scale. Obtain the current block length (l) from the random variable (l') with the Poisson distribution with the parameter (λ) and $l=l'+1$.

B. For the first generate block $\tilde{\mathbf{Y}}_{B(1)} = \{\tilde{Y}_{1,j}\}_{j\in[1,l]}$, the candidate blocks are the historical values constrained to $\{\tilde{\mathbf{y}}_{B(i)}\}_{i\in[1,N]}$ and $\tilde{\mathbf{y}}_{B(i)} = \{\tilde{y}_{B(i,j)}\}_{j\in[1,l]} = \{\tilde{y}_{i,j}\}_{j\in[1,l]}$, e.g. if l=4 and i=10, then $\tilde{\mathbf{y}}_{B(10)}$=$\{\tilde{y}_{10,1}, \tilde{y}_{10,2}, \tilde{y}_{10,3}, \tilde{y}_{10,4}\}$. Select a block randomly with equal probability $1/N$ among N (the number of record length) candidates. For the other generate block and the current generate month τ , the potential generate blocks are restricted to $\{\tilde{\mathbf{y}}_{B(i)}\}_{i\in[1,N]}$ and the elements of each block $\tilde{\mathbf{y}}_{B(i)}$ are $\{\tilde{y}_{B(i,j)}\}_{j\in[1,l]} = \{\tilde{y}_{i,j}\}_{j\in[\tau,\tau+l-1]}$. For example, if i=6, τ=10, and l=5, then $\tilde{\mathbf{y}}_{B(6)}$ =$\{\tilde{y}_{6,10}, \tilde{y}_{6,11}, \tilde{y}_{6,12}, \tilde{y}_{7,1}, \tilde{y}_{7,2}\}$. Notice that if j>ω where ω is the number of seasons (e.g. ω=12 for monthly) then i=i+1 and j=j-ω. Among N blocks, one block is selected such that one of the first elements of the candidate blocks $\tilde{Y}_{t,\tau}$ is selected with the KNNR and the following elements are automatically selected. To do this,

(a) the distances are estimated between the previous generate data $\tilde{Y}_{t,\tau-1}$ and the previous value of the candidate historical blocks $\{\tilde{y}_{i,\tau-1}\}_{i\in[1,N]}$, expressed as:

$$r_i = \left| Y_{t,\tau-1} - \tilde{y}_{i,\tau-1} \right| \qquad (3\text{-}14)$$

where, i=1,…N;

(b) among the k nearest neighbors (i.e. the indexes from the first to the k smallest distances); choose one with the weight probability in Eq.(3-3); k is estimated with the heuristic choice $\sqrt{N}$;

(c) the subsequent block of the selected is assigned as the generated block. Notice that if the block length is always one, the KNNR matched block bootstrapping converges to the original KNNR method.

C. Retrieve and back-transform (according to Eq.(3-11) or (3-12)) $\widetilde{Y}_{t,\tau}$ into the original domain $\{X^s_{t,\tau}\}_{s\in[1,S]}$. It is facile to acquire $\{Y^s_{t,\tau}\}_{s\in[1,S]}$ from $\widetilde{Y}_{t,\tau}$ just by keeping track of the year of the historical data from which $\widetilde{Y}_{t,\tau}$ is originated.

D. The multisite generated data in the original domain $\{X^s_{t,\tau}\}_{s\in[1,S]}$ is blended with the Genetic algorithm for each month and year as follows:

(a) Set the probability of crossover p_c and mutation p_m ; Here 0.333 and 0.01 are used respectively as suggested by Goldberg (1989).

(b) Choose another set of multisite data for t and τ whose summary statistics are close to the one of $\{X^s_{t,\tau}\}_{s\in[1,S]}$ ($\widetilde{Y}_{t,\tau}$) with KNNR, and assign it as $\{X^{*s}_{t,\tau}\}_{s\in[1,S]}$. Here, k is estimated with $\sqrt{N}$. The closeness is defined with the absolute distance between $\widetilde{Y}_{t,\tau}$ and $\widetilde{y}_{\nu,\tau}$, where $\nu = 1,..N$. One from k nearest neighbors is found in Eq.(3-3).

(c) The elements of the two sets, $\{X^s_{t,\tau}\}_{s\in[1,S]}$ and $\{X^{*s}_{t,\tau}\}_{s\in[1,S]}$, are exchanged or not with the crossover process in Genetic Algorithm as follows. If a generated uniform random number (u_c) is smaller than p_c ($u_c < p_c$), alternate the elements of $\{X^s_{t,\tau}\}_{s\in[1,S]}$ into $\{X^{*s}_{t,\tau}\}_{s\in[1,S]}$. Whether each element will be altered or not is decided randomly with p_c. As shown in Figure 3-A.3, tournament selection can be employed to select one of the two exchange data. With interchanging the values of $\{X^s_{t,\tau}\}_{s\in[1,S]}$ and $\{X^{*s}_{t,\tau}\}_{s\in[1,S]}$ and ending up with two sequences, the favorable one can be selected. The measurement of the preference is varied. One possibility is to choose the one that is yielding higher positive temporal crosscorrelation in case an applied algorithm underestimates the serial correlation through months. A drawback of tournament selection with the criterion of the higher serial correlation might be the underestimation of the mean in highly skewed data because the values of extreme cases tend not to be selected in highly skewed data. Therefore, the original sequence $\{X^s_{t,\tau}\}_{s\in[1,S]}$ with the crossover from the other set $\{X^{*s}_{t,\tau}\}_{s\in[1,S]}$ and the following mutation for the elements $\{X^s_{t,\tau}\}_{s\in[1,S]}$ is employed, but without employing the tournament selection.

(c) If $u_c \geq p_c$, mutate the $\{X^s_{t,\tau}\}_{s\in[1,S]}$. The mutation is performed independently for each element (s=1,...S). One of the main objectives, in time series modeling, is to preserve the temporal dependence structure such as lagged cross-correlation. Therefore, instead of applying the mutation for all elements

(s=1,…S) with p_m, only the data values not to be crossovered from the previous Step (c) are mutated with the probability p_m. From a generated random number u_m, if $u_m < p_m$, then substitute the current element $X^s_{t,\tau}$ with one of the historical candidates $\{x^s_{v,\tau}\}_{v\in[1,N]}$. Instead of randomly selecting any $x^s_{v,\tau}$ (where v=1,…,N), choose a value close to the generated value $X^s_{t,\tau}$ with KNNR where k (the number of nearest neighbors) is obtained from $N^{4/5}$ suggested by Fukunaga (1990) not from $N^{1/2}$ so that the candidates to be replaced have a wide range.

E. From repeating the steps above A to D, attain the generated data set with the target length T. The Gamma kernel perturbation is performed to the resampled and blended data with MBB and GA mixture. The process is independent on the other process and simply applied with substituting the resampled data with the gamma generated data with parameter $\alpha = (X^s_{t,\tau}/h)^2$ and $\beta = h^2/X^s_{t,\tau}$, or for highly skewed data, $\alpha = (h)^2$ and $\beta = \left(X^s_{t,\tau}\right)^2/h$.

3.5 Data Description and Test Statistics

To verify the suggested model, a portion of the Colorado River system was utilized. The Colorado River system (CRS) portrays the river flow with 29 selected stations. The historical gaged data has been naturalized for these 29 stations through 2003. Part of the data has been extended by Lee and Salas (2006) back to 1906, employing the

combination of the parametric linear regression and the nonparametric bootstrapping with a trace selection method.

In application, two sets of simulation studies were performed. Firstly, three non-intermittent sites, non-zero values in the data set, are selected among 29 stations of the CRS, such as sites 8, 16, and 20. Those are the most vital sites in managing the Colorado River system. The hybrid model (Srinivas and Srinivasan, 2005) and the moving block bootstrapping with genetic algorithm (GAMBB) model, developed in this paper, were applied to these sites. Since the hybrid model does not have the ability for intermittent data, three non-intermittent sites are selected to compare with the GAMBB model. Secondly, the combined sites with non-intermittent (Sites 21 and 24) and intermittent (Sites 22 and 27) were applied only into the GAMBB model from the reason above. The tributaries of the lower basin of the system include the arid and semi-arid region such as Nevada, Arizona, and New Mexico. The monthly streamflow of the tributaries, especially Site 22 and 27, has intermittency, defined as zero streamflow between the flows greater than zero. The exact location of those sites is displayed in Figure 3-A.4.

The one hundred set of the data set with the same length as the historical are generated for each experimented model. Several statistics are estimated from the historical and generated data to verify the model performance such as mean, standard deviation, skewness, maximum and minimum, and lag-1serial correlation in seasonal and yearly time scale. A boxplot is employed to show the estimated statistics from the generated data. The end line of the box implies the 25 and 75 percent quantile, while the cross line above the box on the whisker does the 90 percent quantile and maximum, below the box on the whisker 10 percent quantile and minimum. And the segment line

with the 'x' mark presents the historical values. The preservation of the cross or serial relation in the generated data is checked through comparing a scatterplot. Half of the generated data sets (50 set) were used, as well as the historical data. Furthermore, the drought statistics with the yearly historical and generated data were compared with the boxplot. The employed drought statistics are the maximum drought and surplus amount, the longest drought and surplus length, and storage capacity with the historic mean as demand level.

Moreover, multisite drought event statistics of the yearly data explained in Haltiner (1985) were calculated for the observed and generated data. The estimated statistics are mean run length (MERL), mean run-sum (MERS), max run length (MARL), and max run-sum (MARS), and storage capacity. Mean and max run length is the mean and maximum value of $l(i)$ defined as the length of deficit at the i^{th} drought event. Mean and max run-sum is the mean and maximum value of $S(i)$, that is the length of deficit at the i^{th} drought event where $S(i)$ is the summation of the deficit of all sites $S(i) = \sum_{k=1}^{S} S^k(i)$ for the i^{th} drought event. Storage capacity can be estimated through the deficits described in Figure 3-A.5 with sequence peak algorithm (Louks et al., 1981). A different threshold level is considered for the water demand D_0^k for site k and unvaried through time. D_0^k is defined as the historic mean of site k ; multiplying threshold level (TL) ranged from 0.7 to 1.0 with 0.05 interval.

3.6 Results

As it is mentioned, two sets of simulation tests were performed in order to verify the suggested GAMBB model. First, non-intermittent sites are applied to the hybrid and GAMBB model. Second, the combined sites are applied only to the GAMBB model. The results of each application are explained in the following two subsections.

Before full application, to observe the effect of the Genetic Algorithm, MBB with KNNR matched block and variable block length ($E[l]=12$) with GA and without GA were tested. Gamma KDE perturbation is not employed in this experiment. The one hundred set of the same record length multisite ([8, 16, and 20]) CRS data were simulated. The scatterplot between Moth8 of Site 8 and Site 20 is presented in Figure 3.3 for the model without GA (left) and Figure 3.3 for the model with GA(right). The significant difference can be monitored between two figures. It is obvious that the generated data without GA rarely simulate the new combination between sites. This implies that the multisite KNN models (Buishand and Brandsma, 2001; Yates et al., 2003) in literature will show the same feature and cannot produce the new combination of the generated data between sties. This behavior might be undesirable in that a data simulation model is built in order to explore any possible events that are unprecedented from the observed data.

3.6.1 Model comparison for non-intermittent case

For the first data set (non-intermittent sites: 8, 16 and 20 of CRS), the hybrid model (Srinivas and Srinivasan, 2005) and GAMBB model were applied. The hybrid model is the combining model with the lag-1 PAR and the bootstrapping of innovation as

it is explained in the review section. The applied block length for the innovation is 24 months (two year period). For the GAMBB model, 12 months (1 year period) of the expected length of E[l] is employed. Site 20 data are presented as the representative result site if a statistic is separately estimated. Completed results can be found in Appendix C.

The basic monthly statistics of site 20 are displayed in Figure 3.4 for the hybrid model and Figure 3.5 for the GAMBB model. Those statistics are shown from Figure 3-A.6 to Figure 3-A.15. Most of the monthly statistics such as mean, standard deviation, skewness, and lag-1 correlation are well preserved in both models. However, minimum and maximum are not preserved in the hybrid model. In detail, the generated maximum can not be higher than the historical maximum especially in the later months of the seasons (i.e. after February). And the generated minimum can not be lower than the historical, especially right after February. To investigate the behavior of the hybrid model, the scatterplots are illustrated in Figure 3.6. In Figure 3.6, the historical data points for site 8 for month 2 and month 3 are presented with triangles and 50 sets of the simulated data from (a) the hybrid model and (b) the GAMBB model with gray circles and the same plot but with the month 8 and month 9 for (c) the hybrid model and (d) GAMBB model. The linear directional shape is shown in Figure 3.6 (a), which is the general characteristic in the hybrid model. Similar behavior is also shown in the local regression with KNN innovations developed by Prairie et al. (2007) and further investigated by Lee and Salas (2008). The synthetic data in this case are only generated from the directional lines. The GAMBB model, however, aptly reproduces the historical relation with local non-linearity (Figure 3.6 (b)). More importantly, the reasons of the underestimation of the minimum and the overestimation of the maximum are revealed here in Figure 3.6 (c). The generated

data from the hybrid model are not much different from the historical data in the later part of the season, here Month 8, as shown in Figure 3.6 (c). The reason is because the fixed block innovation of the hybrid model fits the monthly data into lag-1 PAR model such that:

$$X^s_{\nu,\tau} = \phi^s_{1,\tau} X^s_{\nu,\tau-1} + \varepsilon^s_{\nu,\tau}$$

where $\phi^s_{1,\tau}$ and $\varepsilon^s_{\nu,\tau}$ is the parameter and random component at month τ and site *s*, respectively, and the stored random components $\varepsilon^s_{\nu,\tau}$ are resampled at the generation procedure. Since the innovation is resampled with a two-year block, the whole months of the innovation in a year are generated from the same year of the historical data. The only difference of $X^s_{\nu,\tau}$ in generated data can be achieved from $X^s_{\nu,\tau-1}$ because $\varepsilon^s_{\nu,\tau}$ is taken from the resampling of the stored innovation. However, the synthetic data is recursively generated with the previous value; the difference from the historical data will be diminished along with the later portion of the months. In case of the higher $\phi^s_{1,\tau}$ value that is the lag-1 serial correlation in method of moment parameter estimation, the difference of the generated data from $X^s_{\nu,\tau}$, the historical value, might propagate further down to the later months of the season. To the extent of the extreme case of $\phi^s_{1,\tau}$ is zero, the generated data is no different from the historical data through the next months of the year. Figure 3.6 (d) shows that the GAMBB model properly preserves the locality and dispersion as the historical.

The monthly cross correlations are well preserved at both models, as shown at Figure 3.7 (left) and (right), respectively. Lag-1 cross-correlations are also well preserved, as shown Figure 3-A.10 and Figure 3-A.11. The yearly key statistics for both models in

Figure 3.8 (Hybrid) and Figure 3.9 (GAMBB), are well reproduced in both models, except lag-1 correlation. Even if the block length of the hybrid model is considered as a two-year period to preserve the interannual variability, there is still some underestimation on the lag-1 correlation (Figure 3.8). Meanwhile, the lag-1 correlation for the GAMBB model is better preserved, though the mean block length is half of the hybrid model (Figure 3.9). This is the effect of the KNN matched block selection. The underestimation of the lag-1 yearly serial correlation in the hybrid model affects the slight underestimation of the storage capacity in Figure 3.10(a). The drought, surplus, and storage statistics are slightly better preserved by the GAMBB model than the hybrid model (Figure 3.10 (a) and Figure 3.10 (b) for site 20). The other stations (site 8 and 16) behave the same as site 20. A reader refers to the figures from Figure 3-A.16 to Figure 3-A.19. The multisite monthly and yearly drought statistics with different threshold levels were estimated and presented in Figure 3-A.20 to Figure 3-A.25 for both models. The only storage capacity at 1.0 TL had some underestimation in the generated data from the hybrid model.

Furthermore, negative values rarely occurred in the generated data from the hybrid model. It might not be significant in this case since the frequency of being negative is very small. However in case of the highly skewed data, this might be a serious drawback in that the streamflow data cannot be physically negative. Cross-correlation pairs of historical and generated data are shown in Figure 3-A.26 to Figure 3-A.31.

3.6.2 Application to the combined sites with intermittent and non-intermittent

For the multisite model application, the intermittent sites 22 and 27 were applied, as well as the non-intermittent sites 21 and 24. The portion of the results sites 21 and 22 (one for non-intermittent site and one for non-intermittent site) are presented here. Completed results can be obtained in Appendix C. The specification of the applied GAMBB model was the moving block bootstrapping model with variable block length E[*l*]=12. KNN matched block selection is also used in the model procedure. The applied Gamma KDE formulation is shown in Eq.(3-6), and Eq.(3-9) was employed for smoothing parameter estimation since the applied dataset is highly skewed. For the GA algorithm, tournament selection was applied, such that one with higher lag-1 correlation was selected with the probability 0.8.

The basic monthly statistics displayed in Figure 3.11 and Figure 3.12 show that the model reproduced those statistics reasonably well for Sites 21 and 22. For Sites 24 and 27, refer to Figure 3-A.32 and Figure 3-A.33. All the basic statistics are fairly well preserved through the GAMBB model for both sites. Also, the statistics of Sites 24 and 27 are preserved well. The monthly minimum of Site 22 was always zero at each month in the historical and almost in simulated data except for a few cases in August and September of the generated data since the site is intermittent for all months, including zero values (Figure 3.12). This indicates that the model reproduced the intermittency in the historical data well. The lag-0 cross correlation was well preserved as illustrated in Figure 3.13, as well as lag-1 cross correlation (referred to Figure 3-A.34). The pair cross-correlations, the correlation between a pair of monthly or annual data, were estimated and shown for Site 21 and 22 in Figure 3.14 (a) and (b), respectively. Most of all pair

correlations were well preserved as shown. The yearly statistics of the two sites were fairly preserved with some minor bias, as well as the other sites. Related results are shown from Figure 3-A.35 to Figure 3-A.38. The yearly drought statistics for each site displayed in Figure 3.15 for Sites 21 and 22 were well preserved with the GAMBB model. Further results for the yearly drought statistics are shown in Figure 3-A.39 and Figure 3-A.40. Figure 3-A.41 and Figure 3-A.42 illustrate the pair correlation. The multisite drought statistics were also well preserved as shown in Figure 3.16. Figure 3-A.43 to Figure 3-A.48 shows the further multisite monthly and yearly drought, surplus, and storage statistics. The yearly cross-correlation in Table 3.1 was reproduced well in the generated data from the GAMBB model

3.7 Summary and Conclusions

In this paper, we made an effort to build the stochastic simulation model of the multivariate seasonal streamflow time series with the combination of intermittent and non-intermittent sites. So far, there is not much development for this study except in the generation model of the multivariate weather variable. The nonparametric technique, the moving block bootstrapping procedure, was employed for the suggested model in this paper. To this end, we developed some new features in order to yield more variable sequences, since one of the critical drawbacks for the nonparametric generation model is to generate only the same value as the historical, the repetition of the same seasonal pattern, and no variation spatially (the values are exactly the same as the historical site-by-site). The new features were: (1) the variable block length – the aggregated values to annual or seasonal (in case of monthly generation) will be different from historical, (2)

KNNR block selection – the connection between blocks will be preserved, and (3) Genetic Algorithm mixture – spatially different sequences to historical will be generated, and (4) Gamma KDE perturbation – unprecedented values from historical will be generated. Overall, the developed model was built in order to generate as many unprecedented sequences as possible while preserving the statistical behaviors embedded in the observed data, such as key basic statistics and drought statistics.

In application, the suggested model was compared with the hybrid model at first with the non-intermittent case since the hybrid model does not have the adoptability of the intermittency. The hybrid model has undesirable features, such as the directional relation in the generated data and the generation of the almost the same sequences as the historical, especially during the later part of the seasons. The suggested model, GAMBB, reproduces the basic and drought statistics that are estimated with various synthetic data sets that are unprecedented in the observed data.

In case of the combination cases, the GAMBB model reproduces well the statistical features of the observed data such as the basic key statistics and drought, surplus, and storage statistics. It suggests that the developed model might be an attractive model for the combined case of the intermittency and non-intermittency in a reasonable manner.

Table 3.1 Initialized Population for GA with six strings and four populations

No	String	Z	f(Z)=Z^2	% of Total
1	011010	26	676	9.90
2	110100	52	2704	39.6
3	101000	40	1600	23.4
4	101011	43	1849	27.1
Total			6829	100

Table 3.2 Cross-correlation of Historical and Generated Yearly Streamflow

His	21	22	24	27
21	1.00	0.51	0.60	0.59
22	0.51	1.00	0.63	0.62
24	0.60	0.63	1.00	0.68
27	0.59	0.62	0.68	1.00
GAMBB	21	22	24	27
21	1.00	0.53	0.58	0.55
22	0.53	1.00	0.68	0.65
24	0.58	0.68	1.00	0.70
27	0.55	0.65	0.70	1.00
GAMKNN	21	22	24	27
21	1.00	0.63	0.66	0.59
22	0.63	1.00	0.77	0.73
24	0.66	0.77	1.00	0.77
27	0.59	0.73	0.77	1.00

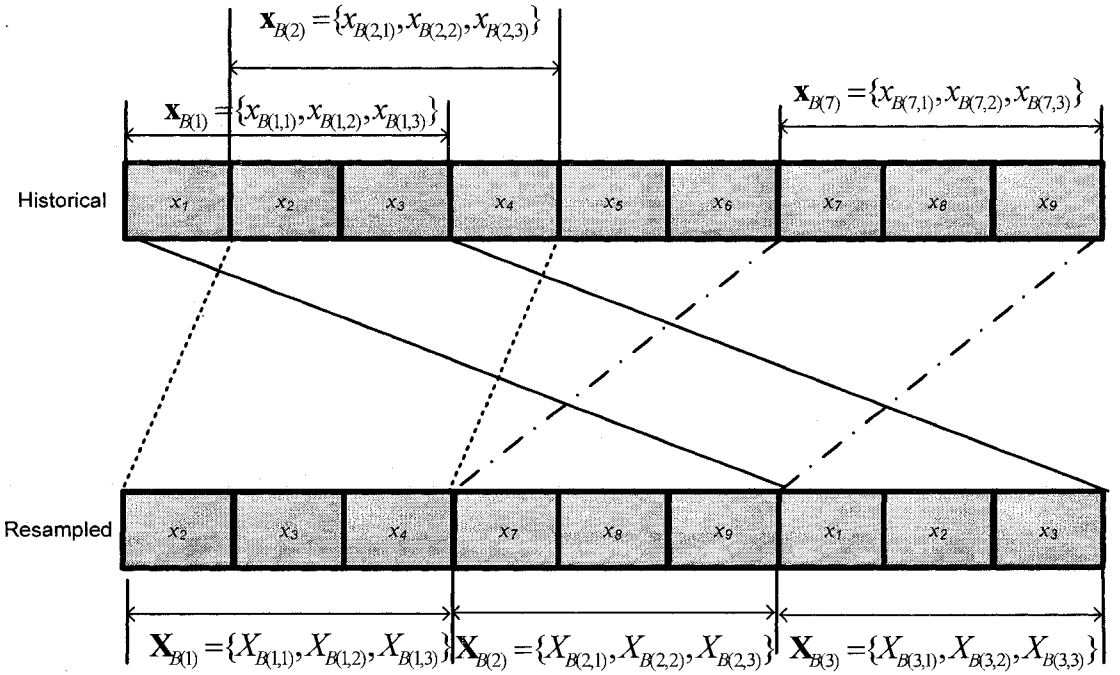
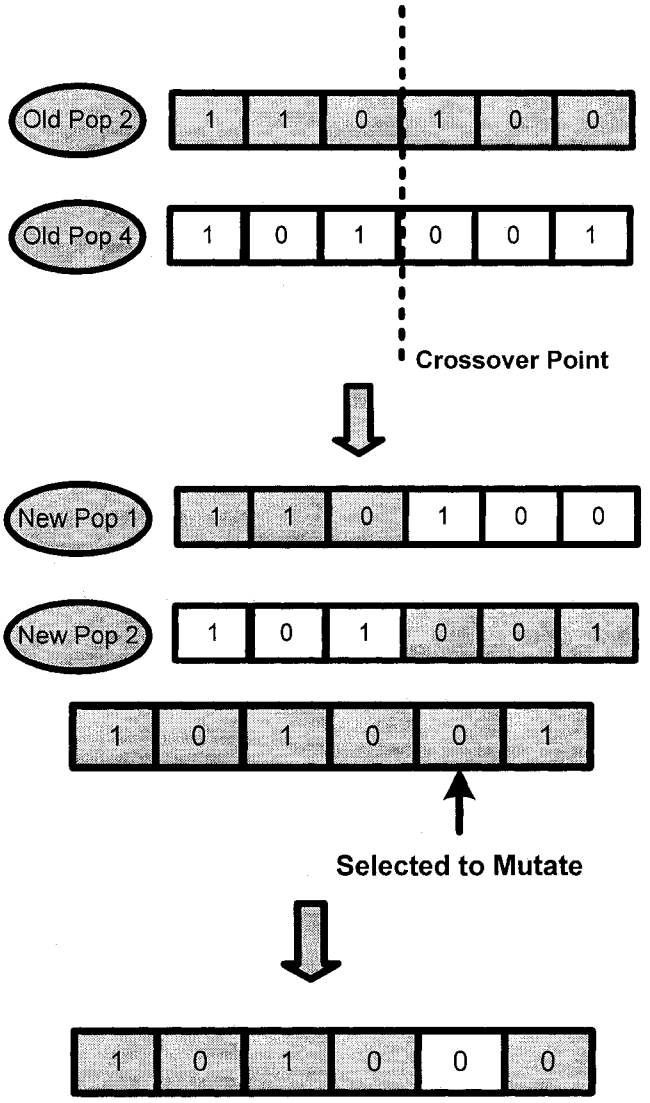


Figure 3.1 Schematic representation of Block Bootstrapping with overlapping; the number inside each box presents time index; $x_{B1,2}$ - the value of the first block and second element;



(a) (b)

Figure 3.2 Sketch of the crossover process (a) and the mutation process (b) for GA with six strings and four populations; (a) From the original population (Table 3-1), population 2 and population 4 is selected as reproduction and the items are exchanged from zero element to the crossover point. (b) from the new population, the second of the elements of the second new population is mutated

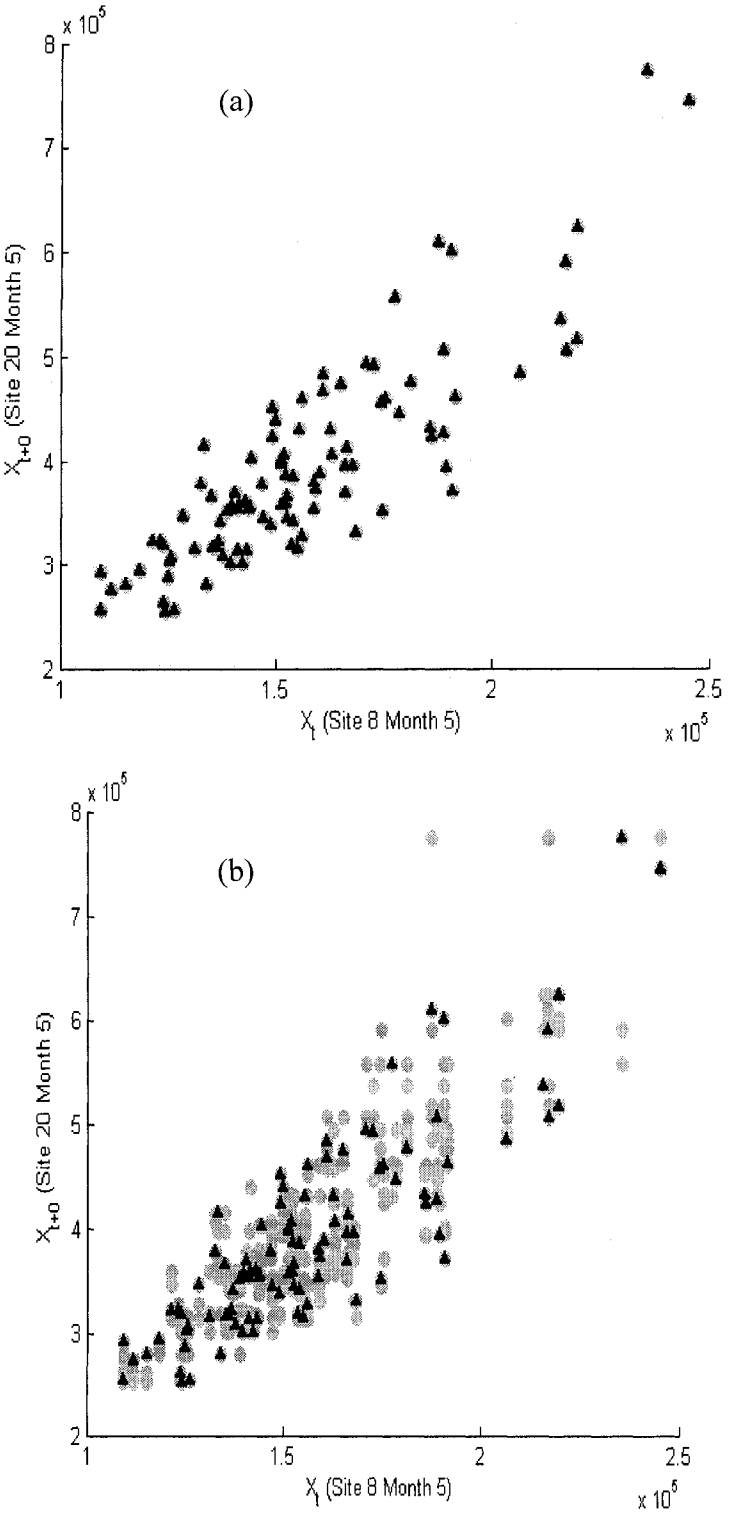


Figure 3.3 Scatterplot of one hundred generated data set (gray filled circle) with the same length as historical (triangle) from MBB without Genetic Algorithm (a) and with Genetic Algorithm (b), Site 8 (x-coordinate) and Site 20 (y-coordinate) for month 5

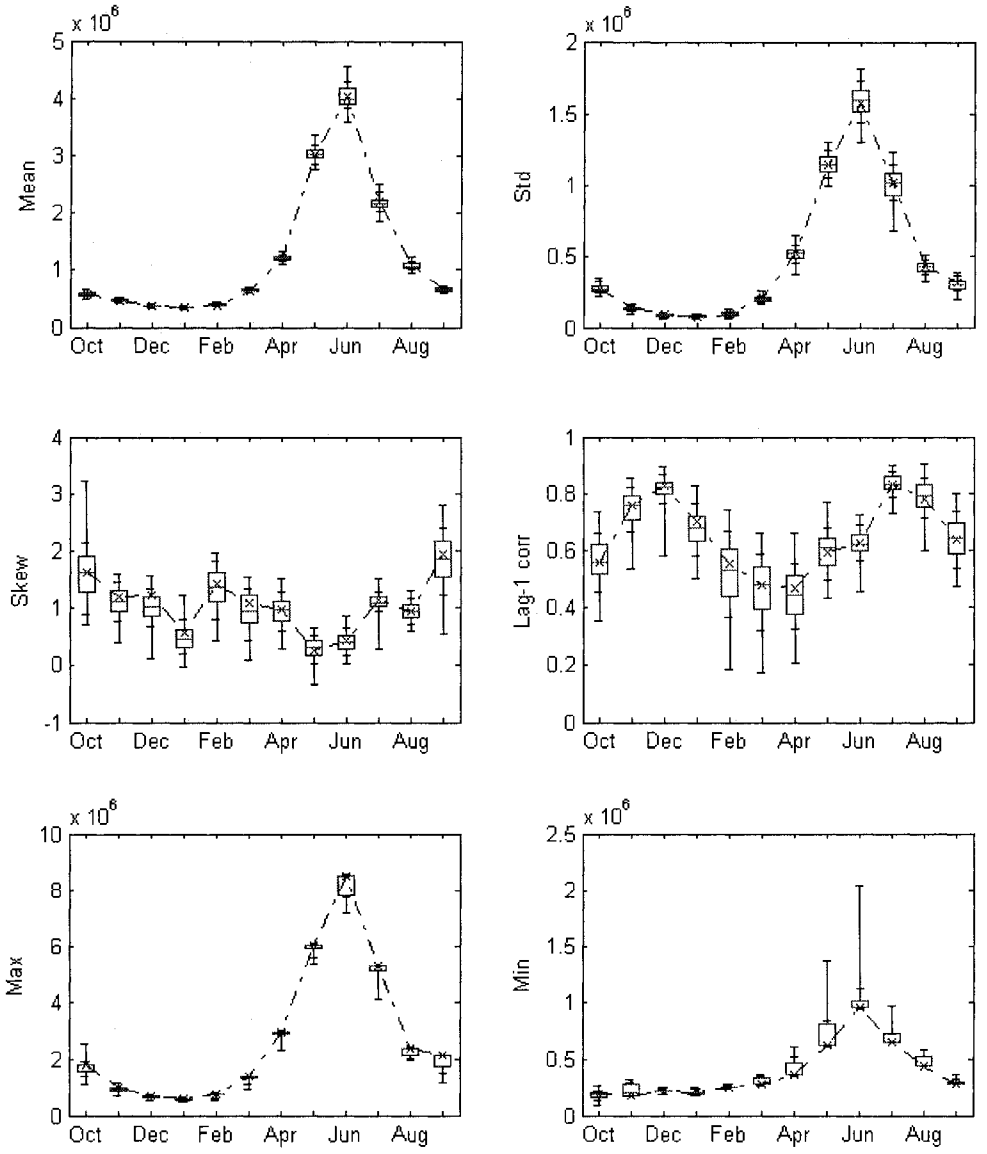


Figure 3.4 Key Statistics of Historical (dot line) and simulations (boxplot) with Hybrid for Site 20 of the Colorado River monthly streamflow Unit : Acre-feet

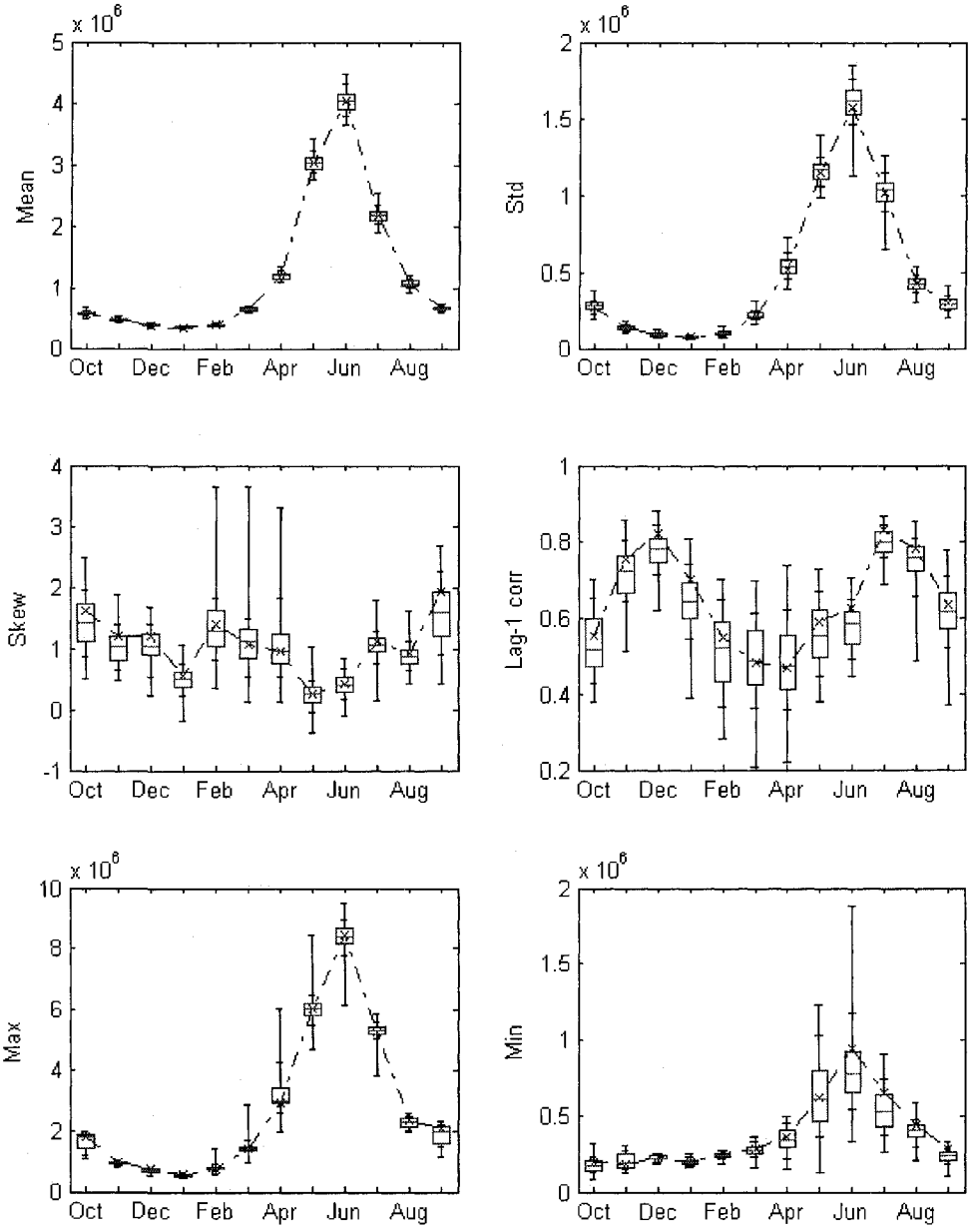


Figure 3.5 Key Statistics of Historical (dot line) and simulations (boxplot) with GAMBB for Site 20 of the Colorado River monthly streamflow Unit : Acre-feet

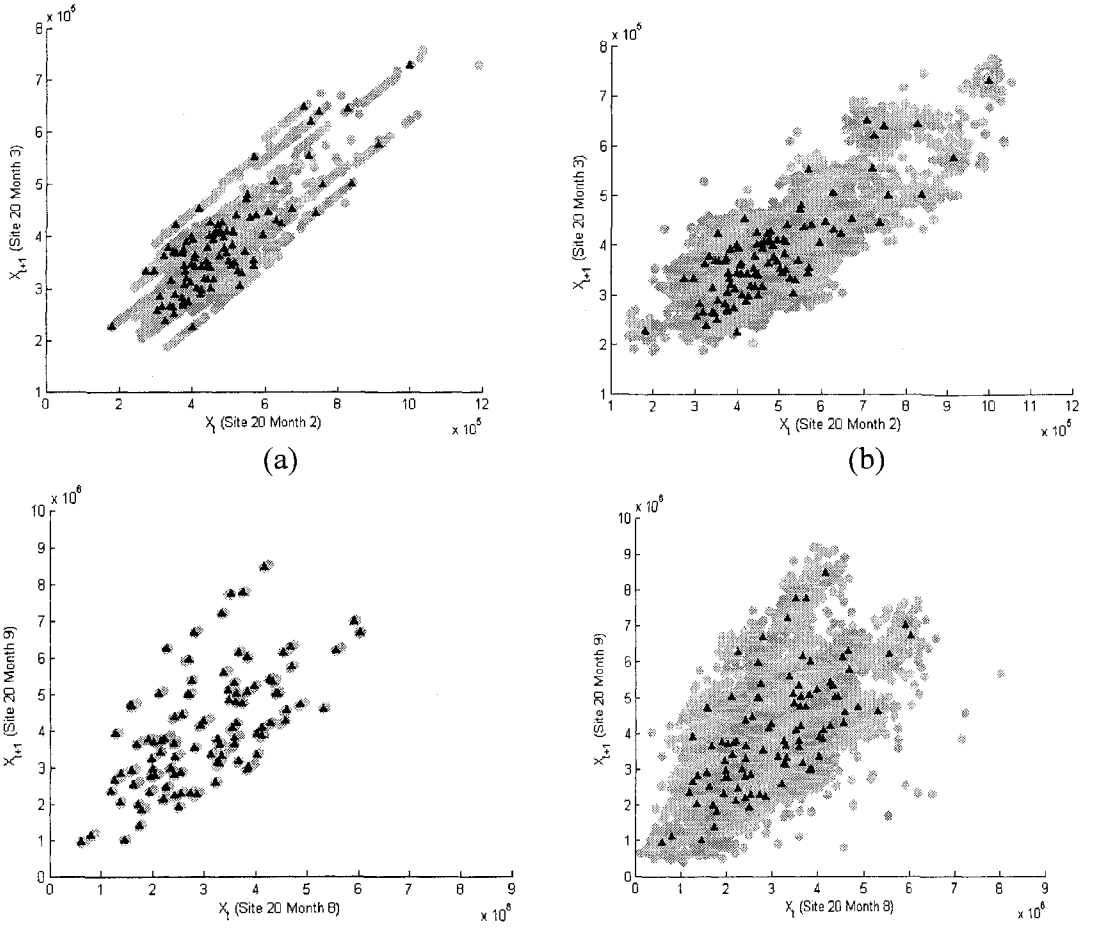


Figure 3.6 Scatterplot from Historical (triangle) and (a) Hybrid, (b) GAMBB simulations (gray filled circle) for month 2 and month3, (c) Hybrid, (d) GAMBB simulations for month 7 and month 8 (gray circle) of Colorado River site 20

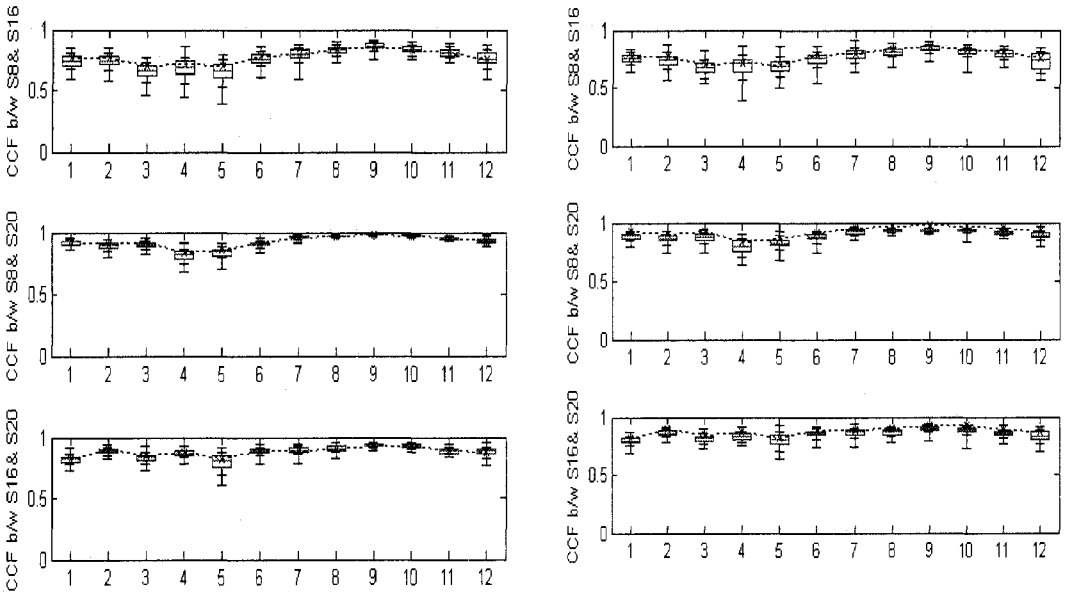


Figure 3.7 Lag-0 cross-correlation between sites from Historical (dot line) and simulations (boxplot) with Hybrid (left) and GAMBB(right) of the Colorado River monthly streamflow

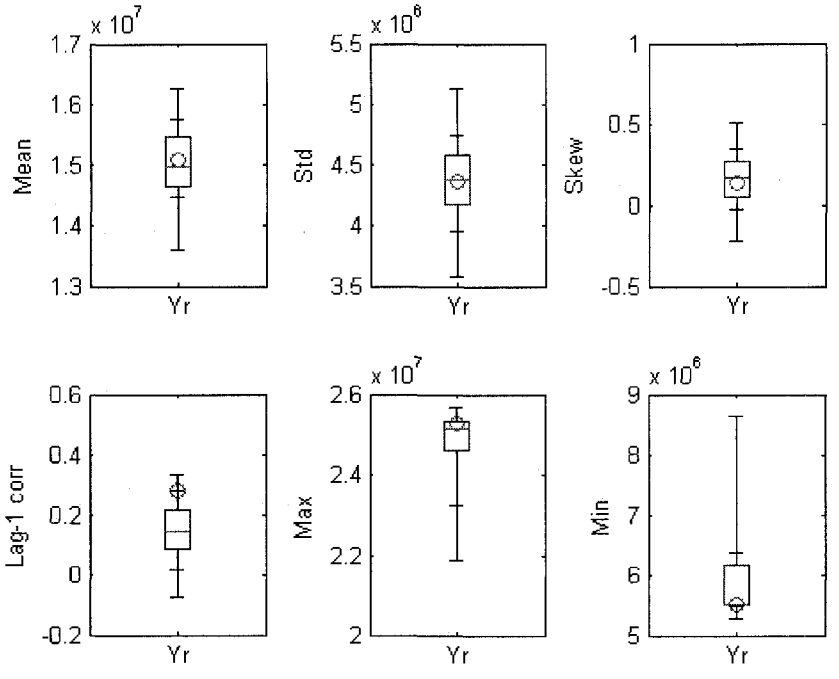


Figure 3.8 Key Statistics of Historical (dot line) and simulations (boxplot) with Hybrid for Site 20 of the Colorado River yearly streamflow Unit : Acre-feet

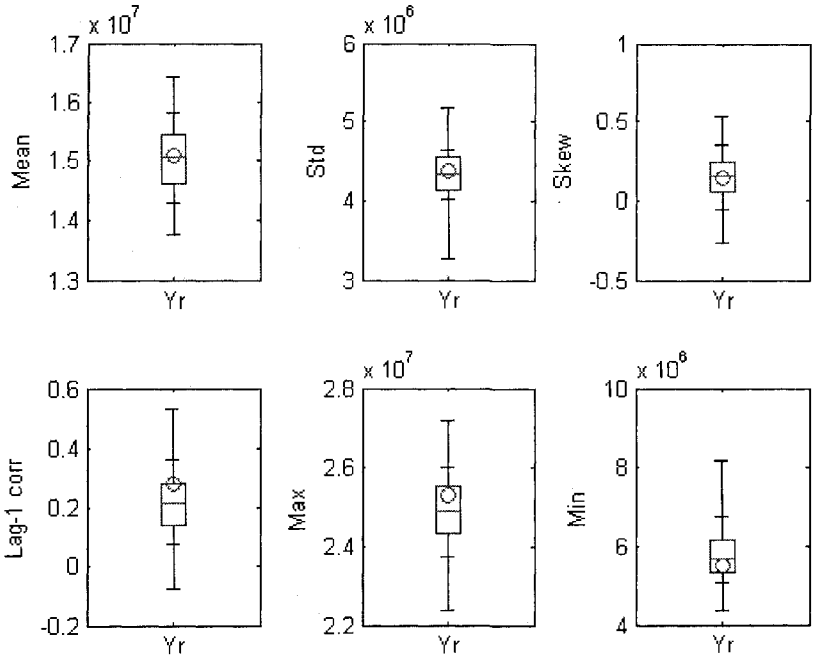
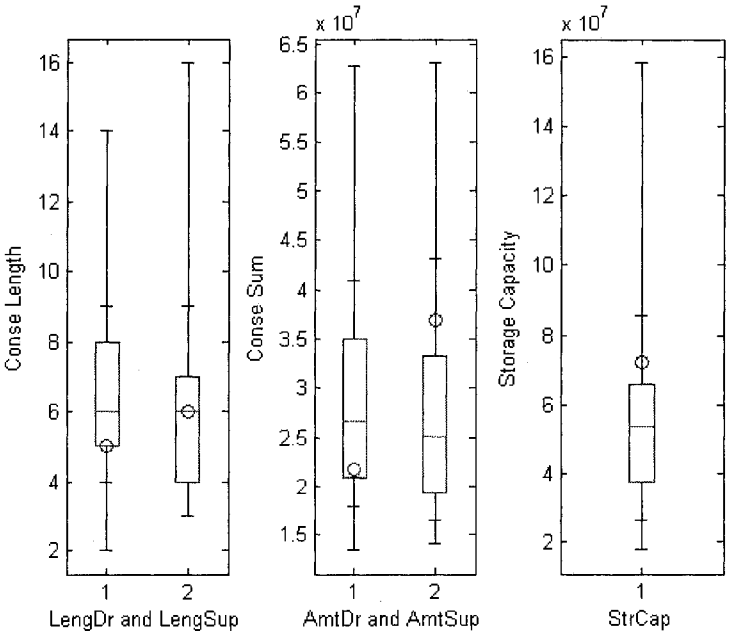
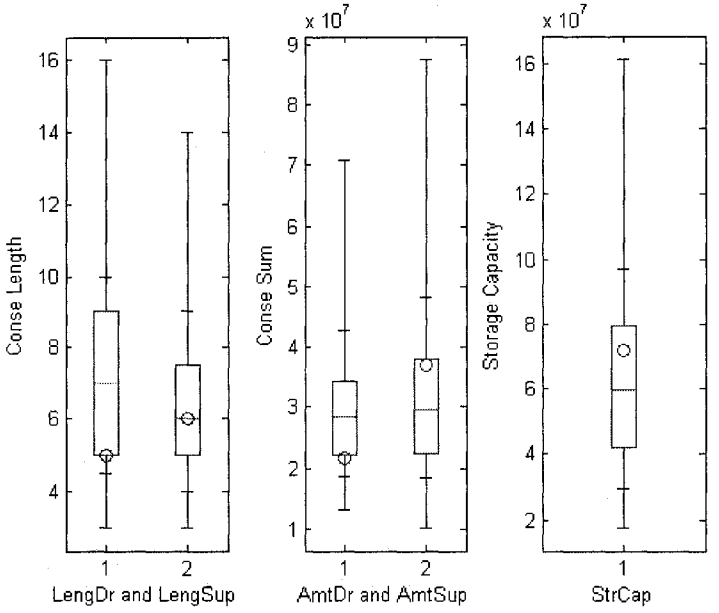


Figure 3.9 Key Statistics of Historical (dot line) and simulations (boxplot) with GAMBB for Site 20 of the Colorado River yearly streamflow Unit : Acre-feet



(a)



(b)

Figure 3.10. Reservoir-related statistics from Historical (dot line) and simulations (boxplot) with (a) Hybrid and (b) GAMBB for Site 20 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (Unit : Acre-feet)

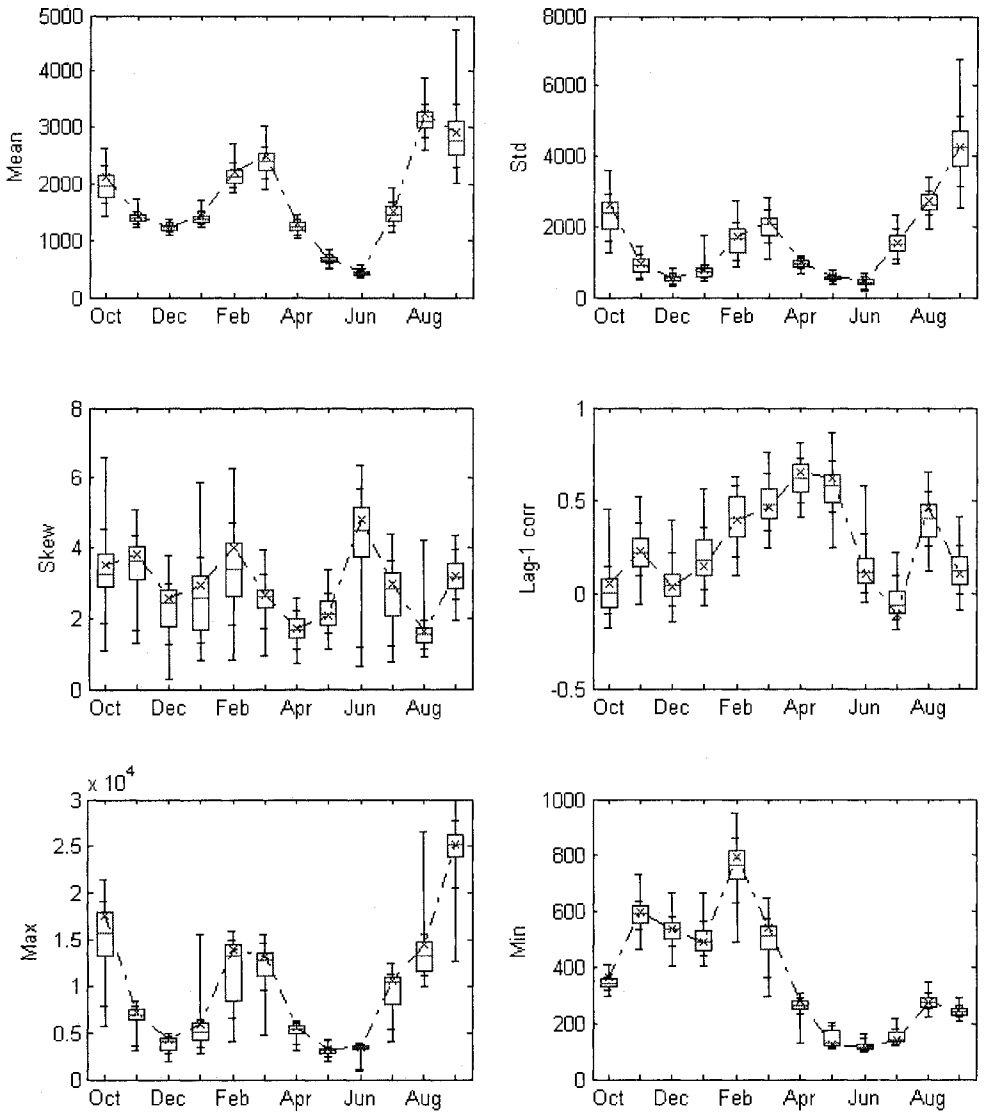


Figure 3.11 Key Statistics of Historical (dot line) and GAMBB simulations (boxplot) for Site 21 of the Colorado River monthly streamflow Unit : Acre-feet

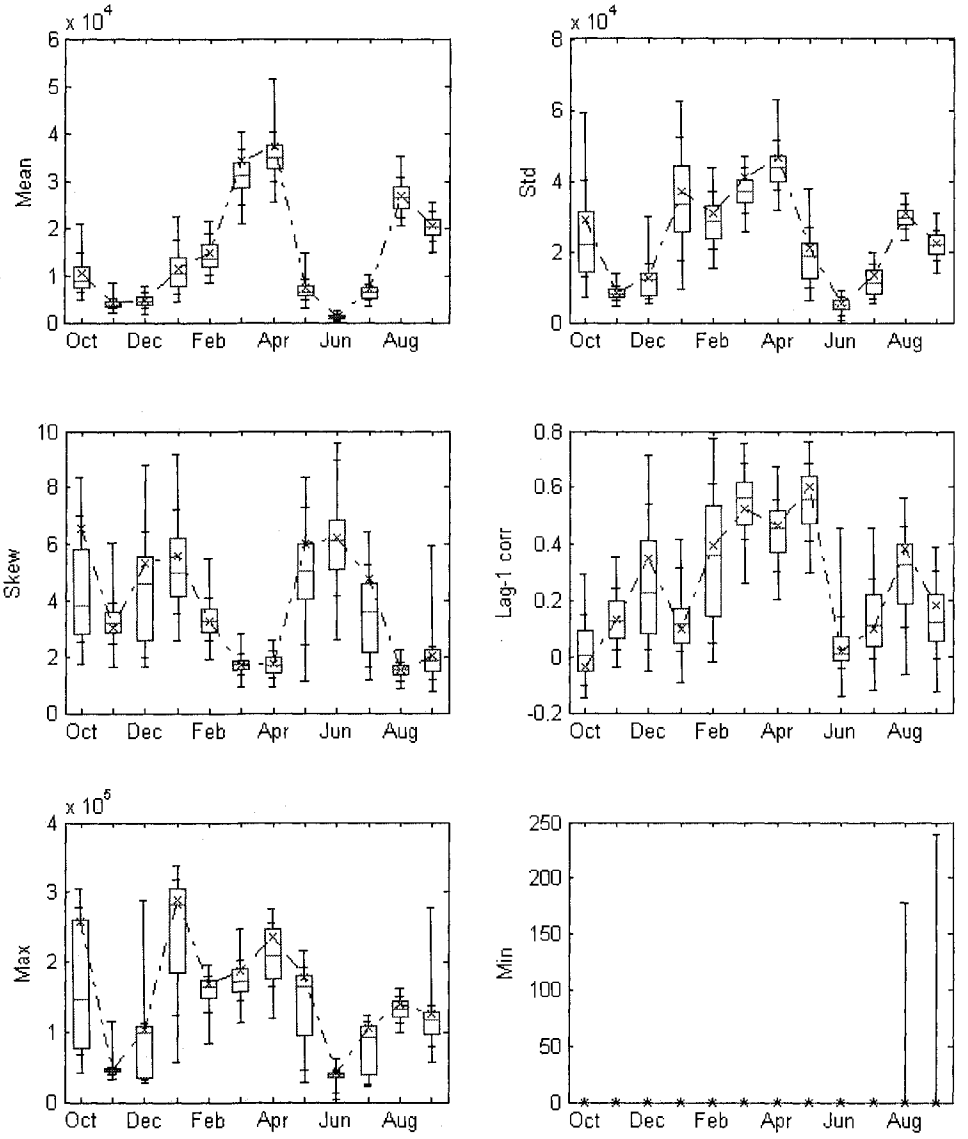


Figure 3.12 Key Statistics of Historical (dot line) and GAMBB simulations (boxplot) for Site 22 of the Colorado River monthly streamflow Unit : Acre-feet
With GAMBB

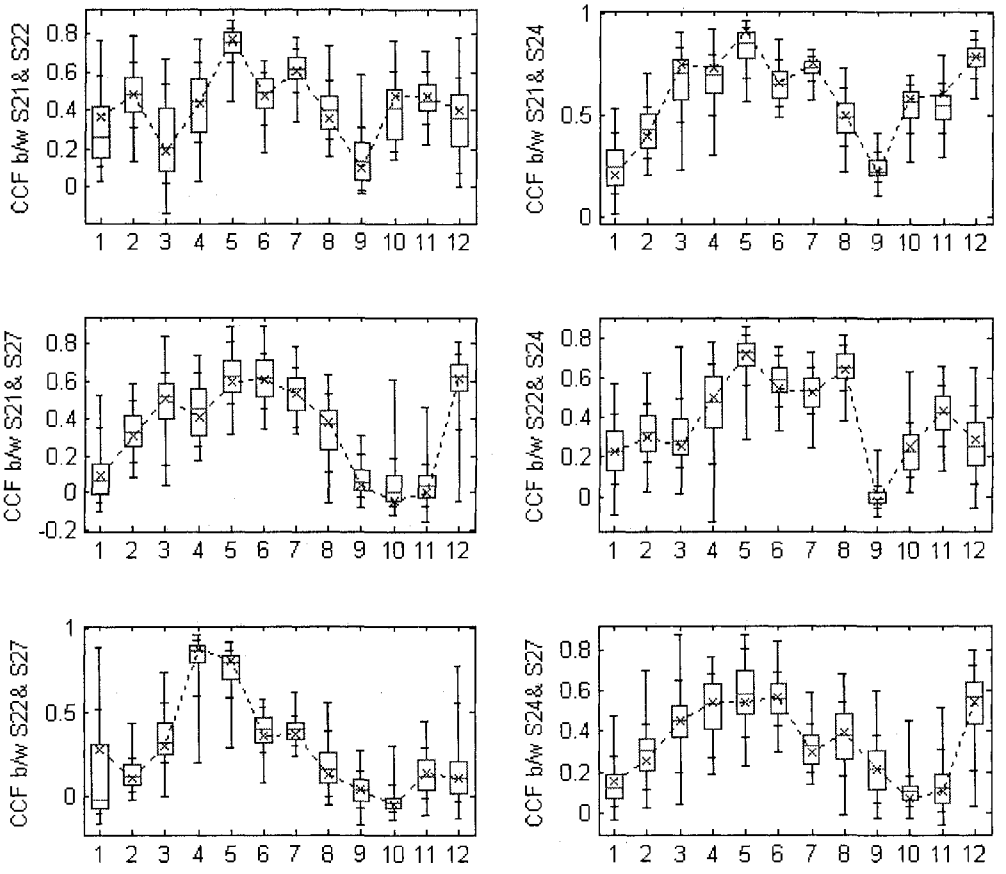
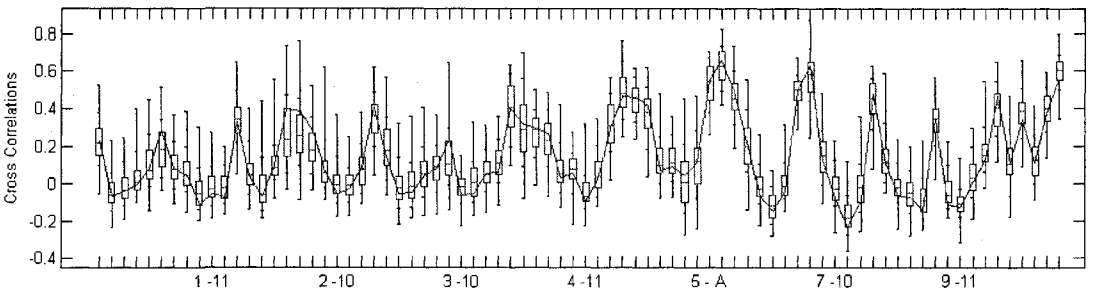
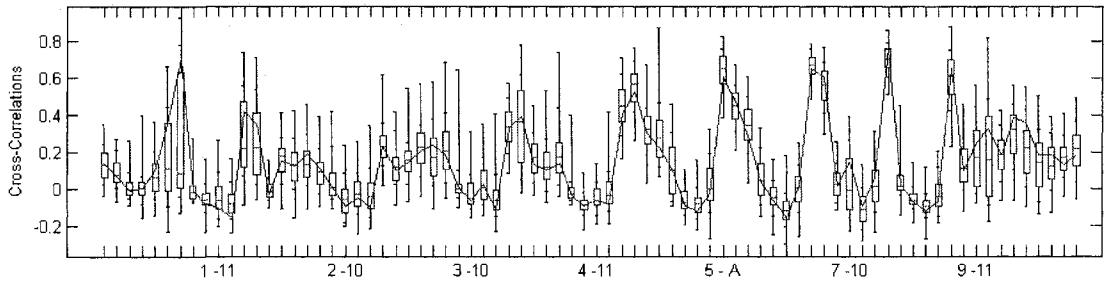


Figure 3.13 Lag-0 cross-correlation between sites from the historical (circle) and GAMBB simulations (boxplot) of the Colorado River monthly streamflow

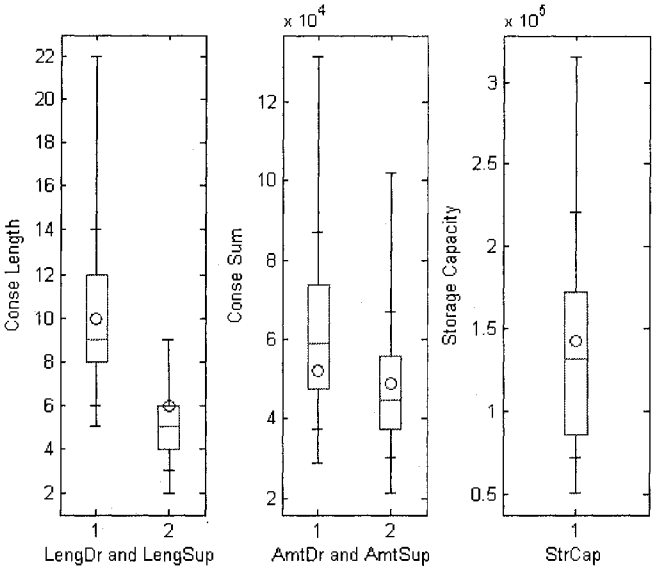


(a)

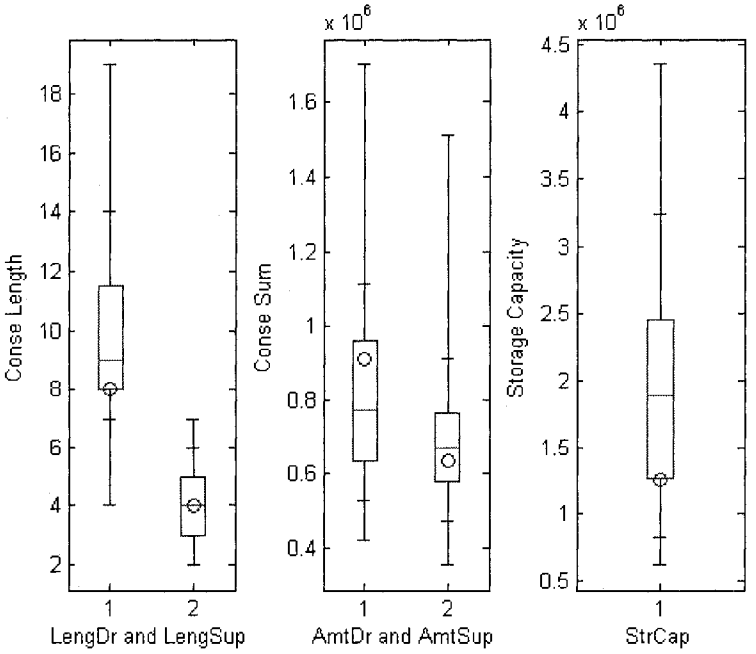


(b)

Figure 3.14 cross-correlation pairs of the historical and simulated data from GAMBB simulations (boxplot) and E[*l*]=*12* of (a) the site 21 and (b) the site 22 at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data



(a)



(b)

Figure 3.15 Reservoir-related statistics from historical (circle) and GAMBB simulations (boxplot) for (a) Site 21 and (b) Site 22 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (Unit : Acre-feet)

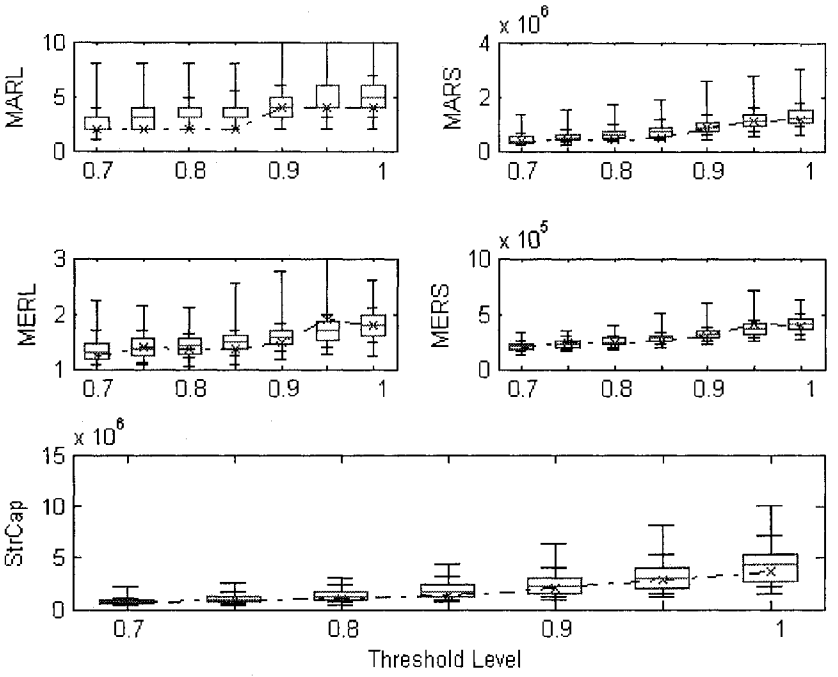


Figure 3.16 Multisite Yearly Drought Statistics of Historical (circle) and GAMBB simulations (boxplot) of the Colorado River streamflow

3.8 References

Beard, L. R. (1973), Transfer of streamflow data with Texas, 1-24 pp, Texas Water Development Board.

Bureau of Reclamation (2007). "Final Environmental Impact Statement Colorado River Interim Guidelines for Lower Basin Shortages and Coordinated Operations for Lakes Powell and Lake Mead", U. S. Department of the Interior, Bureau of Reclamation, Upper and Lower Colorado Regions.

Buishand, T. A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resources Research, 37*, 2761-2776.

Carlstein, E., et al. (1998), Matched-block bootstrap for dependent data, *Bernoulli, 4*, 305-328.

Chebaane, M., et al. (1995), Product Periodic Autoregressive Processes for Modeling Intermittent Monthly Streamflows, *Water Resources Research, 31*, 1513-1518.

Dasgupta, D., and Z. Michalewicz (1997), *Evolutionary Algorithms in Engineering Applications*, Springer Verlag.

Fukunaga, K. (1990), Introduction to Statistical Pattern Recognition, 2 ed., Academic Press.

Goldberg, D. E. (1989), *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley Pub. Co.

Haltiner, J. P. (1985), Stochastic Modeling of Seasonal And Daily Streamflow, Colorado State University, Fort Collins.

Kunsch, H. R. (1989), The Jackknife and the Bootstrap for General Stationary Observations, *Annals of Statistics, 17*, 1217-1241.

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

Lee, T., and J. D. Salas (2008), Periodic Stochastic Model for Simulating Intermittent Monthly Streamflows of the Colorado River System, paper presented at World Environmental & Water Resources Congress 2008, Honolulu, Hawaii.

Lee, T., and J. D. Salas (2008a), Seasonal Streamflow Simulation Techniques to preserve seasonal and yearly dependency, *(in review)*.

Loucks, D. P., Stedinger J.R., and Haith D.A. (1981), *Water Resources Systems Planning And Analysis*, Prentice-Hall.

Porter, J. W., and B. J. Pink (1991), A method of synthetic fragments for disaggregation in stochastic data generation, paper presented at Int. Hydrol. and Water Resour. Symp, The Institution of Engineers, Australia, Canberra.

Prairie, J., et al. (2007), A stochastic nonparametric technique for space-time disaggregation of streamflows, *Water Resources Research, 43*, -.

Prairie, J. R., et al. (2006), Modified K-NN model for stochastic streamflow simulation, *Journal of Hydrologic Engineering, 11*, 371-378.

Press, W., et al. (2002), *Numerical Recipes in C++*, 2 ed., Cambridge University Press.

Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other weather variables, *Water Resources Research, 35*, 3089-3101.

Sharif, M., and D. H. Burn (2006), Simulating climate change scenarios using an improved K-nearest neighbor model, *Journal of Hydrology, 325*, 179-196.

Sharif, M., and D. H. Burn (2007), Improved K-nearest neighbor weather generating model, Journal of Hydrologic Engineering, 12, 42-51.

Sharma, A., and O'Neill, R. (2002). "A nonparametric approach for representing interannual dependence in monthly streamflow sequences." *Water Resources Research*, 38(7), 5.1-5.10.

Sharma, A., Tarboton, D. G., and Lall, U. (1997). "Streamflow simulation: A nonparametric approach." *Water Resources Research*, 33(2), 291-308.

Srikanthan, R. (1979), Stochastic generation of annual and monthly flow volumes, Monash University, Clayton, Vic.(Australia).

Srikanthan, R., and T. A. Mcmahon (1980), Stochastic Generation of Monthly Flows for Ephemeral Streams, *Journal of Hydrology, 47*, 19-40.

Srinivas, V. V., and K. Srinivasan (2001), A hybrid stochastic model for multiseason streamflow simulation, *Water Resources Research, 37*, 2537-2549.

Srinivas, V. V., and K. Srinivasan (2005), Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows, *Journal of Hydrology, 302*, 307-330.

Srinivas, V. V., and K. Srinivasan (2006), Hybrid matched-block bootstrap for stochastic simulation of multiseason streamflows, *Journal of Hydrology, 329*, 1-15.

Stedinger, J. R., et al. (1985), A Condensed Disaggregation Model for Incorporating Parameter Uncertainty into Monthly Reservoir Simulations, *Water Resources Research, 21*, 665-675.

