APPLICATION OF ENTROPY THEORY

IN HYDROLOGIC ANALYSIS AND SIMULATION

A Dissertation

by

ZENGCHAO HAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012

Major Subject: Biological & Agricultural Engineering

APPLICATION OF ENTROPY THEORY

IN HYDROLOGIC ANALYSIS AND SIMULATION

A Dissertation

by

ZENGCHAO HAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Vijay P. Singh
Committee Members,	Yongheng Huang
	Ralph A. Wurbs
	Hongbin Zhan
	Mohsen Pourahmadi
Head of Department,	Stephen W. Searcy

May 2012

Major Subject: Biological & Agricultural Engineering

ABSTRACT

Application of Entropy Theory in Hydrologic Analysis and Simulation.

(May 2012)

Zengchao Hao, B.S., China Agricultural University, China;

M.S., Tsinghua University, China

Chair of Advisory Committee: Dr. Vijay P. Singh

The dissertation focuses on the application of entropy theory in hydrologic analysis and simulation, namely, rainfall analysis, streamflow simulation and drought analysis.

The extreme value distribution has been employed for modeling extreme rainfall values. Based on the analysis of changes in the frequency distribution of annual rainfall maxima in Texas with the changes in duration, climate zone and distance from the sea, an entropy-based distribution is proposed as an alternative distribution for modeling extreme rainfall values. The performance of the entropy based distribution is validated by comparing with the commonly used generalized extreme value (GEV) distribution based on synthetic and observed data and is shown to be preferable for extreme rainfall values with high skewness.

An entropy based method is proposed for single-site monthly streamflow simulation. An entropy-copula method is also proposed to simplify the entropy based method and preserve the inter-annual dependence of monthly streamflow. Both methods

are shown to preserve statistics, such as mean, standard deviation, skenwess and lag-one correlation, well for monthly streamflow in the Colorado River basin. The entropy and entropy-copula methods are also extended for multi-site annual streamflow simulation at four stations in the Colorado River basin. Simulation results show that both methods preserve the mean, standard deviation and skewness equally well but differ in preserving the dependence structure (e.g., Pearson linear correlation).

An entropy based method is proposed for constructing the joint distribution of drought variables with different marginal distributions and is applied for drought analysis based on monthly streamflow of Brazos River at Waco, Texas. Coupling the entropy theory and copula theory, an entropy-copula method is also proposed for constructing the joint distribution for drought analysis, which is illustrated with a case study based on the Parmer drought severity index (PDSI) data in Climate Division 5 in Texas.

DEDICATION

To my family

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely thank my advisor, Dr. Vijay P. Singh, for his invaluable guidance, support and encouragement throughout my doctoral work. I also would like to thank Dr. Yongheng Huang, Dr. Ralph A. Wurbs, Dr. Hongbin Zhan, and Dr. Mohsen Pourahmadi for their constructive suggestions to improve the preliminary proposal and final dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Characterization of hydrologic events, such as rainfall, streamflow and drought, is needed for water resources planning and management. Due to the stochastic nature of hydrologic phenomena, stochastic methods are commonly used. For rainfall analysis, a proper distribution is generally needed to investigate statistical properties of rainfall quantiles and extrapolate beyond the available data for engineering purposes. For streamflow simulation, synthetic streamflow with statistical properties similar to those historical streamflows are required for evaluation of alternative designs and policies against the range of sequences that are likely to occur in the future. A joint distribution with different marginal distributions is generally needed to characterize the correlation between drought variables and distribution property of individual drought variables to analyze return periods corresponding to some occurrence levels of drought events.

A proper characterization of hydrologic events necessitates the consideration of uncertainty in the estimation from limited observations. Entropy theory defines a measure of uncertainty or information and thus provides a proper way to characterize hydrologic events with stochastic nature. Application of entropy theory to rainfall analysis, streamflow simulation and drought analysis constitutes the objective of this study.

Rainfall frequency analysis is needed for the construction of intensity-duration-

---

This dissertation follows the style of *Water Resources Research*.

frequency (IDF) curves which are used for engineering design of drainage systems, culverts, roadways and parking lots. Extreme values, such as the annual rainfall maxima, are generally used for frequency analysis. The generalized extreme value (GEV) distribution, which is based on the extreme value theory, has been commonly used for modeling extreme rainfall in different states. However, there are a variety of studies for extreme rainfall analysis in which other distributions have often been employed. Extreme rainfall exhibits different properties for different durations and in different regions. The question is: what is the effect of time duration, climate zone and the distance from the Gulf of Mexico on the frequency distribution of annual rainfall maxima? In this study, the State of Texas was selected as the study area and we try to answer these three questions to provide an insight into the analysis of extreme rainfall.

In chapter II, the change in the form of the annual rainfall maximum frequency distribution with changes in time duration, climate zone, and distance from the Gulf of Mexico is investigated. An entropy based distribution is then proposed to model the annual rainfall maxima. The performance of the proposed method is compared with the commonly used GEV distribution based on the synthetic data and real observations.

Streamflow is a component of a variety of hydrologic analysis, such as the reservoir planning and operation. Since historical streamflow does not allow for the evaluation of alternative designs and policies against the range of sequences that are likely to occur in the future, synthetic streamflow data are useful in water resources studies. It is desired that synthetic streamflow is similar to historical streamflow and preserves moment statistics (such as mean, standard deviation, and skewness), and

dependence structure (such as lag-one correlation). For traditional methods, the Gaussian assumption is generally needed for parametric methods in which transformation techniques are employed. However, some problems arise, such as the generation of negative values due to the Gaussian assumption and the bias in simulated statistics due to the transformation techniques. In addition, the lag-one correlation (or Pearson product-moment correlation coefficient) only measures the linear dependence of random variables, which may not be adequate in reality. Moreover, some unusual features, such as the bimodality, may exist in the probability density function of streamflow data. It is difficult for the commonly used parametric approach to represent these features. Though the mixed distribution can be used to resolve the bimodality, bias in the statistics of streamflow may occur.

In Chapter III, an entropy based method is proposed for monthly streamflow simulation. With the joint distribution of monthly streamflows of two adjacent months derived using the entropy theory, monthly streamflow is then generated by sequential sampling from the conditional distribution. The proposed entropy-based method does not rely on the assumption of the marginal distributions to be normal and data transformation is not needed. Therefore, issues with the data transformation existing in the commonly used parametric approaches can be avoided. This method can be extended to model more statistics of the underlying streamflow data if needed. The disadvantage of the entropy based method is that the method will be computationally cumbersome when more statistics need to be modeled.

In Chapter IV, an entropy-copula method is proposed for single-site monthly streamflow simulation in which the joint distribution is constructed using the copula theory with the marginal distribution derived using the entropy theory. The entropy-copula method simplifies the entropy method for monthly streamflow simulation in that less number of parameters needs to be estimated simultaneously. Furthermore, the entropy-copula method is also capable of modeling the nonlinear dependence of streamflow between different months due to the copula component. The proposed entropy-copula method is also extended with an aggregated variable to guide the sequential simulation to improve the preservation of high-order correlation and preserve the inter-annual dependence of monthly streamflow.

In Chapter V, both the entropy method and entropy-copula methods are extended to higher dimension for multi-site annual streamflow simulation. The difference between two methods lies in modeling the dependence structure of streamflow. For the entropy method, the joint constraints are used for modeling the dependence while the copula is used for modeling the dependence for the entropy-copula method. Application of the proposed method based on annual streamflow from four stations in Colorado River basin illustrates the effectiveness and difference of the entropy method and entropy-copula method for streamflow simulation.

Drought analysis is important for water resources planning and management. A drought event can be characterized with certain properties, such as duration and severity. Drought duration and severity, assumed as random variables, have been commonly used for drought analysis and a traditional way for characterizing drought is fitting an

empirical distribution to drought duration and severity. The joint distribution is needed to model the correlation between drought variables. Traditional joint distributions that have been applied for drought analysis generally assume that the marginal distribution is of the same type. The copula method has been employed extensively for modeling drought duration and severity with the attractive property that the marginal distribution can be of different forms. However, the marginal distributions are often derived by empirically fitting to the data.

In Chapter VI, an entropy based distribution is proposed for constructing the joint distribution of drought variable. The feature of the proposed entropy-based distribution is that the marginal distributions can be of different forms. The advantage of the proposed method is the marginal distribution can be derived with whatever is known from observations and is not restricted by the empirical forms of distributions.

In Chapter VII, an entropy-copula method is proposed for constructing a joint distribution for drought analysis. Flexible distribution forms can be derived with the entropy method and the commonly used distributions can also be derived as special cases of the entropy based distribution. A variety of copulas have been proposed that are capable of modeling different dependence structures. The joint distribution constructed with the copula method with the marginal distributions derived from the entropy theory is expected to be capable of modeling drought variables separately and jointly.

The general conclusions of this study are covered in Chapter VIII.

CHAPTER II

ENTROPY BASED METHOD

FOR RAINFALL ANALYSIS

2.1 Introduction

Rainfall frequency analysis is used for constructing intensity-duration-frequency (IDF) curves which are needed for a range of hydrologic designs, including drainage systems, culverts, roadways, parking lots, runways, and so on. Extreme rainfall values, such as annual rainfall maxima, are of interest in modeling floods and quantifying the effect of climate change. From the fitted distribution, statistical properties of extreme rainfall values can be investigated and extrapolated beyond the available data for engineering purposes.

The generalized extreme value (GEV) distribution is one of the frequently employed probability distributions for modeling and characterizing extreme values. Derived from the extreme value theory, it is a three-parameter distribution encompassing three classes of distributions, namely, Gumbel, Frechet and Weibull. This distribution has been used for extreme rainfall frequency analysis in different areas of the world. *Schaefer* [1990] used the GEV distribution for frequency analysis of annual rainfall maxima of durations of 2 h, 6 h and 24 h for the state of Washington. *Huff and Angel* [1992] selected the GEV distribution to model the distribution of annual rainfall maxima for durations from 5 minutes to 10 days in mid-western United States. *Parrett* [1997] also used the GEV distribution to construct dimensionless frequency curves of annual

rainfall maxima of durations of 2 h, 6 h and 24 h within each region in Montana. Using the L-moment ratio diagram, *Asquith* [1998] determined that the GEV distribution was an appropriate distribution for modeling the distribution of annual maxima for durations from 1 to 7 days. *Alila* [1999] showed that the annual rainfall extremes for durations from 5 minutes to 24 hours in Canada were better described by the GEV distribution than other distributions, such as the generalized logistic (GLO), Pearson type 3 (P3) and EV1 distributions.

Extreme rainfall exhibits different properties for different durations in different regions. Analysis of rainfall characteristics is important for choosing a suitable rainfall distribution and consequently estimating rainfall quantiles. Therefore, the objective of this study is to investigate the change in the form of the annual rainfall maxima frequency distribution with changes in time duration, climate zone, and distance from the Gulf of Mexico; and then derive an entropy-based distribution that is sufficiently flexible for characterizing rainfall distributions for different durations in different climatic regions or at different distances from the sea. The performance of the proposed entropy based distribution is assessed using synthetic data through Monte Carlo simulation and real observations and is shown to be a promising alternative distribution to the commonly used GEV distribution for modeling extreme rainfall values, especially for observations with high skewness.

The study is organized as follows. In section 2.2, the change in form of empirical distribution of annual rainfall maxima is investigated. Using the entropy theory, a generalized distribution is derived in section 2.3 and the performance of this distribution

is assessed by comparing with the GEV distribution in Section 2.4. After the application of the proposed entropy based distribution in section 2.5, conclusions are given in section 2.6.

2.2 Empirical frequency distribution

2.2.1 Study area

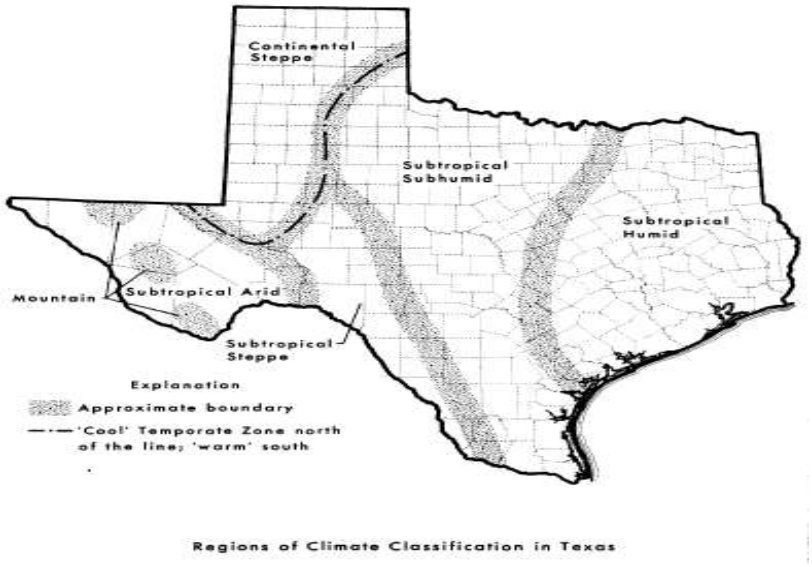

Figure 2. 1 Regions of climate zones in Texas ([Larkin and Bomar, 1983]).

The area selected for this study is the state of Texas (longitude: 93 °31' W to 106 °38' W, latitude: 25 °50' N to 36 °30' N). The climate of Texas is strongly influenced by physical features including the Gulf of Mexico. The passage of frontal systems from northwest and the moist air moving inland from the Gulf of Mexico are the two competing influences that dominate the climate of Texas while proximity to the coast is

the most important factor that determines the regional climatic differences in Texas [*North et al*., 1995].

There are three major types of climate in Texas which are classified as Continental, Mountain and Modified Marine with no clearly distinguishable boundaries, while the modified marine zone is further classified into four "subtropical" zones [*Larkin and Bomar*, 1983; *Narasimhan et al*., 2008], as shown in Figure 2. 1.The Mountain climate is dominant in several mountains of the Trans-Pecos region and is not included in this study. The different climate zones of the Continental and Modified Marine climate are abbreviated as Continental Steppe (CS), Subtropical Arid (SA), Subtropical Humid (SH), Sub-tropical Sub-Humid (SSH) and Sub-Tropical Steppe zone (SST). In addition, the U.S. National Weather Service divided Texas into 10 climate divisions (including Upper Coast, East Texas, High Plain, Trans-Pecos and so on) which are used accordingly in this study.

2.2.2 Data description

Data for 15-minute, hourly, and daily duration for National Weather Service (NWS) stations, as shown in Figure 2. 2, were obtained from the National Climatic Data Center (http://www.ncdc.noaa.gov). The 15 and 45-minute annual maxima were compiled from the 15-minute data. Likewise, the rainfall data for different hourly durations (1-hour and 12-hour) and daily durations (1-day, 7-day and 30-day) were compiled from hourly and daily data, respectively. Annual rainfall maxima data were then obtained from these rainfall data for different durations.

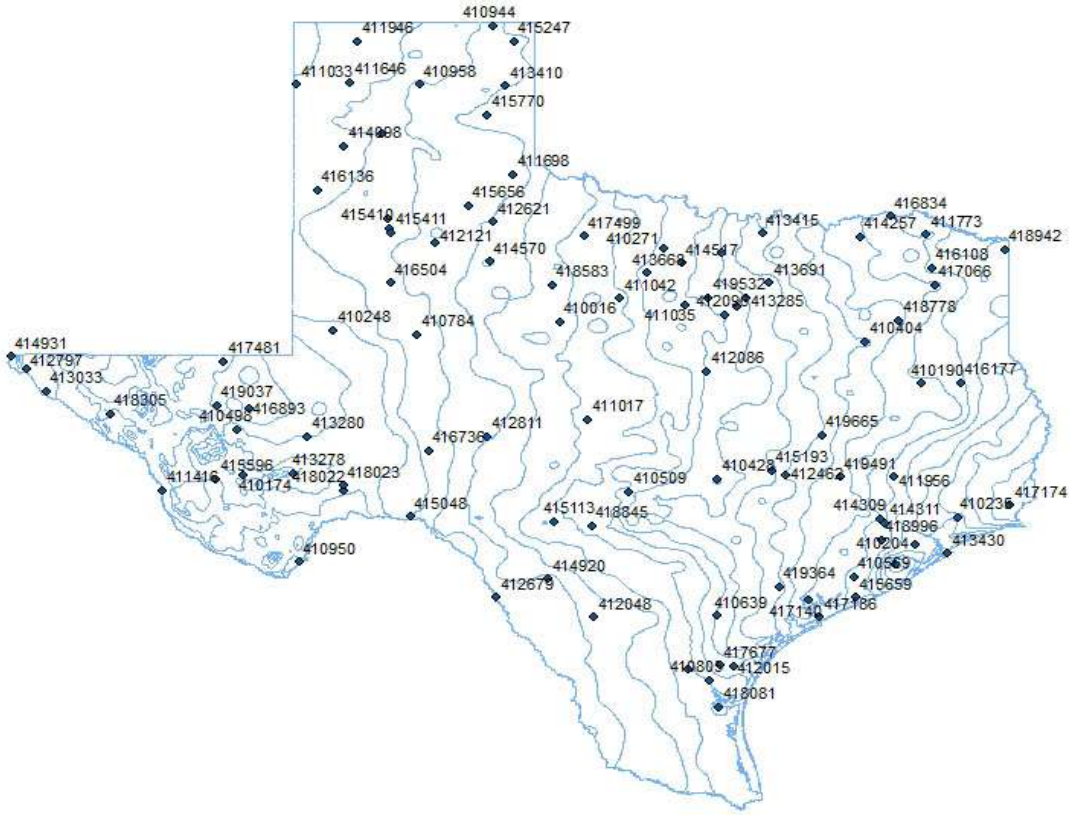

Figure 2. 2 Rainfall stations used in this study.

2.2.3 Change in distribution form with time duration

Histograms of annual rainfall maxima of different durations were prepared for all raingage stations used in this study and those for a sample station (411956) are shown in Figure 2. 3. It was observed that frequency distributions for short durations were more skewed with sharp peaks but tended to be less skewed with increase in the duration. For example, annual rainfall maxima data for station 411956 had a skewness value of 2.7 for 15-minute data but 1.1 for 30-day data (not shown).

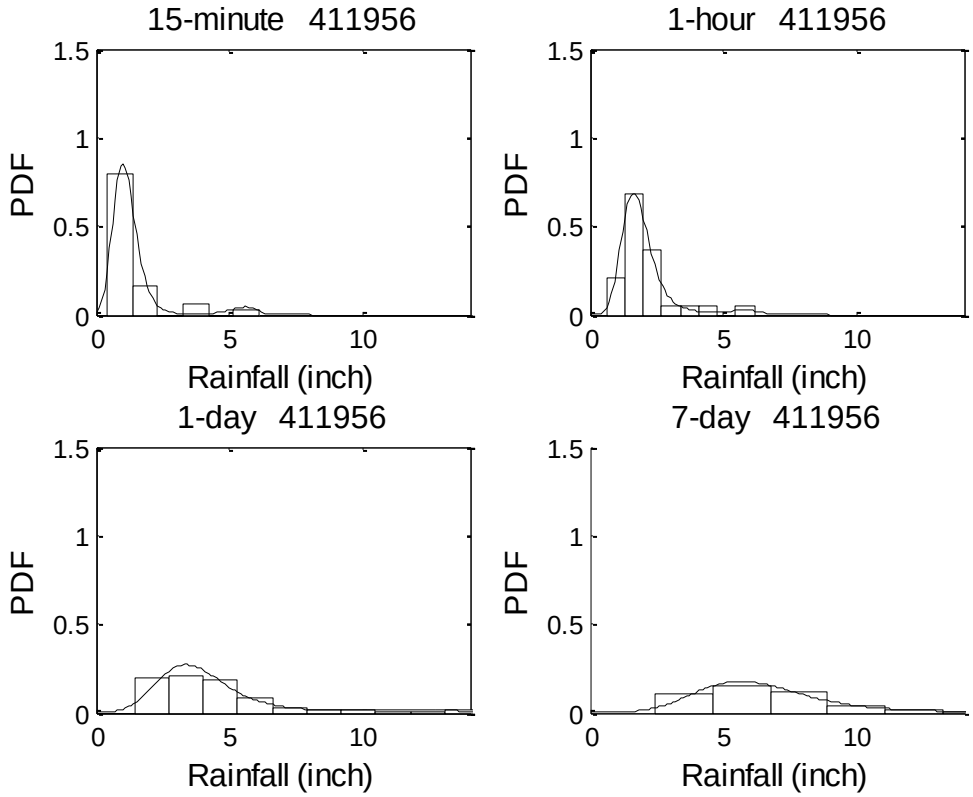

Figure 2. 3 Histograms and probability density functions of rainfall data of different durations (for station 411956 in the Subtropical Humid (SH) climate zone).

To further show this characteristic, the boxplot of skewness values for 40 datasets of different durations is demonstrated in Figure 2. 4. For example, the 75 percentile of skewness of the 15 minute duration was around 3.2 while that for the 30-day duration was 1.2. This is partly because for short duration like 15 minutes, a large amount of rainfall may occur within a short time in certain cases exhibiting large skewness while for long durations, like 30 days, the data is averaged and thus it exhibits less skewness.

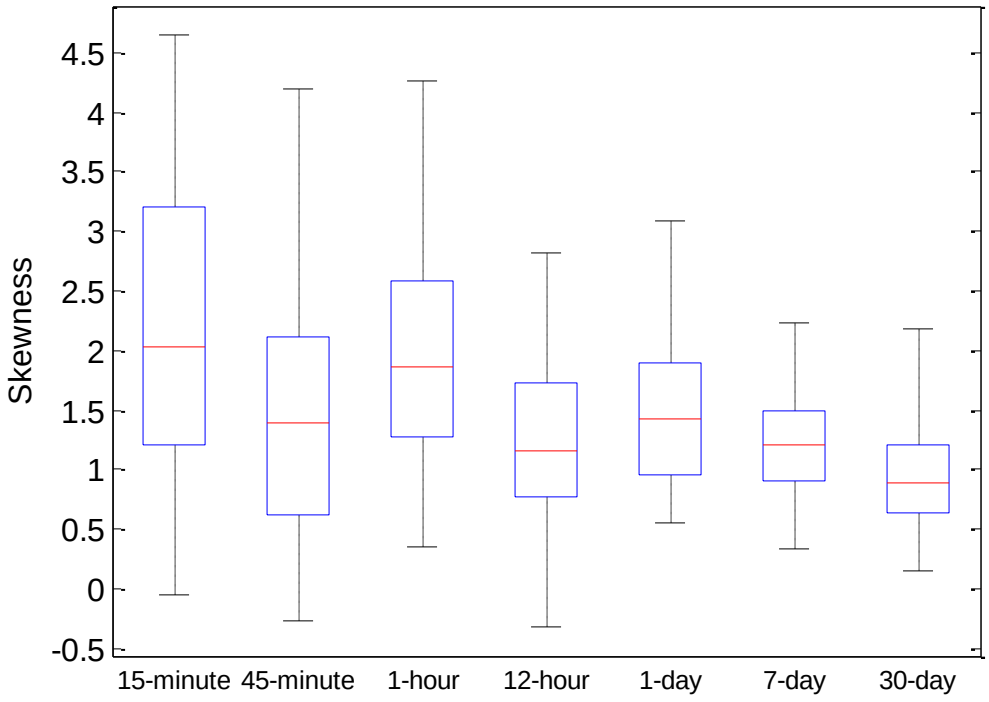

Figure 2. 4 Skewness of annual rainfall maxima of different durations (40 datasets for each duration).

2.2.4 Change in distribution form with climatic zone

Subtropical humid zone (SH)

The subtropical humid (SH) zone lies in the eastern part of Texas which is mostly noted for warm summers [*Larkin and Bomar*, 1983]. Ten stations were selected for the study. This zone includes most parts of Upper Coast and East Texas division. There are four rainfall generating mechanisms that exist in the Upper Coast area, leading to varying patterns from year to year as one or more of these controls change: in May the

typical thunderstorm pattern is expected slightly inland while the belt of maximum activity is along the coast by July; in September tropical disturbances can cause very heavy rains for some years, while in December frontal activity affects the region [*National Fibers Information Center*, 1987]. The East Texas division is characterized by a fairly uniform seasonal rainfall with slight maxima occurring in May and December and there is little variation in the weather in the summer season, because the influence of the Gulf of Mexico is dominant [*National Fibers Information Center*, 1987]. The most widespread and lengthy precipitation periods in East Texas during spring and autumn occur when the cold air forms a barrier, forcing the overriding moist Gulf air to be deflected upward where it cools and condenses [*Carr*, 1967].

For two stations 411956 and 410569, the histograms are shown in Figure 2. 5 (a) and (b) for 12-hour annual rainfall maxima. It can be seen that frequency distributions are smooth for the data of this duration. This region is along the coast and the rainfall pattern is affected by the Gulf of Mexico. Since the proximity to the coast is the most determining factor for regional climate differences [*North et al*., 1995], the reason for this frequency distribution pattern may be due to the moderating moisture from the Gulf of Mexico.

Subtropical sub-humid zone (SSH)

The subtropical sub-humid (SSH) zone is located in the central part of Texas which is characterized by hot summers and dry winters [*Larkin and Bomar*, 1983]. No clear pattern was discernible from the frequency distribution of several stations in this climate zone.

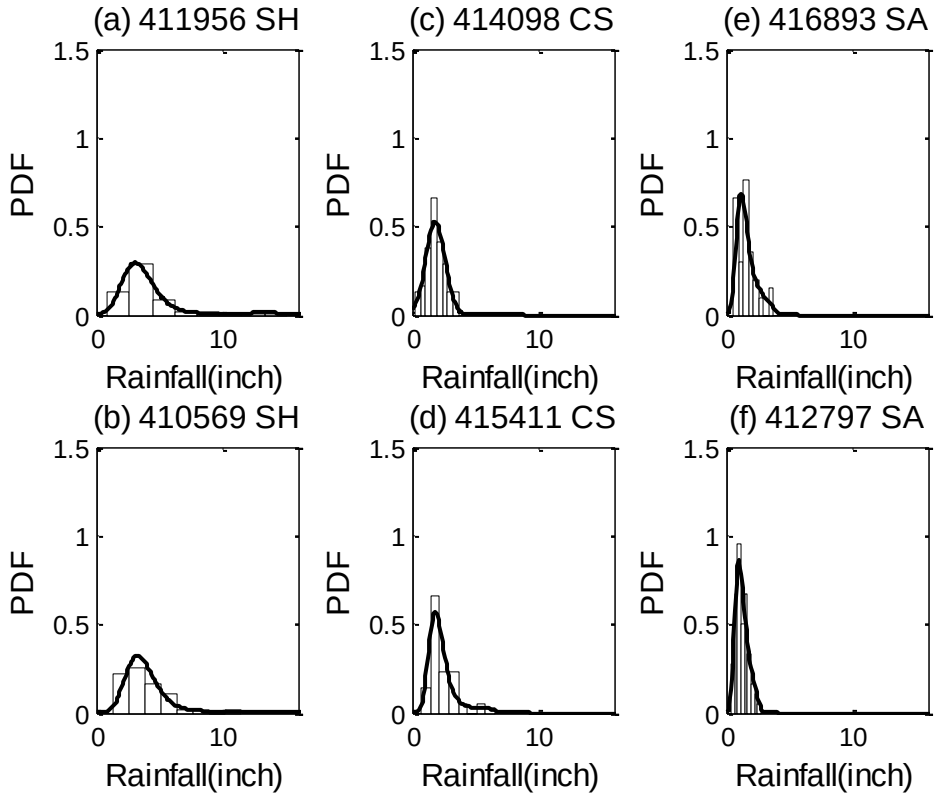

Figure 2. 5 Histograms and probability density functions of 12-hour rainfall data of different climate zones.

Continental steppe zone (CS)

The continental steppe (CS) zone lies in the northwestern part of Texas and includes the regions similar to the High Plain division. The rainfall amount increases steadily through spring and reaches a maximum in May or June, while the thunderstorm activity is also on the rise during the spring season [*National Fibers Information Center*, 1987]. In this region, summer is the wet season and thunderstorms are numerous in June and July but begin to decrease in August. Two stations 414098 and 415411 were used for analysis and the histograms for 12-hour annual rainfall maxima are shown in Figure

2. 5 (c) and (d). The frequency distributions in this part are relatively sharp, compared with those from the SH climate zone. The reason may be that the maximum rainfall mainly comes from the thunderstorms during the summer season.

Subtropical Arid zone (SA)

The subtropical arid zone lies in the extreme western part of Texas and includes the region similar to the Trans-Pecos division. The basin and plateau region of the Trans-Pecos features a subtropical arid climate, which is marked by summertime rainfall anomalies of the mountain relief [*Larkin and Bomar,* 1983]. Rainfall reaches its maximum in July and in summer, where the rain comes mainly from thunderstorms, often affected by local topography [*National Fibers Information Center,*1987]. In the Trans-Pecos region, the biggest percentage of rainfall occurring in this area is due to convective showers and thundershower activity, while the thundershower activity is the primary contributor of rainfall during late summer and early autumn months [*Carr,* 1967]. Two stations 416893 and 412797 were selected for analysis and the histograms for the 12-hour annual rainfall maxima are shown in Figure 2. 5 (e) and (f). The frequency distributions were relatively sharp compared with those from the SH climate zone. The reason for the variation of rainfall may be that the heavy rainfall in SA is mainly produced due to the convective shower and thundershower activity.

Subtropical steppe zone (SST)

From the mid-Rio Grande Valley to the Pecos Valley, the broad swath of Texas has a subtropical Steppe (SST) climate and is typified by semi-arid to arid conditions

[*Larkin and Bomar*, 1983]. No clear pattern of frequency distributions in this zone was found from the data of several stations.

In general, frequency distributions for regions in extremely northern and western parts (or the CS and SA climate zones) were sharp; however, those for the regions in the southeast near the Gulf Mexico (or the SH climate zones) were rather smooth. In general, frequency distributions became smoother from northwest to southeast. Although only a few of the possible mechanisms of rainfall in each region were investigated, the analysis provided an insight into the reason for the specific rainfall frequency distribution pattern in each climate region.

2.2.5 Influence of the distance from the sea (or the Gulf of Mexico)

The Gulf of Mexico is particularly important for the climate of Texas, as it provides the source of moisture and modulates the average seasonal and diurnal cycles, particularly in the coastal regions [*North et al.*, 1995]. In general, the average annual rainfall decreases with increasing distance from the Gulf of Mexico.

To assess the effect of the Gulf of Mexico on the distribution of annual rainfall maxima, 20 stations were selected and divided into two groups each with 10 stations according to the distance from the Gulf of Mexico. The histograms of 12-hour maximum rainfall for four sample stations are shown in Figure 2. 6. It can be seen that the frequency distributions in group II (more than 250 miles away from the Gulf) are not as smooth as those in group I (within 60 miles from the Gulf), which are located along the coast. The smoothness of frequency distributions in Group I is partly due to the closeness of rainfall stations to the Gulf of Mexico. The effect of Gulf of Mexico is

reduced with the distance and the topology factor may also play an important role for the rainfall generating mechanism. The frequency distribution pattern for the two stations in Group II may be due to the mixed effect of the Gulf of Mexico and topology.

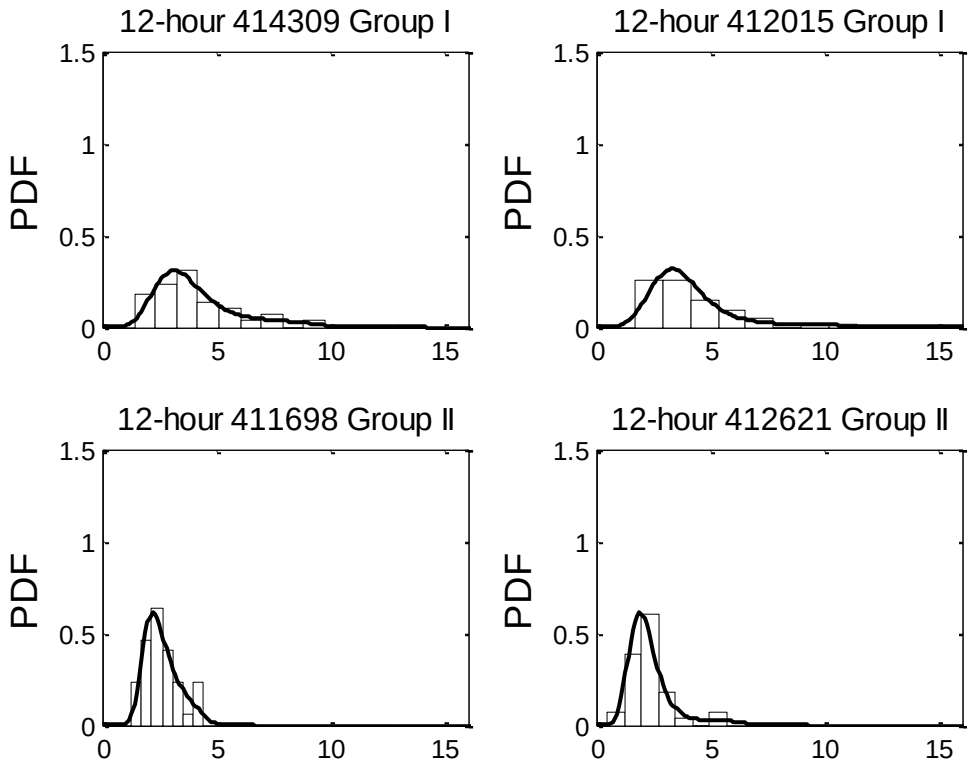

Figure 2. 6 Histograms and probability density functions of 12-hour rainfall data of different distances from the Gulf of Mexico (414309, 60 miles; 412015, 20 miles; 411698, 480 miles; 412621, 450 miles).

It is clear that the probability distribution varies with time duration, climate zone and distance from the sea (or Gulf of Mexico). The question arises if a probability

distribution that can accommodate the effect of these factors. This is addressed in what follows.

2.3 Annual maximum rainfall distribution using entropy theory

2.3.1 Derivation of distribution

Let the annual maximum rainfall for a given duration be represented as a continuous random variable $X \in [a, b]$ with a probability density function (PDF), $f(x)$. For $f(x)$, the Shannon entropy E can be defined as [*Shannon*, 1948; *Shannon and Weaver*, 1949]:

$$E = -\int_a^b f(x) \ln f(x)\, dx \tag{2.1}$$

where x is a value of random variable X with lower limit a and upper limit b. *Jaynes* [1957] developed the principle of maximum entropy (POME) which states that the probability density function should be selected among all the distributions with the maximum entropy subjected the given constraints. The constraints can be expressed in general form as:

$$\int_a^b g_r(x) f(x)\, dx = E(g_r) \quad r=1; 2,\ldots, m \tag{2.2}$$

where the function $g_r(x)$ in equation (2.2) is the known function with $g_0(x)=1$; $E(g_r)$ is the r-th expected value obtained from observations with $g_0=1$; m is the number of constraints.

The maximum entropy based probability density function can then be obtained by maximizing the entropy in equation (2.1), subject to equations (2.2) using the method of Lagrange multipliers, as [*Kesavan and Kapur*, 1992]:

$$f(x) = \exp[-\lambda_0 - \lambda_1 g_1(x) - \lambda_2 g_2(x) ... - \lambda_m g_m(x)] \quad (2.3)$$

where λ_r (r=0, 1,…, m) are the Lagrange multipliers.

2.3.2 Maximum entropy distribution with moments as constraints

Moments can be used for the reconstruction of density based on maximum entropy [*Mead and Papanicolaou*, 1984]. With the first four moments as constraints, the maximum entropy-based probability density function (denoted as ENT4) defined on the interval [a, b], with the function $g(x)$ in equation (2.2) expressed as $g_i(x)=x^i$ (i=1, 2, 3 and 4) , can be expressed as:

$$f(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3 - \lambda_4 x^4) \quad (2.4)$$

In this study, the lower bound of the interval a is set to be zero, while the upper bound b was set to be 20 times the observed maximum value. Since higher moments are involved in this distribution, a relatively large datasets would be needed for the accuracy of the moment estimation.

With the first four moments as constraints, the skewness, kurtosis and multiple modes can be included in the resulting maximum entropy-based distribution [*Zellner and Highfield*, 1988]. Each maximum of the polynomial inside the exponential corresponds to one mode and thus the multiple modes may exist in the maximum distribution [*Smith*, 1993]. *Matz* [1978] developed a new algorithm for the maximum likelihood estimate of this distribution and showed its good performance in characterizing features of empirical distributions, including the bi-modal. Comparing this distribution with the Pearson distribution, *Zellner and Highfield* [1988] showed that it was comparable with the

Pearson distribution while provided a better fit for small sample size, especially at the tails. *Smith* [1993] used the maximum entropy-based distribution with moments as constraints for decision analysis to construct the distribution of value lottery and showed the distribution with first four moments as constraints performed well.

In this study, the entropy based distribution in equation (2.4) was proposed as an alternative for modeling extreme rainfall values. In addition, the entropy distribution with the first three moments as constraint was also selected as the candidate for modeling extreme rainfall values. From equation (2.3), this distribution with three parameters (denoted as ENT3) can be expressed as:

$$f(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3) \tag{2.5}$$

2.3.3 Estimation of parameters

The Lagrange multipliers of equation (2.4) has to be determined using equations (2.2) where $E(g_r)(r=1, 2, 3, 4)$ are the expectation of the first four non-central moments. Generally the analytical solution does not exist and the numerical estimation of the Lagrange multipliers is needed. To that end, one can maximize the function [*Mead and Papanicolaou*, 1984; *Wu*, 2003]:

$$\Gamma = \lambda_0 + \sum_{r=1}^{4} \lambda_r \bar{g}_r = \ln \int_a^b \exp\left[-\sum_{r=1}^{4} \lambda_r g_r(x) \right] dx + \sum_{r=1}^{4} \lambda_r \bar{g}_r \tag{2.6}$$

The maximization can be achieved by employing Newton's method. Starting from some initial value $\boldsymbol{\lambda}_{(0)}$, one can solve for Lagrange parameters by updating $\boldsymbol{\lambda}_{(1)}$ through the equation given below:

$$\boldsymbol{\lambda}_{(1)} = \boldsymbol{\lambda}_{(0)} - H^{-1}\frac{\partial \Gamma}{\partial \lambda_i} \quad r=1, 2, 3, 4 \tag{2.7}$$

where the gradient Γ is expressed as:

$$\frac{\partial \Gamma}{\partial \lambda_i} = \overline{g_i} - \int_a^b \exp\left[-\sum_{r=0}^{4}\lambda_r g_r(x)\right] g_i(x)dx, \quad r=1, 2, 3, 4 \tag{2.8}$$

and H is the Hessian matrix whose elements are expressed as:

$$\begin{aligned} H_{i,j} &= \int_a^b \exp\left(-\sum_{r=0}^{4}\lambda_r g_r(x)\right) g_i(x) g_j(x)dx \\ &- \int_a^b \exp\left(-\sum_{r=0}^{4}\lambda_r g_r(x)\right) g_i(x)dx \bullet \int_a^b \exp\left(-\sum_{r=0}^{4}\lambda_r g_r(x)\right) g_j(x)dx, \quad i, j = 1, 2, 3, 4 \end{aligned} \tag{2.9}$$

2.4 Model evaluation

2.4.1 Performance measure

To quantify the performance of the proposed distribution in modeling the extreme rainfall quantiles, the root mean square error (RMSE) was used defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - o_i)^2} \tag{2.10}$$

where n is the length of the observed data; o_i are the observed quantile; x_i are the estimated quantile from the fitted distribution corresponding to the empirical non-exceedance probabilities estimated from the plotting position formula. In this study, the Gringorten plotting position formula is used defined as [*Gringorten*, 1963]:

$$P = \frac{i - 0.44}{n + 0.12} \tag{2.11}$$

where i is the rank of the observed values and n is the length of the observed data.

2.4.2 Synthetic data from known distribution

Monte Carlo experiments were first carried out to compare the quantiles estimated from the GEV, ENT4 and ENT3 distributions. Two Monte Carlo simulations were conducted with random numbers generated from the known GEV and lognormal distributions. Random numbers of three different lengths (namely, 40, 70 and 100) were generated, which were used to approximate the record length of the 15-minutes, hourly and daily rainfall data in this study. For the first simulation (S_1), the quantiles corresponding to different return periods (T = 5, 10, 25, 50, 100, 200 years) were first assessed with the random number generated from the GEV distribution. For the second simulation (S_2), the quantiles corresponding to relatively long return period (T=100 and 200 years) from the three distributions were assessed with the synthetic data generated from log-normal distribution with different skewness values.

Random number from Generalized Extreme Value distribution (GEV)

The generalized extreme value (GEV) distribution has been applied extensively in hydrology for extreme rainfall analysis. Its probability density function is defined as:

$$f(x;\mu,\sigma,k)=\begin{cases}\frac{1}{\sigma}\left[1+k(\frac{x-u}{\sigma})\right]^{-1/k-1}\exp\left\{-\left[1+k(\frac{x-u}{\sigma})\right]^{-1/k}\right\}, k\neq 0\\ \frac{1}{\sigma}\exp\left[-\frac{x-u}{\sigma}-\exp(-\frac{x-u}{\sigma})\right], k=0\end{cases} \quad (2.12)$$

where k, σ and u are the shape, scale and location parameter. In this study, the MATLAB function *gevfit* was used for the parameter estimation of the GEV distribution with maximum likelihood method.

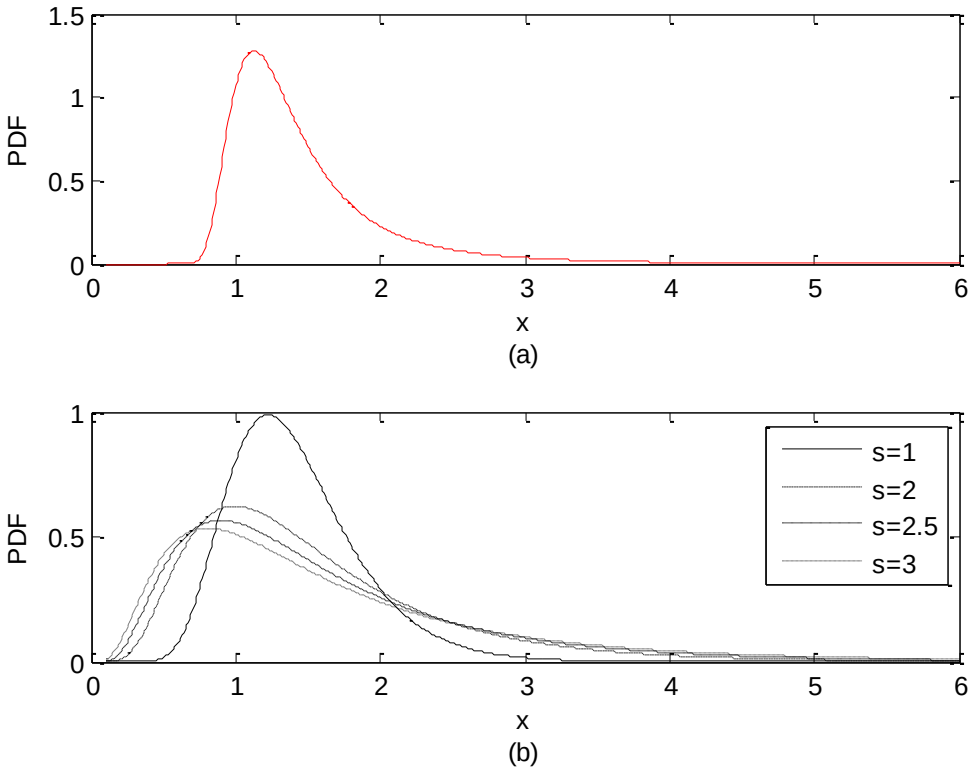

Figure 2. 7 Parent distributions for Monte Carlo simulation. (a) GEV distribution; (b) Lognormal distribution with different skewness (*s*).

1000 datasets of random numbers with different sample sizes (n=40, 70, 100) were generated from this parent distribution. The GEV, ENT4 and ENT3 distributions were then fitted to these datasets and the quantiles corresponding to different return periods were obtained. Parameters (k, σ , u) of the parent distribution were assigned as (0.3,0.3,1.2) and the probability density function is shown in Figure 2. 7

The median and the RMSE values of the estimated quantiles for simulation S_1 are shown in Table 2. 1. From the median values, it can be seen that for short return periods ($T \leq 50$ years), the median values from the ENT4 and GEV distributions were close to

each other for each sample size. For example, for sample size *n*=100, the median values from GEV and ENT4 for return period 50 years were 3.42 and 3.40, respectively, while the observed value was 3.42.

Table 2. 1 Median of estimated quantiles with random numbers generated from the GEV distribution.

Sample size	Return Period (years)	5	10	25	50	100	200
	Observation	1.77	2.16	2.81	3.42	4.18	5.10
n=40	GEV	1.75	2.14	2.76	3.35	4.11	5.04
	ENT4	1.74	2.09	2.80	3.36	3.57	3.69
	ENT3	1.97	2.22	2.49	2.67	2.83	2.98
n=70	GEV	1.76	2.16	2.78	3.37	4.09	5.03
	ENT4	1.74	2.08	2.77	3.42	3.86	4.05
	ENT3	1.99	2.25	2.54	2.72	2.89	3.04
n=100	GEV	1.76	2.15	2.80	3.42	4.16	5.06
	ENT4	1.75	2.07	2.71	3.40	4.21	4.42
	ENT3	2.02	2.30	2.60	2.79	2.96	3.12

The RMSE values of the estimated quantiles for simulation S_1 are shown in Table 2. 2. Generally the RMSE values of the ENT4 distribution were slightly larger than those of the GEV distribution, however, these results were acceptable. For the quantiles corresponding to the relatively long return periods (100 and 200 years), the median quantile from the ENT4 distribution is slightly underestimated, while that from the GEV distribution was close to the true value. This is not unexpected, since the random numbers were generated from the GEV distribution and then the GEV distribution was fitted. Generally ENT4 modeled the data generated from the GEV distribution well,

especially when the sample size was relatively large. The ENT3 distribution also estimated the quantiles relatively well for short periods ($T \leq 25$ years), while it did not model the quantiles well corresponding to relatively long return periods ($T \geq 50$ years).

Table 2. 2 RMSE of estimated quantiles with random numbers generated from the GEV distribution.

Sample size	Distribution	Return Period (years)					
		5	10	25	50	100	200
n=40	GEV	0.14	0.25	0.55	0.96	1.64	2.73
	ENT4	0.15	0.28	0.77	1.70	1.74	2.00
	ENT3	0.38	0.47	0.66	0.95	1.43	2.12
n=70	GEV	0.10	0.19	0.40	0.69	1.15	1.86
	ENT4	0.12	0.22	0.52	1.02	2.28	2.38
	ENT3	0.35	0.41	0.59	0.91	1.45	2.36
n=100	GEV	0.09	0.16	0.34	0.57	0.93	1.46
	ENT4	0.11	0.18	0.42	0.97	2.67	2.74
	ENT3	0.36	0.44	0.62	0.90	1.37	2.05

Random number from log-normal distribution

The probability density function of the log-normal distribution can be expressed as:

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e - \frac{(\ln x - u)^2}{2\sigma^2} \tag{2.13}$$

where u is the mean in the log-scale and σ^2 is the variance in the real scale. The skewness coefficient s is related with the variance σ^2 as $s=[\exp(\sigma^2)+2][\exp(\sigma^2-2)]^{0.5}$. In

this study, the MATLAB function *lognfit* was used for the parameter estimation of the log-normal distribution with maximum likelihood method.

Table 2. 3 Median of estimated quantiles with random numbers generated from the log-normal distribution with different skewness (*k*).

Sample Size	Skewness	*k*=1		*k*=2		*k*=2.5		*k*=3	
	Quantile	x_{100}	x_{200}	x_{100}	x_{200}	x_{100}	x_{200}	x_{100}	x_{200}
	Observation	2.80	3.03	4.87	5.59	5.99	7.03	7.13	8.53
n=40	GEV	2.73	2.95	5.12	6.02	6.48	7.94	7.84	10.00
	ENT	2.61	2.72	4.34	4.54	5.15	5.39	5.93	6.21
	ENT3	2.44	2.55	3.85	4.11	4.62	4.97	5.35	5.81
n=70	GEV	2.77	2.99	5.07	5.98	6.57	8.09	8.25	10.57
	ENT	2.69	2.83	4.60	4.88	5.58	5.91	6.83	7.26
	ENT3	2.46	2.58	3.88	4.14	4.75	5.12	5.78	6.30
n=100	GEV	2.80	3.01	5.13	6.08	6.60	8.14	8.34	10.66
	ENT	2.74	2.88	4.77	5.05	5.92	6.30	7.19	7.70
	ENT3	2.48	2.59	3.90	4.17	4.81	5.21	5.85	6.42

1000 datasets of random numbers with different sample sizes (*n*=40, 70 and 100) with different skewness 1, 2 , 2.5 and 3 were generated from log-normal distribution and used for comparison. Parameter *u* is assigned as 0.3 while the standard deviations corresponding to different skeweness values were assigned as 0.31, 0.55, 0.64 and 0.72, respectively. The PDFs for the parent distributions with these parameters are shown in Figure 2. 7 (b). The objective of this simulation was to show the performance of these distributions in modeling data with different skewness. The median and RMSE values of the estimated quantiles for return period 100 and 200 years (denoted as x_{100} and x_{200}) corresponding to non-exceedance probability 0.99 and 0.995 are shown in Table 2. 3 and Table 2. 4.

Table 2. 4 RMSE of estimated quantiles with random numbers generated from the log-normal distribution with different skewness (*k*).

Sample Size	Distribution	*k*=1		*k*=2		*k*=2.5		*k*=3	
		x_{100}	x_{200}	x_{100}	x_{200}	x_{100}	x_{200}	x_{100}	x_{200}
n=40	GEV	0.45	0.61	1.80	2.82	2.65	4.37	3.95	6.83
	ENT	0.43	0.51	1.41	1.64	1.92	2.30	2.62	3.14
	ENT3	0.41	0.52	1.23	1.64	1.71	2.34	2.43	3.28
n=70	GEV	0.32	0.43	1.18	1.79	1.92	3.06	3.05	5.13
	ENT	0.36	0.45	1.26	1.44	1.95	2.22	2.75	3.00
	ENT3	0.36	0.47	1.09	1.51	1.52	2.12	1.93	2.77
n=100	GEV	0.25	0.33	0.95	1.43	1.63	2.60	2.41	4.02
	ENT	0.30	0.39	1.19	1.38	1.69	1.96	2.70	3.03
	ENT3	0.35	0.46	1.03	1.45	1.36	1.95	1.69	2.57

For the case with skewness *k*=1, the median quantiles from the ENT4 distribution was not as close to the observed values as from the GEV distribution. However, the difference of the estimated median from GEV and ENT4 was relatively small, especially for relatively large sample sizes. For example, for *n*=100, the median values from GEV and ENT4 were 2.80 and 2.74 with the observed value being 2.80. Generally the RMSE values of the two distributions were close to each other. For example, the RMSE of GEV and ENT4 for x_{200} were 0.43 and 0.45, respectively, for *n*=70. The performance of ENT4 is improved with the increase of sample size. Generally the performance of ENT4 and GEV were comparable in this case.

For skewness values of *k*=2 and 2.5, the median values from GEV were overestimated while those from ENT4 were slightly underestimated. When the sample size was relatively small (*n*=40), the GEV distribution performs slightly better than the ENT4 distribution for the median values. However, the RMSE value from the GEV is

higher than the ENT4 distribution. When the sample size was relatively large (n=100), the ENT4 distribution generally performed better than the GEV distribution for the median value, while their performance was comparable for the RMSE values. For example, for the case with k=2.5 and sample size n=100, the median values from GEV and ENT4 corresponding to the 100 year return period were 6.60 and 5.92 while the true value was 5.99. The corresponding RMSE values for GEV and ENT4 were, respectively, 1.63 and 1.69, which are comparable. The performance of the ENT4 distribution improved with the increase of sample size.

For the skewness k=3, the median value estimated from GEV was overestimated significantly, while ENT4 still performed relatively well for estimating quantiles, especially when the sample size was relatively large. For example, the true quantile corresponding to the 100 year return period was 7.13, while the quantiles from GEV and ENT4 with sample size (n=70) were 8.25 and 6.83, respectively. The corresponding RMSE values were 3.05 and 2.75, indicating that ENT4 performed relatively better.

Though the RMSE values from ENT3 distribution was comparable with the ENT4 distribution and sometimes even smaller than ENT4 distribution, generally the median value from ENT3 was underestimated significantly for each sample size with different skewness. These results showed that generally ENT3 did not perform as well as the GEV and ENT4 distributions and was not suitable for modeling extreme values.

Summary

The Month Carlo simulation S_1 showed that generally the ENT4 distribution was comparable to the GEV distribution in modeling extreme rainfall values. Since the GEV

distribution has been extensively applied for modeling extreme values, the results from the first simulation S_1 showed that the ENT4 distribution would also be a candidate for modeling the extreme values. The Monte Carlo simulation S_2 showed that the performance of the ENT4 distribution was comparable with GEV for low skewness , especially when the sample sizes were relatively large ($n \geq 70$). When the skewness was relatively high ($k \geq 2$), the ENT4 distribution performed relatively better than the GEV distribution for estimating quantiles corresponding to relatively long return periods, especially when the sample size was large. *Botero and Francés* [2010] also found that the GEV distribution led to large errors for quantile estimation corresponding to long return periods for high skewness.

Synthetic data from other distributions (e.g., gamma distribution) were also used for comparison and generally similar results were obtained (not presented). Thus it can be concluded from the Monte Carlo simulation that generally the ENT4 distribution provided an alternative to the commonly used GEV distribution and should be preferable for observations with higher skewness. The ENT3 distribution was not suitable for modeling extreme values.

2.4.3 Real rainfall data from observation

To further compare the performance of the GEV distribution and ENT4 distribution, the observed rainfall data from 40 stations for different time duration (15-min, 45-min, 1-hour, 12-hour, 1-day, 7-day and 30-day) were also used. The two distributions were compared based on empirical and theoretical quantiles according to the RMSE measure. The number of stations for each distribution performing the best

(with the least RMSE) is shown in Table 2. 5. For all durations, the ENT4 distribution performed the best for the largest number of stations. For example, for the annual rainfall maxima of the 12-hour duration, the ENT4 distribution performed the best for 36 stations according to RMSE. From these results, it can be seen that the ENT4 distribution would be a good candidate for modeling annual rainfall maxima.

Table 2. 5 Number of stations with the minimum RMSE from each distribution.

Duration	ENT	GEV	ENT3
15-minute	33	7	0
45-minute	36	3	1
1-hour	36	4	0
12-hour	36	4	0
1-day	33	7	0
7-day	32	8	0
30-day	32	8	0

2.5 Application of the entropy based distribution

The entropy-based distribution was used to fit the rainfall data in section 2.2, as shown in Figure 2. 3, Figure 2. 5 and Figure 2. 6, together with the empirical histograms as shown in the previous section. These figures show that the entropy-based distribution (ENT4) fitted the empirical histograms well for the rainfall data of different durations, climate zones and different distances from the Gulf.

The GEV distribution was also applied here for further comparison with the ENT4 distribution. For each duration (15-min, 45-min, 1-hour, 12-hour, 1-day, 7-day and 30-day), 10 stations were used in each climate zone (except that for the SA climate

zone, 6 stations were used for 15-minute and 45-minute duration due to data limitation). The number of stations that ENT4 or GEV performed better in different climate zones is shown in Table 2. 6. Taking the result in the CS climate zone as an example, the ENT4 distribution performed better for all durations for at least 8 out of 10 datasets.

Table 2. 6 Number of stations with the minimum RMSE for different climate zones and durations.

Duration	CS		SA[a]		SH	
	ENT	GEV	ENT	GEV	ENT	GEV
15-minute	9	1	4	2	10	0
45-minutes	10	0	6	0	9	1
1-hour	9	1	9	1	10	0
12-hour	9	1	10	0	8	2
1-day	8	2	7	3	9	1
7-day	8	2	8	2	7	3
30-day	8	2	8	2	7	3

[a] For SA climate region of 15 and 45-minute data, only 6 stations are selected due to data limitation

The ENT4 distribution was also compared with the GEV distribution for different distances from the sea (Group I and Group II) with 10 stations in each group. There were not enough stations with a relatively long record of 15 minutes data in Group I and thus only the hourly (1-hour and 12-hour) and daily data (1-day, 7-day and 30-day) were used for comparison. The number of cases that ENT4 performed better than GEV for the two groups is shown in Table 2. 7. It can be seen that generally the ENT4 distribution performed better than the GEV distribution. Taking the 1 hour data as an

example, the ENT4 distribution had less RMSE for 10 and 8 cases for Groups I and II, respectively.

Table 2. 7 Number of stations with the minimum RMSE for different distances from the sea and durations.

Duration	Group I	Group II
1-hour	10	8
12-hour	5	9
1-day	9	9
7-day	9	9
30-day	10	9

The annual maximum rainfall distribution can then be employed for the construction of intensity-duration-frequency (IDF) curves [*Singh*, 1992], which is defined as a relationship of rainfall intensity occurring over a certain duration d with different return periods. The hourly annual rainfall data for station 418583 were used to construct the IDF curves, as shown in Figure 2. 8. The empirical return period (TE) was obtained from the Gringorten plotting position formula as TE=1/(1-*P*), where *P* is the nonexceedance probability. The empirical return period were also plotted on the IDF curves. Note that the accuracy of the empirical return period for the highest-ranked peak flows is limited [*Stedinger*, 1993; *Beckers and Alila*, 2004]. Generally the return period from the IDF curves fitted the empirical return period well. For example, for the return period 20 years, the theoretical rainfall quantile from the ENT4 distribution was 4.6 inch while the observed quantile was 4.8 inch.

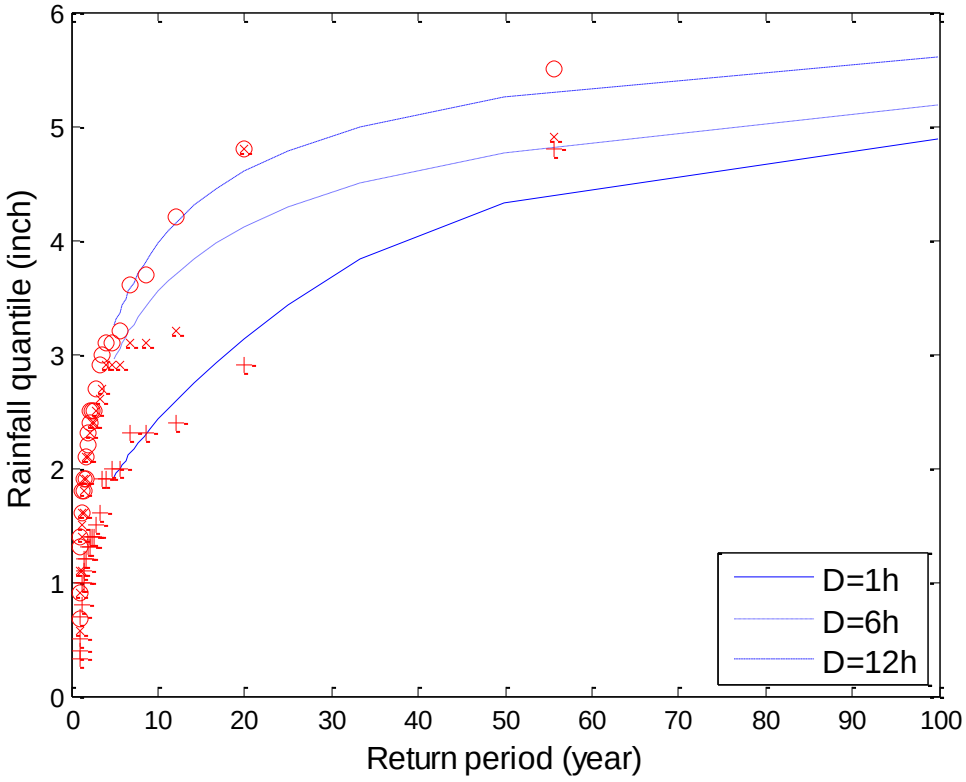

Figure 2. 8 IDF curves for different durations (for station 418583).

2.6 Conclusion

Frequency characteristics of annual rainfall maxima from different stations in Texas are analyzed. Results show that frequency distributions of annual rainfall maxima are highly skewed for short durations, like 15 min, and tend to be smoothed when the duration is relatively long. The distributions also show different patterns across different regions. In northern and western parts, like the CS and SA climate zones, distributions are sharp; however, they are relatively smooth in the southeast, like the SH climate zone. The possible reason is that in the CS and SA climate zones, heavy rainfall is mainly

produced by thunderstorms, while in the SH climate zone, the moisture from the Gulf of Mexico is the moderating factor. For the other climate zones, no clear pattern is found, which may be due to the mixed effect of different rainfall mechanisms. The frequency distribution of rainfall near the Gulf of Mexico is smoother than that far away from the Gulf. The reason may be that the Gulf of Mexico serves as the moisture source.

An entropy based distribution is proposed for frequency analysis of annual rainfall maxima. Monte Carlo simulation based on the synthetic data from different distributions shows that the ENT4 distribution is comparable with the GEV distribution and is preferable for the datasets with high skewness. Furthermore, the ENT4 distribution performs better for most cases than the GEV distribution in the general performance of modeling the quantiles based on the observed rainfall data. These results from the synthetic data and real observations show that the ENT4 distribution is a good candidate to model the annual rainfall maxima of different time scales across Texas.

The ENT4 distribution is applied to the frequency distribution of annual rainfall maxima of different durations, climate zones and distances from the sea, and results show that the ENT4 distribution fits the empirical densities well. Further comparison between the ENT4 and GEV distributions shows that ENT4 performs better than GEV for different durations, climate zones and distances from the sea though the distribution pattern changes. Application of the proposed method for rainfall analysis is illustrated with the construction of IDF curves based on rainfall data of one sample station. Analysis of the changing patterns of rainfall distribution with time duration, climate zone

and distance from the Gulf of Mexico sheds some light on the analysis of rainfall of different durations in Texas.

CHAPTER III

ENTROPY BASED METHOD FOR SINGLE-SITE MONTHLY STREAMFLOW SIMULATION*

3.1 Introduction

Streamflow simulation plays an important role in water resources planning and management. The key requirement for streamflow simulation is that synthetic streamflow sequences preserve key statistical properties of the historical record, such as mean, standard deviation, skewness, and lag correlations. A number of models for streamflow simulation have been proposed and these models can be classified into two groups: parametric and non-parametric.