Svanidze, G. (1978), Mathematical models of streamflow for water power (water management) calculations, paper presented at proceedings second world congress, International Water Resources Association, New Delhi.

Tarboton, D. G., et al. (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resources Research, 34*, 107-119.

Valencia, D., and J. C. Schaake (1973), Disaggregation Processes in Stochastic Hydrology, *Water Resources Research, 9*, 580-585.

Vecchia, A. V., et al. (1983), Aggregation and Estimation for Low-Order Periodic Arma Models, *Water Resources Research, 19*, 1297-1306.

Vogel, R. M., and A. L. Shallcross (1996), The moving blocks bootstrap versus parametric time series models, *Water Resources Research, 32*, 1875-1882.

Yates, D., et al. (2003), A technique for generating regional climate scenarios using a nearest-neighbor algorithm, *Water Resources Research, 39*, -.

Appendix 3-A. Further Detailed Figures

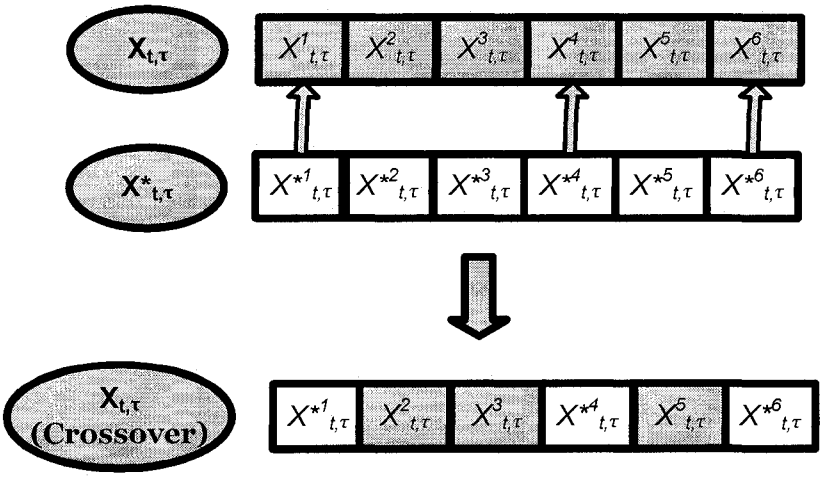


Figure 3-A.1 Sketch of the crossover process applied to MBB method; From the initial $\mathbf{X}_{t,\tau}$, the part of the values is exchanged with the separately selected multisite data set $\mathbf{X}^*_{t,\tau}$. Unlike the basic GA, only one set of data are selected as the generated.

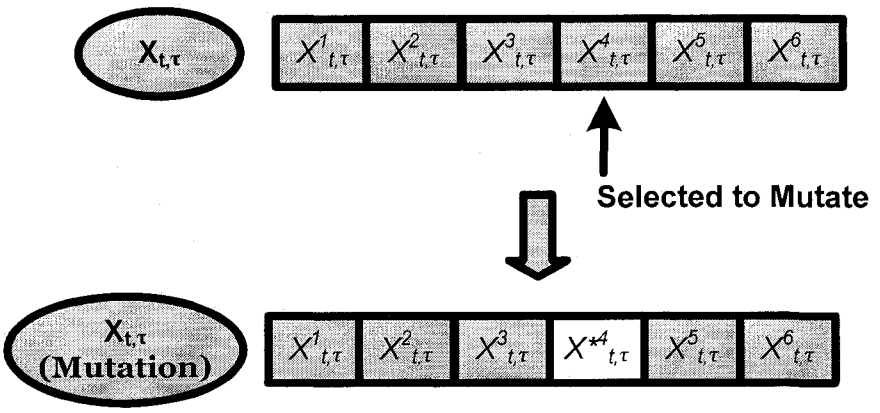


Figure 3-A.2 Sketch of the mutation process applied to MBB method. $X^{*4}_{t,\tau}$ is selected from the k-nearest historical values.

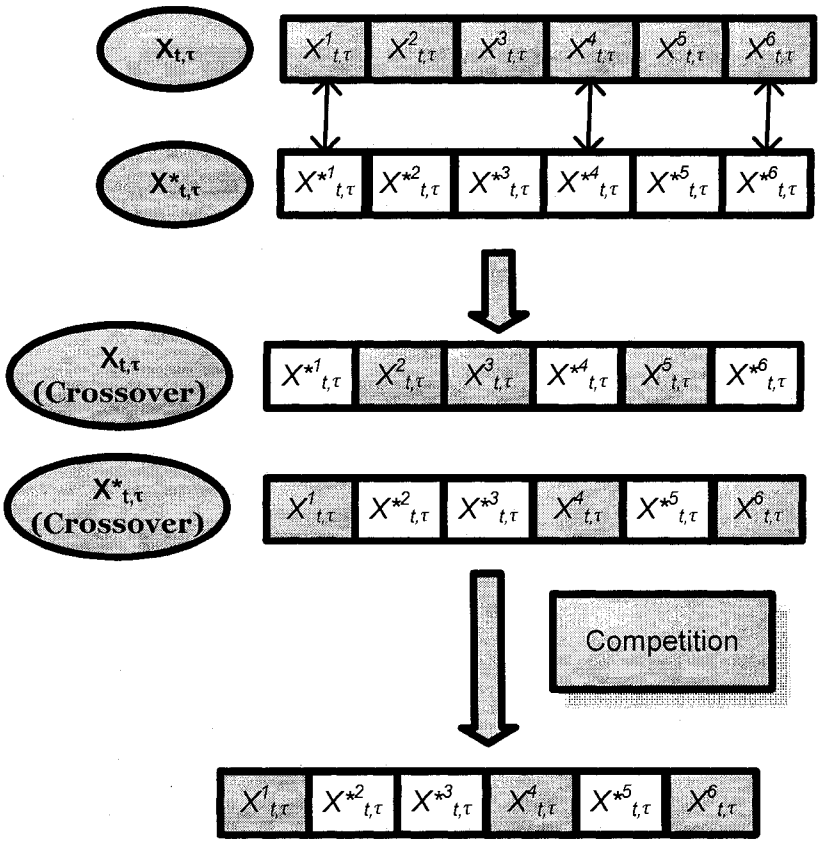


Figure 3-A.3 Sketch of the tournament selection with crossover process applied to MBB method

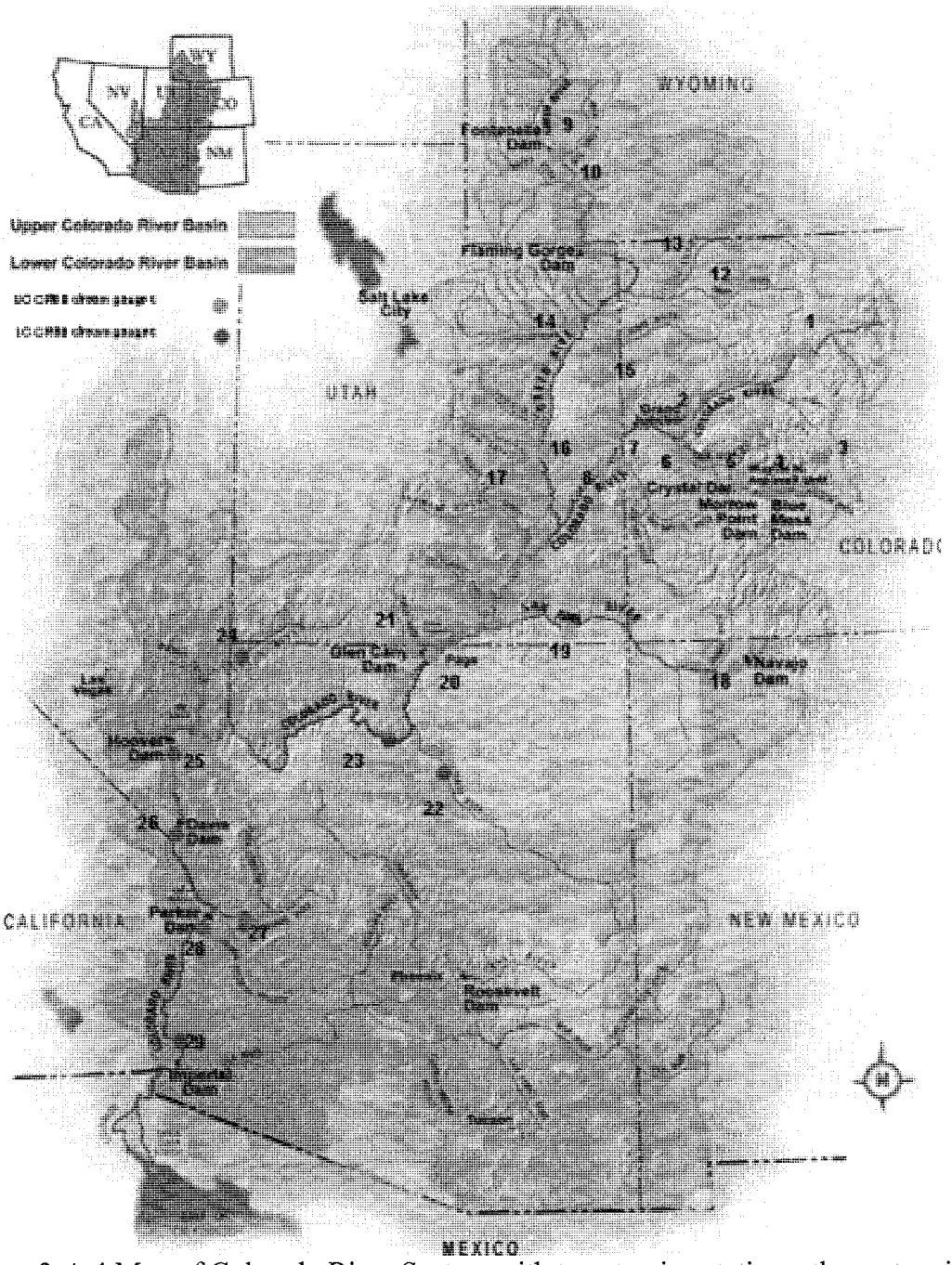


Figure 3-A.4 Map of Colorado River System with twenty nine stations; the system is divided into two as the upper Colorado River basin (1-21) and the lower Colorado River basin (22-29); the map is obtained from Bureau of Reclamation (2007)

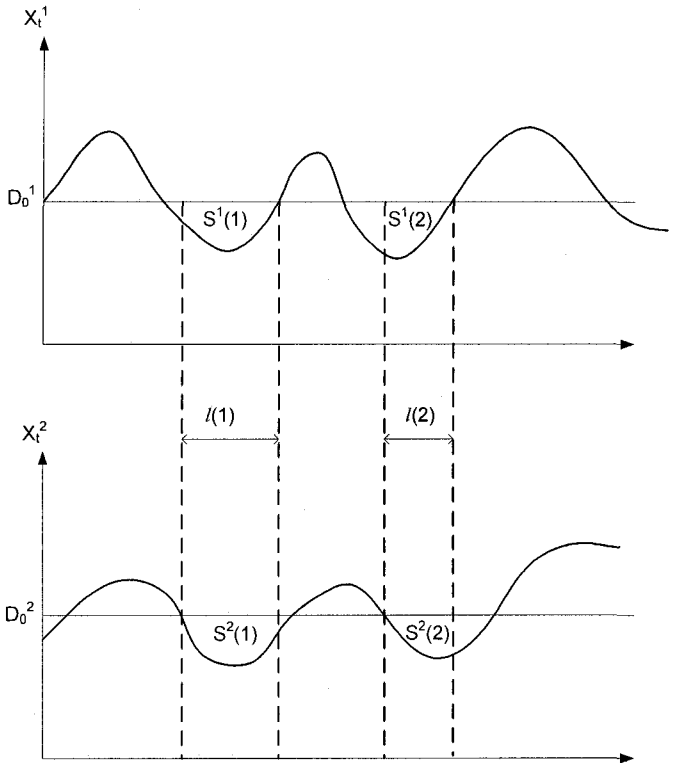


Figure 3-A.5 Graphical Representation of Multisite Drought Statistics : X_t^k presents the time series for site k and time t, D_0^k is the water demand for site k and unvaried through time. This quantity is defined as mean of X_t^k multiplying threshold level (TL) ranged as 0.7 to 1.0 with 0.05 interval. And $S^k(i)$ is the amount of deficit at k^{th} site and the i^{th} drought event and $l(i)$ is the length of the deficit the i^{th} drought event.

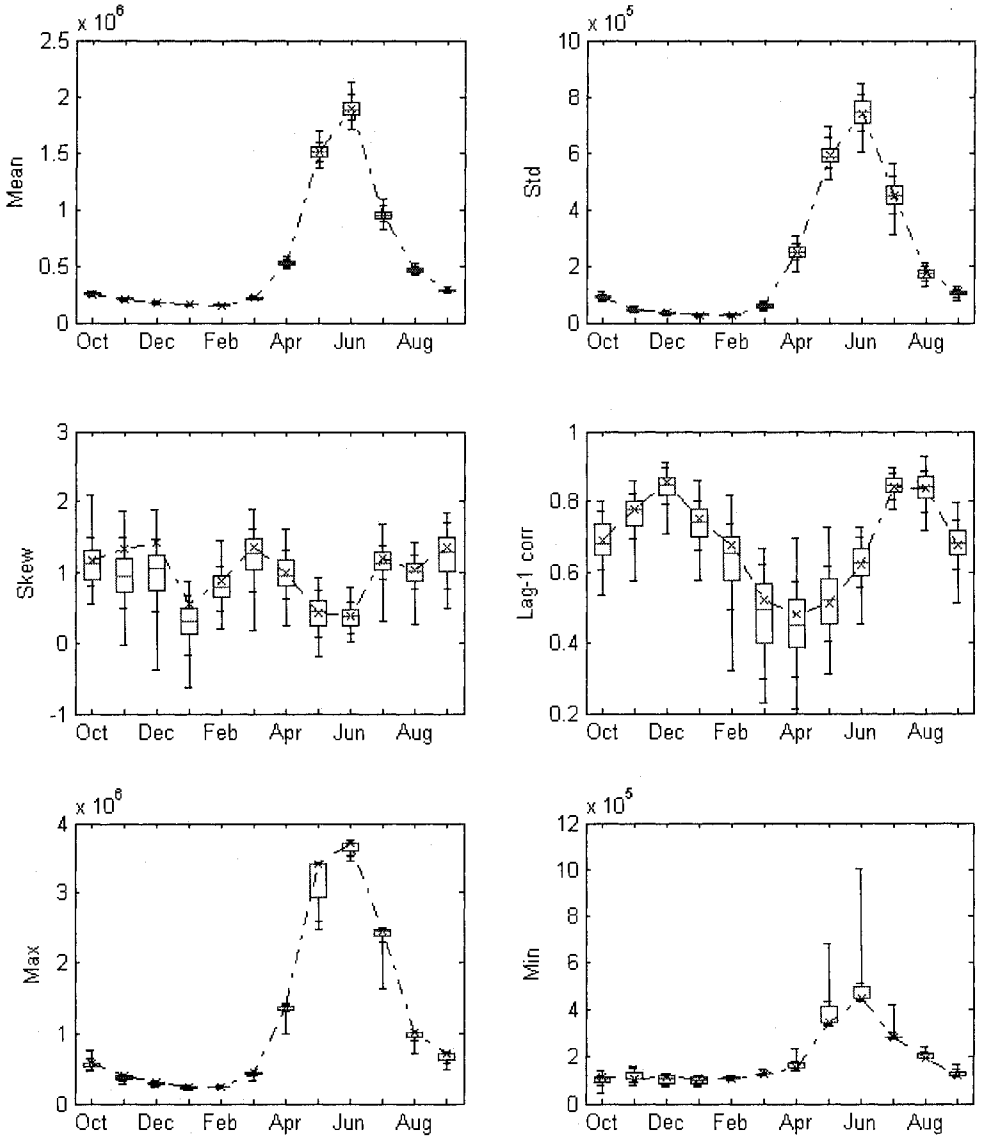


Figure 3-A.6 Key Statistics of Historical (dot line) and simulations (boxplot) with Hybrid for Site 8 of the Colorado River monthly streamflow Unit : Acre-feet

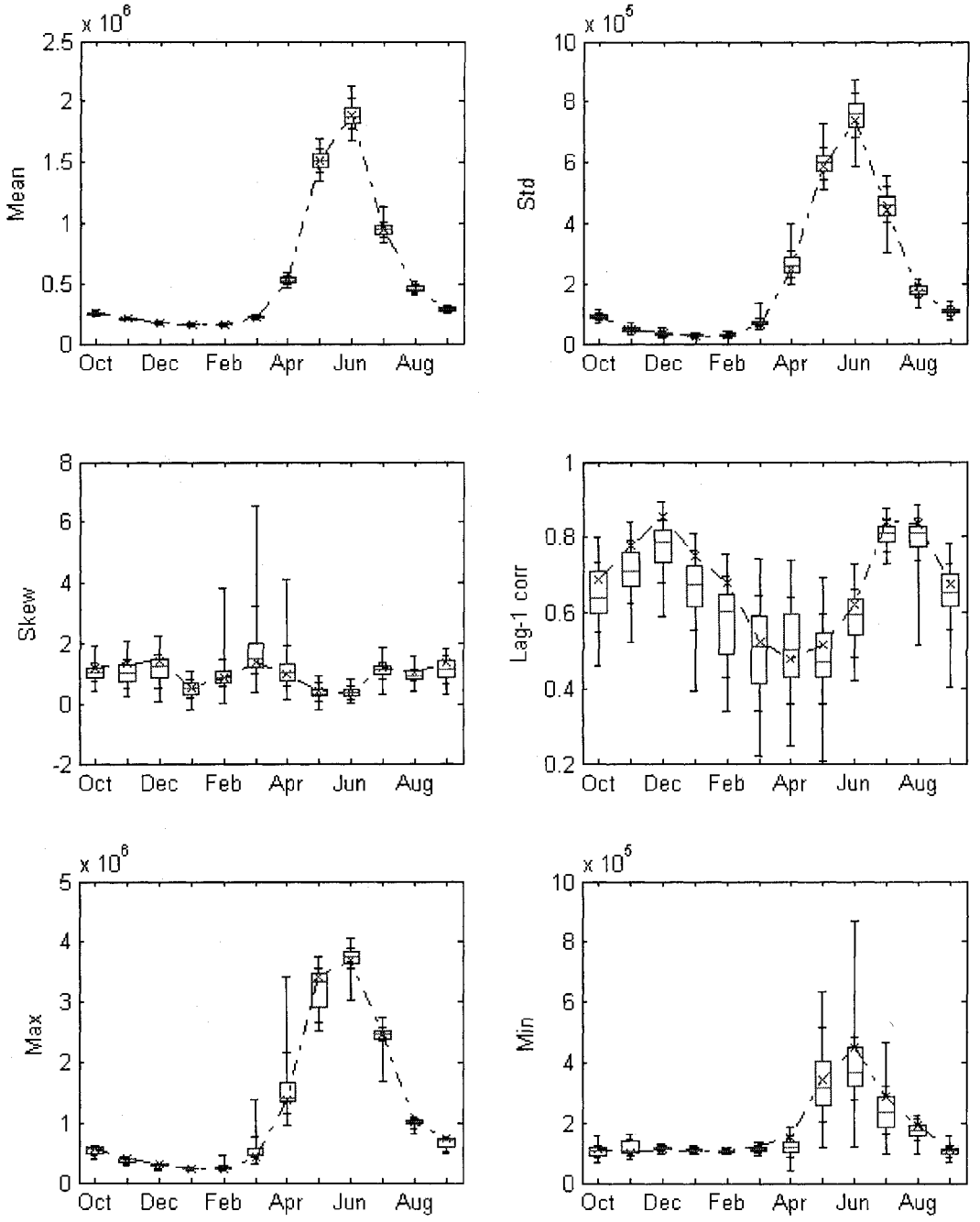


Figure 3-A.7 Key Statistics of Historical (dot line) and simulations (boxplot) with GAMBB for Site 8 of the Colorado River monthly streamflow Unit : Acre-feet

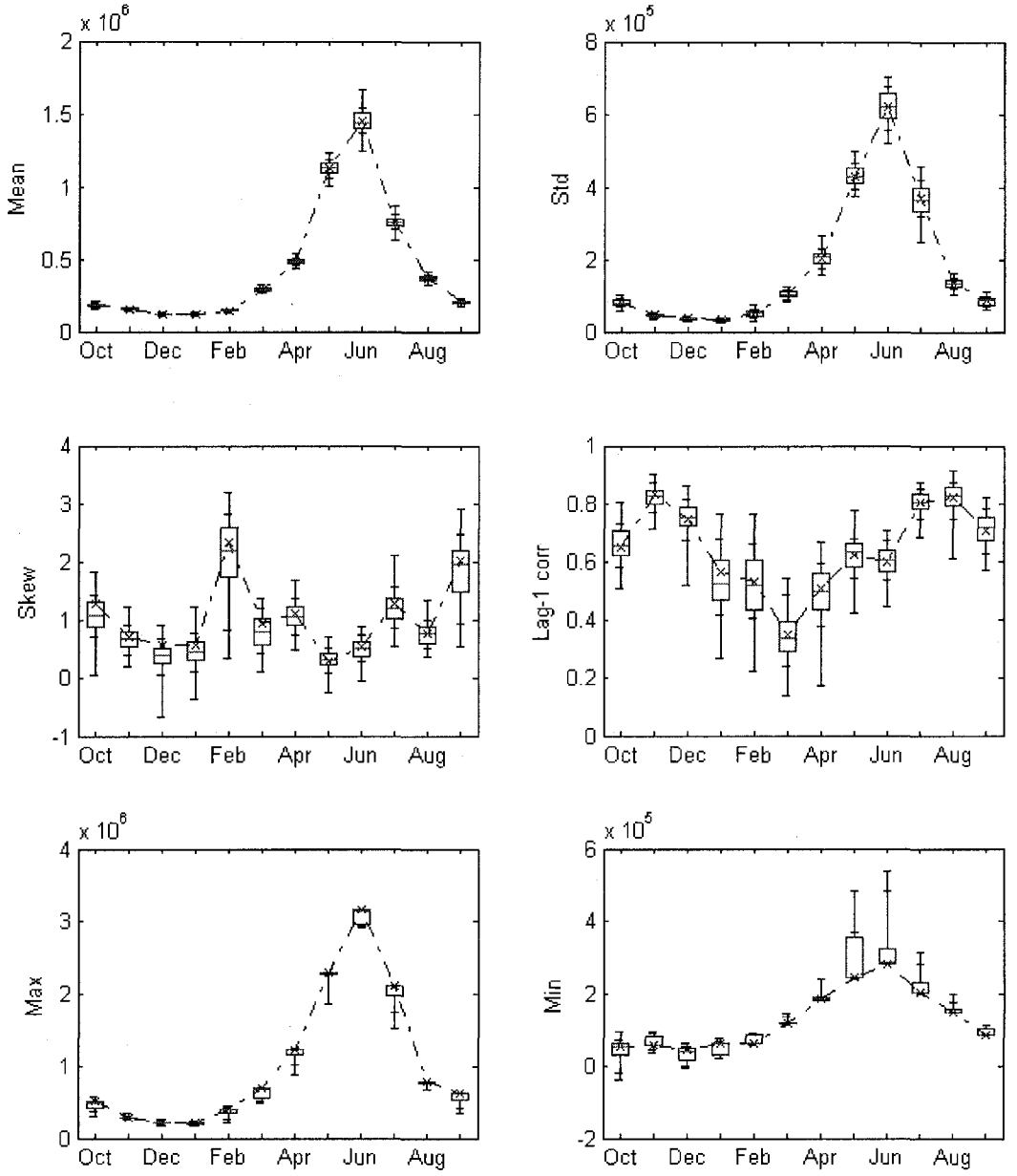


Figure 3-A.8 Key Statistics of Historical (dot line) and simulations (boxplot) with Hybrid for Site 16 of the Colorado River monthly streamflow Unit : Acre-feet

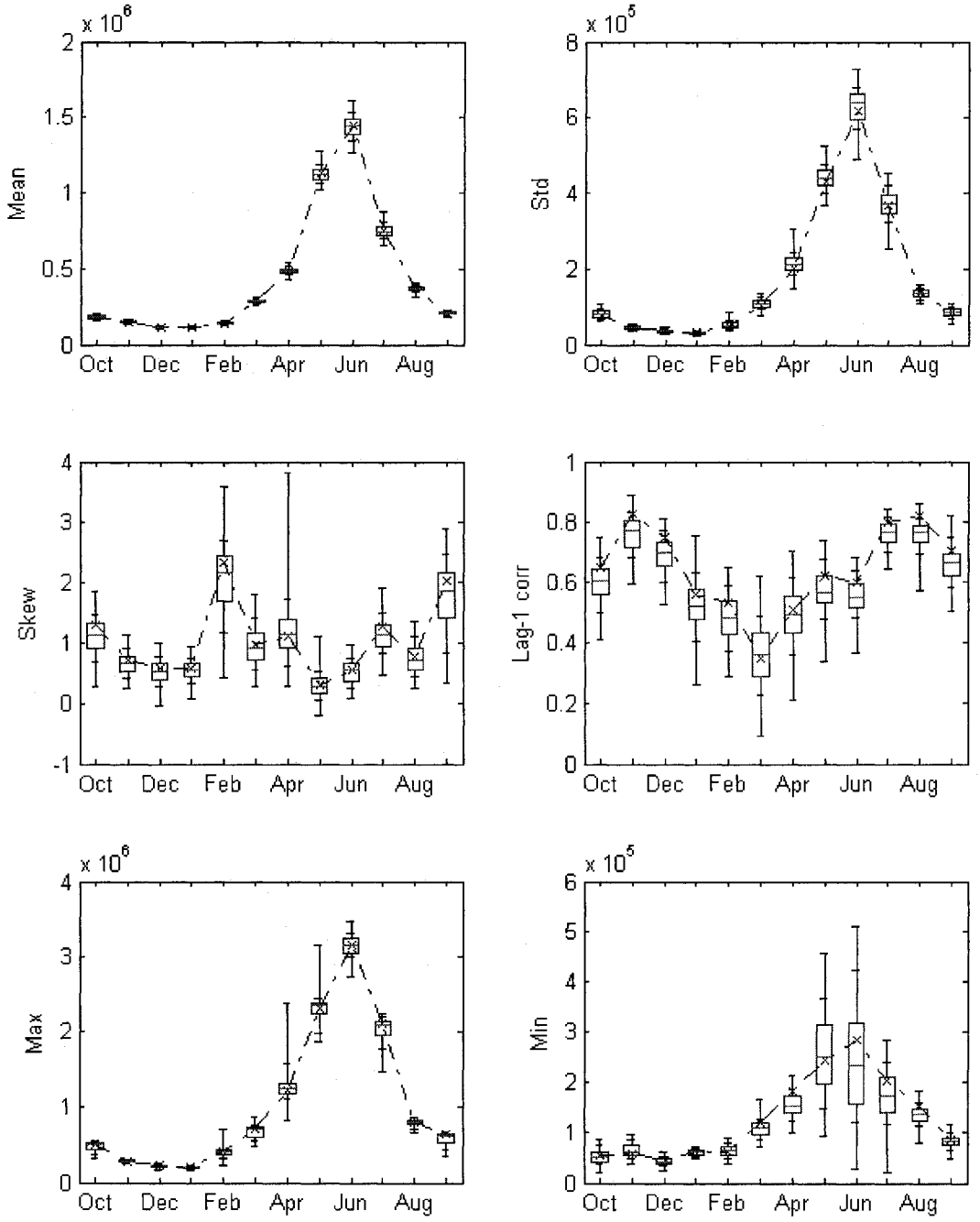


Figure 3-A.9 Key Statistics of Historical (dot line) and simulations (boxplot) with GAMBB for Site 16 of the Colorado River monthly streamflow Unit : Acre-feet

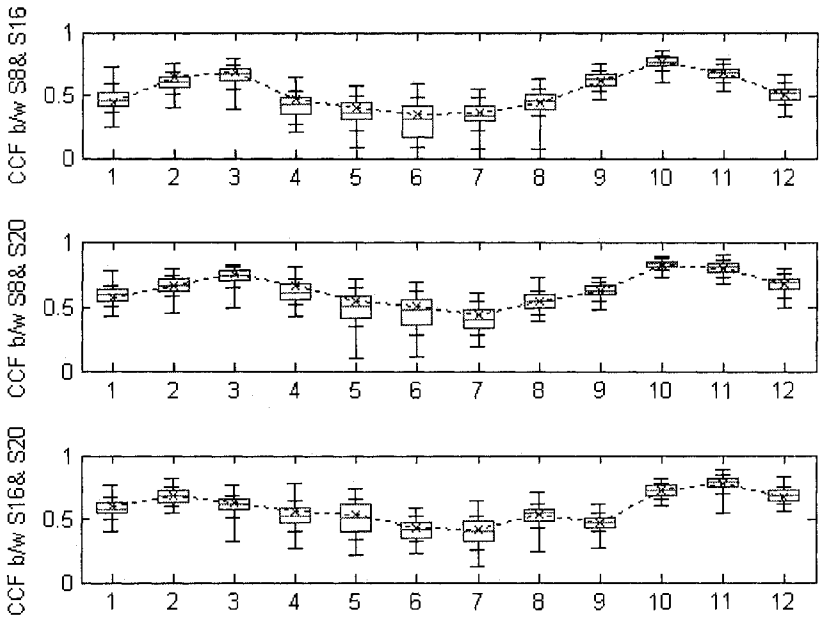


Figure 3-A.10 Lag-1 cross-correlation between sites from Historical (dot line) and simulations (boxplot) with Hybrid of the Colorado River monthly streamflow

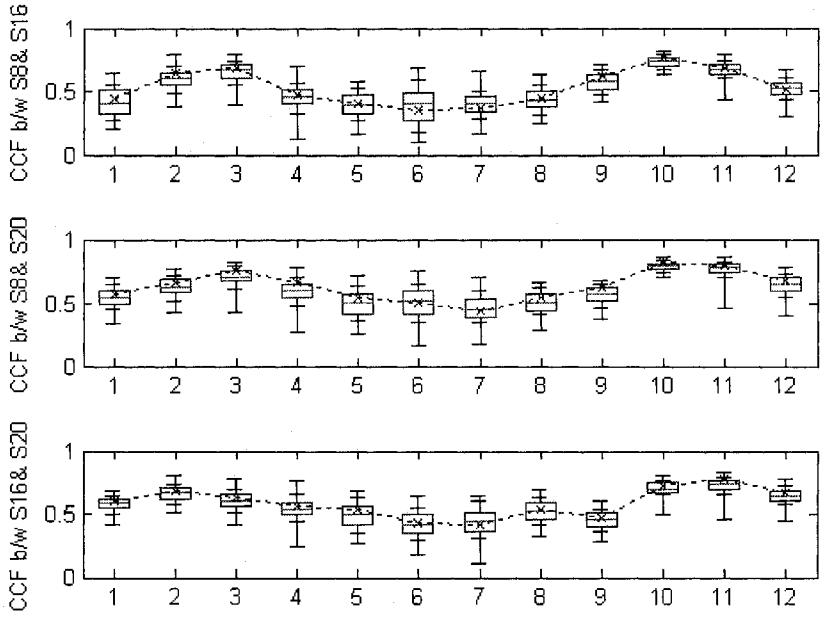


Figure 3-A.11 Lag-1 cross-correlation between sites from Historical (dot line) and simulations (boxplot) with GAMBB of the Colorado River monthly streamflow

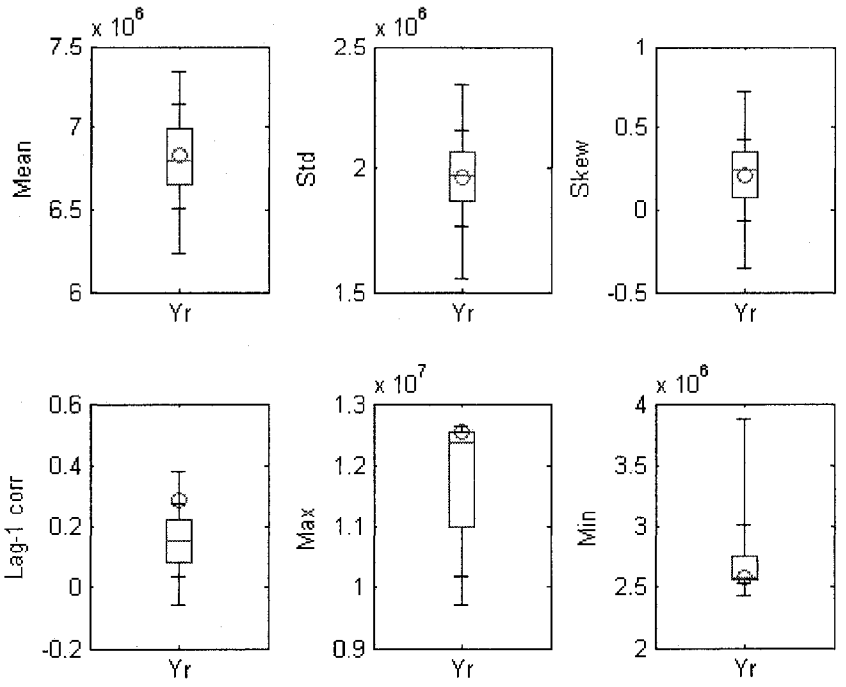


Figure 3-A.12 Key Statistics of Historical (dot line) and simulations (boxplot) with Hybrid for Site 8 of the Colorado River yearly streamflow Unit : Acre-feet

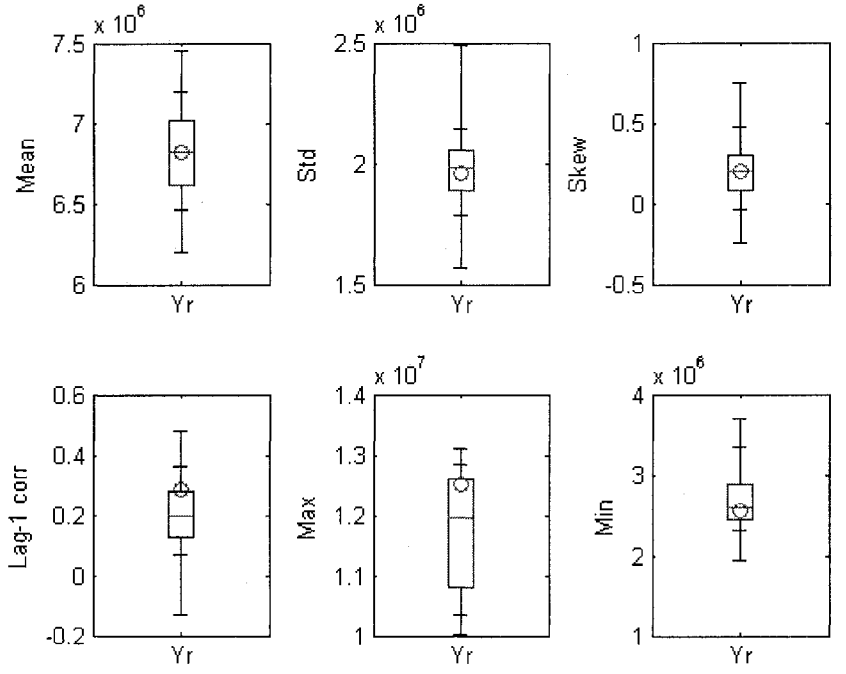


Figure 3-A.13 Key Statistics of Historical (dot line) and simulations (boxplot) with GAMBB for Site 8 of the Colorado River yearly streamflow Unit : Acre-feet

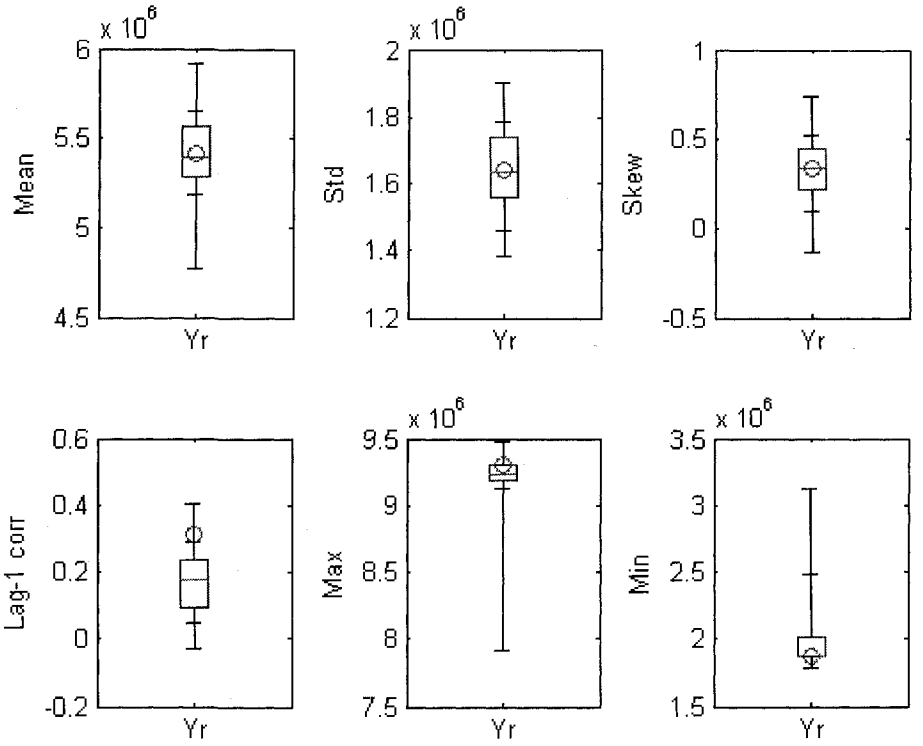


Figure 3-A.14 Key Statistics of Historical (dot line) and simulations (boxplot) with Hybrid for Site 16 of the Colorado River yearly streamflow Unit : Acre-feet

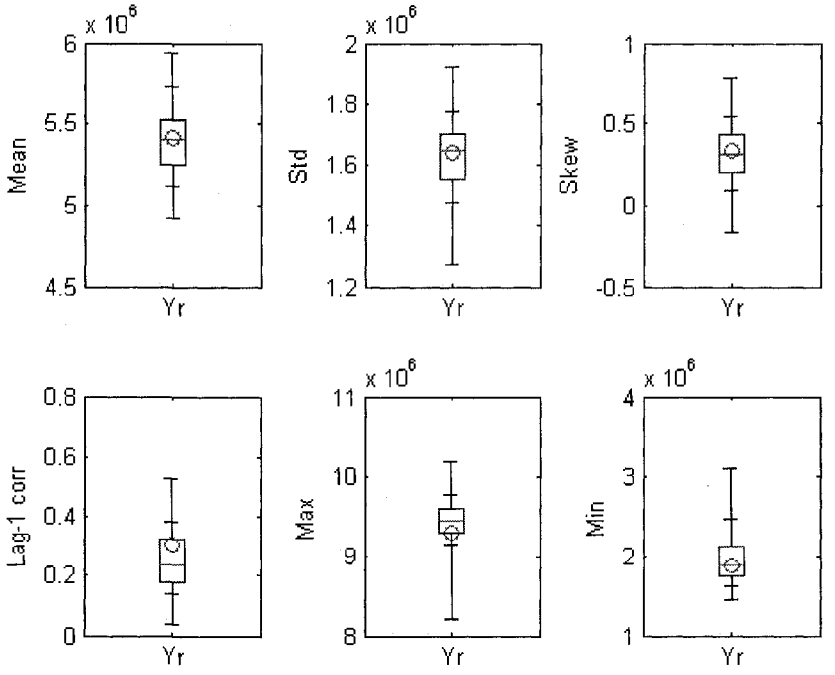


Figure 3-A.15 Key Statistics of Historical (dot line) and simulations (boxplot) with GAMBB for Site 16 of the Colorado River yearly streamflow Unit : Acre-feet

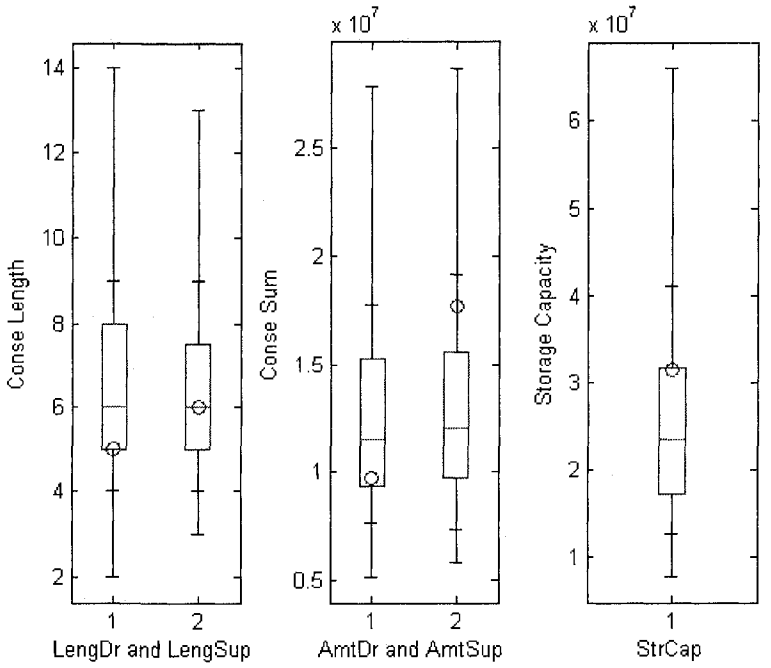


Figure 3-A.16 Reservoir-related statistics from Historical (dot line) and simulations (boxplot) with Hybrid for Site 8 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (Unit : Acre-feet)

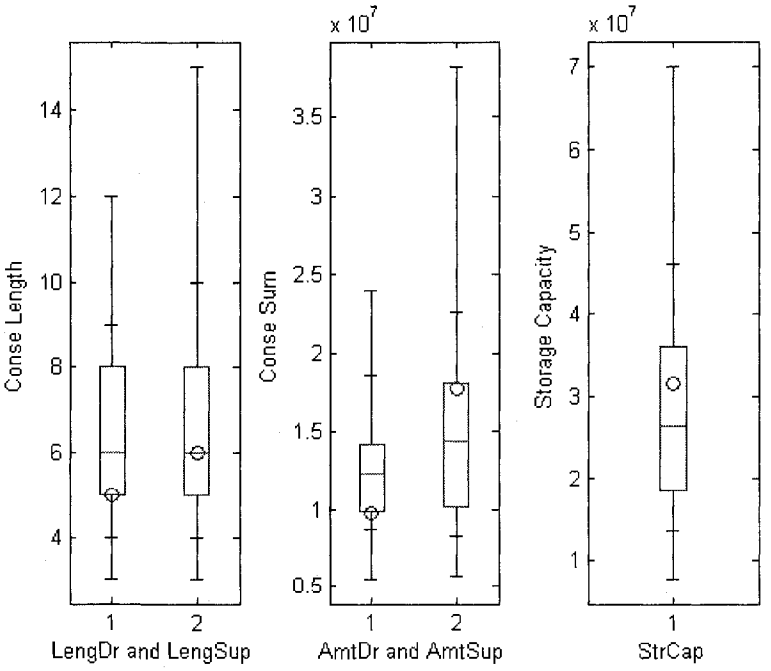


Figure 3-A.17 Reservoir-related statistics from Historical (dot line) and simulations (boxplot) with GAMBB for Site 8 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (Unit : Acre-feet)

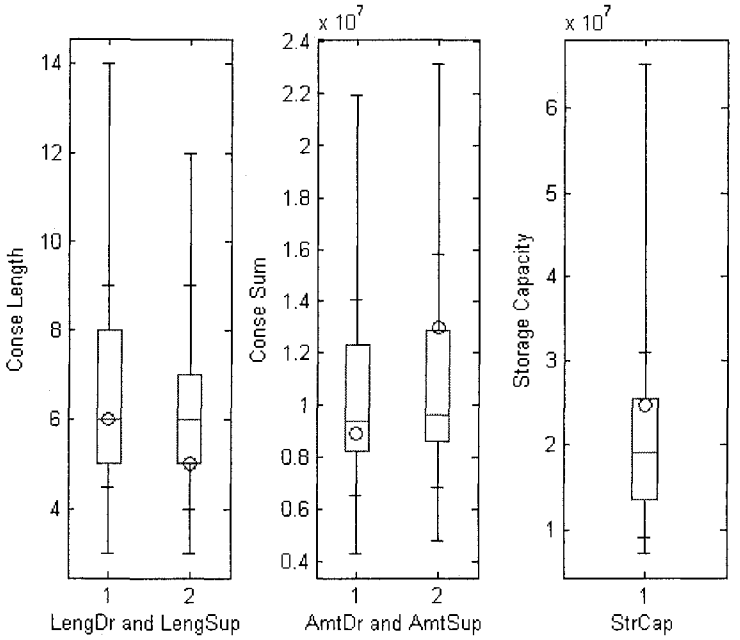


Figure 3-A.18 Reservoir-related statistics from Historical (dot line) and simulations (boxplot) with Hybrid for Site 16 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (Unit : Acre-feet)

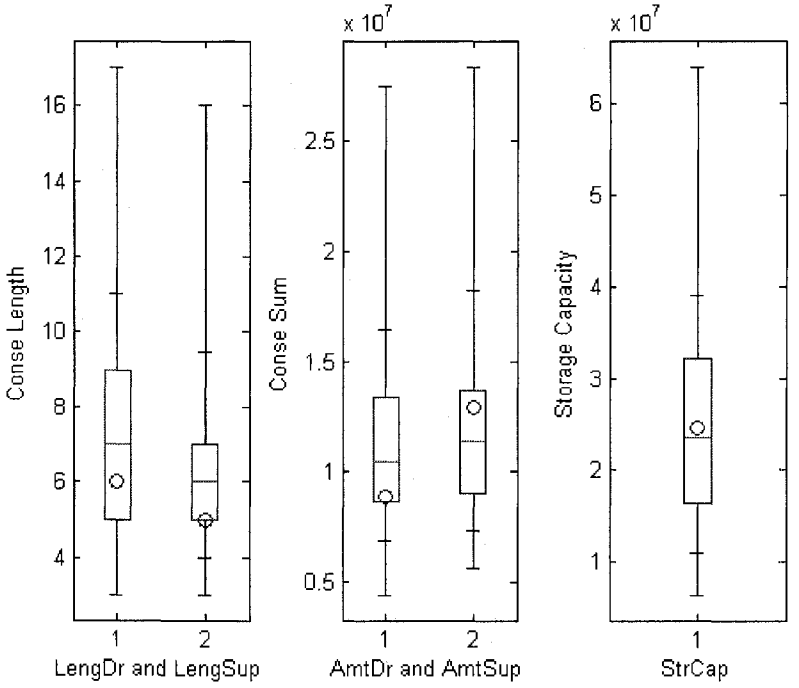


Figure 3-A.19 Reservoir-related statistics from Historical (dot line) and simulations (boxplot) with GAMBB for Site 16 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity (Unit : Acre-feet)

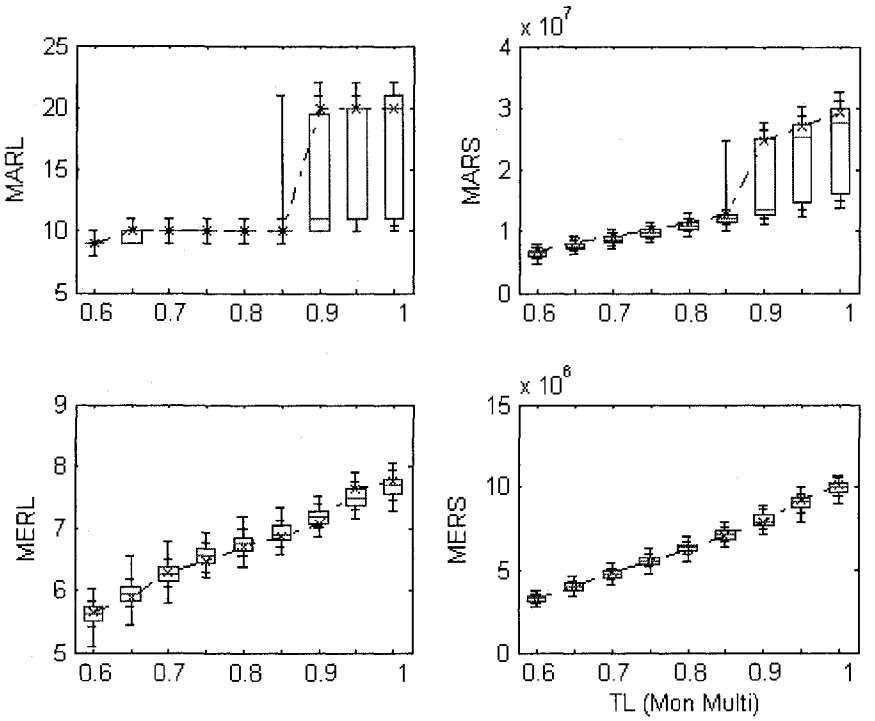


Figure 3-A.20 Multisite Monthly Drought Statistics of Historical (-x-) and Hybrid simulations (boxplot) of the Colorado River streamflow Unit : Acre-feet

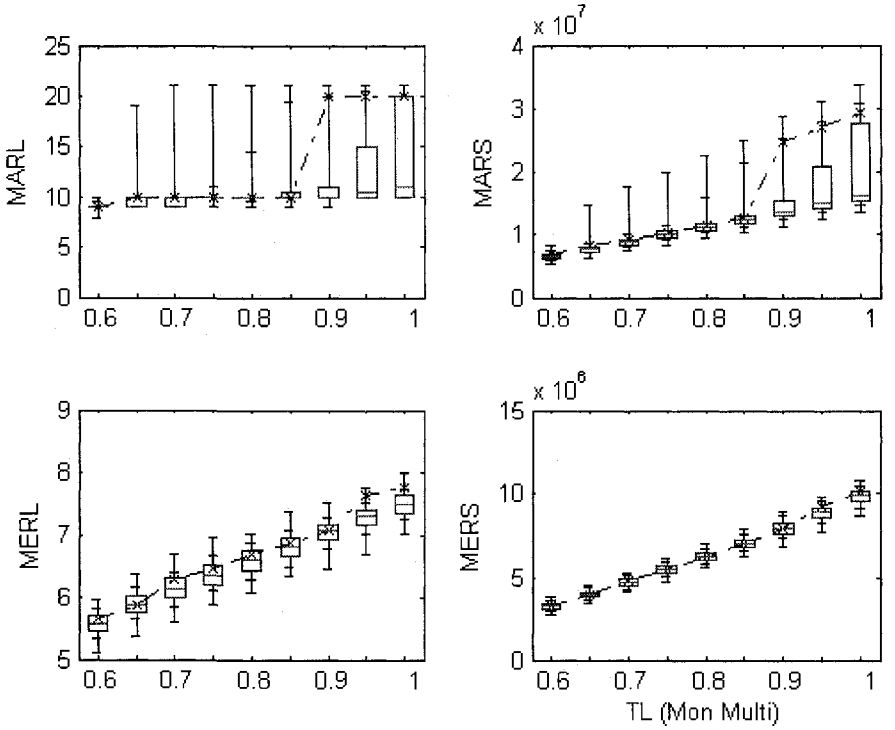


Figure 3-A.21 Multisite Monthly Drought Statistics of Historical (-x-) and GAMBB simulations (boxplot) of the Colorado River streamflow Unit : Acre-feet

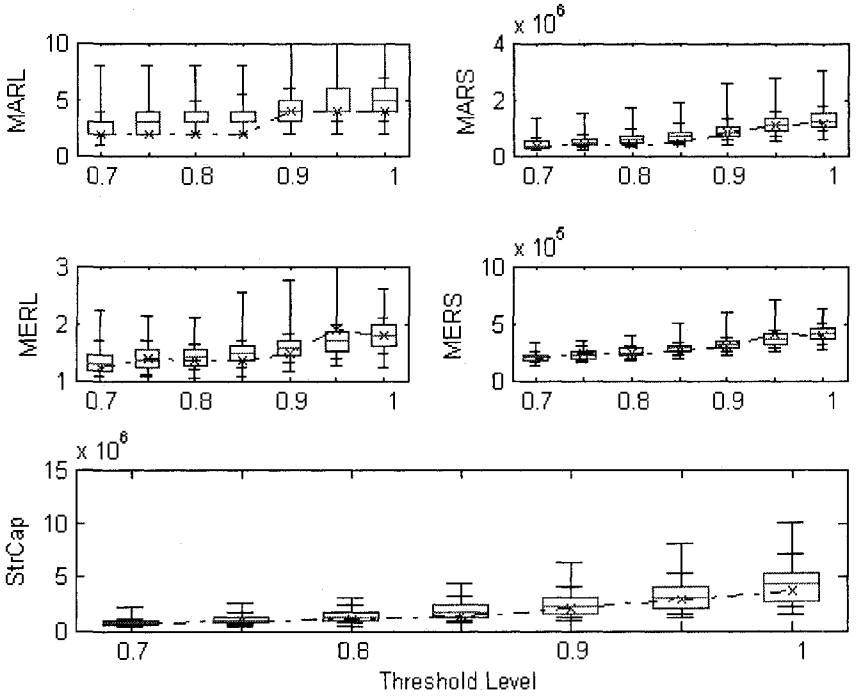


Figure 3-A.22 Multisite Seasonal Drought Statistics of Historical (-x-) and Hybrid simulations (boxplot) of the Colorado River streamflow Unit : Acre-feet

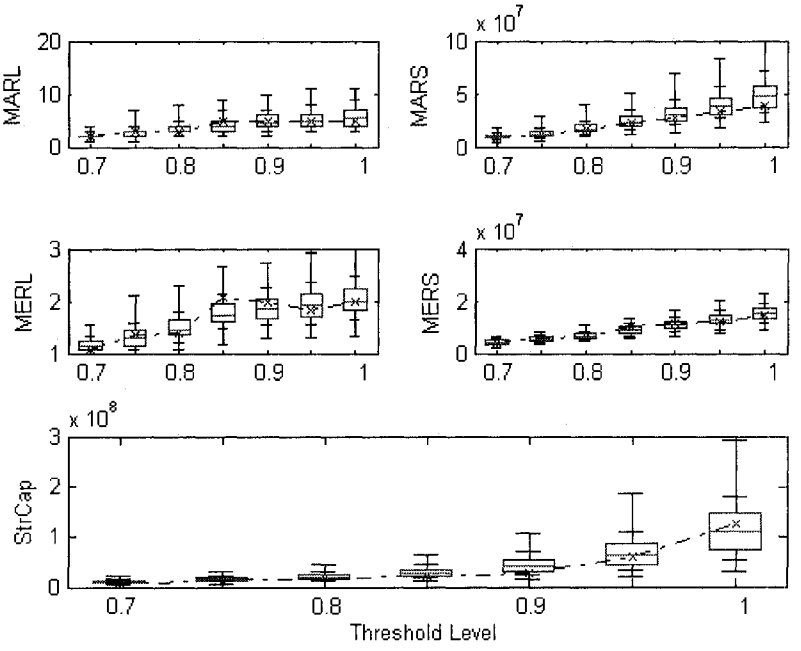


Figure 3-A.23 Multisite Seasonal Drought Statistics of Historical (-x-) and GAMBB simulations (boxplot) of the Colorado River streamflow Unit : Acre-feet

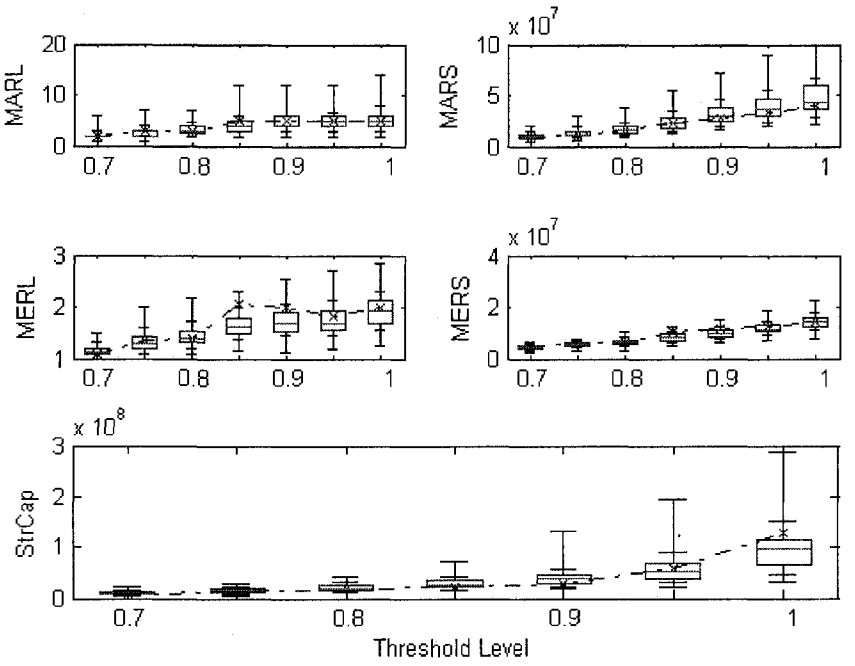


Figure 3-A.24 Multisite Yearly Drought Statistics of Historical (circle) and Hybrid simulations (boxplot) at the Colorado River streamflow Unit : Acre-feet

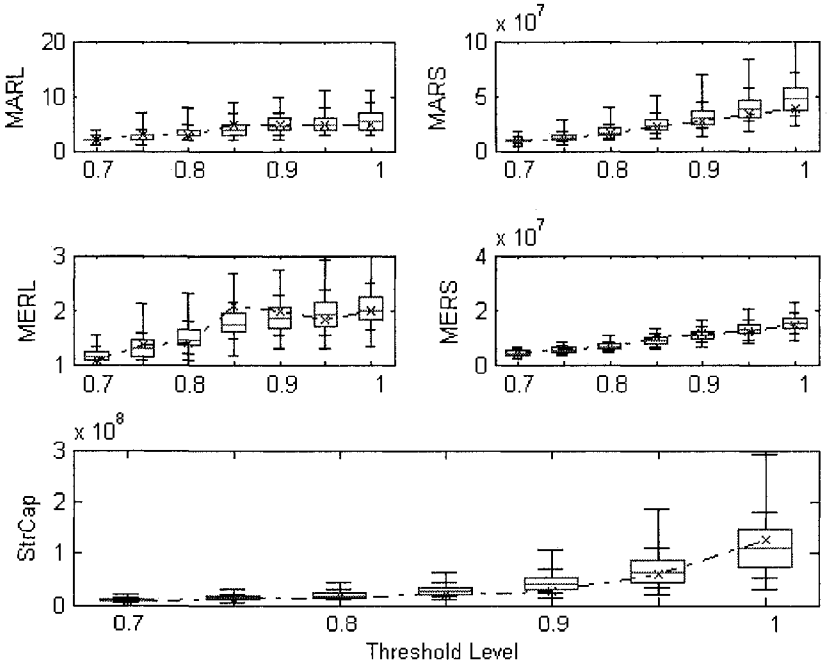


Figure 3-A.25 Multisite Yearly Drought Statistics of Historical (circle) and GAMBB simulations (boxplot) at the Colorado River streamflow Unit : Acre-feet

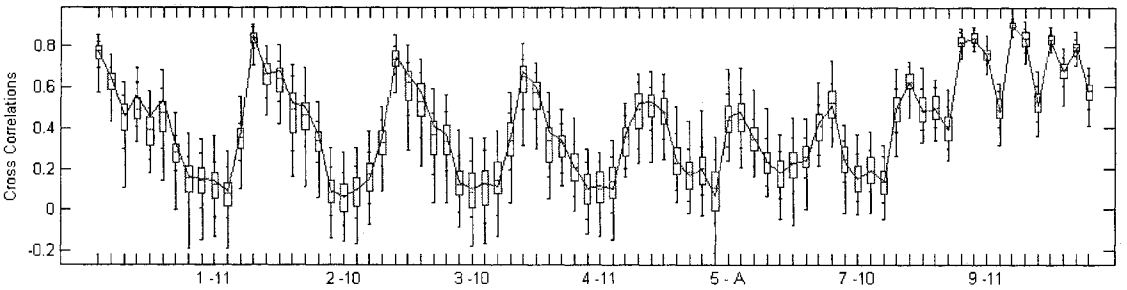


Figure 3-A.26 cross-correlation pairs of Historical (dot line) and Hybrid simulations (boxplot) of Site 8at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

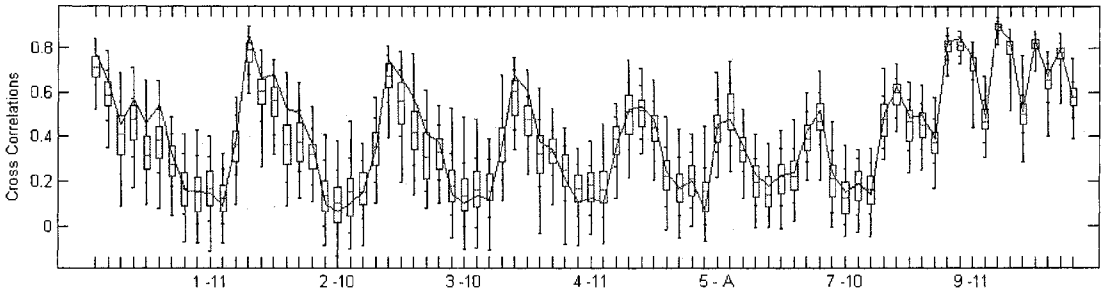


Figure 3-A.27 Cross-correlation pairs of Historical (dot line) and simulations (boxplot) with GAMBB of Site 8at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

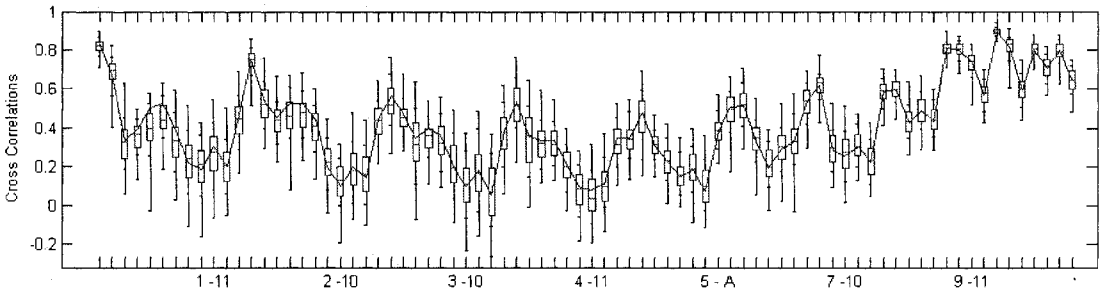


Figure 3-A.28 cross-correlation pairs of Historical (dot line) and Hybrid simulations (boxplot) of Site 16 at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

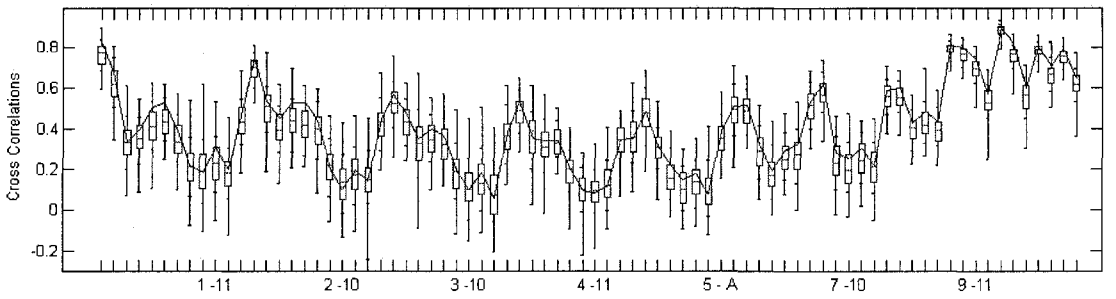


Figure 3-A.29 Cross-correlation pairs of Historical (dot line) and simulations (boxplot) with GAMBB of Site 16 at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

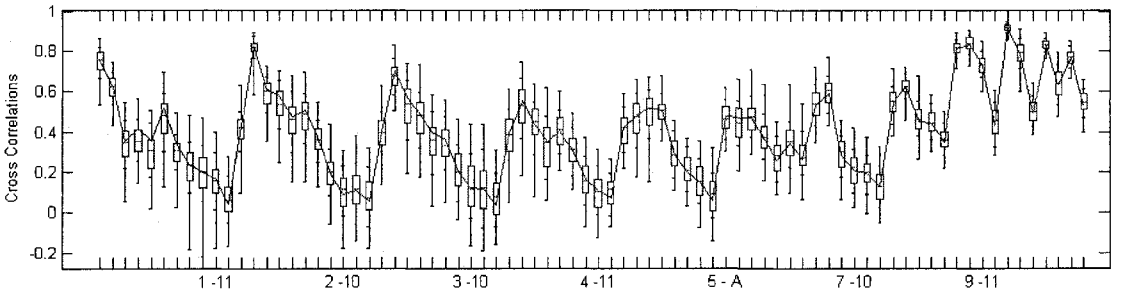


Figure 3-A.30 cross-correlation pairs of Historical (dot line) and Hybrid simulations (boxplot) of Site 20 at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

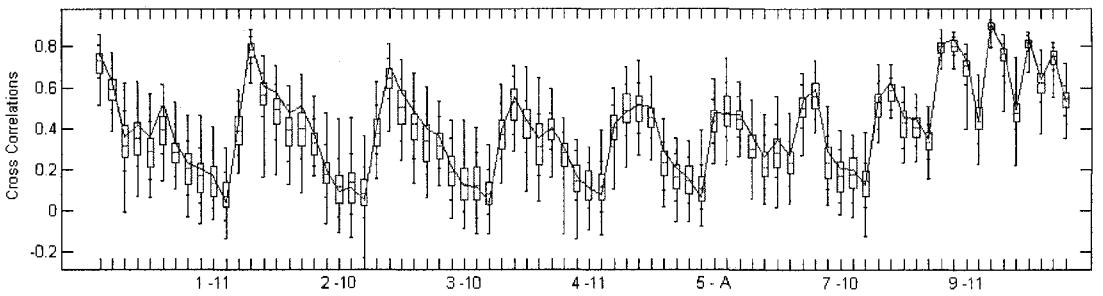


Figure 3-A.31 Cross-correlation pairs of Historical (dot line) and simulations (boxplot) with GAMBB of Site 20 at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

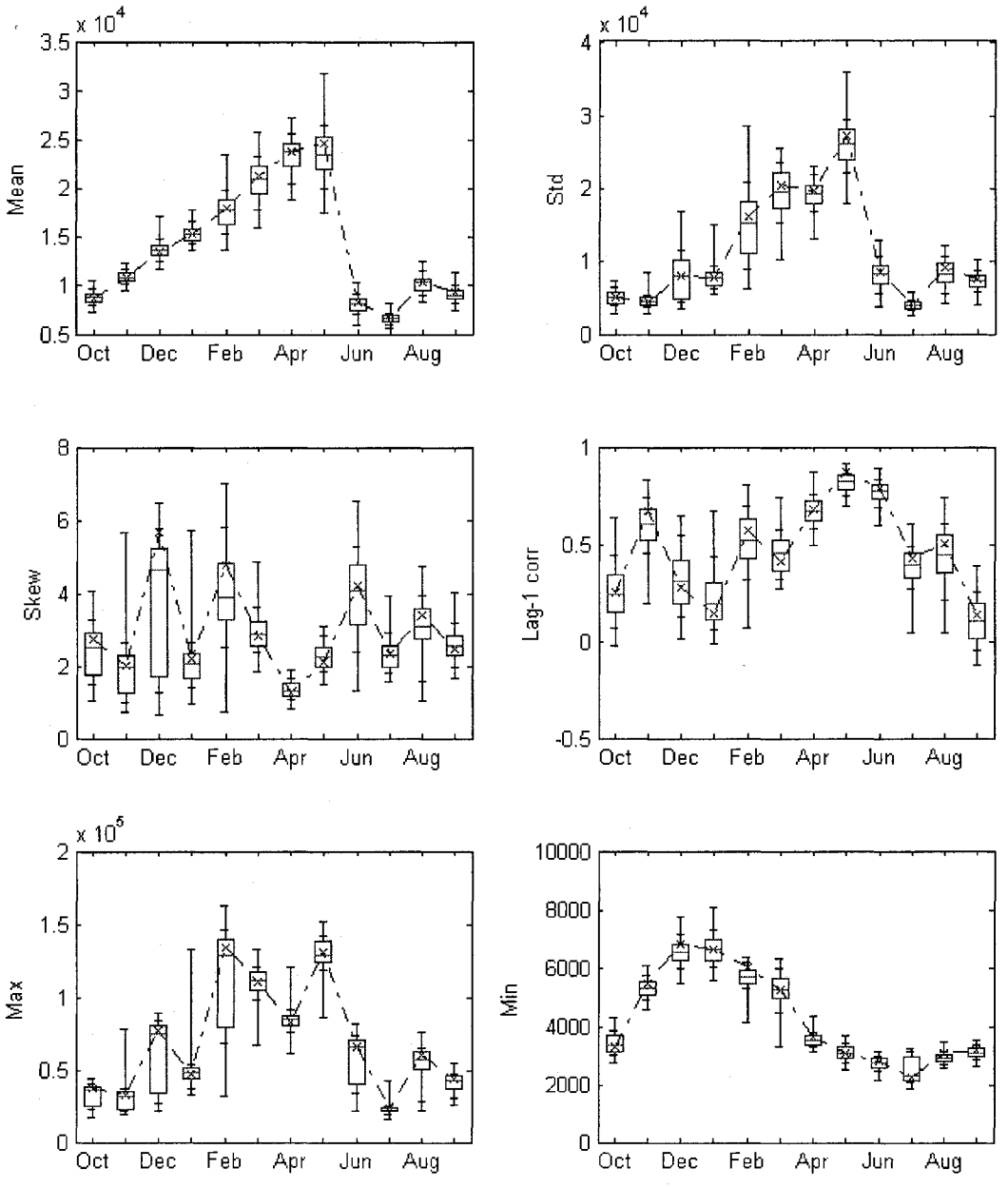


Figure 3-A.32 Key Statistics of Historical (dot line) and GAMBB simulations (boxplot) for Site 24 of the Colorado River monthly streamflow

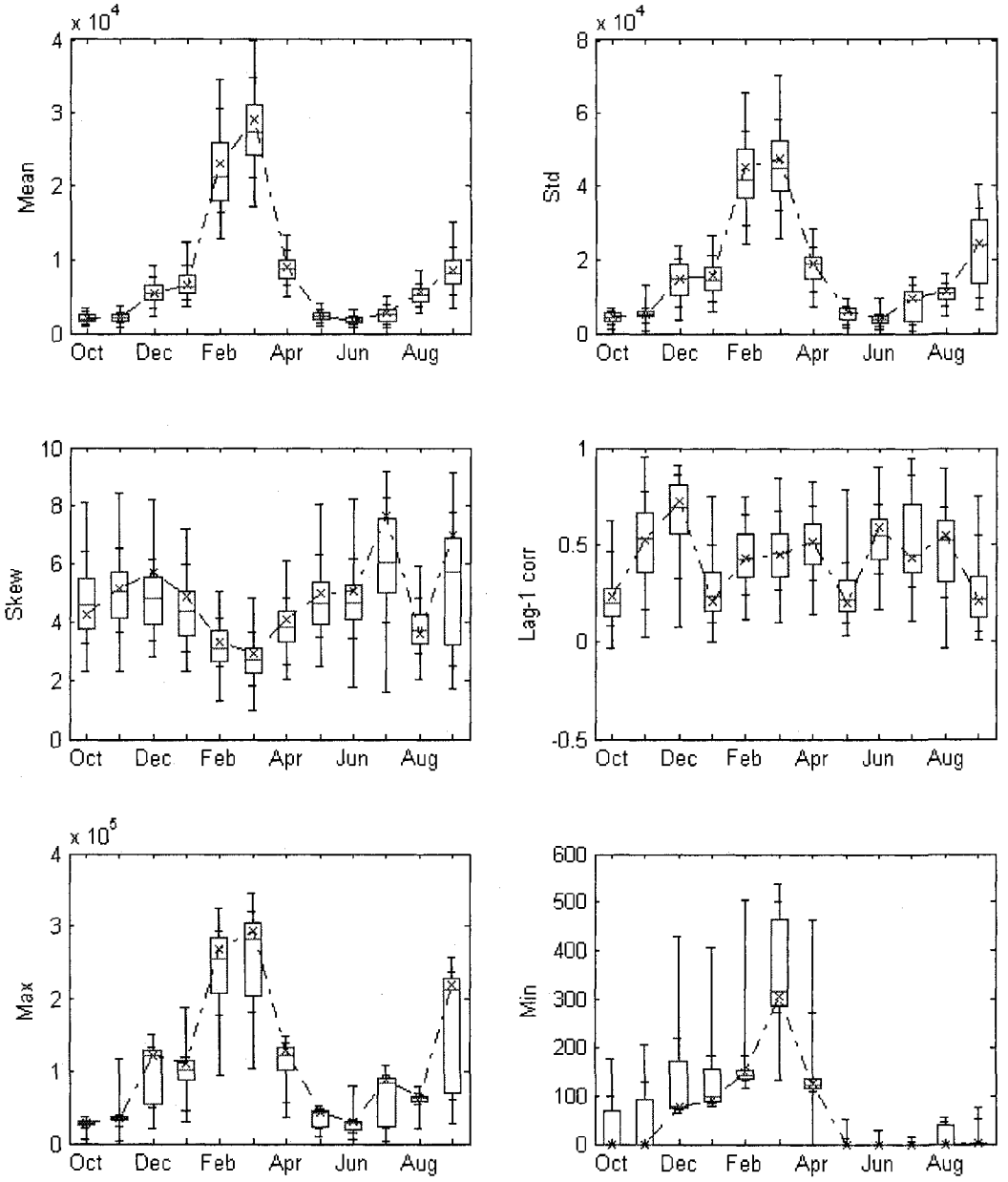


Figure 3-A.33 Key Statistics of Historical (dot line) and GAMBB simulations (boxplot) for Site 27 of the Colorado River monthly streamflow Unit : Acre-feet

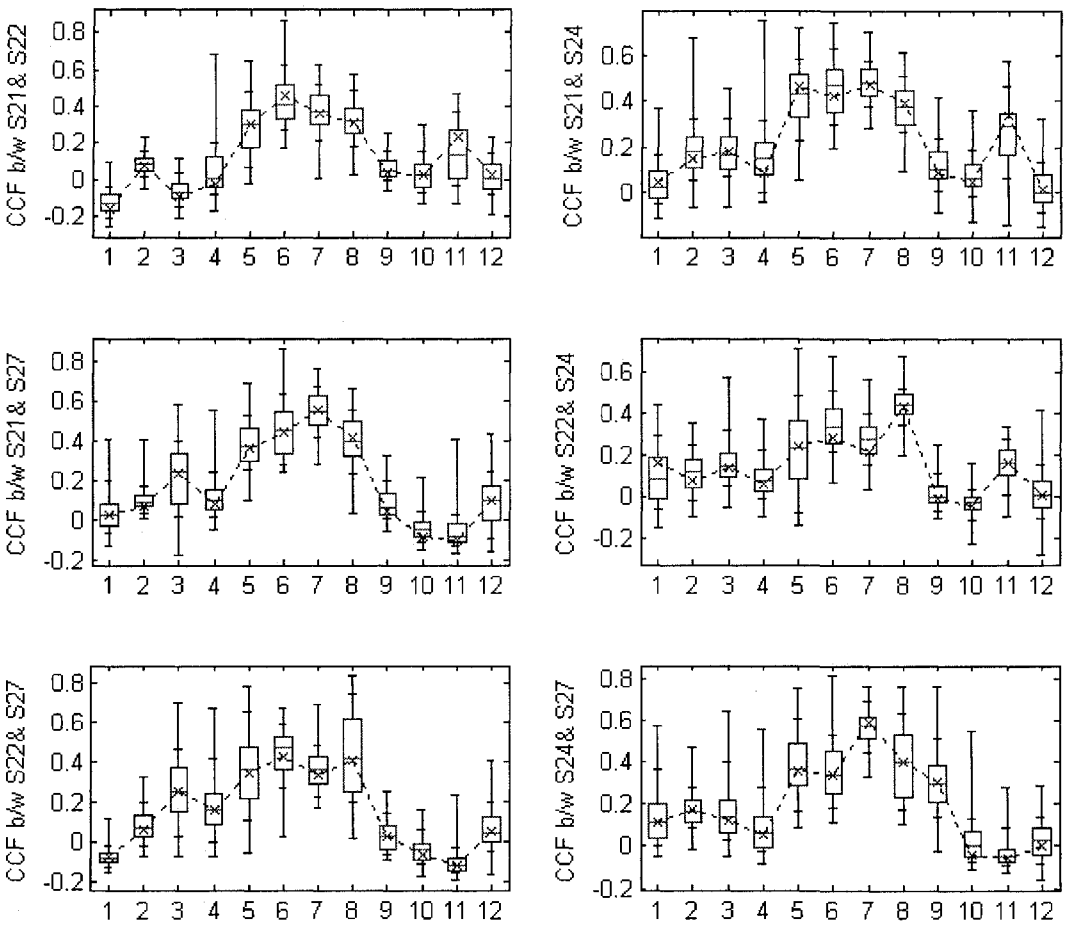


Figure 3-A.34 Lag-1 cross-correlation between sites from the historical (-x-) and GAMBB simulations (boxplot) of the Colorado River monthly streamflow

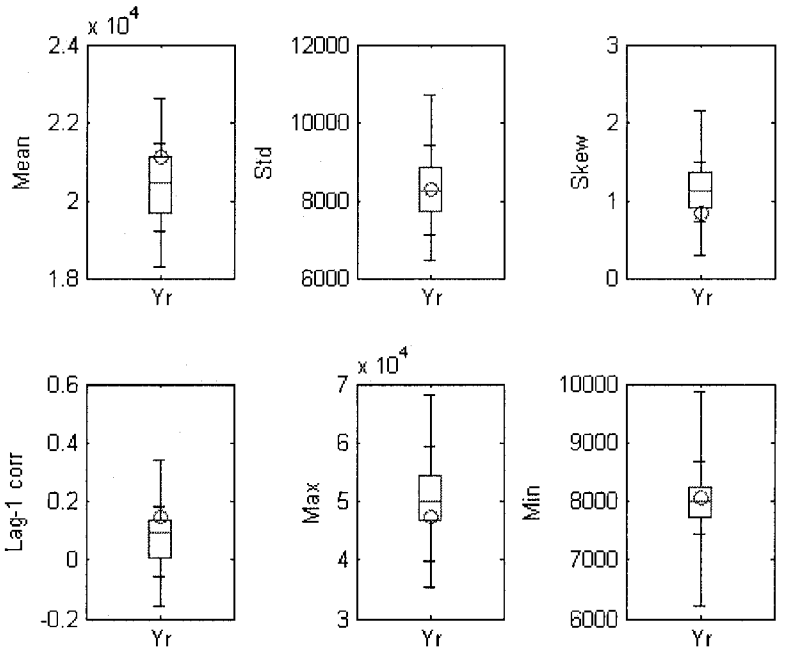


Figure 3-A.35 Key Statistics of Historical (circle) and GAMBB simulations (boxplot) for Site 21 of the Colorado River yearly streamflow Unit : Acre-feet

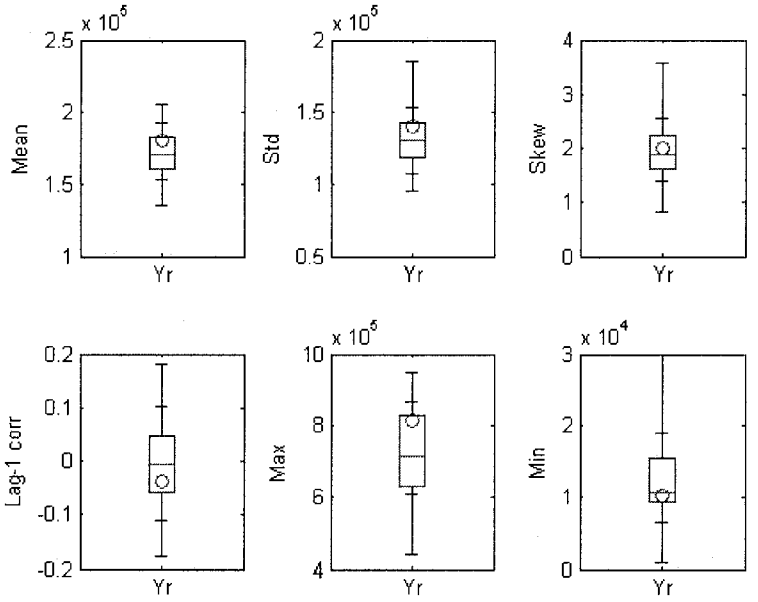


Figure 3-A.36 Key Statistics of Historical (circle) and GAMBB simulations (boxplot) for Site 22 of the Colorado River yearly streamflow Unit : Acre-feet

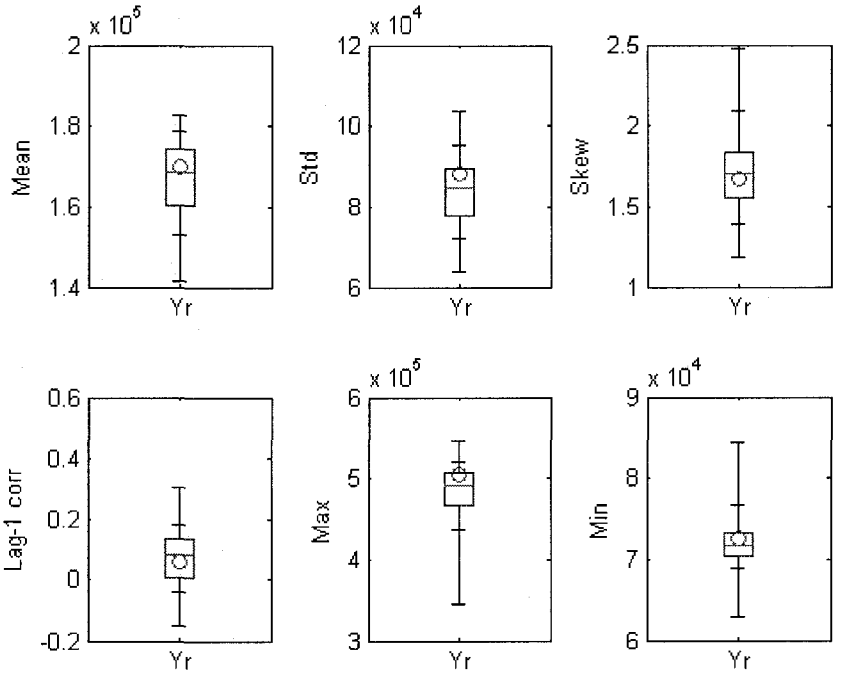


Figure 3-A.37 Key Statistics of Historical (circle) and GAMBB simulations (boxplot) for Site 24 of the Colorado River yearly streamflow

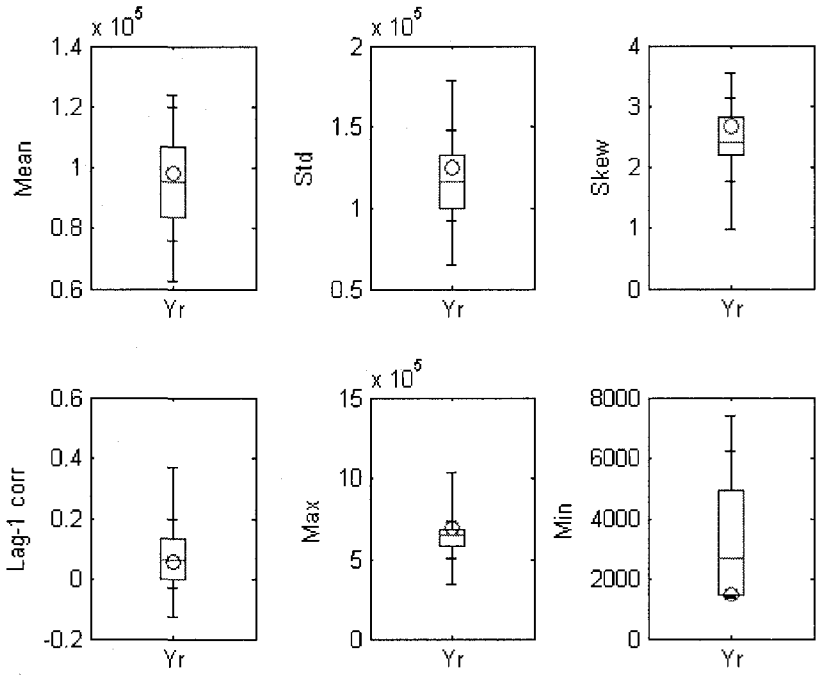


Figure 3-A.38 Key Statistics of Historical (circle) and GAMBB simulations (boxplot) for Site 27 of the Colorado River yearly streamflow

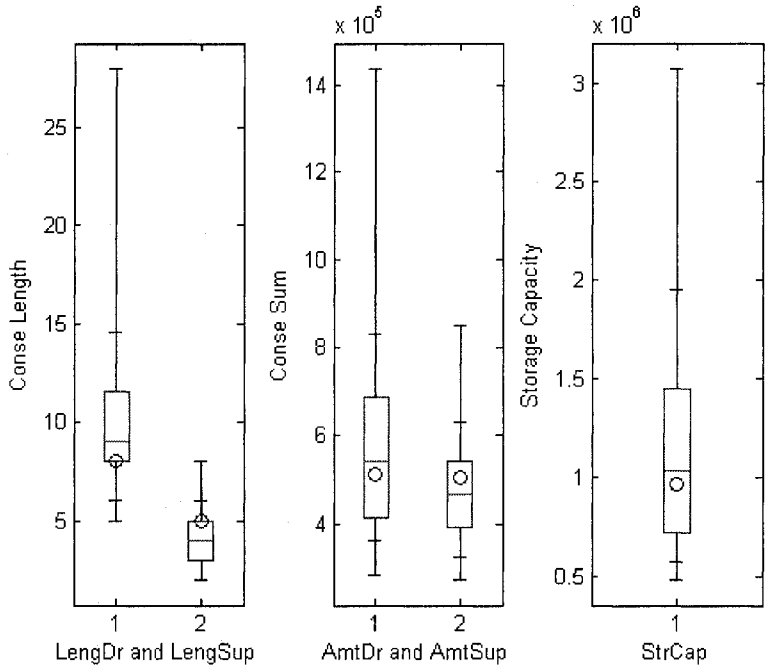


Figure 3-A.39 Reservoir-related statistics from historical (circle) and GAMBB simulations (boxplot) for Site 24 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity

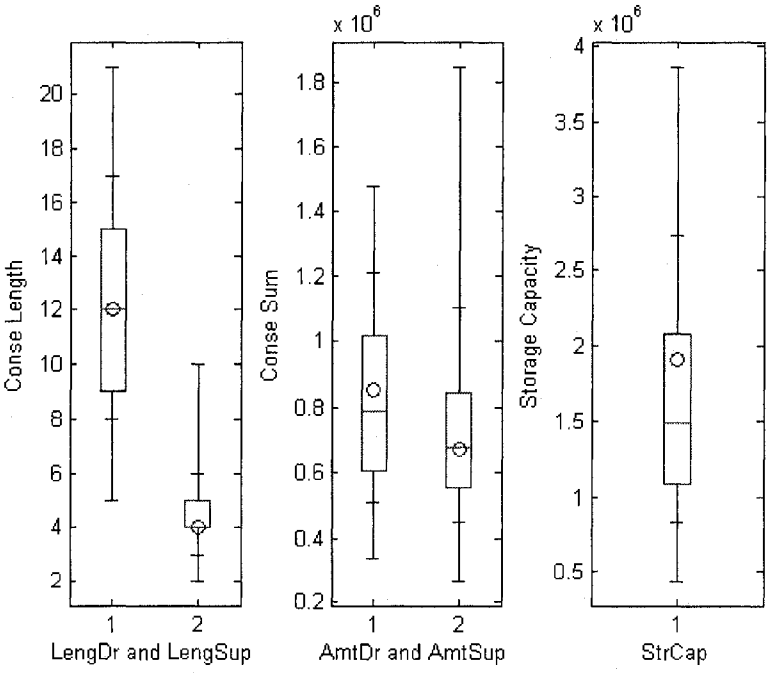


Figure 3-A.40 Reservoir-related statistics from historical (circle) and GAMBB simulations (boxplot) for Site 27 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity

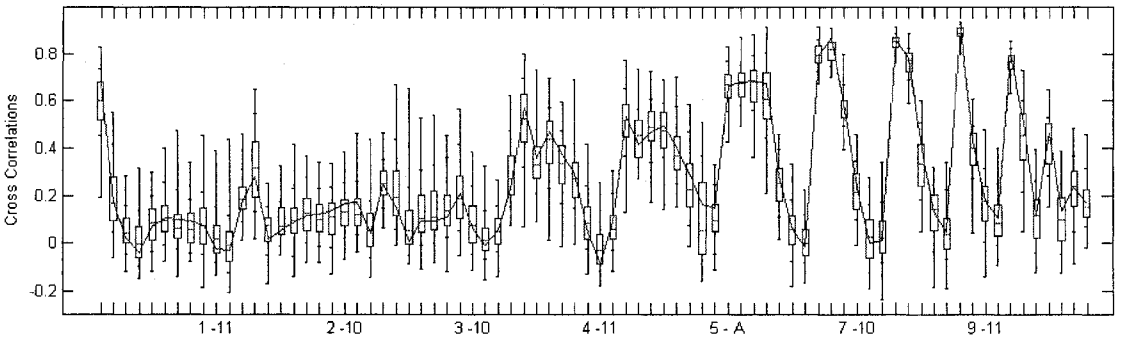


Figure 3-A.41 cross-correlation pairs of the historical and simulated data from GAMBB simulations (boxplot) and E[*1*]=*12* of the site 24 at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

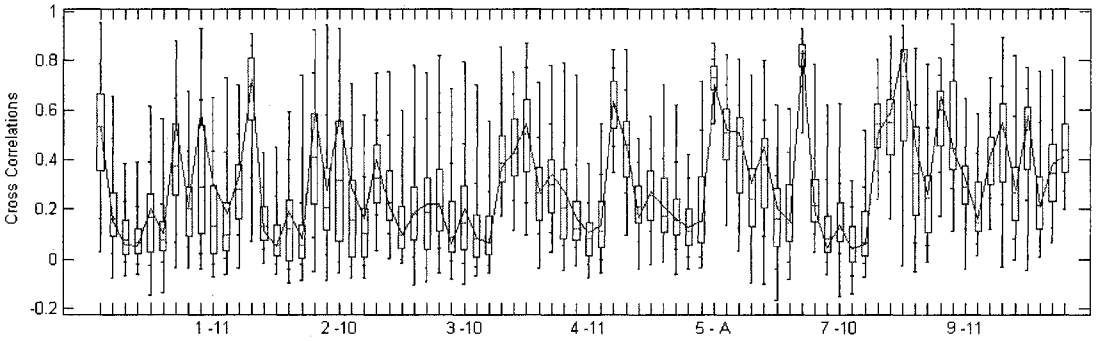


Figure 3-A.42 cross-correlation pairs of the historical and simulated data from GAMBB simulations (boxplot) and E[*1*]=*12* of the site 27at the Colorado River monthly streamflow. The label in x-axis (5-A) indicates the pair between month 5 and annual data

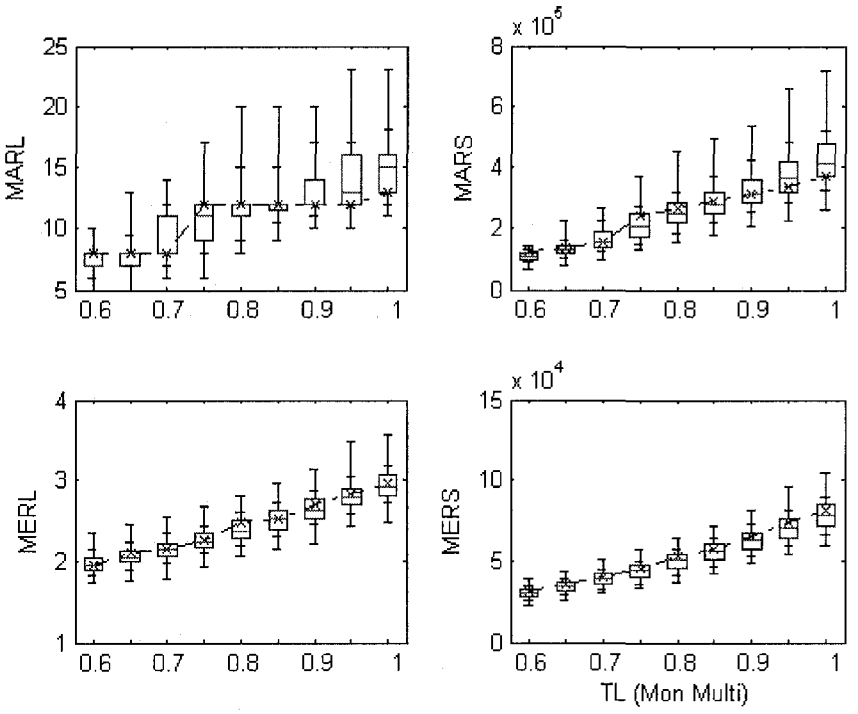


Figure 3-A.43 Multisite Monthly Drought Statistics of Historical (-x-) and GAMBB simulations (boxplot) of the Colorado River streamflow

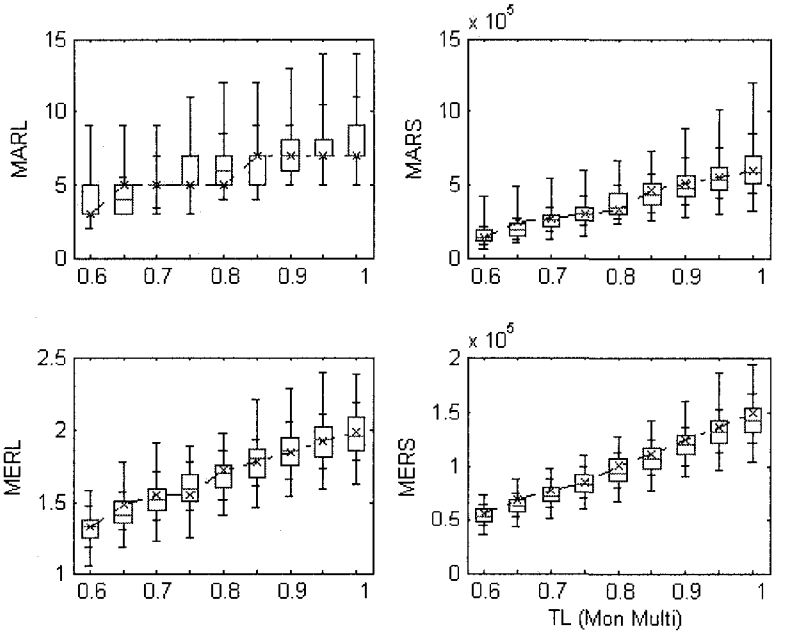


Figure 3-A.44 Multisite Seasonal (4 seasons) Drought Statistics of Historical (-x-) and GAMBB simulations (boxplot) of the Colorado River streamflow

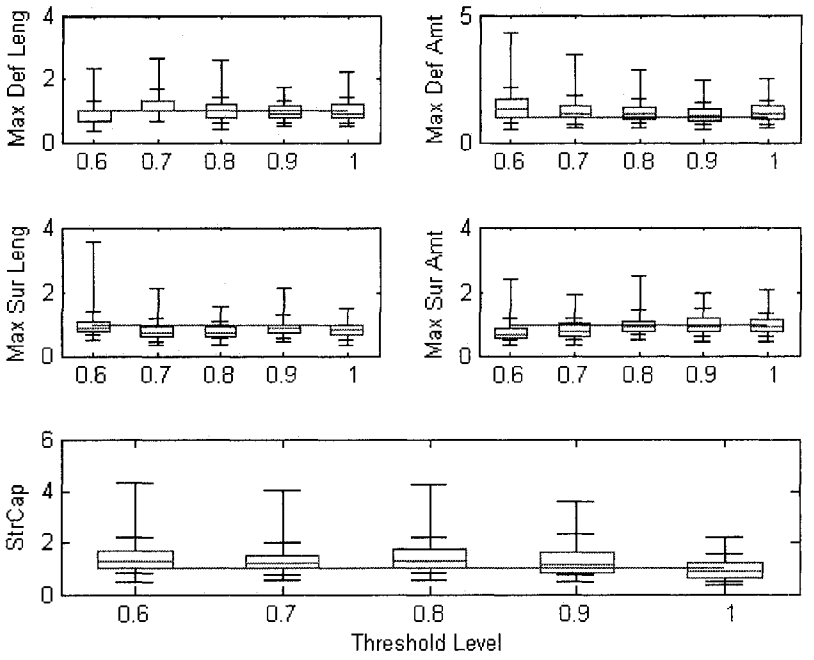


Figure 3-A.45 Multisite Yearly Drought Statistics of Historical (circle) and GAMBB simulations (boxplot) of Site 21 at the Colorado River streamflow

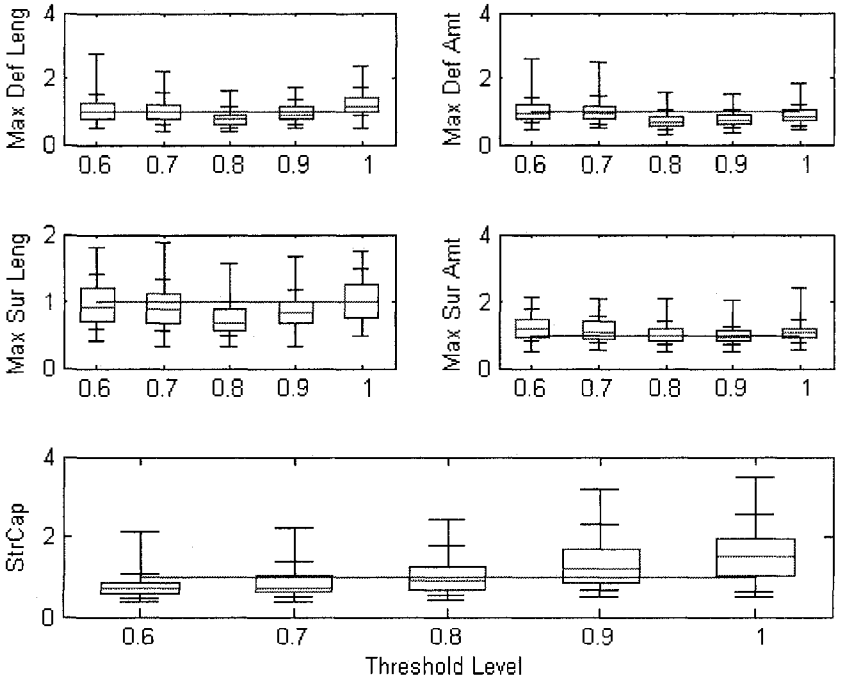


Figure 3-A.46 Multisite Yearly Drought Statistics of Historical (circle) and GAMBB simulations (boxplot) of Site 22 at the Colorado River streamflow

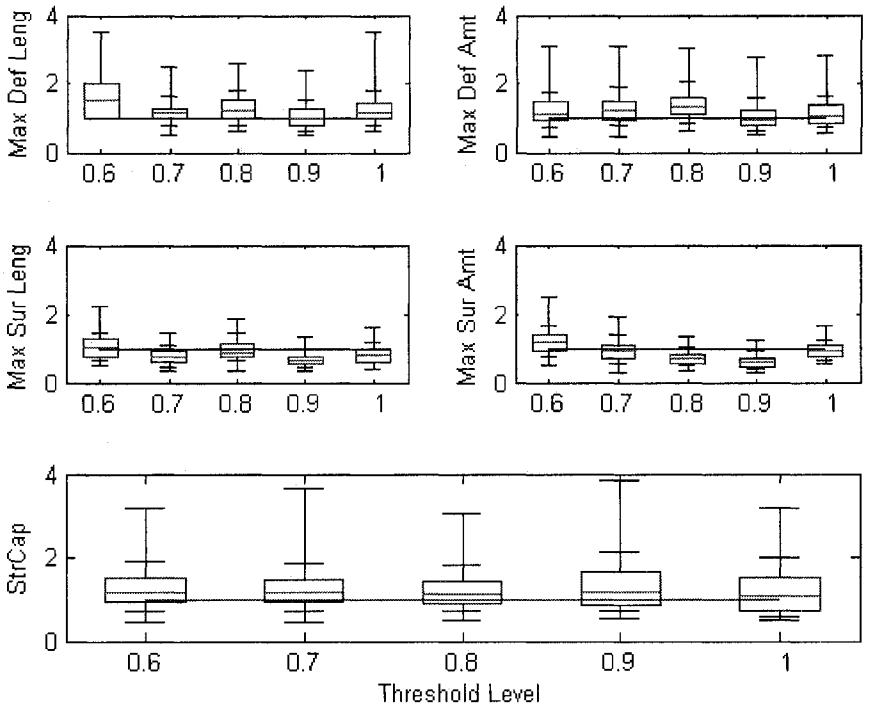


Figure 3-A.47 Multisite Yearly Drought Statistics of Historical (circle) and GAMBB simulations (boxplot) of Site 24 at the Colorado River streamflow

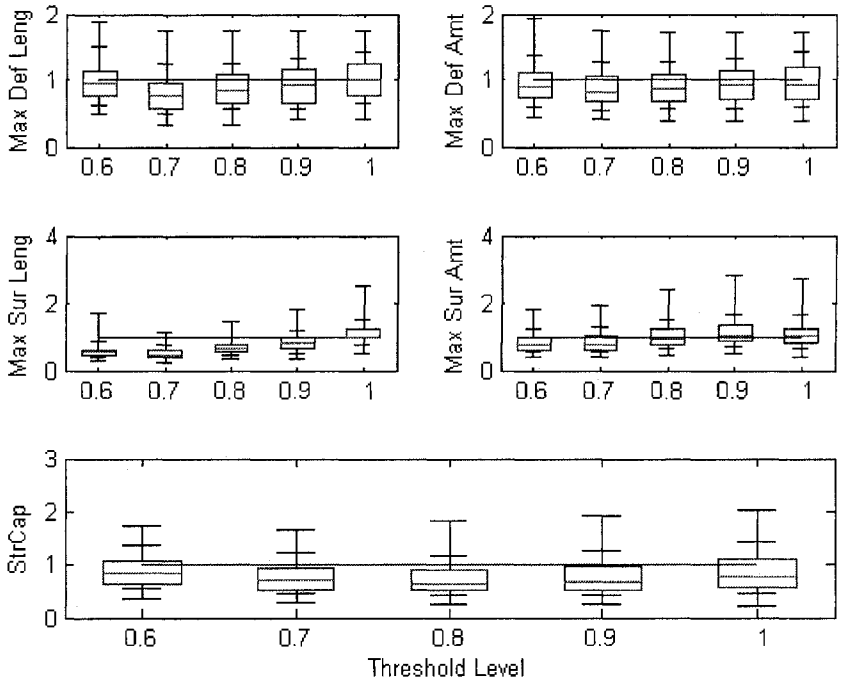


Figure 3-A.48 Multisite Yearly Drought Statistics of Historical (circle) and GAMBB simulations (boxplot) of Site 27 at the Colorado River streamflow

CHAPTER IV

NONPARAMETRIC STREAMFLOW DISAGGREGATION MODEL

4.1 Introduction

Stochastic generation models are required for drought analysis and reservoir planning of a complex river system in the hydrologic field. For analyzing a river system, a generation model of multisite data should be used in order to account for the cross effects among individual sites. Multivariate autoregressive (MAR) time series models have been employed in literature (Salas 1993) and for a seasonal multivariate time series, periodic MAR (PMAR). Since the seasonal time series model cannot reproduce the variability of the aggregated level, disaggregation models have been developed such as Mejia and Rousselle (1976), Santos and Salas (1992), and Valencia and Schaake (1973). The model with single site is mathematically described as

$$\mathbf{Y} = AX + B\mathbf{V} \tag{4-1}$$

where **Y** is the seasonal data and X is the annual data, A and B are parameter matrices and **V** is the independent [d×1] vector with the standard normal distribution, and d is the number of seasons. A final requirement is the model met the additivity condition such that $X = \sum_{i=1}^{d} Y_i$ where $\mathbf{Y} = \{Y_i\}_{i\in\{1,\ldots,d\}}$. This parametric disaggregation model cannot preserve the serial correlation between the last month of the previous year and the first month of the current year. Thus, Mejia and Rousselle (1976) included an additional term to fix the drawback as

$$\mathbf{Y} = AX + B\mathbf{V} + CE \tag{4-2}$$

where C is the parameter matrix and E is the last seasonal value of the previous year. The parametric disaggregation models (Mejia and Rousselle, 1976; Valencia and Schaake, 1973) , however, require estimating a tremendous number of parameters. For this reason, some parsimonious disaggregation models have been proposed by researchers (e.g. Stedinger and Vogel, 1984, Stedinger et al., 1985a and 1985b, and Santos and Salas, 1992. Furthermore, Koutsoyiannis and Manetas (1996) developed a useful algorithm that combines two different models for two time scales, called the accurate adjusting procedure (AAP). For example, if yearly and monthly data are simulated from the lag-1 autoregressive (AR-1) model and lag-1 periodic AR (PAR-1) independently, then this algorithm works to match two different time-scale sequences. Further details will next be described in the review section

Those disaggregation models have significant drawbacks about which many researchers have mentioned (Prairie et al. 2007; Srinivas and Srinivasan 2006; Tarboton

et al. 1998) such as requirements of transformation, along with the assumption of the gaussian marginal distribution, bias of the statistics of the generated data; and generating negative values. To overcome the shortcomings of the parametric disaggregation models, a nonparametric disaggregation (NPD) model has been developed by Tarboton et al. (1998). Rotating disaggregate variables with the Gram-Schmidt orthonormal (GSO) matrix, the rotated data are generated from the kernel density estimate and the scaled aggregate variable is included as the last element of the rotated data. The data are back-rotated to get the original domain. To avoid a massive multivariate kernel density estimate, Prairie et al. (2007) employed KNNB to select the GSO rotated observed data. In this paper, we will investigate the characteristics of the NPD model and reveal the similarity to AAP. Its drawbacks are revealed from the results such as no concern on the variability of each disaggregate variables and the same pattern of disaggregate variables as the historical data. Further detail will be explained later in the result section. To surmount those identified drawbacks, a modification of the NPD model along with a Genetic algorithm is suggested and tested.

In brief summary for the following chapters, two fundamental nonparametric disaggregation models are reviewed in chapter 2. In chapter 3, the suggested model components and procedure are described. The data description and results is shown in chapter 4 followed by the conclusion and summary in chapter 5.

4.2 Review of two existing disaggregation approaches

Among various disaggregation approaches, two existing approaches are reviewed in this section, such as the accurate adjusting procedure (AAP) suggested by

Koutsoyiannis and Manetas (1996) generally used in parametric modeling, and nonparametric disaggregation model with KNN (Prairie et al., 2007), the cutting-edge technique of disaggregation models. At the end of this section, the similarity of these two models is explained followed by suggestions to introduce a new approach according to the weaknesses of these two models.

4.2.1 Notation

Disaggregation in stochastic simulation is a process that splits a higher-level (or aggregate) value into multiple lower-level (or disaggregate) values while preserving the statistics characteristic of both levels. For example, yearly data are disaggregated into monthly data, called temporal disaggregation and main stream station data are disaggregated into multiple substations, called spatial disaggregation. Lower-level variables (e.g. monthly) are denoted as $\mathbf{Y} = (Y_1,...,Y_d)^T$ where d is the number of disaggregate variables and X denotes the upper-level or aggregate (e.g. annual) variable. One of the most important features in disaggregation lexicon is the additivity condition, i.e.

$$Y_1 + Y_2 + ... + Y_d = X \tag{4-3}$$

Also Z^* will denote the generated data for the variable Z. The disaggregation approaches suggested in this paper will require initially choosing a candidate disaggregation variable set. Then the selected disaggregation variables are adjusted to meet the additivity condition. The generated candidate disaggregate variables are denoted as $\widetilde{\mathbf{Y}}^* = [\widetilde{Y}_1^*, \widetilde{Y}_2^*, ..., \widetilde{Y}_d^*]^T$, and their sum denoted as $\widetilde{X}^*$. Note that the candidate lower-

level variables may be generated from parametric models (e.g. MPAR) or from nonparametric procedures, e.g. using KNN. In the nonparametric case, the lower-level sequence candidates are drawn from historical data. In addition, aggregation and disaggregation can be conducted temporally or spatially. Generally our notations here are applicable for both. However, in some cases, we will use $y_{v,\tau}^{s}$ where s=1,…,S represents sites with S total number of sites, $v = 1,..,N$ denotes years with N=total number of years, and $\tau = 1,...,\omega$ represents seasons (or months) with ω =number of seasons. Furthermore, μ_Z and σ_Z are used to represent the mean and standard deviation of the Z, and $\sigma_{Z_1 Z_2}$ represents the covariance between the variable Z_1 and Z_2.

4.2.2 Accurate Adjusting Procedure

Koutsoyiannis and Manetas (1996) developed a useful scheme for coupling two different level models for aggregate and disaggregate variables. The models for the aggregate and disaggregate variables are fitted independently and the data generation procedure proceeds as follows:

(1) The aggregate data X^* are generated from the corresponding higher-level model. Then, the d-dimensional lower-level data $\widetilde{\mathbf{Y}}^* = [\widetilde{Y}_1^*, \widetilde{Y}_2^*, ..., \widetilde{Y}_d^*]^T$ are generated from the lower-level model independently from the aggregate variable.

(2) The sum of the disaggregate values are determined and the distance between the generated aggregate value X^* is calculated as:

$\Delta = \dfrac{\left|X^* - \widetilde{X}^*\right|}{\sigma_X}$ where $\widetilde{X}^* = \sum_{i=1}^{d} \widetilde{Y}_i^*$ and σ_X is the standard deviation of the yearly data X.

(3) If $\varepsilon < \Delta$, where ε is the tolerance level (0.1~1), then regenerate the disaggregate data set $\widetilde{\mathbf{Y}}^*$. Otherwise the disaggregate data are adjusted with one of three adjustments such as proportional, linear, and power adjusting procedures to match up with the additivity condition in Eq.(4-3). These three adjustments are explained below.

(4) The steps (1) ~ (3) will be continued until all the higher-level data are disaggregated. The adjusting procedures referred to in step (3) above may be proportional, linear, and power as:

$$Y_j^* = \widetilde{Y}_j^* + \lambda_j \left(X^* - \widetilde{X}^*\right) \qquad ,j=1,\ldots,d \qquad (4\text{-}4)$$

$$Y_j^* = \widetilde{Y}_j^* \frac{X^*}{\widetilde{X}^*} \qquad ,j=1,\ldots,d \qquad (4\text{-}5)$$

$$Y_j^* = \widetilde{Y}_j^* (X^* / \widetilde{X}^*)^{\lambda_j / \eta_j} \qquad ,j=1,\ldots,d \qquad (4\text{-}6)$$

where $\lambda_j = \sigma_{Y_j,X} / \sigma_X^2$ and $\eta_j = \mu_{Y_j} / \mu_X$

The linear adjustment above preserves the mean and standard deviation as well as the variance-covariance matrix of the lower-level variables (Koutsoyiannis and Manetas 1996). But negative values might be generated and higher order statistics such as skewness might be biased. Therefore, when disaggregate variables exhibit low skewness

(close to normal distribution) it is better to use the linear adjustment. Koutsoyiannis and Manetas (1996) derived the proof that the linear adjusted variables Y_j^* in Eq.(4-4) have the identical mean and variance-covariance matrix assuming that the mean and standard deviation of the aggregate variable are the same as those of the sum of the disaggregate variables. However, it is frequently not true and the variance of the summation of the low-level variables are underestimated when a direct lower-level model is applied (e.g. lag-1 Periodic Autoregressive model:PAR-1) since the covariance matrix of the lower-level variables is not preserved with low-order PARMA model (Bartolini and Salas 1993). This is principle reasons to employ the disaggregation model instead of direct application of the lower-level model.

Koutsoyiannis (1994) showed that the proportional adjusting is appropriate on a gamma marginal distribution with the same common scale parameter and different shape parameters for each disaggregate variable and independent each other. The assumption of independency might be relaxed from numerical tests. Generally monthly streamflow are skewed and can be fitted to a Gamma distribution. This distribution does not produce any negative values in any condition unless an aggregate or a disaggregate variable has negative values, which is the general case that the observed streamflow records are non-negatives. Moreover, proportional adjusting is useful when disaggregate data include intermittent process, zero values between non-zero values. If a disaggregate value is zero, then the proportional adjusting retains zero unlike linear adjustment. The power adjusting procedure is useful in that it is the generalization of the proportional adjustment. But additional repetitions are required to meet the additivity condition since this will not initially preserve the additivity condition and it slows the generation procedure

(Koutsoyiannis and Manetas 1996). Therefore, the linear adjustment and the proportional adjustment are employed with different modeling structure in this study.

The lag-1 Contemporaneous Periodic Autoregressive (CPAR(1)) model is applied to the lower-level data (e.g. monthly) in the paper (Koutsoyiannis and Manetas 1996). Specifically, they propose this model to preserve the skewness coefficient employing the parameterization into the random components. The parameterization for the CPAR(1) model with embedded skewness parameterization is still cumbersome, and easily generates negative values. Since CPAR(1) only accounts for lag-1 serial correlation, the long-term monthly correlation cannot be preserved. Therefore, as Koutsoyiannis and Manetas (1996) mentioned, the CPAR(1) model is not appropriate in cases where the snow-melt dominates streamflows, such as in the Colorado River System. And the lagged cross-correlations between sites are underestimated, since it approximates the dependent structure in variance-covariance matrix during parameter estimation procedure.

4.2.3 Nonparametric Disaggregation model

Tarboton et al.(1998) invented a nonparametric disaggregation (NPD) approach. The NPD model employed the nonparametric conditional density estimate as

$$f(\mathbf{Y} \mid X) = \frac{f(\mathbf{Y}, X)}{\int f(\mathbf{Y}, X) d\mathbf{Y}} \tag{4-7}$$

The coordinates of the disaggregate variable vector are rotated into a new vector space $\mathbf{Z} = (Z_1, ..., Z_d)$ using the rotation matrix ($\mathbf{R}_{d \times d}$) obtained from the Gram Schmidt orthonormalization (GSO) such as:

$$\mathbf{Z} = \mathbf{RY} \tag{4-8}$$

The rotation estimation procedure of GSO guarantees that the last coordinate of a new vector space is aligned perpendicular to the hyperplane in Eq.(4-3). And, the last elements of the rotated variable Z_d are the rescaling of X such as:

$$Z_d = X / \sqrt{d} \tag{4-9}$$

GSO procedure is described in Appendix A with the example on d=2. Tarboton et al. (1998) used the multivariate density estimate of the rotated variable $\mathbf{Z}$. Generating variable X separately from a desirable model, the multivariate kernel density estimate is used to generate the Z variables ($Z_1,...,Z_{d-1}$) with the condition of Z_d, which is obtained from variable X as of Eq.(4-9). Then the generated variables are back rotated by:

$$\mathbf{Y} = \mathbf{R}^{-1}\mathbf{Z} = \mathbf{R}^{T}\mathbf{Z} \tag{4-10}$$

to obtain the original disaggregate level data, where $\mathbf{R}^{-1} = \mathbf{R}^{T}$, using a standard basis (see appendix A).

Since the burdensome feature of the NPD procedure is using the d-dimensional multivariate density estimate, Prairie et al. (2007) employed the k-nearest neighbor bootstrapping (KNNB) technique (Lall and Sharma 1996). KNNB is used to select the $\mathbf{Z}$ variables obtained from the rotated historical data in place of generating from a multivariate density estimate.

The generation procedure of the disaggregation model (Prairie et al. 2007) is summarized as:

(1) Estimate the **R** matrix according to the number of disaggregate variables referring to Eq.(4-A1) and (4-A2), and obtain the $\mathbf{Z}_\nu = (Z_{\nu,1},...,Z_{\nu,d})$ variables from the historical variables as Eq.(4-9), where $\nu = 1,...,N$ and N is the record length.

(2) An aggregate value is generated from the selected model such as ARMA(p,q),(Salas 1980), the modified KNN (Prairie et al. 2006) or KNN bootstrapping model (Lall and Sharma 1996), called X^*.

(3) K-nearest neighbors are obtained from the distance between $Z_d^* = X^* / \sqrt{d}$ and $Z_{\nu,d}$ ($\nu = 1,...,N$). In other words, the K-closest values of $Z_{\nu,d}$ to Z_d^* are chosen among N number of $Z_{\nu,d}$. The K-neighbors are assigned weights as:

$$w(i) = \frac{1/i}{\sum_{j=1}^{k} 1/j} \qquad i=1,2,..,K \tag{4-11}$$

where the number of nearest neighbor is $K = \sqrt{N}$, and N is the sample size (Prairie et al. 2007). Subsequently, one of the weighted K-neighbors is randomly selected. The selection among K-neighbors is the random generation from the discrete weighted distribution from one to K and their probabilities to be selected are as of Eq.(4-11). Roulette sampling can be also applied for this generation (Goldberg 1989).

(4) The remaining elements from $\mathbf{Z}^*$, $Z_1, Z_2, ..., Z_{d-1}$ are taken from the corresponding values to Z_d from step (3). For example, if the j[th] year is selected, set $\mathbf{Z}^* = (Z_{j,1}, Z_{j,2}, ..., Z_{j,d-1}, Z_d^*)$ where $Z_{i,j}$ is the rotated historical value at j[th] year among N years and i[th] variable among d variables. And notice that d[th] element of $\mathbf{Z}^*$ is replaced with Z_d^* obtained from step (3).

(5) Back-rotate the $\mathbf{Z}^*$ vector to original space as:

$$\mathbf{Y}^* = \mathbf{R}^T \mathbf{Z}^*$$

(6) Steps (2) to (5) are repeated until the generation length is met.

This procedure is mathematically investigated for a two dimensional case. Mathematically (readers are referred to Appendix A), Z is described as:

$$\mathbf{Z} = \mathbf{RY} = \begin{pmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{d} & -1/\sqrt{d} \\ 1/\sqrt{d} & 1/\sqrt{d} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Y_1/\sqrt{d} - Y_2/\sqrt{d} \\ Y_1/\sqrt{d} + Y_2/\sqrt{d} \end{pmatrix}$$

And, for replacing Z_d with $Z_d^* = X^*/\sqrt{d}$ X^*, $\mathbf{Z}^* = \begin{pmatrix} Y_1/\sqrt{d} - Y_2/\sqrt{d} \\ X^*/\sqrt{d} \end{pmatrix}$

From back rotation $\mathbf{Y}^* = \mathbf{R}^T \mathbf{Z}^*$

$$\mathbf{Y}^* = \mathbf{R}^T \mathbf{Z}^* = \begin{pmatrix} 1/\sqrt{d} & 1/\sqrt{d} \\ -1/\sqrt{d} & 1/\sqrt{d} \end{pmatrix} \begin{pmatrix} Y_1/\sqrt{d} - Y_2/\sqrt{d} \\ X^*/\sqrt{d} \end{pmatrix} = \begin{pmatrix} \frac{Y_1 - Y_2}{d} + \frac{X^*}{d} \\ \frac{Y_2 - Y_1}{d} + \frac{X^*}{d} \end{pmatrix}$$

By applying d=2,

$$\mathbf{Y}^* = \begin{pmatrix} Y_1 + (X^* - X)/2 \\ Y_2 + (X^* - X)/2 \end{pmatrix} \tag{4-12}$$

From Eq.(4-12), it is noticed that the disaggregation procedure equally distributes the difference between the simulated aggregate value X^* and the historical aggregate value X for each disaggregate values. Further investigation has been performed for the higher dimension of d, not included here. Logically, the procedure can be described differently as:

With the simulated aggregate value X^*, find the historical disaggregate data set $\mathbf{Y} = (Y_1,...,Y_d)$ whose sum is close to X^* employing K-nearest neighboring approach (Lall and Sharma 1996).

The selected historical disaggregate data set is adjusted as

$$\mathbf{Y}^* = \begin{pmatrix} Y_1 + (X^* - X)/d \\ Y_2 + (X^* - X)/d \\ \vdots \\ Y_d + (X^* - X)/d \end{pmatrix} \tag{4-13}$$

This simplified procedure is exactly the same as one from Prairie et al. (2007). This procedure, however, confronts significant consequences as: (a) The generated values might have negative values which are physically impossible to be produced in the directly measured records. Furthermore, it is not applicable for an intermittent streamflow case (zero values between non-zero sequences); (b) If the variance of the disaggregate

variables are significantly different from each other, overestimation of the variability in the disaggregate variables that have small variation might be yielded. And also, the variable with higher variance will be no different from the historical values. This might lead to overestimating minimum and underestimating maximum values. (c) In temporal disaggregation, the disaggregate values are selected only with the condition of the aggregate value of the current year. Subsequently, the correlation between the last month of the previous year and the first month of the current year cannot be preserved.

4.3 Model Description

The two disaggregation approaches presented are similar in that the distance between the aggregate generated value and the summation of the disaggregate generated values are estimated. And, the NPD with KNN (NPDK) performs the linear adjusting procedure (Eq.4-4) with the scaling factor $\lambda_j = 1/d$ for all disaggregation values (j=1,…,d). The major difference is that the NPD uses the KNNB technique (Lall and Sharma 1996) to find a close set of lower-level generated data values whose sum is close to the aggregate value while AAP employs the repetition process. The use of KNNB allows the NPDK model to capture nonlinear data distributions, which the AAP model cannot.

From the investigation of the two disaggregation models in the previous chapter, we propose an algorithm that is able to surmount the shortages of both models such as: (1) not reproducing the long-term monthly correlation from CPAR(1) and degrading the cross-correlation from contemporaneous modeling in AAP, (2) not preserving the

correlation between the first moth of the current year and the last month of the previous year in temporal disaggregation and overestimating the variability in sites with relatively small variation in respect to lower-level data by NPDK, and (3) generating negative values which is not physically plausible for streamflow observation by both AAP and NPDK. Those shortcomings are visibly revealed in the results section.

The main procedure proposed in this paper is the combination of two disaggregation models, AAP and NPD. The KNN matching process is employed to find the candidate disaggregate values, whose sum is close to the generated aggregate value. The adjusting procedure is followed to meet the additivity condition. Since the current disaggregate values are not connected to the previous disaggregated values of the previous year, a remedy for the linkage is made by including the last disaggregate value of the previous year in the lower level sequence selection. Since the current selection algorithm of disaggregate variable only reproduce the same historical pattern in a year, Genetic Algorithm mixture of the disaggregate variables is applied here suggested by Lee and Salas (2008). Overall, the objective of the disaggregation model development in this paper is to develop an algorithm that disaggregates the higher level data preserving both lower and higher level statistics of the historical data and generating new sequences with new seasonal and spatial patterns, as well as new values, not present in historical data.

4.3.1 Combination of the NPD and adjusting procedure

A combination of the NPD and APP disaggregation models is suggested to surmount the shortages of each model. The combined procedure first models the aggregate variable and generating independently and from the historical disaggregate

sequences of which summation is close to the generated aggregate value employing KNNR. Finally, the selected disaggregated values are adjusted to meet the additivity condition. The specific model procedure is as follows:

(1) Generate an aggregate series from a selected model (e.g. KNNR (Lall and Sharma, 1996); the modified K-NN (Prairie et al., 2006); Shifting Mean Level (Sveinsson et al., 2003); ARMA).

(2) The distance between X_t^* and the historical higher-level data X_v is estimated where X_v is the historical aggregate value at year v. The distance is

$$\Delta_v = \left| X_t^* - X_v \right| \qquad v = 1,...,N \qquad (4\text{-}14)$$

(3) Among the smallest K-values of Δ_v, where $K = \sqrt{N}$ (Prairie et al., 2007), one is selected with the random generation from the discrete weighted distribution (with the range from 1 to K) and their probabilities to be selected are the weights as Eq.(4-11). If the j^{th} year is selected, the corresponding historical disaggregate values are assigned as the primary generated disaggregate value $\widetilde{\mathbf{Y}}_t = \mathbf{Y}_j = \{Y_{j,1}, Y_{j,2}, ..., Y_{j,d}\}$. The disaggregate sequences whose sums are closer to X^* have a higher probability to be chosen and vice versa according represented in the weights from Eq.(4-11).

(4) Then, the selected historical lower-level dataset $\widetilde{\mathbf{Y}}_t$ is adjusted with a linear or a proportional adjusting procedure as in Eq.(4-4) or (4-5) to obtain the generated disaggregate set $\mathbf{Y}^* = \{Y_1^*, Y_2^*, ..., Y_d^*\}$ whose sum is equal to X^* from step(1). If a

linear adjustment is used, this disaggregation model is called Disaggregation with KNN and linear adjustment (KLA). And for the proportional adjustment, it is called KPA.

(5) Steps (1) through (4) are repeated until the generation length is met.

The number of nearest neighbors (K) can also be obtained with generalized cross validation. The heuristic method, $K = \sqrt{N}$, has performed also well in the applications (Lall and Sharma, 1996; Yates et al. 2003). K takes important role for the variability of the resampled sequences. When K is smaller, the similar aggregate value of the historical to the generated aggregate variable will be obtained. However, the problem is that the variability of the disaggregate variable over the similar generated aggregate variable gets smaller.

The suggested model in the previous section is not able to preserve the correlation between the disaggregate variables of the current year and those of the previous year. The same problem occurred when the first parametric disaggregation model had been developed by Valencia and Schakke (1973) as described already. A remedy to link the past with the disaggregate values by Mejia and Rousselle (1976) was to include the additional term for the disaggregate value of the previous year.

In this nonparametric disaggregation process, it is easier to include the condition of the last month of the previous year. It only requires replacing the distance measurement in Eq.(4-14) with:

$$\Delta_\nu = \sqrt{\varphi_1\left(X_t^* - X_\nu\right)^2 + \varphi_2\left(Y_{t-1,d}^* - Y_{\nu-1,d}\right)^2} \tag{4-15}$$

where $\nu = 2,...,N$ and $Y^*_{t-1,d}$ is the last generated disaggregate value of the previous generate year (t-1) and $Y_{\nu-1,d}$ is the previous historical disaggregate value of the previous year for year ν. And φ_1 and φ_2 are scaling factor, and since the distance is measured with two different variables, the inverse of variance for each variable are generally employed such as $1/\sigma_X^2$ and $1/\sigma_{Y_d}^2$, respectively(Buishand and Brandsma 2001). This inclusion of the additional term allows preserving the relation between the last month of the previous year and the first month of the current year. Instead of the weighted Euclidean distance, Mahalanobis distance also can be applied (Wójcik et al. 2000). Notice that this distance measurement in Eq.(4-15) is applicable only to temporal disaggregation. Spatial disaggregation does not require this procedure. Instead, the distance in Eq.(4-14) should be use. Even though, the suggested disaggregation approach is explained with the focus on the temporal disaggregation, it is basic to expand the process to comply with the spatial disaggregation.

4.3.2 Mixing with Genetic Algorithm

The suggested model, however, has a critical drawback because the repetitive seasonal patterns from the generated data might lead to a significant mistake in decision making. The repetitive seasonal patterns occur because during the selection procedure from KNN (step(3), chapter 3.1), the entire disaggregate sequence is selected as a block. This argument was previously discussed in Lee and Salas (2008), Porter and Pink (1991), and Srikanthan and Mcmahon (1980). The seasonal repetition is not desirable in that the purpose of stochastic simulation is to analyze the frequency of certain critical events, such as floods or droughts including their pattern which is unprecedented in the historical

dataset. In the paper (Lee and Salas 2008), the mixing process with the Genetic Algorithm has been proposed to overcome this problem. Here we also include this process in the disaggregation algorithm to avoid generating the same pattern as the historical. The cross-over algorithm is only GA process used among the three process, reproduction, crossover, and mutation.

In the selection procedure from the k-values at step (3) of the previous section 3.1, another disaggregated sequence is selected , $\widetilde{\mathbf{Y}}_t^{\,2} = \mathbf{Y}_i = \{Y_{i,1}, Y_{i,2}, ..., Y_{i,d}\}$ assuming that the i^{th} year is selected, so that two sets of lower-level data are chosen, denoted as $\widetilde{\mathbf{Y}}_t^{\,1} = \mathbf{Y}_j = \{Y_{j,1}, Y_{j,2}, ..., Y_{j,d}\}$ and $\widetilde{\mathbf{Y}}_t^{\,2} = \mathbf{Y}_i = \{Y_{i,1}, Y_{i,2}, ..., Y_{i,d}\}$. The cross-over process of the Genetic Algorithm is performed with either random or competition selection. The random selection chooses one of two values for each lower-level data with equal probability while the competition selection chooses the one value having the better statistical characteristics, such as preserving the serial correlation better. For example, the random selection is performed with the disaggregate values by selecting $\widetilde{Y}_{t,l}^{1}$ or $\widetilde{Y}_{t,l}^{2}$ with equal probability where l=1,..,d. This can be done by generating a uniform random number (u), and if $u \le 0.5$ choose $\widetilde{Y}_{t,l}^{1}$ otherwise choose $\widetilde{Y}_{t,l}^{2}$. The competition selection is employed to increase the serial correlation of the generated data. The serial correlation will be increased by one of two values for which the correlation gets higher. In case that:

$$\left[\frac{\widetilde{Y}_{t,l}^{1} - \mu_{Y_l}}{\sigma_{Y_l}} \cdot \frac{Y_{t,l-1}^{*} - \mu_{Y_{l-1}}}{\sigma_{Y_l-1}}\right] \le \left[\frac{\widetilde{Y}_{t,l}^{2} - \mu_{Y_l}}{\sigma_{Y_l}} \cdot \frac{Y_{t,l-1}^{*} - \mu_{Y_{l-1}}}{\sigma_{Y_l-1}}\right] \tag{4-16}$$

choose $\widetilde{Y}_{t,l}^{2}$, otherwise choose $\widetilde{Y}_{t,l}^{1}$ where l=1,..,d.. And the other steps (4) and (5) are the same as the previous chapter 3.1 as well as step (1) through (2).

4.4 Data description and Model Assessment

To verify the suggested model, the Colorado River system is utilized. The Colorado River System is represented with 29 selected stations. The historical gaged data has been naturalized (Prairie and Callejo, 2005; http://www.usbr.gov/lc/region/g4000/NaturalFlow/index.html) for these 29 stations from 1906 to 2003. Part of the data has been extended by Lee and Salas (2006) back to 1906 employing the combination of the parametric linear regression and the nonparametric bootstrapping with trace selection method. The locations of the 29 stations are shown in Figure 4-1.

The temporal and spatial disaggregations are tested separately. For temporal disaggregation, site 20 (Lees Ferry) yearly and monthly data have been used to validate the performance of the suggested model and compare to the model of Prairie et al. (2007). In Table 4-1, the basic monthly and yearly statistics of the historical data are presented. The last row of Table 4-1 illustrates the ratios of standard deviation for each month divided by the yearly data. This value indicates the percent of variance each month has over the total yearly variance. The months in wet seasons (JJA) explain most of the yearly variance while the months in dry seasons contribute little to aggregate (yearly) variance. The KNN with the Gamma kernel density estimate perturbation method (KGK) developed by Lee and Salas (2008) is employed for the yearly data simulation, called X^{*}. The simulated yearly data have been disaggregated with NPDK and the suggested models

in this work. Five types of the models are tested: (1) Nonparametric Disaggregation model with KNN (NPDK). (2) Temporal Disaggregation with KNN selection and the linear accurate adjusting procedure – KLA; (3) Temporal Disaggregation with KNN selection and the linear accurate adjusting procedure with Genetic Algorithm- KLAG; (4) Temporal Disaggregation with KNN selection and the proportional accurate adjusting procedure – KPA; (5) Temporal Disaggregation with KNN selection and the proportional accurate adjusting procedure with the Genetic Algorithm - KPAG.

To demonstrate spatial disaggregation, we use the tributary sites of the lower Colorado River System (sites 21, 22, 24, and 27). An index station is used whose streamflow value is the summation of these four sites. This index station is necessary for the additivity condition in spatial disaggregation. The monthly data of this index station are obtained from the temporally disaggregated data with proportional adjusting and genetic algorithm mixing and the yearly data are generated from KGK (Lee and Salas, 2008) model.

Two model schemes are tested for spatial disaggregation: (1) Spatial Disaggregation with Gram-Schmidt orthonormal rotation and KNN selection – NPDK and (2) Spatial Disaggregation with KNN selection and the proportional accurate adjusting procedure – KPA. Since the linear adjustment procedure employed in the suggested model may produce negative values especially in highly skewed data, only the proportional adjustment model is applied for testing. The former model is the disaggregation model of Prairie et al. (2007). Tributary sites of the lower Colorado River Basin (sites 21, 22, 24, and 27) were chosen to investigate model performance for the intermittent case such that some months have no streamflow (zero value). Those tributary

sites are arid regions where sudden thunderstorms are the main sources of the streamflow. The data are highly skewed, not only in the monthly scale but in the yearly scale and the key yearly statistics are shown in Table 4-2. The ratios in the last row of Table 4-2 signify that site 21 typically has the lowest contribution among those four disaggregate sites.

One hundred sets of the generated series with the same length as the historical data are generated from each model. Various key statistics are estimated from the historical and generated data to verify the model performance such as mean, standard deviation, skewness, maximum and minimum, and lag-1 serial correlation at the seasonal and yearly time scale. A boxplot is employed to show the estimated statistics from the generated data. The end line of the box indicates the 25 and 75 percent quantile while the cross line above the box on the whisker denotes the 90 percent quantile and maximum, and the cross line below the box on the whisker denotes the 10 percent quantile and minimum. The segment line with the 'x' or 'o' mark presents the historical statistics. The preservation of the cross or serial relation in the generated data is checked through using a scatterplot. The generated data sets (100 sets) are used as well as the historical data, to display the shape of two relations, such as temporal and spatial relations. Furthermore, drought statistics with the historical and generated data are compared with the boxplot for yearly and monthly data. The employed drought statistics are the maximum drought and surplus amount, the maximum drought and surplus length, and storage capacity with the historical mean water demand multiplying various demand levels from 0.6 to 1.0.

4.4.1 Temporal Disaggregation

From the five indicated temporal disaggregation models, the various test statistics are estimated and compared for the suggested and existing nonparametric disaggregation model. The results are followed. As mentioned, the aggregate variable is modeled with KGK. The basic and drought statistics of the historical and generated (SM model) yearly data are presented in Figure 4-C.1 and Figure 4-C.2. The basic statistics of the simulated and historical data in the lower-level are shown in Figure 4-2 and Figure 4-3 for NPDK and KLA, respectively. The minimum and lag-1 correlation of KLAG and KPAG are shown in Figure 4-4. Full results of the basic statistics for KLAG, KPA, and KPAG are shown in Figure 4-C.3, Figure 4-C.4, and Figure 4-C.5. In Figure 4-2, the characteristics of the NPDK model (Prairie et al. 2007) is well presented there. The first significant aspect of the figure is the underestimation of the lag-1 serial correlation of the first month since the model has no model structure to link the past of the previous year with the disaggregate values. This shortcoming will be easily fixed by adding one more term in the disaggregate value selection in the suggested model, as in Eq.(4-15). The improvement of this feature on the suggested models such as KLA, KLAG, and KPA is clearly shown in the lag-1 correlation of Figure 4-3 and Figure 4-4. The correlation between the last month of previous year and the first month of the current year is fairly well preserved in the suggested models compared to NPDK model. The slight underestimation of the lag-1 correlation can be observed for the KLAG model in Figure 4-4 (also in KPLAG, referred to Figure 4-C.5). The underestimation is because the Genetic Algorithm mixture disturbs the historical correlation with small magnitude and is the price to pay for employing GA mixture. The effect of GA mixture will be discussed

more on the later in this section. Also, the minimum values are sometimes negative even if it infrequently happens in NPDK model as shown in Figure 4-2. That is because NPDK uses linear adjustment with $\lambda_j=1/d$ and $j=1,\ldots,d$ as in Eq.(4-4). The months with low variability (NDJF) will be highly affected resulting in negative values. The KLA and KLAG models, however, do not produce any negative values in this case as in Figure 4-3 and Figure 4-4 since the difference of the historical and generated yearly are proportionally distributed with the contribution of covariance. Linear adjustment is not preferable when the data is highly skewed because in that case negative values are highly likely. The KPA model guarantees that no negative values will be generated unless there are negative values in the aggregate variable or in the historical data. As shown in Figure 4-4, the minimum value is better preserved with the KPA model. Thirdly, the minimum values in low flow months are underestimated, such as months NDJF, while some overestimation is observed in the higher flow months MJJ. Generally, overestimation of the minimum occurs when the simulation model cannot reach the historical minimum in generation and underestimation of the minimum occurs when the model has higher variability than the historical data. This is the nature of the NPDK model since the difference between the generated aggregate value and the summation of the selected disaggregated values is distributed equally without considering the degree of the variability of individual lower-level variables as shown in Table 4-1. Therefore, the higher flow months are not affected much from the adjustments while the lower flow months are highly affected and result in higher variability. This is the leading factor of the bias in the NPDK model. Some underestimations are shown with the KLA and KLAG models (Figure 4-3 and Figure 4-4) especially for June and July (JJ) with very

rare chance of generating negative values. Those months have higher variance so that the difference from the historical and generated data is weighted into these months. The KPA and KPAG models show better preservation of the minimum values (Figure 4-4 and Figure 4-C.5).

In Figure 4-5, the standard deviation is closely investigated for the NPDK and KLA models. It is obvious that the standard deviation in NPDK is overestimated in the low-flow months, while it is properly reproduced in the KLA model. Underestimation of the variability is not present in high flow months MJJ although the historical mean is slightly above the median; therefore the variability is preserved appropriately with the KNNB procedure. Figure 4-6 indicates the evidence of low variation from the historical data pattern used by NPDK. The scatterplot in Figure 4-6 displays the relationship between month 8 (X-coordinate) and month 9 (Y coordinate), which are high flow months. The 100 generated data sets are marked with gray circles while the triangles represent the historical values. The generated data is always extremely close to the historical values in NPDK model. The scatterplot (Figure 4-6) reveals a weakness of the KLA model. The generated values have directional patterns induced from the linear adjustment. A similar feature is also observed in the KPA model, not shown. The Genetic Algorithm mixture suggested by Lee and Salas (2008) is employed to remedy this weakness. The generated data of the GA applied models on KLA and KPA (KPAG and KPAG) has appropriate spread through the data region while containing the historical relationships (Figure 4-6). The inclusion of the GA induces some underestimation of the lag-1 serial correlation as mentioned (Figure 4-4).

The densities of the historical and disaggregated data set are estimated with normal kernel and the asymptotic optimal bandwidth (Simonoff 1996) shown in Figure 4-7 for Month 5 (February). In NPDK model, the density around the mode is underestimated while it is overestimated outside of mode, especially lower part. This indicates the overestimation of variance as mentioned for the NPDK model in Figure 4-5. Smaller magnitude of underestimation is observed in the KLAG model, which the Genetic Algorithm mixture causes. The density estimate of the other months is relatively well preserved for all models.

Temporal pair cross- correlation of the historical and the models (NPDK, KLA, and KLAG) are shown in Figure 4-8. Some significant overestimation of the cross correlations are revealed for NPDK model (Figure 4-8 (a)), especially Months 1-6 which are low flow months. Notice that this is inconsistent with Figure 6 of the paper of Prairie et al. (2007), but the aggregate variable is generated from Shifting Mean model while in the paper the modified KNN model was employed. A parametric model (e.g. ARMA and SM) might generate values smaller than the historical minimum while a nonparametric model (especially employing resampling technique) is limited to generate the smaller than the historical minimum. Employment of the SM model propagates the variance of the cross correlation in low flow months different from Figure 6 of Prairie et al. (2007). The pair cross correlations are well preserved in the KLA model. Some significant bias is shown for the KLAG model. As mentioned, the GA mixture process disturbs the temporal cross-correlation even if the competitive selection with Eq.(4-16) is used. This is a shortcoming of the GA mixture, but the GA mixture yields more variable sequences

as shown in Figure 4-6 and the long-term persistency is preserved with the aggregate variable.

The ratios of monthly drought, surplus, and storage statistics for the historical and generated data are estimated to represent the long-term variability of time series. The behavior of those statistics depends highly on the aggregate variable (yearly data in this temporal disaggregation). The generated data of the temporal disaggregation model employs the same yearly data as the aggregate variable. In other words, there are not many differences that can be observed on the drought, surplus, and storage statistics even if improved representation is found with the basic statistics of both the KLA and KPA models. Detailed graphs related to these statistics are found from Figure 4-C.6 to Figure 4-C.10.

4.4.2 Spatial Disaggregation

The mean and minimum of the generated and historical data for site 21 (the statistics for site 22 are shown in Figure 4-C.11) are illustrated in Figure 4-9 for NPDK and KPA. The whole statistics are shown at Figure 4-C.13, Figure 4-C.14, and Figure 4-C.15 for NPDK, KPA, and KPAG respectively. Significant biases are observed in all of the basic statistics of the generated data for site 21 from NPDK model (Figure 4-9). That effect is induced from the low-variability relative to the other sites highlighted in Table 4-2. Among the four sites, site 21 has the lowest yearly variance. Also, a significant number of negative values are generated as shown in the minimum of the plot since the monthly streamflow data is highly skewed in this semi arid region for sites 21 and 22. Meanwhile, the key statistics for sites 21 and 22 are well preserved with the KPA model except for a

slight bias in the lag-1 serial correlation and minimum. Similar behavior is obtained in the statistics of yearly time scale for site 21. There is no significant difference for the yearly key statistics between the two models for site 22 (Reader are referred to Figure 4-C.17 and Figure 4-C.20). The cross-correlation of the historical and generated data is presented in Figure 4-10. The KPA model better preserves the cross-correlation than then NPDK model. In particular, the cross-correlation between site 21 and the other stations (site 22, site 24, and site 27) is not reproduced in many months for the NPDK model.

Monthly and yearly drought statistics are estimated for historical and generated data with different demand levels (0.7-1.0). The ratios of the statistics for the generated data divided by historical data are illustrated for monthly (referred to Figure 4-C.23 and Figure 4-C.26) and for yearly (Figure 4-11 and Figure 4-12), respectively. The maximum deficit and surplus amount as well as the storage capacity at 0.7 and 0.8 demand levels of monthly drought statistics for site 21 are highly overestimated in the NPDK model while the KPA model preserves these statistics reasonably well. The same behavior is also indicated in the yearly drought statistics (Figure 4-11 and Figure 4-12). For site 22, there are no significant differences in the monthly and yearly drought statistics between the NPDK and KPA models (Reader are referred to the figures from Figure 4-C.25 to Figure 4-C.28).

4.5 Summary and Conclusions

The stochastic disaggregation modeling is inevitable to analyze critical events such as drought for an entire river system. From reviewing the existing disaggregation models and uncovering the pros and cons of the current models, we suggest a useful

disaggregation model to overcome the shortcomings of these models. The suggested modeling procedures are: (1) to model the aggregate variable independently and generate the sequences; (2) to find two disaggregated data sets whose summation is close to the current aggregate value with KNNB; (3) to cross-over the two data sets with the random or competition selection from GA process and select one data set; and (4) to adjust the selected disaggregated sequences with linear or proportional adjusting according to the characteristics of the historical data set. We recommend the following: (1) in the case of data with high skewness and no negative values, the proportional adjustment should be used, such as the tributary stations of the lower Colorado River System; (2) in the case of data with small skewness and negative values, the linear adjustment is recommended, such as the intervening flows of the Colorado River System (Lee and Salas 2006).