A commonly used parametric model for synthetic streamflow generation is the autoregressive moving average (ARMA) model [*Lettenmaier and Burges*, 1977; *Hipel and McLeod*, 1978; *Hipel et al*., 1979; *Salas and Delleur*, 1980; *Loucks et al.*, 1981; *Vogel and Stedinger*, 1988; *Savic et al*., 1989], which is quite flexible and can be used for annual as well as seasonal streamflow simulation. The ARMA model is based on the Gaussian assumption which is not usually satisfied by streamflow data. An alternative to the ARMA model for simulating seasonal streamflow is the disaggregation model which has been widely applied [*Valencia and Schaake*, 1973; *Mejia and Rousselle*, 1976]. For

---

*Reprinted with permission from "Single-site monthly streamflow simulation using entropy theory" by Hao, Z. and V. P. Singh (2011), *Water Resources Research, 47*, W09528, doi:10.1029/2010WR010208, Copyright [2011] by American Geophysical Union.

the disaggregation model, annual or aggregated streamflow is generated with an appropriate model and then the generated streamflow is disaggregated to obtain monthly or seasonal streamflow. The disaggregation model ensures the sum of low time scale streamflow values (e.g., monthly) adds up to high time scale streamflow values (e.g., yearly), but has many parameters that need to be estimated. To reduce the number of parameters, several parsimonious models have been proposed, such as condensed disaggregation model [*Stedinger et al*., 1985; *Grygier and Stedinger*, 1988] and stepwise disaggregation model [*Santos and Salas*, 1992]. *Koutsoyiannis and Maneta* [1996] proposed a simple disaggregation model that combines models of lower scale (e.g., monthly) and higher scale (e.g., yearly) with the accurate adjusting procedure.

Parametric models generally require the assumption regarding the marginal distribution of underlying streamflow data. However, the Gaussian assumption usually made may not hold in reality. Therefore, transformation techniques to render the data to be normal are often applied, which in turn give rise to several potential drawbacks. First, some bias of the statistical properties in the original domain may be caused when data is simulated in the transformed domain. Second, negative values may be generated. Third, non-Gaussian features, such as skewness and bimodal, cannot be captured and reproduced efficiently [*Prairie et a*l., 2006]. The autoregressive model with gamma distribution has been proposed to avoid the data transformation [*Fernandez and Salas*, 1990] , though the bimodal property cannot be reproduced. Furthermore, it is hard for a usual parametric model to capture the nonlinear relationships that may be observed in the historical record [*Salas and Lee*, 2010].

An attractive alternative is nonparametric models and *Lall* [1995] provided a review of the application of non-parametric models in hydrology. Nonparametric models are often based on bootstrap techniques or kernel density estimation and they avoid model selection, minimize (or avoid) parameter estimation, and do not make any assumption about the probability distribution. *Lall and Sharma* [1996] proposed a nearest neighbor bootstrap method for re-sampling monthly streamflow, while probabilistically preserving the dependence structure. To reproduce the serial correlation of historical data, *Vogel and Shallcross* [1996] suggested the moving block bootstrap (MBB) by resampling the observed time series in approximately independent blocks, and compared the method with parametric methods for generating annual streamflow series. *Sharma et al.* [1997] proposed a nonparametric method for monthly streamflow simulation applying the conditional density function with Gaussian kernel, and *Sharma and O'Neill* [2002] extended that method to impose a long-term dependence in the simulated streamflow by incorporating an aggregated variable (denoted as NPL model). *Salas and Lee* [2010] developed a nonparametric method using the K-Nearest Neighbor (KNN) resampling technique with gamma kernel perturbation that can generate data different from the historical record for single site seasonal streamflow simulation. For this method, two approaches, one with the aggregate variable (denoted as KGKA model) and another with the pilot variable (denoted as the KGKP model), were developed to preserve the annual variability.

Nonparametric methods have also been applied for seasonal streamflow simulation with disaggregation approach. *Tarboton et al.* [1998] developed a

nonparametric disaggregation model for simulation based on the conditional distribution obtained by a kernel density estimation method. To address the issue of inefficiency of kernel density estimation method in higher dimensions, *Prairie et al.* [2007] applied a fast KNN based bootstrap approach to construct and simulate from the conditional distribution. *Lee et al*. [2010] proposed a space-time disaggregation model based on KNN coupled with a genetic algorithm that can overcome the shortcomings of the models proposed by *Prairie et al.* [2007] and *Koutsoyiannis and Manetas* [1996]. Based on KNN re-sampling, *Nowak et al*. [2010] proposed a space-time disaggregation algorithm for disaggregating annual flow to daily flows at different sites.

To simulate streamflow, an assumption about the marginal distribution is often made, especially for parametric models. However, many streamflow records cannot be characterized by commonly assumed probability distributions [*Sharma and O'Neill,* 2002]. The ability to preserve the cross boundary relation (e.g., the correlation between the last season of the previous year and the first season of the current year) and the generation of negative values are two issues that emerge for both parametric and nonparametric models [*Lee et al.,* 2010]. To address the first issue, *Mejia and Rousselle* [1976] made a modification to link past and present values being disaggregated. A practical way to address this problem is to start the generation from a season where the correlation is small. However, this does not work when all correlations between seasons are high. The issue of negative values arises due to the use of normal transformation in parametric models and the application of the Gaussian kernel in nonparametric models.

Generally, negative values generated during simulation can be disregarded. However, this solution may not be appropriate when too many negative values are generated.

This study proposes a new model for simulating monthly streamflow at a single site which is capable of preserving key statistics, such as mean, standard deviation, skewness and lag-one correlation. The model is based on entropy theory, wherein a probability distribution function (PDF) is derived without the assumption of normality or the use of a normal transformation. Moreover, the model can preserve the cross-correlation and avoids generation of negative values. It can also be extended to incorporate higher-order moments and more lag correlations if needed. With the specified statistical properties, such as mean, standard deviation, skewness, and lag-one correlation as constraints, the joint probability density function of streamflow of two adjacent months is constructed by maximizing entropy, and the conditional density function is derived from the joint PDF, from which streamflow can be generated .

The paper is organized as follows. Describing the framework of the method in section 3.2, the proposed method is tested using a synthetic example with known underlying model in section 3.3, followed by an application to the Colorado River basin for streamflow simulation in section 3.4. Conclusions along with a summary of the main features of the proposed method are given in section 3.5.

3.2 Method

The first step in the streamflow simulation is the derivation of joint and conditional probability density functions of streamflow. The derivation involves the expression of the joint Shannon entropy, specification of constraints based on the

statistics to be preserved, maximization of the entropy subject to the specified constraints, and determination of the Lagrange multipliers. Then the monthly streamflow is simulated from the conditional distribution sequentially.

3.2.1 Shannon entropy

For a bivariate case involving two continuous random variables X and Y or random vector (X , Y) with joint probability density function $f(x, y)$ defined over the space $[a, b] \times [c, d]$, the Shannon entropy can be defined as:

$$E = -\int_c^d \int_a^b f(x,y) \ln f(x,y)\, dxdy \tag{3.1}$$

3.2.2 Specification of constraints

For streamflow simulation, it is desired to preserve such statistics as mean, standard deviation, skewness and lag-one correlation. These statistics can be regarded as constraints for deriving the distribution of streamflow. Then, sampling from the distribution can be expected to preserve these required statistics. The mean, standard deviation, and skewness of streamflows can be determined through the first three moments. In order to preserve the correlation between streamflows of two adjacent months (say, January and February), the joint PDF of the continuous random vector (X, Y) is needed for which constraints in general form can be stated as:

$$\int_c^d \int_a^b g_i(x,y) f(x,y) dxdy = E(g_i) \quad i=1, 2,\ldots, m \tag{3.2}$$

Specifically,

$$\int_c^d \int_a^b f(x,y) dxdy = 1 \quad i=0 \tag{3.3}$$

$$\int_c^d \int_a^b x^i f(x,y)dxdy = E(x^i) \quad i=1, 2, 3 \tag{3.4}$$

$$\int_c^d \int_a^b y^{i-3} f(x,y)dxdy = E(y^{i-3}) \quad i=4, 5, 6 \tag{3.5}$$

$$\int_c^d \int_a^b xyf(x,y)dxdy = E(xy), \quad i=7 \tag{3.6}$$

where x and y are streamflow values of adjacent months; $g_i(x, y)$ (or g_i) is a known function of random vector (X, Y), which can be specified as $g_0=1$, $g_1=x$, $g_2=x^2$, $g_3=x^3$, $g_4= y$, $g_5=y^2$, $g_6= y^3$ and $g_7=xy$ for the proposed constraints; $E(g_i)$ is the expected value of the function $g_i(x, y)$ (e.g., if $g_1(x, y)=x$, then $E(x)$ is the mean of X) estimated from the historical record; $E(x)$ and $E(y^{i-3})$ are the first to third non-central moments of random variables X and Y, respectively; E(xy) is the expectation of XY and m is the number of constraints (m=7 in this case). The constraint in equation (3.3) assures that the integration of the probability density function over the whole interval should be unity, which is often termed as the "normalization condition" or the "total probability theorem."

3.2.3 Maximization of entropy and derivation of probability distributions

According to the principle of maximum entropy , formulated by *Jaynes* [1957], the least biased probability distribution will be the one that maximizes the Shannon entropy subject to the given constraints. To derive the joint PDF of streamflows of two adjacent months (say, January and February), the entropy given by equation (3.1) is maximized, subject to the constraints given by equation (3.3)-(3.6). The maximization can be performed using the method of Lagrange multipliers.

Denoting the Lagrange multipliers for the joint PDF of January and February streamflows as $\Phi_{1,2}=[\lambda_0, \lambda_1,\ldots, \lambda_7]$, where $\lambda_0, \lambda_1,\ldots, \lambda_7$ are the Lagrange multipliers, the Lagrangian function L, using equation (3.2) can be expressed as [*Kapur*, 1989]:

$$L = -\int_c^d \int_a^b f(x,y) \ln f(x,y) dxdy - (\lambda_0 - 1)(\int_c^d \int_a^b f(x,y) dxdy - 1) \\ - \sum_{i=1}^{m} \lambda_i \left[\int_c^d \int_a^b g_i(x,y) f(x,y) dxdy - E(g_i) \right] \tag{3.7}$$

Differentiating L with respect to f and setting the derivative to zero, the maximum entropy-based joint probability density function is obtained with representation of $g_i(x, y)$ by their specific values as [*Kesavan and Kapur*, 1992]:

$$f(x,y) = \exp\left[-\sum_{i=0}^{m} \lambda_i g_i(x,y) \right] = \exp(-\lambda_0 - \sum_{i=1}^{3} \lambda_i x^i - \sum_{i=4}^{6} \lambda_i y^{i-3} - \lambda_7 xy) \tag{3.8}$$

Substituting equation (3.8) in the "normalization condition" in equation (3.3), one can obtain the zeroth Lagrangre multiplier λ_0 as a function of other Lagrange multipliers as:

$$\exp(\lambda_0) = \int_c^d \int_a^b \exp\left[-\sum_{i=1}^{m} \lambda_i g_i(x,y) \right] dxdy \tag{3.9}$$

The joint PDF given by equation (3.8) has unknown Lagrange multipliers, λ_i (i= 1,…, 7), that need to be determined.

For monthly streamflow simulation, 12 joint density functions with random vector ($X_{t,n}$, $Y_{t,n}$) have to be estimated from the historical data, where t is the year and n (n=1, 2, …, 12) is the month. For the joint distribution of December and January streamflows, the random vector has to be replaced by ($X_{t\text{-}1,12}$, $Y_{t,1}$) to preserve the cross-

correlation between streamflow in December of the previous year ($X_{t\text{-}1,12}$) and that in January of the current year ($Y_{t,1}$).

The marginal density function for X can be obtained by integrating the joint PDF $f(x, y)$ given by equation (3.8) over Y as:

$$\begin{aligned} f(x) &= \int_c^d f(x,y)dy \\ &= \exp(-\lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3 - \lambda_0)\int_c^d \exp(-\lambda_4 y - \lambda_5 y^2 - \lambda_6 y^3 - \lambda_7 xy)dy \end{aligned} \tag{3.10}$$

The conditional density function of Y given $X=x$ can now be obtained as:

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{\exp(-\lambda_4 y - \lambda_5 y^2 - \lambda_6 y^3 - \lambda_7 xy)}{\int_c^d \exp(-\lambda_4 y - \lambda_5 y^2 - \lambda_6 y^3 - \lambda_7 xy)dy} \tag{3.11}$$

The conditional cumulative distribution function $F_{Y|X}(y|x)$ of Y given $X=x$ can be written as:

$$F_{Y|X}(y|x) = \int_c^y f(z|x)dz \tag{3.12}$$

3.2.4 Parameter estimation

The Lagrange multipliers contained in equation (3.8) are now determined. Substitution of equation (3.8) in equation (3.2) results in a set of nonlinear equations whose solution results in the Lagrange multipliers:

$$\int_c^d \int_a^b g_i(x,y)\exp\left[-\sum_{k=0}^{m}\lambda_k g_k(x,y)\right]dxdy = E(g_i) \;\; i=1, 2,\ldots, m \tag{3.13}$$

In general, an analytical solution for obtaining the Lagrange multipliers (for $m >$ 2) does not exist and numerical solution is the only resort. It has been shown that the

problem of solving the set of nonlinear equations is equivalent to finding the minimum of a convex function Γ expressed as [*Mead and Papanicolaou*, 1984; *Kapur*, 1989]:

$$\Gamma = \lambda_0 + \sum_{i=1}^{m} \lambda_i E(g_i) = \ln \int_c^d \int_a^b \exp\left[- \sum_{i=1}^{m} \lambda_i g_i(x,y) \right] dxdy + \sum_{i=1}^{m} \lambda_i E(g_i) \;\; i=1,2,\ldots, m \quad (3.14)$$

The Newton-Raphson method can be applied to achieve the minimization of the convex function yielding the Lagrange multipliers $\boldsymbol{\lambda}=[\lambda_0, \lambda_1,\ldots, \lambda_7]'$ as follows. Starting from some initial value $\boldsymbol{\lambda}_{(0)}$, one updates $\boldsymbol{\lambda}_{(1)}$ using the equation:

$$\boldsymbol{\lambda}_{(1)} = \boldsymbol{\lambda}_{(0)} - H^{-1} \frac{\partial \Gamma}{\partial \lambda_i} \quad i=1, 2,\ldots, 7 \quad (3.15)$$

where the gradient of the convex function is expressed as:

$$\frac{\partial \Gamma}{\partial \lambda_i} = E(g_i) - \int_c^d \int_a^b \exp\left[- \sum_{k=0}^{m} \lambda_k g_k(x,y) \right] g_i(x,y) dxdy \;\; i=1, 2,\ldots, 7 \quad (3.16)$$

and the Hessian matrix H is expressed as:

$$H = \begin{bmatrix} \text{var}(x) & \text{cov}(x, x^2) & \ldots & \text{cov}(x, xy) \\ \text{cov}(x^2, x) & \text{var}(x^2) & \ldots & \text{cov}(x^2, xy) \\ \ldots & \ldots & \ldots & \ldots \\ \text{cov}(xy, x) & \text{cov}(xy, x^2) & \ldots & \text{var}(xy) \end{bmatrix}$$

where elements $H_{i,j}$ (i, j=1, 2,…, 7) of the Hessian matrix are expressed as:

$$H_{i,j} = \text{cov}(g_i, g_j) = \int_c^d \int_a^b \exp\left[- \sum_{k=0}^{m} \lambda_k g_k(x,y) \right] g_i(x,y) g_j(x,y) dxdy$$

$$- \int_c^d \int_a^b \exp\left[- \sum_{k=0}^{m} \lambda_k g_k(x,y) \right] g_i(x,y) dxdy \bullet \int_c^d \int_a^b \exp\left[- \sum_{k=0}^{m} \lambda_k g_k(x,y) \right] g_j(x,y) dxdy$$

where cov(g_i, g_i)=$var(g_i)$ and H^{-1} is the inverse of Hessian matrix H. In this study, the MATLAB function *fminsearch* was used to obtain the minimum of equation (3.14) and hence the Lagrange multipliers.

For the generation of monthly streamflow, 12 joint PDFs of streamflows of two adjacent months are needed and the corresponding Lagrange multiplier sets ($\mathbf{\Phi}_{1,2}$, $\mathbf{\Phi}_{2,3},\ldots, \mathbf{\Phi}_{12,1}$) of each joint PDF have to be estimated. Each Lagrange multiplier in the joint PDF in equation (3.8) is related to one statistic that is to be preserved. For instance, in parameters $\Phi_{1,2}=[\lambda_0, \lambda_1,\ldots,\lambda_7]$ of the joint PDF of streamflow of January and February, λ_1 , λ_2 and λ_3 relate to the mean, standard deviation and skewness of the January streamflow, λ_4 , λ_5 and λ_6 relate to the mean, standard deviation and skewness of the February streamflow, and λ_7 is the parameter relating to the lag-one correlation of streamflows of two adjacent months. Likewise, parameters $\mathbf{\Phi}_{2,3}$ of the joint PDF of streamflow of February and March relate to the required statistics for the February and March streamflows and so on. If more statistics (e.g., kurtosis) need to be preserved, one can incorporate the corresponding Lagrange multipliers in the joint PDF. Thus, the entropy-based formulation is quite flexible and can be extended to incorporate more statistical properties, if needed.

3.2.5 Generation

There are several techniques that can be employed for the generation of random values from the bivariate distribution, such as the conditional distribution method, the transformation method, the acceptance/rejection method, and the composition method [*Johnson*, 1987; *Balakrishnan and Lai*, 2009]. In order to sample from the continuous

joint PDF $f(x, y)$ to obtain the random values of (X, Y), the conditional distribution method was employed in this study. For the generation of streamflow while preserving the correlation between adjacent months, streamflow values of one month can be generated from the conditional distribution given the streamflow value of its previous month. When the method is applied to the generation of monthly data of each year, 12 conditional distributions have to be used sequentially. To illustrate this method, it is assumed that the simulation starts from January and n year data is to be generated. Let $x_{t,s}$ denote streamflow for month s of year t (s=1, 2,…,12, t=1, 2,…, n). The step by step simulation procedure for generating random values of each month can now be summarized as:

(1) Generate a random vector $(x_{1,1}, x_{1,2})$ from the joint density function given by equation (3.8) with parameters $\Phi_{1,2}$. The random values $x_{1,1}$ and $x_{1,2}$ are the January and February streamflows of the first year.

(2) With the initial value $x_{1,2}$ generated in step (1), one can generate the March streamflow of the first year $x_{1,3}$ from the conditional cumulative distribution function in equation (3.12) with parameters $\Phi_{2,3}$. To that end, generate a uniform distributed random value w_1 between [0, 1] which can be done with the use of random number generator function rand in MATLAB. This w_1 value can be considered to be the conditional cumulative probability corresponding to a specific value $x_{1,3}$, given the initial value $x_{1,2}$. This can be expressed with equation (3.12) as:

$$F_{Y|X}(x_{1,3}|x_{1,2}) = w_1$$

Then, $x_{1,3}$ can be generated by solving the above equation. Similarly, monthly streamflows $x_{1,4}$,..., $x_{1,12}$ can be generated while parameters $\Phi_{3,4}$, …, $\Phi_{11,12}$ are used sequentially. Then monthly streamflow of the first year can be generated.

(3) With $x_{1,12}$ (December streamflow of the first year) generated in step (2), $x_{2,1}$ (January streamflow of the second year) can be generated with the parameter $\Phi_{12,1}$ similarly. In this manner, monthly streamflow of the second year, $x_{2,1}$,…, $x_{2,12}$, can be generated.

(4) Repeat step (3) until monthly streamflow of the nth year is generated.

In the above steps, numerical integration is performed to generate random values from the inverse (conditional) cumulative distribution.

3.3 Test with synthetic data

In order to test the performance of the proposed method to approximate the density function of the known model and reproduce the statistics of samples from it, the bivariate gamma distribution was selected. The gamma marginal distribution of a random variable z with scale parameter β and shape parameter γ is defined as:

$$f(z) = \frac{z^{\gamma-1}}{\Gamma(\gamma)\beta^{\gamma}} \exp(-z/\beta) \tag{3.17}$$

The five-parameter bivariate gamma distribution $f(x,y)$ by *Smith et al*. [1982] can be expressed as:

$$f(x,y)=\frac{(\beta_1 x)^{\gamma_1-1}(\beta_2 y)^{\gamma_2-1}\exp\left[-(\beta_1 x+\beta_2 y)/(1-\eta)\right]}{(1-\eta)^{\gamma_1}\Gamma(\gamma_1)\Gamma(\gamma_2-\gamma_1)} \times\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\frac{\eta^{j+k}}{(1-\eta)^{2j+k}}\frac{\Gamma(\gamma_2-\gamma_1+k)}{\Gamma(\gamma_2+j+k)}\frac{(\beta_1 x\beta_2 y)^j(\beta_2 y)^k}{j!k!} \tag{3.18}$$

where β_1 and β_2 are the scale parameters; γ_1 and γ_2 are the shape parameters ; $\eta=\rho(\gamma_2/\gamma_1)0.5$, where ρ is the correlation coefficient between x and y. Parameters of the bivariate gamma distribution were specified as: β_1=1, γ_1=6, β_2 =5, γ_2 =9 and ρ=0.25.

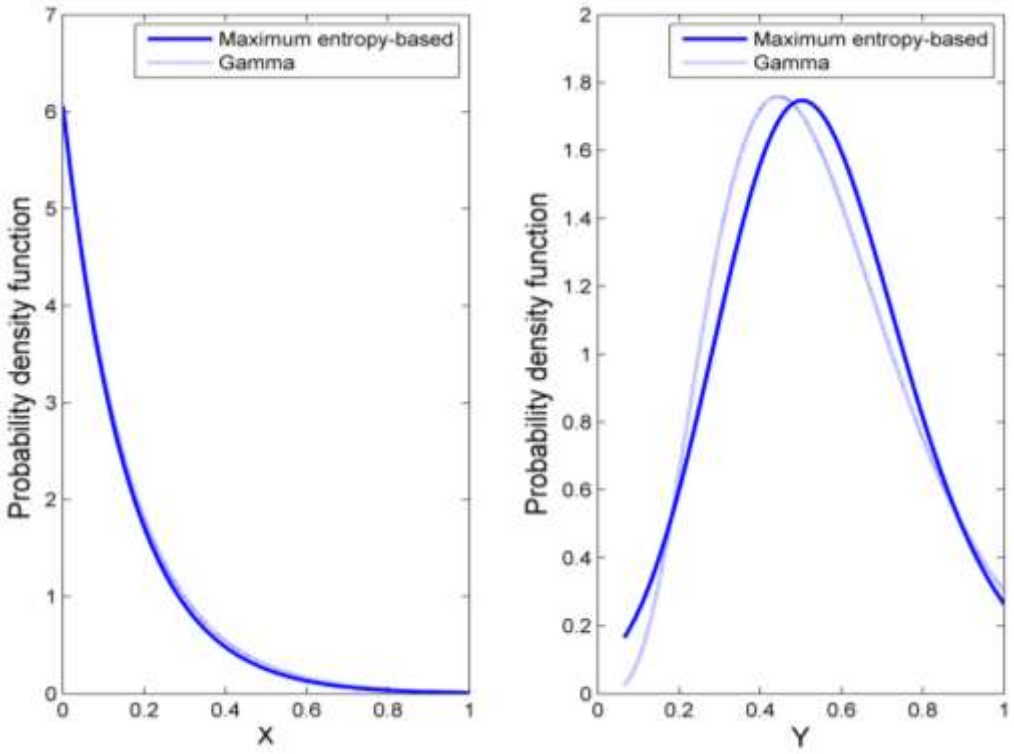

Figure 3. 1 Maximum entropy-based marginal PDFs and gamma marginal PDFs for variables X and Y.

One sample consisting of 3000 data pairs was drawn from the bivariate gamma distribution, which is regarded as the calibration sample for fitting and evaluating the entropy-based method. To quantify the performance of the proposed method in approximating the marginal and bivariate gamma PDFs, the root mean square error (RMSE) was computed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - p_i)^2} \tag{3.19}$$

where n is the length of the data, s_i and p_i (i=1,…, n) are the maximum entropy-based probability densities of equations (3.8) and (3.10) and densities from the marginal and bivariate gamma distribution of equation (3.17) and (3.18) corresponding to the i^{th} value.

The maximum entropy-based marginal PDFs of random variable X and Y in equation (3.10) estimated from the calibration sample together with the gamma marginal PDFs in equation (3.17) were plotted, as shown in Figure 3. 1. As can be seen the maximum entropy-based density of X was virtually indistinguishable from that of the gamma density of X. Generally the maximum entropy-based density approximates the gamma density of Y relatively well, though some discrepancies exist. The RMSE values between the maximum entropy-based density and gamma density were 0.062 for variable X and 0.14 for variable Y, respectively. Thus, the maximum entropy-based marginal PDFs estimated from the calibration sample can approximate the gamma marginal PDFs relatively well.

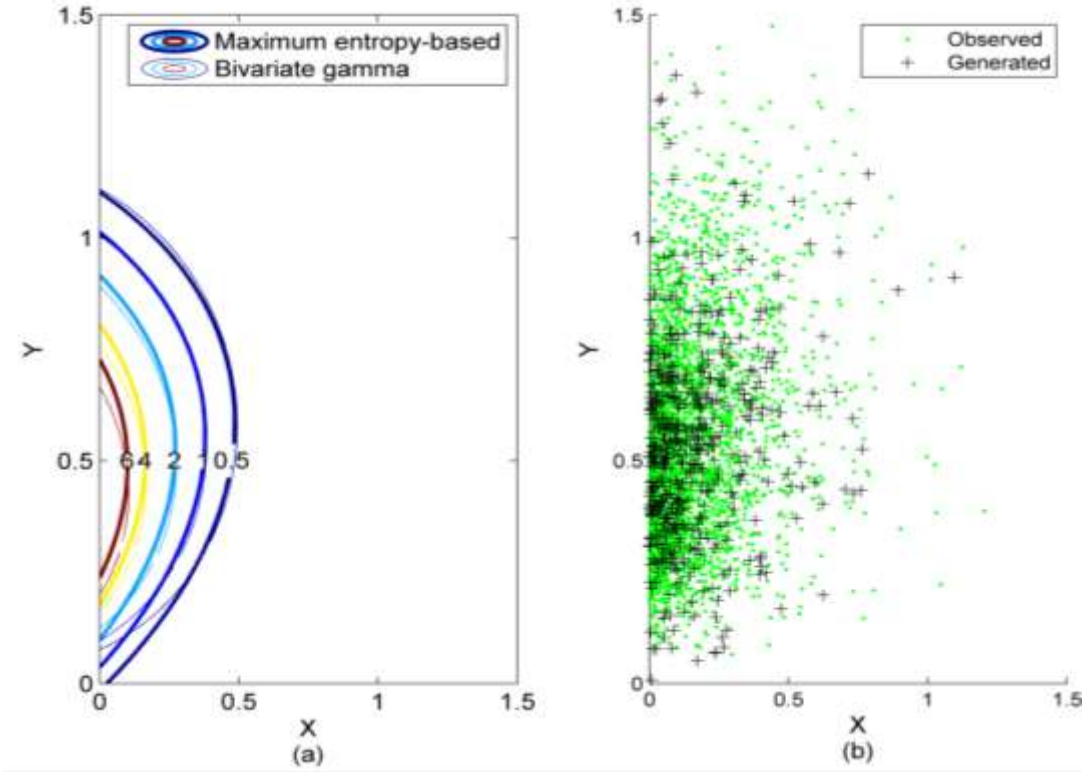

Figure 3. 2 Comparison of maximum entropy-based joint distribution and bivariate gamma distribution. (a) Contours of maximum entropy-based joint PDF and bivariate gamma PDF; (b) Comparison of generated data pairs with the calibration sample.

Contours of the maximum entropy-based joint PDF in equation (3.8) estimated from the calibration sample and the underlying bivariate gamma PDF in equation (3.18) were plotted, as shown in Figure 3. 2 (a). The contour lines of the maximum entropy-based PDF approximate those of the underlying gamma PDF relatively well. The RMSE value between the maximum entropy-based joint density and underlying bivariate gamma density was 0.74. A bivariate sample with 500 data pairs was generated and is shown together with the calibration sample in Figure 3. 2 (b). It is seen that generally the

spreading pattern of the generated data pairs matches that of the calibration sample well. This shows that the proposed method approximated the underlying bivariate gamma PDF relatively well.

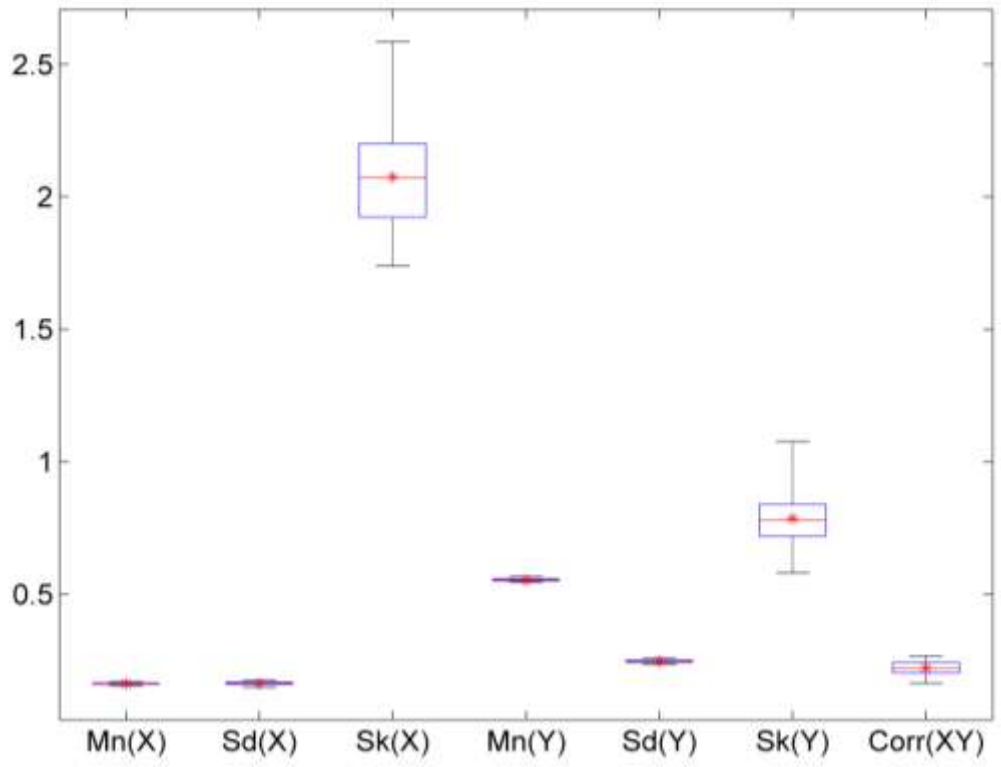

Figure 3. 3 Boxplots of statistics of the calibration sample and generated data pairs. (Mn, Sd and Sk represent the mean, standard deviation and skewness. Corr represents the correlation. Star marks represent statistics of the calibration sample.)

100 bivariate samples each consisting of 3000 data pairs were generated from the maximum entropy-based joint PDF in equation (3.8) estimated from the calibration

sample. Statistics of generated data pairs and the calibration sample were compared using box plots, including the mean, standard deviation, skewness and lag-one correlation. The central mark of the box is the median and the end lines of the box represent 25th and 75th percentiles. The whiskers are the maximum and minimum values of the simulated statistics. A wide box plot signifies large variability. When a statistic falls in the box plot, the performance is considered to be good [*Prairie et al*., 2007; *Nowak et al.*, 2010; *Salas and Lee*, 2010]. Statistics of the generated data pairs and the calibration sample were compared with box plots as shown in Figure 3. 3. All statistics fell in the box plots and this showed that the proposed method can preserve the mean, standard deviation, skewness and lag-one correlation well.