The temporal and spatial disaggregation has been tested using data from the Colorado River System. The testing results indicate that the suggested modeling procedure is reasonable at both sites with lower skewness and sites with high skewness and zero values (intermittent process). The specific conclusion from the results is that the proposed models overcome the drawbacks mentioned in the paper of Prairie et al. (2007) such as the inability to capture the correlation between the first month of the current year and the last month of the previous year and the proper preservation of the extrema (minimum and maximum). The former is overcome by including the variable of the last month of the previous year on KNN selection and the latter is done by the accurate adjusting procedure. Furthermore, the proposed disaggregation models have the ability to model the intermittent and non-intermittent variables jointly with the proportional adjustment. More variable sequences can be obtained using the Genetic Algorithm

mixture. A drawback employing this algorithm is the underestimation of cross correlation. But the aggregate variable holds the dependency structure so that the GA mixture is useful to apply in case more variable sequences are needed.

The disaggregation procedure suggested in this paper can be used differently in the schematic disaggregation. For example, the CRS contains 29 stations. One single direct multivariate model is not a reasonable approach to use for the entire large river basin. The schematic approach would generate data of a few index sites with multivariate modeling, and then these index sites are spatially disaggregated to the tributary stations.

Table 4-1 Basic monthly and statistics of site 20 (month 1~ month 6) Unit: Acre-Feet

	Mon 1	Mon 2	Mon 3	Mon 4	Mon 5	Mon 6	Mon 7
Mean	580893	480821	382530	356611	393775	645201	1199946
Std	272006	141531	95859	78632	97576	211390	512460
Skewness	1.6408	1.2147	1.2225	0.59	1.4188	1.0814	0.9605
Lag-1 Corr	0.5575	0.7577	0.8255	0.7032	0.5515	0.4819	0.4699
$\mu_y/\mu_{X(\%)}$	6.2	3.2	2.2	1.8	2.2	4.8	11.7

	Mon 8	Mon 9	Mon 10	Mon 11	Mon 12	Yearly
Mean	3037199	4054340	2190444	1083174	671371	15076306
Std	1146760	1572353	1012249	423971	309698	4365301
Skewness	0.2713	0.4266	1.1327	0.9464	1.9532	0.1402
Lag-1 Corr	0.5923	0.6251	0.8311	0.7815	0.6373	0.283
$^{\dagger}\sigma_y/\sigma_{X(\%)}$	26.3	36.0	23.2	9.7	7.1	100

† σ_y/σ_X represents the standard deviation of each monthly data over the one of the yearly data

Table 4-2 Basic yearly statistics of tributary sites of lower Colorado River basin (sites 21~24) Unit: Acre-Feet

	Site 21	Site 22	Site 24	Site 27
Mean	21118	180415	169968	98190
Std	8313	140404	88275	125025
Skewness	0.8392	2.0084	1.6774	2.6731
Lag-1 Corr	0.1465	-0.0384	0.0607	0.0608
$\mu_y/\mu_{X(\%)}$	4.5	38.4	36.2	2.1

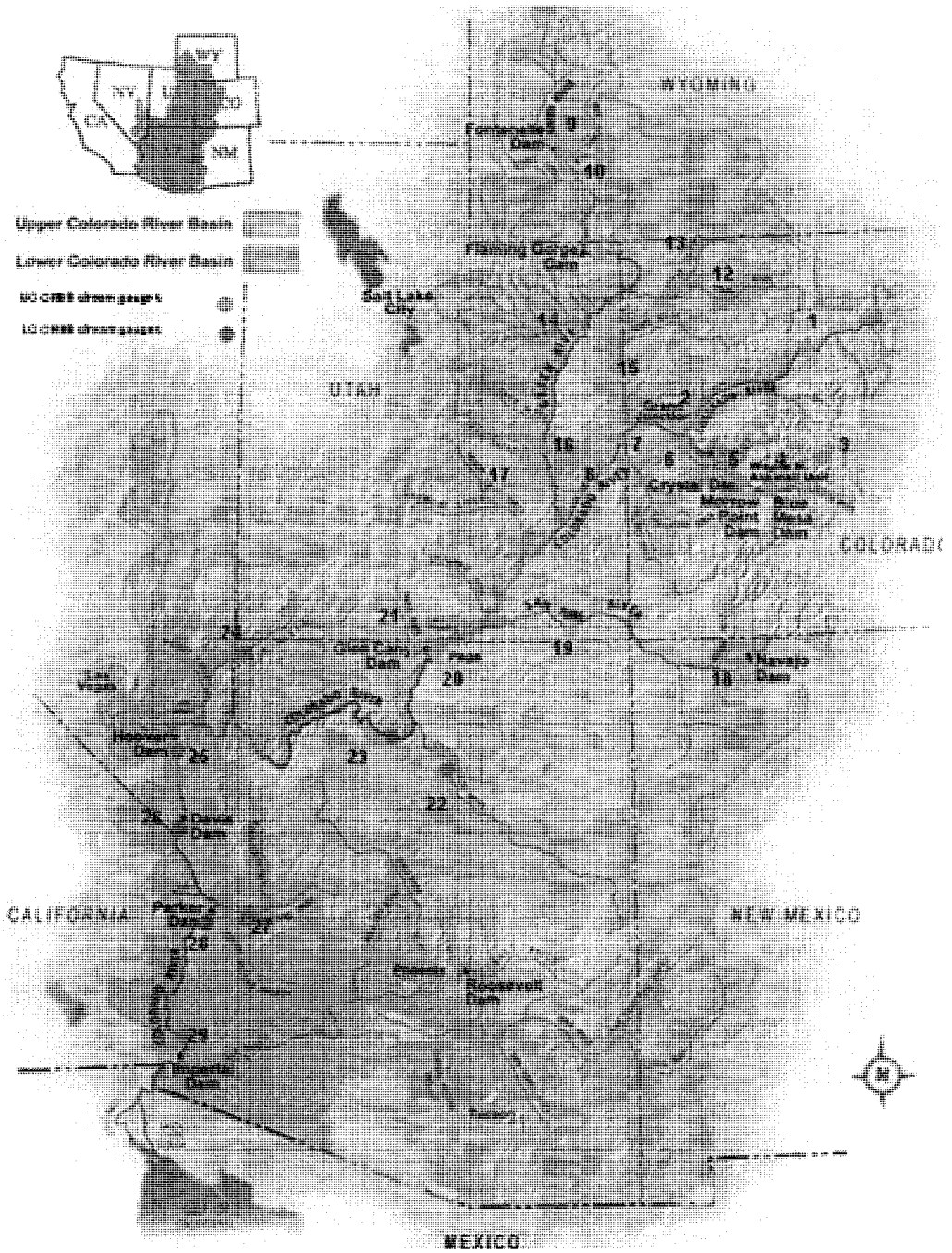


Figure 4-1 Map of Colorado River System with twenty nine stations; the system is divided into two as the upper Colorado River basin (1-20) and the lower Colorado River basin (21-29): Map from Bureau of Reclamation (2007)

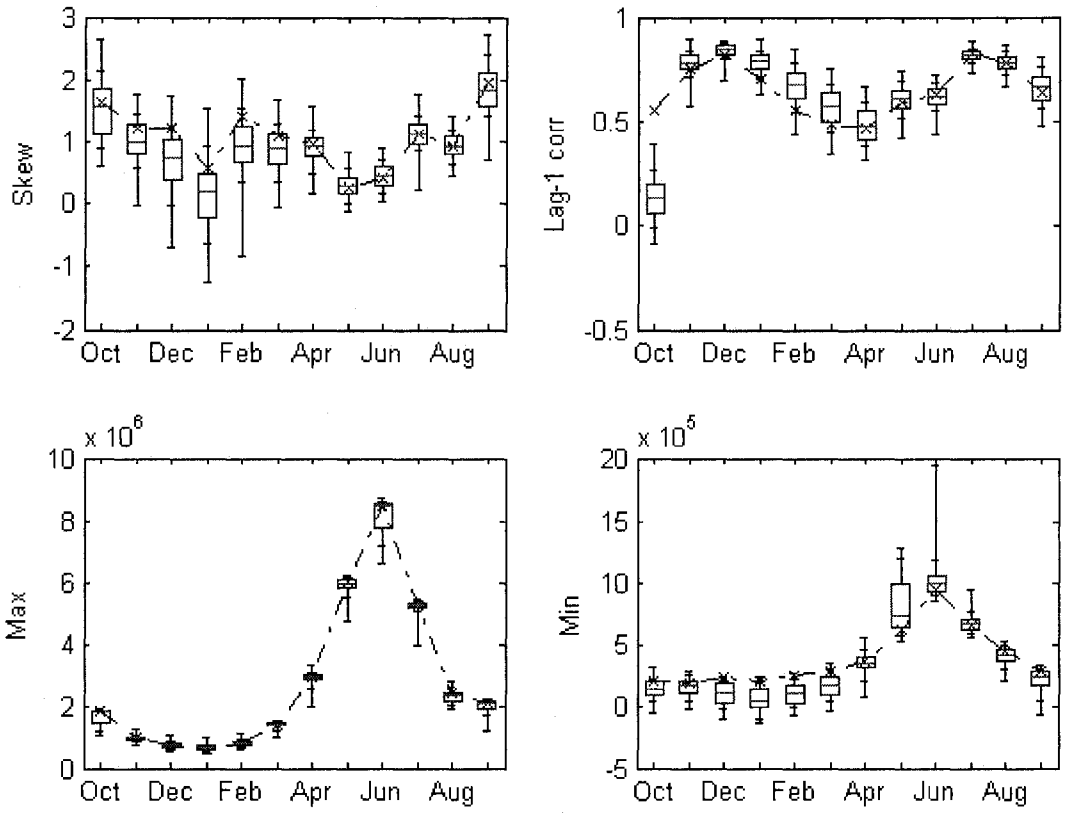


Figure 4-2 Key Statistics of Historical (dot line) and NPDK simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

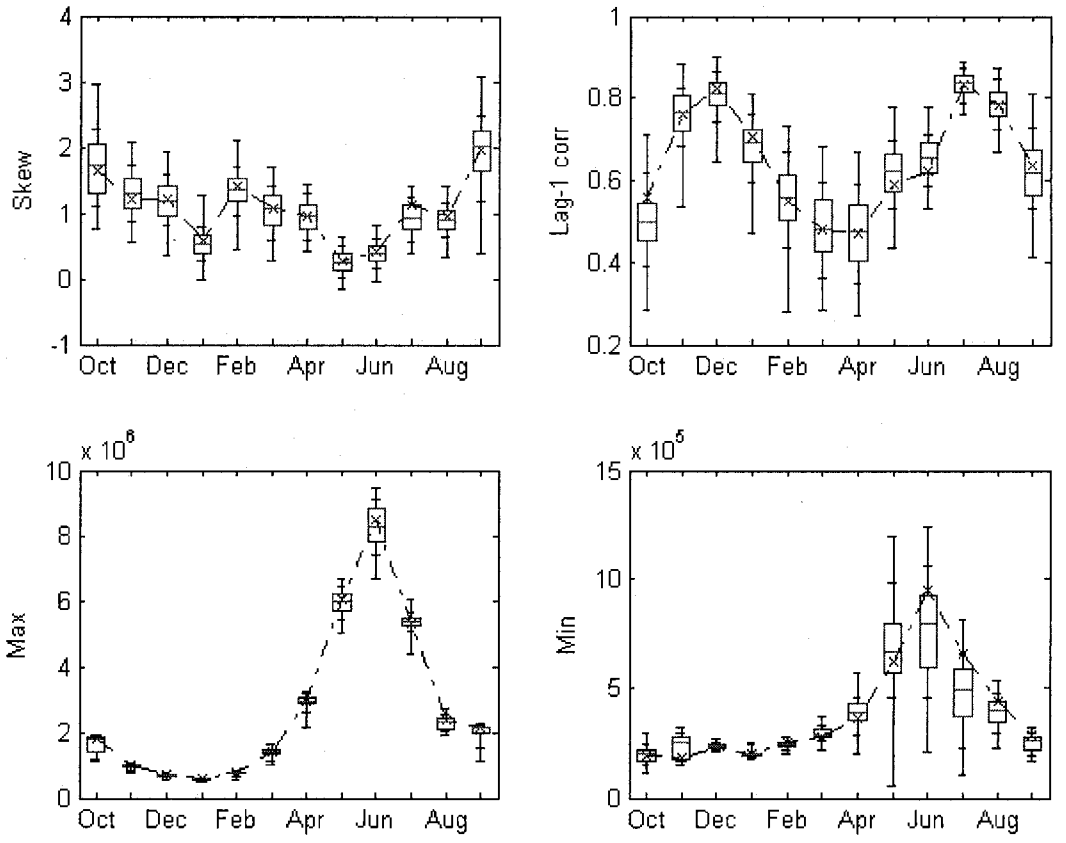
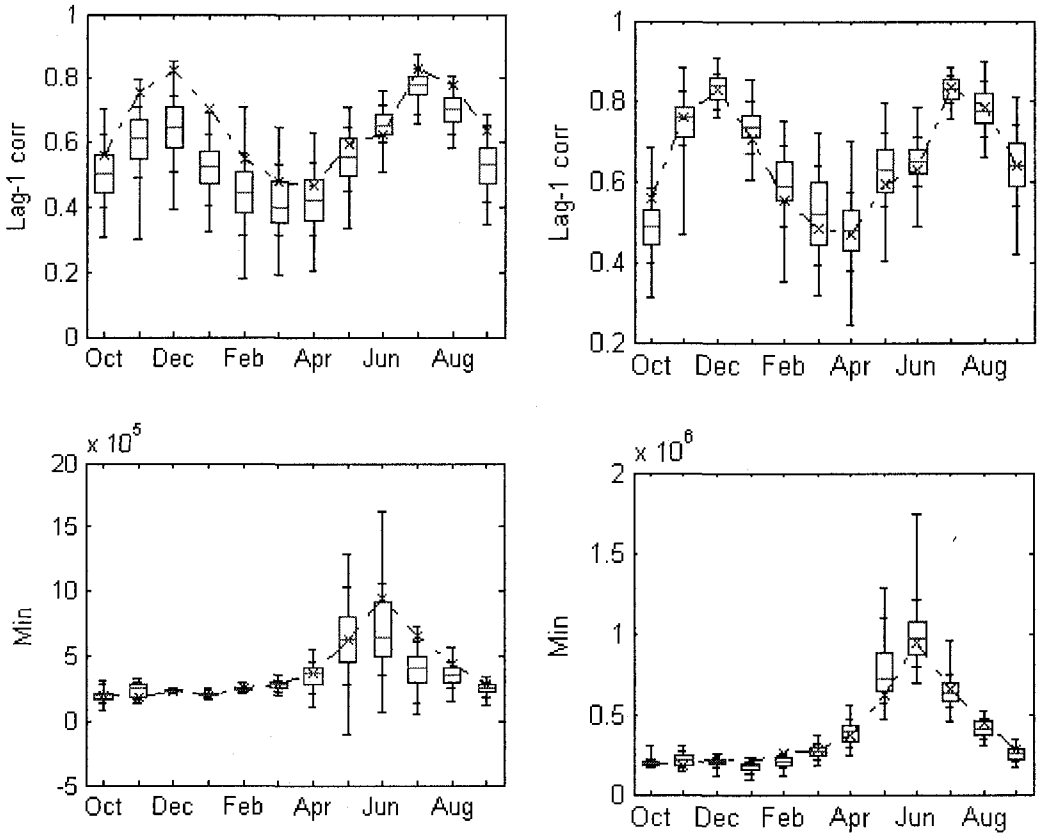


Figure 4-3 Key Statistics of Historical (dot line) and KLA simulations (boxplot) for Site 20 of the Colorado River monthly streamflow



(a) KLAG (b) KPA

Figure 4-4 Lag-1 correlation and minimum of Historical (dot line) and simulated data of KLAG (left) and KPA (right) for Site 20 of the Colorado River monthly streamflow

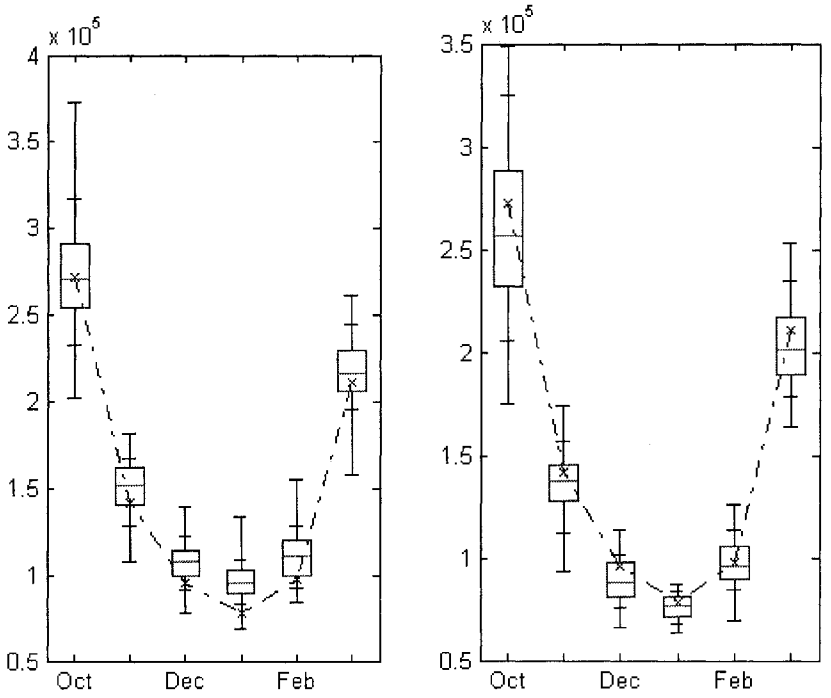


Figure 4-5 Standard deviation of Historical (dot line) and NPDK(left) and KLA (right) simulations (boxplot) for the dry months (October-March) Site 20 of the Colorado River monthly streamflow

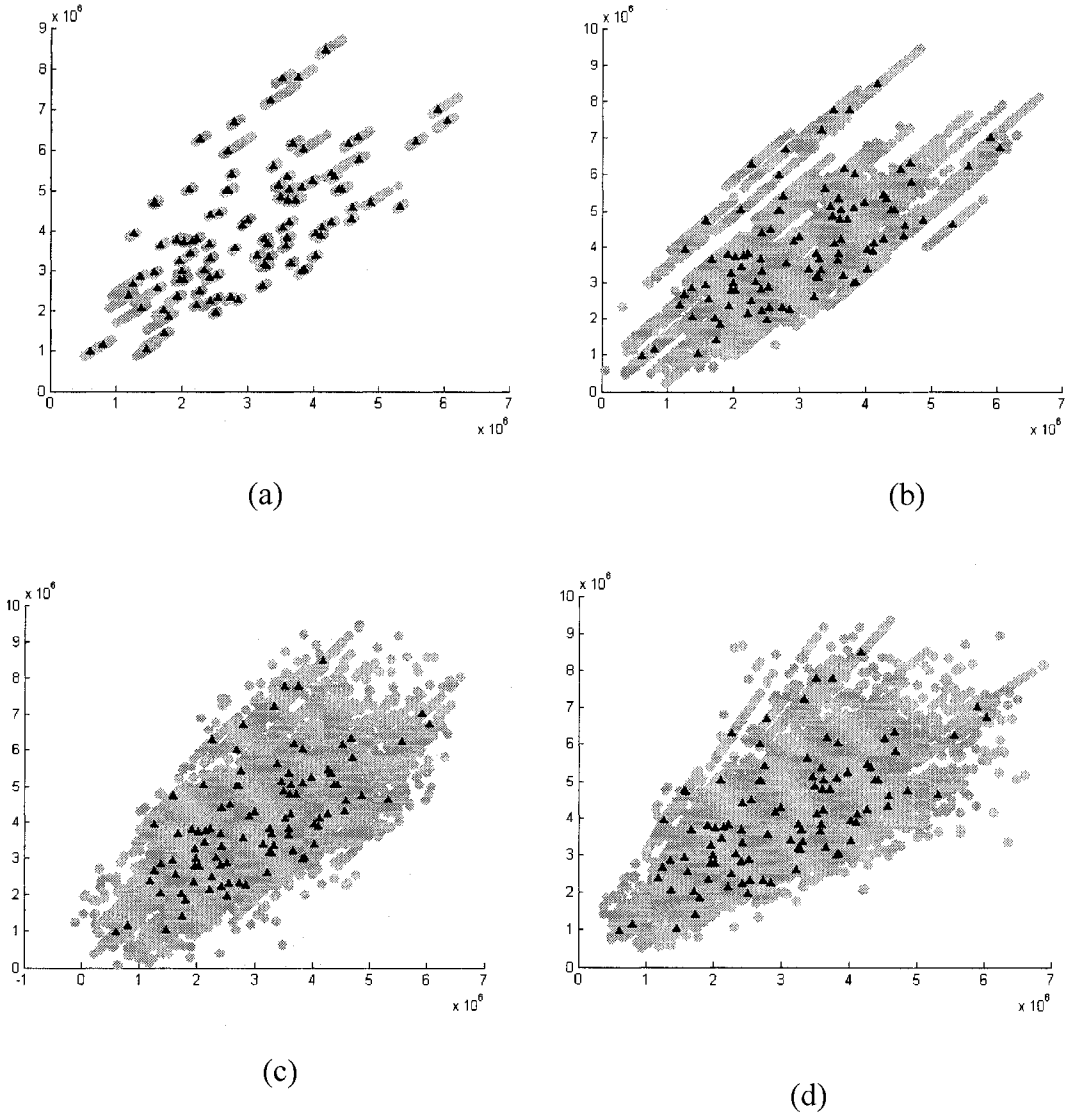


Figure 4-6 Scatterplot from Historical (dot line) and (a)NPDK, (b)KLA, (c) KLAG, and (d) KPAG simulations (boxplot) at the Colorado River monthly streamflow; X-coordinate for $X_{t,Mon8}^{Site20}$ and Y-coordinate $X_{t,Mon9}^{Site20}$

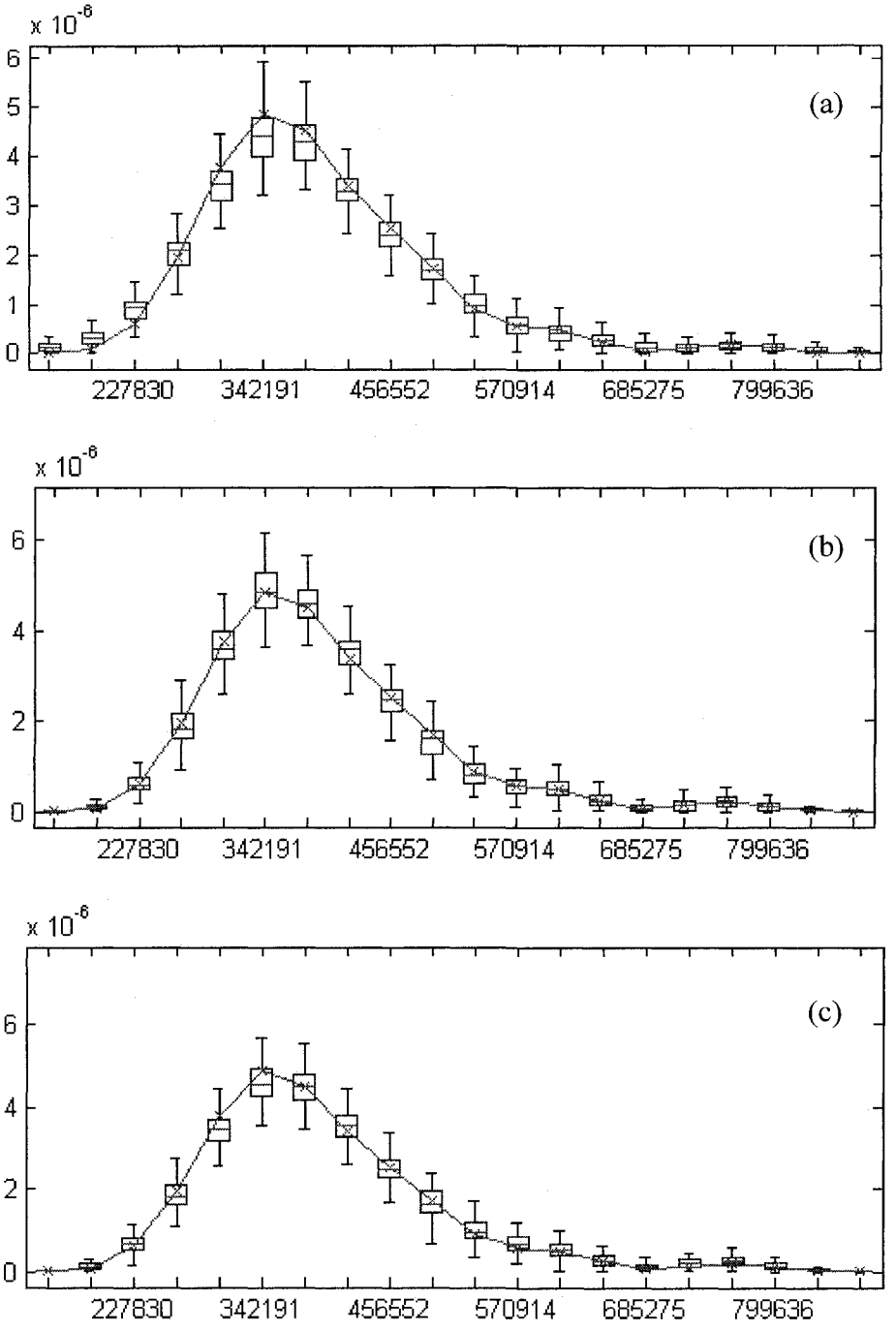


Figure 4-7 Kernel Density Estimate with normal kernel of Month 5 of Station 20 of Colorado River system for (a) NPDK, (b) KLA, and (c) KLAG

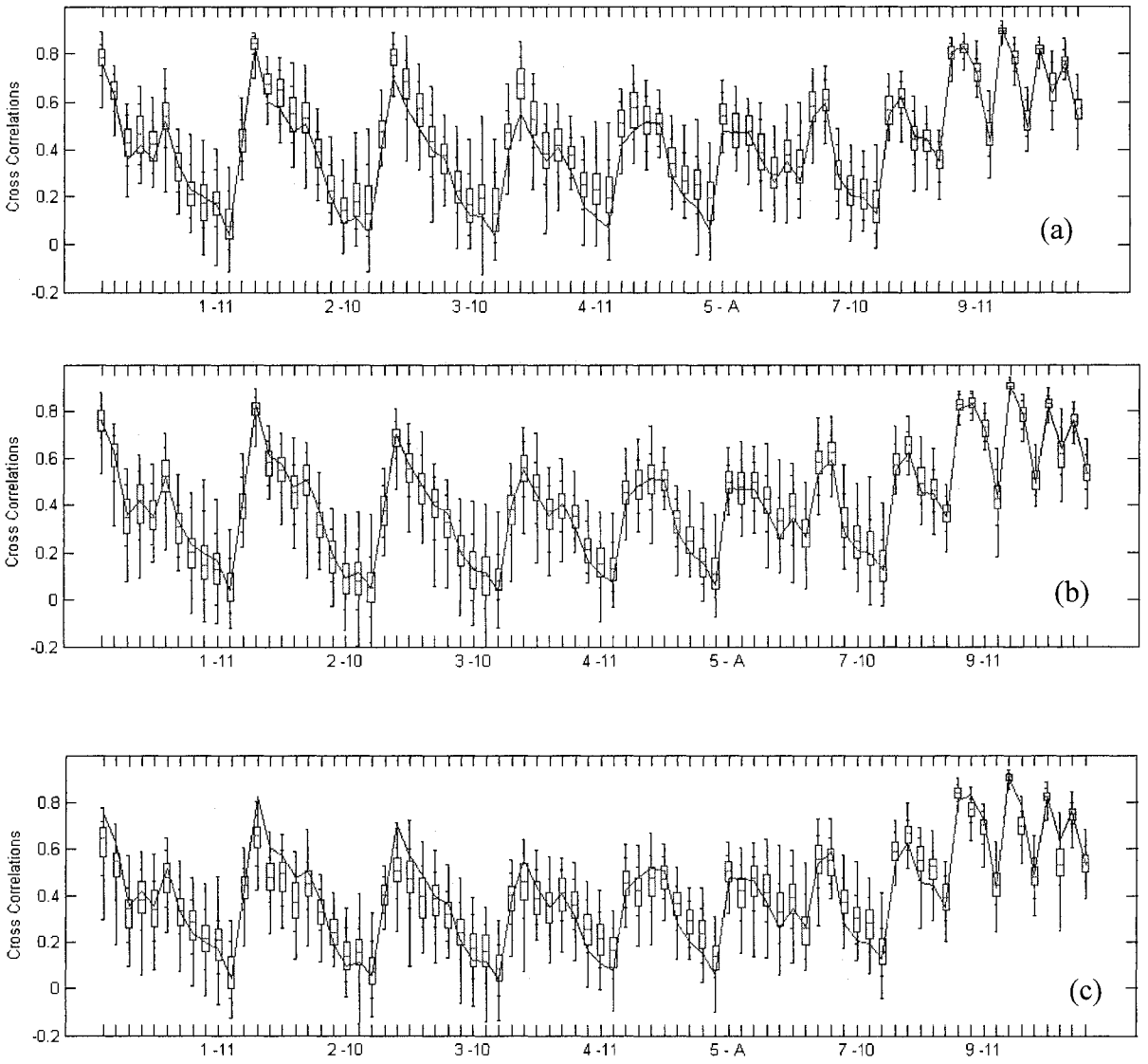
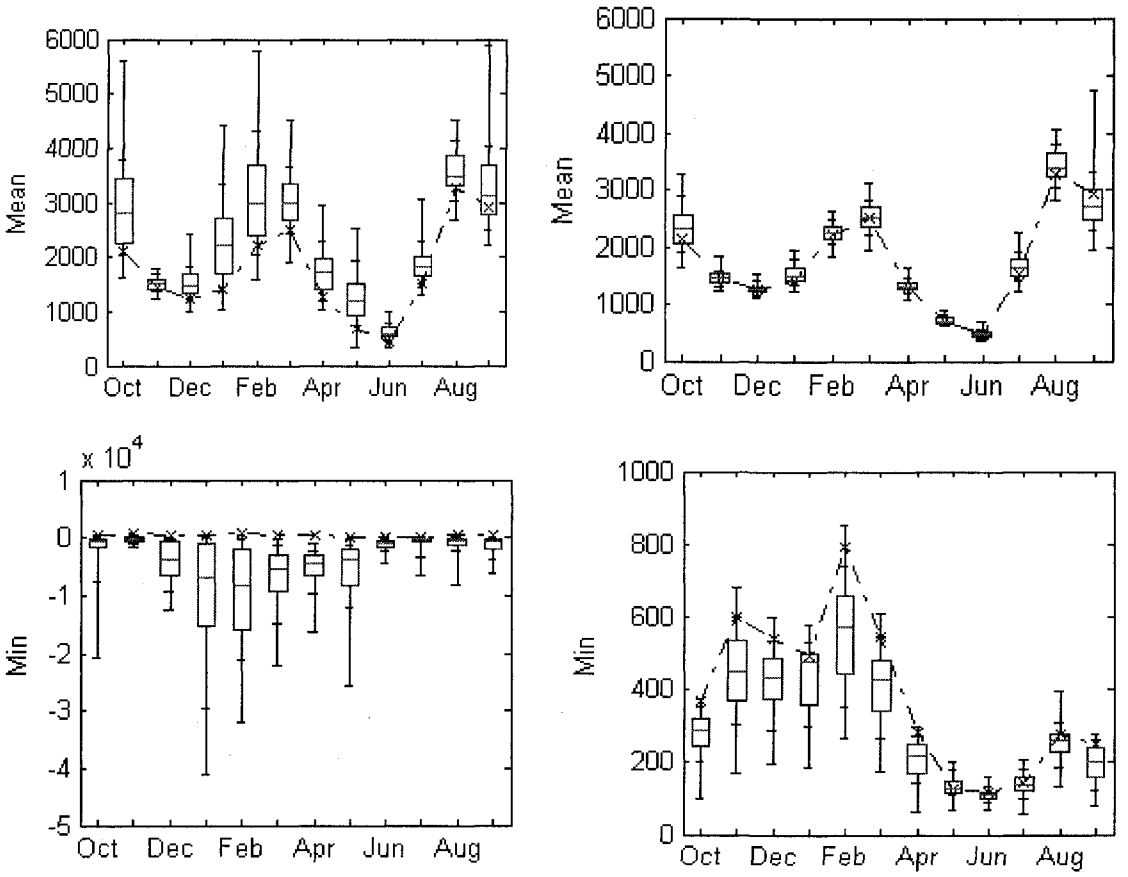
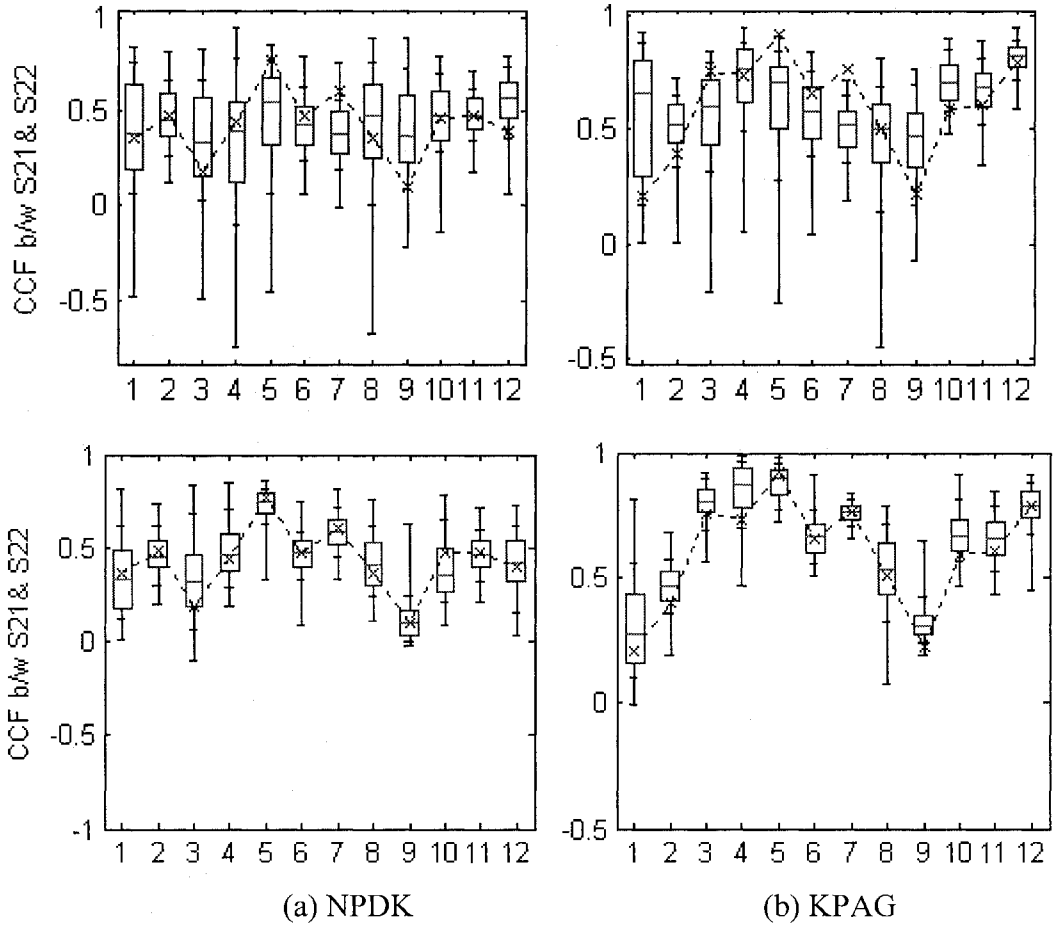


Figure 4-8 Temporal pair cross correlations of historical (segment line) and generated data (boxplot) of (a) NPDK, (b) KLA, and (c) KLAG models for Station 20 of Colorado River. The x-axis sequence is 1-2,1-3,…,1-12,1-A, 2-3,… Months are numbered according to the water year and A represents annual.



(a) NPDK (b) KPAG

Figure 4-9 Key Statistics of Historical (dot line) and (a) NPDK and (b) KPAG simulations (boxplot) for Site 21 of the Colorado River monthly streamflow



(a) NPDK (b) KPAG

Figure 4-10 Lag-0 cross-correlation between sites from the historical (circle) and NPDK and KPAG simulations (boxplot) of the Colorado River monthly streamflow

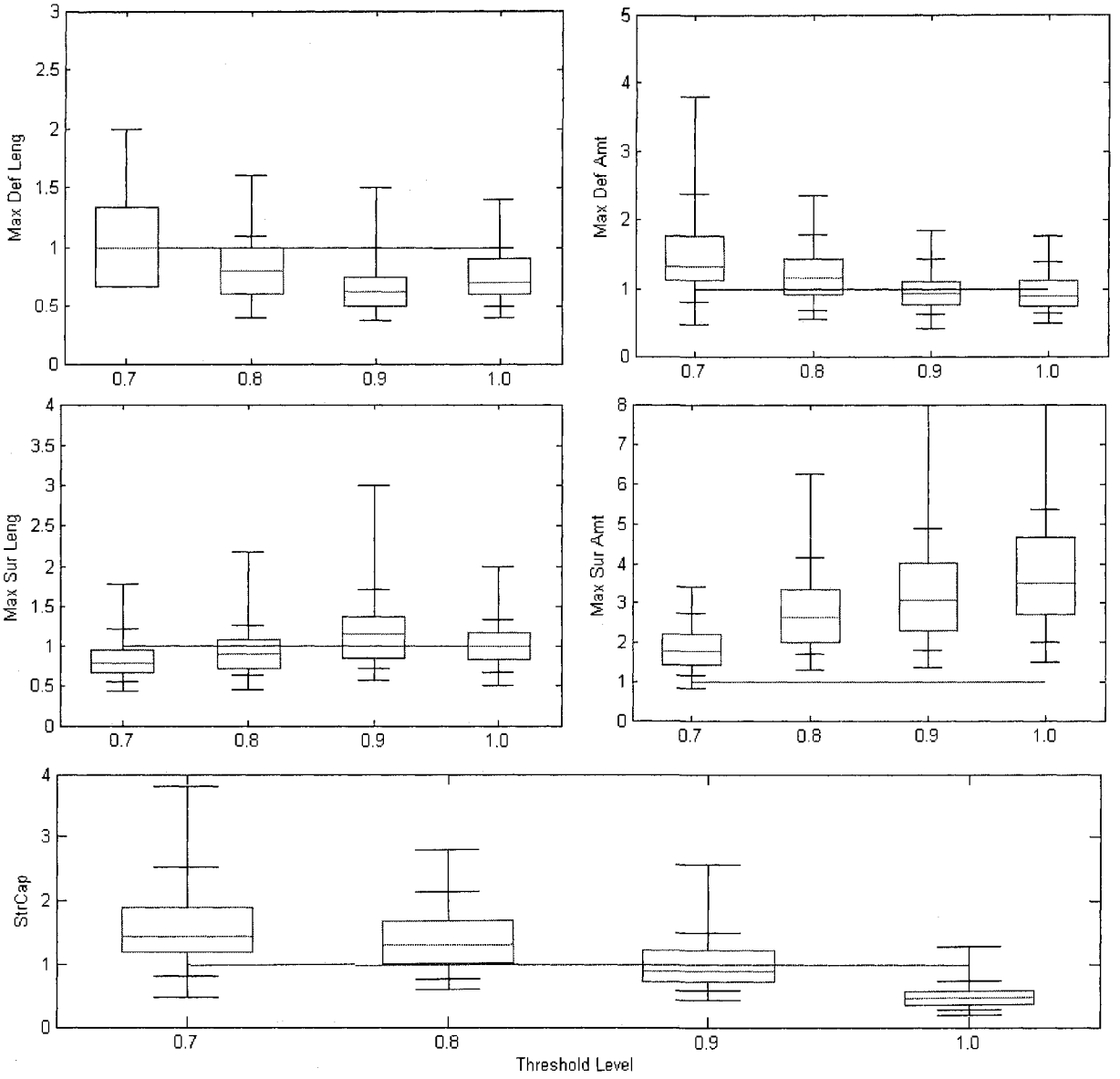


Figure 4-11 Yearly drought statistics at different threshold levels of Historical (dot line) and NPDK simulations (boxplot) for Site 21 of the Colorado River monthly streamflow

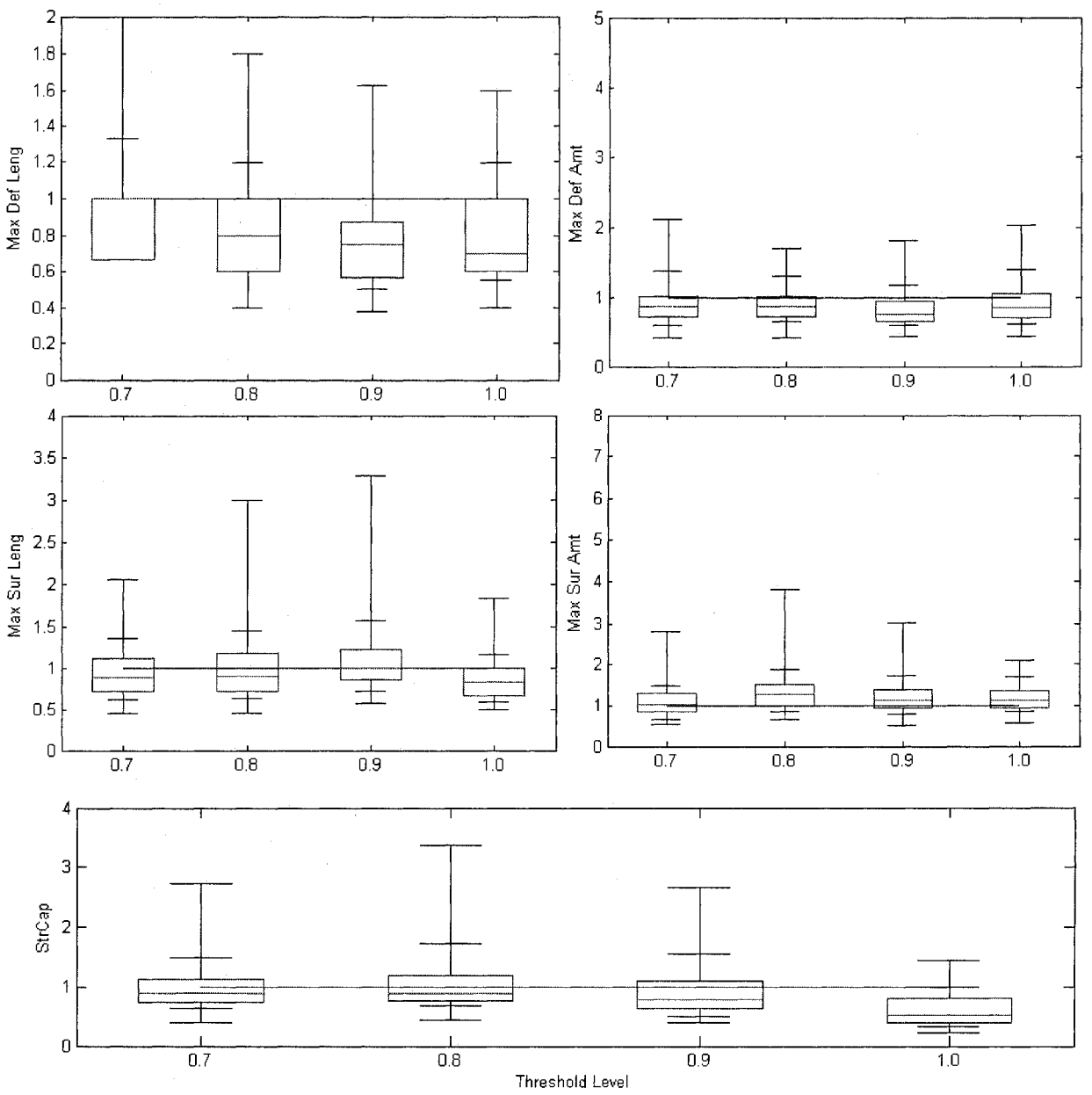


Figure 4-12 Yearly drought statistics at different threshold levels of Historical (dot line) and KPA simulations (boxplot) for Site 21 of the Colorado River monthly streamflow

4.6 References

Bartolini, P., and Salas, J. D. (1993). "Modeling of Streamflow Processes at Different Time Scales." *Water Resources Research*, 29(8), 2573-2587.

Buishand, T. A., and Brandsma, T. (2001). "Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling." *Water Resources Research*, 37(11), 2761-2776.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley Pub. Co.,.

Koutsoyiannis, D., and Manetas, A. (1996). "Simple disaggregation by accurate adjusting procedures." *Water Resources Research*, 32(7), 2105-2117.

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

Lee, T., and Salas, J. D. (2006). "Record Extension of Monthly Flows for the Colorado River System." Bureau of Reclamation, U.S. Department of Interior.

Lee, T., and Salas, J. D. (2008). "Multivariate Simulation Monthly Streamflows of Intermittent and Non-intermittent."

Prairie, J., Rajagopalan, B., Lall, U., and Fulp, T. (2007). "A stochastic nonparametric technique for space-time disaggregation of streamflows." *Water Resources Research*, 43(3), -.

Salas, J. D. (1993). "Analysis and Modeling of Hydrologic Time Series

" In: *Handbook of Hydrology*, M. D.R., ed., McGraw-Hill.

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

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer.

Srinivas, V. V., and Srinivasan, K. (2006). "Hybrid matched-block bootstrap for stochastic simulation of multiseason streamflows." *Journal of Hydrology*, 329(1-2), 1-15.

Tarboton, D. G., Sharma, A., and Lall, U. (1998). "Disaggregation procedures for stochastic hydrology based on nonparametric density estimation." *Water Resources Research*, 34(1), 107-119.

Valencia, D., and Schaake, J. C. (1973). "Disaggregation Processes in Stochastic Hydrology." *Water Resources Research*, 9(3), 580-585.

Wójcik, R., Beersma, J. J., and Buishand, T. A. (2000). "Rainfall generator for the Rhine basin; multi-site generation of weather variables for the entire drainage area ", KNMI publication: PUBL-186-IV.

Appendix 4-A. Gram Schmidt Orthonormalization (GSO)

Gram Schmidt orthonormalization procedure to obtain the rotation matrix with unit orthogonal basis vectors, $\mathbf{R} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_d)$ is summarized as

$$\mathbf{e}_d = \left(1/\sqrt{d}, ..., 1/\sqrt{d}\right)^T \tag{4-A1}$$

For $1 \le j \le d-1$ (4-A2)

$$\mathbf{e}_j = \frac{\mathbf{i}_j - \sum_{k=j+1}^{d} (\mathbf{e}_k \cdot \mathbf{i}_j)\mathbf{e}_k}{\left|\mathbf{i}_j - \sum_{k=j+1}^{d} (\mathbf{e}_k \mathbf{i}_j)\mathbf{e}_k\right|}$$

where $\mathbf{i}_1 = (1,0,...,0)^T, \mathbf{i}_2 = (0,1,...,0)^T ..., \mathbf{i}_d = (0,0,...,1)^T$, $|\mathbf{a}|$ is the norm of vector **a**, and $\mathbf{a} \cdot \mathbf{b}$ is the inner product of vector **a** and **b**.

For example, if d=2 then

$$\mathbf{e}_2 = \left(1/\sqrt{2}, 1/\sqrt{2}\right)^T$$

$$\mathbf{e}_1 = \frac{\begin{pmatrix}1\\0\end{pmatrix} - \left(1/\sqrt{2}, 1/\sqrt{2}\right)\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}}{\left|\begin{pmatrix}1\\0\end{pmatrix} - \left(1/\sqrt{2}, 1/\sqrt{2}\right)\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}\right|} = \frac{\begin{pmatrix}1/2\\-1/2\end{pmatrix}}{1/\sqrt{2}} = \begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\end{pmatrix}$$

Therefore,

$$\mathbf{R} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \text{ and } \mathbf{R}^T = \mathbf{R}^{-1} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

Appendix 4-B. Example of Disaggregation with KNNR and linear or proportional adjustment

A simple example of the suggested disaggregation model with KNNR and linear adjustment (KLA) or proportional adjustment (KPA) is shown below. The process is taken from the model procedure in Chapter 4-3. The employed generated data is the monthly streamflow of Lees Ferry at Colorado River for the first 10 years as shown in Table 4-B1.

(1) Fit a model for X_t^* and generate from the fitted model (KGK model is employed), say the X_1^*=179.5 is generated at first .

(2) The distances $\Delta_\nu = \left| X_1^* - X_\nu \right|$ and ν=1,…,10 is estimated as shown in Table 4-B2.

(3) Among the smallest K-values of Δ_ν, where $K = \sqrt{N} = 3$, one is selected with the random generation from the discrete weighted distribution and its cumulative distribution {0.55, 0.82, 1.00} is compared with the generated value from uniform distribution, say 0.62. And the monthly data set of second year is selected as $\tilde{\mathbf{Y}}_1 = \mathbf{Y}_{1914} = \{6.7, 5.4, ..., 7.8\}$.

(4) Then, the selected historical lower-level dataset $\tilde{\mathbf{Y}}_1$ (the first second row of Table B3) is adjusted with a linear (the fourth row of Table 4-B3) or a proportional (the fifth row of Table 4-B3) adjusting procedure. For example, for linear adjustment 11.4+0.023*(179.5-190.8)=11.1 and for proportional adjustment 11.4*179.5/190.8=10.7

(5) For the next year, say X_2^*=214.9, generated from KGK. The distances from Eq.(4-15) is estimated as

$$\Delta_\nu = \sqrt{\varphi_1\left(X_t^* - X_\nu\right)^2 + \varphi_2\left(Y_{t-1,d}^* - Y_{\nu-1,d}\right)^2}$$
$$= \sqrt{1/35.6^2\left(214.9-212.3\right)^2 + 1/3.7^2\left(5.9-15.0\right)^2}$$
$$= 2.46$$

Notice that the first historical year is excluded since $Y_{\nu-1,d} = Y_{0,12}$ cannot be obtained. One is selected from the k-nearest neighbors {1909,1912,1911}. The monthly data set of the selected year is adjusted linearly or proportionally as shown in Step (4).

(6) Step(5) is repeated until the generation length is met.

The below steps are necessary when Genetic Algorithm mixture is employed. If from the steps (1)~(3), $\tilde{\mathbf{Y}}_1 = \mathbf{Y}_{1914} = \{6.7,5.4,...,7.8\}$ is taken, then

(3-1) One more monthly set is obtained with KNNR that is close to $X_1 = \sum_{i=1}^{d} \widetilde{Y}_{1,i}$.

The distances for KNNR are $\Delta_\nu = |X_1 - X_\nu|$ where ν=1,..,N. Notice that X_1 is also included since the same monthly set as the current set $\widetilde{\mathbf{Y}}_1$ is possible to be selected.

(3-2) From the cumulative weighted probability with $K = \sqrt{N} = 3$ as {0.55, 0.82, 1.00}, one is selected from uniform random number [0,1]. Assume that the uniform random number 0.93 is generated, then the monthly data set of the third order (1914) is chosen for $\widetilde{\mathbf{Y}}_1^{\,2} = \mathbf{Y}_{1914}$ and set $\widetilde{\mathbf{Y}}_1 = \widetilde{\mathbf{Y}}_1^{\,1}$.

(3-3) Two data set ($\widetilde{\mathbf{Y}}_1^{\,1}$ and $\widetilde{\mathbf{Y}}_1^{\,2}$) are mixed with GA as follows and create the new monthly data set, say $\widetilde{\mathbf{Y}}_1^{\,GA}$. Set the cross probability as 0.5 and a new data set is obtained with one by one. If the uniform random number u_i where i=1,…,d

$$\widetilde{Y}_i^{GA} = \begin{cases} \widetilde{Y}_i^1 & u_i < p_c \\ \widetilde{Y}_i^2 & \text{otherwise} \end{cases}$$

For example, $u_1 = 0.59$ then $\widetilde{Y}_i^{GA} = 6.7$. All the new data set $\widetilde{\mathbf{Y}}_1^{\,GA}$ is presented in Table 4-B6 and Figure 4-B1. This new data set is linearly or proportionally adjusted with the Steps (4) and (5) and repeating as Step(6).

Table 4-B1. The monthly data of Lees Ferry at Colorado River for the first 10 years (10^5 AF)

	1	2	3	4	5	6	7	8	9	10	11	12	Yr
1906	4.6	4.0	2.3	2.4	2.9	6.8	12.0	36.4	50.1	29.5	16.1	15.0	182.1
1907	7.4	5.0	3.5	3.6	3.8	7.9	14.7	27.0	59.7	51.0	19.2	9.6	212.3
1908	6.1	3.8	2.7	2.8	3.8	6.6	10.4	16.0	29.2	19.2	11.2	6.0	117.7
1909	4.8	4.0	3.1	3.8	3.2	7.6	11.2	33.5	72.0	41.1	18.8	15.3	218.4
1910	6.8	4.9	3.8	2.9	4.9	14.0	17.3	33.0	31.0	13.7	8.7	6.3	147.4
1911	6.2	4.5	3.5	3.7	4.8	9.0	9.5	29.2	41.2	23.5	10.2	5.9	151.3
1912	11.4	4.4	3.5	3.5	3.3	5.4	9.0	36.8	61.5	32.1	13.6	6.3	190.8
1913	6.4	5.3	3.1	3.5	3.1	5.2	18.3	32.7	31.4	19.8	8.7	7.0	144.7
1914	6.7	5.4	3.3	3.7	4.0	8.8	15.9	46.9	63.0	31.2	14.1	7.8	210.7
1915	9.6	5.3	3.3	3.0	4.0	5.3	14.8	24.3	36.4	21.5	8.5	5.3	141.4
Mean	7.0	4.7	3.2	3.3	3.8	7.7	13.3	31.6	47.6	28.3	12.9	8.5	171.7
Stdev	2.1	0.6	0.4	0.5	0.7	2.6	3.3	8.3	15.7	11.3	4.1	3.7	35.6

Table 4-B2. The estimated distance between the historical yearly data and X_1^* and its order

	1906	1907	1908	1909	1910	1911	1912	1913	1914	1915
Δ_v	2.68	32.87	61.73	38.95	32.10	28.22	11.35	34.75	31.20	38.09
Order	1	6	10	9	5	3	2	7	4	8

Table 4-B3. The selected historical monthly data $\tilde{\mathbf{Y}}_t$ and linear adjustment coefficient λ_j and linearly and proportionally adjusted data

Mon	1	2	3	4	5	6	7	8	9	10	11	12
His	11.4	4.4	3.5	3.5	3.3	5.4	9.0	36.8	61.5	32.1	13.6	6.3
λ_j	0.023	0.011	0.005	0.004	0.007	0.023	0.060	0.210	0.323	0.189	0.072	0.036
L.Adj	11.1	4.3	3.5	3.4	3.2	5.1	8.3	34.5	57.8	29.9	12.8	5.9
P.Adj	10.7	4.2	3.3	3.3	3.1	5.1	8.5	34.6	57.9	30.2	12.8	5.9

Table 4-B4. The estimated distances and its order

	1907	1908	1909	1910	1911	1912	1913	1914	1915
Δ_v	2.46	1.04	0.10	2.55	0.25	0.14	0.26	0.30	0.57
Order	8	7	1	9	3	2	4	5	6

Table 4-B5. The estimated distance between the historical yearly data and X_1 and its order

	1906	1907	1908	1909	1910	1911	1912	1913	1914	1915
Δ_v	0	30.2	64.4	36.3	34.7	30.8	8.7	37.4	28.6	40.7
Order	1	4	10	7	6	5	2	8	3	9

Table 4-B6. Monthly streamflow of year 1912, 1914, and the selected set from GA

Month	1	2	3	4	5	6	7	8	9	10	11	12
1912	11.4	4.4	3.5	3.5	3.3	5.4	9	36.8	61.5	32.1	13.6	6.3

1914	6.7	5.4	3.3	3.7	4	8.8	15.9	46.9	63	31.2	14.1	7.8
Rand	0.59	0.33	0.39	0.67	0.77	0.87	0.45	0.36	0.88	0.58	0.96	0.23
Selected	6.7	4.4	3.5	3.7	4	8.8	9	36.8	63	31.2	14.1	6.3

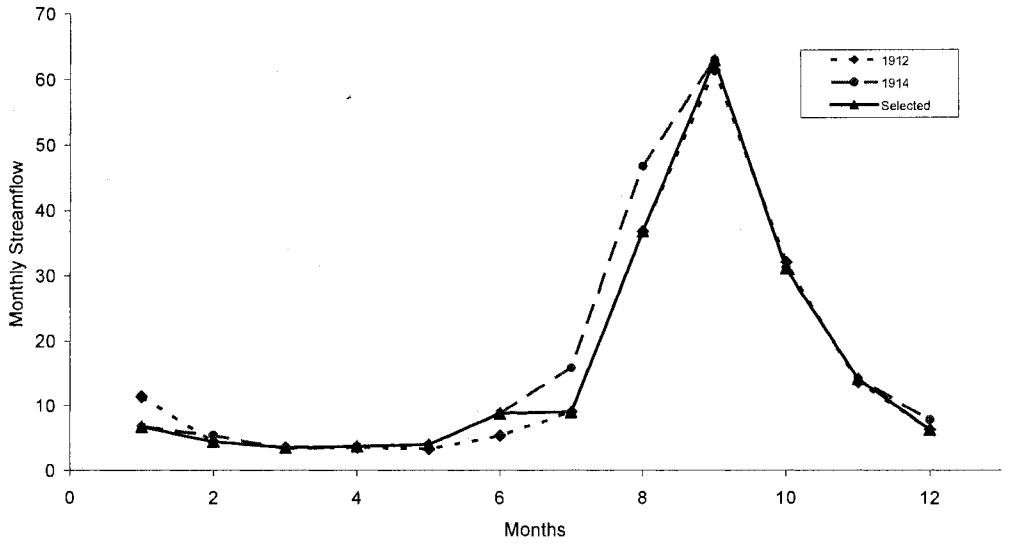


Figure B1. Monthly streamflow of year 1912, 1914, and the selected set from GA

Appendix 4-C. Detailed Figures

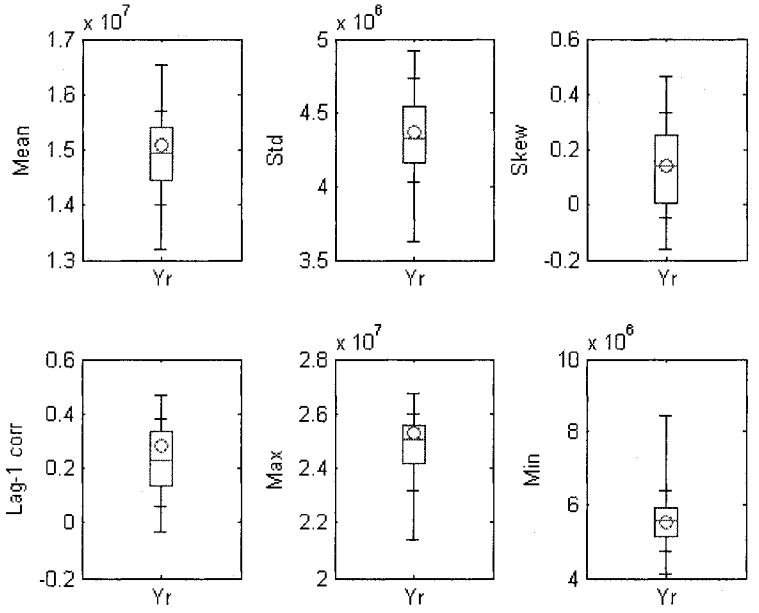


Figure 4-C.1 Key Statistics of Historical (dot line) and KPAG simulations (boxplot) for Site 20 of the Colorado River yearly streamflow

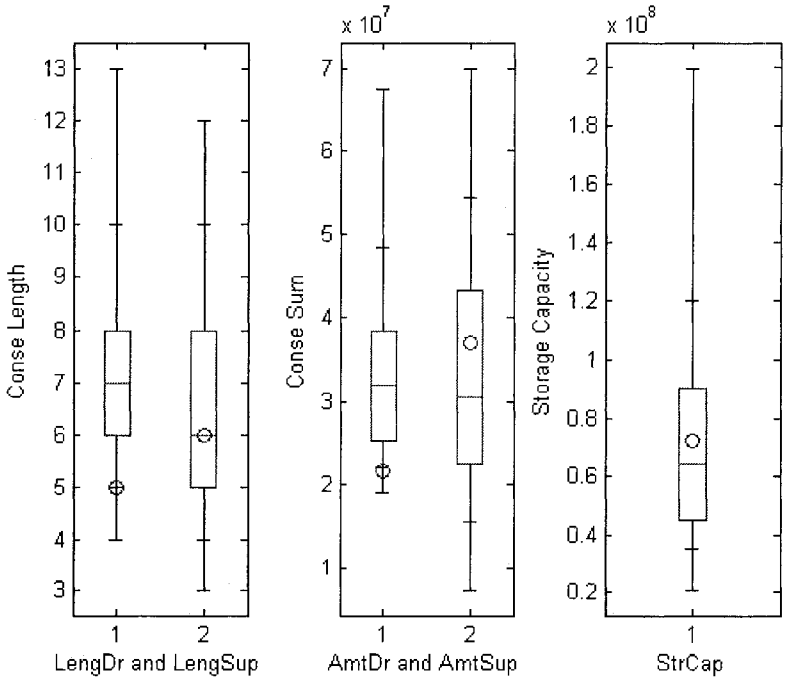


Figure 4-C.2 Drought, Surplus, Storage Statistics of site 20 at Colorado River yearly streamflow

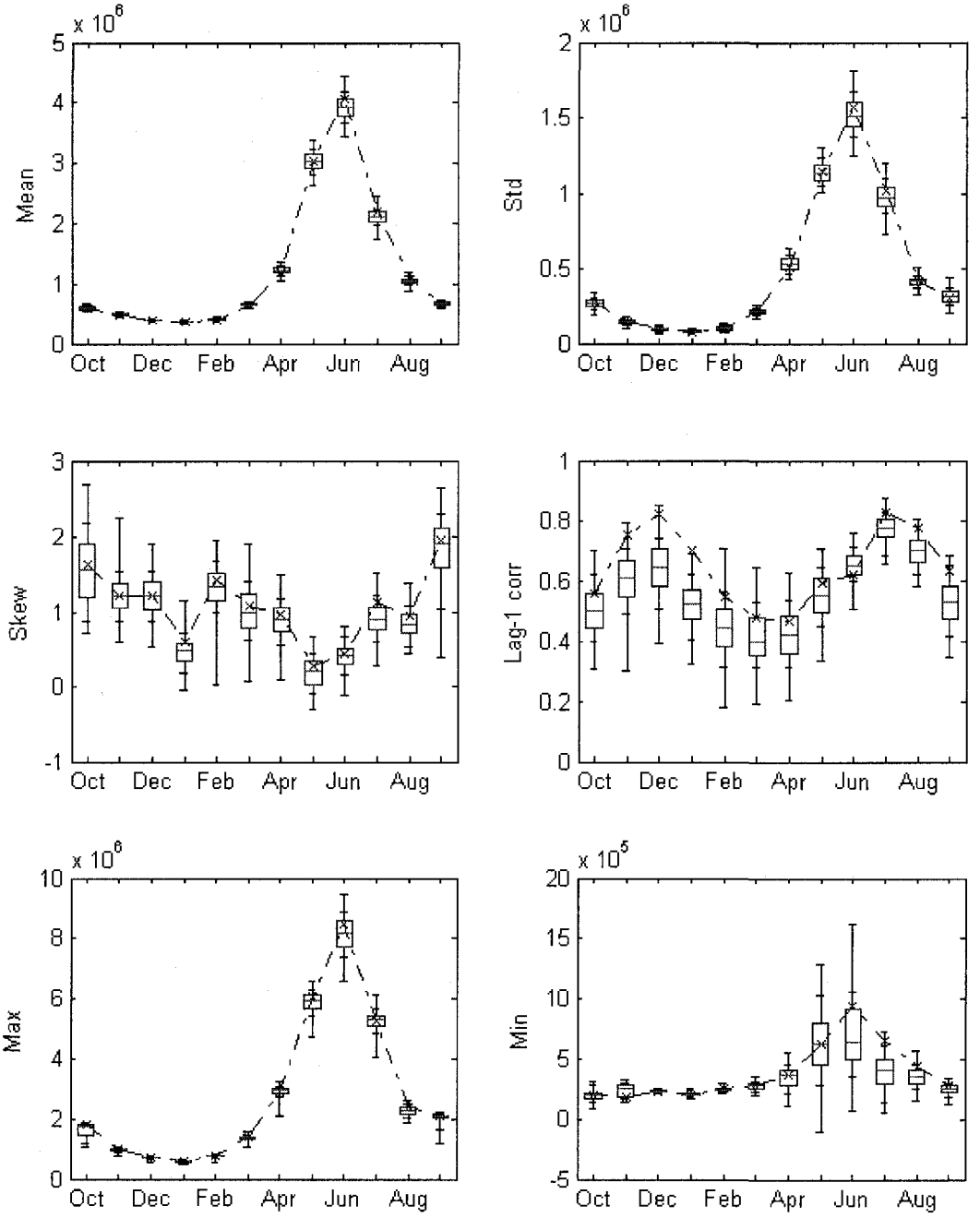


Figure 4-C.3 Key Statistics of Historical (dot line) and KLAG simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

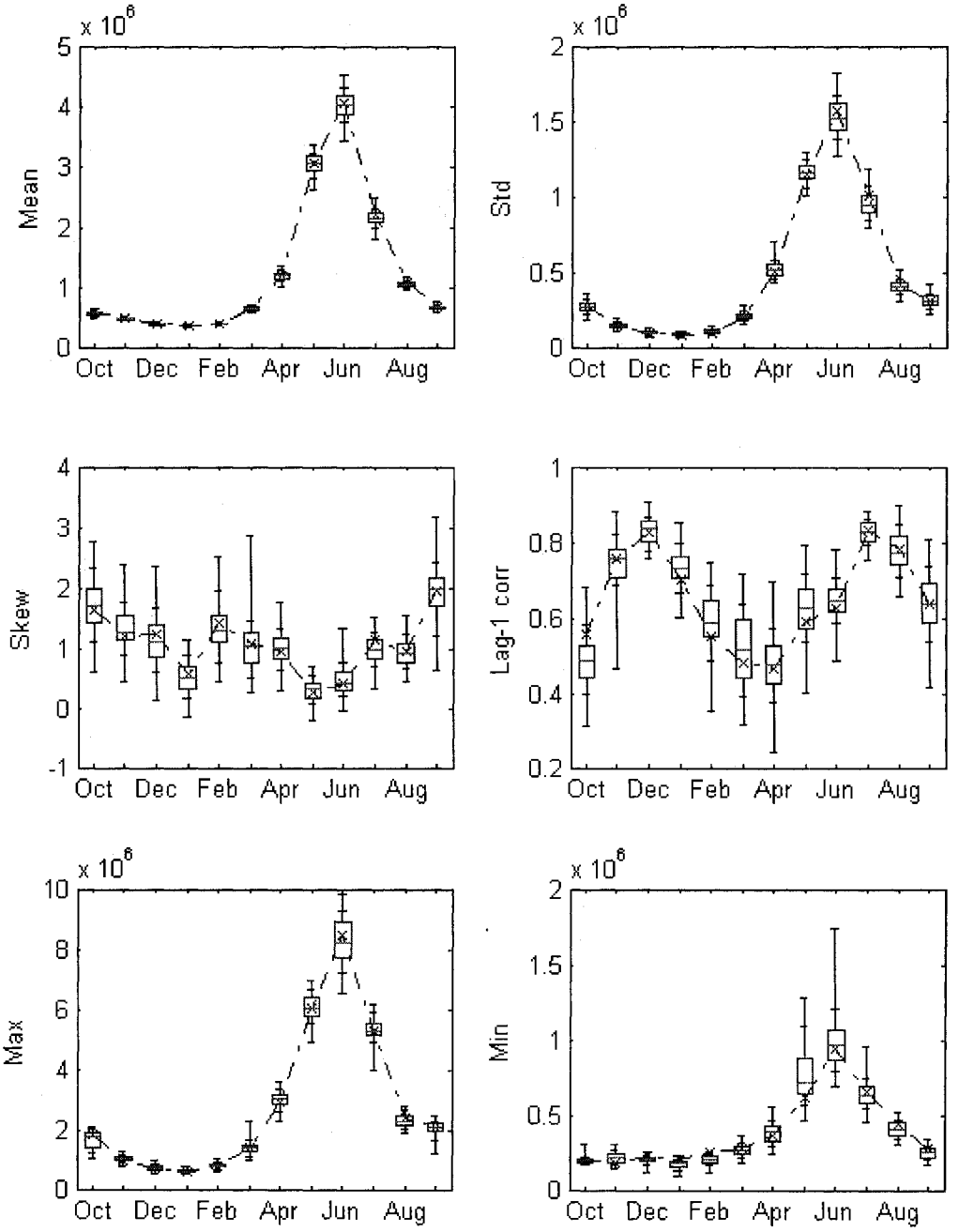


Figure 4-C.4 Key Statistics of Historical (dot line) and KPA simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

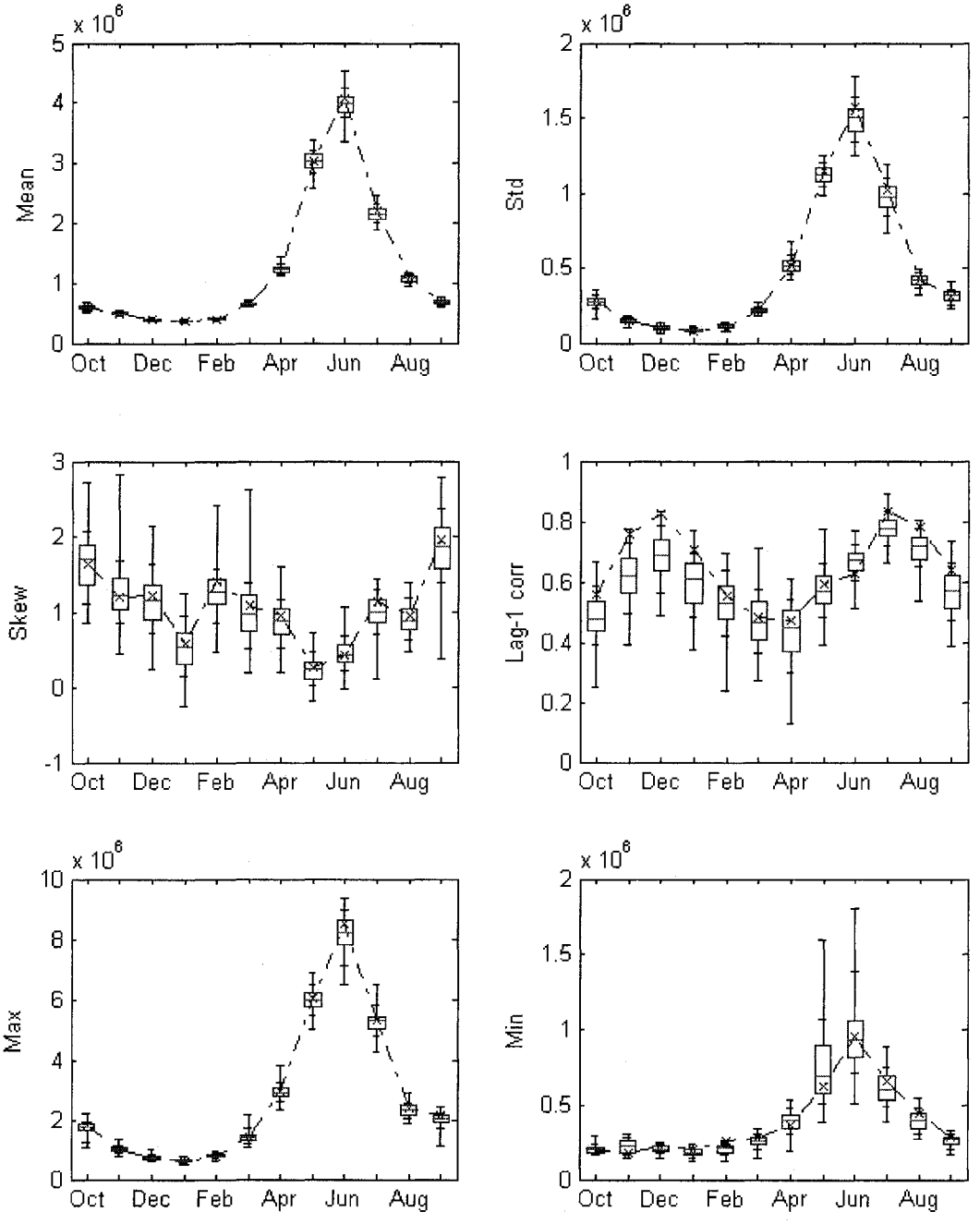


Figure 4-C.5 Key Statistics of Historical (dot line) and KPAG simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