3.4 Application

The entropy-based method was applied to monthly streamflow at 10 sites in the Colorado River basin from 1906-2003 [*Lee and salas*, 2006]. These data can be found at the website: http://www.usbr.gov/lc/region/g4000/NaturalFlow/previous.html. Without loss of generality, the monthly streamflow data was scaled to [0, 1] for computational convenience. For the original data (OD) of each month with maximum value MX and minimum value MN, the scaled data (SD) of each month was expressed as: SD=[OD-(1-*d*)MN)]/[(1+*d*)MX-(1-*d*)MN],where *d* is a scale parameter, which was selected as 0.05 in this study. With the use of constraints in equations (3.3) to (3.6), parameters ($\Phi_{1,2}$, $\Phi_{2,3}$, … ,$\Phi_{12,1}$) of each joint PDF in equation (3.8) were first estimated. Then, the conditional distribution was derived from the known joint PDF using equations (3.8) and (3.10). Thereafter, samples were drawn sequentially using the procedure outlined in

Section 3.5 and then transformed back to the original domain. From the scaling expression, the values of 0 and 1 in the scaled domain corresponded to the (1-*d*)MN and (1+*d*)MX in the original domain and thus the values outside the observed streamflow range can be generated.

Simulation results were satisfactory for all stations, as shown for the station, Lees Ferry, Arizona, on the Colorado River (U.S. Geological Survey station number 09380000), which has been used in earlier studies [*Prairie et al*., 2006; *Salas and Lee*, 2010]. 100 flow sequences, each of 100 and 400 years long, termed as S_1 and S_2, respectively, were generated to test the proposed method. Statistics of generated and historical data, including the mean, standard deviation, skewness, lag-one correlation, maximum and minimum values, were compared using box plots. Furthermore, other statistics pertaining to low values reflecting drought conditions, such as maximum drought length, maximum drought amount, maximum surplus length, maximum surplus amount, and storage capacity, were also compared for generated and historical data.

The box plots was used to measure the performance of the proposed method and the performance was considered to be good when a statistic fell in the box as described in the previous section. To quantify the performance of the entropy-based method, absolute error (AE) and relative error (RE) of the simulated statistics were computed as $AE=S_m-X_o$ and $RE=(S_m-X_o)/X_o$, where S_m is the median of simulated statistic for the generated data, and X_o is the statistic for the historical data.

3.4.1 Validation of Marginal PDF and joint PDF

Maximum entropy-based marginal PDFs and empirical histograms of scaled streamflows were constructed and compared, as shown for two sample months of May and June in Figure 3. 4. Note that the marginal entropy based PDF in equation (3.10) of streamflow of a specific month, say May, can either be derived from the joint density function of April and May streamflow with parameter $\Phi_{4,5}$ (denoted as $f_{5,45}$) or from the joint density function of May and June streamflows with parameter $\Phi_{5,6}$ (denoted as $f_{5,56}$). Though the PDF of May streamflows can be derived from different joint distributions with different Lagrange multipliers, densities $f_{5,45}$ and $f_{5,56}$ should be close to each other, which is verified in Figure 3. 4.

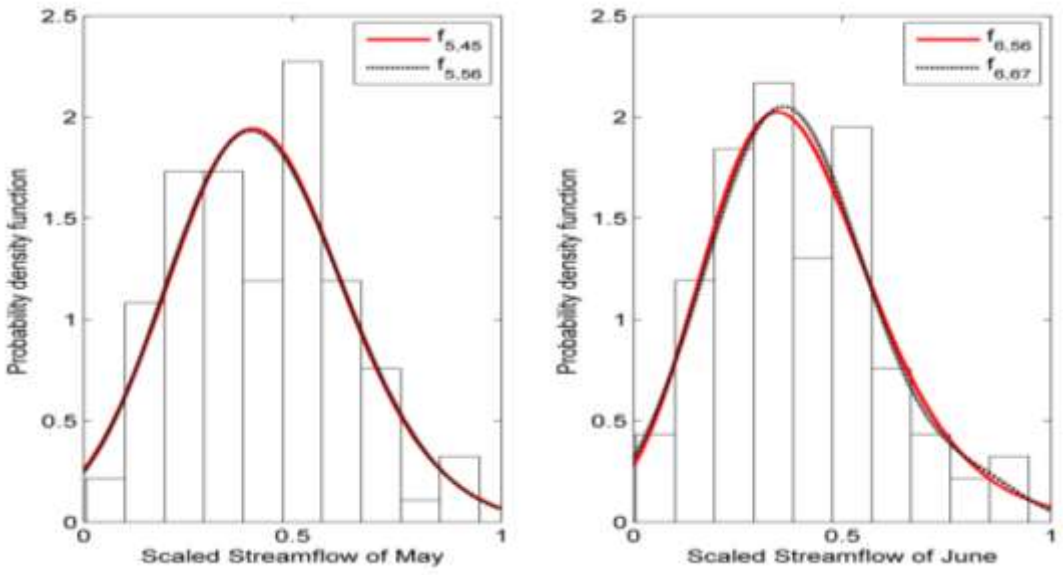

Figure 3. 4 Maximum entropy-based marginal PDFs and empirical histograms for scaled May and June streamflow ($f_{5,45}$: marginal PDF for May streamflow with parameter $\Phi_{4,5}$; $f_{5,56}$: marginal PDF for May streamflow with parameter $\Phi_{5,6}$; $f_{6,67}$, marginal PDF for June streamflow with parameter $\Phi_{6,7}$).

The probability density function of the May streamflow was bimodal, which has been shown by *Prairie et al.* [2006]. The maximum entropy-based PDFs fitted the empirical histograms relatively well, except that the bimodality in the density of the scaled May streamflow could not be resolved. Contours of maximum entropy-based and empirical joint densities are shown in Figure 3. 5 (a, b). The historical data spread along the contours as seen in Figure 3. 5 (a). The maximum entropy-based joint densities matched the empirical densities well for most parts as shown in Figure 3. 5 (b). For example, the maximum entropy-based joint density values near the empirical contour line with a density of 2 were 1.5, 2 and 2.6.

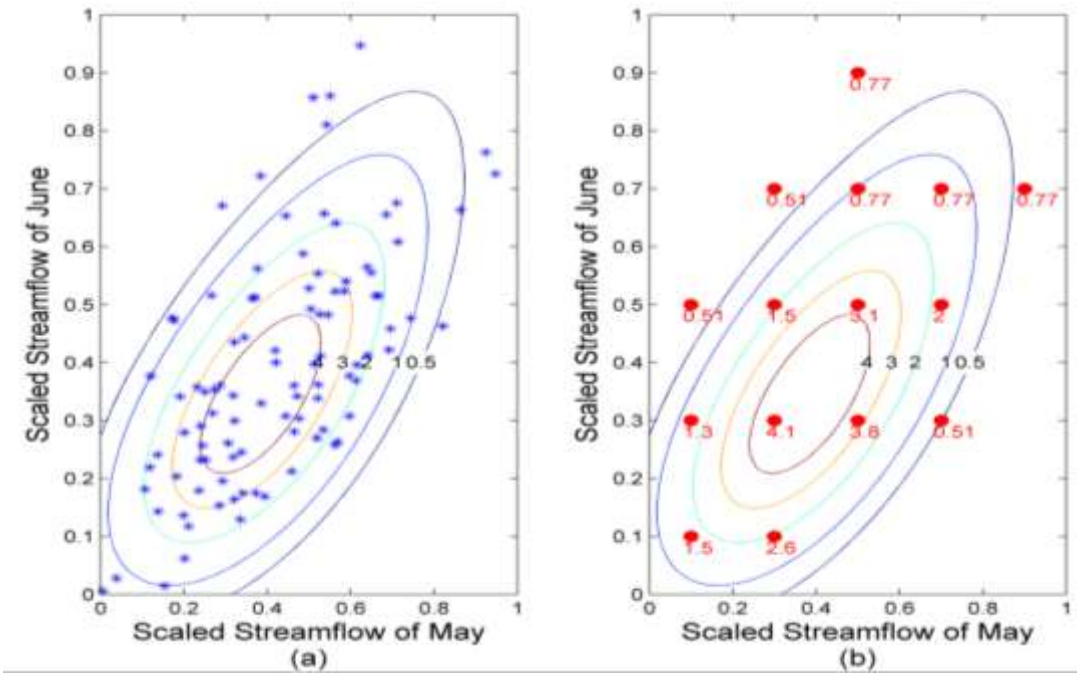

Figure 3. 5 Contours of the maximum entropy-based PDF of the scaled May and June streamflow. (a) historical data plotted as stars; (b) empirical joint PDFs plotted as points.

3.4.2 Monthly mean, standard deviation, skewness and lag-one correlation

Statistics of generated and historical data for all months of simulation S_1 were computed, as shown in Figure 3. 6. The median values of simulated mean, standard deviation, skewness and lag-one correlation were close to those of the historical data. All statistics of mean, standard deviation, skewness and lag-one correlation fell in the box plots, indicating the goodness of the entropy-based method.

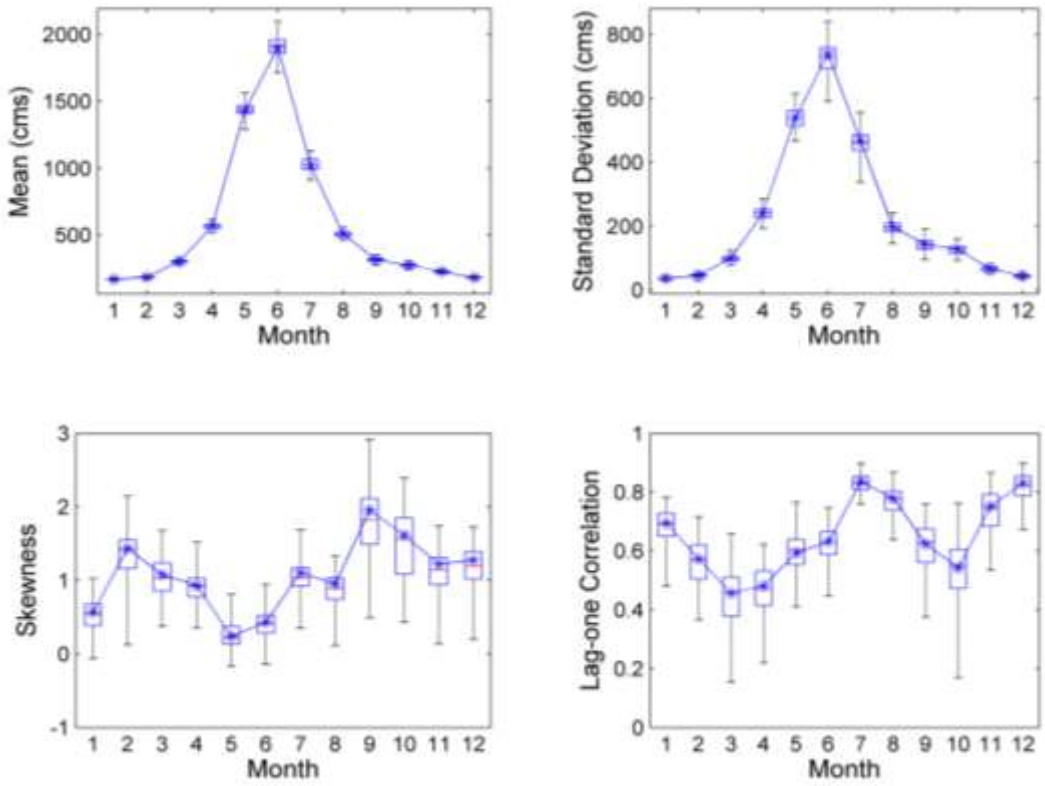

Figure 3. 6 Boxplots of mean, standard deviation, skewness and lag-one correlation of generated and historical data for simulation S_1. (Continuou lines with star marks for each month represent statistics of the historical data.)

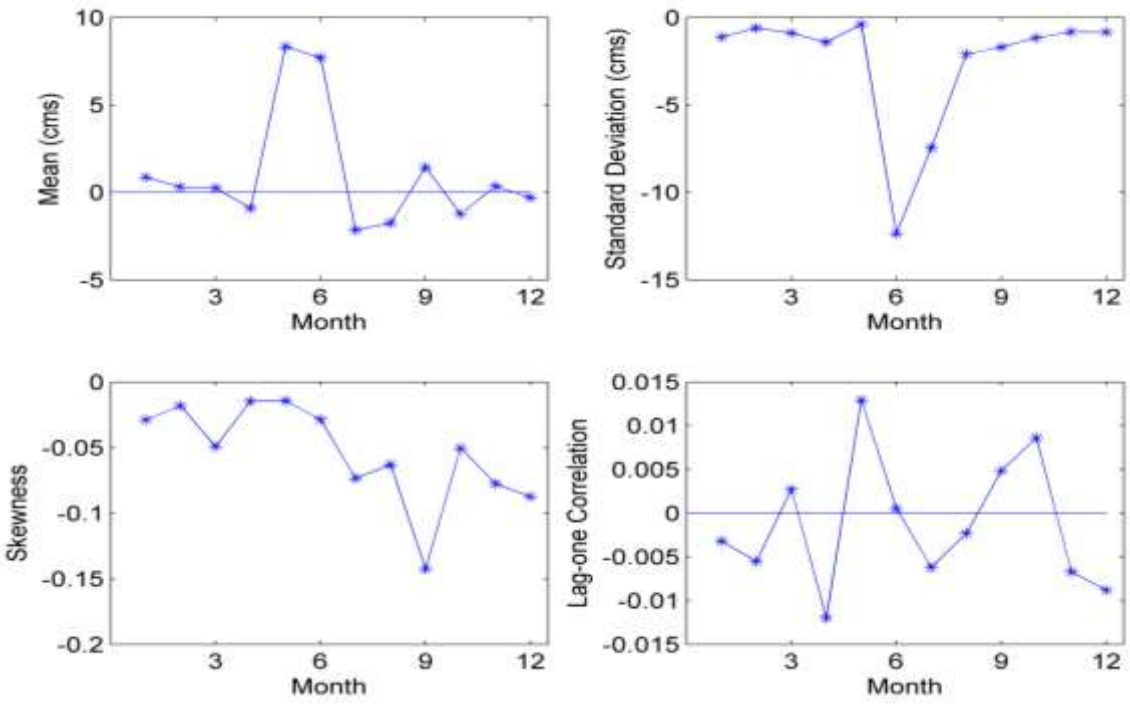

Figure 3. 7 Absolute errors of mean, standard deviation, skewness and lag-one correlation for simulation S_1.

The absolute error and relative error for each statistic were calculated, as shown in Figure 3. 7 and Table 3. 1. Even though the absolute error was relatively large for several months, like that for the mean of May and standard deviation of June, as seen from Figure 3. 7 , the result was satisfactory based on the relative error in Table 3. 1. The relative error of mean, standard deviation and lag-one correlation was lower than 5% and that of skewness was lower than 10% for all months. The relative error of simulated skewness was relatively high and was not preserved as well as other statistics. The lag-one correlation between the December streamflow of the previous year and the January

streamflow of the current year was also preserved well, as seen from Table 3. 1 in that the relative error was -0.5%.

Table 3. 1 Relative error (%) of statistics for each month for simulation S_1. (Units for the mean and standard deviation are in cubic meter per second (cms).)

Month / Statistics	1	2	3	4	5	6	7	8	9	10	11	12
Mean	0.52	0.17	0.08	-0.16	0.58	0.41	-0.21	-0.35	0.46	-0.46	0.16	-0.17
Standard Deviation	-3.04	-1.31	-0.90	-0.59	-0.07	-1.67	-1.59	-1.06	-1.18	-0.91	-1.23	-1.89
Skewness	-5.07	-1.27	-4.57	-1.58	-6.03	-6.60	-6.65	-6.63	-7.30	-3.12	-6.36	-6.86
Lag-one Correlation	-0.46	-0.96	0.60	-2.49	2.18	0.08	-0.74	-0.30	0.78	1.59	-0.89	-1.06

Salas and Lee [2010] showed that nonparametric model with the long-term dependence (NPL) model underestimated the skewness throughout the year and overestimated the standard deviation for wet months, while the *K*-Nearest Neighbor (KNN) resampling technique with gamma kernel perturbation with the aggregate variable (KGKA) and the pilot variable (KGKP) underestimated lag-one correlation. Since all these statistics were preserved well and no underestimation or overestimation existed, the entropy-based method performed better in preserving the four statistics.

Statistics of generated and historical data for all months of simulation S_2 are shown in Figure 3. 8. It is seen that all statistics fell in the box plots, indicating satisfactory model performance for simulation S_2. In addition, the box plots became narrower and thus the variability of simulated statistics was reduced, as shown in Figure 3. 8.

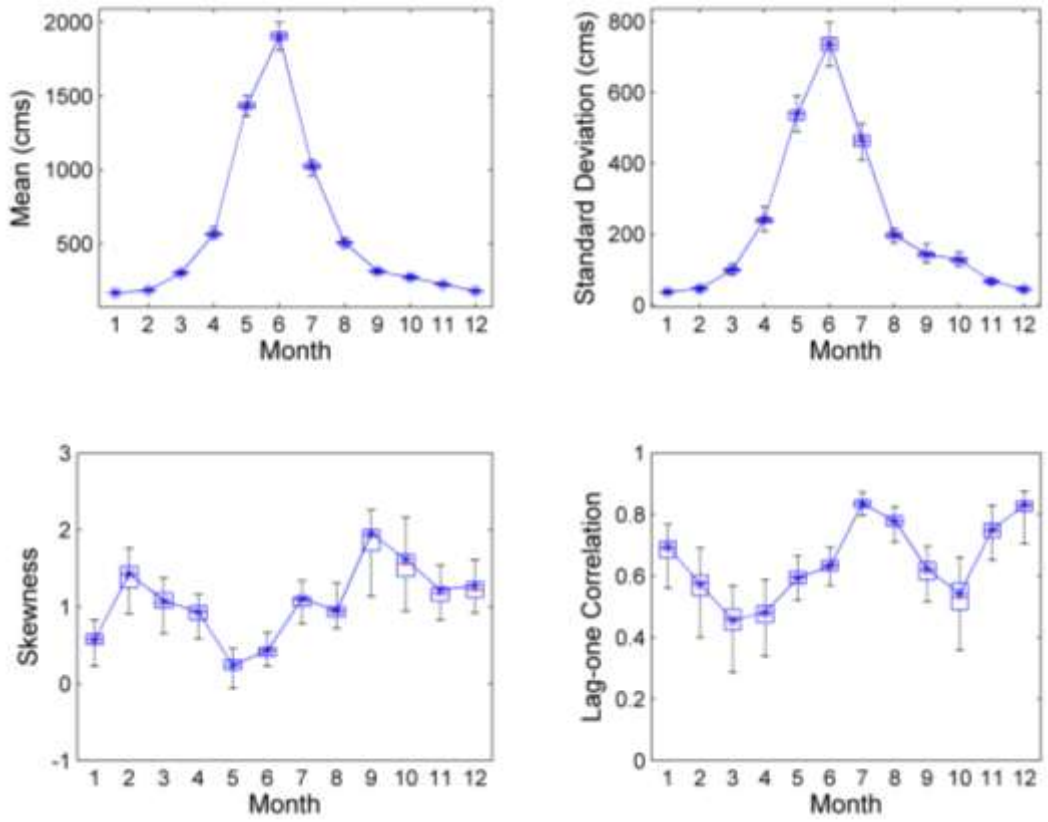

Figure 3. 8 Boxplots of mean, standard deviation, skewness and lag-one correlation of generated and historical data for simulation S_2. (Continuous lines with star marks for each month represent statistics of the historical data.)

Specifically, comparison of statistics of the generated and historical data of selected months, January and May, for simulation S_1 and S_2 with different percentiles is shown in Table 3. 2. For the simulation of skewness of the May streamflow, 25th, 50th and 75th percentiles of simulated skewness were 0.15, 0.24, 0.36 in simulation S_1 and 0.19, 0.24, 0.31 in simulation S_2, respectively. The interquartile range (distance between 25th percentile and 75th percentile) is 0.12 in simulation S_2 which was smaller than 0.21 in simulation S_1.

Table 3. 2 Comparison of statistics of generated and observed streamflow of January and May for simulation S_1 and S_2. (Units for the mean and standard deviation are in cubic meter per second (cms).)

Month	Statistics	Simulation S_1	Simulation S_2	Percentile	Observation
January	Mean	165	166	25th	167
		167	168	50th	167
		170	169	75th	167
	Standard Deviation	33	35	25th	37
		36	36	50th	37
		38	37	75th	37
	Skewness	0.39	0.51	25th	0.57
		0.58	0.58	50th	0.57
		0.69	0.64	75th	0.57
	Lag-one Correlation	0.63	0.66	25th	0.70
		0.68	0.69	50th	0.70
		0.72	0.71	75th	0.70
May	Mean	1408	1416	25th	1433
		1437	1436	50th	1433
		1472	1454	75th	1433
	Standard Deviation	512	523	25th	539
		538	540	50th	539
		560	548	75th	539
	Skewness	0.15	0.19	25th	0.24
		0.24	0.24	50th	0.24
		0.36	0.31	75th	0.24
	Lag-one Correlation	0.55	0.58	25th	0.59
		0.59	0.60	50th	0.59

Generally, the median values of statistics of simulated data matched those of historical data well when the length of generated annual streamflow was longer than 100 and simulation of these statistics could be further improved by generating a longer record. Nevertheless, for a streamflow record with a length of annual streamflow around

100, the proposed method satisfactorily preserved the mean, standard deviation, skewness and lag-one correlation.

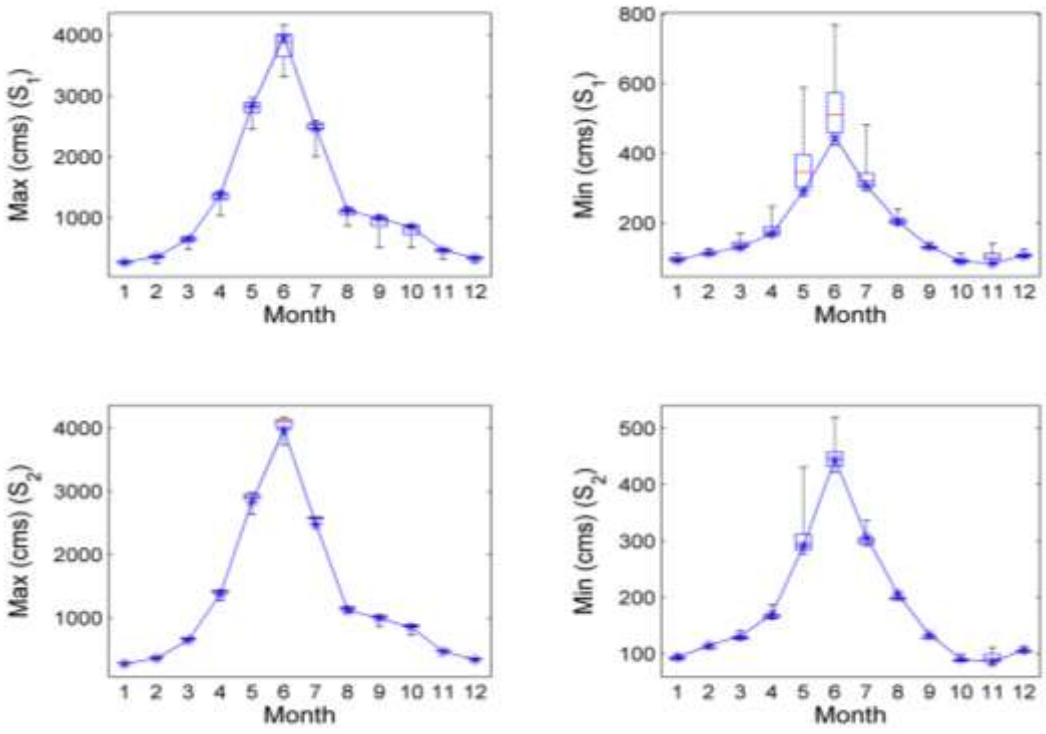

Figure 3. 9 Boxplots of maximum and minimum values of generated and historical data for simulation S_1 and S_2. (Continuous lines with star marks for each month represent historical data.)

3.4.3 Monthly maximum and minimum values

The maximum and minimum values of generated data and historical data for all months of simulation S_1 and S_2 were obtained, as shown in Figure 3. 9 . There was no significant overestimation or underestimation of the maximum and minimum values for most months of simulation S_1. However, for many months of simulation S_2, the

maximum values were overestimated and the minimum values were underestimated. The value of scale parameter *d* affects the generated maximum and minimum values and when *d* equates zero the generated values are bounded by the observed maximum and minimum values. In both simulation S_1 and S_2, no negative values were generated.

Salas and Lee [2010] showed that the nonparametric NPL model preserved the maximum values well, although the minimum values were underestimated, whereas both the KGGA and KGKP models preserved the maximum and minimum values well. Since the overestimation of maximum values and underestimation of minimum values occurred when a relatively long record of annual streamflow were generated, the proposed method did not perform as well.

3.4.4 Extension to higher-order moments

The entropy-based method can be extended to incorporate higher-order moments and more lag correlations, if needed. For example, in order to preserve kurtosis in the simulation, two Lagrange multipliers associated with the fourth non-central moments of variables X and Y would be added in equation (3.2). Then, streamflow would be generated based on the corresponding conditional distribution as illustrated in section 3.5. Although the preservation of the kurtosis may not be essential and the sample instability problems with the estimation of higher moments may exist [*Fiering*, 1967], one simulation of this extension demonstrated the performance of the proposed method. Comparison of simulated kurtosis between the proposed method and the extended method for 100 sequences with 400 years of annual streamflow generated in each

sequence, as shown in Figure 3. 10 (a, b), showed that the kurtosis was preserved better when the fourth moment was also incorporated as a constraint.

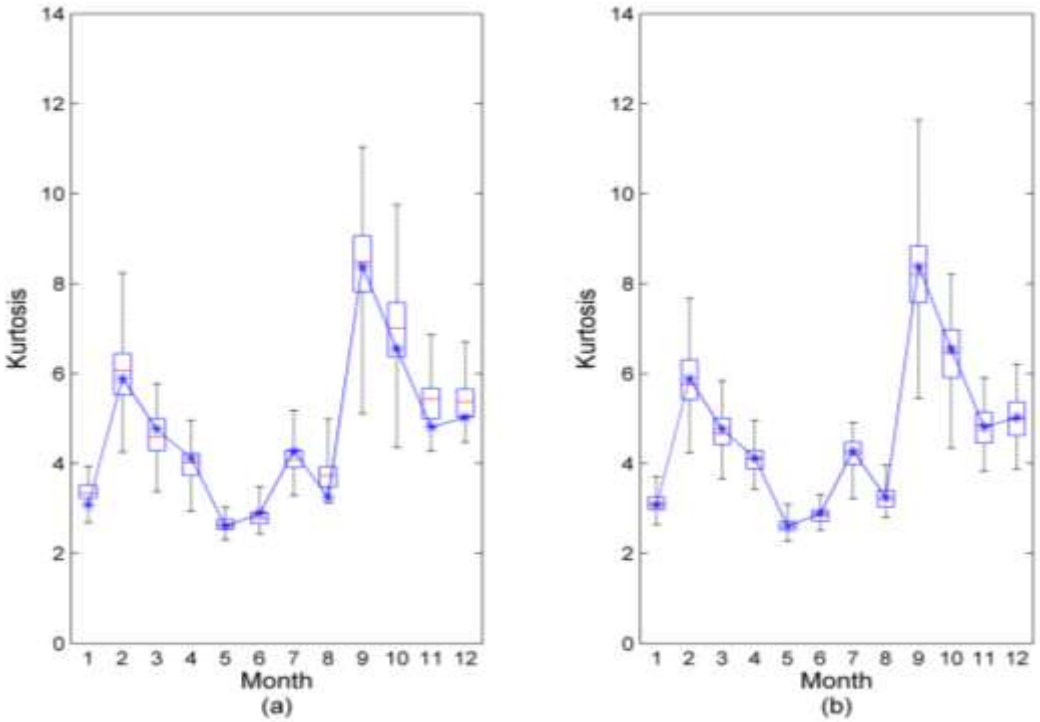

Figure 3. 10 Boxplots of kurtosis of historical and generated data. (a) proposed method; (b) extended method.

3.4.5 Drought, surplus and storage statistics

Box plots of the drought, surplus and storage statistics (ratio of generated over historical) were constructed for simulation S_1 and S_2 and only the result for simulation S_2 is presented as shown in Figure 3. 11. The water demand level was selected as a fraction of the historical mean and in this study it was selected as 0.7, 0.8, 0.9 and 1.0. For

simulation S_2, as shown in Figure 3. 11, the maximum deficit length and amount for the water demand level 1.0 were overestimated somewhat, but in general these statistics were preserved well. However, for simulation S_1 these statistics were not preserved as well.

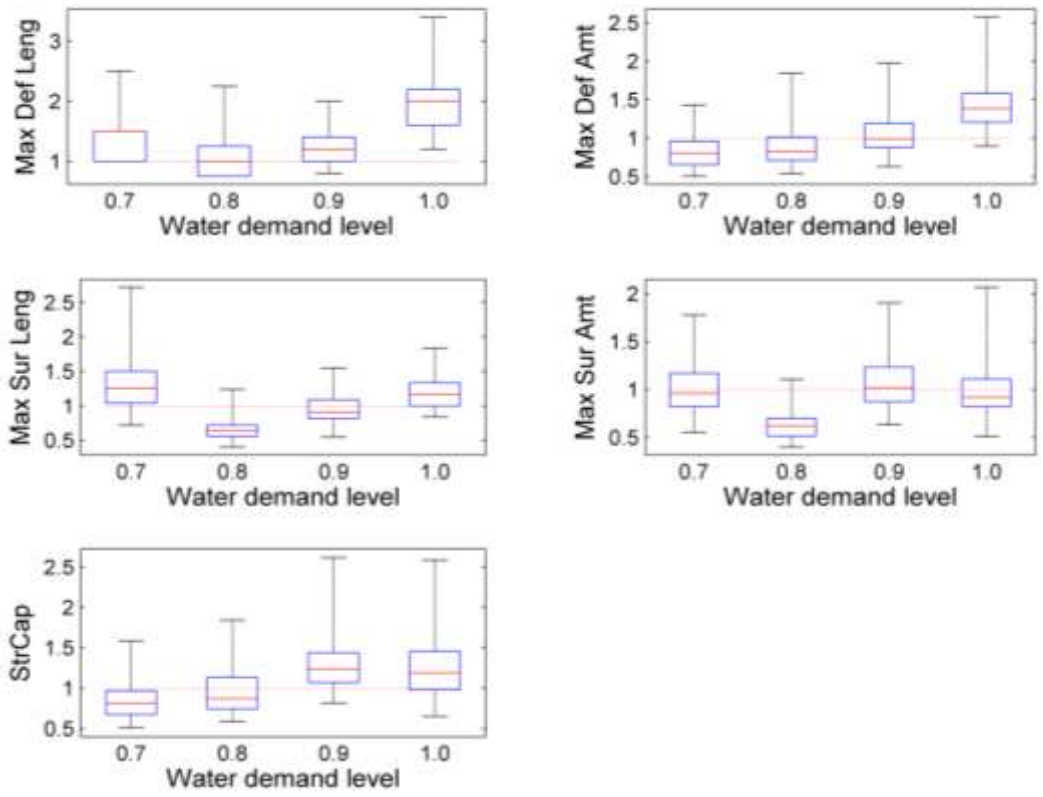

Figure 3. 11 Boxplots of ratio of drought, surplus and storage capacity statistics (Max Def Leng, maximum deficit length; Max Def Amt, maximum deficit amount; Max Sur Leng, maximum surplus length; Max Sur Amt, maximum surplus amount).

3.5 Conclusion

A new model, based on entropy theory, for single-site monthly streamflow simulation is developed. Streamflow data is generated by sampling from the conditional distribution derived from the joint probability density function of streamflow of two adjacent months. The entropy-based model is applied to 10 sites in the Colorado River basin and results indicate that it satisfactorily preserves mean, standard deviation, skewness and lag-one correlation. Streamflow outside the observed streamflow range can be generated, though overestimation of the maximum values and underestimation of minimum values can occur when a relatively long record of annual streamflows is generated. Generally, drought, surplus and storage statistics can be preserved well with the generation of a relatively long record.

The advantage of the proposed method is that no assumption is made about the marginal distribution of the historical data. Therefore, the method can be applied to non-normal streamflow and the transformation of streamflow to be normal is not needed. In addition, it can preserve the cross-correlation between streamflow in December (or last season) of the previous year and that in January (or first season) of the current year and avoid negative values in the generation. Further, if more statistical characteristics (e.g., kurtosis and more lag correlations) are needed to be preserved, the entropy-based method can also be applied by incorporating these statistics as constraints. The disadvantage of the method is that it will be computationally cumbersome when more statistics are to be preserved and determination of more Lagrange multipliers is involved. This would be the case if the method were applied to multi-site streamflow simulation,

since statistics of streamflow at different stations would be used as constraints and integration in higher dimension will be involved in the determination of more Lagrange multipliers and streamflow simulation. However, this should not be an insurmountable difficultly, given the available numerical tools and computer progress. In addition, the bimodality that may exist in the empirical probability density function cannot yet be resolved with the proposed model.

CHAPTER IV

ENTROPY-COPULA METHOD FOR SINGLE-SITE MONTHLY STREAMFLOW SIMULATION

4.1 Introduction

For steamflow simulation, it is desired that synthetic streamflow is similar to historical streamflow and preserves moment statistics (such as mean, standard deviation, and skewness), and dependence structure (such as lag-one correlation).

The preservation of inter-annual statistics is one of the difficulties in streamflow simulation. The inter-annual statistics are important for the simulation of long wet and dry periods that are critical for drought management and planning [*Sivakumar and Berndtsson*, 2010]. Generally a lag-one seasonal model is not sufficient for the preservation of the inter-annual statistic. *Sharma and O'Neill* [2002] proposed a nonparametric approach for monthly streamflow simulation that is capable of preserving the inter-annual dependence by using an aggregate variable as a conditional variable in the simulation. *Salas and Lee* [2010] proposed two approaches based on the *K*-Nearest Neighbor resampling techniques that are capable of preserving annual variability by introducing an aggregate variable and pilot variable.

In addition, the lag-one correlation (or Pearson product-moment correlation coefficient) only measures the linear dependence of random variables, which may not be adequate in reality. Some nonlinear dependence is also desirable to characterize the dependence of streamflow. Generally it is difficult for the conventional parametric

approaches to represent the nonlinear dependence. The bimodality is one of the "unusual features" that may exist in the probability density function of streamflow data that is difficult for the conventional parametric approach to represent [*Sharma and O'Neill*, 2002]. Though the mixed distribution can be used to resolve the bimodality, bias in the statistics of streamflow may occur. The nonparametric approaches are generally needed to model the nonlinear dependence and resolve the bimodality in the probability density function of streamflow.

The entropy-based streamflow distribution can be derived from the moments and thus the sampling from the resulting distribution is expected to preserve these moment statistics. *Hao and Singh* [2011]applied the entropy theory for the single site monthly streamflow simulation that is capable of preserving mean, standard deviation, skewness (and kurtosis if needed). However, the model requires specification of constraints for the statistics to be preserved and that leads to as many Lagrange multipliers as statistics that need to be estimated which may be tedious. The copula based joint distribution can be applied to model the dependence of streamflow. *Lee and Salas* [2011] proposed the copula method for annual streamflow simulation. Therefore, combining the concepts of entropy and copula, an entropy-copula method is proposed for streamflow simulation in which the joint distribution is constructed using the copula method, where the marginal distributions are constructed using the entropy method. In this method, less parameters are needed to be estimated as compared with the method proposed by *Hao and Singh* [2011]. The entropy based marginal distributions do not rely on the Gaussian assumption and are able to model the asymmetry property of the streamflow without data

transformation. The entropy-based method also has the potential ability to resolve bimodality. Furthermore, the proposed method is able to model the nonlinear dependence of streamflow due to the inclusion of the copula component. The entropy-copula method can be extended to preserve the inte-annual dependence. Application in the Colorado River basin illustrates the effectiveness of the proposed method for monthly streamflow simulation.

4.2 Method

4.2.1 Entropy theory and marginal distribution

Using the principle of maximum entropy proposed by *Jaynes* [1957] with the first four moments as constraints, the maximum entropy-based probability density function defined on the interval [*a*, *b*] can be obtained as:

$$f(x) = \exp\left[-\lambda_0 - \lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3 - \lambda_4 x^4\right] \tag{4.1}$$

where λ_i, i=0,1,…, 4, are the Lagrange multipliers. The cumulative distribution function (CDF) of the maximum entropy-based distribution in equation (4.1) can be expressed as:

$$F_X(x) = \int_a^x f(t)dt \tag{4.2}$$

The Lagrange multipliers in equation (4.1) can be determined in terms of the specified constraints using the method presented [*Hao and Singh*, 2011]. For streamflow simulation, the preservation of mean, standard deviation, and skewness of monthly streamflow is needed. If the mean, standard deviation, and skewness (or the first three moments) are used as constraints for deriving the distribution of streamflow of each month, then samples from the distribution can be expected to preserve these statistics.

Certain monthly streamflows exhibit bimodality, and studies have shown that the maximum entropy-based distribution in equation (4.1) can model the bimodality [*Matz*, 1978]. The fourth moment (related to kurtosis) was also used to derive the distribution. Thus, the distribution in equation (4.1) can be employed to generate streamflow that preserves the mean, standard deviation, skewness and kurtosis. In addition, it also possesses the potential ability to resolve the bimodality.

4.2.2 Copula concept and joint distribution

For the continuous random vector (X,Y) with marginal cumulative distribution functions (CDF) $F_X(x)$ and $F_Y(y)$, respectively, the bivariate probability distribution of random vector (X,Y) can be expressed with its marginal CDFs and the copula C as [*Nelsen*, 2006; *Salvadori*, 2007]:

$$P(X \le x, Y \le y) = C[F_X(x), F_Y(y); \theta] = C(u, v; \theta) \tag{4.3}$$

Copula C with the parameter θ represents the dependence structure linking the marginal distributions and maps the two marginal distributions into the joint distribution as $[0,1]^2 \rightarrow [0,1]$. The conditional distribution can be derived from the copula in equation (4.3). Let $U = F_X(x)$ with u denoting a realization of random variable U and $V = F_Y(y)$ with v denoting a realization of random variable V. The conditional distribution can be defined as:

$$P(Y \le y | X = x) = \mathrm{C}_{2|1}(v|u) = \frac{\partial C(u, v; \theta)}{\partial u} \tag{4.4}$$

Four commonly used copula, namely Clayton, Frank, Gumbel and Gaussian, are listed in Table 4. 1. These copulas are capable of modeling random variables with

different dependence structures. For example, random variables from both the Gaussian copula and Frank copula exhibit symmetric dependence, while those from the Clayton copula exhibits asymmetric dependence strong in the left tail and weak in the right tail [*Trivedi and Zimmer*, 2005].

Table 4. 1 Copulas with associated parameter space and Kendall's tau.

Copula	C(u,v)	Parameter space	Kendal's tau
Clayton	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$\theta \in (0,\infty)$	$\frac{\theta}{\theta+2}$
Frank	$-\frac{1}{\theta}\ln\left[1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right]$	$\theta \in (-\infty,\infty)\setminus\{0\}$	$1-\frac{4}{\theta}+\frac{4D(\theta)}{\theta}$
Gumbel	$\exp\left\{-\left[(-\log u)^{-\theta}+(-\log v)^{-\theta}\right]^{-1/\theta}\right\}$	$\theta \in [1,\infty)$	$1-\frac{1}{\theta}$
Gaussian	$\Phi_2(\Phi^{-1}(u),\Phi^{-1}(v))$	$\theta \in (-\infty,\infty)$	$\frac{2}{\pi}\arcsin(\theta)$

The parameter of the copula θ in equation (4.3) has to be estimated. Utilizing the relationship of the copula parameter with some association measure (e.g., Kendall's tau, as shown in Table 4. 1, the parameter θ can be obtained through the association measure estimated from the observation. The exact maximum likelihood (EML) method and the inference functions for marginal (IFM) are two other methods that can be used for parameter estimation [*Joe*, 1997]. For the EML method, the likelihood including parameters of the marginal distributions and those of the copula can be maximized to estimate the parameters simultaneously. For the IFM method, parameters of the marginal distributions and those of the copula can be split, while the respective maximum

likelihood functions can be estimated separately. In this study, parameters of the copula were estimated using the IMF method in this study

4.2.3 Entropy-Copula method

The maximum entropy-based distribution in equation (4.1) can be used to preserve moment statistics (e.g., mean, standard deviation, skewness and kurtosis). The copula-based joint distribution in equation (4.3) can be used to model the dependence structure (e.g., lag-one correlation) between streamflows of two adjacent months. Therefore, samples from the copula based joint distribution in equation (4.3) (or the conditional distribution in equation (4.4))with the maximum entropy based marginal distribution $F_X(x)$ and $F_Y(y)$ in equation (4.2) can be expected to preserve the mean, standard deviation, skewness, kurtosis and lag-one correlation. Moreover, an important feature of the copula-based joint distribution is that it can model the nonlinear dependence of streamflows. The proposed method incorporates the entropy and copula concepts (denoted as EC method) and is used for monthly streamflow simulation hereafter.

The conditional distribution can be employed for the generation of random values from the joint distribution [*Johnson*, 1987; *Joe*, 1997]. For two adjacent months (say, January and February), the February streamflow can be generated from the conditional distribution in equation (4.4) given the January streamflow. Totally 12 conditional distributions have to be used sequentially to generate monthly streamflows. The cumulative distributions F_j and $F_{j\text{-}1}$ [corresponding to v and u in equation (4.4)] of streamflows of month j and j-1 are needed for the generation of streamflow of month j.

Let $x_{t,s}$ be the streamflow for month s of the year t (s=1, 2…, 12, t=1, 2…, n). The simulation steps are summarized as follows:

(1) Generate a uniform random value $u_{1,1}$ between [0, 1], which can be considered as the cumulative probability corresponding to a specific value of x, say $x_{1,1}$ (January streamflow of the first year). Then one obtains $x_{1,1}$ from the inverse cumulative probability distribution function as: $x_{1,1}=F_1^{-1}(u_{1,1})$, where F_1 is the cumulative distribution of the January streamflow.

(2) Generate another uniform distributed random value w between [0, 1], which is considered to be the conditional cumulative probability corresponding to a specific value $x_{2,1}$ (the February streamflow of the first year) with cumulative probability $u_{2,1}$, given the initial value $x_{1,1}$. Then, one obtains: $w=C_{2|1}(u_{2,1}|u_{1,1})$ from equation (4.4). $u_{2,1}$ can be obtained as: $u_{2,1}=C^{-1}_{2|1}(w)$ and then $x_{2,1}$ can be solved from: $x_{2,1}=F_2^{-1}(u_{2,1})$ accordingly, where F_2 is the cumulative distribution of the February streamflow.

(3) Repeat step (2) above until $x_{12,1}$ (the December streamflow of the first year) is generated. Thus, monthly streamflows of the first year are generated.

(4) With the generated $x_{12,1}$, $x_{1,2}$ (the January streamflow of the second year) can be generated with step (2) and similarly streamflows of other months of the second year $x_{3,2}$, $x_{4,2}$,…, $x_{12,2}$ can be generated.

(5) Repeat step (4) until $x_{12,n}$ is generated.

4.2.4 Extended Entropy-Copula method

The proposed entropy-copula method (denoted as EEC method) can be extended to preserve the inter-annual statistics by introducing an aggregate variable in the conditional distribution similar to the framework developed by *Sharma and O'Neill* [2002]. For monthly streamflow denoted as $X_1, X_2,\ldots, X_{12}, X_{13}, X_{14},\ldots, X_n$, where $X_1, X_2,\ldots, X_{12}$ are the monthly streamflows of the first year and so on, an aggregated streamflow can be defined as the summation of the previous m monthly streamflows (m=12 in this study):

$$Z_{t-1} = \sum_{i=1}^{m} X_{t-j} \tag{4.5}$$

Denoting streamflows of two adjacent months as X_t, X_{t-1} and the corresponding aggregate variable as Z_{t-1} with the cumulative distribution functions $F(X_{t-1})$, $F(X_t)$ and $G(Z_{t-1})$ respectively, the joint distribution of the random vector (Z_{t-1}, X_{t-1} , X_t) can be expressed by the copula method as:

$$P(Z_{t-1} \le z_{t-1}, X_{t-1} \le x_{t-1}, X_t \le x_t) = C(v_1, v_2, v_3; \alpha) \tag{4.6}$$

where α is the parameter that can be a scalar or vector depending on the copula family; v_1,v_2, and v_3 are the realizations of the random variables $V_1=G(Z_{t-1})$, $V_2=F(X_{t-1})$ and $V_3=F(X_t)$.The conditional distribution of monthly streamflow X_t given the previous monthly streamflow X_{t-1} and the aggregated streamflow Z_{t-1} can be expressed as:

$$\begin{aligned} &P(X_t \le x_t \,|\, Z_{t-1} = z_{t-1}, X_{t-1} = x_{t-1}) \\ &= C_{3|12}(v_3 \,|\, v_1, v_2) = \frac{\partial C^2(v_1, v_2, v_3; \alpha)}{\partial v_1 \partial v_2} \bullet \left[\frac{\partial C^2(v_1, v_2; \alpha)}{\partial v_1 \partial v_2} \right]^{-1} \end{aligned} \tag{4.7}$$

The cumulative distribution F_r, F_{r-1} and G_{r-1} [corresponding to v_3, v_2 and v_1 in equation (4.7)] where F_r (or F_{r-1}) and G_{r-1} are the CDFs of the monthly streamflow and the corresponding aggregate streamflow, are needed for the generation of streamflow of month r. The procedure for generating monthly streamflow by the EEC method is now summarized as follows:

(1) Assign random values to Z_{t-1} and X_{t-1}. Compute the corresponding cumulative probabilities $G_{t-1}(z_{t-1})$ (denoted as v_1) and $F_{t-1}(x_{t-1})$ (denoted as v_2).

(2) Generate a uniform random value η between [0, 1], which is considered to be the conditional cumulative probability corresponding to a specific value x_t, given the initial value z_{t-1} and x_{t-1} (or v_1 and v_2). From equation (4.7), one obtains: $C_3|_{1,2}(v_3|v_1,v_2)= \eta$. The cumulative probability $F_t(x_t)$ (denoted as v_3) can be obtained as: $v_3=C^{-1}{}_3|_{1,2}(\eta)$ and then x_t can be obtained as: $x_t=F_t^{-1}(v_3)$.

(3) Increase time step t and update the random values of X_{t-1} and Z_{t-1}.

(4) Repeat steps (1)-(3) until the required length of monthly streamflow is generated.

4.3 Application

Monthly streamflow of the Colorado River at Lees Ferry, Arizona from 1906-2003 was used for the application of the proposed method. More detail about the datasets are given by *Hao and Singh* [2011]. 100 flow sequences with 100 years of streamflow in each sequence were generated to assess the performance of the proposed method. The basic statistics (mean, standard deviation, skewness, lag-one correlation, maximum and minimum values), higher-order correlation and inter-annual statistics from generated

data were compared with those from the historical data using the box plots. The performance was considered to be good when a statistic fell in the box plot.

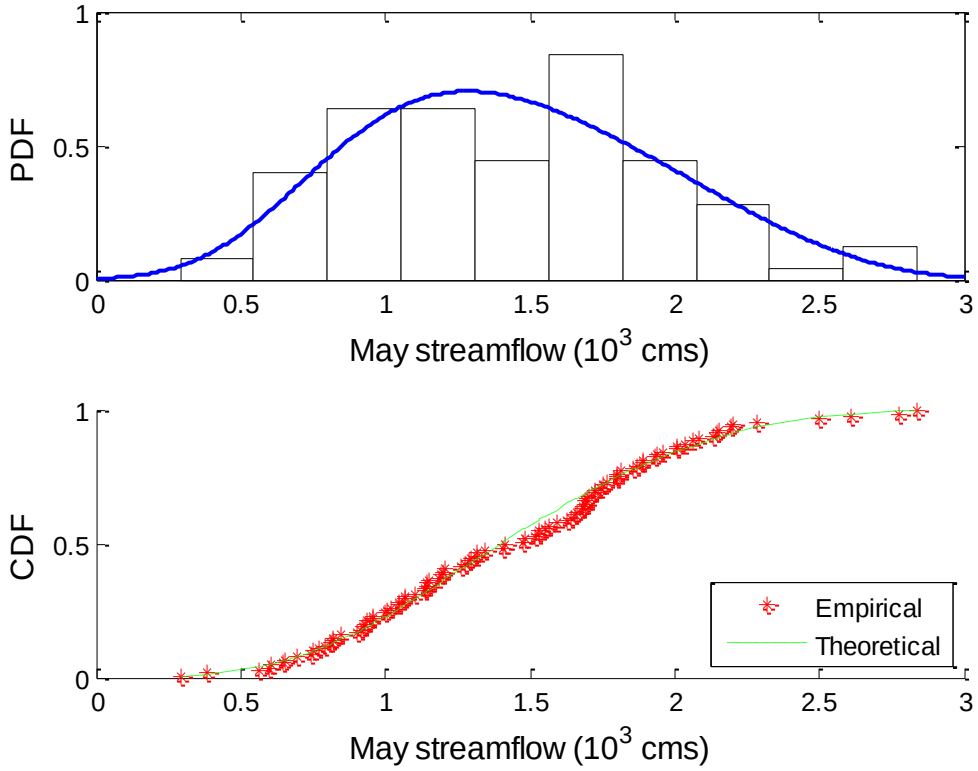

Figure 4. 1 Empirical and theoretical distribution for May streamflow. (a)Empirical histogram and theoretical PDF; (b) Empirical and theoretical CDF.

4.3.1 Marginal PDF

Maximum entropy-based marginal distributions for each month were compared with empirical histograms and the cumulative distributions estimated from the Gringorten plotting position formula. The empirical histograms and the theoretical

entropy-based PDFs for May and September streamflow were shown in Figure 4. 1 and Figure 4. 2, respectively. It can be seen that the entropy-based PDF fitted the empirical histogram well and theoretical CDF also fitted the empirical CDF well. Note that the skewness of September streamflow is relatively high (1.96). These results showed that the maximum entropy-based marginal distribution modeled the underlying streamflow well, even though high skewness was involved.

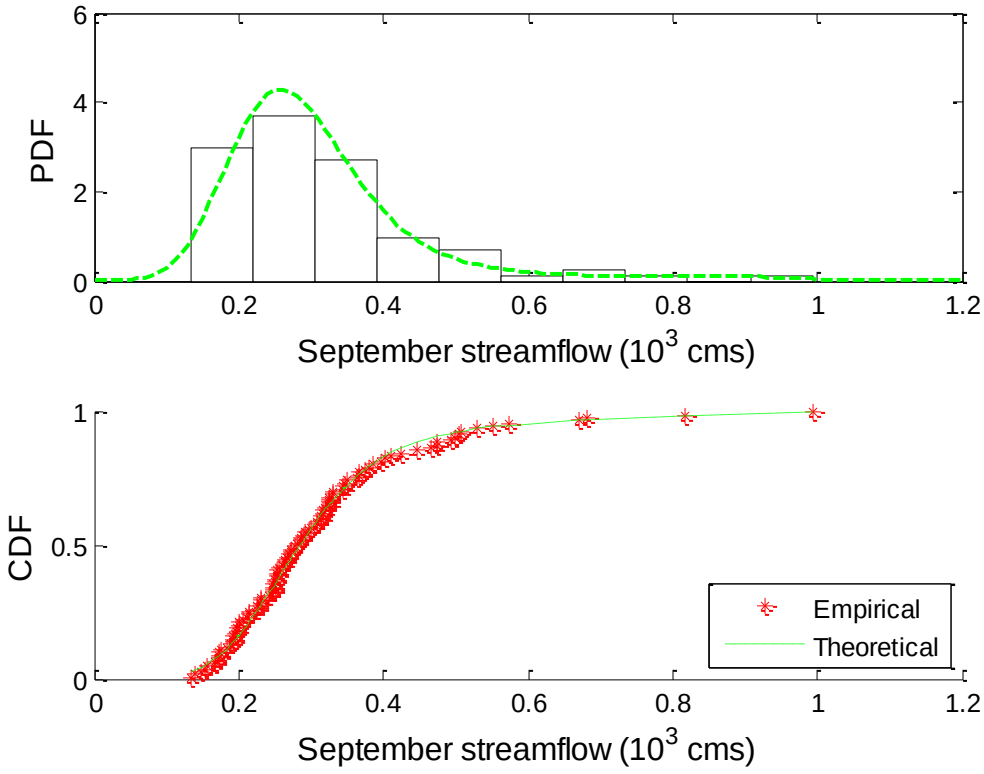

Figure 4. 2 Empirical and theoretical distribution for September streamflow. (a)Empirical histogram and theoretical PDF; (b) Empirical and theoretical CDF.

The bimodality in the PDF of the May streamflow, which has been found in several studies [*Prairie et al.*, 2006], was not resolved with the PDF in equation (4.1). However, this can be overcome when more moments are used to derive the maximum entropy-based distribution (not presented).

4.3.2 Copula selection

The Clayton, Frank, Gumbel and Gaussian copula were selected to construct the joint distribution. The suitability of different copulas was assessed with graphical method and goodness of fit test.

Graphical method

For the random samples $X=(x_1, x_2,\ldots, x_n)$ and $Y=(y_1, y_2,\ldots, y_n)$, the pseudo-observations $U=(u_1, u_2,\ldots, u_n)$ and $V=(v_1, v_2,\ldots, v_n)$ can be obtained as $U_i=R_i/(n+1)$ and $V_i=S_i/(n+1)$, $i=1, 2,\ldots, n$, where R_i and S_i are the ranks of the random samples X and Y, respectively.

The empirical copula $C(u_i,v_i)$, $i=1, 2,\ldots, n$, can be defined with the pseudo-observations U and V as:

$$C(u_i, v_i) = \frac{1}{n}\sum_{j=1}^{n} I\left(U_j \le u_i, V_j \le v_i\right) \tag{4.8}$$

The empirical Kendall distribution K_n of the empirical copula $w_i=C(u_i, v_i)$ can be defined as:

$$K_n(w) = \frac{1}{n}\sum_{i=1}^{n} I\left(w_i \le w\right) \tag{4.9}$$

For the continuous random vector (X, Y) with marginal cumulative distribution functions (CDF) $F_X(x)$ and $F_Y(y)$, the theoretical Kendall distribution function $K\theta(w)$ can be defined as:

$$K_\theta(w) = P(C_\theta(F_X(x), F_Y(y))) = P(C_\theta(U,V) \le w) \tag{4.10}$$

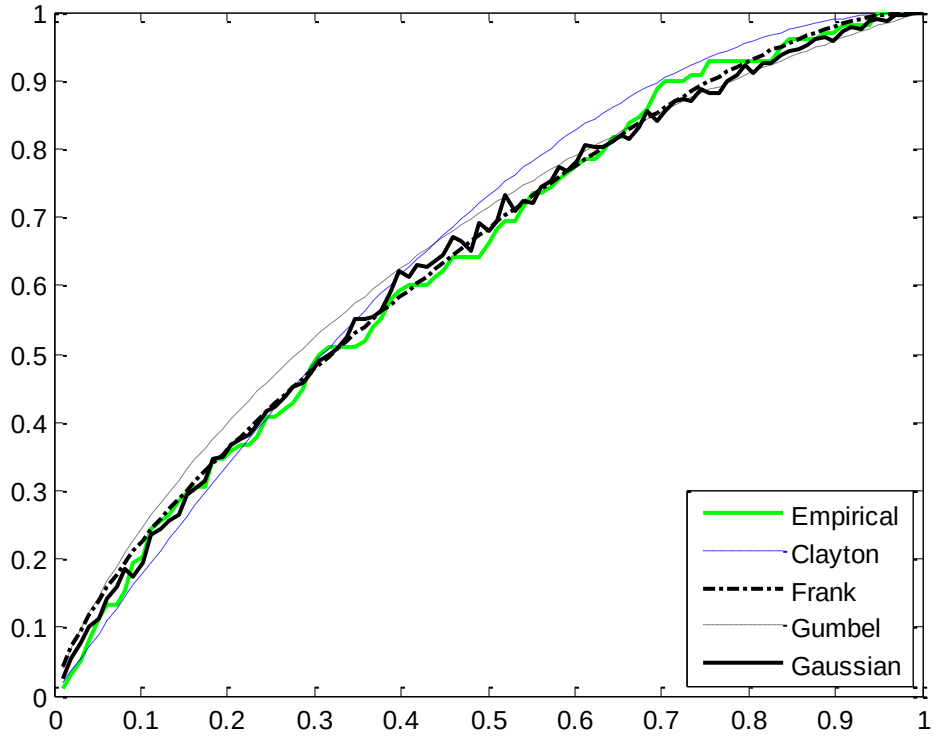

Figure 4. 3 K-Plot of different copulas for May-June streamflow pairs.

The graphical method is based on the comparison of the empirical distribution Kn(w) in equation (4.9) and the fitted distribution $K\theta(w)$ in equation (4.10) (or K-Plot).

The K-plot for the streamflow pairs of May-June and October-November are shown in Figure 4. 3 and Figure 4. 4. Generally there is not a single copula that fit the empirical copula better than other copulas for all streamflow pairs. It can be seen that the Frank copula seems to fit the empirical distribution better for most part then other copulas from Figure 4. 3 while the Gaussian copula seems to perform better from Figure 4. 4.

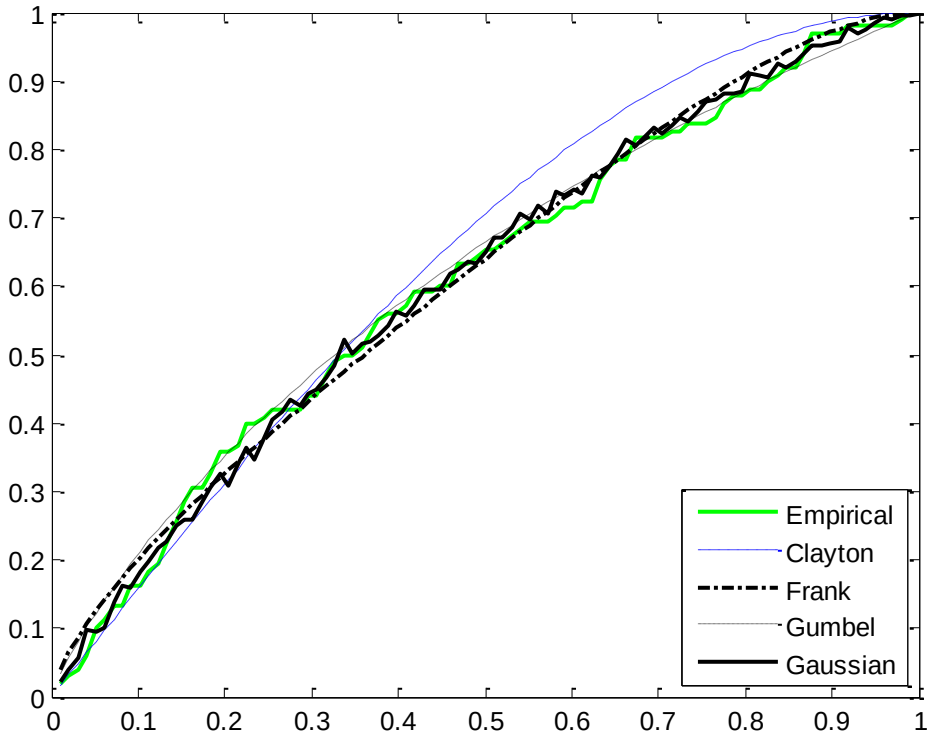

Figure 4. 4 K-Plot of different copulas for October-November streamflow pairs.

Goodness of fit test

The formal goodness of fit test was used to determine whether a specific copula is suitable to model the dependence structure of underlying data. The Cramér—von

Mises statistic (S_n) and Kolmogorov-Smirnov statistic (T_n) were used for the goodness of fit test for the four copulas [*Genest et al.*, 2006; *Genest and Favre*, 2007; *Genest et al.*, 2007]. These statistics S_n and T_n are the variants of those proposed by *Wang and Wells* [2000] and are based on the process:

$$\mathbf{K}n(w) = \sqrt{n}\left\{Kn(w) - K_{\theta_n}(w)\right\} \tag{4.11}$$

where $Kn(w)$ and $K\theta_n(w)$ are the empirical and theoretical probability distributions defined in equation (4.9) and (4.10). These two statistics are defined as:

$$\begin{aligned} S_n &= \int_0^1 \left|\mathbf{K}_n(w)\right|^2 k_{\theta_n}(w)dw \\ &= \frac{n}{3} + n\sum_{j=1}^{n-1} K^2 n\left(\frac{j}{n}\right)\left\{K_{\theta_n}\left(\frac{j+1}{n}\right) - K_{\theta_n}\left(\frac{j}{n}\right)\right\} \\ &- n\sum_{j=1}^{n-1} Kn\left(\frac{j}{n}\right)\left\{K^2{}_{\theta_n}\left(\frac{j+1}{n}\right) - K^2{}_{\theta_n}\left(\frac{j}{n}\right)\right\} \end{aligned} \tag{4.12}$$

$$T_n = \sup_{0\le w\le 1}\left\{\mathbf{K}_n(w)\right\} = \sqrt{n}\max_{i=0,0\le j\le n-1}\left\{K_{\theta_n}\left(\frac{j}{n}\right) - K_{\theta_n}\left(\frac{j+i}{n}\right)\right\} \tag{4.13}$$

The *p*-values of statistics (S_n and T_n) at the 5% significance level based on a run of 5000 samples were obtained using the parametric bootstrap procedure [*Genest et al.*, 2006]. The results for the statistic S_n and the associated *p*-value are shown in Table 4. 2. The very low *p*-value (<5%) in Table 4. 2 signified that the null hypothesis the copula was a valid model should be rejected. Take the streamflow pairs of July-August for example. The Clayton and Gumble copulas were rejected since the *p*-values were lower than 5% while the Frank and Gaussian copulas were valid models. The number of streamflow pairs that a copula was rejected for the Clayton, Frank, Gumbel and Gaussian was 6, 4, 6 and 2, respectively.

Table 4. 2 Statistics S_n and associated *p*-values for different streamflow pairs.