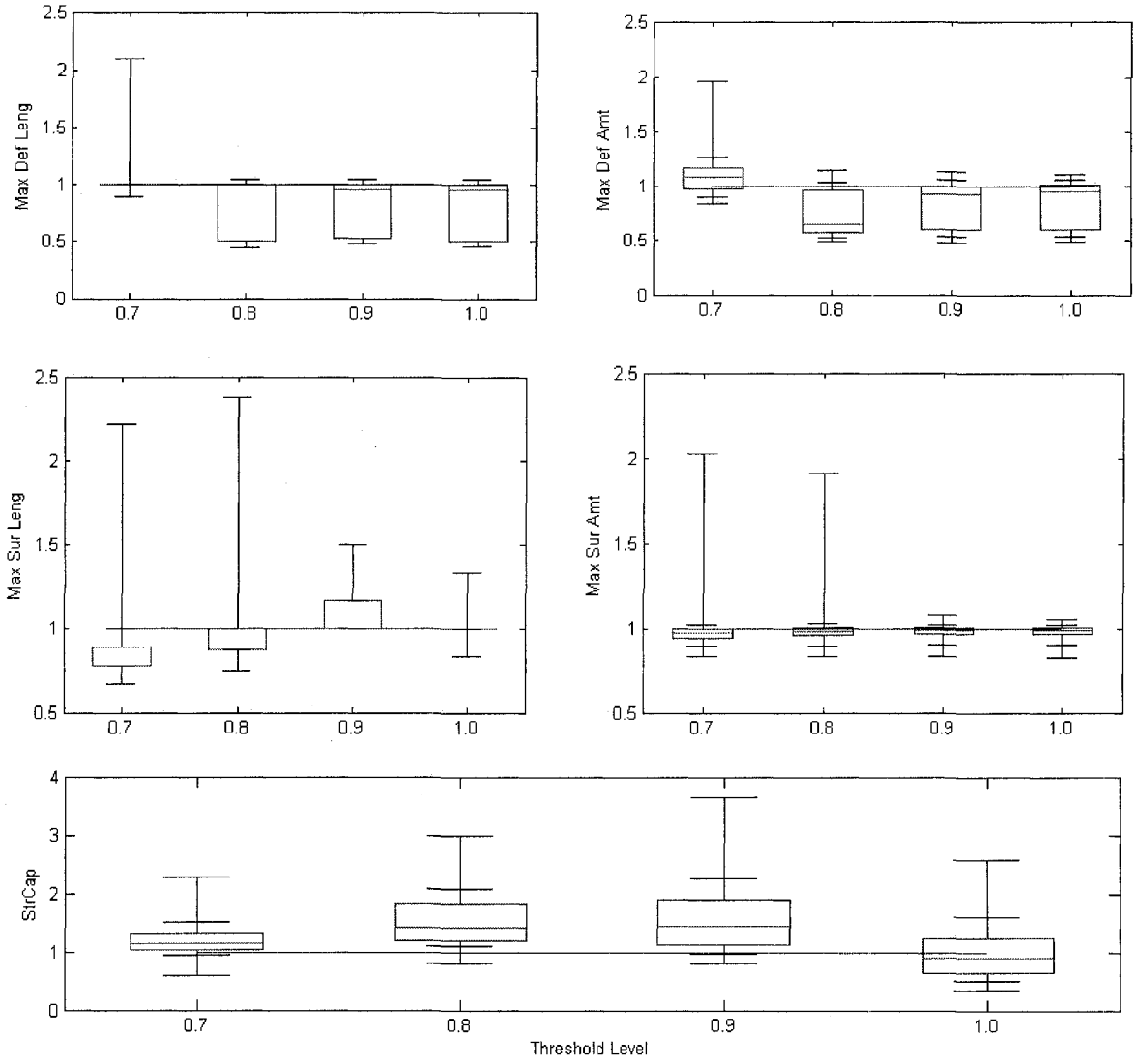


Figure 4-C.6 Monthly drought statistics at different threshold levels of Historical (dot line) and NPDK simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

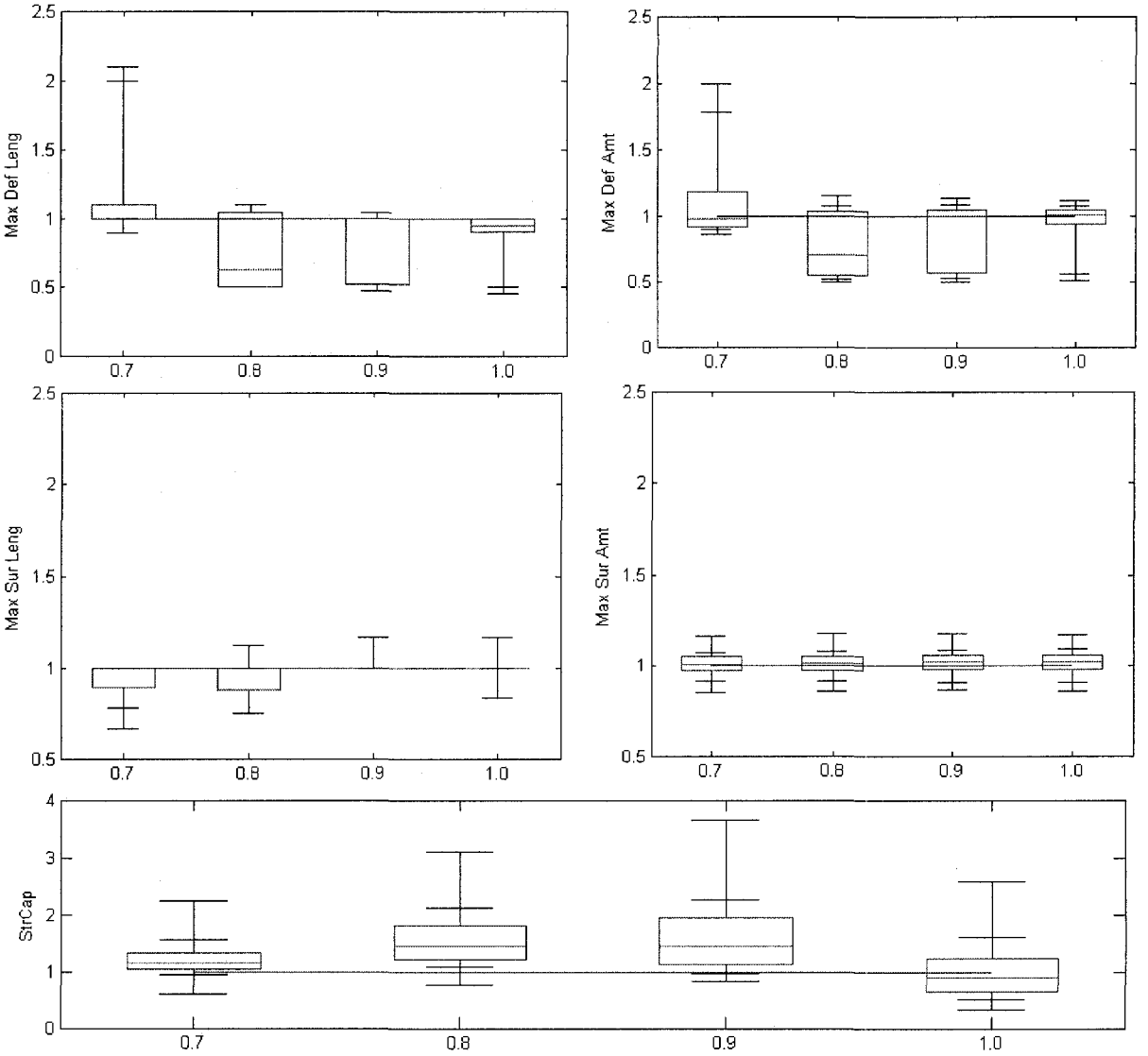


Figure 4-C.7 Monthly drought statistics at different threshold levels of Historical (dot line) and KLA simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

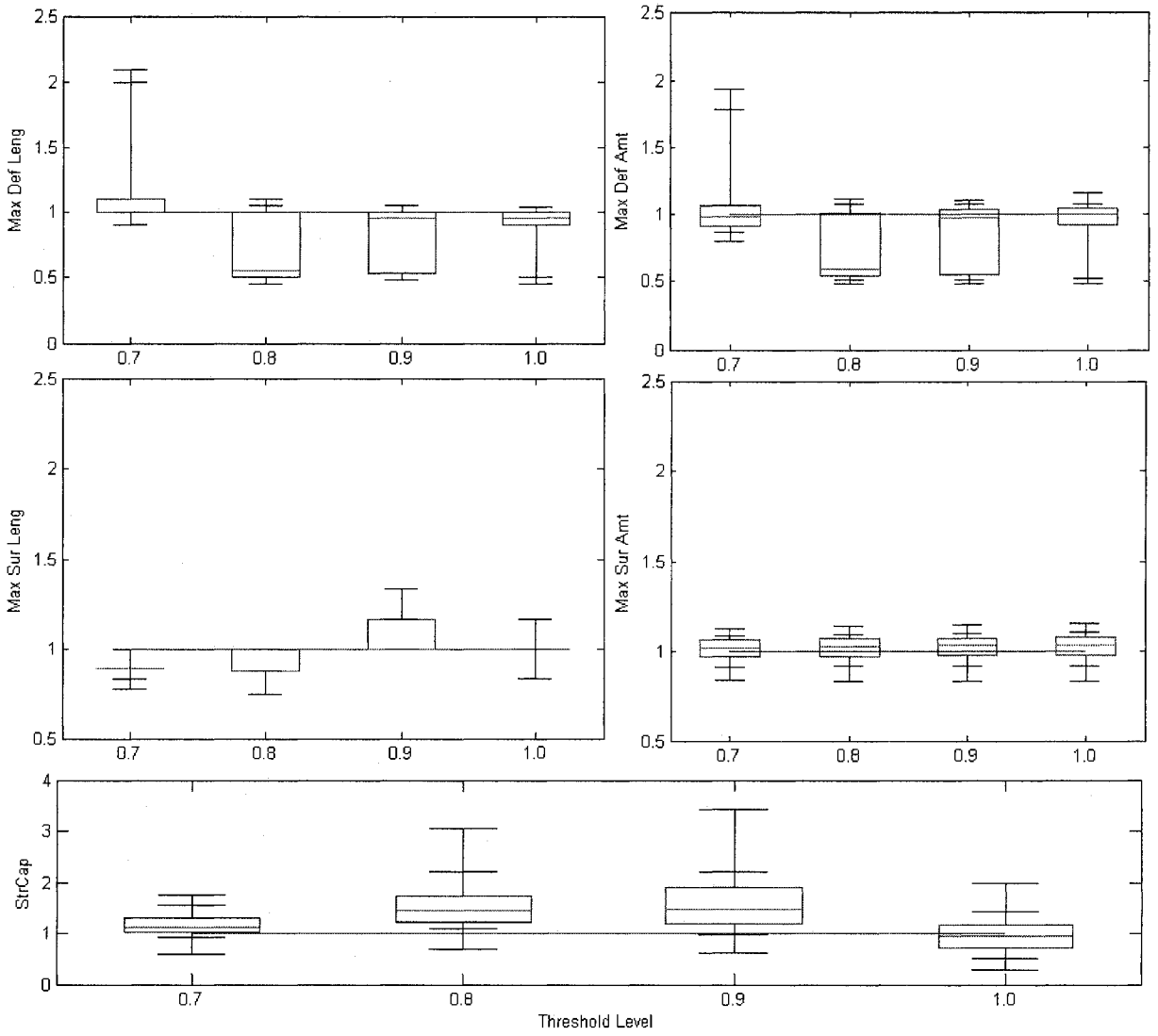


Figure 4-C.8 Monthly drought statistics at different threshold levels of Historical (dot line) and KLAG simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

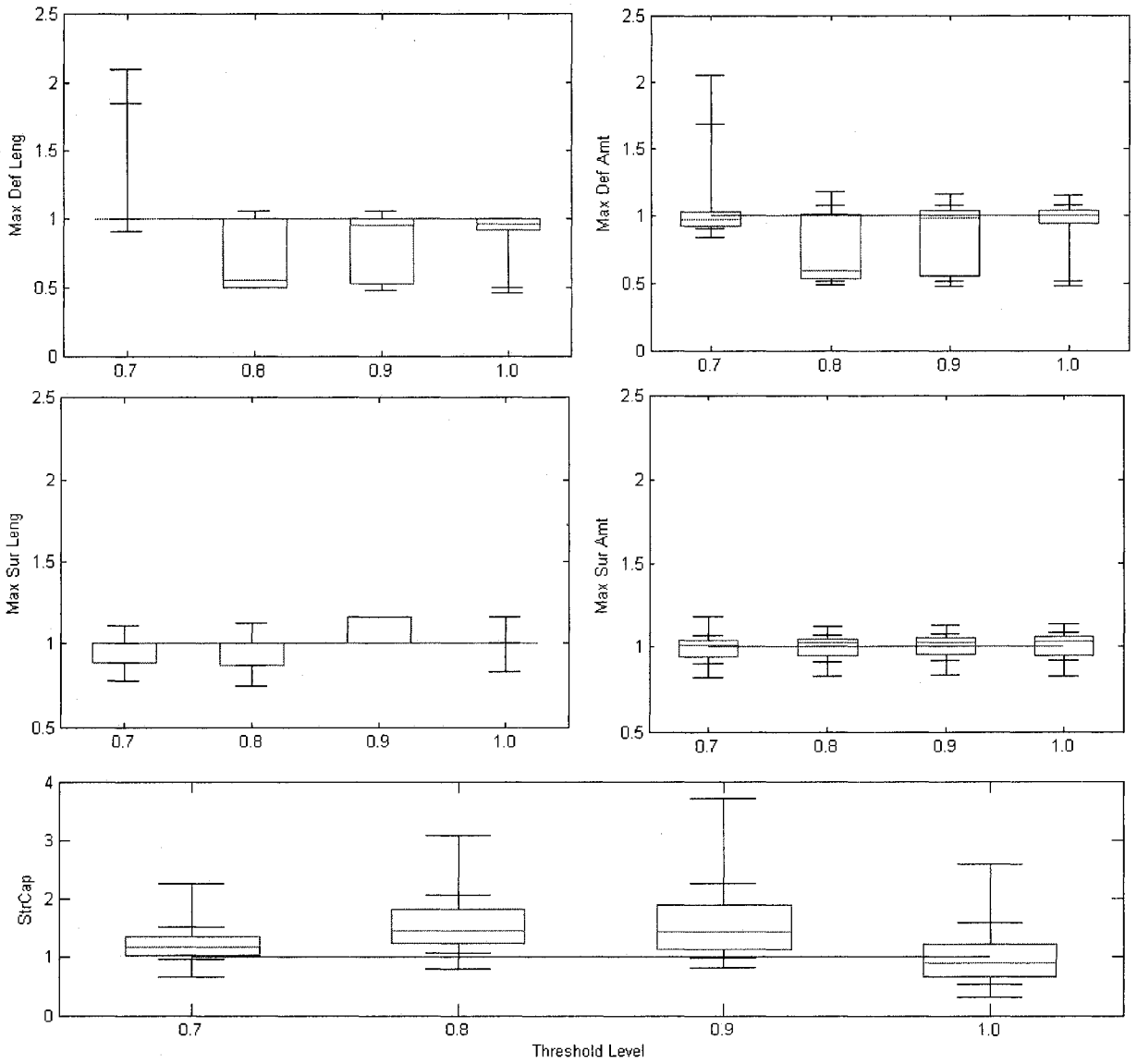


Figure 4-C.9 Monthly drought statistics at different threshold levels of Historical (dot line) and KPA simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

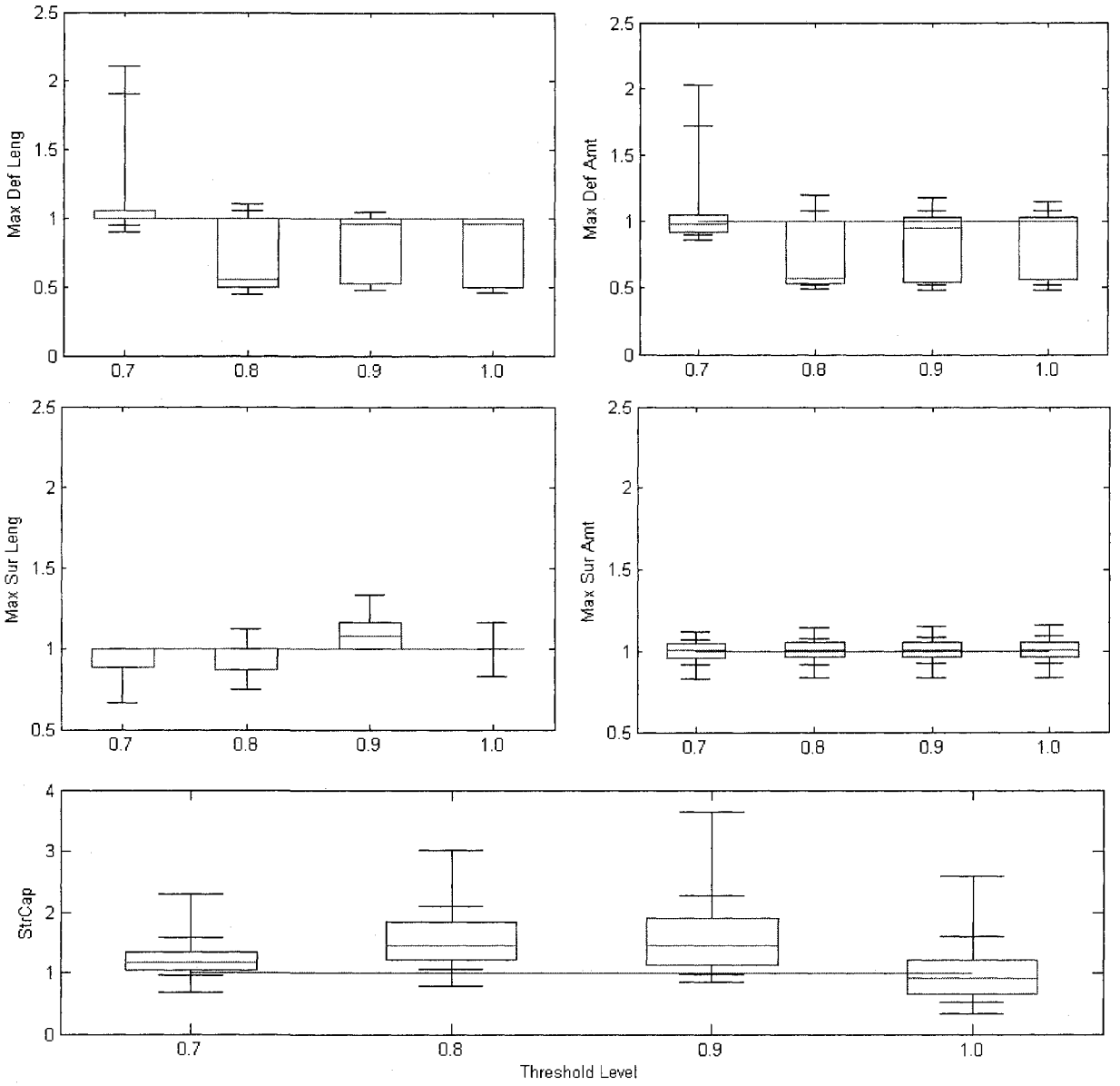


Figure 4-C.10 Monthly drought statistics at different threshold levels of Historical (dot line) and KPAG simulations (boxplot) for Site 20 of the Colorado River monthly streamflow

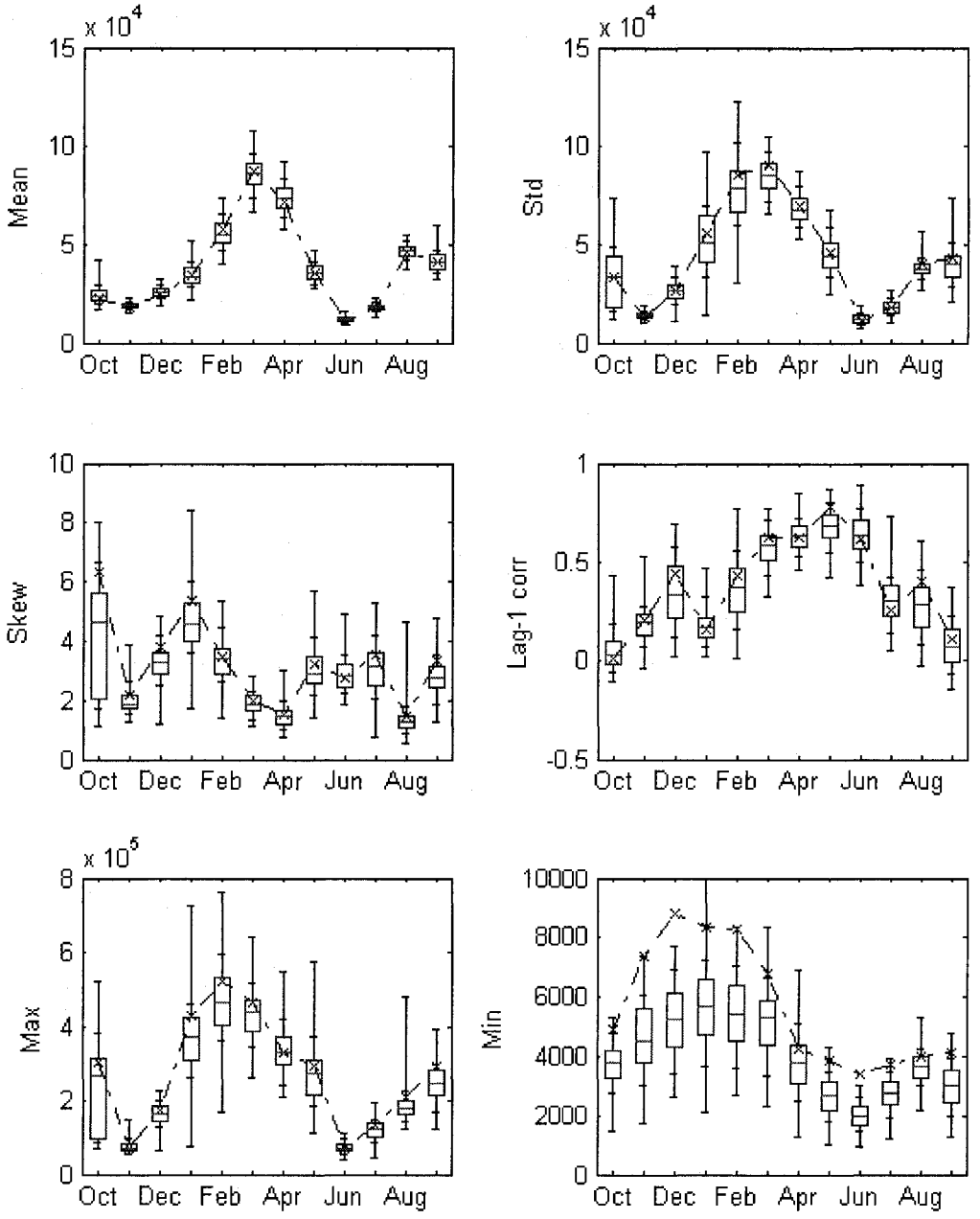


Figure 4-C.11 Key Statistics of Historical and the model generated monthly data of the aggregate variable (KGK and disaggregation) for the tributary stations of Lower Colorado River

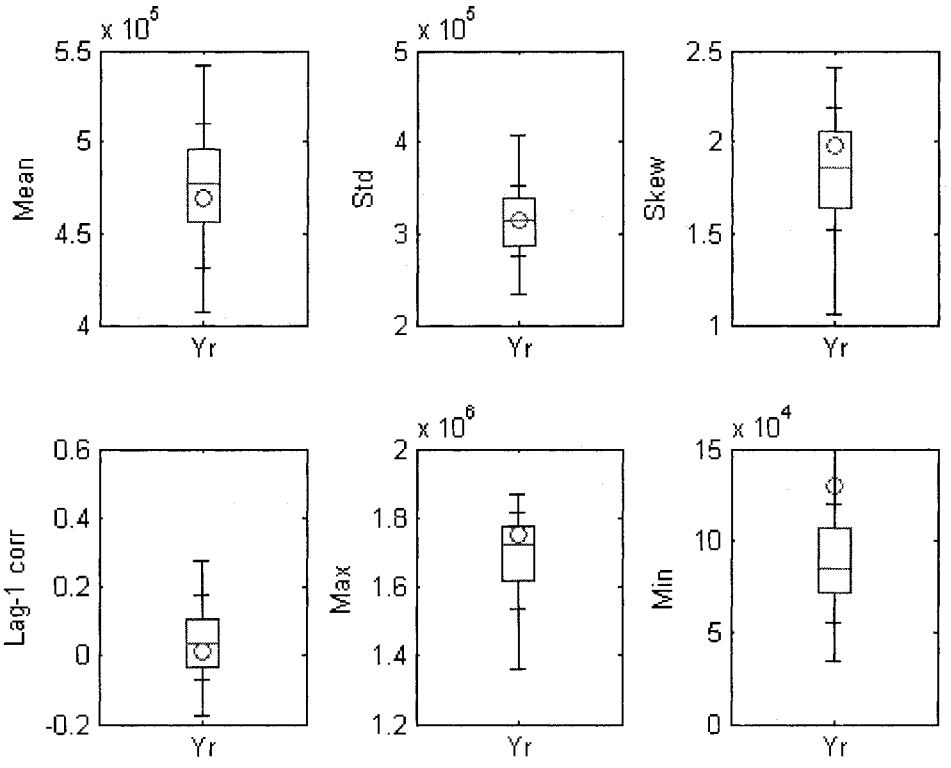


Figure 4-C.12 Key Statistics of Historical and the model generated yearly data of the aggregate variable (KGK and disaggregation) for the tributary stations of Lower Colorado River

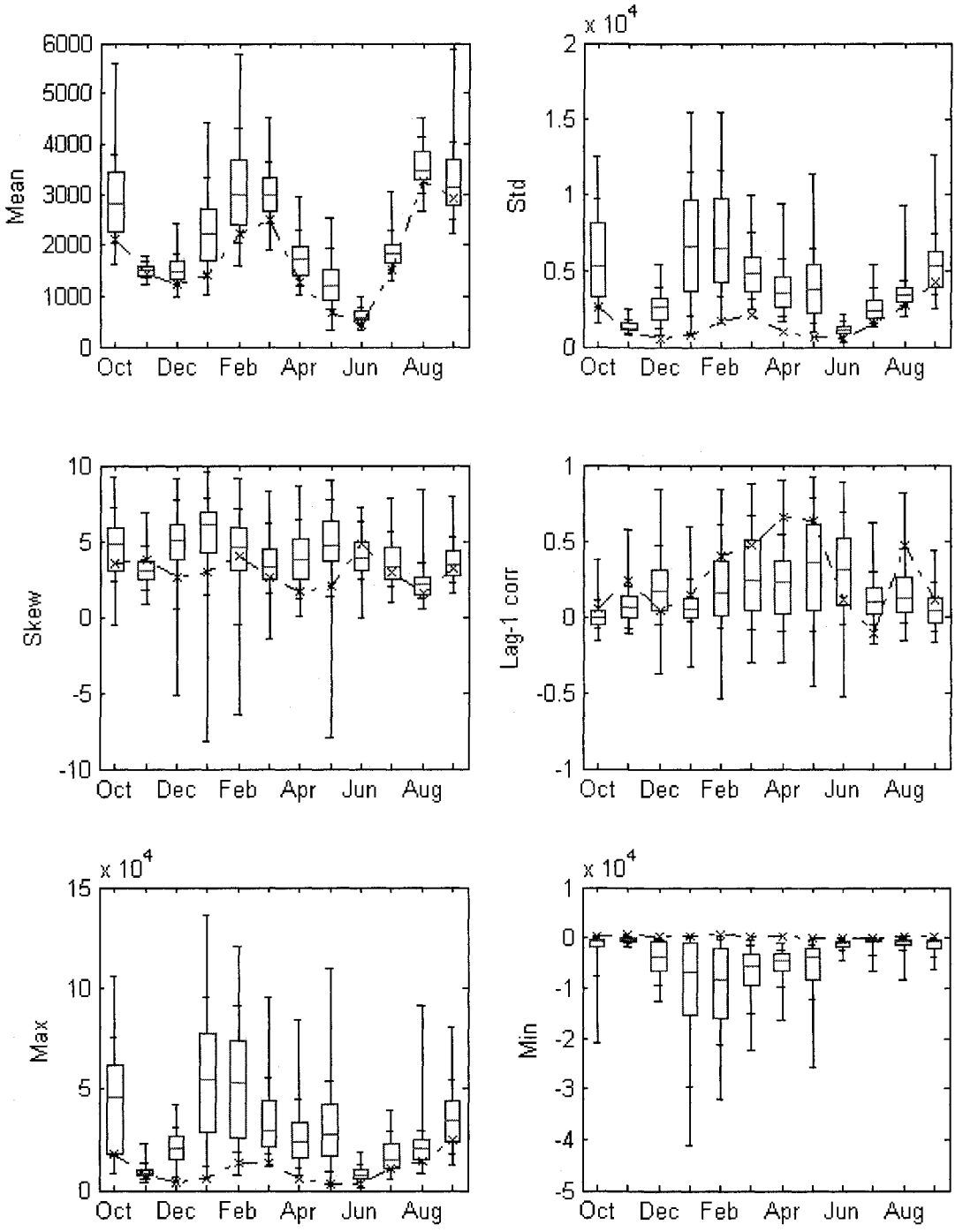


Figure 4-C.13 Key Statistics of Historical (dot line) and NPDK simulations (boxplot) for Site 21 of the Colorado River monthly streamflow

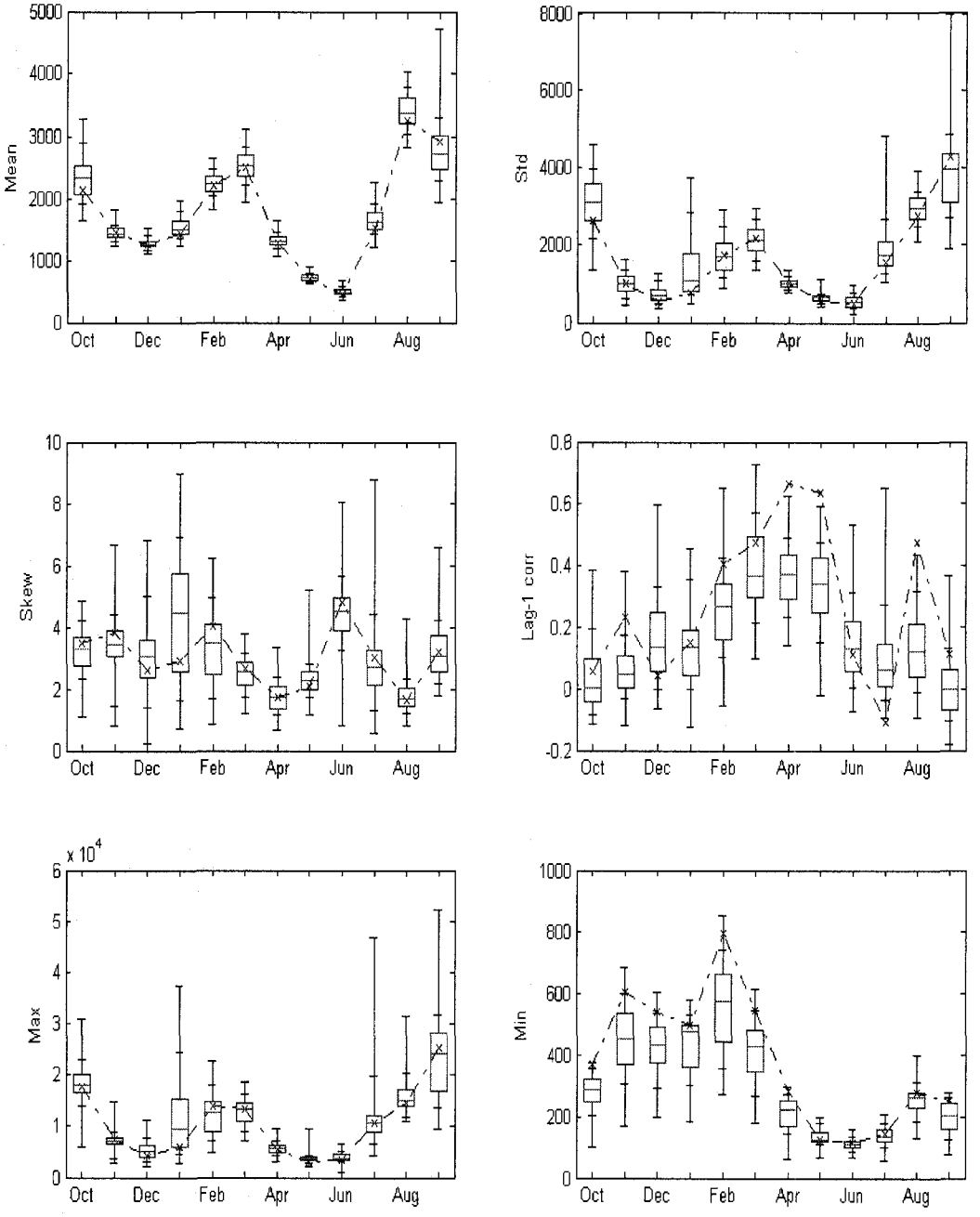


Figure 4-C.14 Key Statistics of Historical (dot line) and KPA simulations (boxplot) for Site 21 of the Colorado River monthly streamflow

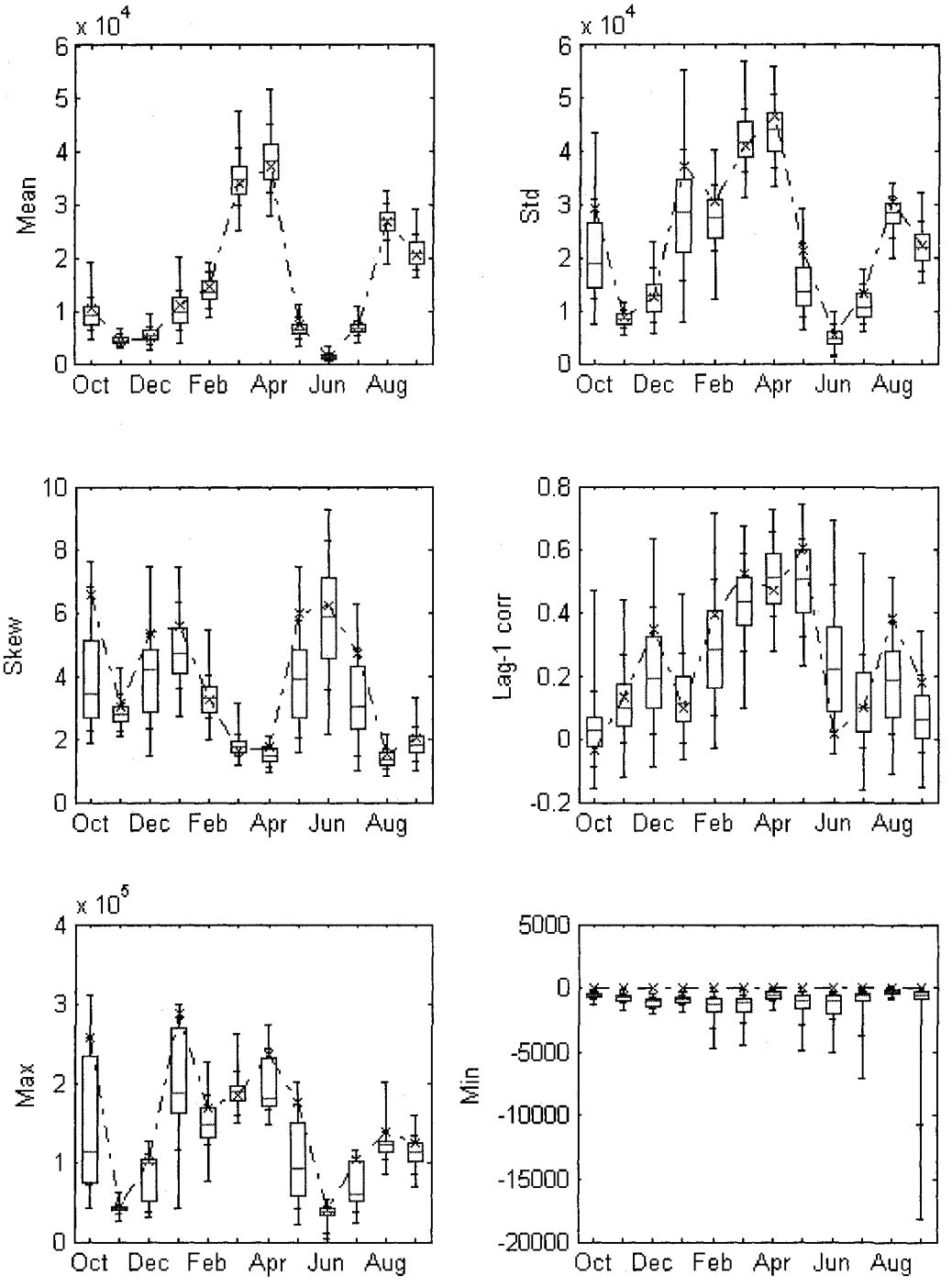


Figure 4-C.15 Key Statistics of Historical (dot line) and NPDK simulations (boxplot) for Site 22 of the Colorado River monthly streamflow

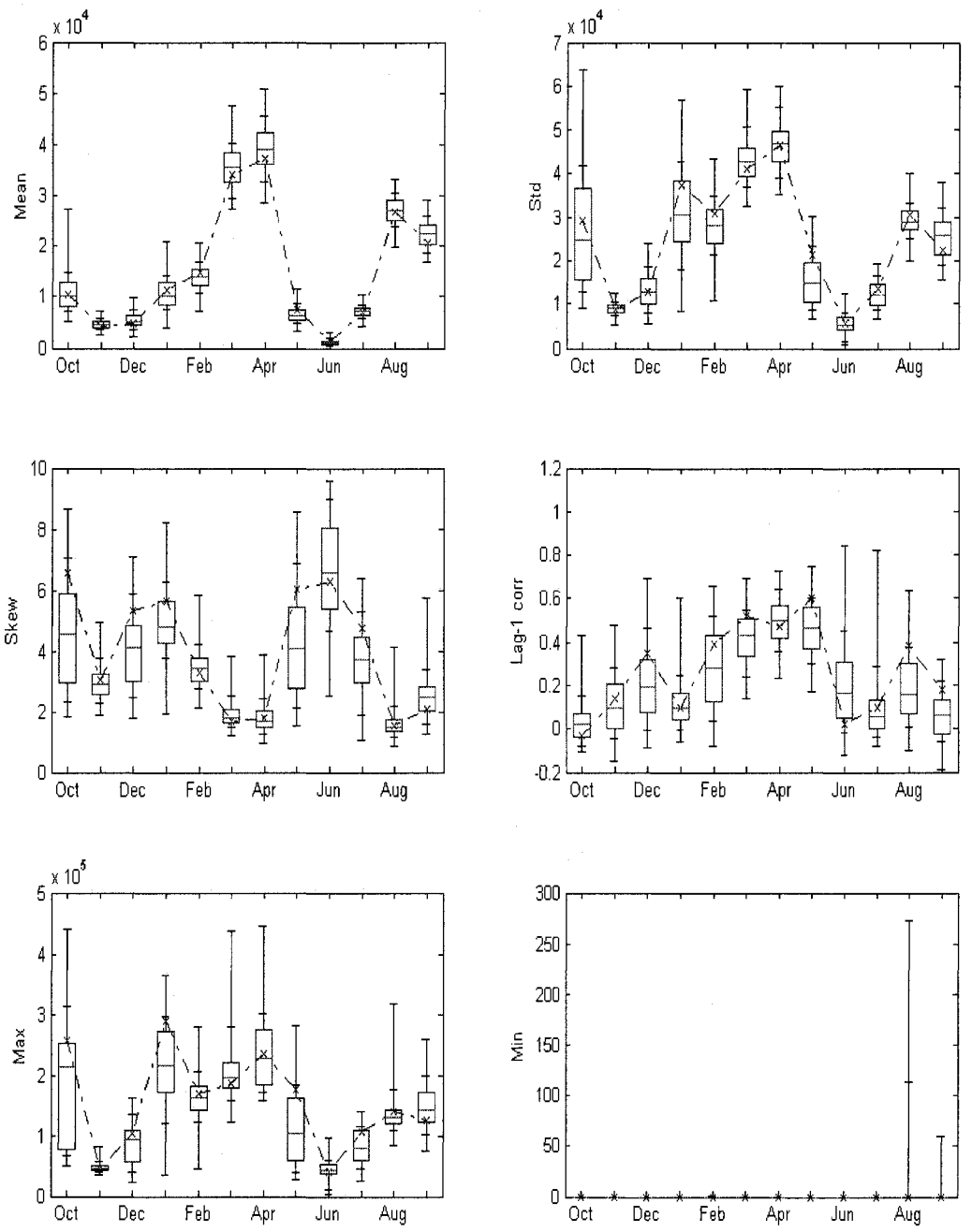


Figure 4-C.16 Key Statistics of Historical (dot line) and KPA simulations (boxplot) for Site 22 of the Colorado River monthly streamflow

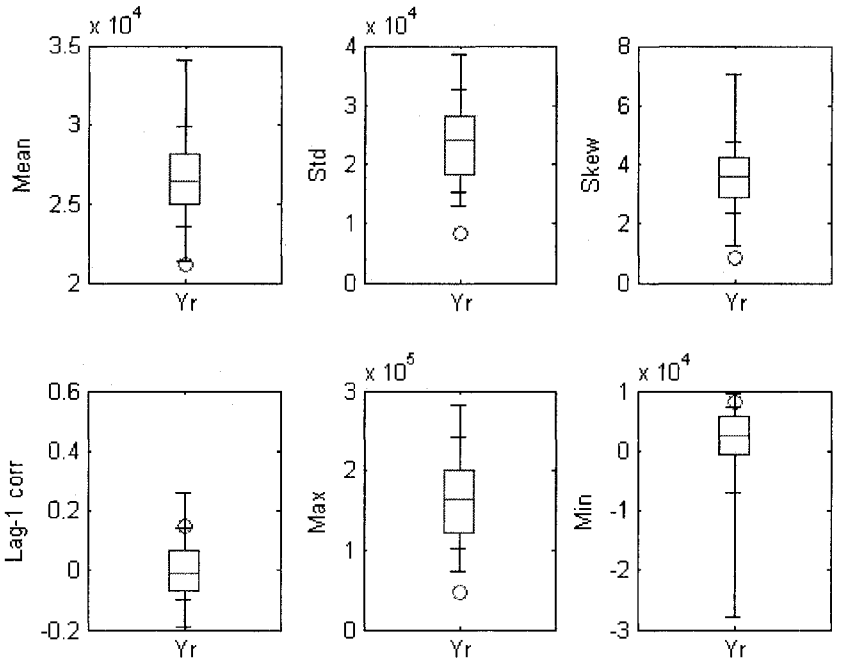


Figure 4-C.17 Key Statistics of Historical (circle) and NPDK simulations (boxplot) for Site 21 of the Colorado River yearly streamflow

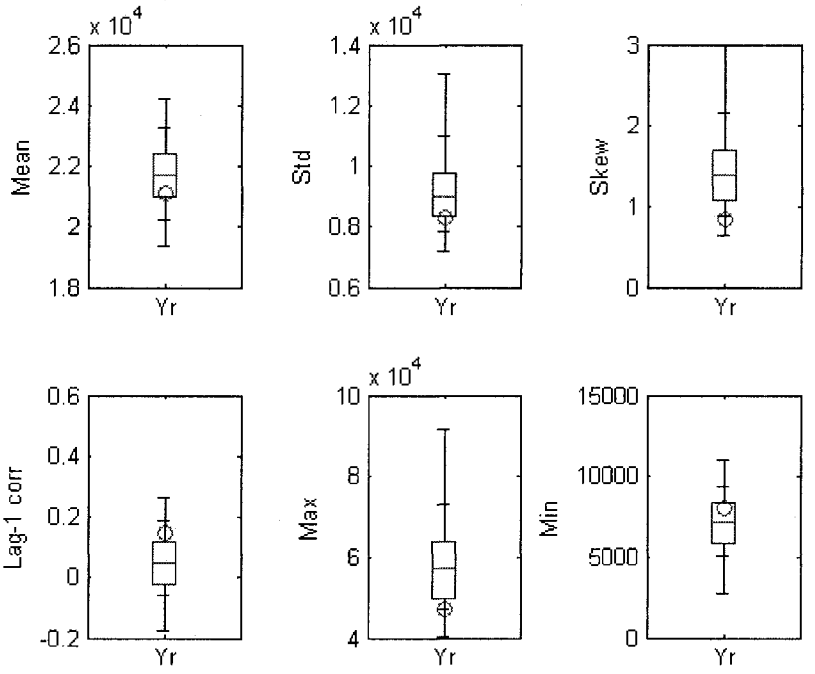


Figure 4-C.18 Key Statistics of Historical (circle) and KPA simulations (boxplot) for Site 21 of the Colorado River yearly streamflow

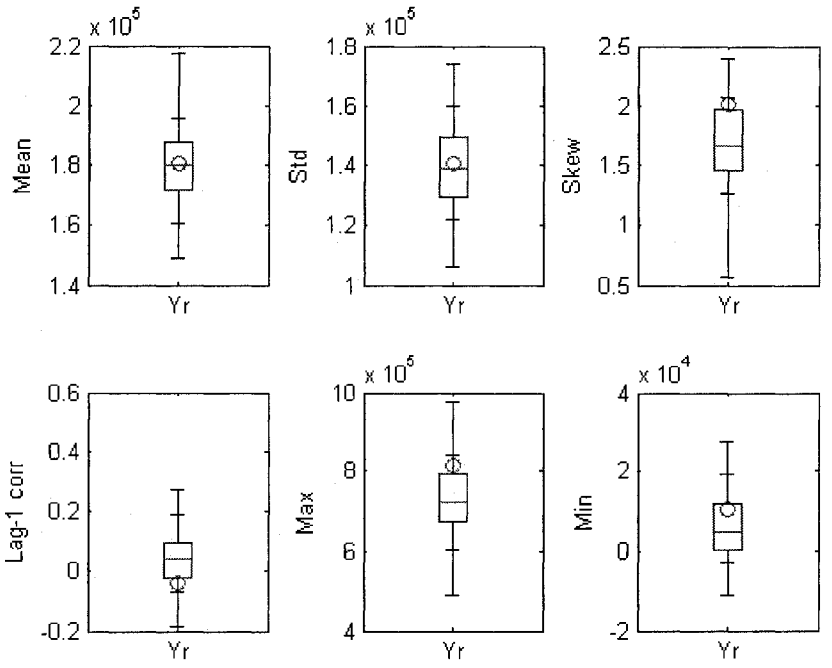


Figure 4-C.19 Key Statistics of Historical (circle) and NPDK simulations (boxplot) for Site 22 of the Colorado River yearly streamflow

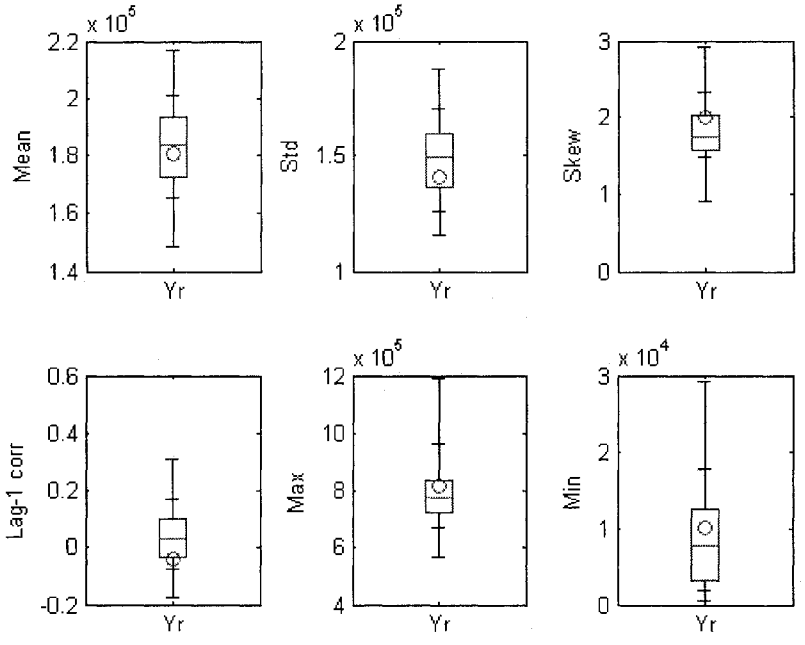


Figure 4-C.20 Key Statistics of Historical (circle) and KPA simulations (boxplot) for Site 22 of the Colorado River yearly streamflow

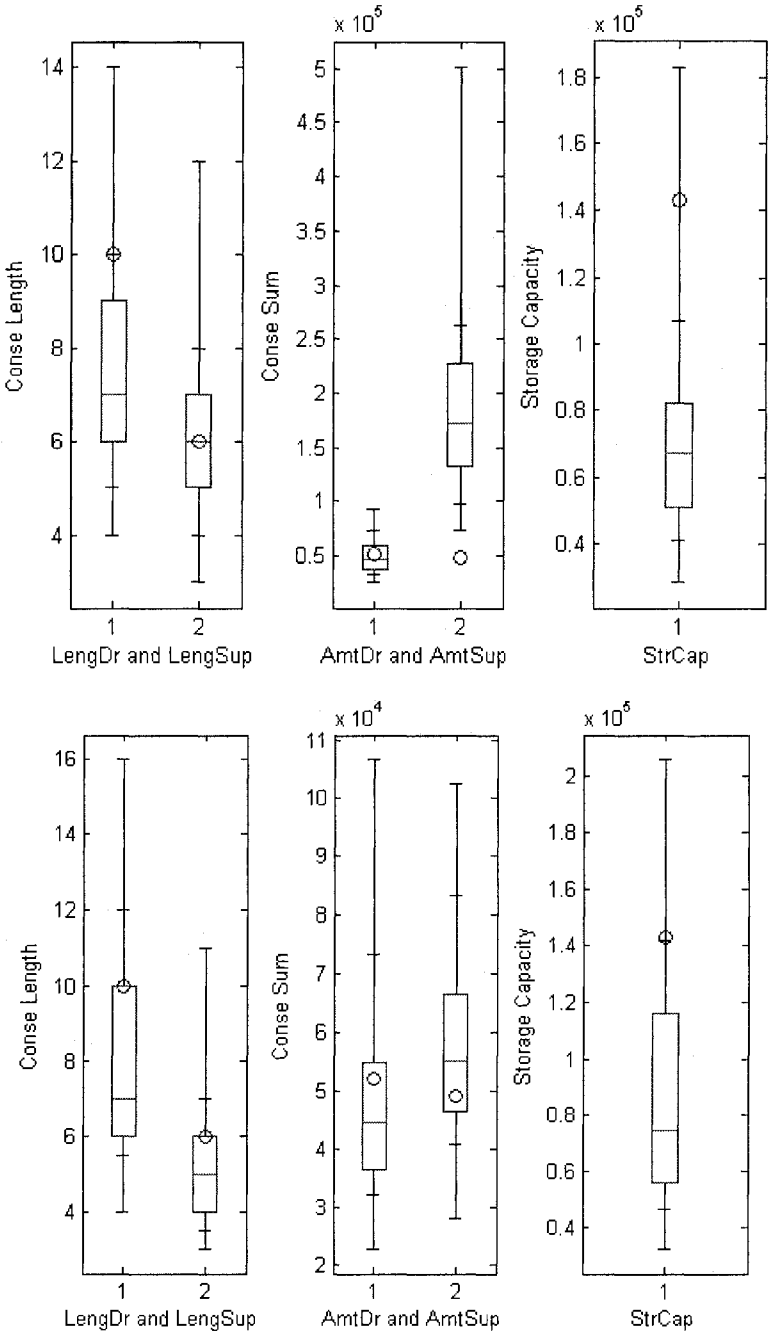


Figure 4-C.21 Reservoir-related statistics from historical (circle) and NPDK (up) and KPA (bottom) simulations (boxplot) for Site 21 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity

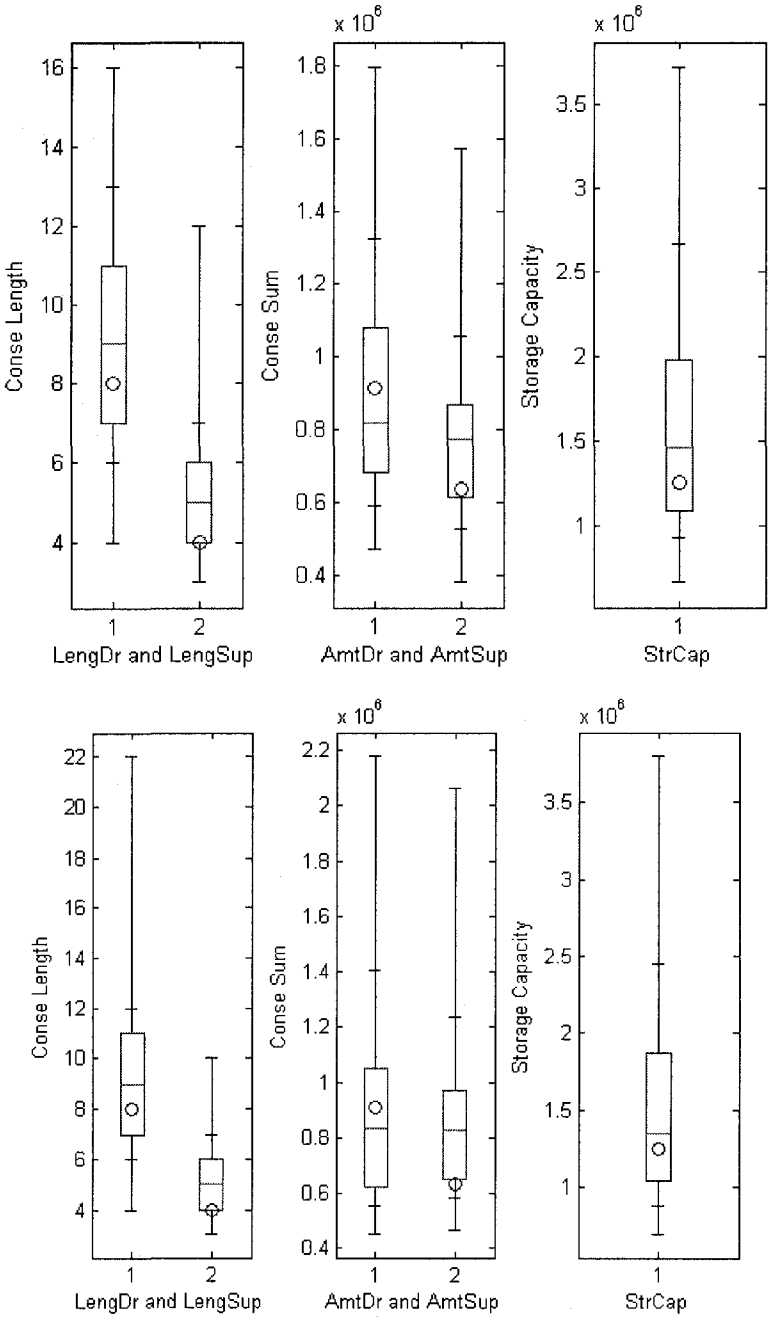


Figure 4-C.22 Reservoir-related statistics from historical (circle) and NPDK(up) and KPA (bottom) simulations (boxplot) for Site 22 of the Colorado River yearly streamflow – maximum drought length, maximum surplus length, maximum drought amount, maximum surplus amount, and storage capacity

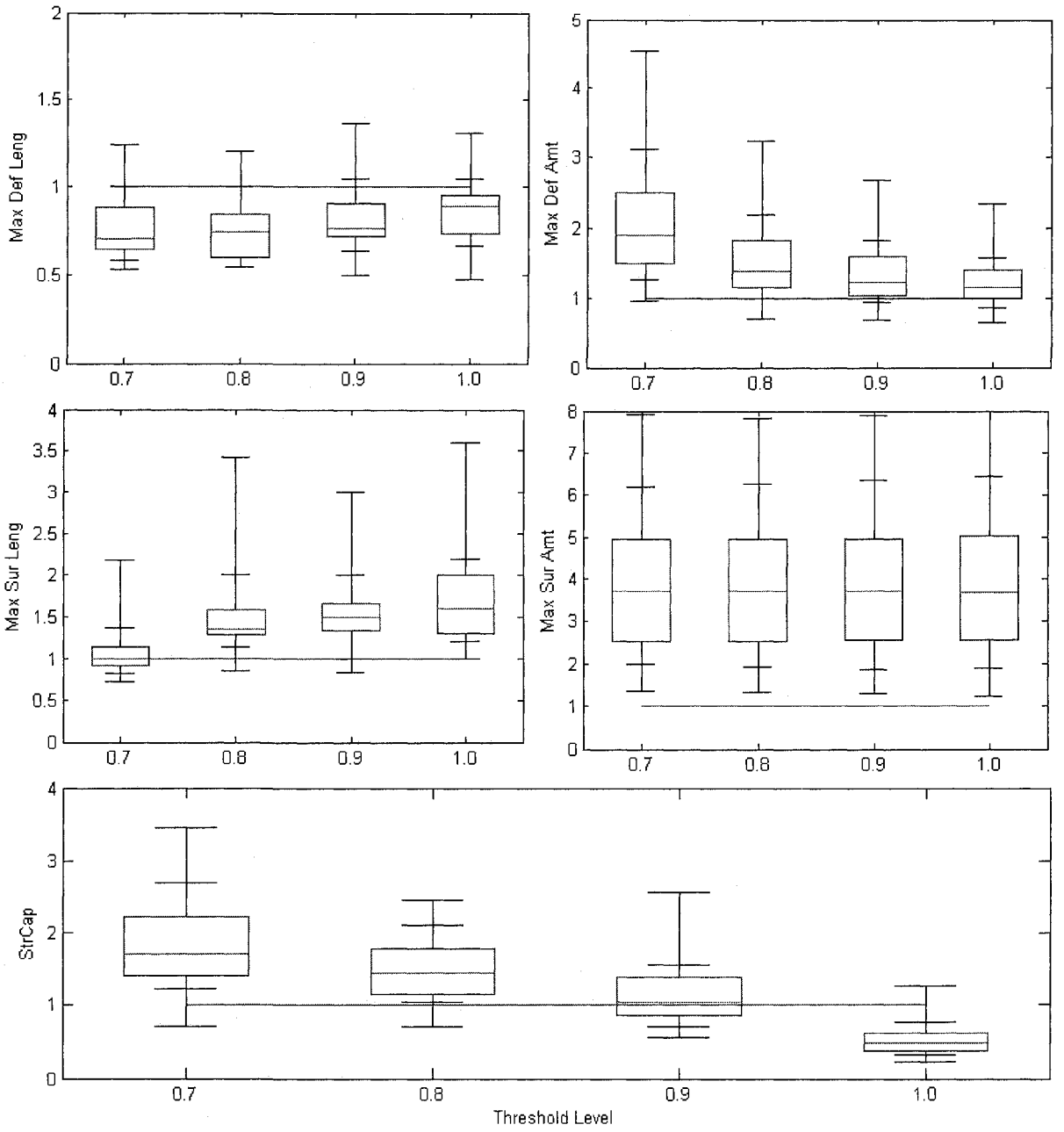


Figure 4-C.23 Monthly drought statistics at different threshold levels of Historical (dot line) and NPDK simulations (boxplot) for Site 21 of the Colorado River monthly streamflow

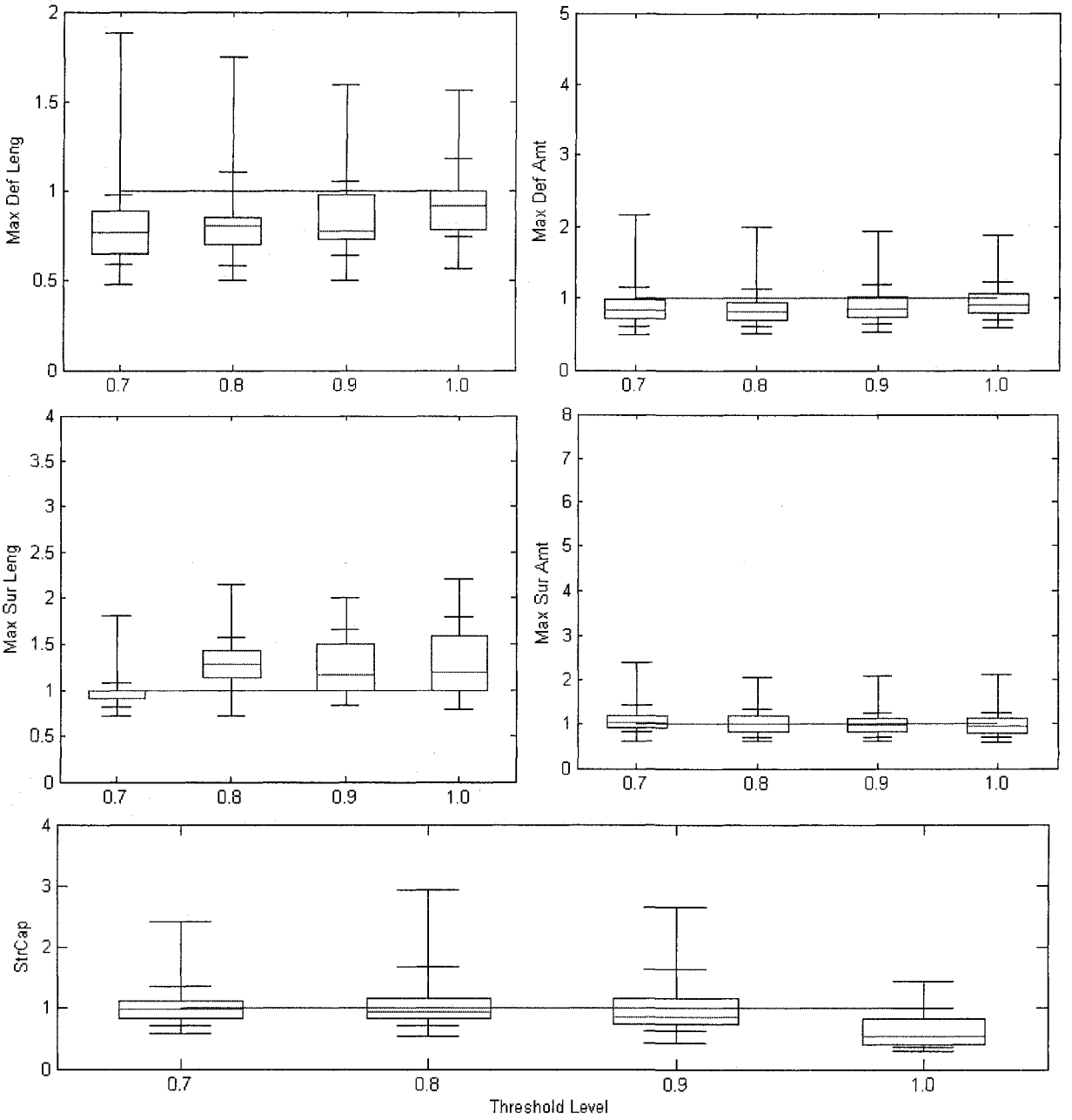


Figure 4-C.24 Monthly drought statistics at different threshold levels of Historical (dot line) and KPA simulations (boxplot) for Site 21 of the Colorado River monthly streamflow

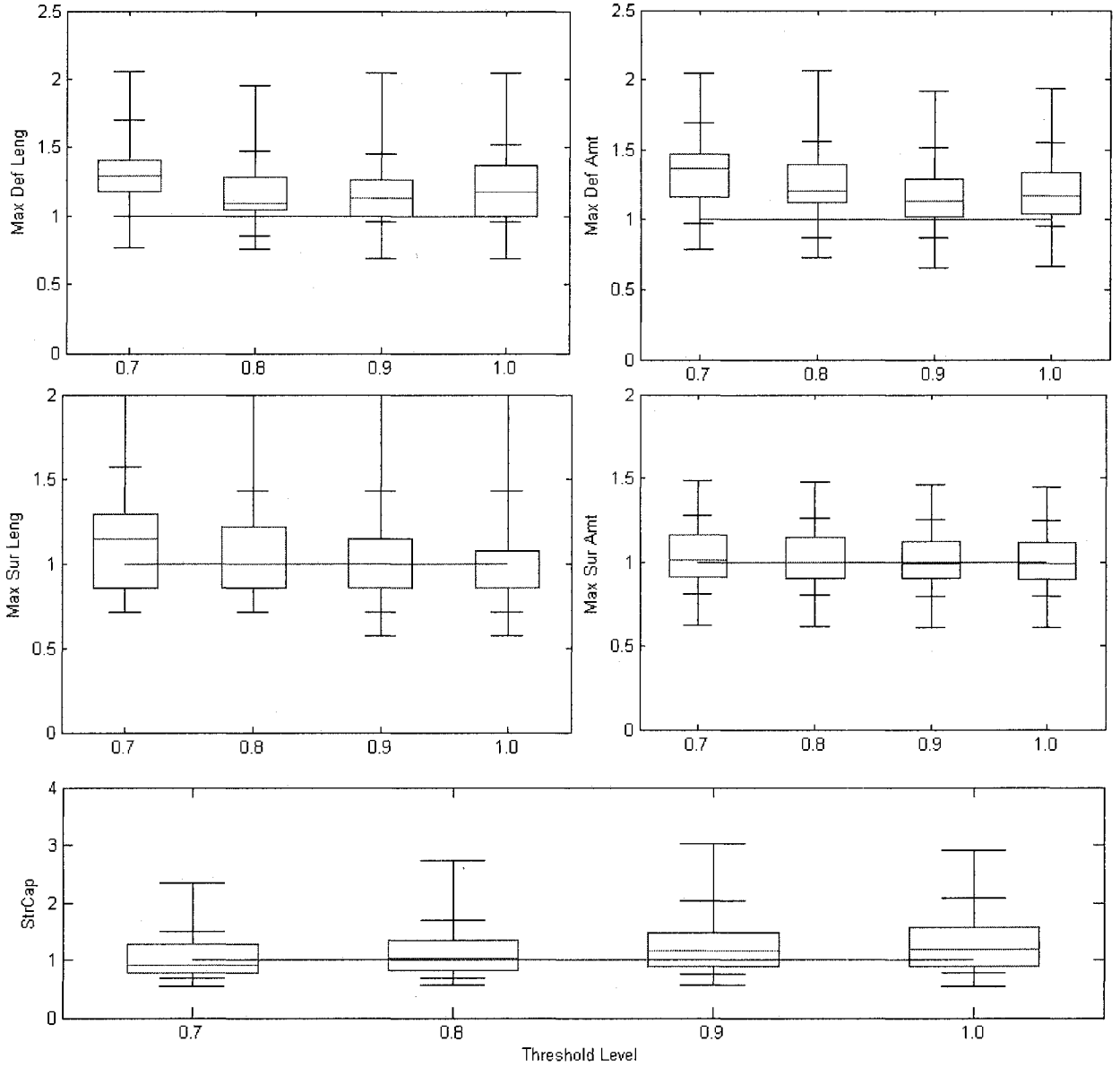


Figure 4-C.25 Monthly drought statistics at different threshold levels of Historical (dot line) and NPDK simulations (boxplot) for Site 22 of the Colorado River monthly streamflow

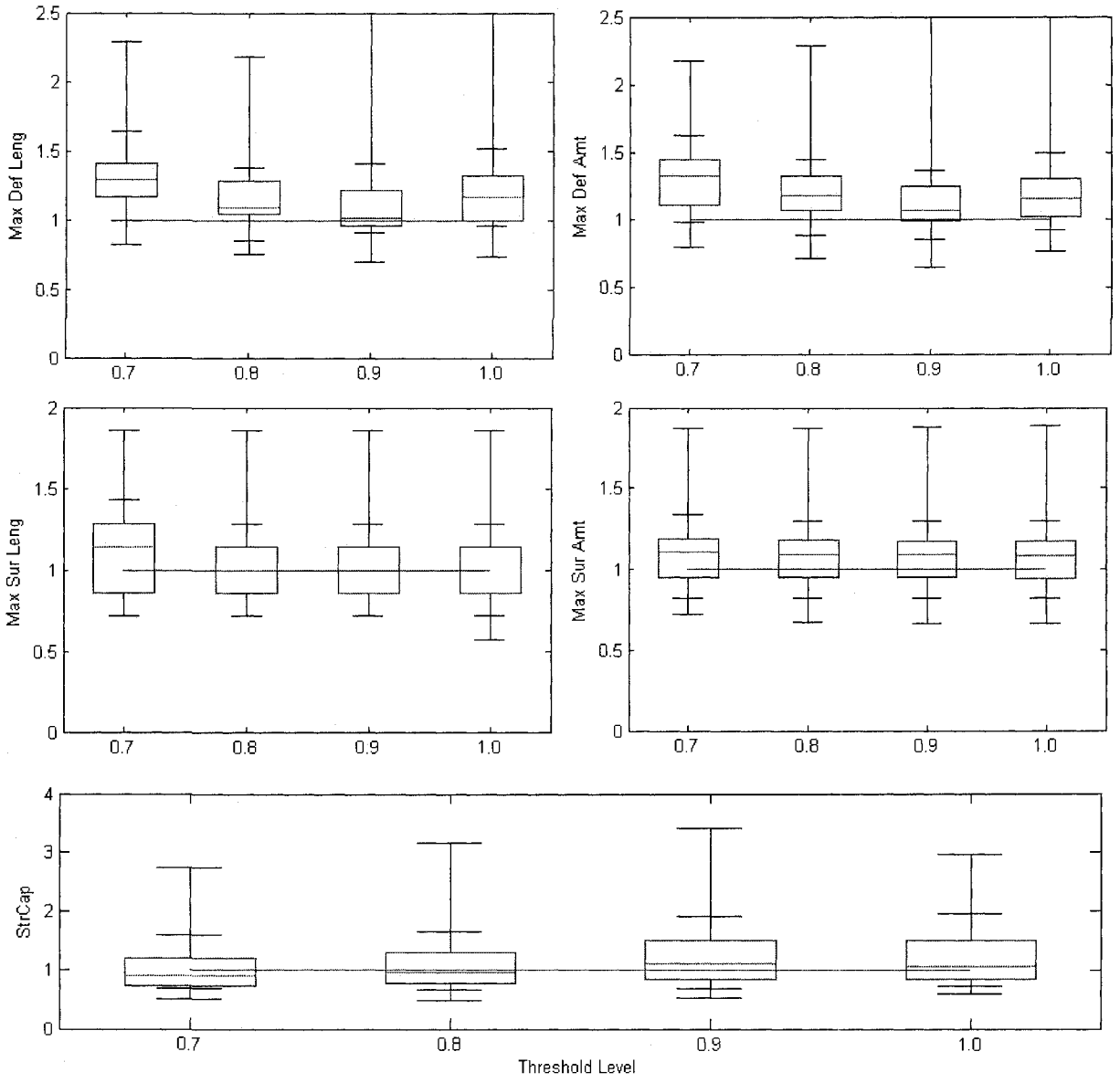


Figure 4-C.26 Monthly drought statistics at different threshold levels of Historical (dot line) and KPA simulations (boxplot) for Site 22 of the Colorado River monthly streamflow

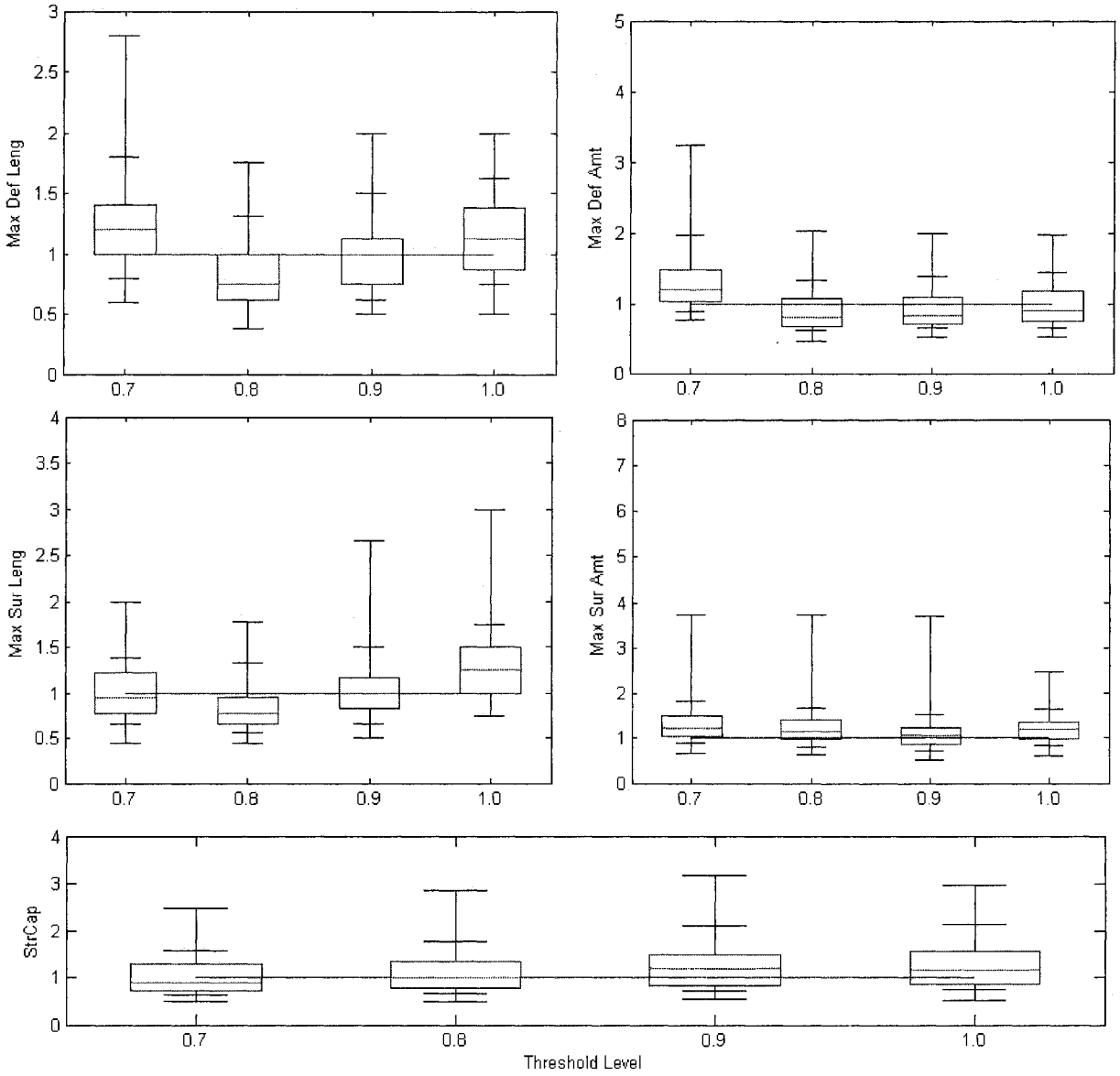


Figure 4-C.27 Yearly drought statistics at different threshold levels of Historical (dot line) and NPDK simulations (boxplot) for Site 22 of the Colorado River monthly streamflow

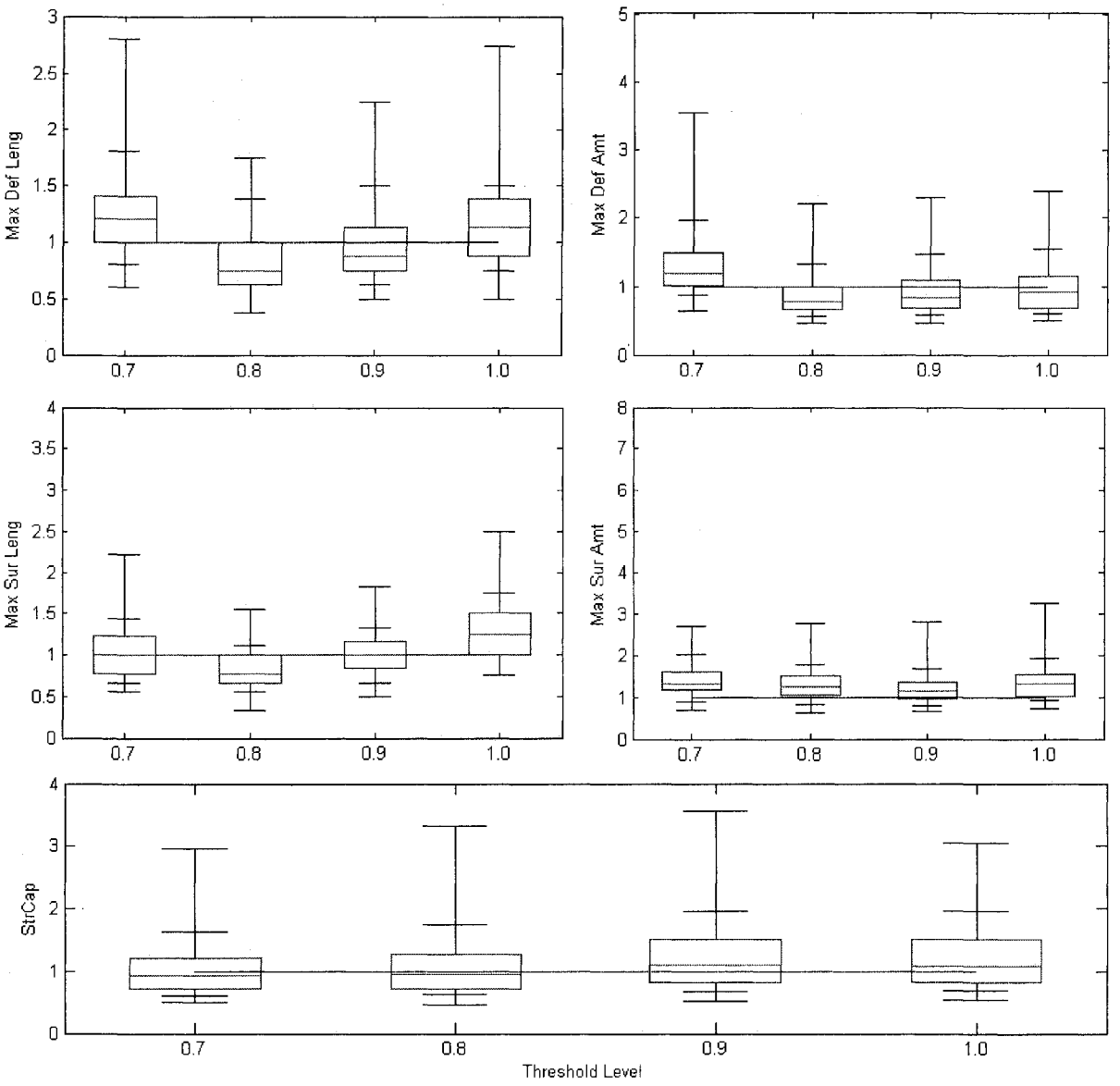


Figure 4-C.28 Yearly drought statistics at different threshold levels of Historical (dot line) and KPA simulations (boxplot) for Site 22 of the Colorado River monthly streamflow

CHAPTER V

DISAGGREGATION OF DAILY TO HOURLY PRECIPITATION

5.1 Introduction

The requirement of rainfall data of lower-level time resolution (daily, hourly) has been increased for the hydrological modeling and prediction, such as flood prediction and water quality assessment. Most precipitation data are measured with daily scale while the requirement of the detailed scale, such as hourly for detailed models, has been amplified. Relative shortage of hourly time scale hinders the task. Using measured daily precipitation data and their characteristics, some researchers have developed disaggregation schemes to fulfill the requirement (Hershenhorn J. and Woolhiser, 1987; Econonpouly et al., 1990; Bardossy, 1999). The developed approaches are applicable on the stationary basis through a day (24hours). But some sites, such as the Denver Airport site, have a significant diurnal cycle from a convective cyclic weather system through a day. The disaggregation scheme to consider the diurnal cycle would not have much attention. By the author's knowledge, no applicable disaggregation schemes with diurnal cycle have yet been presented in the literature. Therefore, the objective of this paper is to

develop and present the appropriate disaggregation schemes for daily rainfall to hourly rainfall with the diurnal cycle, to apply the schemes on different cases, and to investigate and compare the statistical behaviors of the schemes.