Copula	S_n and *p*-value	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-1
Clayton	S_n	0.25	0.19	0.09	0.06	0.13	0.11	0.15	0.44	0.17	0.22	0.29	0.15
	p-value	0.01	0.05	0.32	0.65	0.13	0.09	0.03	0.00	0.04	0.01	0.00	0.10
Frank	S_n	0.05	0.08	0.33	0.25	0.06	0.07	0.08	0.12	0.11	0.07	0.13	0.15
	p-value	0.72	0.31	0.00	0.00	0.45	0.16	0.13	0.05	0.07	0.30	0.03	0.01
Gumbel	S_n	0.06	0.22	0.51	0.54	0.21	0.14	0.19	0.11	0.16	0.08	0.06	0.10
	p-value	0.64	0.02	0.00	0.00	0.01	0.01	0.00	0.15	0.04	0.30	0.52	0.13
Gaussian	S_n	0.06	0.12	0.22	0.30	0.04	0.04	0.07	0.14	0.08	0.03	0.08	0.08
	p-value	0.69	0.15	0.01	0.00	0.83	0.62	0.24	0.06	0.35	0.94	0.24	0.24

The results for the statistic T_n and the associated *p*-value are shown in Table 4. 3. The number of month pairs of rejecting the copula was 6, 2, 4 and 2, respectively. Generally there was not a single copula that is valid for modeling all streamflow pairs. The Gaussian copula seems to preforms slightly better than other copulas from the number of rejections.

Table 4. 3 Statistics T_n and associated *p*-values for different streamflow pairs.

Copula	T_n and *p*- value	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-1
Clayton	T_n	1.06	0.92	0.70	0.56	0.87	0.83	0.96	1.30	1.12	1.10	1.39	0.88
	p-value	0.03	0.12	0.47	0.80	0.15	0.11	0.04	0.00	0.01	0.02	0.00	0.15
Frank	T_n	0.56	0.73	1.22	1.09	0.67	0.64	0.62	0.74	0.74	0.66	0.80	0.84
	p-value	0.75	0.26	0.00	0.00	0.39	0.25	0.37	0.22	0.22	0.36	0.08	0.05
Gumbel	T_n	0.52	1.03	1.52	1.34	0.79	0.82	0.94	0.77	0.86	0.65	0.56	0.65
	p-value	0.88	0.04	0.00	0.00	0.22	0.06	0.02	0.25	0.11	0.46	0.71	0.45
Gaussian	T_n	0.63	0.86	1.10	1.10	0.51	0.56	0.62	0.71	0.74	0.47	0.68	0.65
	p-value	0.61	0.15	0.01	0.01	0.89	0.61	0.47	0.37	0.30	0.93	0.35	0.44

A practical way for the simulation of the monthly streamflow may be to choose different copulas in modeling different streamflow pairs. In this study, the Gaussian copula was selected hereinafter for the illustration of the proposed entropy-copula method for monthly streamflow simulation. The entropy-copula (EC) method and extended entropy-copula (EEC) method with the Gaussian copula were denoted as ECG and EECG method.

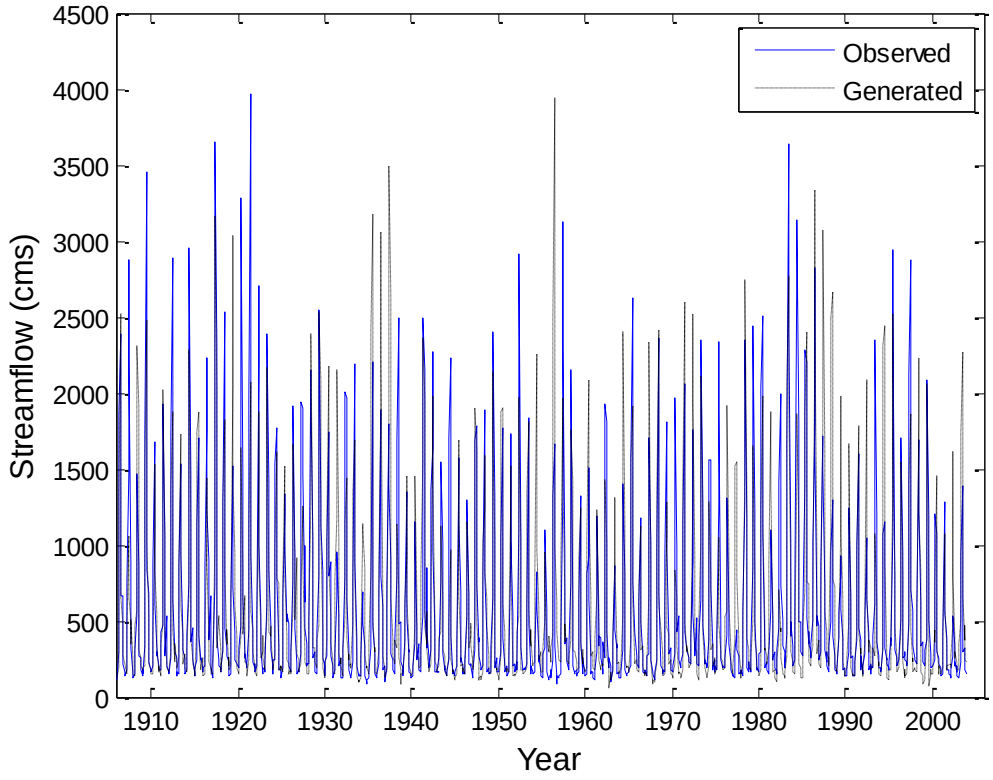

Figure 4. 5 Comparison of observed monthly streamflow and a sequence of generated monthly streamflow. (Unites for streamflow are in cubic meters per second (cms).)

4.3.3 Variability of streamflow

The variability of the generated streamflow was first compared with the historical streamflow. The plot for observed streamflow and a sequence of simulated monthly streamflow for 98 years from the ECG method is shown in Figure 4. 5. Generally the variability of generated streamflow is similar to that of the observed streamflow. The maximum and minimum values of simulated streamflow matched that of observed streamflow. Similar results were obtained from the generated streamflow from the EECG method (not shown).

4.3.4 Basic statistics

The observed statistics and the median values of the generated basic statistics from the ECG method are shown in Figure 4. 6. The relative error (RE) is also used for assessing the performance, which is defined as RE = $(S_m - X_o)/X_o$, where S_m is the median of simulated statistic and X_o is the observed statistic. The relative error for each statistic, including the mean (Mn), standard deviation (Sd), skewness (Sk), lag-one correlation (L1), maximum values (Max) and minimum values (Min) , is shown in Table 4. 4.

The ECG method performed well in preserving the mean, standard deviation, and skewness, since all the statistics fell in the box plot. The RE for mean and standard deviation was under 5% and that for skewness was under 10% for all months. Generally the lag-one correlation was preserved well, though for certain months, such as October (with RE 14.4%), the observed statistics did not fall in the box.

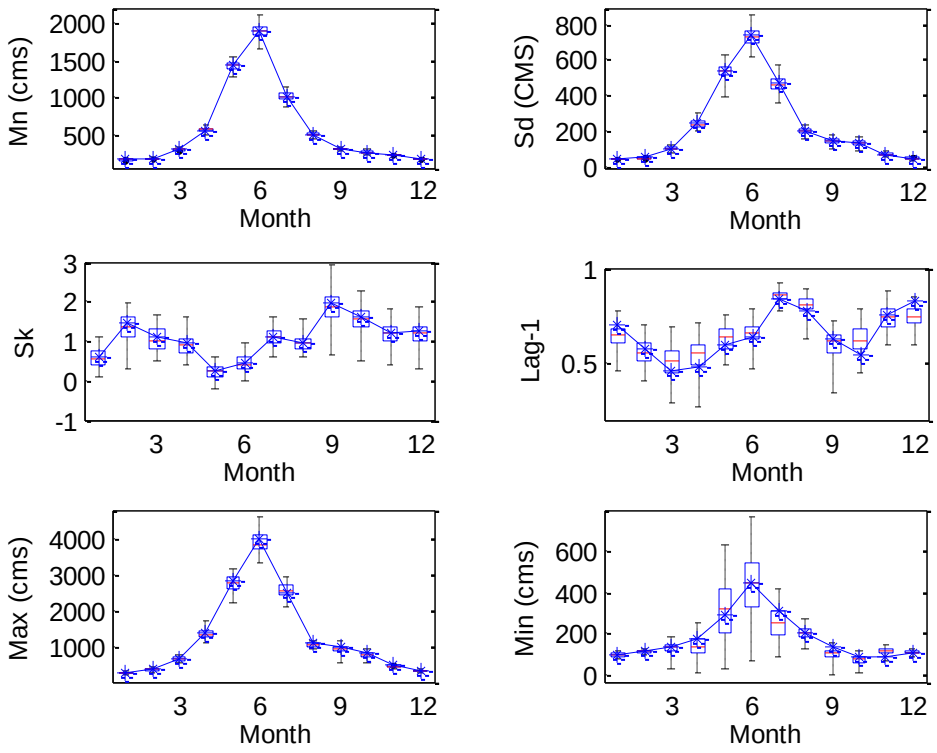

Figure 4. 6 Boxplots of basic statistics of generated and historical monthly streamflow from the ECG method.

Table 4. 4 Relative error (%) for simulated statistics of each month.

statistics	1	2	3	4	5	6	7	8	9	10	11	12
Mn	-0.2	-0.2	0.1	0.5	-0.2	-0.1	-0.1	-0.6	-0.5	-0.4	-0.3	-0.5
Sd	-1.6	-2.2	-1.0	-2.0	-0.7	-1.9	-1.1	-1.3	-1.4	-1.8	-3.9	-3.8
Sk	-0.8	-5.8	-5.7	-4.3	1.8	-4.5	1.7	-2.4	-4.5	-3.3	-2.3	-5.9
L1	-7.0	-3.0	12.0	13.7	7.0	4.3	2.3	3.8	-1.8	14.4	-0.9	-10.2
Max	-2.2	-1.5	-0.6	-3.3	-2.7	-2.9	3.5	-6.1	-6.4	-10.3	-3.6	-3.2
Min	4.2	-2.9	1.4	-18.3	8.3	0.9	-16.9	-2.1	-23.4	-18.4	38.0	1.4

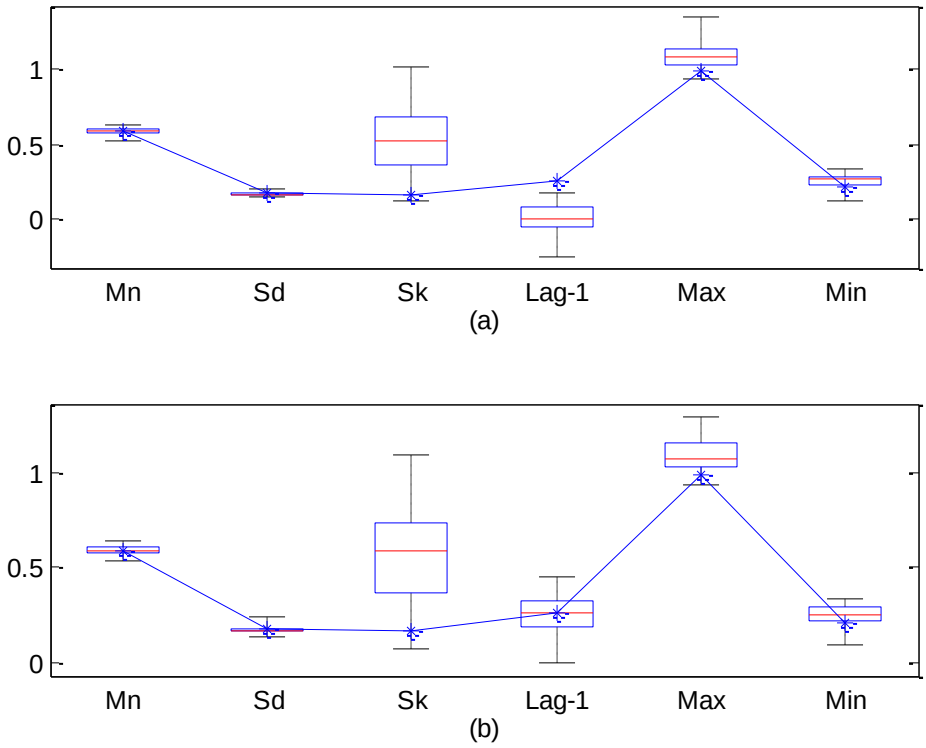

Figure 4. 7 Boxplots of basic statistics of generated and historical annual streamflow from two methods. (a) ECG method; (b) EECG method. (Mn, Sd, Sk,Lag-1, Max and Min represent the mean, standard deviation, skewness, lag-one correlation, maximum and minimum values, respectively. Unites for Mn, Sd, Max and Min are in 10^3 cms.)

The maximum and minimum values were generally preserved well for most months (with RE under 5% and 20% for maximum and minimum values for most months), though over-estimation or under-estimation for certain months occurred. Results of the EECG method in preserving these statistics of each month were similar to those from the ECG method and thus are not shown.

To assess the performance in the preservation of these statistics at the annual level, the annual streamflow is obtained by adding the generated monthly streamflow for each year. Generated and observed basic statistics of annual streamflow from the ECG and EECG methods are shown in Figure 4. 7. The mean and standard deviation were preserved well by both methods. Neither ECG nor EECG method preserved the skewness well. However, the two methods differed significantly in preserving the lag-one correlation. The EECG method preserved the lag-one correlation well, while the ECG method did not perform as well. For the preservation of maximum and minimum values, the ECG and EECG performed relatively well, though some overestimation occurred in the maximum values.

4.3.5 Higher-order correlation

Sharma and O'Neill [2002] proposed a nonparametric method for monthly streamflow simulation for preserving long-term dependence (denoted as NPL) and showed that the NPL model improved higher-order correlation of monthly streamflows. To assess the performance of the proposed ECG and EECG method in preserving the long-term dependence, a relatively higher order correlation (lag-four in this study) was selected for comparison. Boxplots of lag-four correlations of observed and generated and monthly streamflow from the ECG and EECG methods are shown in Figure 4. 8. In general the ECG method did not preserve these higher order correlations well. This is not unexpected, since only lag-one correlation was included in the ECG method. Generally the median values of generated statistics were closer to the observed statistics for the EECG method than those for the ECG method, as seen in Figure 4. 8 (b). Thus,

the preservation of the higher order correlation can be improved by EECG method, although the higher order correlation was not included directly.

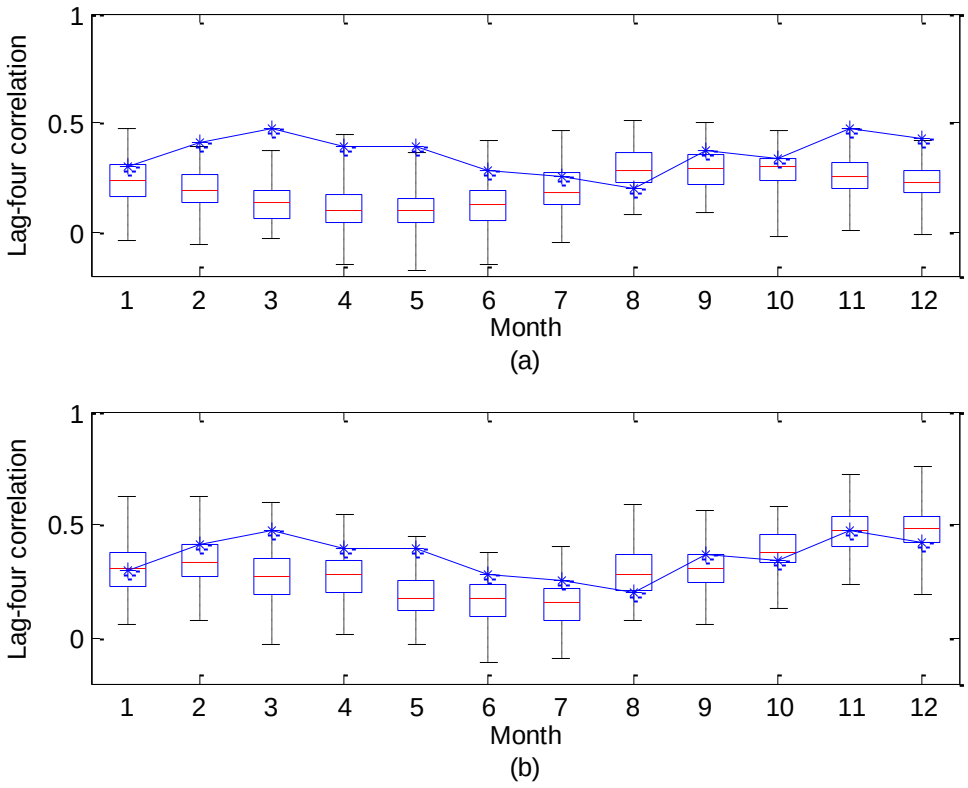

Figure 4. 8 Boxplots of lag-four correlation of generated and historical monthly streamflow from two methods. (a)ECG method; (b) EECG method.

4.3.6 Inter-annual statistics

The inter-annual dependence between the streamflows of seasonal and annual timescales was also assessed for the EECG method. Box plots of lag-one and lag-four correlation between streamflow of a specific month (seasonal time scale) and

streamflows of the previous 12 months (annual time scale) of the generated and historical data are shown in Figure 4. 9 (a) and (b). It is seen that the lag-one correlation was preserved well for all months except for February, as expected from the structure of the EECG model. The lag-four correlation, although not directly included, was also preserved well as shown in Figure 4. 9(b).

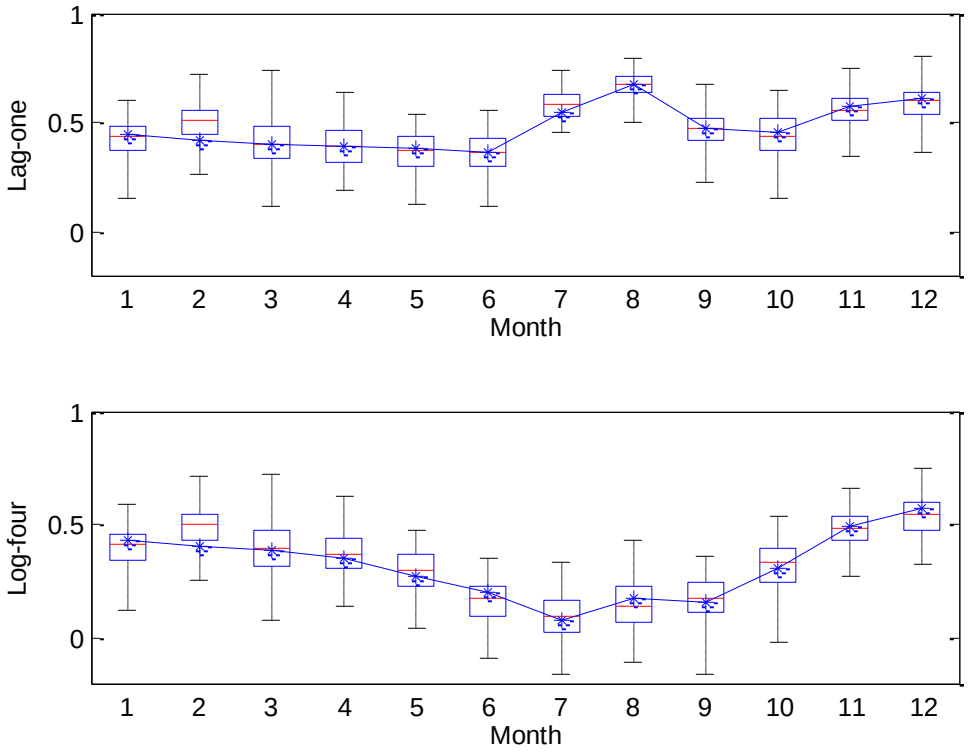

Figure 4. 9 Boxplots of inter-annual dependence of generated and historical monthly streamflow from the EECG method. (a) lag-one; (b) lag-four.

4.4 Conclusion

An entropy-copula method is proposed for single-site monthly streamflow simulation and is shown to preserve statistics of monthly streamflow well. The mean and standard deviation of the generated streamflow at the annual scale can also be preserved well. The extended entropy-copula method is shown to preserve inter-annual dependence well and improve the higher-order correlation of monthly streamflow. The preservation of the lag-one correlation at the annual scale can also be improved by the extended method. The marginal distribution derived with the entropy theory with the first four moments as constraints is capable of modeling the complex properties (such as high skewness) of the underlying streamflow data.

The possible limitation of the entropy-copula method may be that many Lagrange multipliers may be needed for modeling certain properties (e.g., multi-mode in the distribution) of the underlying data. The entropy-copula framework can be applied and extended to higher dimensions for hydrologic modeling or simulation with the copula to model the dependence structure and entropy-based marginal distributions to model the underlying data in the univariate case.

CHAPTER V

MULTI-SITE ANNUAL STREAMFLOW SIMULATION WITH ENTROPY AND COPULA METHODS

5.1 Introduction

For streamflow simulation in a river basin, it is desired that statistical properties of streamflow at an individual site and dependence properties of streamflow among different sites are preserved. The autoregressive moving average (ARMA) framework has been commonly used for multi-site streamflow simulation, in which the streamflow series are converted to a sequence of normally distributed random variables [*Loucks et al.*, 1981]. The general model for the streamflow simualtion at *n* sites for any season *t* with lag-one correlation can be expressed as [*Matalas*, 1967; *Finzi et al*., 1975; *Salas and Delleur*, 1980]:

$$Z_t = AZ_{t-1} + B\varepsilon_\tau \tag{5.1}$$

where $\mathbf{Z}_t=[Z^1_t,\ldots, Z^n t,]^T$ and $\mathbf{Z}_{t\text{-}1} =[Z^1_{t\text{-}1},\ldots, Z^n_{t\text{-}1}]^T$ are the vectors of standardized flows (e.g., annual) with length *n* as the number of sites; $\boldsymbol{\mathcal{E}}_t=[\,\mathcal{E}^1_t,\ldots, \mathcal{E}^n_t]^T$ is the vector of normal random variables that are independent of $\boldsymbol{Z}$; *A* and *B* are the parametric matrices expressed as:

$$A=\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \qquad B=\begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{n1} & b_{n2} & \ldots & b_{nn} \end{pmatrix}$$

where the elements $a_{i,j}$ or $b_{i,j}$ specify the temporal and spatial dependence. The multisite streamflow of different seasons can be generated sequentially with the increase of t by using equation (5.1).

Parametric and nonparametric disaggregation methods have also been used for multi-site streamflow simulation. A parametric model for temporal disaggregation of annual to monthly streamflow for n sites can be expressed as [*Valencia and Schaake*, 1973; *Loucks et al*., 1981]:

$$X_t = CY_t + DV_t \tag{5.2}$$

where $Y_t = [Y^1_t, \ldots, Y^n_t]^T$ is the vector of the aggregated streamflow (annual) in the year t; $X_t = [X^1_{1,t}, \ldots, X^1_{s,t}, \ldots, X^n_{1,t}, \ldots, X^n_{s,t}]^T$ is the vector of the disaggregated streamflow (seasonal) for the season s ($s \leq m$) in the year t; $V_t = [V^1_t, \ldots, V^n_{mt}]^T$ is a vector of nm normal random variables; C and D are the matrices expressed as:

$$C = \begin{pmatrix} c_{1,1} & c_{1,2} & \ldots & c_{1,n} \\ c_{2,1} & c_{2,2} & \ldots & c_{2,n} \\ \ldots & \ldots & \ldots & \ldots \\ c_{nm,1} & c_{nm,2} & \ldots & c_{nm,n} \end{pmatrix} \qquad D = \begin{pmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,nm} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,nm} \\ \ldots & \ldots & \ldots & \ldots \\ d_{nm,1} & d_{nm,2} & \ldots & b_{nm,nm} \end{pmatrix}$$

where the elements $c_{i,j}$ or $d_{i,j}$ specify the dependence of the aggregated and disaggregated streamflows. This method was first developed by *Valencia and Schaake* [1973] and then modified to reduce parameters [*Mejia and Rousselle*, 1976; *Santos and Salas*, 1992]. Likewise, the spatial disaggregation of the key station streamflow (aggregated streamflow) to the substation streamflow (disaggregated streamflow) can also be carried out with the model in equation (5.2).

The nonparametric model for disaggregation is based on the kernel density estimate or re-sampling techniques. A typical characteristic of the nonparametric model is that it does not make assumptions about the dependence structure or probability density function of the underlying data. For nonparametric disaggregation, streamflow simulation is cast as a sampling from the conditional probability density function [*Tarboton et al.*, 1998]:

$$f(X|Z) = f(X,Z) / \int f(X,Z)dX \tag{5.3}$$

where X is the disaggregated variable (monthly or tributary streamflow); Z is the aggregate variable (annual or main streamflow); and $f(X, Z)$ is the joint probability density function that can be obtained from the kernel density estimation. To overcome the difficulty that the simulation in higher dimension is inefficient and cumbersome for the kernel method, *Prairie et al.* [2007] developed a nonparametric disaggregation method with *K*-Nearest Neighbor based bootstrap for streamflow simulation from the conditional distribution. *Lee et al.* [2010] proposed another nonparametric space-time disaggregation method that further improves the method by *Prairie et al.* [2007] . *Srinivas and Srinivasan* [2005] proposed a hybrid moving block bootstrap method for multi-site streamflow simulation that incorporates the properties of parametric and nonparametric method. In this method, the parametric model (e.g., autoregressive model) is fitted to the data and then the residuals are resampled with block bootstrapping.

Recently, a number of new approaches were proposed for streamflow simulation. *Nowak et al.* [2011] proposed a wavelet auto-regressive method for multi-site annual streamflow simulation that is capable of capturing spectral and distributional properties

of the streamflow data. *Hao and Singh* [2011] introduced the entropy theory for single-site monthly streamflow simulation, in which the required statistics are used as constraints to derive the joint distribution with the entropy theory. The copula theory has been used extensively in hydrology to model the joint distribution of hydrologic variables. *Lee and Salas* [2011] proposed the copula method for annual streamflow simulation of the Nile River in which the joint distribution of streamflow is derived with the copula. *Hao and Singh* [2012] proposed an entropy-copula method for single site monthly streamflow simulation of the Colorado River at Lees Ferry, Arizona , in which the joint distribution is constructed using the copula method with the marginal distributions derived using the entropy method.

This study extends the entropy method and entropy-copula method for multi-site annual streamflow simulation. The joint distributions of annual streamflow at different sites are constructed with the entropy method and the entropy-copula method and then annual streamflow at different sites is generated from the conditional distribution. The proposed methods are applied to the annual streamflow simulation of four sites in Colorado River basin and the performance of the two methods in preserving the statistics of annual streamflow are compared.

5.2 Method

5.2.1 Entropy method

For continuous random variables X_1 and X_2 with joint probability density function $f(x_1, x_2)$ defined over the space $[a, b]\times[c, d]$, the Shannon entropy can be defined as [*Shannon*, 1948; *Shannon and Weaver*, 1949]:

$$E = -\int_c^d \int_a^b f(x_1, x_2) \ln f(x_1, x_2)\, dx_1 dx_2 \tag{5.4}$$

According to the principle of maximum entropy [*Jaynes*, 1957], the joint probability density function $f(x_1, x_2)$ with the maximum entropy should be selected subject to the constraints (or known information).

Hao and Singh [2011] proposed the entropy method for single-site monthly streamflow simulation in which the joint distribution $f(x_1, x_2)$ of monthly streamflow of adjacent months is derived using the principle of maximum entropy. The statistics to be preserved can be used as constraints to derive the joint distribution $f(x_1, x_2)$. This method can also be used for streamflow simulation at two sites. For preserving the moment statistics (mean, standard deviation, and skewness), the first three moments of X and Y can be used as constraints while the cross product XY can be used as a constraint for preserving the lag-one correlation. The fourth moment can also be used as the constraint to characterize the distribution property of the underlying data. The general form of the constraints can be specified as:

$$\int_c^d \int_a^b g_i(x_1, x_2) f(x_1, x_2) dx_1 dx_2 = E(g_i) \quad i=0, 1, 2,\ldots, m \tag{5.5}$$

where $g_i(x_1, x_2)$ is a function of random vector (X, Y); $E(g_i)$ is the expected value of the function $g_i(x_1, x_2)$. The maximum entropy distribution $f(x_1, x_2)$ can then be obtained by maximizing the entropy in equation (5.4) subject to the constraints in equation (5.5) as [*Kesavan and Kapur*, 1992]:

$$f(x_1, x_2) = \exp\left[-\sum_{i=0}^{m} \lambda_i g_i(x_1, x_2)\right]$$
$$= \exp(-\lambda_0 - \sum_{i=1}^{4} \lambda_i {x_1}^i - \sum_{i=1}^{4} \lambda_i {x_2}^i - w x_1 x_2) \quad (5.6)$$

where λ and w (λ_m) are the Lagrange multipliers corresponding to the marginal and joint constraints, respectively; m is the number of constraints (m=9 for the bivariate case).

The zeroth Lagrangre multiplier λ_0 can be expressed as a function of other Lagrange multipliers as:

$$\exp(\lambda_0) = \int_c^d \int_a^b \exp\left[-\sum_{i=1}^{m} \lambda_i g_i(x, y)\right] dx dy \quad (5.7)$$

From the joint distribution in equation (5.6), the marginal distribution of random variable X_1 and X_2 can be derived as:

$$f(x) = \int_c^d f(x_1, x_2) dx_2$$
$$= \exp(-\sum_{i=1}^{4} \lambda_i {x_1}^i - \lambda_0) \int_c^d \exp(-\lambda_5 x_2 - \lambda_6 {x_2}^2 - \lambda_7 {x_2}^3 - \lambda_8 {x_2}^4 - w x_1 x_2) dx_2 \quad (5.8)$$

The conditional distribution $f(x_2|x_1)$ can then be derived accordingly as:

$$f(x_2|x_1) = \frac{\exp\left(-\lambda_5 x_2 - \lambda_6 {x_2}^2 - \lambda_7 {x_2}^3 - \lambda_8 {x_2}^4 - w x_1 x_2\right)}{\int_c^d \exp\left(-\lambda_5 x_2 - \lambda_6 {x_2}^2 - \lambda_7 {x_2}^3 - \lambda_8 {x_2}^4 - w x_1 x_2\right) dx_2} \quad (5.9)$$

The annual streamflow of two sites can then be generated sequentially from the conditional density function in equation (5.9).

The entropy method can be extended for multi-site streamflow simulation. Denote the joint distribution of annual streamflow at sites 1, 2 and 3 as $f(x_1, x_2, x_3)$

defined on the interval $[a, b] \times [c, d] \times [e, f]$. The entropy E' of the joint distribution $f(x_1, x_2, x_3)$ can be defined as:

$$E' = -\int_e^f \int_c^d \int_a^b f(x_1, x_2, x_3) \ln f(x_1, x_2, x_3)\, dx_1 dx_2 dx_3 \tag{5.10}$$

The first three moments are used as separate constraints to preserve the mean, standard deviation and skewenss of steamflow at each site. The joint constraints can be specified as the cross product between any two random variables (namely X_1X_2, X_1X_3 and X_2X_3). The maximum entropy based distribution $f(x_1, x_2, x_3)$ can be derived with the separate and joint constraints by maximizing the entropy in equation (5.10) expressed as:

$$f(x_1, x_2, x_3) = \exp\left(-\lambda_0 - \sum_{i=1}^{4} \lambda_i x^i{}_1 - \sum_{i=1}^{4} \lambda_{i+4} x^i{}_2 - \sum_{i=1}^{4} \lambda_{i+8} x^i{}_3 - w_{12}x_1x_2 - w_{13}x_1x_3 - w_{23}x_2x_3 \right) \tag{5.11}$$

where λ and w are the Lagrange multipliers corresponding to the separate and joint constraints, respectively. The conditional distribution $f(x_3|x_1, x_2)$, which can be used for generation of the streamflow at site 3 given the streamflow at site 1 and 2, can be derived from equation (5.11) as:

$$f(x_3 \mid x_1, x_2) = \frac{\exp\left(-\lambda_9 x_3 - \lambda_{10} x_3{}^2 - \lambda_{11} x_3{}^3 - \lambda_{12} x_3{}^4 - w_{13}x_1x_3 - w_{23}x_2, x_3\right)}{\int \exp\left(-\lambda_9 x_3 - \lambda_{10} x_3{}^2 - \lambda_{11} x_3{}^3 - \lambda_{12} x_3{}^4 - w_{23}x_2, x_3\right) dx_3} \tag{5.12}$$

Similarly, the joint distribution streamflow at n sites ($X_1, X_2, \ldots, X_n$) can be expressed as:

$$f(x_1, x_2, ..., x_n) = \exp\left(-\lambda_0 - \sum_{i=1}^{4} \lambda_i x_1 - \sum_{i=1}^{4} \lambda_{i+4} x^i{}_2 - ... - \sum_{i=1}^{4} \lambda_{i+4(n-1)} x^i{}_n - \sum_{1 \le i < j \le n} w_{i,j} x_i x_j \right) \tag{5.13}$$

The conditional distribution f($x_n|x_1, x_2,\ldots, x_{n-1}$), which can be used for generation of streamflow at site n (X_n) given the streamflow at other n-1 sites ($X_1, X_2,\ldots, X_{n-1}$), can be derived as

$$f(x_n \mid x_1, x_2, ..., x_{n-1}) = \frac{\exp\left(-\sum_{i=1}^{4} \lambda_{i+4(n-1)} x^i{}_n - \sum_{i=1}^{n-1} w_{i,n} x_i x_n\right)}{\int \exp\left(-\sum_{i=1}^{4} \lambda_{i+4(n-1)} x^i{}_n - \sum_{i=1}^{n-1} w_{i,n} x_i x_n\right) dx_n} \tag{5.14}$$

5.2.2 Entropy-copula method

For continuous random variables X and Y with the cumulative distribution functions (CDF) $F_X(x)$ and $F_Y(y)$, the bivariate probability distribution function can be expressed with copula C as [*Nelsen*, 2006; *Salvadori*, 2007]:

$$P(X \le x, Y \le y) = C(u, v; \theta) \tag{5.15}$$

where u and v are realizations of the random variables $U= F_X(x)$ and $V=F_Y(y)$; θ is the copula parameter that measures the dependence between marginals. The copula C maps the two marginal distributions into the joint distribution as $[0,1]^2 \rightarrow [0,1]$. A number of copula families have been developed, such as the elliptical copula (Gaussian and *t*), Archimedean copula (Clayton, Gumbel, Frank and Ali-Mikhail-Haq) and other copula families [*Nelsen*, 2006]. The conditional distribution can be obtained from the joint distribution in equation (5.15) as:

$$P(Y \le y \mid X = x) = \frac{\partial C(u, v; \theta)}{\partial u} \tag{5.16}$$

Hao and Singh [2012] proposed the entropy-copula for single site monthly streamflow simulation with the marginal distribution derived with entropy method. In

this study, the first four moments can be used as constraints to derive the entropy-based marginal distribution expressed as:

$$f(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3 - \lambda_4 x^4) \tag{5.17}$$

The monthly streamflow is then generated sequentially by using equation (5.16) with the entropy-based distribution as the marginal distribution (5.17). The mean, standard deviation, and skewness can be expected to be preserved with the marginal distribution in equation (5.17).

The entropy-copula method can also be extended for the simulation of annual streamflow at multi-sites. The joint distribution of annual streamflow at sites 1, 2 and 3 with a copula C can be expressed as:

$$F(X_1, X_2, X_3) = C[F_1(x_1), F_2(x_2), F_3(x_3); \alpha] \tag{5.18}$$

where $F_1(x_1)$, $F_2(x_2)$, and $F_3(x_3)$ are the cumulative distribution of annual streamflow at site 1, 2 and 3; α is the parameter of the copula.

The conditional distribution of X_3 given X_1 and X_2 which can be used for the generation of the streamflow at site 3 given that from site 1 and site 2, can be derived as:

$$P(X_3 \le x_3 | X_2 = x_2, X_1 = x_1) = \frac{\partial C^2(v_1, v_2, v_3; \alpha)}{\partial v_1 \partial v_2} \bullet \left[\frac{\partial C^2(v_1, v_2; \alpha)}{\partial v_1 \partial v_2} \right]^{-1} \tag{5.19}$$

where v_1, v_2, and v_3 are the realizations of the random variables $V_1=F_1(x_1)$, $V_2=F_2(x_2)$, and $V_3=F_3(x_3)$. Similarly, the joint distribution for the simulation of streamflow at n sites (X_1, X_2,…, X_n) can be expressed with copula C as:

$$F(X_1, X_2, ..., X_n) = C[F_1(x_1), F_2(x_2), ..., F_n(x_n); \beta] \tag{5.20}$$

where $F_n(x_n)$ is the cumulative distribution function estimated from the entropy method of the annual streamflow at sites n; β is the parameter of the copula.

A specific copula has to be employed to construct the multivariate distribution in equation (5.15), (5.16), (5.18) and (5.20). The Gaussian copula has the property that it is easily extended to higher dimension [*Clemen and Reilly*, 1999; *Schölzel and Friederichs*, 2008]. The multivariate Gaussian copula is therefore used here to illustrate the application of the proposed method for annual streamflow simulation.

The Gaussian copula with marginal CDFs U_1, U_2,…, U_n can be defined as:

$$C(u_1, u_2, ..., u_n) = \Phi_n\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_n)|\Sigma\right) \tag{5.21}$$

where Φ is the CDF of the standard normal distribution function and Φ_n is the CDF of a multivariate normal distribution function with mean 0 and $n \times n$ covariance matrix $\sum$ whose i, j entry is corr($\Phi^{-1}(u_i)$, $\Phi^{-1}(u_j)$). The multi-variate cumulative distribution of the annual streamflow data at n sites can then be expressed as:

$$F(x_1, x_2, ..., x_n) = \Phi_n\left(\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_2(x_2)), ..., \Phi^{-1}(F_n(x_n))\right) \tag{5.22}$$

where $F_1(x_1)$, $F_2(x_2)$,…, $F_n(x_n)$ are the cumulative distribution of X_1, X_2,…, X_n estimated from the entropy based marginal density function in equation (5.17).

5.2.3 Comparison of two methods

For multi-site streamflow simulation, the cross correlation between annual streamflow of different site has to be preserved, apart from the preservation of individual statistics of each site. Accordingly, the joint distribution in higher dimension is needed for the entropy method and entropy-copula method. In the entropy method, all these required statistics can be used as constraints to derive the maximum entropy joint

distribution in higher dimension and then used for streamflow simulation. In the copula method, a copula in higher dimension is used, while the entropy based distribution in equation (5.17) can be used as marginal distributions.

The entropy method and entropy-copula method are similar in modeling the streamflow at a single site. In other words, the first four moments are used as constraints to derive the marginal or joint distribution. The difference between the entropy method and entropy-copula method for annual streamflow simulation lies in modeling the dependence structure of streamflow at different sites. For the entropy method, the joint constraints in the form of X_iX_j $(1 \leq i < j \leq n)$, are used to characterize the dependence with the random variables. In the entropy-copula method, the copula $C(F_1, F_2,\ldots, F_n)$ is used to model the dependence with the cumulative distribution function of random variables.

5.2.4 Simulation methodology

Suppose the stations from upstream to downstream are denoted as 1, 2,…, n with the corresponding annual streamflow donoted as $X_1, X_2,\ldots, X_n$. The generation of multi-site annual streamflow can be performed based on the conditional distribution (or the joint distribution). Both the entropy and entropy-copula method can be used for the construction of the joint distribution $P(X_1, X_2)$, $P(X_1, X_2, X_3)$, …, $P(X_1, X_2,\ldots, X_n)$ (likewise for $P(X_2|X_1)$, $P(X_3|X_1, X_2)$, …, and $P(X_n|X_1, X_2,\ldots, X_{n-1})$ for the simulation of X_2, X_3 and X_n, respectively. The simulation methodology can be summarized as follows:

(1) Initialize the annual streamflow at site 1 in the first year, i.e., x^1_1, by sampling from the marginal distribution $F_1(x)$ or assigning random values from the historical record.

(2) Generate the annual streamflow at site 2 in the first year, i.e., x^2_1, with the conditional distribution $P(X_2|X_1)$. Generate annual streamflow at site 3 in the first year, i.e., x^3_1 with the conditional distribution $P(X_3|X_1, X_2)$. Similarly, annual streamflow at site n in the first year, i.e.,x^n_1, can be generated with the conditional distribution $P(X_n|X_1, X_2,\ldots, X_{n-1})$ with the previously generated $x^1_1, x^2_1,\ldots, x^{n-1}_1$.

(3) With steps (1) and (2), annual streamflow in other years at each site until the required length *n* can be generated.

5.3 Application

5.3.1 Data description

The annual streamflow from 1906-2003 of four stations in Colorado River basin, namely sites 21 (Paria River at Lees Ferry, AZ), 22 (Little Colorado River near Cameron, AZ), 24 (Virgin River at Littlefield, AZ), and 27 (Bill Williams River below Alamo Dam, AZ) are used for the illustration of the proposed method. The positions of these sites are illustrated in Figure 5. 1 and the annual streamflow at each site can be downloaded from the website (http://www.usbr.gov/lc/region/g4000/NaturalFlow/previous.html).

The mean, standard deviation, skewness, auto-correlation, maximum and minimum values of the annual streamflow at each site are listed in Table 5. 1. The autocorrelation of annual streamflow at each site is relatively low (≤ 0.13) and thus this

statistics is not taken into account in the simulation. The preservation of statistics of each site (such as mean, standard deviation, and skewness) and cross-correlation between different sites for annual streamflow is required for simulation.

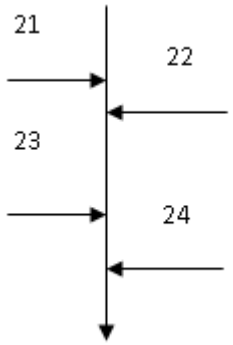

Figure 5. 1 Illustration of four stations in Colorado River basin.

Table 5. 1 Statistics of annual streamflow at four sites.

Statistics	Station 1	Station 2	Station 3	Station 4
Mean	0.80	7.19	7.23	3.98
Standard deviation	0.31	4.76	3.88	4.76
Skewness	1.29	1.62	1.43	2.46
Autocorrelation	0.11	0.03	0.13	0.10
Maximum value	1.88	25.13	19.80	27.42
Minimum value	0.35	0.66	2.86	0.05

In this study, both the entropy method and the entropy-copula method were used for the multi-site annual streamflow simulation at these four sites. For computational convenience, the annual streamflow data were scaled to the interval (0, 1) before parameter estimation for the entropy method. For the original annual streamflow data

(OD) of each station with maximum value MX and minimum value MN, the scaled annual streamflow data (SD) of each station was expressed as: SD = [(OD-MN)]/[(1+ *d*)MX-(1-*d*)MN], where *d* is a scale parameter. The scale parameter *d* (=0.05 in this study) relates to the maximum and minimum value and is for the generation of streamflow values beyond the observaiton. For the entropy-copula method, the scaled data were also used for comparison with the entropy method. The generated streamflow was than rescaled to its original domain after generation.

5.3.2 Performance measure

One hundred sequences of annual streaflow with the similar length as the historical record (98 years) are generated. Basic statistics of annual streamflow at an individual site, including the mean, standard deviation, skewness, maximum and minimum values, and dependence structure between different sites from the generated sequences are compared with those from historical records. Box plots are used to display the generated and simulated statistics. The performance of the proposed method is judged to be good when a statistic falls within the boxplot.

Three measures are used for characterization of the dependence structure of streamflows of different sites, namely, the Pearson correlation coefficient, Spearsman's Rho and Kendall's Tau. For the spatial correlation of streamflow between different sites, these three kinds of correlation are used for comparison. The Pearson correlation coefficient has been traditionally used for measuring the dependence structure of streamflows [*Sharma et al.*, 1997; *Prairie et al*., 2007; *Lee et al.*, 2010; *Nowak et al.*,

2010; *Salas and Lee*, 2010]. For continuous random variables X and Y, the sample estimate of the Pearson correlation coefficient ρ_{xy} can be defined as:

$$\rho_{xy} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{5.23}$$

where $E(X)$, $E(Y)$ and $E(X^2)$, $E(Y^2)$ are the mean and variance of random variables X and Y; E(XY) is the expectation of the cross product of random variable X and Y. The Pearson correlation coefficient measures the linear dependence of two random variables.

The Spearman's Rho is defined in a similar way as the Pearson correlation coefficient but use the rank of the observations instead. For the observed streamflow pairs (X_i, Y_i), i=1, 2,…, n, define R_i as the rank of X_i and S_i as the rank of Y_i. The Spearman's Rho can be defined as [*Genest and Favre*, 2007]:

$$\rho_s = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \tag{5.24}$$

where:

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{n+1}{2} = \overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

The Kendall's tau is defined as [*Genest and Favre*, 2007]:

$$\rho_\tau = \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4}{n(n-1)} P_n - 1 \tag{5.25}$$

where P_n and Q_n are the concordant and discordant pairs of observations. The pairs of observations (x_i, y_i) and (x_j, y_j) are said to be concordant if $(x_i-y_i)(x_i-y_i)>0$ while are said to be discordant if $(x_i-y_i)(x_i-y_i)<0$.

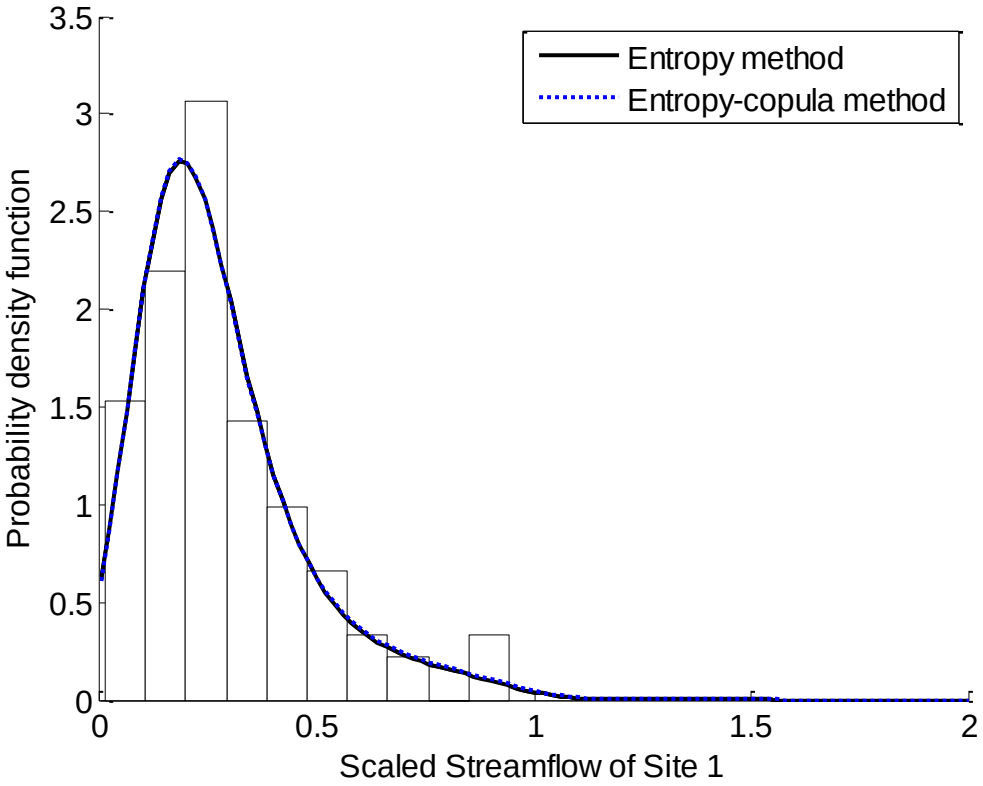

Figure 5. 2 Marginal PDF of the annual streamflow at site 1 from entropy method and entropy-copula method.

5.3.3 Marginal and Joint PDF

The probability density function (PDF) was first assessed. The marginal PDFs of annual streamflow at site 1 from the entropy method in equation (5.8) and entropy-

copula method in equation (5.17) are shown in Figure 5. 2. It can be seen that the PDFs from both method are indistinguishable, both of which fitted the histogram well. This can be interpreted in that the marginal PDF in equation (5.8) derived from the joint distribution is also characterized by the first four moments, which is similar to the PDF derived directly from the first four moments in equation (5.17).

Table 5. 2 Goodness of fit test for statistics S_n and T_n with associated *p* values for different streamflow pairs.

Streamflow Pairs	1-2	1-3	1-4	2-3	2-4	3-4
S_n	0.05	0.16	0.75	0.59	0.97	0.74
p-value	0.05	0.17	0.77	0.65	1.01	0.62
T_n	0.12	0.14	0.10	0.81	0.94	0.16
p-value	0.09	0.13	0.25	0.83	0.90	0.11

The goodness of fit test was conducted for assessing the validity of the Gaussian copula for the entropy-copula method. The Cramér—von Mises statistic (S_n) and Kolmogorov-Smirnov statistics (T_n) were computed and the associated *p*-values based on a run of 2000 samples were obtained using the parametric bootstrap procedure [*Genest et al*., 2006; *Genest and Favre*, 2007], as shown in Table 5. 2. The *p*-values from S_n and T_n are higher than 5% for all streamflow pairs. All the *p*-values are higher than 5% from the statistics T_n. It can be seen that generally the Gaussian copula is a valid model for constructing the joint distributions for different streamflow pairs.

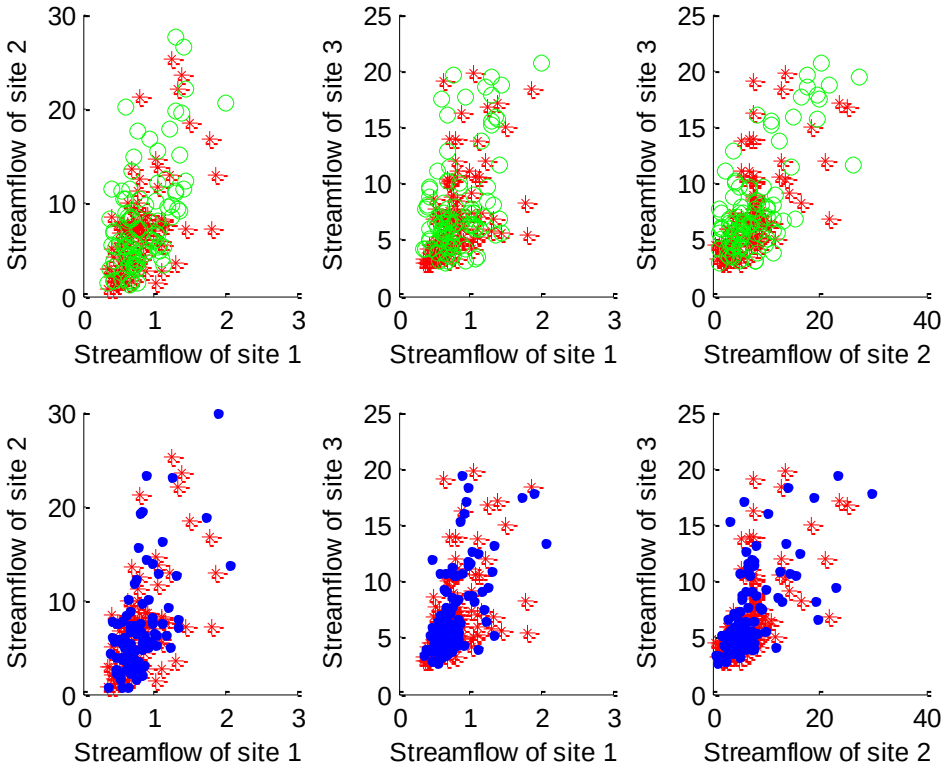

Figure 5. 3 Scatter plot of observed streamflow (star) and generated streamflow from entropy method (open circle) and entropy-copuala method (dot).

The scatter plot was used to compare observed streamflow and generated streamflow. For the annual streamflow at site 1, site 2 and site 3, the joint distribution in equation (5.11) by the entropy method or equation (5.18) by the entropy-copula method is needed and the corresponding conditional distribution can be used for streamflow generation. Three sequences of 100 annual streamflow pairs from site 1, site 2 and site 3 were generated. The scatter plot of the generated streamflow pairs for the three stations compared with observed streamflow pairs are shown in Figure 5. 3. Generally the

spread pattern of the generated streamfllow pairs from both methods matches the observed streamflow pairs well.

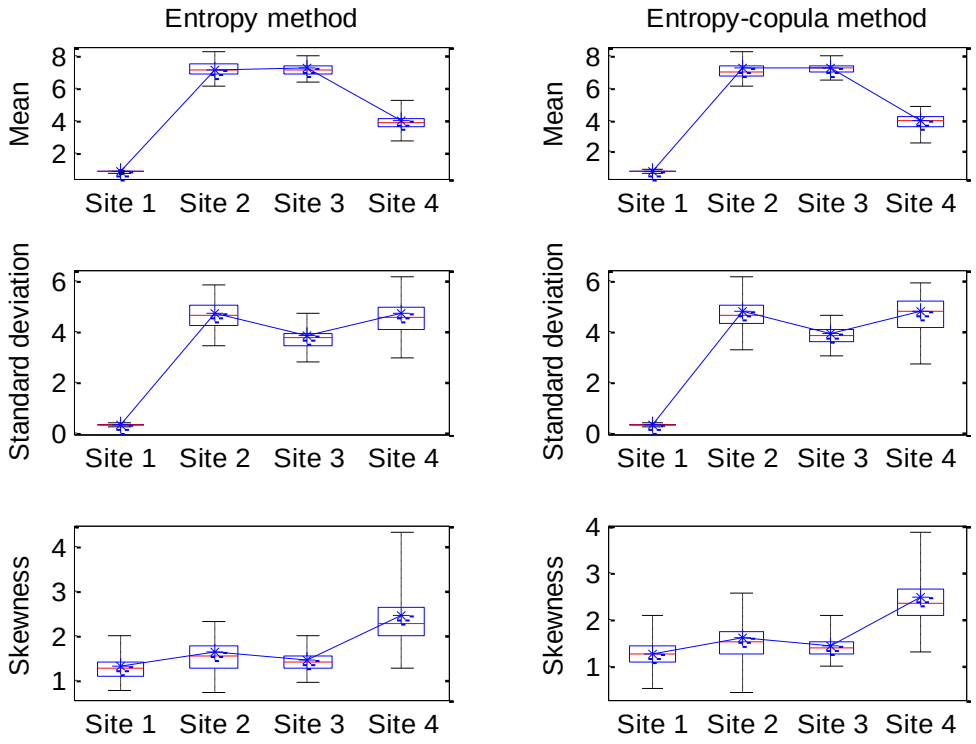

Figure 5. 4 Boxplots of mean, standard deviation and skewness of annual streamflow from entropy method and entropy-copula method. (The unit for mean and standard deviation is cubic meter per second.)

5.3.4 At-site properties

The mean, standard deviation and skewness of streamflow at each site with the entropy method and the entropy-copula method are shown in Figure 5. 4 . No significant difference is found in the preservation of these statistics. All these statistics fell in the boxes and both methods preserved the statistics well. The relative error (RE), defined as

RE = $(S_m - X_o)/X_o$, where S_m is the median of simulated statistic and X_o is the observed statistic of annual streamflow, for each site is shown in Table 5. 3 for the entropy method and entropy-copula method. The relative error (absolute value) of mean and standard deviation of four sites was below 5% and skewness was below 10% for both methods. The relative error of the mean, standard deviation, and skewness of the statistics from the two methods was comparable.

Table 5. 3 Relative error (%) of statistics generated from entropy method and entropy-copula method.

Method	Statistics	Station 1	Station 2	Station 3	Station 4
Entropy method	Mean	0.1	0.0	-0.3	-2.9
	Standard deviation	-1.1	-2.4	-3.2	-4.4
	Skewness	-1.4	-5.0	-1.5	-7.6
	Maximum	0.8	2.0	-1.9	-10.2
	Minimum	-0.5	34.7	-2.7	40.5
Entropy-copula method	Mean	-0.8	-2.3	-0.8	-2.2
	Standard deviation	-0.7	-3.5	-0.7	-0.1
	Skewness	-1.4	-6.0	-2.2	-5.0
	Maximum	0.1	0.6	-1.5	-6.4
	Minimum	-0.9	29.7	-2.3	40.4

The maximum and minimum values for the entropy and entropy-copula method are shown in Figure 5. 5. It can be seen from the boxplot that generally all these statistics were preserved well, though the minimum values were rather underestimated for the entropy-copula method. The relative error for the generated maximum and minimum values are also shown in Table 5. 3. For the maximum values, the relative errors (absolute value) were under 15% for both the entropy and entropy-copula method.

The relative errors for the minimum values were relatively large compared with other statistics. For example, the relative errors (absolute value) of generated minimum values for site 4 were 40.5% and 40.4%, respectively. The reason may be that the observed minimum values are relatively small which will cause a large relative error. The median values of the generated minimum values were 0.07 and 0.07 for the entropy and entropy-copula methods, while the observed minimum value was 0.05. From the median of generated minimum values, the minimum values were also preserved relatively well.

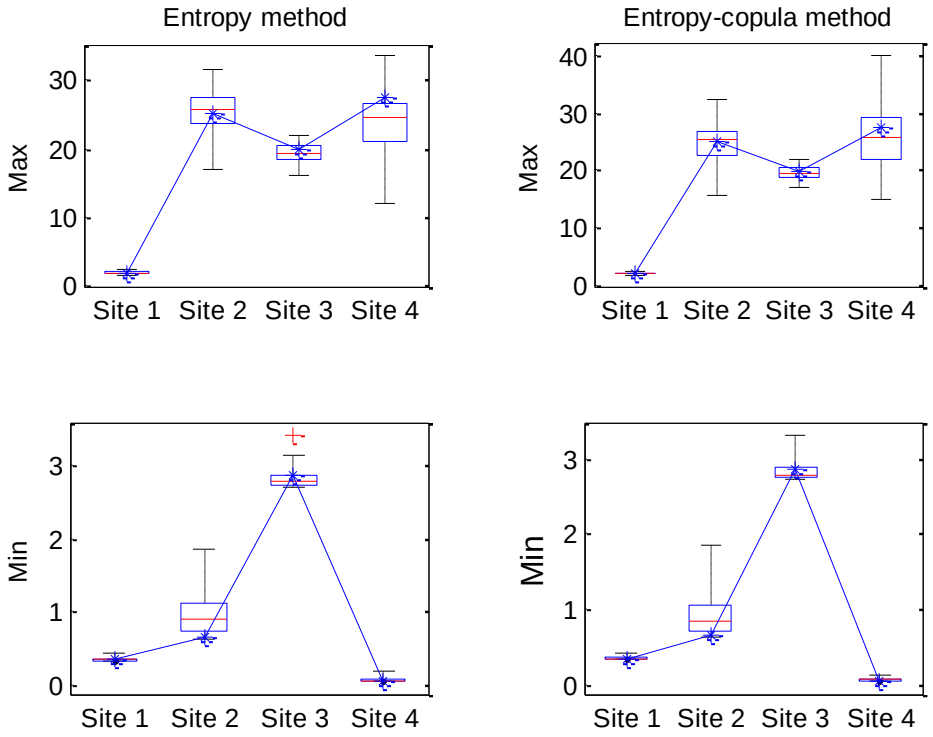

Figure 5. 5 Boxplots of maximum and minimum values of annual streamflow from entropy method and entropy-copula method. (The unit for maximum and minimum is cubic meter per second.)

5.3.5 Multi-site properties

The cross correlation between different sites from the entropy method and entropy-copula method is shown in Figure 5. 6. It can be seen that the observed Pearson correlation falls in the box for all streamflow pairs and thus the Pearson correlation of annual streamflow among different sites can be preserved well by the entropy method. For the entropy-copula method, the observed Pearson correlation falls in the box for all streamflow pairs except for that between site 1 and site 4.

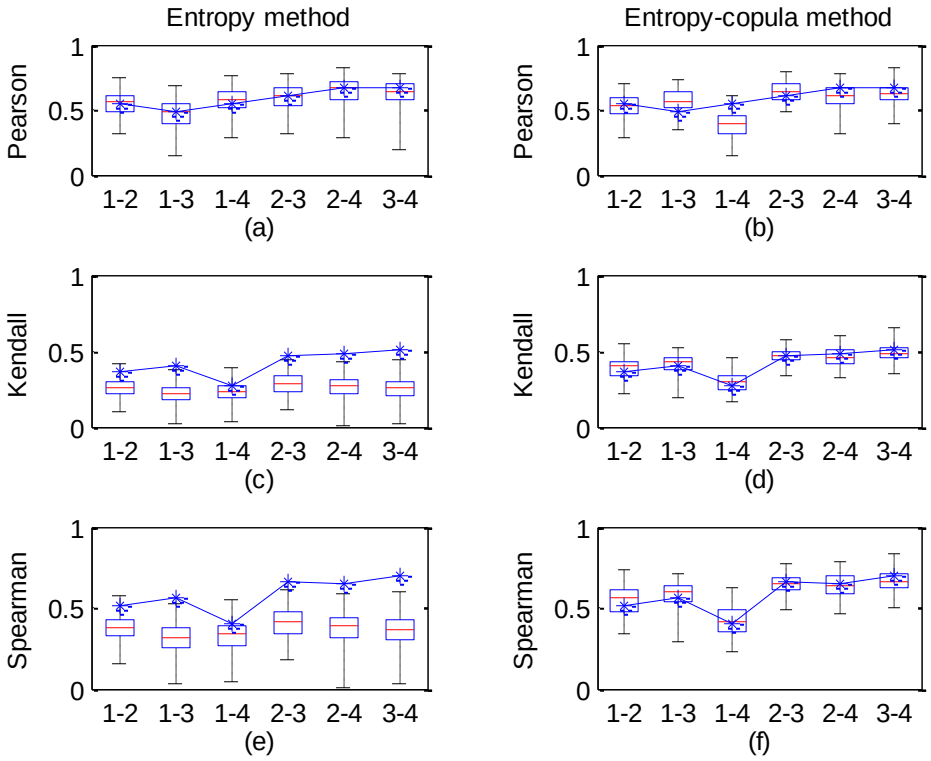

Figure 5. 6 Boxplots of Pearson, Keandall, and Spearman correlations of annual streamflow pairs from entropy method and entropy-copula method. (The symbol *i*-*j* (1≤*i*<*j*≤4) represents the annual streamflow pairs of station *i* and *j*.)

Table 5. 4 Relative error (%) of different dependence measure from entropy method and entropy-copula method.

Method	Dependence Measure	Streamflow pairs					
		1-2	1-3	1-4	2-3	2-4	3-4
Entropy Method	Pearson	1.3	1.0	5.6	-0.5	0.8	-5.5
	Kendall	-29.2	-46.9	-14.3	-38.0	-43.7	-49.7
	Spearman	-26.9	-44.4	-13.3	-36.4	-40.2	-47.5
Entropy-Copula Method	Pearson	-3.4	17.9	-26.9	2.9	-8.3	-6.5
	Kendall	7.5	6.2	7.4	1.4	-3.6	-5.5
	Spearman	10.8	5.9	5.6	-0.8	-0.7	-5.0

The relative error is shown in Table 5. 4. The relative error (absolute value) is within the range 0.5%-5.6% for all streamflow pairs by the entropy method. The relative error (absolute value) is within the range 2.9%-26.9% for all streamflow pairs by the entropy-copula method. Thus the entropy method outperforms the entropy-copula method in preserving the Pearson correlation. The reason is that the cross correlation of random variables X and Y (annual steamflow) is modeled directly with the entropy method. From the definition of the Pearson correlation in equation (5.23), when the sample estimation of $E(X)$, $E(Y)$, $E(X^2)$ and $E(Y^2)$ were used as separate constraints in the form of the first two moments and $E(XY)$ is used as the joint constraint, the pearson correlation can be modelled derectly and is expected to be preserved well, while the cross correlation is modeled by the copula through the CDF of the random variables X and Y with the entropy-copula method.