Daily rainfall disaggregation models can be applied to three cases in terms of the availability of daily rainfall data. The first case is for simulation. In this case, the complete hourly data exist for the entire year, obviously, as well as daily. Therefore, we can simply generate the daily rainfall and disaggregate the data into the hourly rainfall. The purpose of this case is rather different than the others. The second case is that only some portion of the hourly data is missed, while the entire daily data subsist. The goal is to disaggregate the daily data that the hourly data are missing. The third case is that only the daily rainfall data exist so that the disaggregation should be performed using the hourly rainfall characteristics of other sites, which have climatologically and regional proximity to the target site. In the application, we assumed that a portion (the second case) or all data (the third case) are missing so that we could check the model performance employing the difference between the model value synthesized from the proposed disaggregation model and the historical value and their statistical characteristics.

In Section 2 the history and developments of the disaggregation models are discussed and in Section 3 and 4 data and model description are shown, respectively. In Section 5 model performance and comparison are presented and followed by conclusive remarks in Section 6.

5.2 Literature Review

Daily rainfall disaggregation models have been developed and tested by researchers. The main categories could be as follows : the distribution-based approach, point process based approach, stochastic precipitation method using stored data, adjusting schemes, neural networks, and scaling cascade models, etc.

Betson et al.(1980) described a model to disaggregate daily rainfall into hourly rainfall. It requires a large number of transition probabilities. For the parameter-efficient model, Hershenhorn and Woolhiser (1987) and Econonpouly et al. (1990) developed a distribution-based approach. The method disaggregates daily rainfall into individual storms and simulates the number of rainfall events in a day and the amount, duration, and starting time of each event conditioned on the amount of that day and the preceding and following days. The model is fitted on the Walnet Gulch Experimental Watershed No5. The estimated parameter was used to fit the model for the other sites near the site presenting the applicable results according to 5% of the Kolmogorov-Smirnov significant test for the distributions. Arnold and Williams (1989) and Lane and Nearing(1989) proposed a simple model to simulate half hourly rainfall intensity from daily rainfall using a double exponential function to determine breakpoint. The disaggregation model assumes daily rainfall falls in only one event. Connolly et al (1998) used a similar methodology as Hershenhorn and Woolhiser(1987) but with a different distribution function. For a number of events Poisson distribution was applied and event duration with gamma distribution, event amount with exponential distribution, event starting time with beta distribution, and break point intensity within each event- double exponential, respectively. The approach was fitted on Biloela and Katherine data in Australia.

Accuracy of the model was determined by comparing the measured and simulated event characteristics and cumulative rainfall kinetic energy given in Rosewell(1986). The model was a reasonably accurate prediction of peak rainfall, so that it was adequate for input to infiltration and runoff models, while the accurate parameterization required enough high quality data.

A stochastic precipitation disaggregation method was developed to enforce the Upper Charles River Watershed by Socolofsky et al.(2001) using an hourly gauge near the watershed, the Logan Airport gauge. The method relies on measured hourly data in the same climatological regimes as the daily data to be disaggregated and samples the measured hourly data directly applied from the Logan Airport gauge. Therefore, the main task of this approach is to select appropriate event statistics from the nearby Logan hourly events database so that they sum to the daily total rainfall recorded within the watershed. It concludes that the technique performs well in supplying hourly rainfall data for use by continuous simulation watershed models and disaggregating distant gauges in a similar climate regime without any concern on diurnal cycle. Further, Choi et al. (2008) tested this model through the Texas region with slight modification.

Some researchers have developed the disaggregation scheme of daily rainfall based on the point process model (Glasbey et al, 1995; Koutsoyiannis and Onof, 2001; Cowpertwait et al., 1996). Glasbey et al. (1995) studied the disaggregation of daily rainfall by simulating long sequences of hourly data based on the Rodrigues-Iturbe et al. (1998) model and comparing the daily totals between all generated days and choosing the best match and then rescaling to match the daily total. The method applied to Edinburgh

(Turnhouse) and it was concluded that the model simulated realistic hourly data, while it does not involve the diurnal cycle.

Koutsoyiannis and Onof (2001) developed the model based on the Bartlett-Lewis process, adjusting the hourly values to obtain the required daily values. The adjusting method has been studied by Koutsoyiannis (1994), and Koutsoyiannis and Manetas (1996) such as proportional adjusting, linear adjusting, and power adjusting for different time scales. They also applied the lag-1 Gamma Autoregressive (GAR(1)) model to test the performance of the adjusting method revealing that, by some distant allowance between the value of the total sum of hourly generated rainfall and the target daily value to disaggregate, the process worked reasonably well. The performed scaling analysis with different aggregated level showed that the method reproduced most of the important statistics like variance, skewness, lag-1 autocorrelation, and dry probability, as well as mean, but no consideration on diurnal cycle.

Bo et al. (1994) performed the disaggregation of rainfall time series using Bartlett-Lewis rectangular pulses fitting on central Italy and Kentucky in the U.S. and remarked that the upper limit for the disaggregation scale for the model would be two days and these characteristics are related to the power law dependence of the power spectrum for timescales smaller than two days.

Bardossy (1999) developed a disaggregation scheme with three steps. First, the number of wet sub-periods is generated, conditioned on the total daily amount using the Polya distribution. Second, with Aitchisons relative distributions the relative precipitation amounts are generated arbitrarily. Finally, the generated amounts are rearranged to match

the autocorrelation function and the scaling properties using Markov Chain Monte Carlo (MCMC) based on Metropolis Hastings algorithm. Furthermore, Additional information such as atmospheric circulation patterns was taken into account for the disaggregation to improve the reproduction properties. The model was applied into Essen and Hennetalsperre in the Ruhr catchments (German). It disclosed that the scheme reproduced the autocorrelation function and the scaling properties properly. Furthermore, Bardossy and his colleagues have been using the Simulated Annealing (SA) to generate a precipitation time series (Bardossy 1997, 1998), and to disaggregate monthly to daily (Guenni and Bardossy 2002). In the papers, the objective was focused on fitting the scaling characteristics into historical data rather than periodicity or cycle.

The Neural Networks technique has been used to disaggregate the hourly rainfall data into sub-hourly time increments by Burian et al (2000,2001,2002). The model performed 15-min rainfall depth by training with performance measures such as signal-to-noise ratio. But this method does not include the intermittency in subdividing the hourly rainfall. Furthermore, Olsson J. et al (2004) used the Artificial Neural Network (ANN) to forecast a 12-hr period mean rainfall with wind speed at 850 hpa and predictable water. They separated the Neural Network approaches into two parts such as the intermittency and variability, and after each prediction, they were combined. The approach was applied into the Chikugo River basin in Japan. It concluded that two NNs in series improved the reproduction of intermittency and better performance during winter and spring than summer and autumn.

Scaling cascade models are applied for the rainfall disaggregation by Olsson (1998), Olsson and Berndtsson(1998), and Guntner et al. (2001). This model operates by

dividing each rainy time period into halves of equal length and distributing the rainfall volume between the halves. It was shown to reproduce the important fundamental characteristics such as the division of rainy and dry periods and the scaling behavior using the power spectrum, respectively. But the diurnal cycles could not be taken into account by the scale invariant properties, and for the halving characteristics of the model, the starting time interval should be not be the day but the value corresponding to power of 2 requiring a smaller scale data set. These multifractal random cascade approaches treat the data as a realization of a stochastic process possibly not being able to account for the uniqueness of the data set used. To overcome this problem, the notion of deterministic chaos and the related methods of data processing has been developed by Sivakumar et al.(2001)

In this study, among many of daily rainfall disaggregation models, the stochastic disaggregation methods are investigated and enhanced in order to preserve the diurnal cycle in hourly data as well as the basic statistics.

5.3 Model Description

Three stochastic models for disaggregating daily rainfall data into hourly are utilized and compared extensively for estimating hourly precipitation from daily. They are; (1) Conditional Markov Chain and Simulated Annealing based method (CMSA); (2) mixed periodic discrete autoregressive with gamma autoregressive PDAR(1)-GAR(1) model with Accurate Adjusting (PGAA); and (3) Stochastic selection method with Weighted Storm Distribution (SSMW). In describing the methods, we will use the same (or similar) notation as much as possible.

We denote by D_d the daily precipitation amount and $y_{d,\tau}$ is the hourly precipitation (amount) where d represents any particular day and τ represents any particular hour (of the day). They are related as:

$$D_d = \sum_{\tau=1}^{24} y_{d,\tau} \tag{5-1}$$

Also, we will use D_d^* to denote the precipitation occurrence of a given day, i.e. $D_d^* = 1$ if precipitation occurs, otherwise $D_d^* = 0$.

5.3.1 Conditional Markov Chain and Simulated Annealing (CMSA)

This method consists of two basic components. The first one accounts for the occurrence of hourly precipitation using conditional Markov Chain, that is, we use a transition probability for hourly precipitation conditioned on the daily state. The second component determines the amount of hourly precipitation. For this part we modified substantially the simulated annealing approach utilized by Bardossy(1997). His approach uses simulated annealing for determining the hourly precipitation occurrence and amount geared to preserving the source type of the scaling feature of precipitation. However, because our objective is to estimate hourly precipitation where the daily cycle may be a relevant feature, Bardossy's simulated annealing approach had to be modified. First of all, as indicated above, we used a different approach for determining the occurrence of hourly precipitation. Secondly, we used more realistic probabilistic model for fitting the transformed ratios of precipitation amounts and different objective functions so that the daily cycle can be accounted for. Thirdly, Bardossy (1997) used Polya distribution to define the number of wet hours connected with the aggregated daily amount. We tried the

same procedure but it was hard to deal with the daily cycle later on the simulated annealing. Instead, the following occurrence procedure was developed to handle the number and time of wet hours.

Occurrence Process

Natural hourly precipitation occurs whenever some amount of precipitation occurs in a given day (and vice versa). Because hourly precipitation is auto-correlated, we will use a Markov Chain for the modeling occurrence of precipitation at a certain hour conditioned on the state in the previous hour and the state of the corresponding day as follows:

$$P_{ij}(\tau) = P\{x_{d,\tau} = j \mid x_{d,\tau-1} = i, D_d^* = 1\} \tag{5-2}$$

where i, j =0,1 and $D_d^* = 1$ if D_d>0, otherwise, $D_d^* = 0$. $x_{d,\tau}$ is whether the current day (d) at a certain hour (τ) is rainy or not In addition, the conditional limiting distribution is denoted as

$$P_j(\tau) = P\{x_{d,\tau} = j \mid D_d^* = 1\} \tag{5-3}$$

The referred probabilities are estimated by counting. For instance, $P_{ij}(\tau)$ is estimated by

$$\hat{P}_{ij}(\tau) = \frac{n\{x_{d,\tau} = j \mid x_{d,\tau-1} = i, D_d^* = 1\}}{n\{x_{d,\tau-1} = i \mid D_d^* = 1\}} \tag{5-4}$$

where for example $n\{x_{d,\tau} = j \mid D^*_d = 1\}$ is the number of times in a certain day d where precipitation occurred, the precipitation during the hour τ is in state j given that the precipitation in the previous hour was in state i. Thus, the generation of the precipitation occurrence can be executed without trouble using the foregoing conditional transition probabilities.

Precipitation Amount

The daily precipitation amount must be divided into hourly quantities in such a way that their sum add up to the daily quantity. For this purpose, a transformation procedure (partition) suggested by Aitchison (1982) is employed. Three logistic transformations such as additive, multiplicative, and hybrid were described by Aitchison for compositional data. Here, the hourly precipitation data are transformed using additive logistic transformation. The hourly data are described as the ratios of daily data with the condition that the sum of ratios adds to unity in each day. Bardossy(1997) first applied Aitchison's procedure by fitting the normal distribution to the log-transformed region. Instead we fit a gamma distribution because of the fact that the ratios and the logs of the ratios are skewed. In our application, the maximum among the ratios in a certain day is specified so that the transformed data are bounded in a negative side and possible distributions to fit such bounded data are the gamma or log-normal. Obviously, the mathematical description as shown below applies only for the case where $D^*_d = 1$.

Assume that in a given day there are k hours of precipitation (not necessarily continuous). For example, Fig 1. shows a precipitation occurrence where k=3 and

precipitation occurrence at times τ =12, 17, and 18. In this case there will be three non-zero precipitation amounts $y_{d,12}, y_{d,17}, y_{d,18}$. Consider the ratios

$$R_j = \frac{y_{d,\tau}}{D_d} \text{ , j=1,..., k} \tag{5-5}$$

where R_j is defined only when $y_{d,\tau} > 0$. Then,

$$\sum_{j=1}^{k} R_j = 1 \tag{5-6}$$

In addition, let

$$W_j = \log \frac{R_j}{R_{\max}} \quad \text{, j=1,...,k} \tag{5-7}$$

where $R_{\max} = \max(R_1,....,R_k)$ and $W_j \le 0$. Although the normal distribution has been used to represent W_j (e.g. Bardossy, 1997), it is restricted because W_j is bounded (i.e. $W_j \le 0$). Thus, instead of the normal, the gamma distribution will be used (i.e. $-W_j \sim Gamma(\alpha, \beta)$). The parameters may be estimated from the data based on the method of moments or maximum likelihood (Kottegoda and Rosso, 1997). Notice that among W_j (j=1,...,k) there is one zero value when $R_{\max} = R_j$. This zero is always occurred at the hourly rainfall in a rainy day. This zero should be excluded in fitting and generation. Therefore, the number of the fitted k values for W_j (j=1,...,k) are k-1 at each rainy day.

From the model of the precipitation occurrence process, it is clear that occurrence and the specific times (hours) of such occurrence on a certain day are defined. Then, each portion $-W_j$ can be generated from the gamma distribution and retransformed to get R_j. Note that only k-1 values of R need to be generated because $R_{\max}$ can be acquired with the condition at (5-6). For example, if k=3, $-W_1$ and $-W_2$ are generated from gamma distribution, then set R_1, R_2, and $R_3 = R_{\max}$ and from Eq.(5-6)

$$\frac{R_j}{R_{\max}} = \exp(W_j) \qquad \text{j=1,...,k-1} \tag{5-8}$$

and we need one more equation to obtain the values (R_1, R_2, and $R_3 = R_{\max}$). The unity condition in Eq.(5-6) can do this role. Therefore from simple mathematics,

$$R_{\max} = \frac{1}{1+\sum_{j=1}^{k-1}\exp(W_j)} \tag{5-9}$$

and

$$R_j = R_{\max}\exp(W_j) \tag{5-10}$$

The procedure is illustrated in Figure 5.1, where a unit daily amount is divided into three hourly portions. Notice that R_3 is not necessarily $R_{\max}$. It is just ordered for convenience. For example, one can distribute $R_{\max}$ in any place (e.g. R_2 can be $R_{\max}$).

A suitable distributing scheme of the hourly precipitation amount is needed to account for the diurnal cycle in statistics such as the mean, variance, and skewness. This

will be accomplished by using Simulated Annealing (technique). The values of hourly precipitation will be rearranged and successively verified so that the final arranged values will preserve the main statistics and the diurnal cycle. Simulated annealing (SA) is based on the Metropolis algorithm. The following will give a brief sense on the Metropolis-Hastings and the Metropolis algorithm, which will be used in simulated annealing.

Let us assume that $f(z)$ is the distribution (density) of interest, where z is the value of the variable such that generally $z \in \overline{R}$ and $\overline{R}$ represents the real domain. The goal will be sampling z from the distribution $f(z)$. Once the random variable is generated, further statistical analysis can be performed over the generated values such as mean, standard deviation, and skewness. Commonly, the distribution $f(z)$ knows the analytical form but the distribution $f(z)$ cannot be generated from a general method. Therefore, an indirect type of approach is suggested to generate the variable z from the density $f(z)$. From the general generation method, such as the rejection method, (1) a random variable is generated from another distribution, say a proposal distribution ; (2) the generated value is kept or rejected with certain probability; and (3) if it is rejected, then the previous value is selected instead for current generation value. The procedure can be described as follows:

(a) Initialize the iteration counter to i=1 and the chain to z^0. The initial value z^0 is selected within the domain $\overline{R}$.

(b) Generate a proposed (or candidate) value z^* from the proposal distribution denoted as $q(z^* \mid z^{i-1})$. Notice that we use notation z^* instead of z^i because

it is not fixed at this point. In other words, the value is just a candidate at this step. The proposal distribution should be a distribution easy to generate. A uniform distribution with some range at the center of $z^{(i-1)}$ is generally used, i.e. $q(z^* \mid z^{i-1}) \sim Unif[z^{i-1}+\alpha, z^{i-1}-\alpha]$, where α is the certain range such as 1.

(c) Evaluate the acceptance probability as follows to determine whether to choose the candidate or not.

$$P(z^*, z^{i-1}) = \min\left(1, \frac{f(z^*)q(z^{i-1} \mid z^*)}{f(z^{i-1})q(z^* \mid z^{i-1})}\right) \tag{5-11}$$

Here, if the proposal distribution $q(z^* \mid z^{i-1})$ is symmetric (i.e. $q(z^{i-1} \mid z^*) = q(z^* \mid z^{i-1})$, typically it is true (e.g. uniform case)), then the acceptance probability is reduced to:

$$P(z^*, z^{i-1}) = \min\left(1, \frac{f(z^*)}{f(z^{i-1})}\right) \tag{5-12}$$

If Eq.(5-12) is used for the acceptance probability instead of Eq.(5-11), then this total procedure (a)-(e) is called Metropolis algorithm instead of M-H algorithm. In this case, the acceptance probability depends only on the ratio of the target distribution $f(z^*)/f(z^{(i-1)})$.

(d) Set $z_i = z^*$ with probability $P(z^*, z^{(i-1)})$ and $z_i = z^{(i-1)}$ otherwise

(e) Iterate (b) to (d) until the desired number of sampling values are obtained

From this algorithm, the sampling values from the desired density $f(z)$ of interest are obtained. More detail will be found at Press (2003) or Gelman et al. (2003).

Different from Metropolis or M-H algorithm, simulated annealing is used as an algorithm to find a value or location to minimize or maximize a certain objective function instead of generating a complex density. But the procedure is almost identical to the Metropolis algorithm. From the example in Figure 5.1, since we already know in which hour the rain occurs, the only necessary step is to rearrange those three values sot that the diurnal cycle in the key statistics is preserved from the generated data. To achieve this, two values among three as in Figure 5.1 are selected. Say, R_1 and R_3 are selected. Then, decide whether ' $R_1 = R_3$ and $R_3 = R_1$ ' or leave it as it is ' $R_1 = R_1$ and $R_3 = R_3$ '. Simulated annealing is employed here to determine whether two selected precipitation amounts greater than zero are switched or not. In a certain rainy day for more than one occurrence hour, two precipitation amounts are selected randomly and determined from simulated annealing whether the value is switched between each other or not, probabilistically. The specified target distribution for the simulated annealing is aimed at maximizing an objective function. The target distribution commonly utilized has the form

$$f_O(O) = K(T)\exp(-O/T) \tag{5-13}$$

where O is an objective function and T is the temperature T, and K(T) is the constant that makes the objective function $f_O(O)$ to be unity, i.e. $\int_O f_o(O) = 1$ (Ingber, 1993). The scheme is built to minimize the objective function O. In the foregoing formulation, T plays a key role in the algorithm for "cooling". The algorithm starts with a "hot"

temperature that gets cooler as the annealing continues until the objective function is stabilized. The term annealing is from a smith heating a steel material and hammering (annealing) until it is cooled to get the targeted shape. More details on MCMC and Simulated Annealing can be found in Robert and Casella (1999), Press (2003), and Ingber (1993). In applications, several values of temperatures are applied so that the temperature is hot enough (easy to change a shape) at the starting place and gets cooler with the appropriate speed to have enough time for shaping. There is no rule of thumb for those temperatures. Those temperatures are selected subjectively satisfying the necessary condition as the starting temperature T=1000 and the ending temperature T=0.001, and the diminution (or decreasing) factor DF= 0.99. The starting temperature gets cooler as the process continues with $T_i = T_{i-1} \times DF$ where i represents the iteration step. If T becomes smaller than the ending temperature, the process is stopped. In this study, only one set of values (initial temp, end temp, and diminution factor) are used as described above.

The objective function used herein is geared to preserve the variation of the hourly statistics of the precipitation amount :

$$O = w_1 \sum_{\tau=1}^{24} (\mu_\tau^* - \mu_\tau)^2 + w_2 \sum_{\tau=1}^{24} (\sigma_\tau^* - \sigma_\tau)^2 + w_3 \sum_{\tau=1}^{24} (\gamma_\tau^* - \gamma_\tau)^2 \qquad (5\text{-}14)$$

where $\mu_\tau, \sigma_\tau, \gamma_\tau$ are the historical hourly mean, standard deviation, and skewness, respectively and the statistics with the asterisk are the calculated value at the current state. w_1, w_2, and w_3 are weighting factors. One might try a different weighting on objective function. But the proper controlling of temperature in simulating annealing (slow

decreasing) will lead to make no difference by different weighting from the characteristics of the approach. Therefore, the unity weighting factor is used in application, i.e. $w_1 = w_2 = w_3 = 1$.

As mentioned, the main focus for the disaggregation problem is to decide whether the selected two values switch or not. For this swapping process, simulated annealing is employed. At first, the objective function $f_O(O)$ is estimated with before and after switching two values. If the after-switching objective function is larger, then switch the value. Otherwise only switch the value with the probability of $P_s = \frac{f(O_{after})}{f(O_{before})} = \frac{K(T)\exp(-O_{after}/T)}{K(T)\exp(-O_{before}/T)}$ and $P_s = \exp[(O_{before} - O_{after})/T]$ from Eq.(5-12) and (10). From this scheme, the minimization of the objective function is achieved, since the objective is to rearrange the hourly amounts partitioned from the daily amount in order to preserve the hourly statistics with the diurnal cycle.

Overall procedure for the disaggregation may be summarized as follows:

(1) The hourly occurrence process for a day is taken using Eqs.(5-2) and (5-3) if a certain day has some amount of rainfall. A number of occurrence events and the specific times are determined in this process.

(2) (k-1) number of W_j are generated from the gamma distribution as $-W_j \sim Gamma(\alpha, \beta)$.

(3) Initially, the generated hourly rainfall amount obtained from Eq.(5-8) should be randomly or subjectively distributed at first on the event occurred hours

determined from the stage (1). This arrangement of the hourly rainfall quantities is not important at this point, since it will be exchanged.

(4) Execute the swapping procedure:

i. Set up the initial Temperature (T=1000)

ii. Select two hourly amount values out of k amounts which are greater than zero

iii. Calculate the objective function before and after swapping as P_s

iv. Swapping or keeping the values according to the objective function and the acceptance probability (i.e. P_s>1), then switch those selected values, otherwise decide with generating u~unif[0,1] (i.e. if u< P_s, then swap.)

v. Cooling the temperature with the diminution factor (DF=0.99)

vi. Do (ii~v) until the critical low temperature (T=0.001)

(5) Repeat (1)~(4) until all daily data are disaggregated

The graphical description is displayed in Figure 5.1.

5.3.2 Product model with Accurate Adjusting (PGAA)

Koutsoyiannis and Manetas (1996) and Koutsoyiannis and Onof (2002) developed a simple and useful disaggregation strategy. The main constraint in

disaggregation is the additivity, thus, they suggested an adjusting scheme in which the lower scale process is generated (using a specific lower scale model and estimated parameters) e.g. a Periodic-Autoregressive(1) model to generate monthly flows, until their sum, i.e. the value of the higher scale process(e.g. annual flow volume) is within the range close to the higher target scale amount, and then the generated values are adjusted to procure the additivity condition. In our study, the lower temporal scale is hourly and the upper scale is daily, so two different scale models are selected, and then the disaggregation scheme of Koutsoyiannis and Manetas (1996) is applied to fulfill the additivity condition.

The model selected for the lower scale is chosen to preserve the hourly (periodic) statistics. The model is the product of a Periodic Discrete Autoregressive and a Gamma Autoregressive model denoted as PDAR(1)-GAR(1). The product model is denoted as

$$y_{d,\tau} = x_{d,\tau} z_{d,\tau} \tag{5-15}$$

And the PDAR(1) is a binary process ($x_{d,\tau}$) that represents the intermittency of the daily rainfall data. It is equivalent to a Periodic Markov Chain (Chebaane et al., 1995) i.e. the occurrence probability of the state at present time depends only on the state at the previous hour. But the occurrence probability varies throughout the day. The varying occurrence probability during the day is an important feature of the model. The stationary Gamma-Autoregressive(1) model is used after eliminating the periodicity of the daily cycle of the mean and standard deviation for the amount process ($z_{d,\tau}$). It implies constant skewness and lag-1 autocorrelation throughout the day. Typically, hourly rainfall is highly skewed and autocorrelated. The log-transformed AR(1) model can be

applied. But the transformation can lead the bias of the key statistics such as mean, standard deviation, and skewness.

Periodic Discrete First order Autoregressive and Gamma First order Autoregressive (PDAR(1)-GAR(1)) model is used for hourly generation, and the proportional adjusting procedure is applied for additivity constraints. The model description is separated with PDAR(1) for the occurrence process that is equivalent to the periodic Markov chain and GAR(1) for the amount process in Eq.(5-15).

PDAR(1) for occurrence process

The model can be described simply using PMC as follows:

$$\begin{cases} x_{d,\tau} = 1, \text{ if it rains during the hour } \tau \text{ of day d} \\ x_{d,\tau} = 0, \text{otherwise} \end{cases}$$

The periodic Markov Chain defined by the transition probabilities may be estimated using maximum likelihood estimation.

$P_{ij}(\tau) = P[x_{d,\tau} = j \mid x_{d,\tau-1} = i]$ where, i, j=0,1 and τ =1,..,24

and they may be estimated by maximum likelihood such as

$$\hat{P}_{ij}(\tau) = \frac{n_{ij}(\tau)}{n_i(\tau - 1)} \tag{5-16}$$

where $n_{ij}(\tau)$ implies that the number of data to change the state from i at τ-1 time to j at time τ, and $n_i(\tau - 1)$ is the number of the data at the state i at time τ-1. The limiting distribution can be defined for the starting generation as :

$$\hat{P}_i(\tau) = \frac{n_i(\tau)}{n(\tau)}, \text{ here } n(\tau) \text{ is the number of data at } \tau. \tag{5-17}$$

GAR(1) for amount process

The GAR-1 model has been applied on streamflow generation (Salas, 1993). It is non-Gaussian process so that it does not need to transform the data. In fact, the marginal distribution is gamma, and the model has been applied by standardization as $z'_{d,\tau} = \frac{z_{d,\tau} - \mu_\tau}{\sigma_\tau}$, where μ_τ and σ_τ are the hourly mean and standard deviation. After the generation from the GAR(1) model, the generated data should be reformatted with $z_{d,\tau} = z'_{d,\tau}\sigma_\tau + \mu_\tau$. Thus, the stationary GAR(1) is

$$z'_{d,\tau} = \phi_1 z'_{d,\tau-1} + \varepsilon_{d,\tau} \tag{5-18}$$

The marginal distribution would be $z'_{d,\tau} \sim \text{Gamma}(z_0, \alpha, \beta)$.

$$\varepsilon_{d,\tau} = z_0(1-\phi_1) + \eta_{d,\tau} \tag{5-19}$$

$$\begin{cases} \eta = 0 & \text{if M} = 0 \\ \eta = \sum_{j=1}^{M} E_j(\phi_1)^{U_j} & \text{if M} > 0 \end{cases} \tag{5-20}$$

where z_0, α, β are the location, scale, and shape parameters, respectively, ϕ_1 is the autoregression coefficient, M is distributed with Poisson, mean $-\beta\ln(\phi_1)$, E_j is the iid exponential variable with mean $1/\alpha$, and U_j is uniform(0,1). Parameter estimation is available based on the method of moments (Fernandez and Salas, 1990)

$$\hat{\beta} = \left(\frac{2}{\hat{\gamma}}\right)^2,\ \alpha = \frac{\hat{\sigma}\hat{\gamma}}{2},\ z_0 = \hat{\mu} - \hat{\alpha}\hat{\beta},\ \hat{\phi}_1 = r_1 \tag{5-21}$$

where $\hat{\mu}, \hat{\sigma}, \hat{\gamma}, r_1$ are the estimated mean, standard deviation, skewness, and lag-1 autocorrelation from the data in non-zero values (Salas et al. , 1980).

Since Katz and Parlange (1995) recommended to smooth out the diurnal cycle in the key statistics, the key statistics applied in the GAR(1) model including mean and standard deviation of the rainfall amount data, are smoothed with Fourier transformation as described in that paper.

Adjusting Procedure

Koutsoyiannis and Manetas(1996) proved that the proportional adjusting procedure might lead an accurate same marginal distribution for Gamma distribution with higher and lower scale. The amount procedure is a Gamma based model, GAR(1), even if the autocorrelation is concerned. Therefore, the proportional adjusting is backed up to apply rather than the power and linear adjusting procedures.

If the generated values are satisfied with the condition below Eq.(5-23), a proportional adjusting procedure is applied to the generated lower scale (hourly) values according to

$$y_{d,\tau} = y^*_{d,\tau}\left(\frac{D_d}{\sum_{\tau=1}^{24} y^*_{d,\tau}}\right) \tag{5-22}$$

where $y^*_{d,\tau}$ is the generated hourly value from the product model and D_d is the known daily value.

$$\left| \ln\left(\sum_{\tau=1}^{24} y^{*}_{d,\tau} \Big/ D_d \right) \right| \le \varepsilon^{adj} \tag{5-23}$$

where ε^{adj} is a tuning parameter compromising with time consumption and bias reduction. If ε^{adj} is too small, the acceptance chance, as in Eq. (5-23), is very rare so that the computation time gets longer. If it is too large, the final values are biased on the statistics, such as mean and standard deviation. Here the study for the tuning parameter is out of the scope for this paper. Therefore, only one small value (0.1) is used for this parameter.

The procedure can be summarized as:

(1) Estimate the parameters for the hourly model by using the methods and equations outlined above.

(2) Generate the hourly values from the model.

(3) If the aggregated amount of the candidate is within the critical value ε^{adj} ,as in Eq.(5-23), then adjust the current generated value using Eq.(5-22), otherwise repeat Step (2).

(4) Repeat (2)~(3) until the target data are wholly disaggregated.

5.3.3 Stochastic Selection Method with Weighted Storm Distribution (SSMW)

This method was developed by Socolofsky et al.(2001). It uses the measured lower scale (hourly) data set with the same climatological regime to disaggregate. The

method is based on separating the measured daily precipitation into events that are distributed uniformly with non-overlapping constraints (Socolofsky et al., 2001).

Without any adjustment, the method would preserve the stationary hourly statistics. The model, however, should concern the diurnal cycle so that the model should be periodic-based. The adjustment has been applied for the distribution of the storm events for the diurnal cycle developed by the authors of this paper. To account for the diurnal cycle, the storm events are distributed with the weights of the hourly occurrence probability $P_1(\tau)$ of Eq.(5-17). Therefore, the events will more likely occur where the probability is higher than other hours. This is the modification from the original model in Socolofsky et al.(2001) by the authors of this paper.

An event is defined as a continuous sequence of hourly precipitation separated by at least one dry hour period (i.e. separated from the non-event period). The step is summarized as :

(1) The hourly precipitation events in a day can be separated. The depth and duration (hours) of each event can be defined from the hourly data set. The events are ordered as the increasing depth of each event (e.g. $y_2 > y_1$). Set a table of the ordered events based on a depth of each event with the duration of each event, like the table on the right side of Figure 5.2, where n is the total number of events in the hourly data set. The events will be obtained from the whole historical data set. As mentioned, an event is a continuous sequence of hourly rainfall until it gets dry period.

(2) Get a cumulative density function F_j (CDF) of event depth employing the order and the depth in Figure 5.2 as $F_j = j/(n+1)$.

(3) The objective is to separate a daily precipitation amount D_d into several precipitation events. The separated events should be selected and defined from the historical data set which is on the table (Figure 5.2). To begin with, get the depth of daily precipitation D_d from a generated time series or from the current (historical) data and search the CDF ordinate ξ corresponding to D_d as shown in Figure 5.2 (right side).

(4) Generate ζ ~unif[0, ξ] and find the corresponding event amount from the CDF ς (Figure 5.2 – right side),which is y'_d. The event corresponding or closest to the depth y'_d should be defined from the table and stored as an event with its duration.

(5) Set $D_d = D_d - y'_d$ and repeat (3)~(4) until $D_d \le \varepsilon^{dat}$, where ε^{dat} is the stopping critical value. This critical value should be larger than, or at least the same as, the smallest value of the events in the table. If the condition $D_d \le \varepsilon^{dat}$ is met, then the repetition should be stopped. The residual is distributed uniformly through hours . ε^{dat} can be used as a calibration parameter using different values of ε^{dat}, but here the effect of the calibration parameter is not checked.

(6) Distribute the event according to the hourly occurrence probability ($P_1(\tau)$,Eq.(5-17)) with a constraint that the current value should not overlap to the next day. The probability that an event starts at a certain time τ is:

$$\mathrm{P}\{\text{an event starts at a certain hour } \tau\} = \frac{P_\tau(1)}{\sum_{j=1}^{24} P_j(1)}, \quad \tau = 1,\ldots,24$$

If the events overlap inside, they should be added together. The final remaining value, which is less than ε^{dat} , is added for one of 24 hours with discrete uniform selection.

More details are referred to by Socolofsky et al. (2001).

5.4 Applications and Model Performance Criteria

5.4.1 Applications

Three cases are illustrated in which disaggregation of daily precipitation may be needed: (1) for simulating hourly precipitation whereby daily precipitation is generated first, which in turn is disaggregated into hourly quantities; (2) for record extension of hourly data where longer daily data are available; (3) for estimating hourly data at a given station where daily data are available given that hourly data are available at another station. For the first case hourly rainfall data, Denver International Airport (DIA), which are available for the period (1949-1990) are employed. For the second case, the hourly and daily data of the DIA station are used. Assuming that hourly data for the record 1979

to 1990 are missing that record extension for that record is needed. For the third case, daily and hourly precipitation at DIA assumed that only daily data at Parker are available, and hourly data will be obtained by disaggregating from the daily data employing the disaggregation for the parameter derived for the data at DIA. DIA is station number 5220, and Parker is 56326 obtained from U.S. Hourly precipitation data from NOAA.

Some occurrence probabilities and basic statistics of the hourly precipitation data for the month of July for Denver Airport for the period 1949-1990 (whole record), 1949-1978, and 1979-1990 are shown in Table 5.1 and Table 5.2, as well as Parker. The same statistics are shown in Table 5.1 and Table 5.2 for the hourly precipitation for the periods 1949-1978 and 1979-1990, respectively. Note that in the period 1979-1990, precipitation has not occurred in the hours 10, 11, and 12. Likewise, the same statistics are shown in Table 5.1 and Table 5.2 for hourly precipitation in July for the period 1950-1994 for Parker near the airport. in Table 5.1 and Table 5.2 unveil the difference and resemblance in statistical behavior of hourly precipitation. The Denver site shows a more frequent precipitation occurrence while the precipitation amount at Parker is generally larger than at Denver (Figure 5.3). Both sites depict the diurnal cycle. The DIA hourly precipitation data has been assessed by Collander et al. (1993) and employed for testing the developed hourly precipitation generation model by Katz and Parlange (1995).

The model validation is performed for the three cases, and the statistical criteria are selected for checking the validity of the models. The investigation is not only on the difference between the real value and the disaggregated value, but also on the difference between the statistics of the historical data and the statistics of the disaggregated data.

5.4.2 Model Performance and Validation Criteria

To judge the performance of a given model and to compare among the models, two types of statistical criteria have been used. The first ones are statistics that compare the observed hourly precipitation values versus the predicted hourly values obtained from a disaggregation model. The second ones are statistics that compare the hourly precipitation statistics (e.g. mean) from the observations versus those estimated from the predicted hourly values (obtained from a disaggregation model). Although the statistical criteria are well-known and may be found in many references, we summarize them below.

A few of model performance criteria used are explained. Mean absolute error (MAE), and root mean square error (RMSE) are among the most popular criteria for checking the model performance. Model 1 may be selected over Model 2 if the MAE and RMSE of model 1 are better than the comparing values for model 2. Another common performance criterion is the coefficient of determination R^2, but it only evaluates the linear relationships between variables. R^2 is insensitive to additive and proportional differences between the model simulations and observations. To overcome these shortcomings, Legates and McCabe(1999) recommended using the baseline-adjusted modified coefficient of efficiency(MCE). It is defined as:

$$MCE = 1.0 - \frac{\sum_{i=1}^{n} |O_i - P_i|}{\sum_{i=1}^{n} |O_i - \bar{O}'|} \qquad (5\text{-}24)$$

where O_i= measured observations; P_i=model predictions; $\bar{O}'$ =baseline value of the observations against which the model is to be compared (e.g. mean) ; and n is the number

of observations to be compared. Values of MCE between $-\infty$ and 0 represent that the predicted values are not better than the baseline values, while values of MCE larger than 0 and approaching to 1 imply better fitting.

In addition, the index of agreement d, (Willmott, 1981) has been used for checking the model performance. It is given by:

$$d = 1.0 - \frac{\sum_{i=1}^{N} |O_i - P_i|}{\sum_{i=1}^{N} |O_i - P_i| + |O_i - \overline{O}'|} \tag{5-25}$$

It varies from 0.0 to 1.0, with higher values indicating better agreement between the model and observations, similar to the interpretation of the coefficient of determination R^2. The index has been modified not to be sensitive to extreme values from the squared differences so it represents an improvement over R^2.

Furthermore, the second type of statistical criteria as referred above is summarized here:

$$MAE_{\theta} = \frac{1}{24} \sum_{\tau=1}^{24} \left| \theta_{\tau}^{O} - \theta_{\tau}^{G} \right|$$

$$RMSE_{\theta} = \sqrt{\frac{1}{24} \sum_{\tau=1}^{24} (\theta_{\tau}^{o} - \theta_{\tau}^{G})^2}$$

$$MCE = 1.0 - \frac{\sum_{\tau=1}^{24} \left| \theta_{\tau}^{O} - \theta_{\tau}^{G} \right|}{\sum_{\tau=1}^{24} \left| \theta_{\tau}^{O} - \overline{O}' \right|}$$

$$d = 1.0 - \frac{\sum_{\tau=1}^{24} \left| \theta_\tau^O - \theta_\tau^G \right|}{\sum_{\tau=1}^{24} \left(\left| \theta_\tau^G - \bar{O}' \right| + \left| \theta_\tau^O - \bar{O}' \right| \right)}$$

where θ_τ^o represents a historical statistic for hour τ estimated from the observed hourly values and θ_τ^G represents a statistic for hour τ estimated from the predicted hourly values obtained from the disaggregated values of a given model. The hourly statistics included mean, standard deviation, skewness, and $P_\tau(1), P_\tau(0,1), P_\tau(1,1)$. For the base-line value, the mean value of observed hourly statistics was used.

Those performance criteria shows employing boxplots. All box plots in this paper have the same representation, such that the box and the middle line in the box depicts the quartiles and median, respectively, while the straight line is stretched until the max and min value from the box.

5.5 Results

Three different cases of disaggregations were tested as mentioned with the developed three models. The simulated, extended, and transferred data for each disaggregation case were investigated and employed to validate three proposed models through comparing the statistical features with those of the historical data.

For the first case, the first part of the Denver hourly record (1949~1990 July) was employed for model fitting and parameter estimation of disaggregation model. The objective was to build up a disaggregation model so that the disaggregated data would preserve the statistical characteristics through the higher time scale (daily), as well as the

lower time scale (hourly). The necessity of this case can be described as follows. One might build a daily model and generate daily data. Then, the daily data are disaggregated to obtain the lower time scale data, such as hourly. This disaggregation is named as a data simulation case. The transition probability for the CMSA model at Eq.(5-2) and Eq.(5-3) is estimated and shown at Table 5.3 for the whole record in Denver. The 100 sets of the same length of the record as the Denver (42yrs, 31days) have been disaggregated into hourly data. One month example is shown at Figure 5.4 (one more example in Figure 5-A.1). The statistics from the disaggregated data sets are estimated and compared with the historical data. In Figure 5.5, Figure 5.6, and Figure 5.7 for mean, skewness and P_1, CMSA and PGAA model reproduces the basic statistics and occurrence probabilities with the diurnal cycle. SSMW underestimate the mean and standard deviation during the frequent rainfall hours (i.e. hours 14-22, referred to Table 5.2), but overestimate the skewness and the transition probabilities (P_{01} and P_{11}) during the infrequent hours (i.e. hours 1-12). The occurrence probabilities (P_1) of the historical through the whole hours are overestimated as shown Figure 5.7. This is the evidence of the model such that the diurnal cycle is not properly reproduced from this model. The reason is that multiple numbers of events selected from the historical data cannot be arranged with preserving the diurnal cycle. The other figures for standard deviation, P_{01}, and P_{11} can be found from Figure 5-A.2 to Figure 5-A.4. Notice that the basic statistics in the PGAA model are reproduced but with smoothed manner as expected from the modeling process (smoothing the key statistics for the diurnal cycle). The mean of the performance criteria of the occurrence probabilities and the key statistics are shown in Table 5.4 and the boxplots of the performance criteria for mean and P_1 are shown at Figure 5.8 and Figure

5.9. Further results are shown from Figure 5-A.5 to Figure 5-A.8. It is shown that the CMSA and PGAA model have similar behavior in the preservation of the occurrence probabilities. It is because those two models employ the transition probability matrix and limiting distribution in generating the occurrence process with a different format (CMSA: conditional transition probability and PGAA: PDAR(1) model). And the basic statistics such as mean, standard deviation, and skewness are well preserved in the CMSA and PGAA. PGAA has the priority to be selected according to the performance criteria as shown in Table 5.4. The historical Denver hourly data can be compared with the disaggregated data. Even if the suggested models are not for forecasting, the comparison can be made to obtain the model performance and features. The results are in Table 5.5. The criteria revealed the priority on the SSMW model. Note that the value comparison should be carefully made on this case because the data are intermittent, which include a lot of zero values. But the models do not capture the specific time of occurrence. The models only disaggregate the daily rainfall into hourly stochastically not forecasting. If the zero values are compared with some amount, the criteria might be biased on a model, which has more frequent events and less peak points on amount. The statistical behavior of the SSMW model is to have more frequent on occurrence and less frequent on large values, which makes significant difference between two other models in the comparison of the performance criteria with the disaggregated values. This feature can be read from Figure 5.10 in that the events spread out through all hours the disaggregated time series for the SSMW model. Therefore, choosing a model with this comparison might not be reliable.

The second case can be explained as follows (named extension case). If one has longer length of the record in a higher time scale (daily) and also has shorter length of the record in a lower time scale (hourly), the shorter record may be extended until the record of the higher time scale is available. The objective of the model for this case is to extend the unmeasured record during the available daily data employing the statistical behavior of the measured hourly record. Half of the data in Denver are used for model fitting and the other half are assumed not to be measured at a lower time scale (hourly) but measured at a daily time scale. The assumed non-measured data are extended using the disaggregation models. The statistics of the disaggregated data and the values are compared with the real historical data. Again, the disaggregated data can be compared with the real historical data. In Table 5.4, the results are shown with similar behavior in the first disaggregation case (simulation). The occurrence probabilities, P1 and P01, are better matched on CMSA, while P11 is better on the CMSA and PGAA. In Figure 5.11, an example month of the disaggregated time series and historical data for daily and hourly are shown (One more example is in Figure 5-A.1). The similar distributional characteristic as the simulation case in SSMW is shown. In the figure, there are two days in which the event occurs. The historical hourly time series has shown that the rainfall events occurred through a small number of hours. This can be explained by the behavior of the diurnal cycle. The rainfall tends to occur in certain frequent hours in a day, such as hours 16 to 22. This could be related with the performance criteria on the comparison of the generated value and historical value which is represented in Figure 5.12. The same conclusion as the simulation case with Figure 5.10 is leaded such that the SSMW model tends to spread the daily rainfall into hourly rainfall uniformly. In Figure 5.13, the

standard deviation of the historical and disaggregated hourly data for the extension case is shown. Completed plots are shown from Figure 5-A.10 to Figure 5-A.15. It is shown that the standard deviation underestimated during the frequent rainfall hours (i.e. 14-21) as in Figure 5.13. This is the same behavior as the simulation case. The other statistics (mean, skewness, and occurrence and transition probabilities) of this extension case behaves similarly as the simulation case. The disaggregated hourly data from SSMW are rather distributed evenly compared with the other models (low mean value and high occurrence probability through the hours). Completed plots are shown from Figure 5-A.16 to Figure 5-A.21.

The third case is to disaggregate the daily precipitation data at another site using a Denver parameter set, named as a data transfer case as presented by Econopouly, et al. (1990). For this purpose, the climatologically and regional proximity station of rainfall data, Parker, was chosen. The statistical characteristics of the hourly rainfall data for this station are shown in Table 5.1 and Table 5.2. The aim of this transfer case is to fit a disaggregation model with the hourly and daily data with the data of a different site (Parker) and the daily rainfall data of the target station (DIA) are disaggregated into hourly data. Two examples of this disaggregation case are shown at Figure 5-A.22. The same criteria and basic statistics are estimated and compared at three models as the previous cases. The performance criteria of the occurrence and transition probabilities and key statistics are shown in Table 5.4. P1 and P01 are better fits on CMSA and P11 on PGAA. There are some negative values which imply that the disaggregated values from the models are not better than the case in which the values are filled with the baseline values such as the historical mean. PGAA seems to be prior to the other models except

the index of agreement d which chooses CMSA as a better one. Figure 5.14 illustrates the performance criteria of the mean. From the figure, the variability of the estimated criteria seems to be lager than the ones in the other cases. It should be noted that the larger uncertainty is shown because the disaggregation was performed using the statistics of another site compared to the first simulation case. Figure 5.15 show the mean of the historical and the disaggregated hourly data for three models and also for the historical mean of the Parker. The others such as standard deviation, skewness, and occurrence and transition probabilities are also estimated. Those are shown from Figure 5-A.23 to Figure 5-A.27. The mean and the variation of the mean through hours are similar between two stations so that the employment of Parker station as the reference site should be feasible. The key statistics and occurrence and transition probabilities are relatively well preserved in the disaggregated data of CMSA and PGAA models while SSMW has the same deviation as the previous cases (i.e. the simulation and extension cases). The similar behavior of the performance criteria is observed as the previous simulation and extension cases. Those are shown at from Figure 5-A.28 to Figure 5-A.34.

5.6 Summary and Conclusions

Three distinctive disaggregation models were developed employing the current available technologies in order to reproduce the diurnal cycle embedded in the key statistics and occurrence probabilities. The first model, CMSA, is based on the conditional Markov Chain for the occurrence process and simulated annealing for the amount arrangement, which combines the binary process with the conditional probability matrix and the amount process rearranging the quantities with simulated annealing. The second model, PGAA, is PDAR(1)-GAR(1) with an accurate adjusting procedure. The

model generated the hourly rainfall data for a day with PDAR(1)-GAR(1) and the generated value was taken or regenerated according to whether the summation of the generated hourly is smaller than the tuning parameter. The third one, SMWSD, is the stochastic selection with weighted storm distribution. From the event table and CDF of event depth produced from the hourly historical data, historical events were selected until the summation of the events was smaller than the calibration parameter.

Three applications were experimented with for three models such as a simulation case, an extension case, and a data transfer case. Four performance criteria over the generated values, occurrence probabilities, and key statistics were estimated and compared with three proposed disaggregation models at each case.

From the results, some critical remarks are presented as follows:

a. PGAA and CMSA well preserve the occurrence probabilities. Two models employ a similar model for occurrence process.

b. PGAA is superior to reproduce the key statistics of hourly data with diurnal cycle.

c. The proposed disaggregation models are not useful for forecasting, since the exact rainfall occurrence in a day cannot be defined. It is not critically important as long as the disaggregated data preserves the diurnal cycle, the key statistics, and the occurrence probabilities.

The developed disaggregation model could be applicable and considerable to split daily rainfall data where a diurnal cycle is predominant.

Table 5.1 Hourly Precipitation occurrence probabilities

	DIA (1949-1990)			DIA (1949-1978)			DIA (1979-1990)			Parker (1950-1994)		
Time	p1	p01	p11	p1	p01	p11	p1	p01	p11	p1	p01	p11
1	0.018	0.006	0.708	0.02	0.008	0.632	0.013	0	1	0.019	0.009	0.538
2	0.016	0.005	0.714	0.016	0.003	0.8	0.016	0.008	0.5	0.019	0.007	0.615
3	0.012	0.002	0.8	0.012	0.003	0.727	0.011	0	1	0.014	0.004	0.737
4	0.009	0.004	0.583	0.009	0.004	0.5	0.011	0.003	0.75	0.006	0.001	0.889
5	0.012	0.005	0.533	0.01	0.003	0.667	0.016	0.011	0.333	0.006	0.003	0.5
6	0.008	0.004	0.545	0.008	0.003	0.571	0.011	0.005	0.5	0.006	0.005	0.222
7	0.009	0.005	0.5	0.01	0.005	0.444	0.008	0.003	0.667	0.006	0.004	0.333
8	0.011	0.005	0.5	0.011	0.007	0.4	0.011	0.003	0.75	0.007	0.004	0.4
9	0.008	0.002	0.8	0.01	0.002	0.778	0.003	0	1	0.009	0.005	0.417
10	0.005	0.001	0.857	0.008	0.001	0.857	0	0	0	0.008	0.007	0.182
11	0.004	0.001	0.8	0.005	0.001	0.8	0	0	0	0.011	0.011	0.063
12	0.003	0.002	0.25	0.004	0.003	0.25	0	0	0	0.008	0.006	0.273
13	0.009	0.009	0.083	0.012	0.011	0.091	0.003	0.003	0	0.012	0.008	0.353
14	0.031	0.029	0.1	0.032	0.03	0.1	0.027	0.025	0.1	0.032	0.026	0.205
15	0.058	0.047	0.227	0.067	0.059	0.177	0.035	0.019	0.462	0.052	0.035	0.361
16	0.068	0.044	0.404	0.068	0.038	0.476	0.07	0.058	0.231	0.049	0.03	0.412
17	0.084	0.055	0.409	0.078	0.051	0.397	0.099	0.063	0.432	0.057	0.034	0.438
18	0.078	0.041	0.52	0.073	0.042	0.471	0.091	0.038	0.618	0.055	0.035	0.403
19	0.082	0.042	0.533	0.077	0.038	0.542	0.094	0.05	0.514	0.043	0.019	0.567
20	0.081	0.043	0.505	0.072	0.038	0.507	0.102	0.057	0.5	0.04	0.024	0.429
21	0.048	0.021	0.587	0.052	0.026	0.521	0.04	0.008	0.8	0.042	0.024	0.448
22	0.042	0.018	0.6	0.041	0.016	0.632	0.046	0.023	0.529	0.034	0.016	0.563
23	0.034	0.01	0.727	0.033	0.01	0.71	0.035	0.008	0.769	0.027	0.011	0.595
24	0.034	0.012	0.659	0.032	0.012	0.633	0.038	0.011	0.714	0.022	0.004	0.8

Table 5.2 Basic Statistics of the Amount for Hourly Precipitation

	DIA (1949-1990)			DIA (1949-1978)			DIA (1979-1990)			Parker (1950-1994)		
	Mean	Std	Skew	Mean	Std	Skew	Mean	Std	Skew	Mean	Std	Skew
1	1.598	1.932	1.863	1.39	2.015	2.262	2.388	1.488	0.057	1.729	1.999	1.578
2	0.581	0.783	3.559	0.542	0.915	3.341	0.677	0.308	-0.06	1.456	1.103	0.082
3	0.66	0.566	0.964	0.716	0.651	0.66	0.508	0.207	0	2.634	4.142	2.343
4	0.783	0.689	1.479	0.603	0.469	1.012	1.143	0.984	0.795	1.101	1.047	0.576
5	0.66	0.643	1.975	0.508	0.402	2.013	0.889	0.891	1.219	1.365	1.309	0.386
6	0.947	1.035	1.117	0.617	0.584	1.299	1.524	1.481	0.058	1.58	1.141	-0.24
7	0.572	0.377	0.598	0.593	0.421	0.485	0	0	0	1.919	0.881	-0.75
8	1.07	1.777	3.087	1.295	2.079	2.502	0.508	0.359	0.817	1.092	1.023	0.736
9	1.219	1.509	1.577	1.326	1.56	1.429	0	0	0	1.63	1.007	-0.28
10	1.27	1.278	1.754	1.27	1.278	1.754	0	0	0	1.663	1.025	-0.28
11	1.727	1.782	0.586	1.727	1.782	0.586	0	0	0	2.635	2.353	2.083
12	0.635	0.328	0	0.635	0.328	0	0	0	0	0.993	0.868	0.868
13	1.609	1.905	1.197	1.709	1.965	1.064	0	0	0	2.809	3.271	2.166
14	2.153	2.946	2	2.32	3.338	1.721	1.651	1.156	0.119	3.331	3.404	2.99
15	2.791	4.913	2.821	2.802	5.169	2.824	2.735	3.607	1.777	3.563	5.756	3.848
16	2.078	3.29	3.19	1.879	2.655	2.678	2.56	4.503	2.852	3.485	4.564	2.145
17	1.942	4.356	6.567	2.06	4.967	6.481	1.709	2.831	2.074	4.099	6.577	3.885
18	3.529	5.995	2.613	4.004	6.317	2.011	2.577	5.251	4.605	3.741	5.135	2.195
19	1.947	3.496	4.207	2.039	3.026	2.61	1.756	4.351	5.092	2.718	4.49	3.896
20	1.933	3.665	3.277	2.051	3.916	3.284	1.725	3.212	3.017	3.207	6.335	3.868
21	2.157	3.782	4.106	2.111	3.912	4.397	2.303	3.455	2.588	4.257	6.655	2.943
22	2.203	3.223	2.517	1.578	1.675	1.123	3.601	5.056	1.374	2.72	3.601	2.298
23	1.351	2.004	3.096	1.098	1.783	3.329	1.954	2.426	2.703	1.977	1.65	1.144
24	1.409	2.58	3.357	1.143	2.285	3.827	1.978	3.14	2.705	2.481	3.552	2.961

Table 5.3 Transition probabilities and limiting probabilities conditioned on $D_d^* = 1$ of CMSA in Eqs.(5-2) and (5-3) for DIA hourly precipitation for the period (1949-1990)

Hour	1	2	3	4	5	6	7	8	9	10	11	12
P1	0.063	0.054	0.039	0.031	0.039	0.028	0.031	0.036	0.026	0.018	0.013	0.010
P01	0.019	0.016	0.008	0.013	0.019	0.013	0.016	0.019	0.005	0.003	0.003	0.008
P11	0.708	0.714	0.800	0.583	0.533	0.545	0.500	0.500	0.800	0.857	0.800	0.250
Hour	13	14	15	16	17	18	19	20	21	22	23	24
P1	0.031	0.103	0.193	0.229	0.283	0.262	0.275	0.270	0.162	0.141	0.113	0.113
P01	0.029	0.103	0.185	0.177	0.233	0.171	0.177	0.183	0.080	0.066	0.035	0.043
P11	0.083	0.100	0.227	0.404	0.409	0.520	0.533	0.505	0.587	0.600	0.727	0.659

Table 5.4 Basic Statistics of the Amount for Hourly Precipitation

		Case1:Simulation			Case2:Extension			Case3:Transfer		
Stat	Method	P1	P01	P11	P1	P01	P11	P1	P01	P11
MAE (0)	CMSA	0.048	0.029	0.755	0.09	0.072	2.146	0.072	0.053	1.823
	PGAA	0.041	0.03	0.754	0.091	0.072	2.141	0.082	0.059	1.746
	SSMW	0.173	0.114	1.875	0.177	0.149	2.264	0.195	0.129	1.823
RMSE (0)	CMSA	0.218	0.171	0.866	0.3	0.267	1.461	0.268	0.229	1.347
	PGAA	0.202	0.172	0.867	0.301	0.267	1.459	0.286	0.243	1.32
	SSMW	0.415	0.338	1.369	0.42	0.386	1.503	0.442	0.359	1.349
MCE (1)	CMSA	0.752	0.761	0.407	0.584	0.466	-0.063	0.566	0.459	-0.542
	PGAA	0.789	0.76	0.408	0.579	0.464	-0.06	0.506	0.392	-0.48
	SSMW	0.111	0.074	-0.474	0.186	-0.113	-0.122	-0.17	-0.323	-0.542
d (1)	CMSA	0.881	0.889	0.752	0.779	0.73	0.471	0.799	0.763	0.426
	PGAA	0.898	0.888	0.753	0.777	0.729	0.471	0.779	0.738	0.437
	SSMW	0.573	0.515	0.343	0.59	0.411	0.352	0.519	0.434	0.35
		Mean	Std	Skew	Mean	Std	Skew	Mean	Std	Skew
MAE (0)	CMSA	4.69	10.019	9.944	7.206	12.855	7.111	7.307	12.587	7.636
	PGAA	3.616	6.964	8.244	6.357	10.047	6.825	6.459	11.178	7.409
	SSMW	4.837	10.265	11.832	6.485	10.811	9.84	6.886	12.277	9.005
RMSE (0)	CMSA	2.158	3.149	3.145	2.674	3.566	2.658	2.694	3.528	2.757
	PGAA	1.896	2.63	2.865	2.511	3.154	2.604	2.532	3.332	2.714
	SSMW	2.199	3.202	3.432	2.544	3.284	3.131	2.623	3.501	2.994
MCE (1)	CMSA	0.03	0.023	-0.093	0.02	-0.004	0.313	0.025	0.09	0.076
	PGAA	0.252	0.321	0.094	0.135	0.215	0.341	0.139	0.192	0.103
	SSMW	0	-0.001	-0.301	0.118	0.156	0.05	0.082	0.113	-0.09
d (1)	CMSA	0.589	0.589	0.482	0.559	0.565	0.616	0.565	0.574	0.545
	PGAA	0.613	0.668	0.543	0.566	0.613	0.631	0.531	0.565	0.558
	SSMW	0.455	0.447	0.33	0.395	0.458	0.449	0.409	0.458	0.447

Note: The number following the statistics is the value when a model has the perfect fit. (e.g. MAE(0)). And the best fit model highlighted with gray color.

Table 5.5 Mean Values of the Performance criteria in comparison between the model value and the historical value at DIA

	Case1:Simulation			Case2:Extension			Case3:Transfer		
	CMSA	PGAA	SSMW	CMSA	PGAA	SSMW	CMSA	PGAA	SSMW
MAE	0.111	0.109	0.104	0.114	0.111	0.106	0.129	0.127	0.123
RMSE	0.334	0.329	0.323	0.338	0.333	0.325	0.359	0.356	0.351
MCE	0.076	0.099	0.134	0.061	0.092	0.133	0.084	0.160	0.123
d	0.539	0.549	0.562	0.535	0.549	0.563	0.522	0.526	0.533

Note: The number following the statistics is the value when a model has the perfect fit. (e.g. MAE(0)). And the best fit model highlighted with gray color.

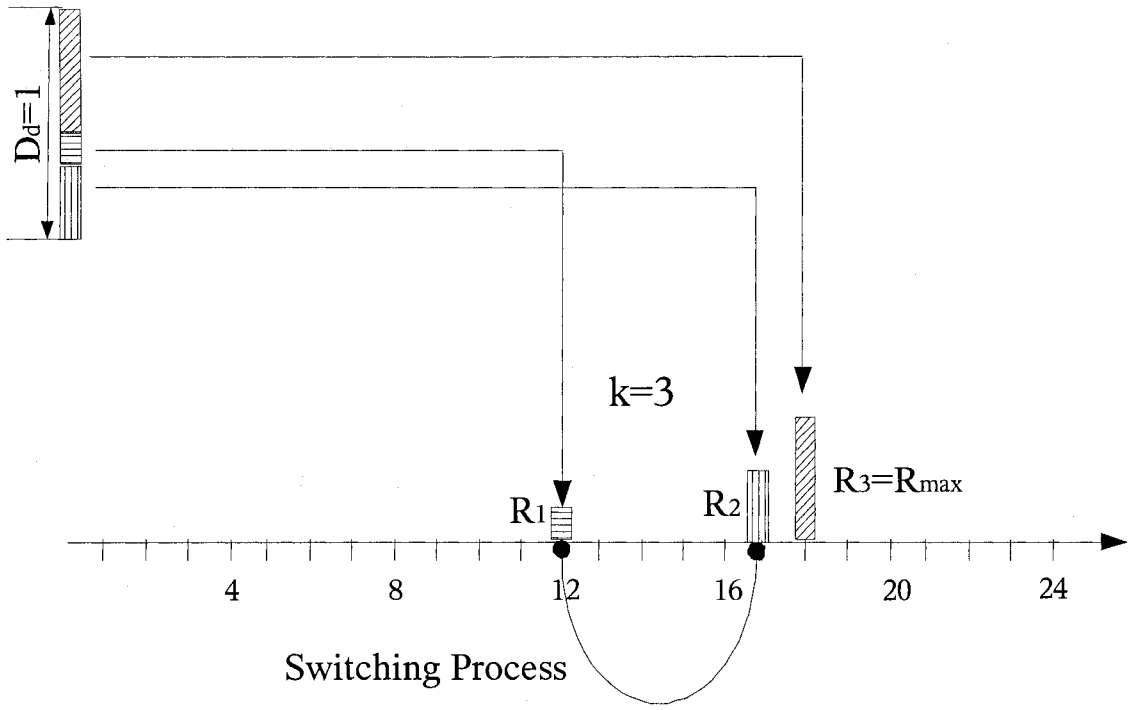


Figure 5.1 Conditional Markov Chain and Simulated Annealing : The upper-left side bar represents the unit daily amount (D_d=1) . The event hours are specified from the conditional Markov Chain (k=3, h=12, 17, 18). And the daily amount should be separated employing the modified Atchison distribution. The separated hourly amounts will be found proper place using the simulated annealing process (switching process).

Order	Duration	Depth
	i	y_i
1	i_1	y_1
2	i_2	y_2
3	i_3	y_3
…	…	…
n	i_n	y_n

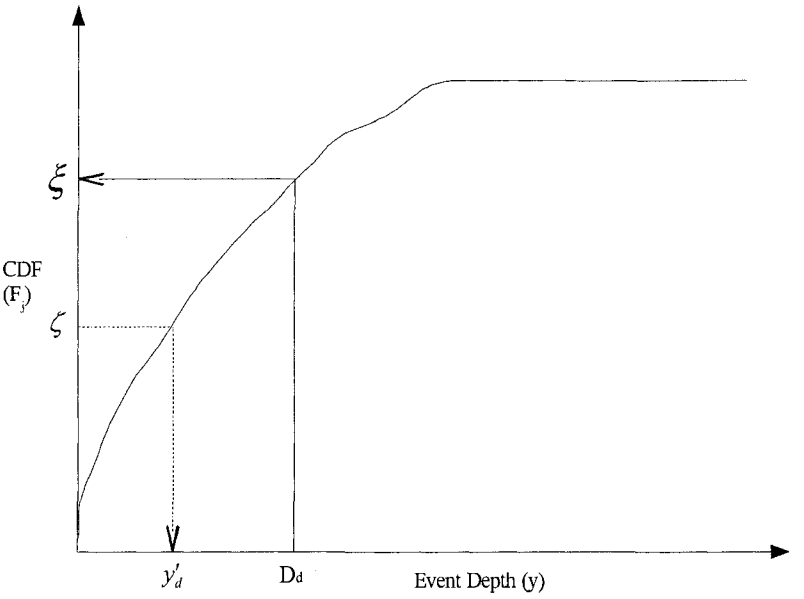


Figure 5.2 Events Table (Left) and CDF (F_j) of Event Depth (Right)

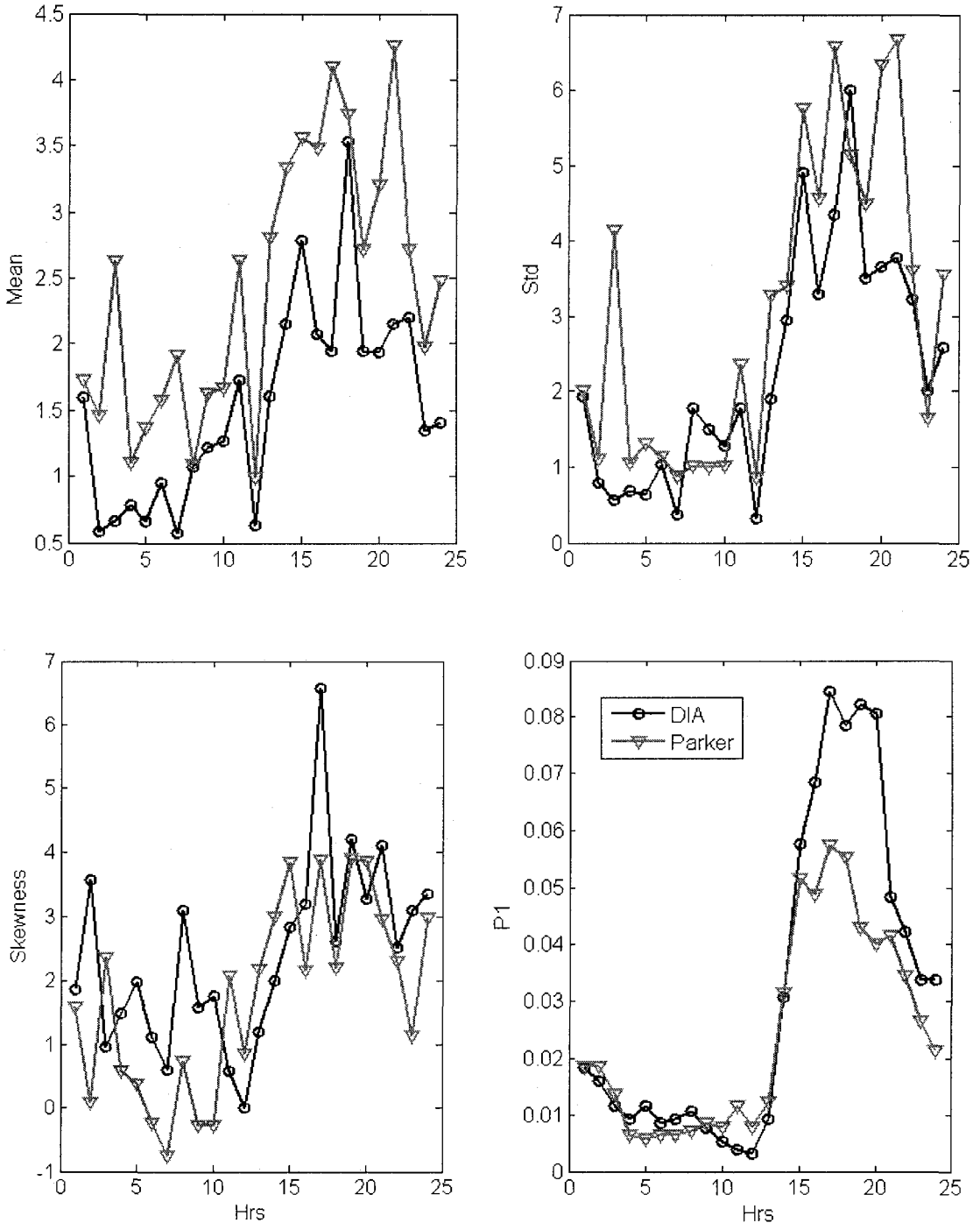


Figure 5.3 Comparison of the hourly statistics for the two sites (Denver International Airport: Segment Line with circle and Parker: Segment Line with triangle)

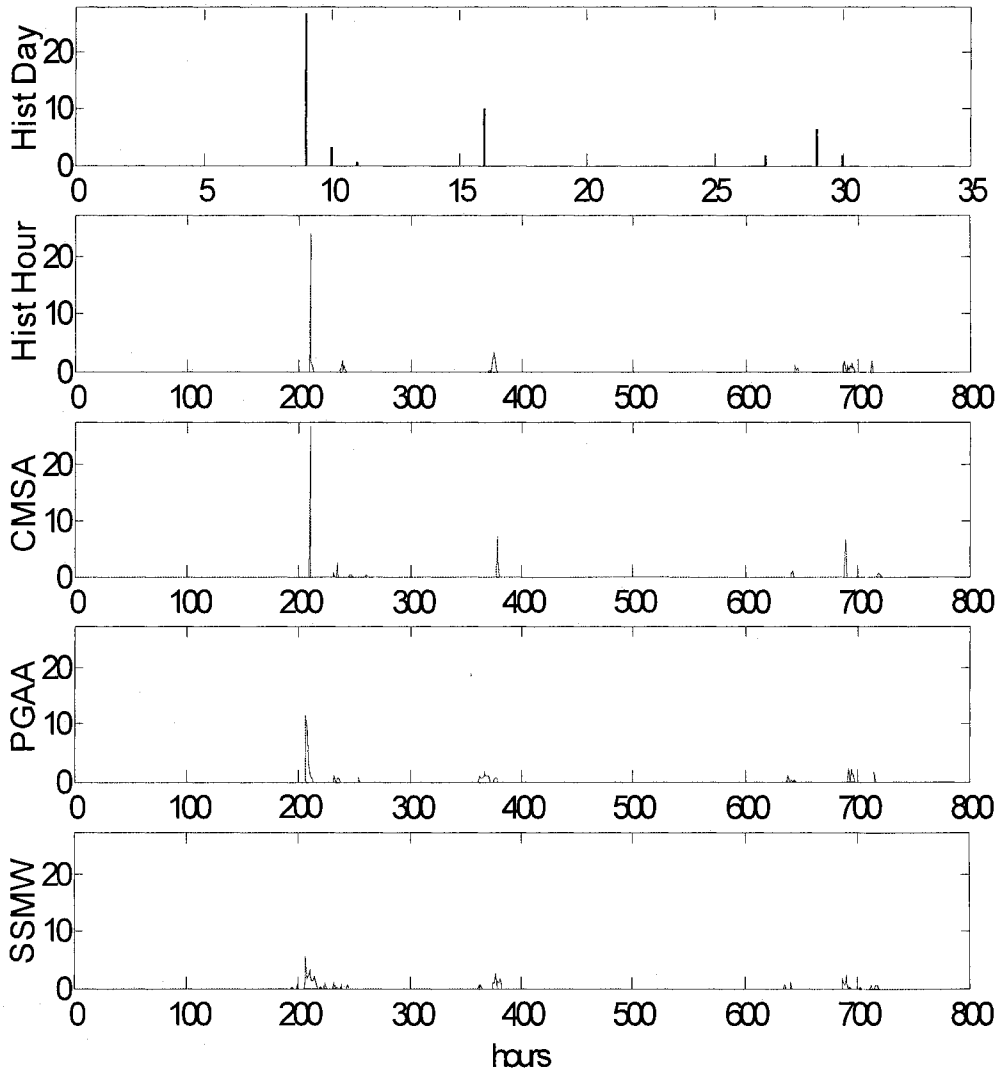


Figure 5.4 Realization of the disaggregation and historical hourly and daily for the simulation case

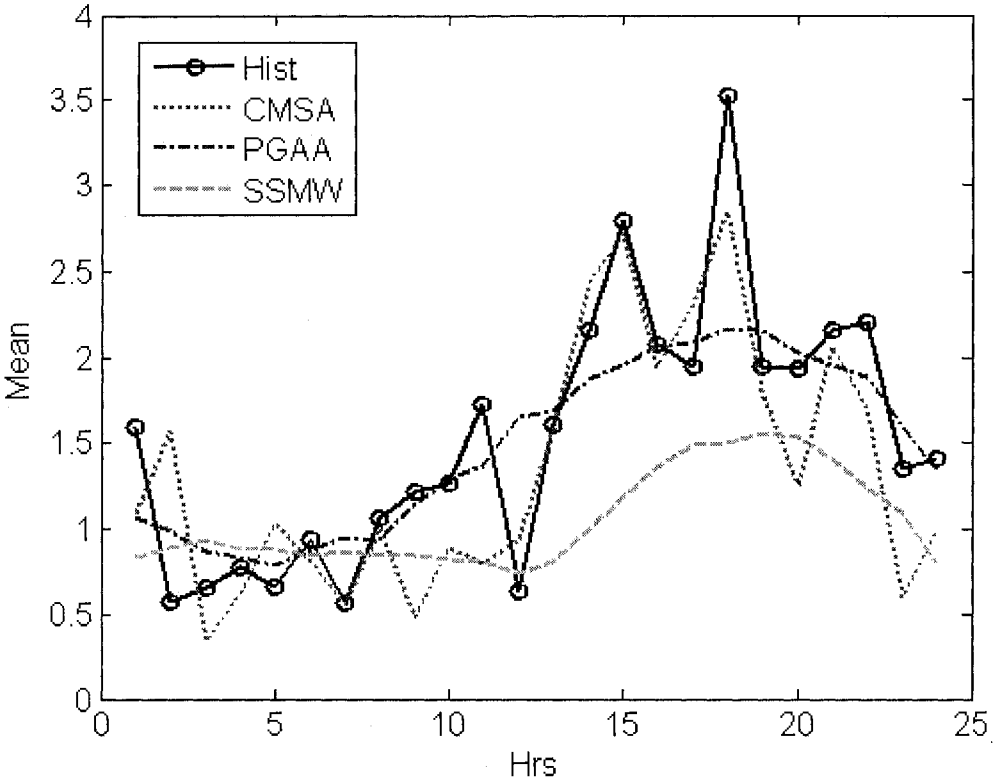


Figure 5.5 Historical and Disaggregated hourly Mean for Simulation Case

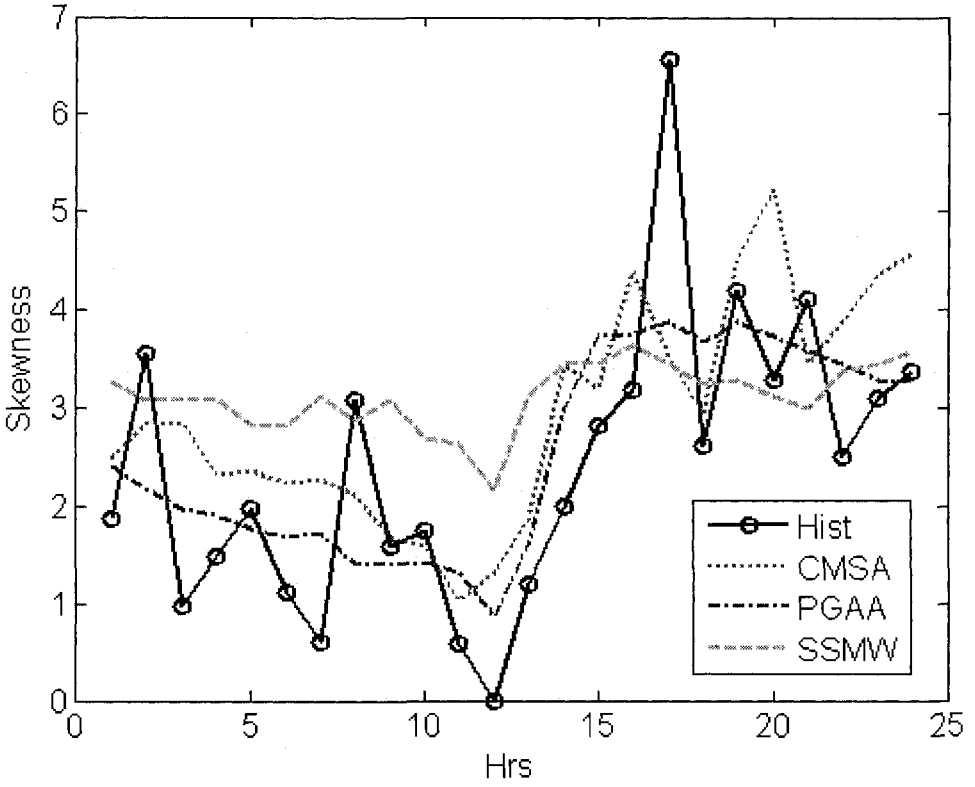


Figure 5.6 Historical and Disaggregated hourly Skewness for Simulation Case

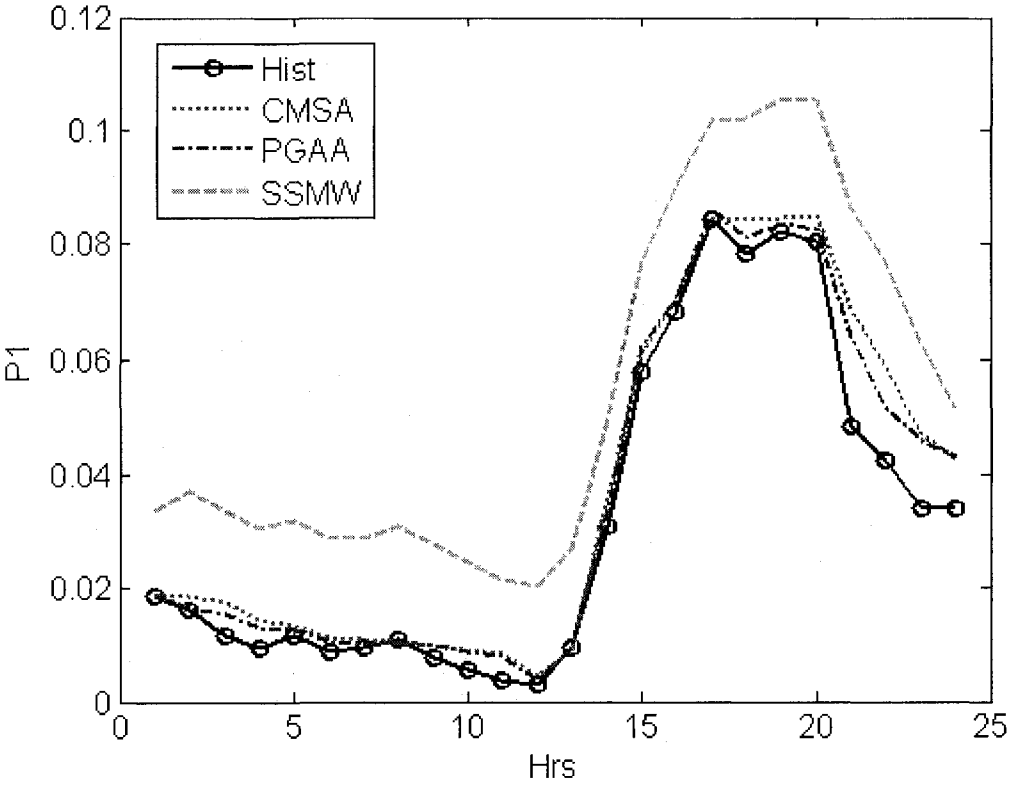


Figure 5.7 Historical and Disaggregated Occurrence probability P1 for Simulation Case

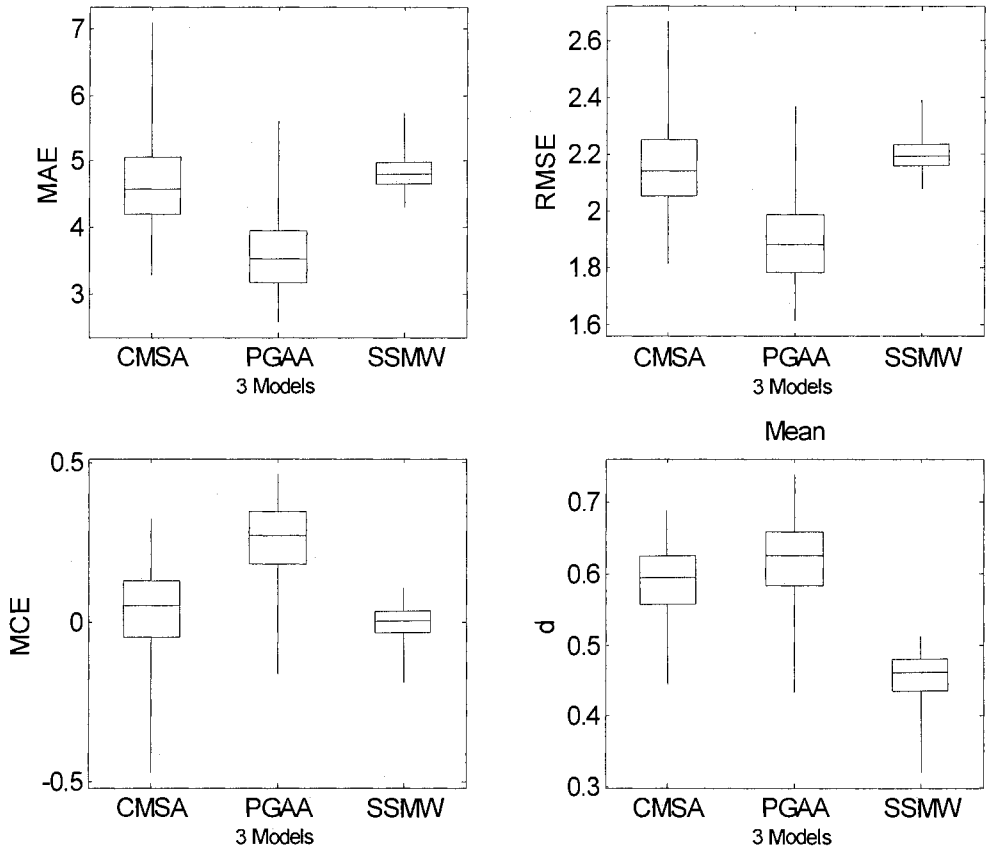


Figure 5.8 Performance criteria of the Mean for Simulation case

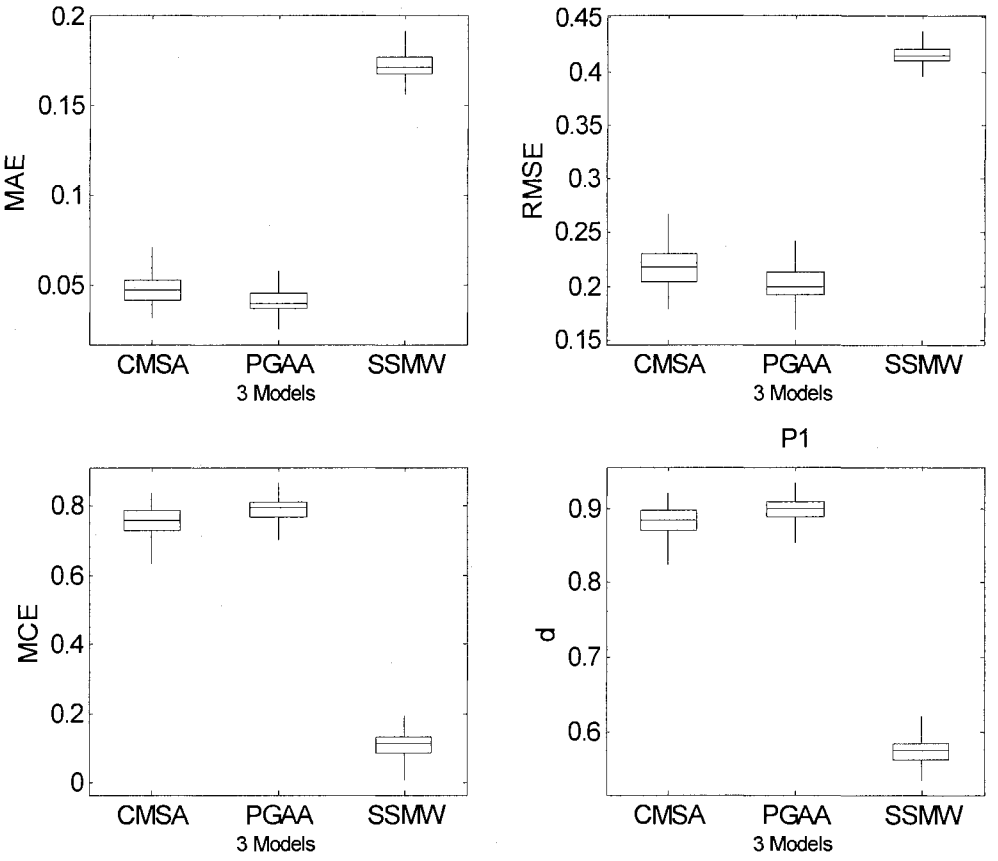


Figure 5.9 Performance criteria of the probability P1 for Simulation case

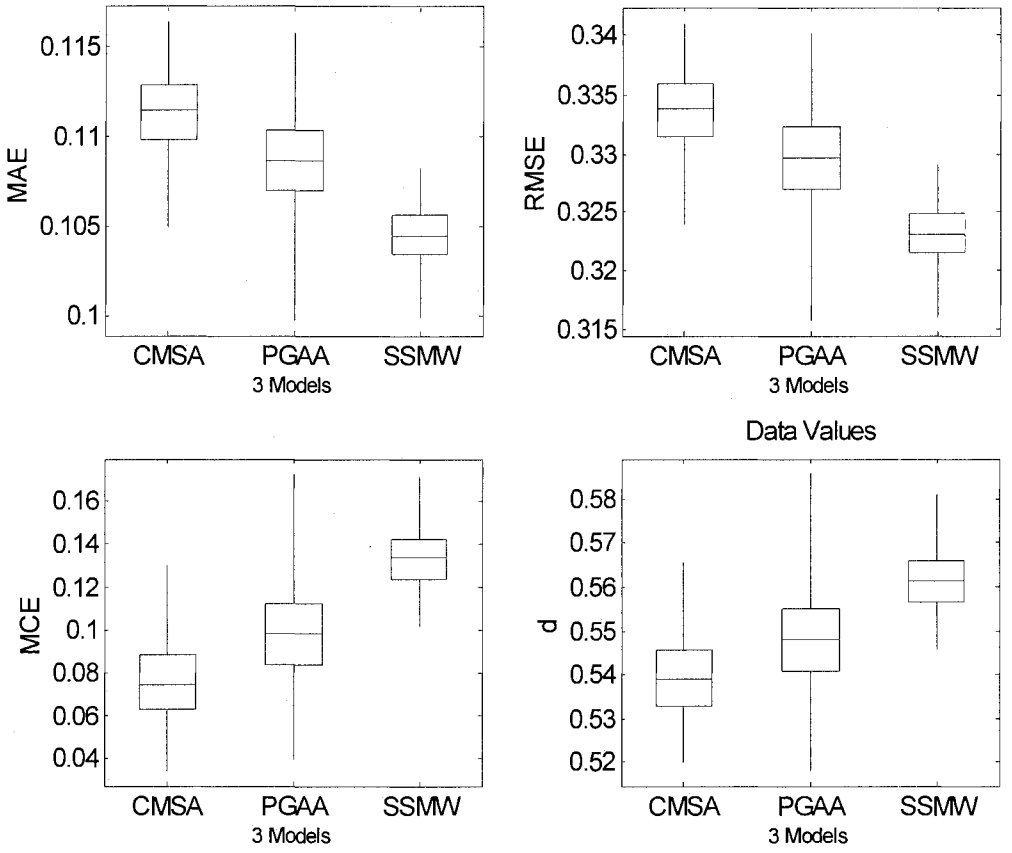


Figure 5.10 Performance criteria of the Direct Data Values for Simulation case

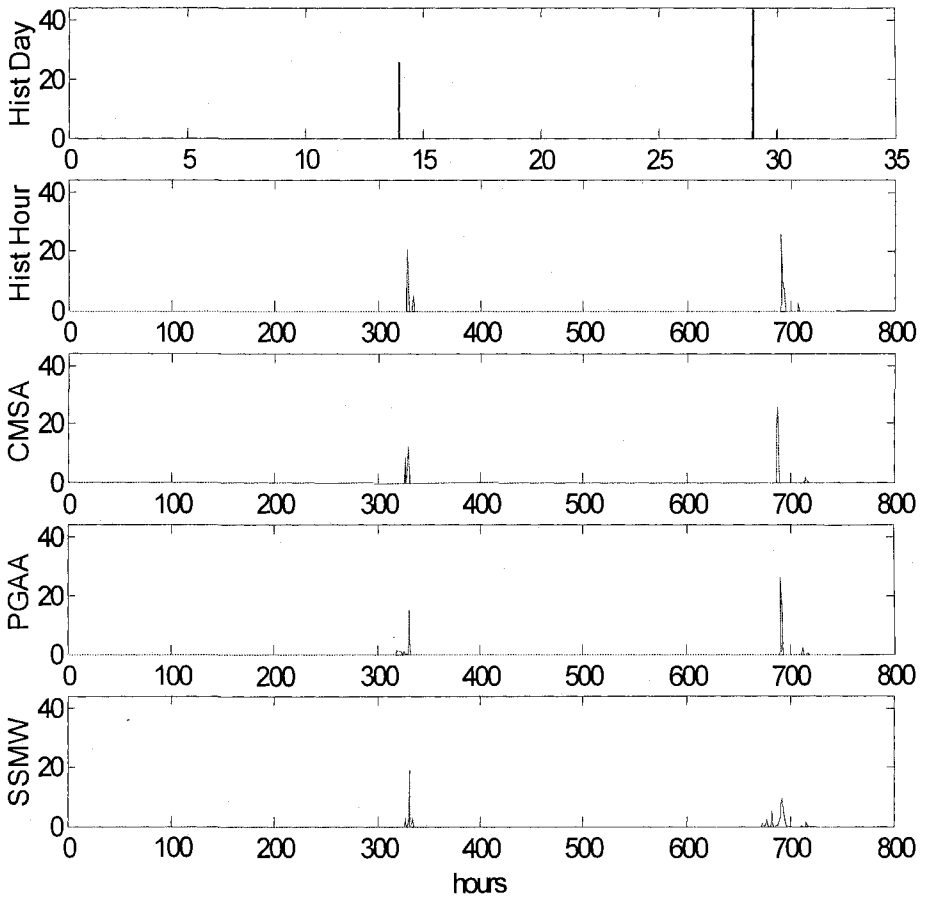


Figure 5.11 Realization of the disaggregation for Extension case

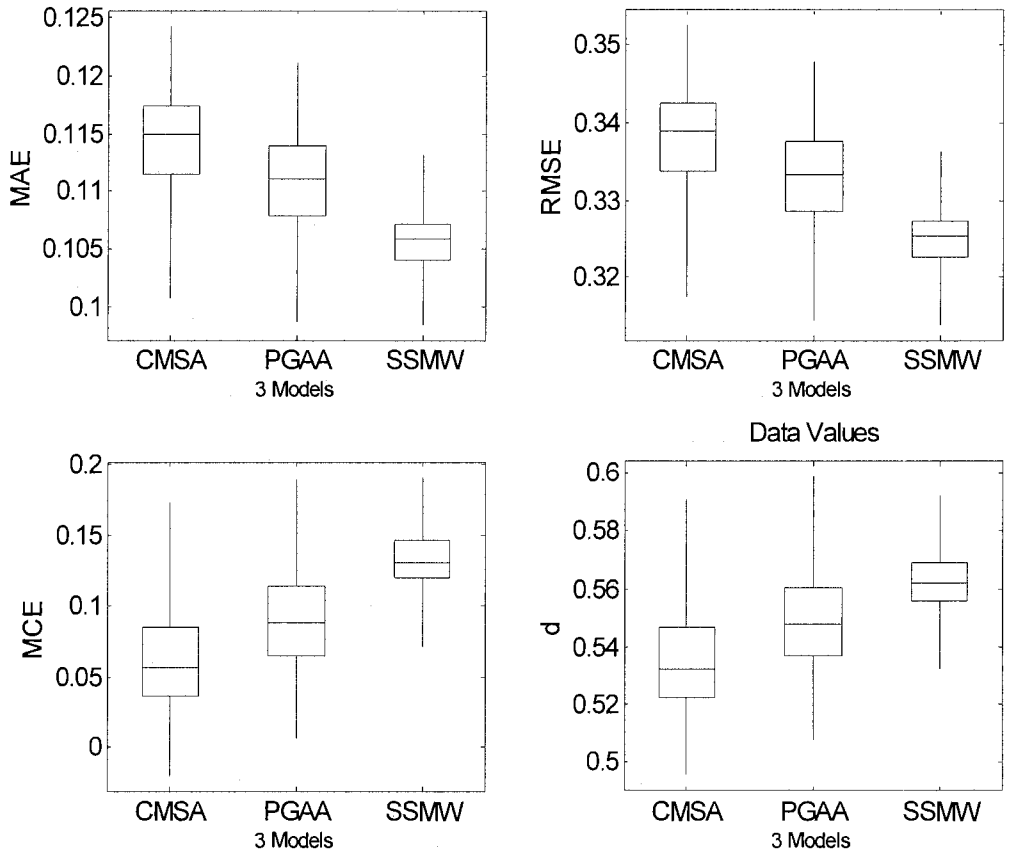


Figure 5.12 Performance criteria of the Direct Data Values for Extension case

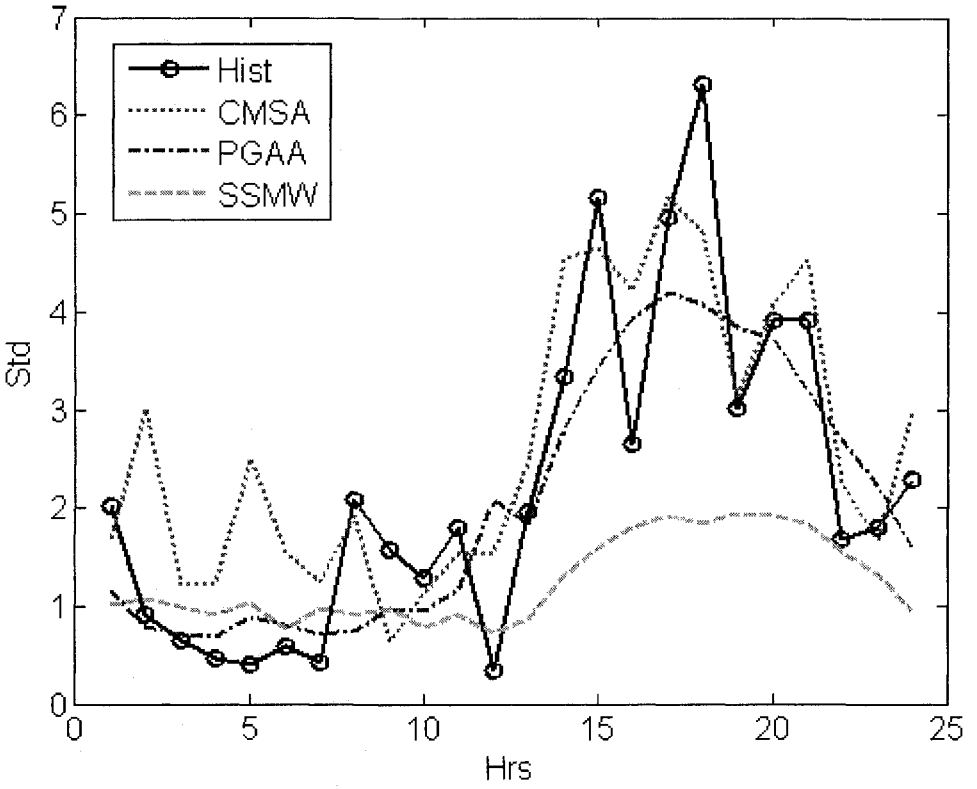


Figure 5.13 Historical and Disaggregated hourly Standard Deviation for Extension Case

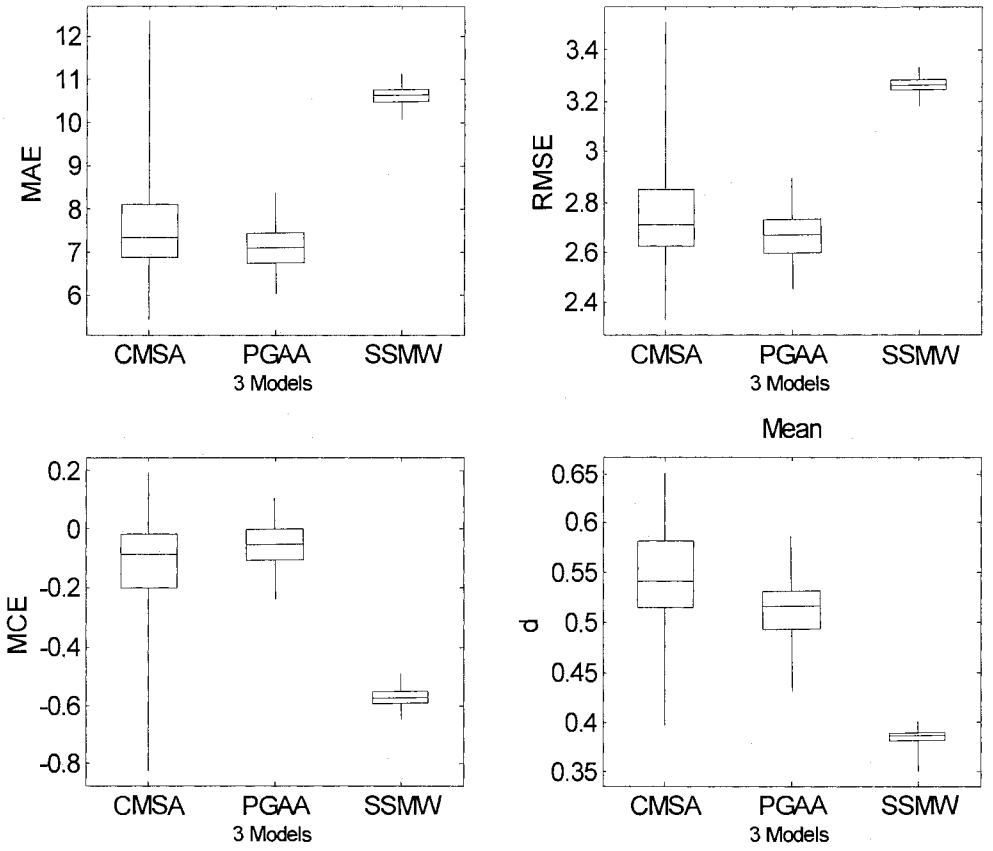


Figure 5.14 Performance criteria of the Mean for Transfer case

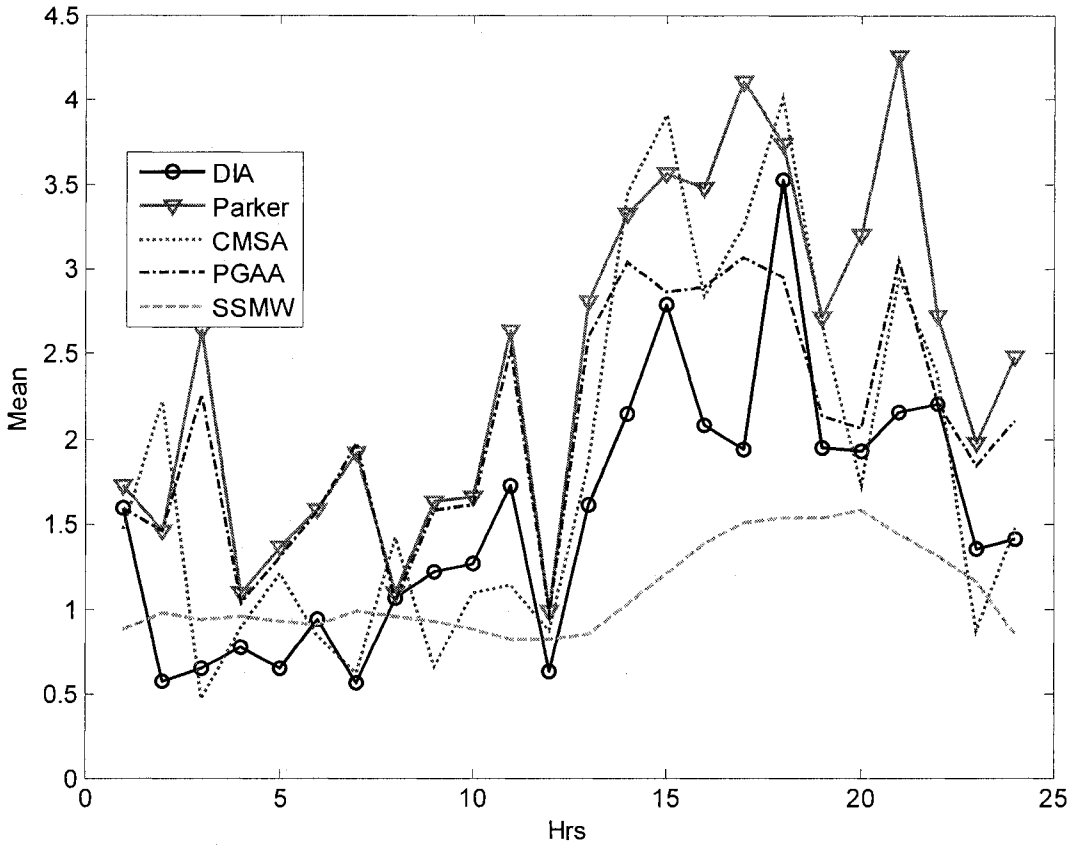


Figure 5.15 Hourly mean of the historical and disaggregated data for Transfer Case

5.7 References

Aitchison J (1982), "The Statistical Analysis of Compositional Data", J.R.Statist. Soc. B, Vol.44(2), pp.139-177

Arnold J. G. and Williams J.R. (1989) "Stochastic generation of internal storm structure", Trans. ASAE 32(1), pp.161-166

Bardossy A. (1998) "Generating precipitation time series", Water Resources Research Vol.34(7) 1737-1744

Bardossy A. (1999) "Disaggregation of Daily Precipitation", River Basin Modeling, Management and Flood Mitigation concerted Action Proceeding.

Betson, R.P., Bales, J. and Pratt,H.E., (1980), User's Guide to TVA-HYSIM. U.S. EPA, EPA-600/7-80, pp40-44

Bo Z., Islam S., and Eltahir EAB (1994), "Aggregation-disaggregation properties of a stochastic rainfall model", Water Resources Research, Vol.30(12) pp.3423-3435

Burian SJ , Durrans R. and Tomic S. (2000), "Rainfall Disaggregation using artificial neural networks", J of Hydrologic Engineering, ASCE

Burian SJ , et al. (2001), "Training Artificial neural networks to perform rainfall disaggregation", J of Hydrologic Engineering, ASCE

Burian SJ , Durrans R. (2002), "Evaluation of Artificial neural networks rainfall disaggregation model", Water Science and Technology, Vol.45(2),pp.99-104

Chebaane, M., Salas, J. D., and Boes, D. C. (1995). "Product Periodic Autoregressive Processes for Modeling Intermittent Monthly Streamflows." *Water Resources Research*, 31(6), 1513-1518.

Collander RS, Tollerud EI, Li L, and Vironf-Lazar A, 1993, Hourly Precipitation data and station histories: A research assessment, in Preprints, Eighth Symposium on Meteorological Observations and Instrumentation, pp. 153-158, American Meteorological Society, Boston.

Connolly R.D. et al (1998) "A daily rainfall disaggregation model", Agricultural and forest meteorology, Vol.92 pp.105-117

Cowpertwait et al. (1996) "Stochastic point process modeling of rainfall II. Rationalization and disaggregation", Journal of Hydrology, Vol.175, 47-65

Econonpouly TW, Davis DR, and Woolhiser DA (1990), "Parameter transferability for a daily rainfall disaggregation model" Journal of hydrology, Vol.118, 209-228

Fernandez, B., and Salas, J. D. (1990). "Gamma-Autoregressive Models for Stream-Flow Simulation." *Journal of Hydraulic Engineering-Asce*, 116(11), 1403-1414.

Glasbey C.A., Cooper G., McGechan M.B., (1995), "Disaggregation of daily rainfall by conditional simulation from a point-process model", Journal of Hydrology, Vol.165, pp1-9

Guenni L and Bardossy A (2002) "A two steps disaggregation method for highly seasonal monthly rainfall", Stochastic Environmental Research and Risk Assessment, Springer-Verlag Vol.16 188-206

Guntner et al. (2001), "Cascade-based disaggregation of continuous rainfall time series: the influence of climate", Hydrology and Earth System Sciences, Vol.5(2),pp 145-164

Hershenhorn J. and Woolhiser D.A. (1987) "Disaggregation of Daily Rainfall", Journal of Hydrology, 95, pp 299-322

Ingber L (1993), "Simulated annealing : practice versus theory", Math. Comput. Modelling Vol.18(11) pp 29-57

Katz RW and Parlange MB (1995), "Generalizations of chain-dependent processes: Application to hourly precipitation", Water Resources Research Vol.31(5) 1331-1341

Kottegoda NT and Rosso R (1997), "Statistics, Probability, and Reliability for Civil and Environmental Engineers", McGraw-Hill

Koutsoyiannis D., and Onof C.(2001), "Rainfall disaggregation using adjusting procedures on a Poisson cluster model", Journal of Hydrology, pp 109-122

Koutsoyiannis D.(1994), "A stochastic disaggregation method for design storm and flood synthesis", Journal of Hydrology, Vol.156, pp 193-225

Koutsoyiannis D., and Manetas A.(1996), "Simple disaggregation by accurate adjusting procedures", Water Resources Research, Vol.32(7), pp 2105-2117

Lane L.J. and Nearing M.A. (1989) USDA- Water Erosion Prediction Project: Hillslope profile model documentation. NSERL Report No.2, USDA-ARS

Legates DR and McCabe GJ (1999), Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation, WRR 33(1) 167-75

Olsson J. et al (2004) "Neural Networks for Rainfall Forecasting by Atmospheric Downscaling", Journal of Hydrologic Engineering, ASCE

Olsson J and Berndtsson R(1998), "Temporal rainfall disaggregation based on scaling properties", Wat. Sci. Tech. Vol.37(11), pp73-79

Olsson J (1998), "Evaluation of a scaling cascade model for temporal rainfall disaggregation", Hydrology and Earth System Science, 2(1) pp19-30

Press SJ (2003), "Subjective and Objective Bayesian Statistics- Principles, Models, and Applications", 2/e, Wiley-Interscience

Robert CP, Casella G (1999), "Monte Carlo Statistical Methods", Springer pp 233-238

Rodriguez-Iturbe I., Cox DR, Isham V., (1988), " A point model for rainfall : further developments", Proc. R. Soc. London, Ser. A, Vol.417 pp283-298

Rosewell C.J. (1986) "Rainfall kinetic energy in eastern Australia", J. Clim. Appl. Meterol. Vol.25 pp.1695-1701

Salas JD et al. (1980), "Applied Modeling of Hydrologic Time Series", Water Resources Publications, LLC

Salas, J. D. (1993). "Analysis and Modeling of Hydrologic Time Series" In: *Handbook of Hydrology*, M. D.R., ed., McGraw-Hill.

Socolofsky S, Adams EE, Entekhabi D. (2001) "Disaggregation of Daily Rainfall for Continuous Watershed Modeling", Journal of Hydrologic Engineering, ASCE

Sivakumar B., Sorooshian S., Gupta HV, and Gao X. (2001), "A chaotic approach to rainfall disaggregation", Water Resources Research, Vol.37(1), pp.61-72

Willmott CJ (1984), " On The Validation of Models", Physical Geography, Vol.2(2), 184-194

Data Acquisition

US Hourly Precipitation Data TD3240, NOAA, Produced by Forecast systems and Laboratory Boulder, CO and National Climatic Data Center Asheville, NC (July 1998) Website : www.ncdc.noaa.gov

Appendix 5-A. Detailed Figures

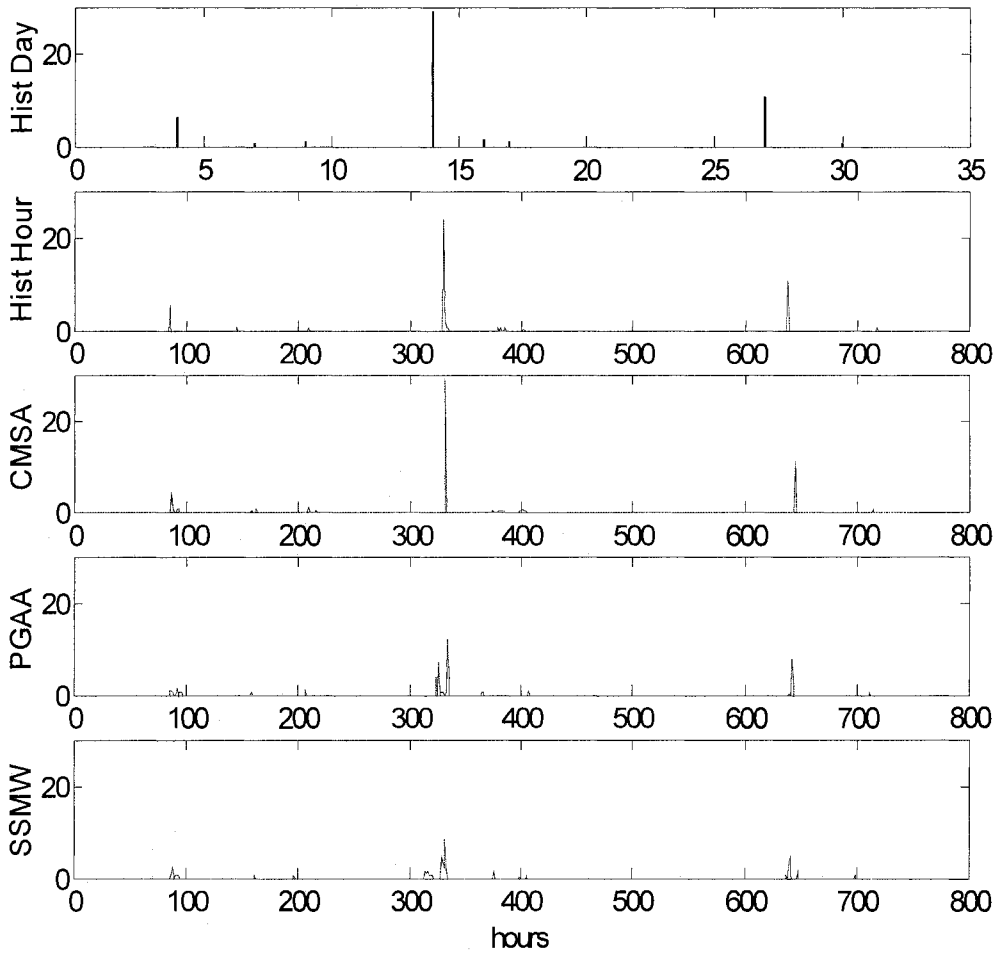


Figure 5-A.1 Realization of the disaggregation and historical hourly and daily for the simulation case

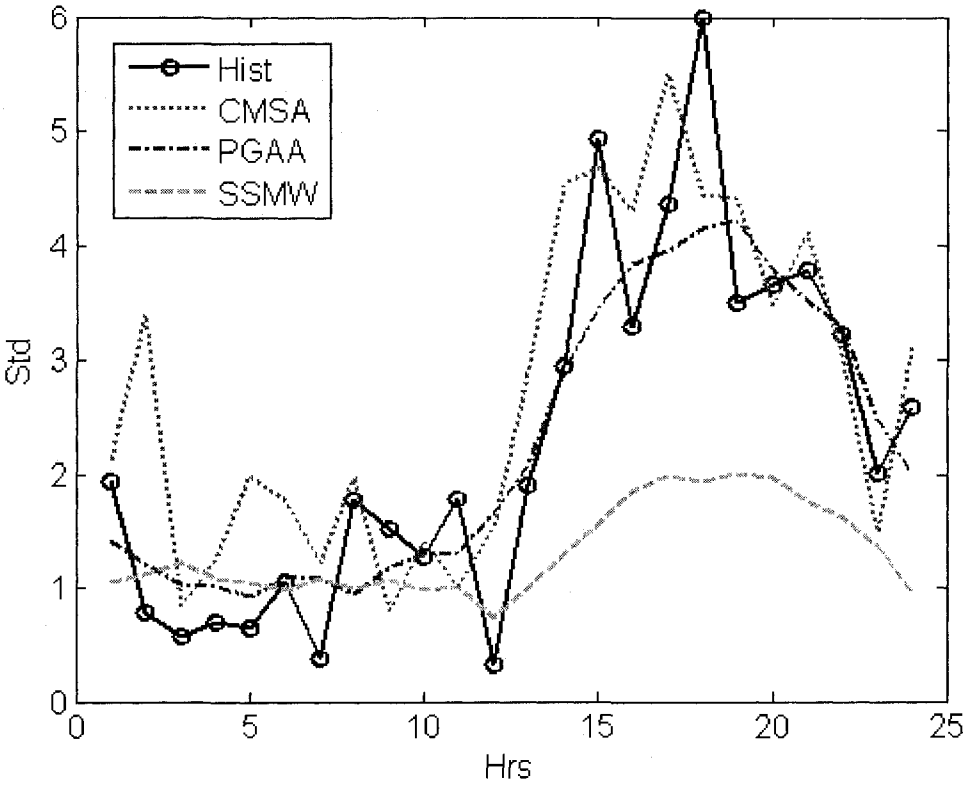


Figure 5-A.2 Historical and Disaggregated hourly Standard Deviation for Simulation Case

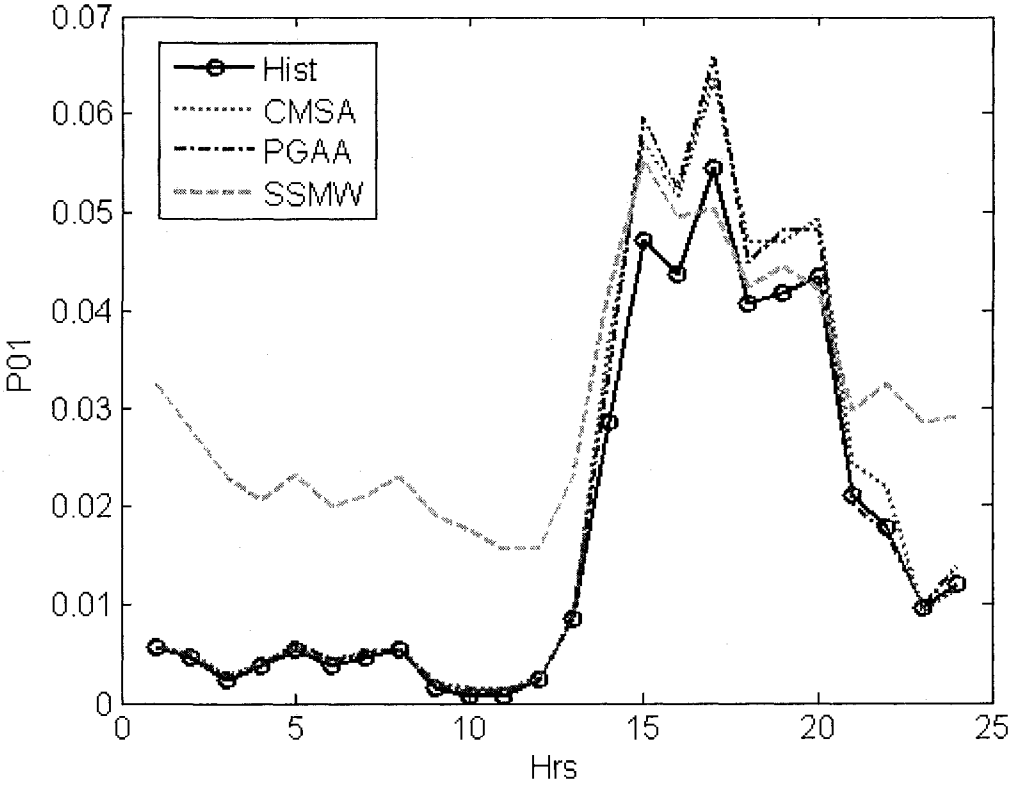


Figure 5-A.3 Historical and Disaggregated Occurrence probability P01 for Simulation Case

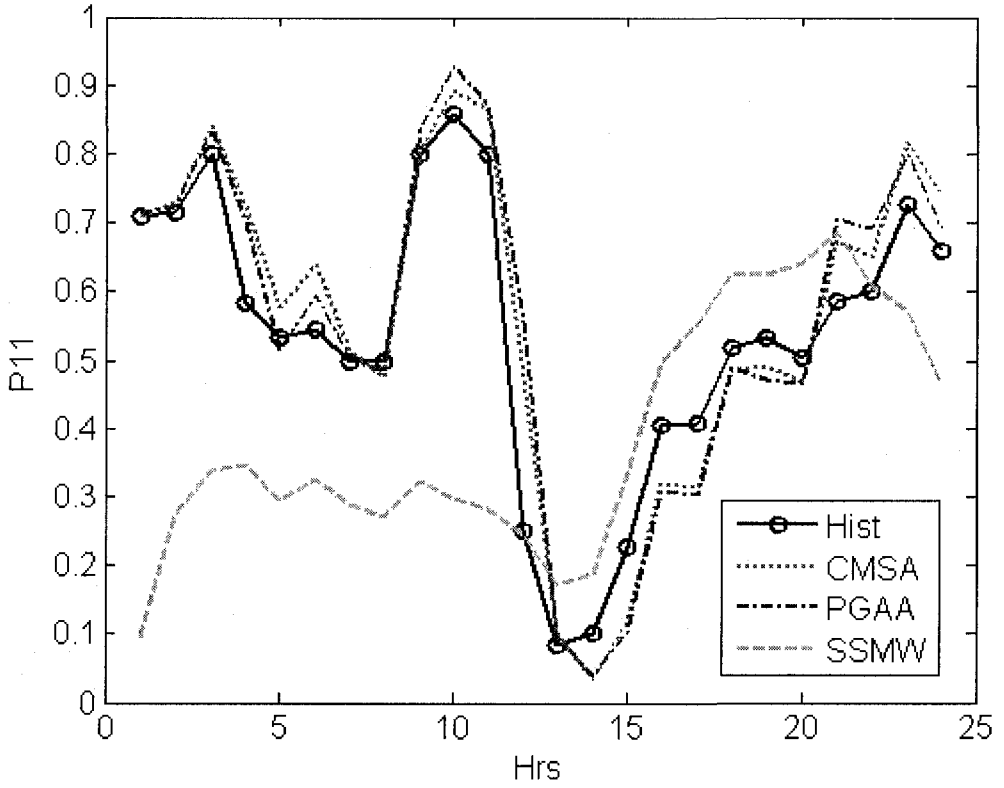


Figure 5-A.4 Historical and Disaggregated Occurrence probability P11 for Simulation Case

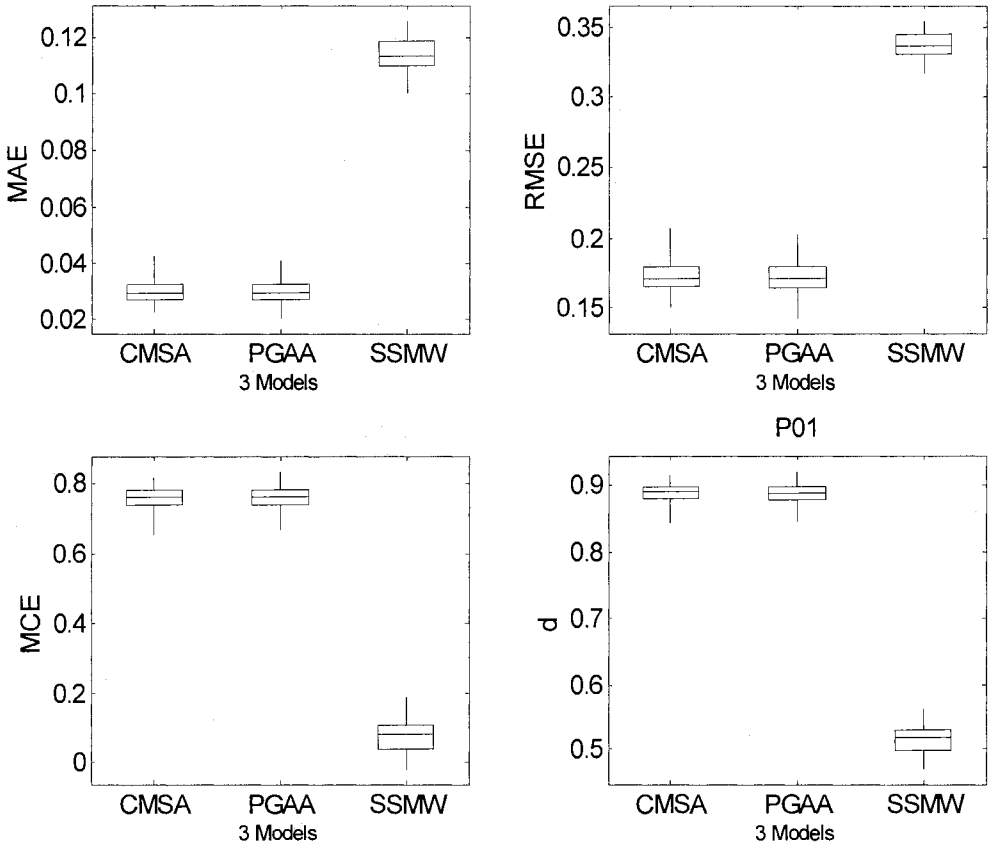


Figure 5-A.5 Performance criteria of the probability P01 for Simulation case

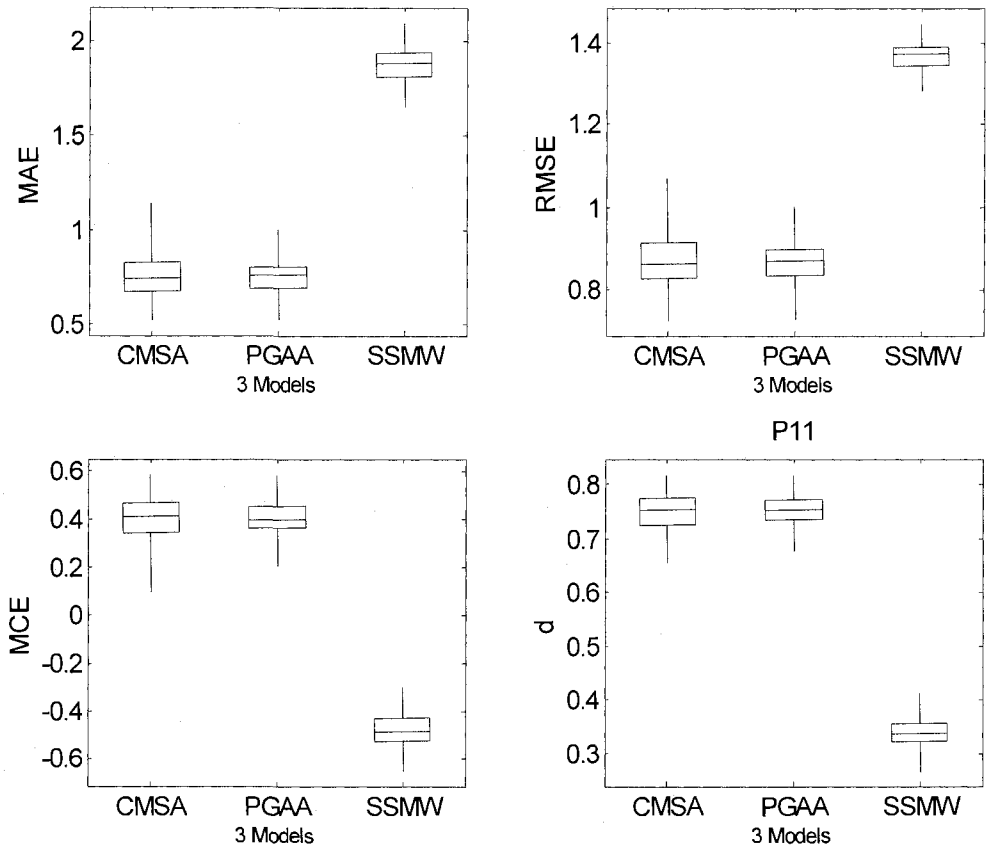


Figure 5-A.6 Performance criteria of the probability P11 for Simulation case

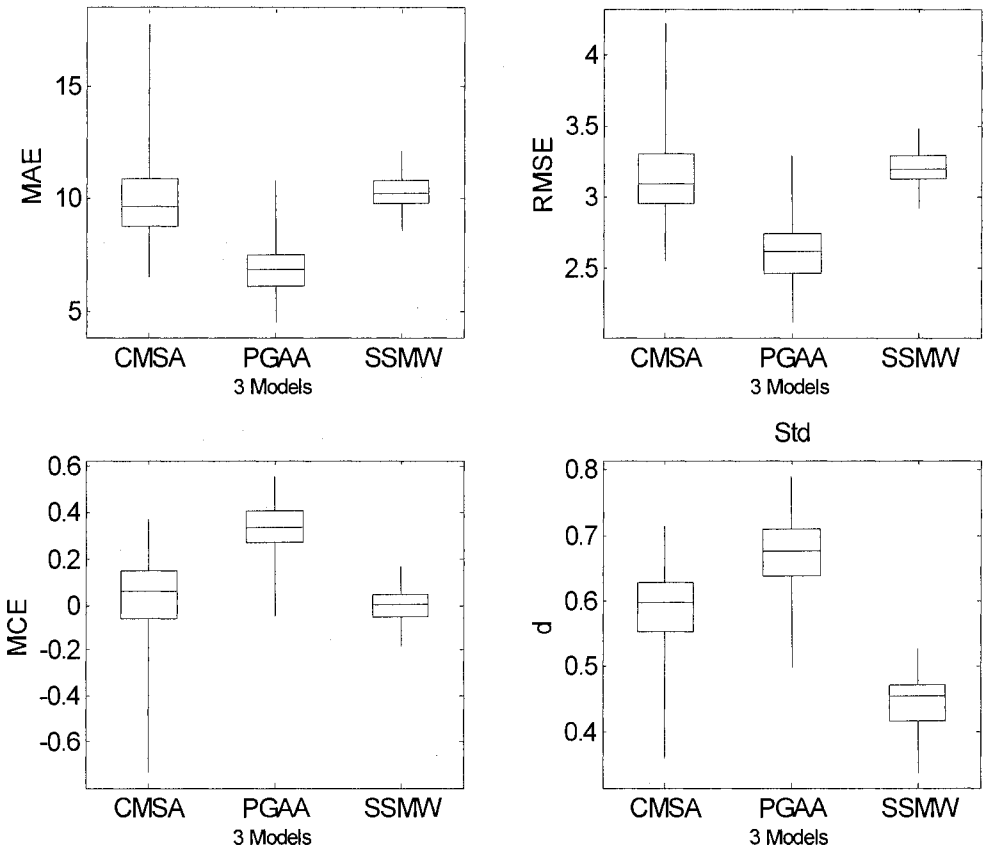


Figure 5-A.7 Performance criteria of the Standard Deviation for Simulation case

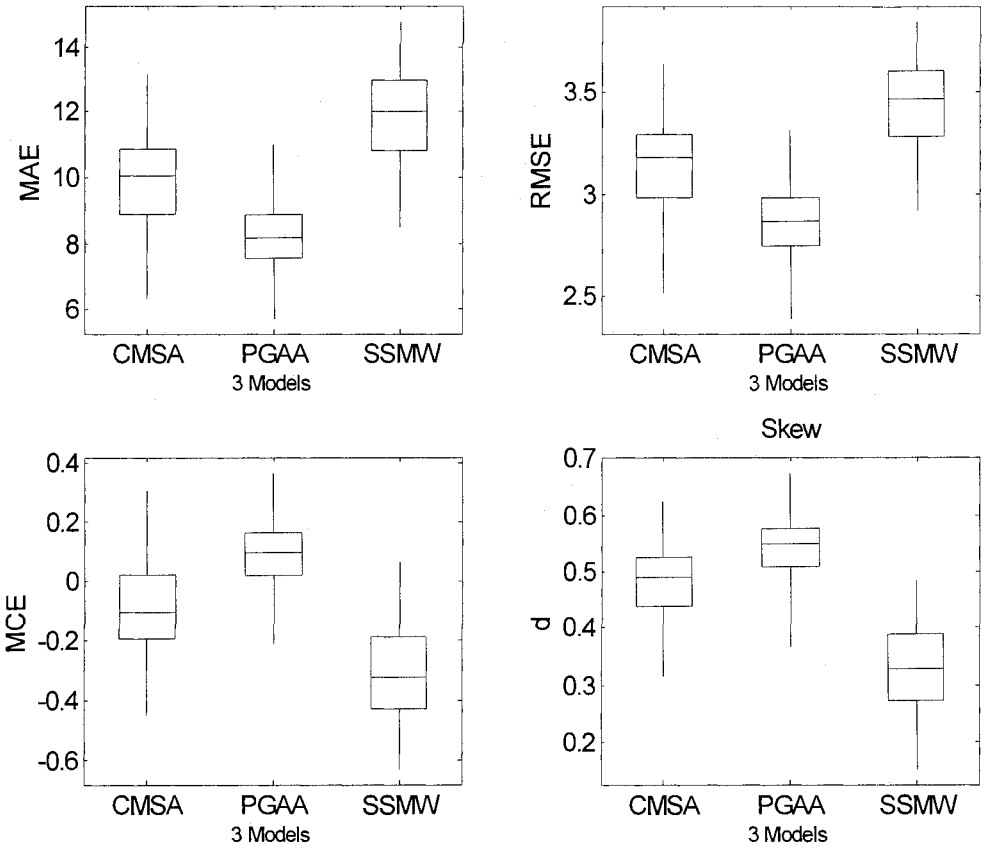


Figure 5-A.8 Performance criteria of the Skewness for Simulation case

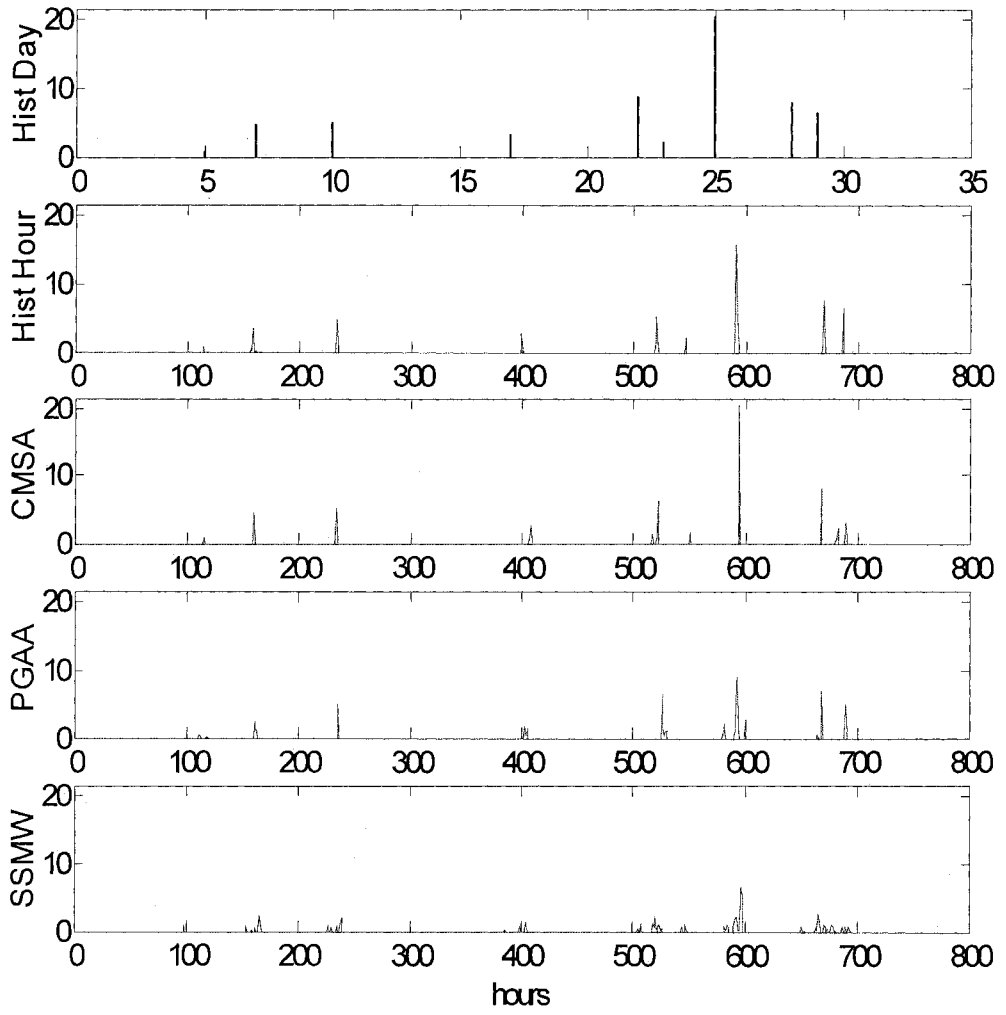


Figure 5-A.9 Realization of the disaggregation for Extension case

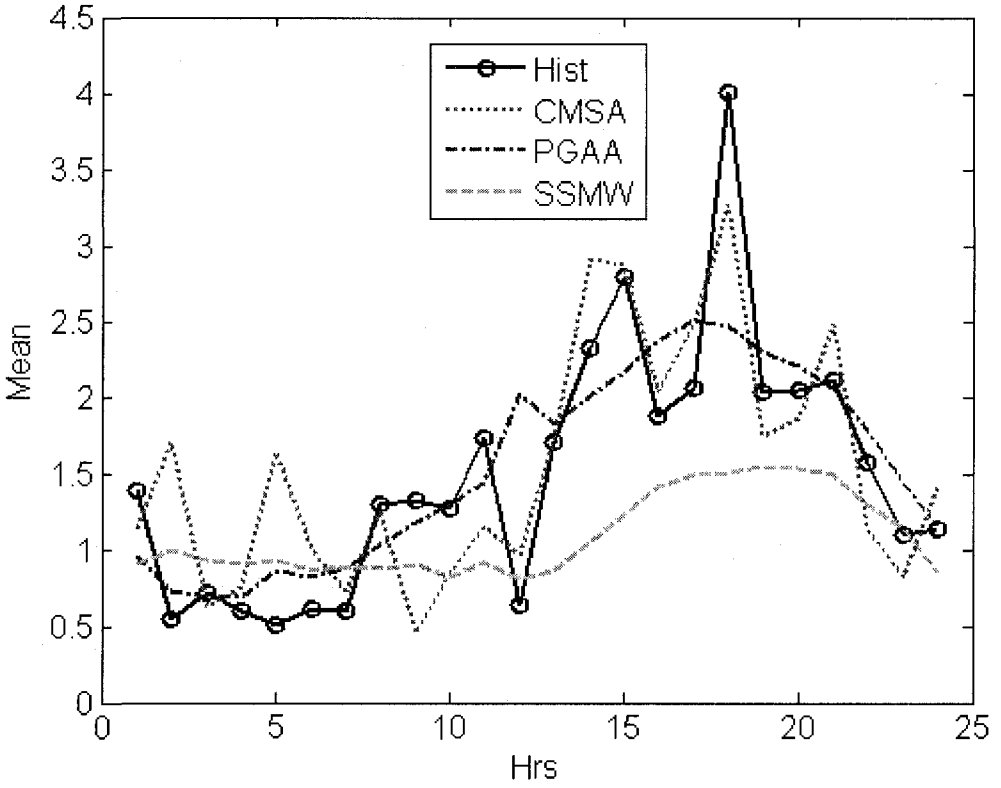


Figure 5-A.10 Historical and Disaggregated hourly Mean for Extension Case

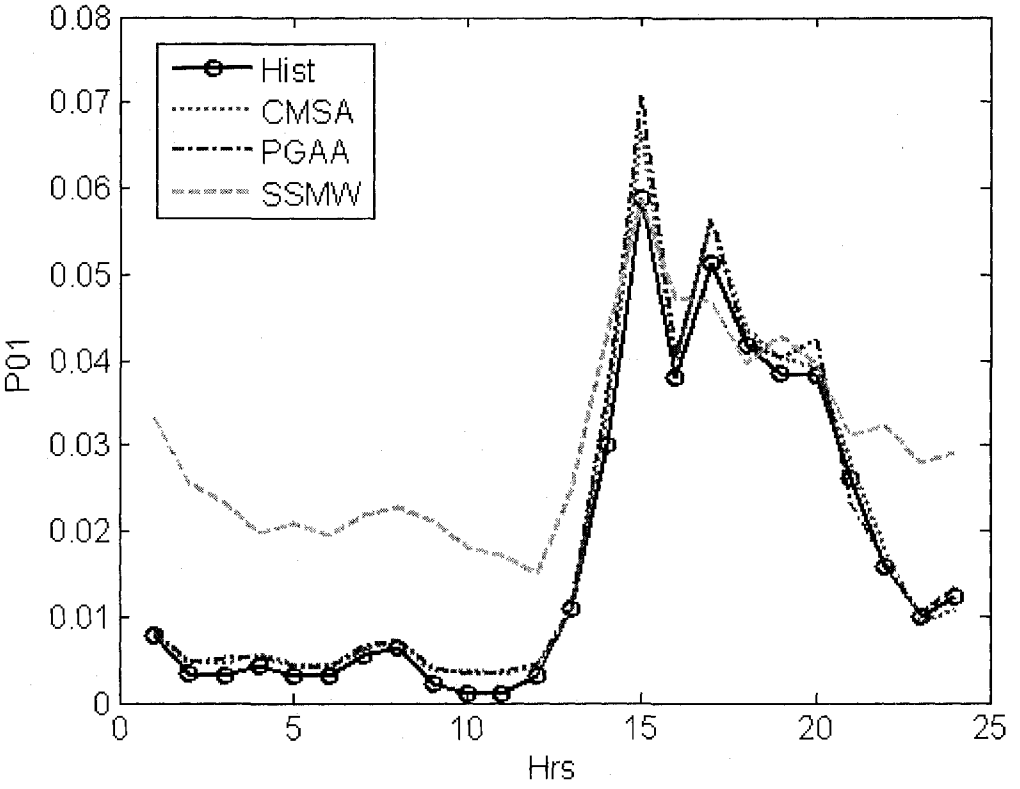


Figure 5-A.11 Historical and Disaggregated Occurrence probability P01 for Extension Case

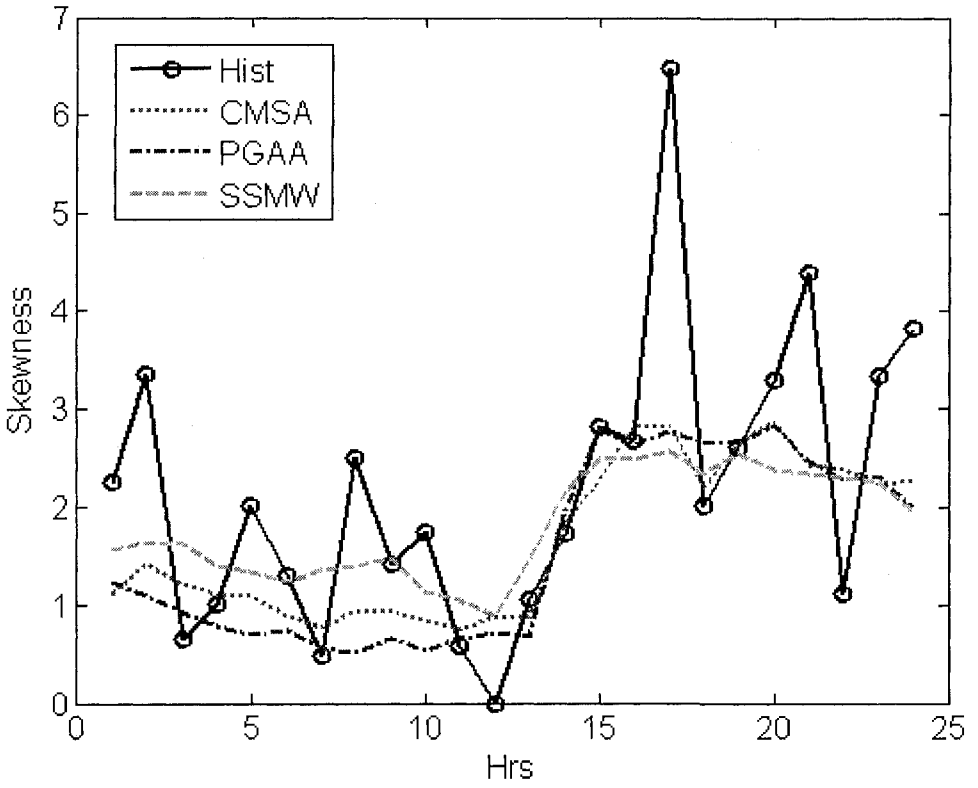


Figure 5-A.12 Historical and Disaggregated hourly Skewness for Extension Case

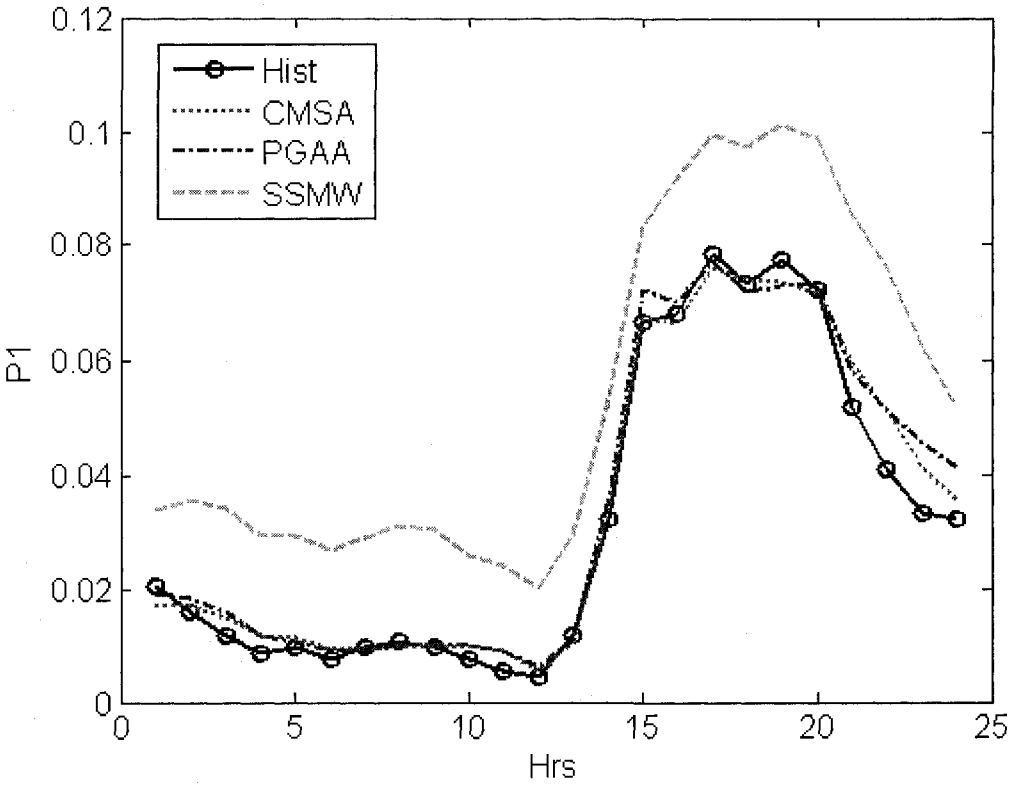


Figure 5-A.13 Historical and Disaggregated Occurrence probability P1 for Extension Case

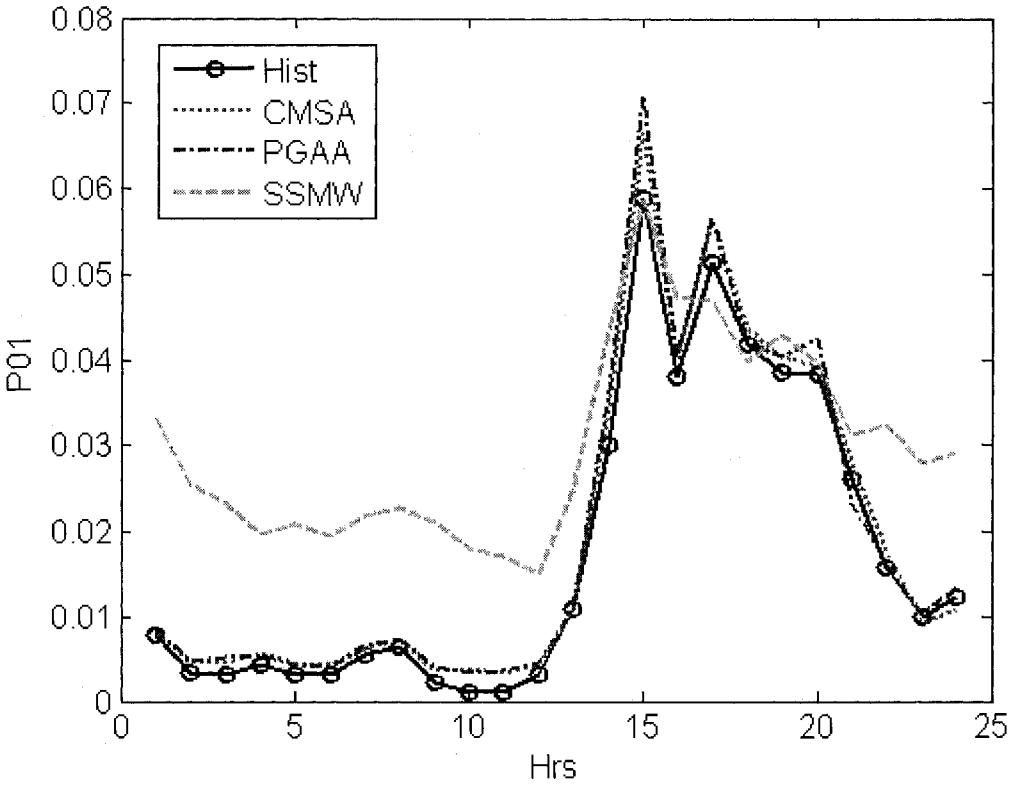


Figure 5-A.14 Historical and Disaggregated Occurrence probability P01 for Extension Case

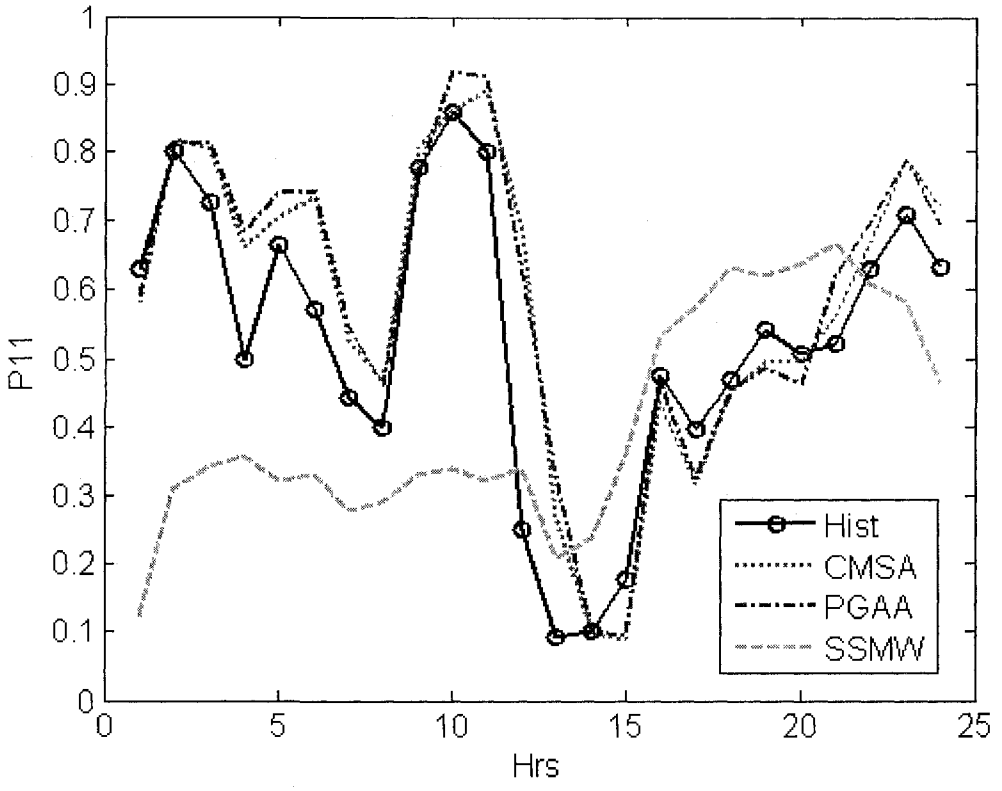


Figure 5-A.15 Historical and Disaggregated Occurrence probability P11 for Extension Case

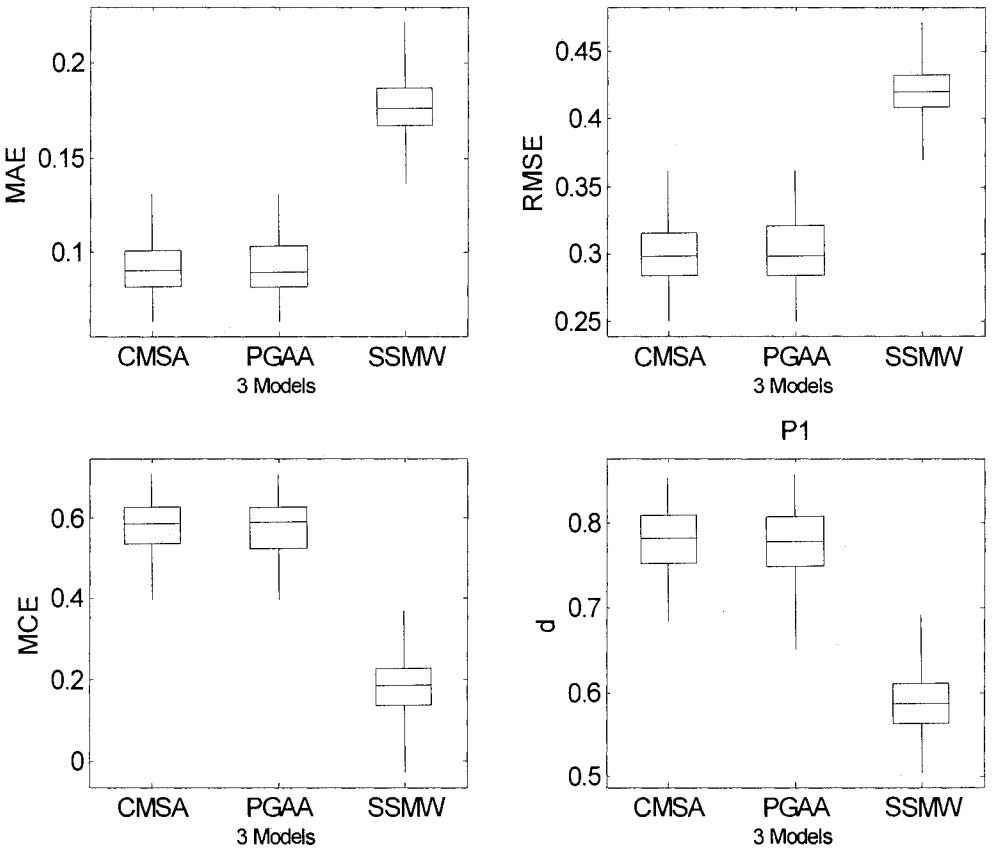


Figure 5-A.16 Performance criteria of the probability P1 for Extension case

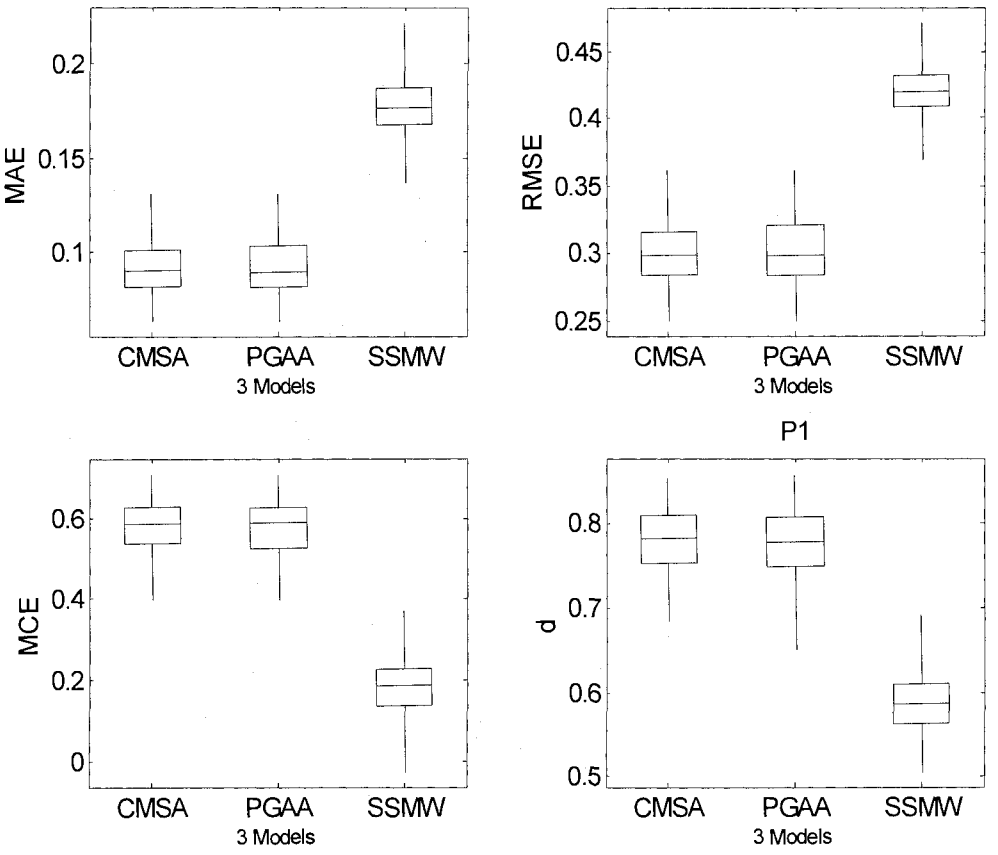


Figure 5-A.17 Performance criteria of the probability P01 for Extension case

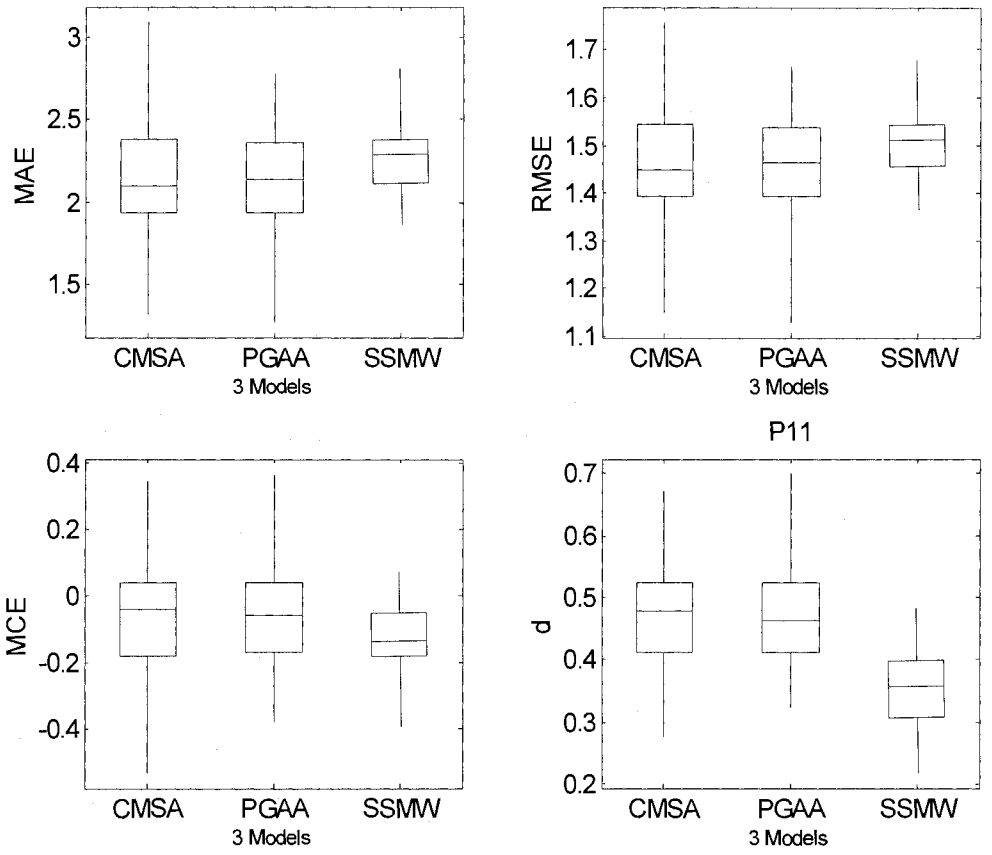


Figure 5-A.18 Performance criteria of the probability P11 for Extension case

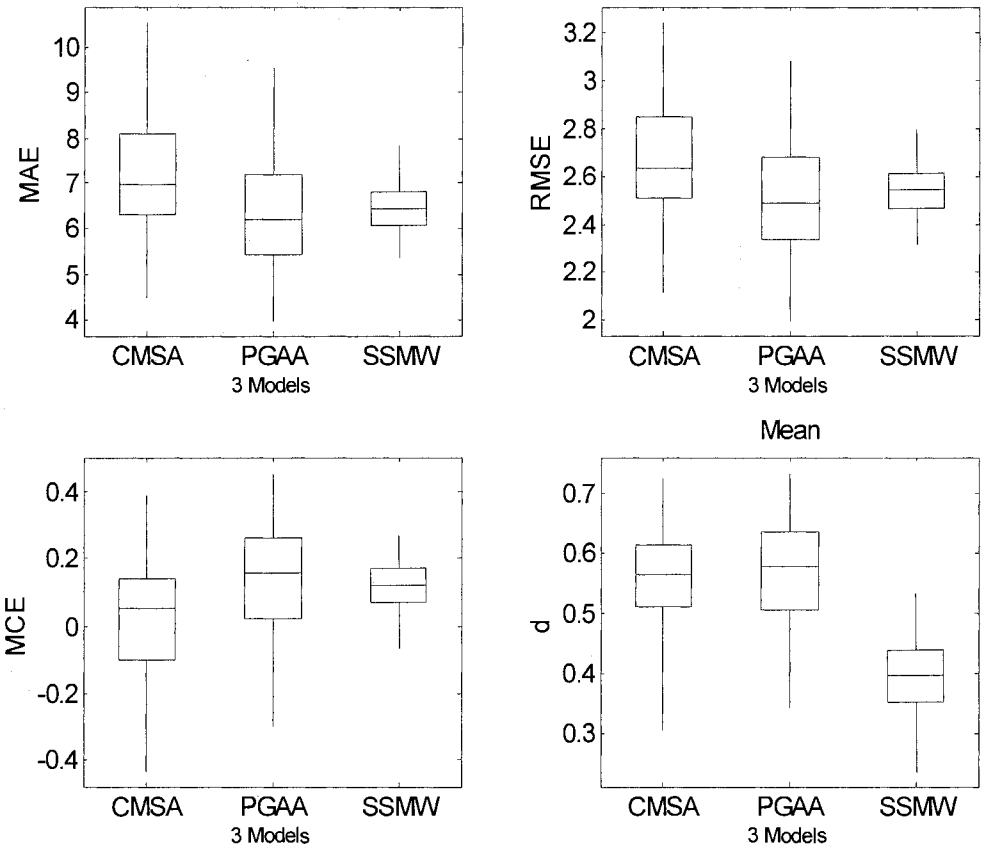


Figure 5-A.19 Performance criteria of the Mean for Extension case

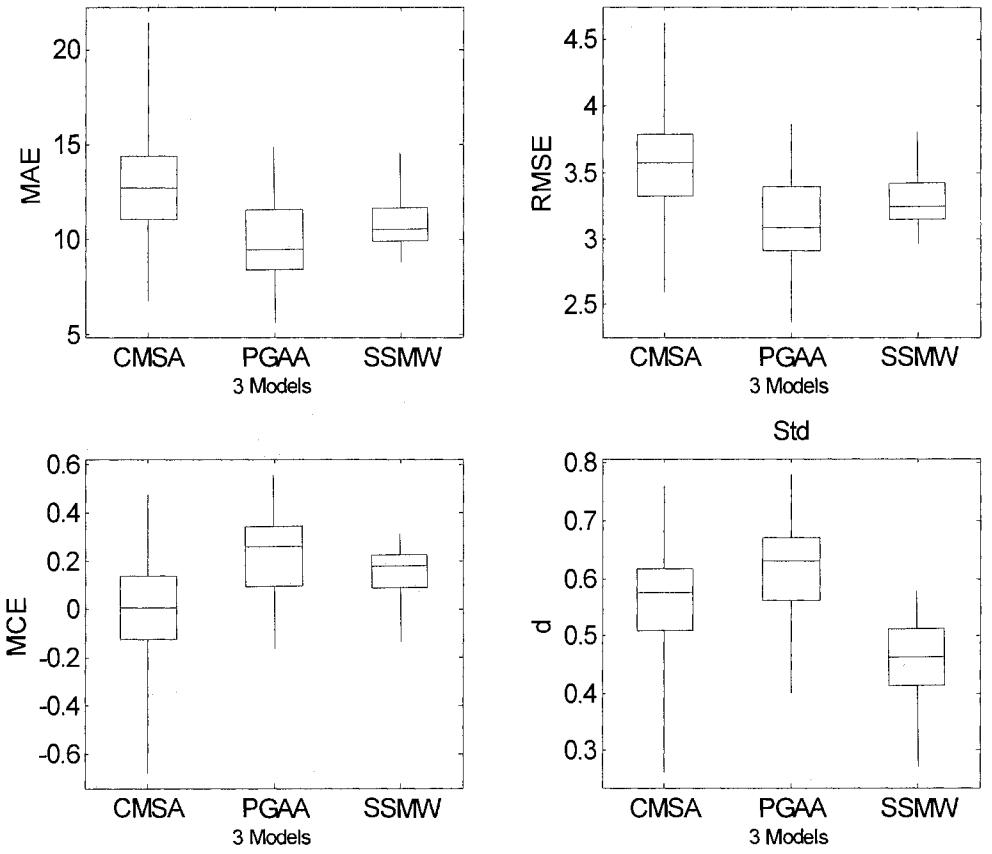


Figure 5-A.20 Performance criteria of the Standard Deviation for Extension case

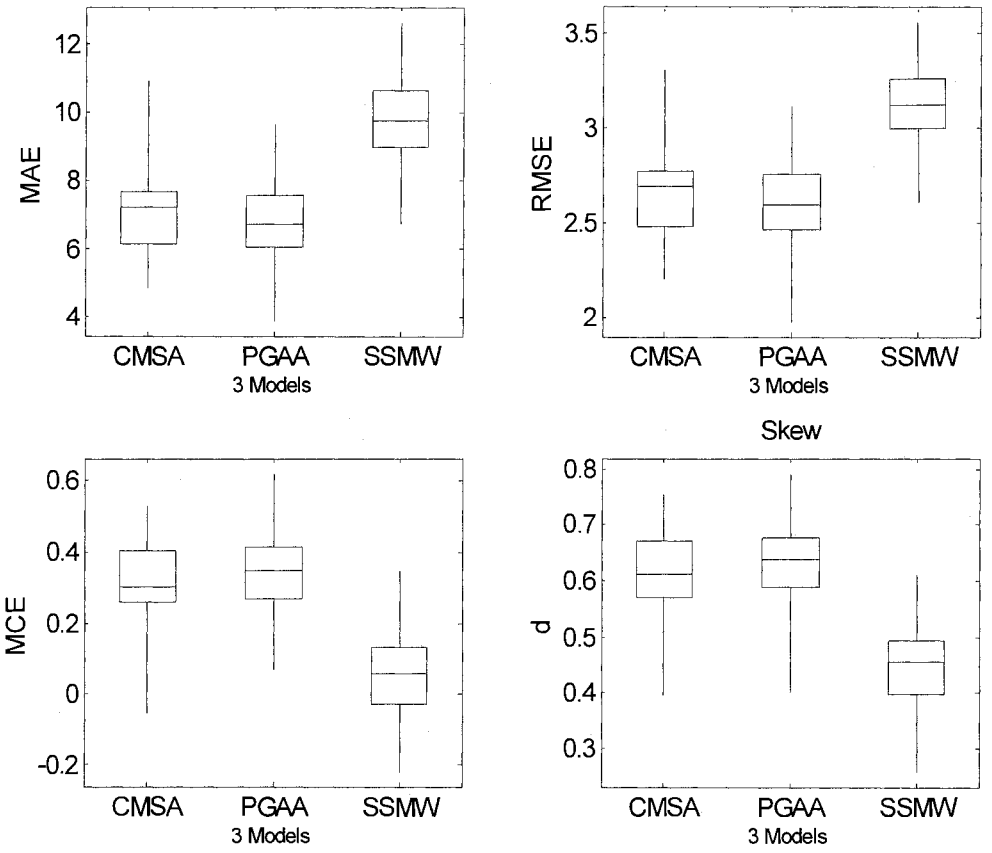


Figure 5-A.21 Performance criteria of the Skewness for Extension case

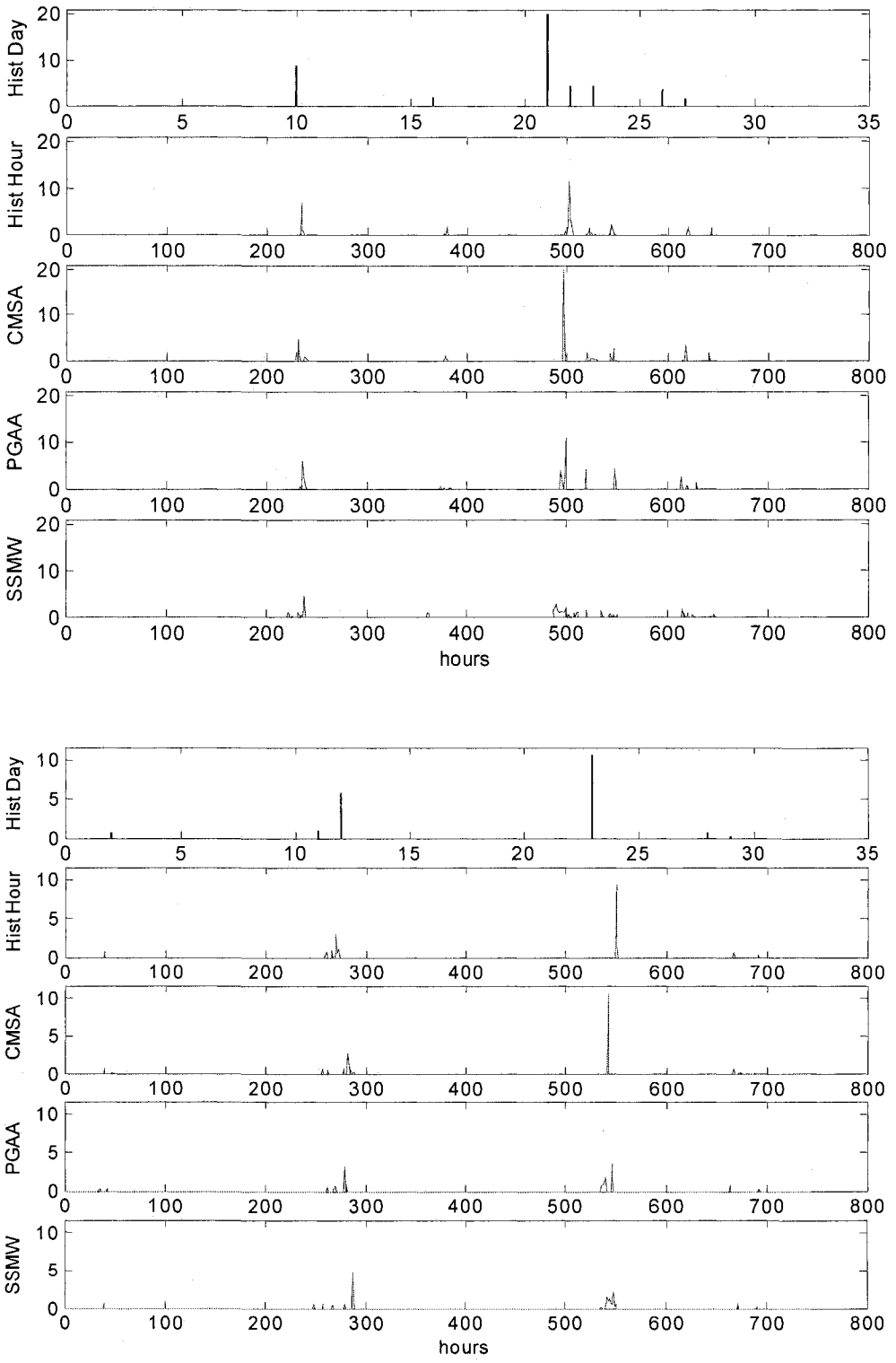


Figure 5-A.22 Realization of the disaggregation and historical hourly and daily for Transfer case

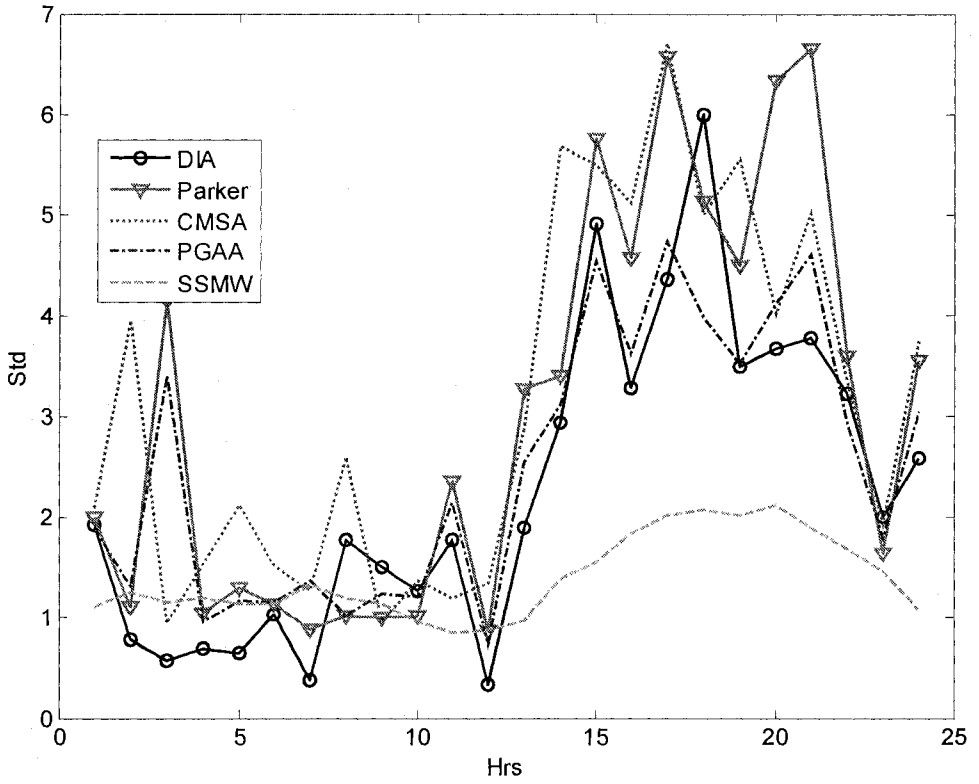


Figure 5-A.23 Historical and Disaggregated hourly Standard Deviation for Transfer Case

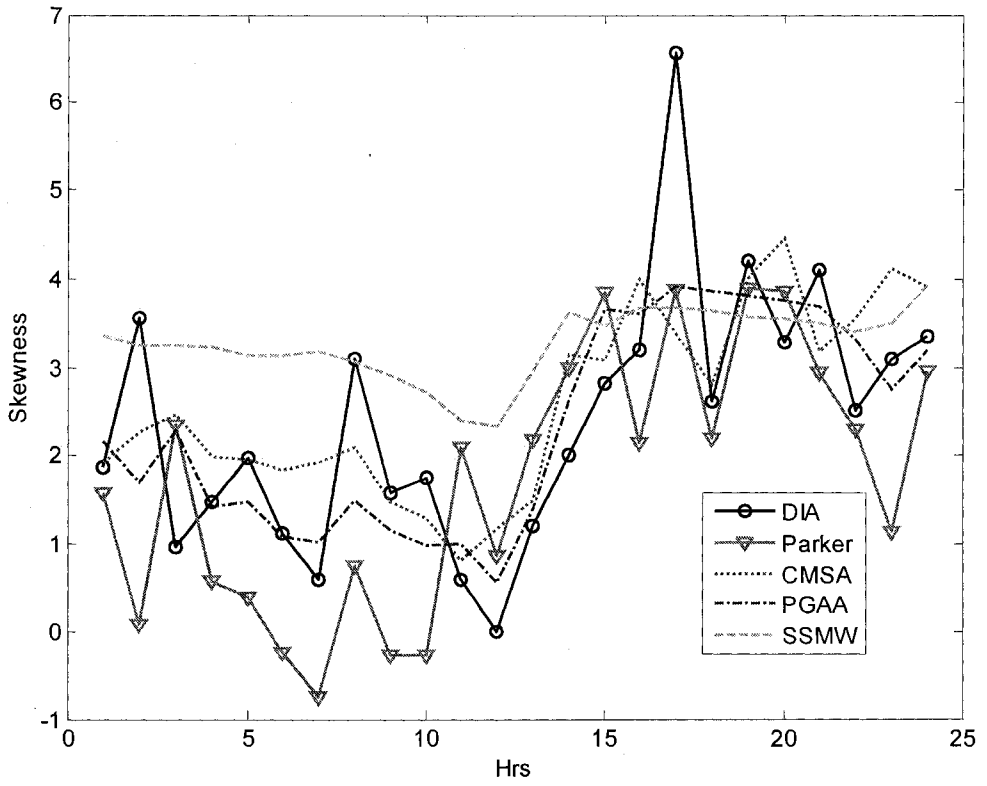


Figure 5-A.24 Historical and Disaggregated hourly Skewness for Transfer Case

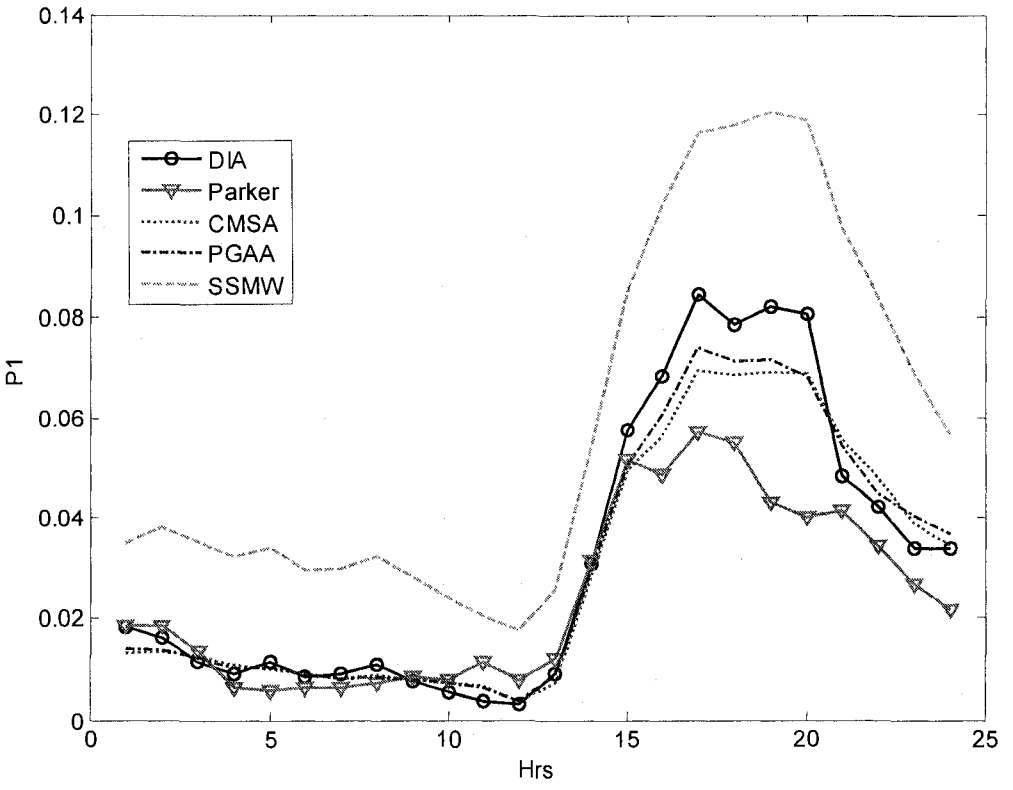


Figure 5-A.25 Historical and Disaggregated Occurrence probability P1 for Transfer Case

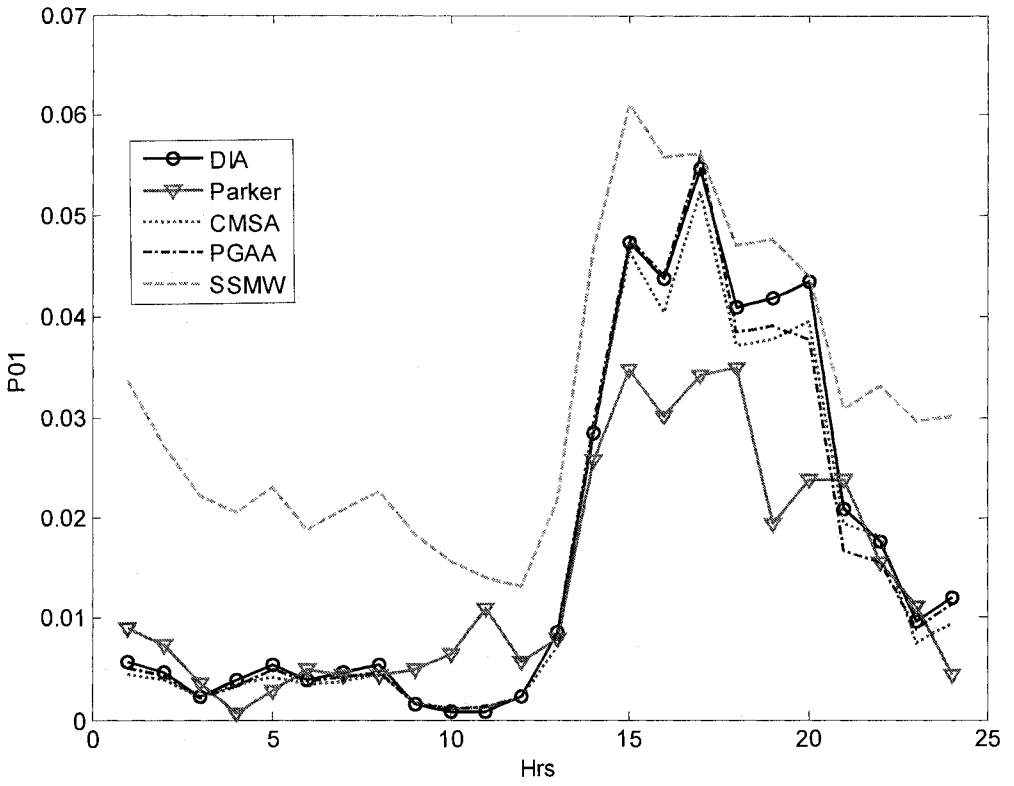


Figure 5-A.26 Historical and Disaggregated Occurrence probability P01 for Transfer Case

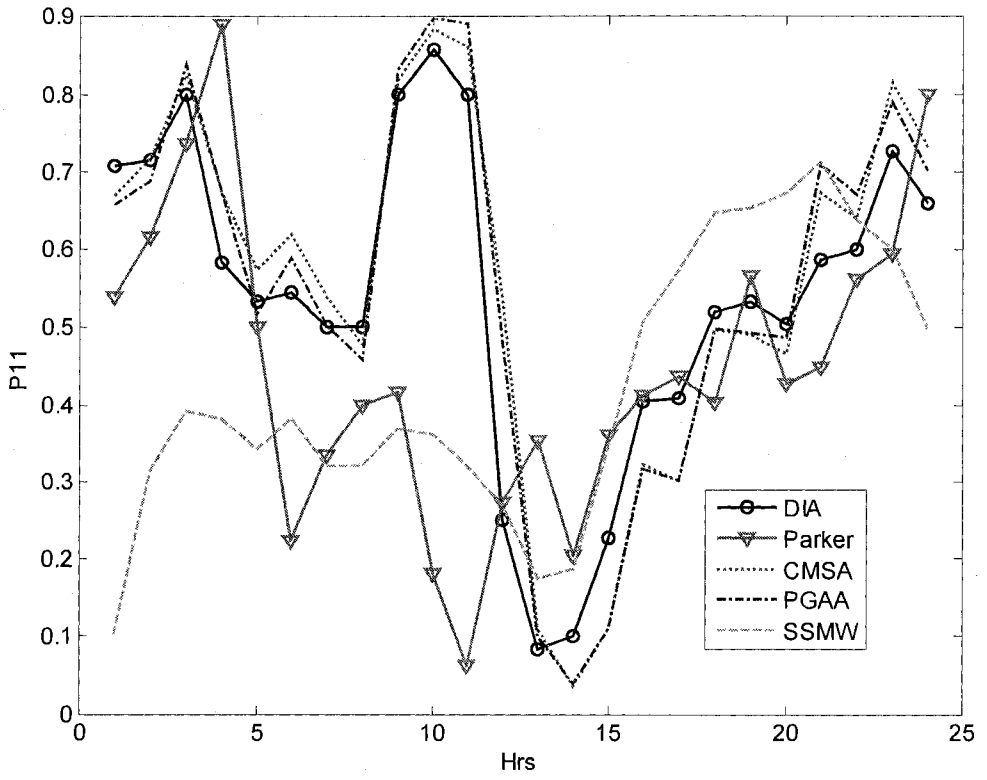


Figure 5-A.27 Historical and Disaggregated Occurrence probability P11 for Transfer Case

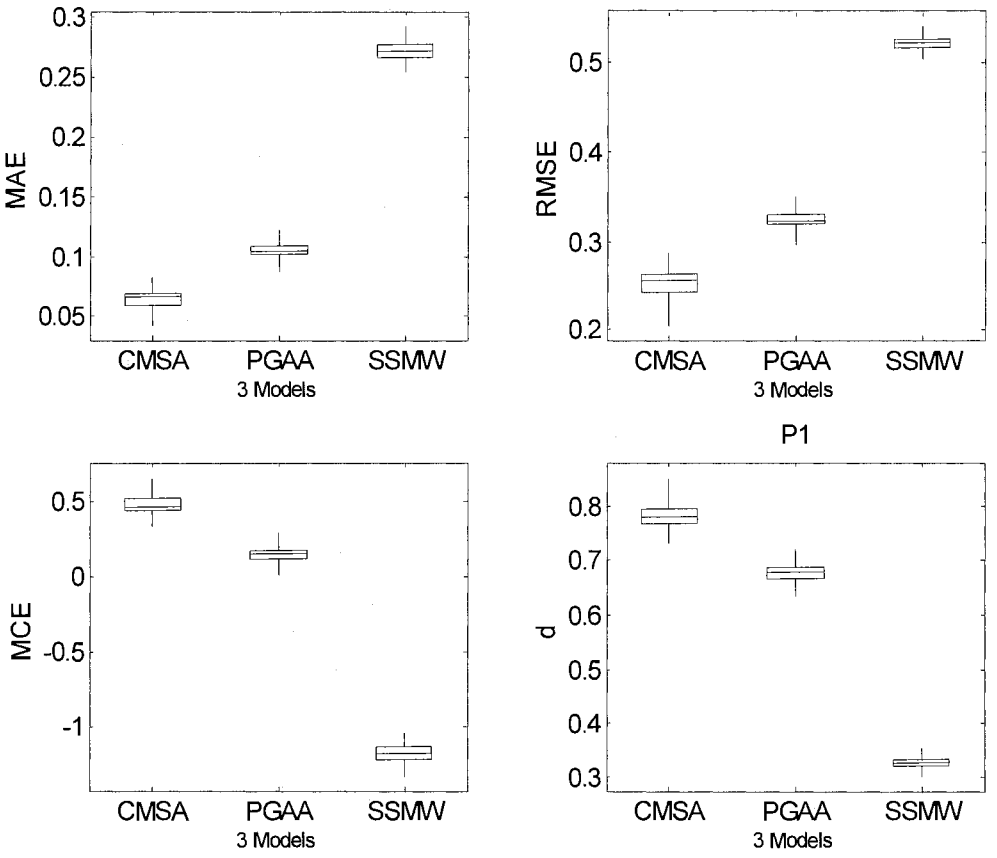


Figure 5-A.28 Performance criteria of the probability P1 for Transfer case

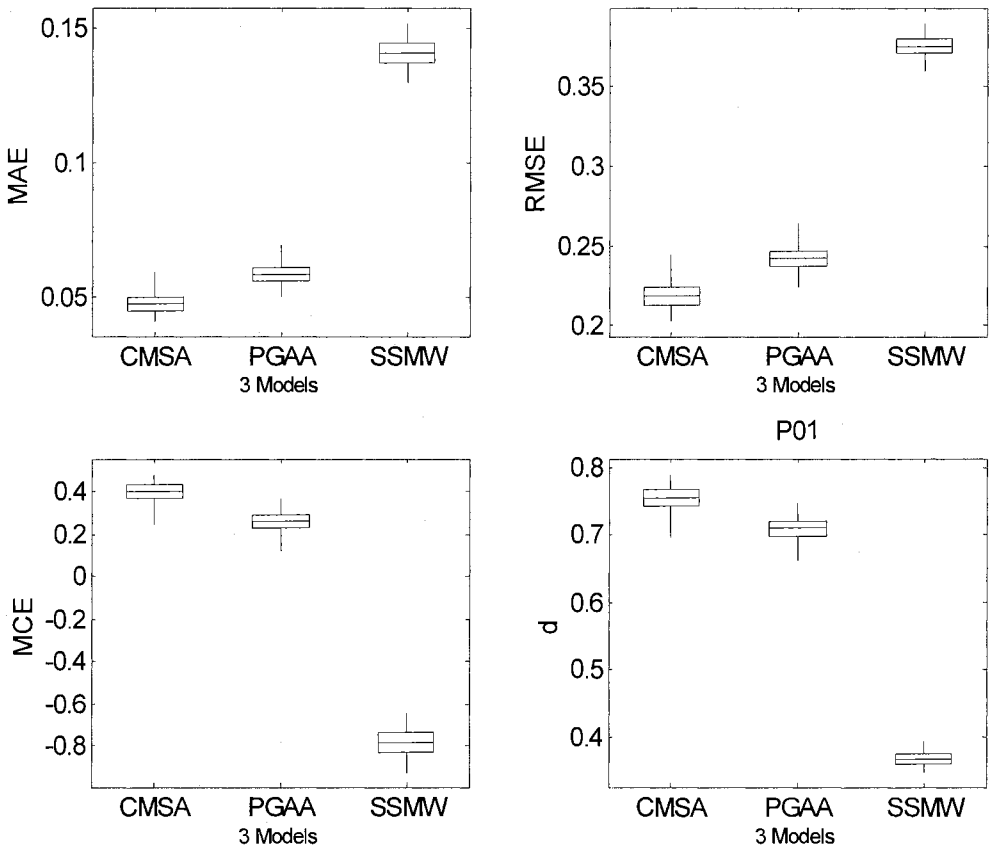


Figure 5-A.29 Performance criteria of the probability P01 for Transfer case

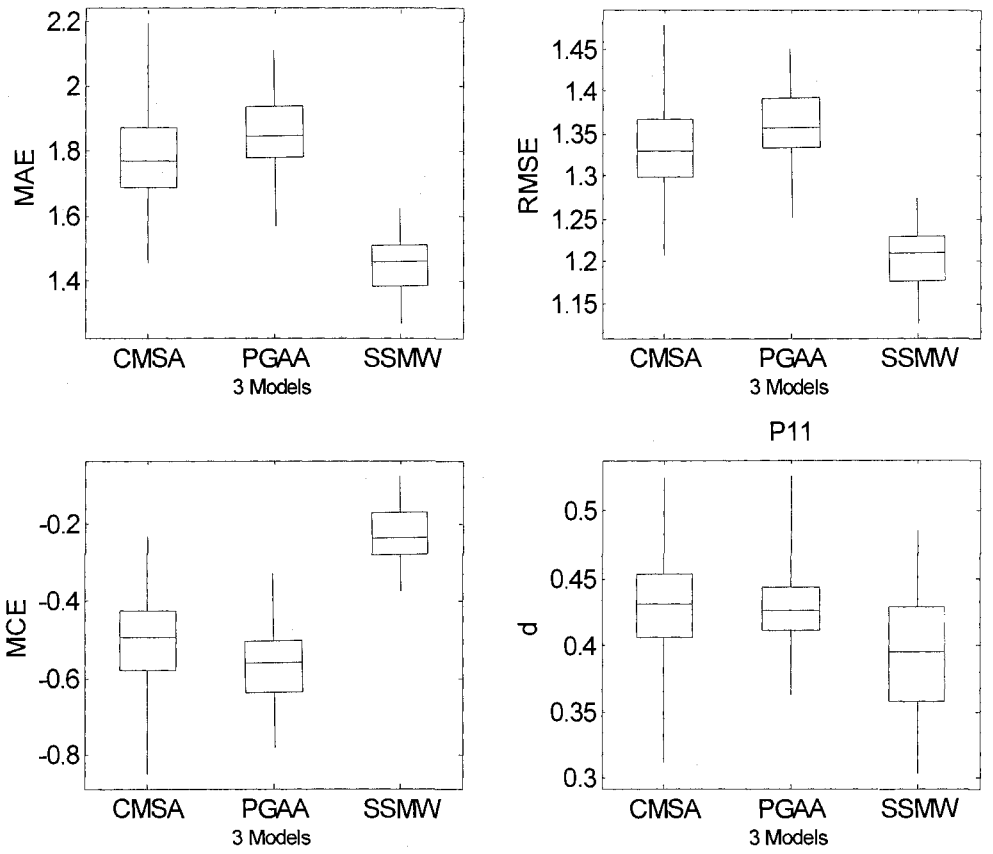


Figure 5-A.30 Performance criteria of the probability P11 for Transfer case

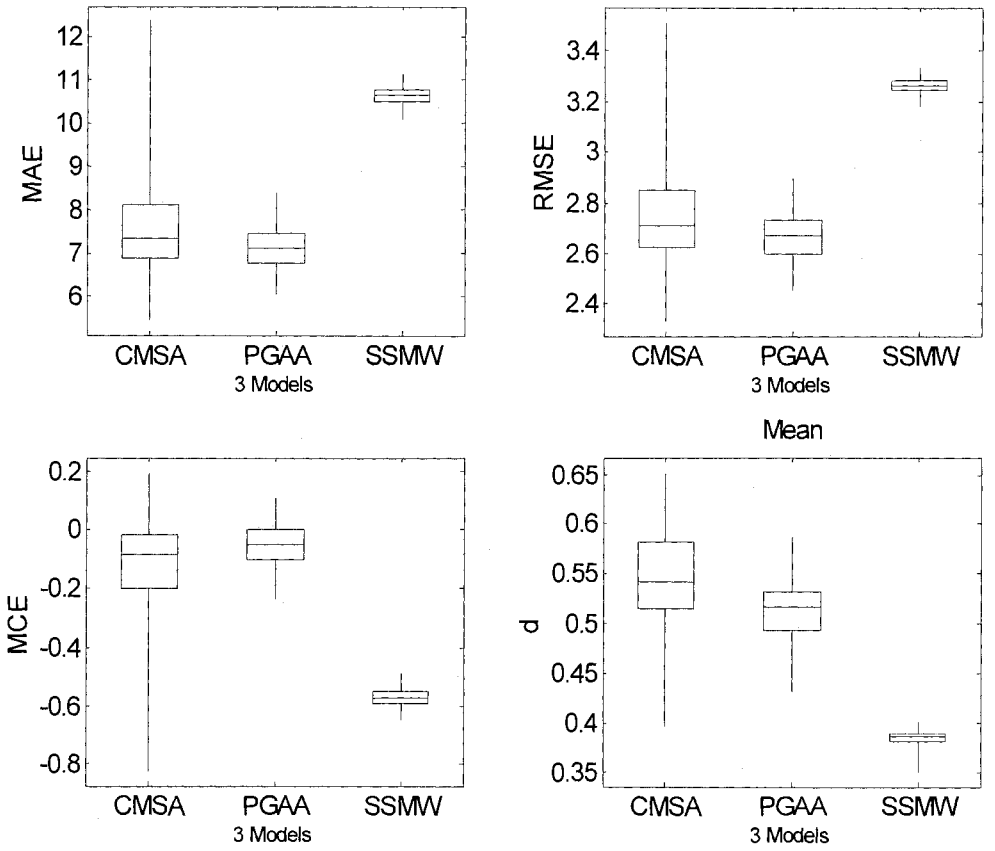


Figure 5-A.31 Performance criteria of the Mean for Transfer case

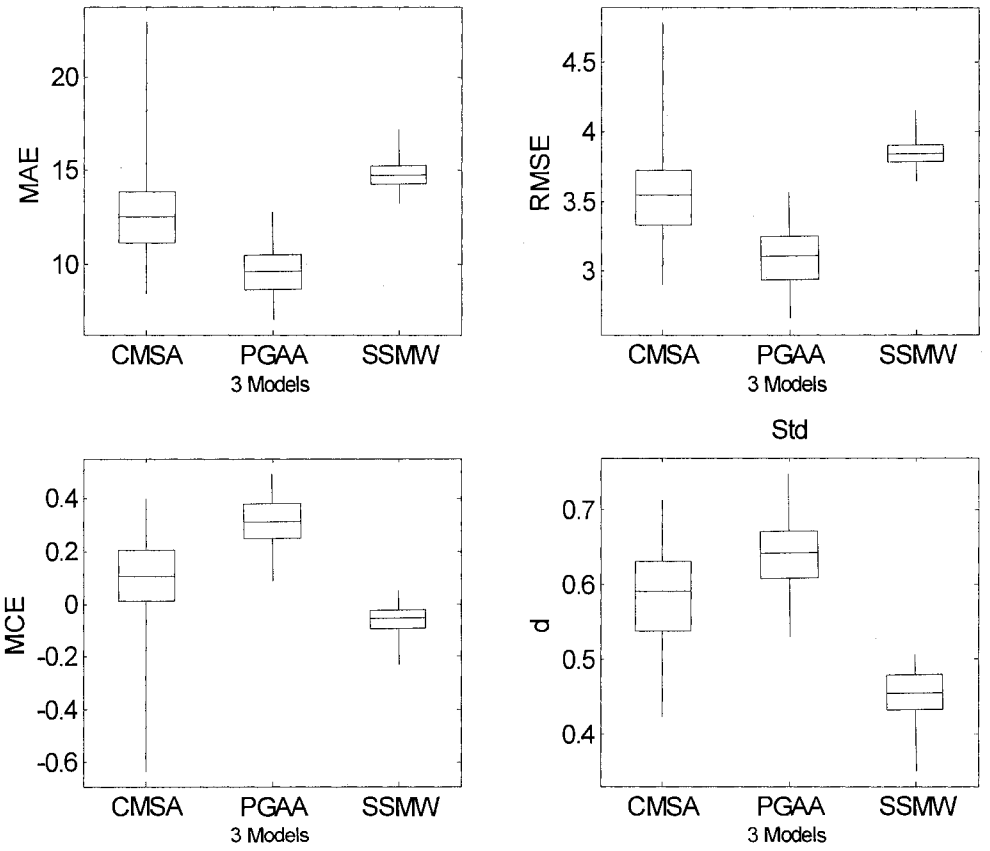


Figure 5-A.32 Performance criteria of the Standard Deviation for Transfer case

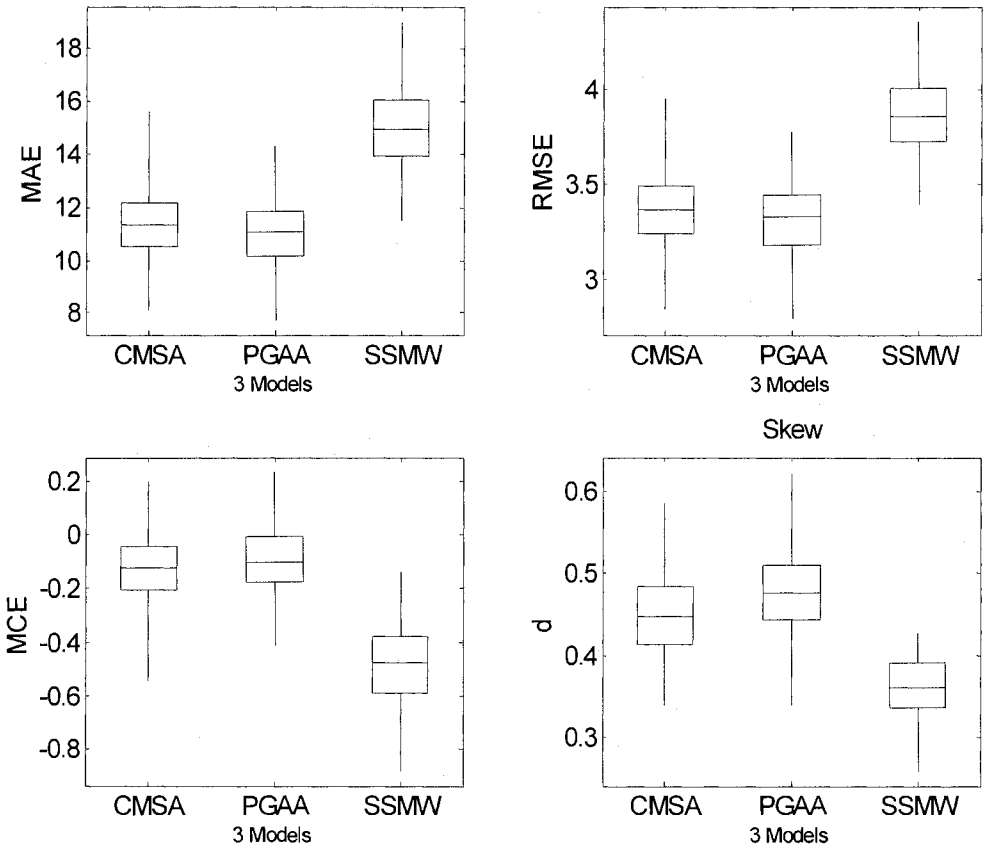


Figure 5-A.33 Performance criteria of the Skewness for Transfer case

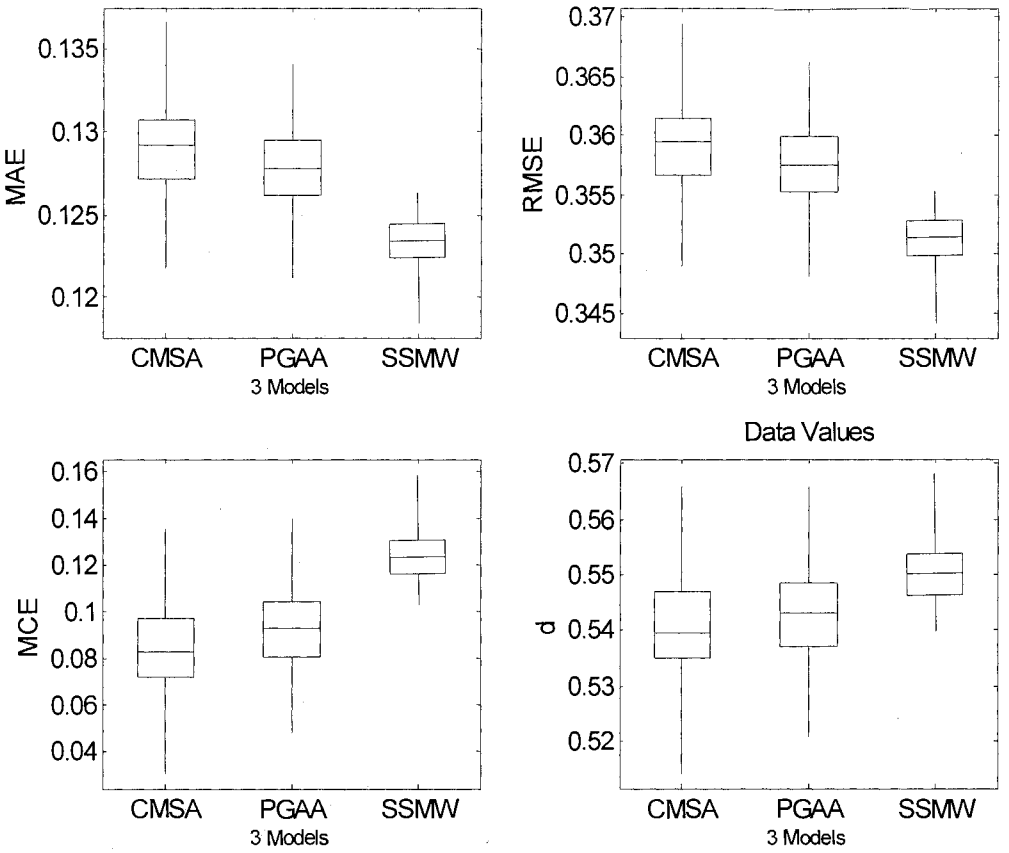


Figure 5-A.34 Performance criteria of the Direct Data Values for Transfer case

CHAPTER VI

CONCLUSIONS, CONTRIBUTIONS, AND RECOMMENDATIONS

6.1 Conclusions

Stochastic generation models are required for various purposes such as drought analysis, reservoir planning of a complex system, and water quality modeling. Over the decades many alternative models have been proposed based on parametric and non-parametric techniques. For streamflow data nonparametric modeling techniques are focused. From the meticulous investigations over the existing models, new models are proposed and some existing models are enhanced. The proposed and enhanced model in this study eliminates the limitations and drawbacks that the existing nonparametric models have. Those developed models are tested with various streamflow data, mainly in the Colorado River system. The results show that they are reliable and useful models to simulate streamflows of a single site and a large river basin even with intermittent and non-intermittent sites jointly. For rainfall data, the existing disaggregation models are improved to account for the diurnal cycle in hourly data. The improved three models are tested with rainfall data in Colorado stations. The results showed that two among three models reproduce the diurnal cycle properly as well as the key statistics of hourly data.

The specific conclusions drawn from the research are:

(1) The model KNNR with Gamma KDE perturbation and aggregate or pilot variable produces feasible generated data that are not par of the historical data and preserve the long-term variability. Extensive comparisons and applications of the proposed techniques show significant improvements over the existing models.

(2) The proposed multivariate bootstrapping model with KNN block selection and Genetic Algorithm mixture yield new spatial patterns different from the part of the historical data and avoid the discontinuity in block selection. Further, the model can be applied for cases of joint intermittent and non-intermittent flows. This capability is very useful when a multivariate simulation is required on a large river basin such as the Colorado River system which includes arid, semi-arid, and moderate zone. The arid and semi-arid zone contains the intermittent data.

(3) The detailed investigation of existing disaggregation models uncovered a number of drawbacks. Appropriate modifications and simpler structure is suggested employing the accurate adjusting and the KNNR selection accompanied by the Genetic Algorithm mixture. The proposed model has been shown to perform more efficiently and better than the existing one.

(4) Three distinctive disaggregation models such as CMSA, PGAA, and SSMW are enhanced to take the diurnal cycle into account in the disaggregated hourly data. From various tests and comparisons, CMSA and PGAA models well

reproduce the diurnal cycle as well as the key statistics. And PGAA is the easiest model to apply.

6.2 Summary of Contributions

Nonparametric models for streamflow data are proposed and some existing models are improved if applicable. Those proposed and improved models has better performance than the existing model such as generation of the values not part of the historical data, preservation of long-term variability, applicability to the sites that intermittent and non-intermittent data are combined. The improvements on the disaggregation models of rainfall data now allow to reproduce the diurnal cycle in hourly data.

(1) Univariate model: The developed approach is based on the k-nearest neighbors resampling (KNNR). The critical drawback of the existing KNNR generation model is that it only produces historical values. This drawback is overcome by using the Gamma Kernel Density Estimate. New parameterization of the Gamma kernel is proposed and compared with the previous one revealing some superior features. Also the reproduction of the long-term variability cannot be achieved with the existing KNNR model. Here the interannual variable and the pilot variable are proposed to reproduce long-term variability. The suggested models have been tested with the data of the Colorado River and Niger River and revealed successful preservation of the key statistics and drought and storage statistics.

(2) Multivariate Model: The critical drawbacks of the existing nonparametric generation model is to generate the same values as the historical, the repetition of the same seasonal pattern, and no variation spatially. The new features of the model proposed herein include (a) the variable block length – the aggregated values to annual or seasonal (in case of monthly generation) will be different from historical, (b) KNNR block selection – the connection between blocks will be preserved, and (c) Genetic Algorithm mixture – spatially different sequences to historical are generated, and (d) Gamma KDE perturbation – different values than the historical data will be generated. The suggested model has been tested using data of Colorado River System and showed better results than those obtained based on the existing model.

(3) Disaggregation Model: The drawbacks of an existing nonparametric disaggregation technique have been examined in some detail. Firstly, the correlation between the first month of the current year and the last month of the previous year is not preserved and the proper spatial or temporal mixing cannot be reproduced. These drawbacks are remedied with the suggested much simpler model. The proposed model uses (a) the KNNR selection of the aggregate variable followed by accurate adjusting for the disaggregate variable data corresponding to the selected aggregate variable (b) the consideration of the correlation between the first month of the current year and the last month of the previous year with including the condition of the last month of the previous year when the lower-level variable is obtained with KNNR selection (c) Genetic Algorithm mixture for obtaining more variable pattern than the

historical data. The suggested model has been tested using data of the Colorado River System. The results are compared with the existing models and showed that the referred drawbacks are all eliminated.

(4) Daily rainfall disaggregation model: The various models for disaggregating daily rainfall data do not consider the diurnal cycle in the hourly data. Three distinctive disaggregation models were improved to disaggregate daily rainfall data into hourly so that the diurnal cycle of hourly data are properly taken into account. The capability of the reproduction diurnal cycle will be useful as input data for dam operation and water quality modeling that diurnal cycle is important feature to be considered.

6.3 Recommendations

Two recommendations for further studies are:

(1) Include exogenous variable into the KGKP model as a pilot variable to reflect global climate effect on streamflow. Wavelet analysis and Hilbert-Huang transformation may be useful techniques to find a component to be used as an exogenous variable from climate variable (e.g. ENSO index).

(2) Currently, a drawback of nonparametric disaggregation model is to perform spatial and temporal disaggregation one-by-one. More than one step of spatial disaggregation after temporal disaggregation for a river basin system induce the underestimation of the seasonal correlation in the spatially lower-level stations.

Therefore, a nonparametric disaggregation model that can implement the spatial-temporal disaggregation at the same time could be useful in this sense.