It should be noted that the Pearson correlation coefficient used here is a measure of linear correlation following the traditional approach for streamflow simulation. The linear correlation may not be sufficient for characterizing the dependence property.

Though the entropy-copula method does not preserve as well as the entropy method in terms of the Pearson correlation, this does not mean that the entropy-copula does not perform as well in modeling streamflow at different sites. The copula method also enables the characterization of nonlinear dependence properties with different copula fucntions.

To further compare the performance in preserving the nonlinear dependence of these two methods, the Kendall and Spearman correlations were also used for comparison. The generated and observed Kendal and Spearman correlations between different sites are shown in Figure 5. 6. The performance of these two methods differs significantly in preserving the dependence. For the entropy method, these two correlations were not preserved well. The relative error (absolute value) for the Kendal and Spearman correlaiton is within the range 13.3%-49.7%. For the entropy-copula method, all these correlations were preserved well, since all the observed statistics fell in the boxes. The relative error (absolute value) for the Kendal and Spearman correlaiton is within the range 0.7%-10.8%. These results showed that the entropy-copula outperformed the entropy method in preserving the Kendall and Spearman correlation (nonlinear dependence).

5.4 Conclusion

The entropy method and entropy-copula method are proposed for the multi-site annual streamflow simulation at different sites. Both methods are capable of preserving mean, standard deviation and skewness well. The relative error (absolute values) of mean and standard deviation was below 5% and that of skewness was below 10% for

both methods. The entropy method preserves the Pearson correlation well for all streamflow pairs, while the entropy-copula preserves the Pearson correlation well for all streamflow pairs except for that between site 1 and site 4. The relative error (absolute value) for the Pearson correlation is within the range 0.5%-5.6% and 2.9%-26.9% for the entropy method and entropy copula method, respectively, and the entropy method outperforms the entropy-copula method in preserving the Pearson correlation. The entropy method does not preserve the Kendall and Spearman correlation well, while the entropy-copula preserves these correlations well. The relative error (absolute value) is within the range 13.3%-49.7% and 0.7%-10.8% for the entropy method and entropy-copula method, respectively and the entropy-copula method outperforms the entropy method in preserving Kendall and Spearman correlation.

The advantage of the entropy method is that the required statistics can be used as constraints to derive the joint distribution and then the random samples generated from the corresponding conditional distribution can preserve these statistics. However, as many parameters as the required statistics need to be estimated. This method would be complicated when a large number of statistics needs to be preserved. The advantage of the entropy-copula method is that it is simpler in terms of parameter estimation, since only a few parameters need to be modeled simultaneously. In addition, different copulas can be used to construct the joint distribution and thus a variety of dependence structures can be modeled.

CHAPTER VI

ENTROPY BASED METHOD FOR DROUGHT ANALYSIS

6.1 Introduction

Drought analysis is important for water resources planning and management. *Yevjevich* [1967] used the run theory to define a drought as a sequence of cumulative intervals where water supply remains below water demand. This enables the characterization of a drought event with certain drought properties, such as duration and severity. These properties, assumed as random variables, have been commonly used for analyzing droughts.

The probabilistic characterization of drought based on drought duration and severity, either separately or jointly, is needed for drought analysis and a traditional way is to fit a probability density function. The drought duration can be modeled by a geometric distribution [*Kendall and Dracup*, 1992; *Mathier et al.*, 1992] when it is treated as a discrete random variable or by an exponential distribution when it is treated as a continuous random variable [*Zelenhasi and Salvai*, 1987]. The gamma distribution is generally used to describe drought severity. However, the correlation between drought duration and severity cannot be characterized by univariate analysis and alternative multivariate approaches have, therefore, been used to model the correlation between drought variables [*González and Valdés*, 2003; *Salas et al.*, 2005; *Kim et al.*, 2006; *Shiau*, 2006; *Nadarajah*, 2007; *Nadarajah*, 2009].

A number of bivariate distributions have been proposed to characterize the joint behavior of drought duration and severity. These distributions have considered the same marginal distributions of drought duration and severity, for example, the bivariate Pareto distribution [*Nadarajah*, 2009]. However, this type of distribution may not work, since in reality the drought duration and severity may have different marginal distributions. To address this issue, the copula method has been applied to construct joint distributions with different marginal distributions [*Shiau*, 2006; *Shiau et al.*, 2007], but the marginal distributions need to be derived which is often done empirically. The joint distribution can also be constructed from the product of conditional distribution of drought severity given drought duration and marginal distribution of drought duration [*Shiau and Shen*, 2001]. Furthermore, nonparametric methods have also been proposed for bivariate drought analysis [*Kim et al.*, 2003; *Kim et al.*, 2006].

The objective of this paper is to propose an alternative method, based on the entropy theory, for constructing a bivariate distribution of drought duration and severity. The advantage of this method is that the marginal distributions can be of different forms and can be derived based on the given information. The proposed method is applied to monthly streamflow of Brazos River at Waco, Texas, for drought analysis.

6.2 Method

The method for deriving an entropy-based bivariate distribution entails: (1) defining the Shannon entropy for univariate and bivariate cases, (2) defining given information in terms of constraints, (3) maximizing entropy using the method of

Lagrange multipliers and deriving the probability density functions (PDFs), and (4) determination of the Lagrange multipliers.

6.2.1 Univariate and bivariate entropy

For continuous random variables X and Y with a joint PDF $f(x, y)$ defined over the space $[a, b]\times[c, d]$, the marginal distribution for random variable X can be obtained by integrating the joint PDF $f(x, y)$ over Y as:

$$f(x)=\int_c^d f(x,y)dy \tag{6.1}$$

Similarly, the marginal distribution for random variable Y can be obtained as:

$$f(y)=\int_a^b f(x,y)dx \tag{6.2}$$

Generally the expression for the marginal distribution cannot be expressed explicitly and numerical solution is needed.

For continuous random variables X and Y with a joint PDF $f(x, y)$ defined over the space $[a, b]\times[c, d]$, the bivariate Shannon entropy H can be defined as:

$$H=-\int_c^d\int_a^b f(x,y)\ln f(x,y)\,dxdy \tag{6.3}$$

The objective is to derive the PDFs $f(x)$, $f(y)$ and joint PDF $f(x, y)$ with entropy theory.

6.2.2 Constraints

It has been shown that many of the commonly used distributions can be derived using entropy theory with different constraints [*Singh*, 1998]. For example, the exponential distribution can be obtained with the constraint in the form of mean, while the gamma distribution can be obtained with the constraints in the form of mean and

logarithmic mean. For the bivariate case, the bivariate normal distribution can be obtained when the first two moments of each variable and the product of the two variables are specified as constraints [*Kapur*, 1989]. Thus, the entropy-based distribution possesses a flexible form and provides an alternative way to derive a bivariate distribution of underlying data by choosing appropriate constraints to accommodate different marginal distribution forms.

For the bivariate case with random vector (X, Y) with a probability density function $f(x, y)$, the constraints can be specified as:

$$\int_c^d \int_a^b g_i(x, y) f(x, y) dxdy = E(g_i) \quad i=0,1, 2,\ldots, m \tag{6.4}$$

where $g_i(x, y)$ is a known function of random vector (X, Y) with $g_0(x, y)=1$; $E(g_i)$ is the expected value of the function $g_i(x,y)$; and m is the number of constraints.

To derive the joint density function $f(x,y)$ of drought duration (denoted as random variable X) and drought severity (denoted as random variable Y), the constraints in equation (6.4) need to be specified separately and jointly. The first constraint is the unity constraint that ensures the integration of the probability density function equals one. This constraint corresponds to equation (6.4) for the case $i=0$ and can be expressed as:

$$\int_c^d \int_a^b f(x, y) dxdy = 1 \quad i=0,1, 2,\ldots, m \tag{6.5}$$

The joint constraint characterizing the dependence structure between the two variables X and Y can be specified as:

$$\int_c^d \int_a^b xyf(x, y) dxdy = E(xy) \quad i=0,1, 2,\ldots, m \tag{6.6}$$

Except for the unity constraint in equation (6.5) and the joint constraint in equation (6.6), separate constraints for random variables *X* and *Y* need to be specified to derive the joint distribution. Two sets of separate constraints are specified in this study to illustrate the proposed method. To preserve the parsimonious property, only one constraint (with one Lagrange parameter) is used for drought duration and two constraints (with two Lagrange parameters) are used for drought severity.

As stated previously, the exponential distribution can be derived with the constraint of mean, while the gamma distribution can be derived with the constraints of mean and logarithm mean. This implies that the mean and logarithm mean of the underlying random variable would be good candidates as constraints. For the first set of separate constraints, it is assumed that the mean is specified as a constraint for random variable *X* (drought duration) and the mean and logarithm mean are specified as constraints for random variable *Y* (drought severity) expressed as:

$$\int_c^d \int_a^b xf(x,y)dxdy = E(x) \tag{6.7}$$

$$\int_c^d \int_a^b yf(x,y)dxdy = E(y) \tag{6.8}$$

$$\int_c^d \int_a^b \ln yf(x,y)dxdy = E(\ln y) \tag{6.9}$$

Constraints expressed by equations (6.5)-(6.9) constitute one set of constraints and will correspond to a joint distribution denoted as ME1.

On the other hand, Pearson's product-moment correlation coefficient has been used to measure the linear dependence which involves the second order moments. The

expectations of the second order moment of drought duration and severity were thus selected for the second sets of separate constraints expressed as:

$$\int_c^d \int_a^b x^2 f(x,y)dxdy = E(x^2) \tag{6.10}$$

$$\int_c^d \int_a^b y^2 f(x,y)dxdy = E(y^2) \tag{6.11}$$

In addition, the logarithm mean of random variable *Y* in equation (6.9) is retained as another separate constraint for drought severity. Constraints expressed by equations (6.5), (6.6) and (6.9)-(6.11) constitute another set of constraints and will correspond to a jont distribution denoted as ME2.

6.2.3 Maximization of entropy

The principle of maximum entropy was proposed by *Jaynes* [1957] which states that the probability density function should be selected among all the distributions with the maximum entropy subject to specified constraints. Thus, the joint PDF can be obtained by maximizing the entropy given by equation (6.3), subject to the constraints given by equation (6.4), which can be done using the method of Lagrange multipliers. Denoting the Lagrange multipliers as $\lambda_0, \lambda_1, \ldots, \lambda_m$, the Lagrangian function *L* with equations (6.3) and (6.4) can be expressed as [*Kapur*, 1989]:

$$L = -\int_c^d \int_a^b f(x,y) \ln f(x,y)dxdy - \sum_{i=0}^{m} \lambda_i \left[\int_c^d \int_a^b g_i(x,y) f(x,y)dxdy - E(g_i) \right] \tag{6.12}$$

The maximum entropy-based joint probability density function is obtained by differentiating *L* in equation (6.12) with respect to *f* and setting the derivative to zero as [*Kesavan and Kapur*, 1992]:

$$f(x,y) = \exp\left[-\sum_{i=0}^{m} \lambda_i g_i(x,y)\right] \quad i=0, 1, 2,\ldots, m \tag{6.13}$$

Each Lagrange multipliers in equation (6.13) is related to each constraint. In real applications, different constraints in equation (6.4) can be chosen and thus different joint distributions from equation (6.13) can be obtained. The suitable joint distribution can be selected according to certain performance measures based on the observations and thus employed for drought analysis.

With the constraints in equations (6.5) to (6.9), the maximum entropy-based joint probability density function (denoted as ME1) can be expressed as:

$$f(x,y) = \exp\left[-\lambda_0 - \lambda_1 x - \lambda_2 y - \lambda_3 \ln y - \lambda_4 xy\right] \tag{6.14}$$

Likewise, the maximum entropy-based joint distribution (denoted as ME2) based on the unity constraint in equation (6.5), joint constraints in equation (6.6) and separate constraints in equations (6.9), (6.10) and (6.11) can then be expressed as:

$$f(x,y) = \exp\left[-\lambda_0 - \lambda_1 x^2 - \lambda_2 y^2 - \lambda_3 \ln y - \lambda_4 xy\right] \tag{6.15}$$

The marginal distribution for random variable X (drought duration) can be obtained by integrating the joint PDF $f(x, y)$ either for ME1 in equation (6.14) or ME2 in equation (6.15) over Y (drought severity) as expressed in equation (6.1). Similarly, the marginal distribution for random variable Y (drought severity) can be obtained as expressed in equation (6.2).

An interesting property of the marginal distributions in equation (6.1) and (6.2) obtained from the joint distribution ME1 or ME2 for drought duration and severity is that the distribution forms of random variable X and Y can be different. What is more,

the marginal distribution can differ from the commonly used distributions, such as exponential or gamma, thus providing more flexible distribution to characterize random variable *X* or *Y*.

6.2.4 Determination of Lagrange multipliers

Generally, the analytical solution for obtaining the Lagrange multipliers in equation (6.13) does not exist for the bivariate case and therefore a numerical solution is needed. For the maximum entropy-based bivariate distribution in equation (6.13), the Lagrange multipliers can be obtained by minimizing the convex function Γ expressed as [*Mead and Papanicolaou*, 1984; *Kapur*, 1989]:

$$\Gamma = \lambda_0 + \sum_{i=1}^{4} \lambda_i \overline{g_i} = \ln \int_c^d \int_a^b \exp\left[-\sum_{i=1}^{4} \lambda_i g_i(x,y) \right] dxdy + \sum_{i=1}^{4} \lambda_i \overline{g_i} \tag{6.16}$$

6.3 Application

6.3.1 Data

Monthly streamflow data of Brazos River at Waco, TX (USGS 08096500) for the period from January 1941 to December 2009 was used for drought analysis. The mean streamflow of each month was used as the truncation level to define the drought event. Drought duration was defined as the number of consecutive months during which streamflow was below the truncation level, while drought severity was defined as the cumulative difference between the truncation level and observed streamflow within the corresponding drought duration.

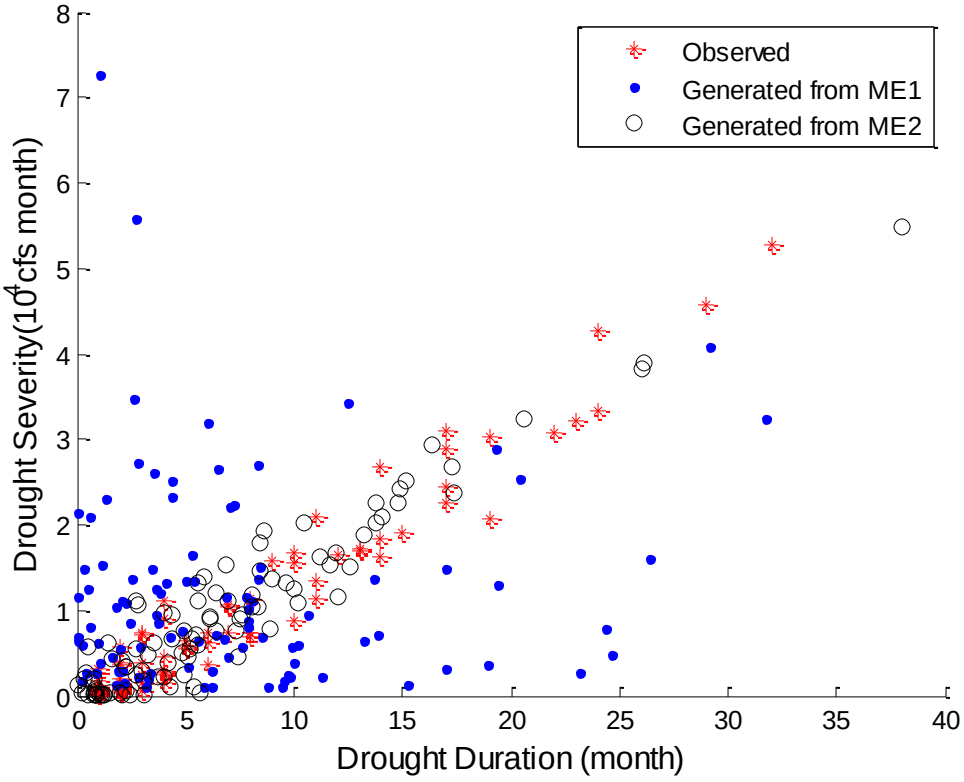

Figure 6. 1 Scatterplot of observed data and generated data from ME1 and ME2 distributions.

The scatter plot of observed data shows strong dependence between drought duration and severity as shown in Figure 6. 1. This indicates that a bivariate distribution is needed to characterize the dependence between drought duration and severity for drought analysis.

6.3.2 Comparison of ME1 and ME2

The root mean square error (RMSE) is used here to assess the performance of the entropy based distribution defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - o_i)^2} \tag{6.17}$$

where n is the number of observed values; x_i and o_i are, respectively, the theoretical and empirical cumulative probability.

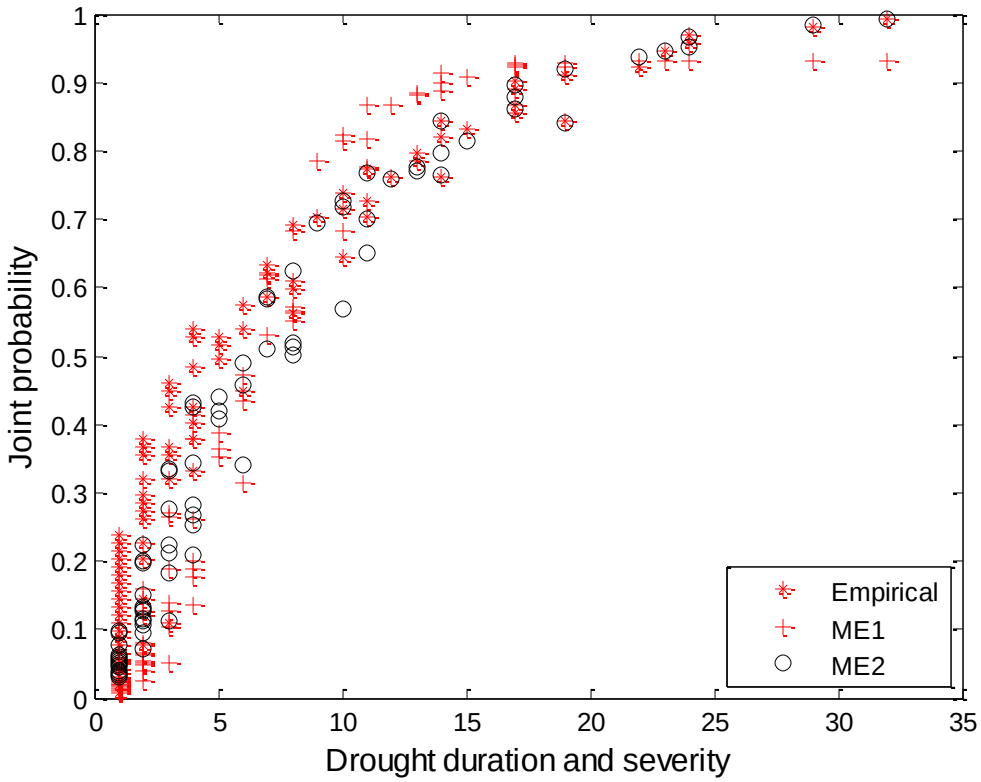

Figure 6. 2 Comparison of empirical and theoretical probability for drought duration and severity.

Two maximum entropy-based joint density functions (ME1 and ME2) in equation (6.14) and (6.15) were obtained with different constraints. The performances of

the distributions (marginal and joint distributions) were assessed based on drought data defined previously. Empirical histograms of drought duration and severity data were first compared with the theoretical probability density functions defined by equations (6.1) and (6.2) for ME1 and ME2. Results showed that generally the theoretical PDFs fitted the empirical histograms well (not presented here). In addition, theoretical cumulative probabilities for drought duration and severity from ME1 and ME2 also fitted the empirical cumulative probabilities estimated from the Gringorten's plotting position formula well (not presented here). These results shown that both the marginal distributions from ME1 and ME2 modeled drought duration and severity well.

The joint distributions from ME1 and ME2 were then compared. The empirical joint cumulative probability corresponding to the combination of drought duration and severity obtained using the approach proposed by *Yue et al.* [1999] was compared with the theoretical joint cumulative probabilities from ME1 and ME2 as shown in Figure 6. 2 The RMSE values of the joint probability for ME1 and ME2 were 0.090 and 0.060, respectively. It can be seen that the theoretical cumulative probabilities from ME2 fitted the empirical values well, better than ME1. Furthermore, a large number of data generated from the two maximum entropy-based joint density functions (ME1 and ME2) were compared with observed data to assess the performance of modeling the dependence structure. For this study, a set of 100 random vectors was generated from each distribution and a scatter plot of generated and observed data is also shown in Figure 6. 1. It can be seen that generally the pattern of spread of generated data from ME2 matched that of the observed data. However, the generated data from ME1 did not

reproduce the dependence structure satisfactorily. The correlation coefficients of the generated data from ME1 and ME2 were 0.14 and 0.93, respectively, whereas for the observed value it was 0.97. More data were generated from each distribution to assess the performance of the two distributions and similar results were obtained. These results showed that the ME2 distribution was suitable for modeling the dependence structure between drought duration and severity of this dataset and was used hereafter for further analysis and application.

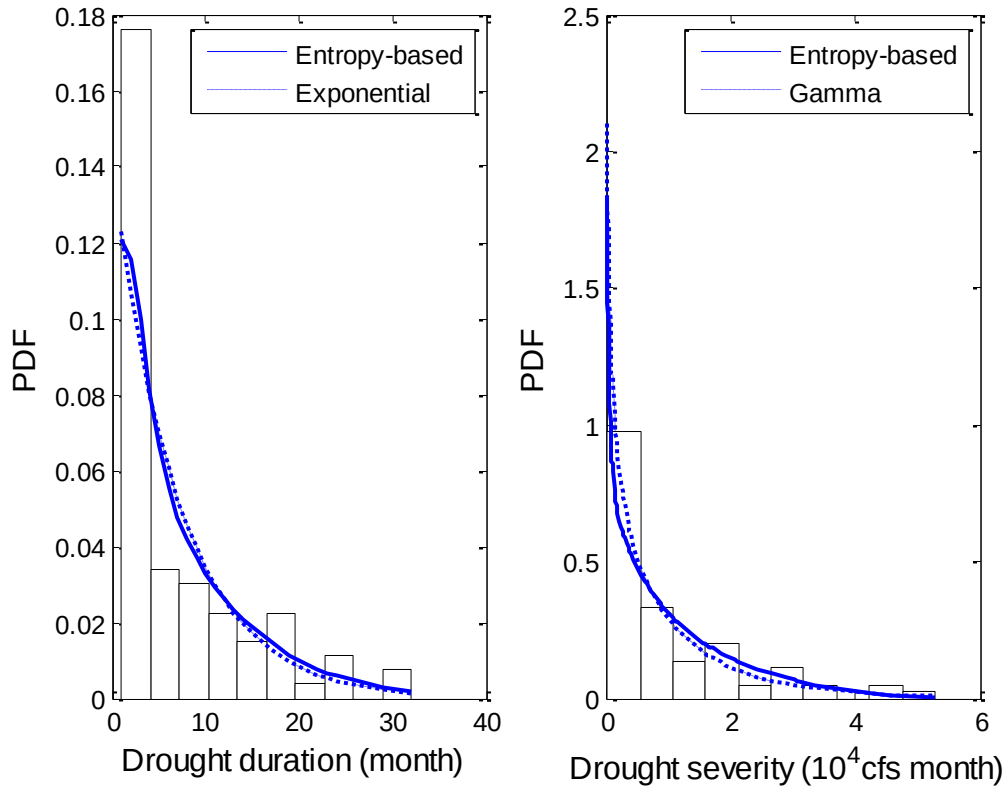

Figure 6. 3 Empirical histograms and marginal PDFs from entropy-based ME2, exponential and gamma distributions.

6.3.3 Drought analysis

The exponential distribution is commonly used for modeling drought duration, while the gamma distribution is commonly used for modeling drought severity. These two distributions were also used for comparison with the marginal distributions derived from ME2 distribution. Empirical histograms and marginal distributions of drought duration with equation (6.1) and of drought severity with equation (6.2) are shown in Figure 6. 3 . The PDFs of the exponential distribution for drought duration and gamma distribution for drought severity are also shown in Figure 6. 3. The PDFs from the entropy-based distribution captured the general pattern of empirical histograms. In addition, the PDF curves from the entropy-based distributions were close to those from exponential and gamma distributions, though some discrepancies existed.

The empirical cumulative probabilities estimated from Gringorten's plotting position formula and theoretical cumulative probabilities are shown in Figure 6. 4. In general theoretical probabilities fitted empirical probabilities well. For drought duration, the entropy-based distribution and exponential distribution were close in the left part (duration $d < 8$ months) while in the right part (duration $d > 8$ months) the entropy based distribution seemed to fit better. The RMSE values of the probabilities estimated from the entropy-based distribution and exponential distribution were 0.059 and 0.058, respectively, which were very close to each other. For drought severity, some underestimation existed for the entropy-based distribution on the left part ($s < 1.5 \times 10^4$ cfs·month) and gamma distribution fitted the empirical probability better. In the middle part (1.5×10^4 cfs·month $< s < 3 \times 10^4$ cfs·month), the entropy-based distribution seemed

to fit the empirical probability distribution better. The RMSE values of the probabilities estimated from the entropy-based distribution and gamma distribution were 0.067 and 0.043, respectively, indicating that gamma distribution performed slightly better.

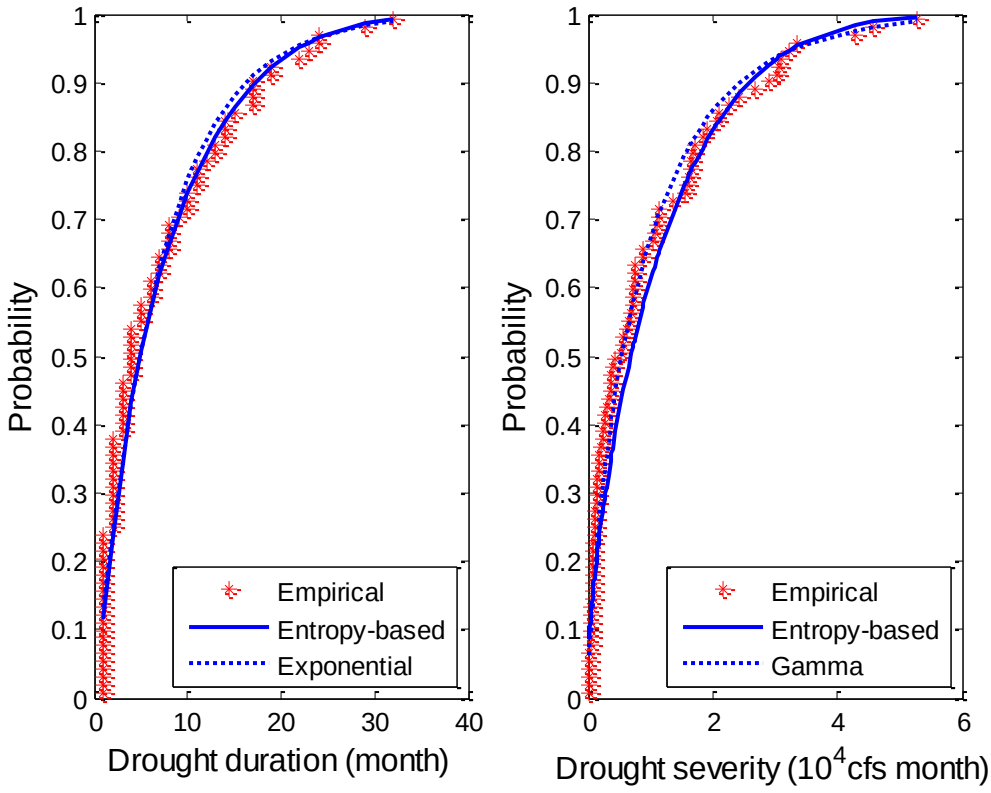

Figure 6. 4 Empirical probabilities and theoretical probabilities from entropy-based ME2, exponential and gamma distributions.

The Kolmogorov-Smirnov (K-S) test was used to further assess the goodness-of-fit of the entropy-based cumulative distribution functions to model drought duration and severity data. Critical values for duration and severity data at a 5% significance level

were 0.26 and 0.46, indicating the hypothesis of the entropy-based distribution to model drought duration and severity data could not be rejected.

Visual and quantitative comparison, together with the results from the exponential and gamma distributions, showed that generally the marginal distribution from the ME2 distribution modeled the drought duration and severity relatively well. Based on the satisfactory results of ME2 distribution in modeling the dependence structure, the proposed ME2 distribution was then applied for drought analysis.

The return period for drought duration D greater than or equal to a certain value d and for drought severity S greater or equal to a certain value s can be defined as [*Shiau*, 2003; *Shiau*, 2006]:

$$T_D = \frac{E(L)}{P_D(D \geq d)} \qquad T_S = \frac{E(L)}{P_S(S \geq s)} \tag{6.18}$$

where $E(L)$ is the expected drought interval time that can be estimated from observed drought data; T_D , T_S are the return periods of drought duration and drought severity, respectively; $P_D(D \geq d)$ and $P_S(S \geq s)$ are the exceedance probabilities of drought duration and drought severity that can be estimated from equations (6.1) and (6.2), respectively. The joint return period of drought duration and severity can be defined by the drought duration and severity exceeding specific values. Specifically, the joint return period T_{DS} of drought duration $D \geq d$ and severity $S \geq s$ can be defined as [*Shiau*, 2003; *Shiau*, 2006]:

$$T_{DS} = \frac{E(L)}{P(D \geq d, S \geq s)} \tag{6.19}$$

where $P(D \geq d, S \geq s)$ is the exceedance probability of drought duration and severity that can be obtained from the joint density function in equation (6.15). The conditional return periods are also needed to assess the risk of water resources systems. The conditional return period $T_{D|S \geq s}$ for drought duration given drought severity exceeding a certain value can be defined as [*Shiau*, 2003; *Shiau*, 2006]:

$$T_{D|S \geq s} = \frac{T_S}{P(D \geq d, S \geq s)} \tag{6.20}$$

Similarly, the conditional return period $T_{S|D \geq d}$ for drought severity given drought duration exceeding a certain value can be defined as [*Shiau*, 2003; *Shiau*, 2006]:

$$T_{S|D \geq d} = \frac{T_D}{P(D \geq d, S \geq s)} \tag{6.21}$$

Table 6. 1 Return period of drought duration and severity.

Return period (Year)	Drought duration (Month)	Drought severity (10^4cfs month)
2	6.9	1.0
5	14.1	2.1
10	19.0	2.9
20	23.4	3.5
50	28.5	4.3
100	32.0	4.8

The univariate return periods of 2, 5, 10, 20, 50 and 100 years defined by separate drought duration and severity were shown in Table 6. 1. For example, the drought duration for the 100 year return period was around 32 months and the drought

severity for the 100 year return period was around 4.8×10^4 cfs month. The joint return periods defined by equation (6.19) for different duration and severity values are shown in Figure 6. 5. Taking the drought during 2005-2007 as an example, this drought lasted for 19 months with a severity 3.0×10^4 cfs month. The univariate return periods were 10.0 years and 11. 9 years. The joint return period estimated from equation (6.19) was 12.7 years. The result from the joint distribution gave a relatively longer return period than that from either drought duration or severity separately.

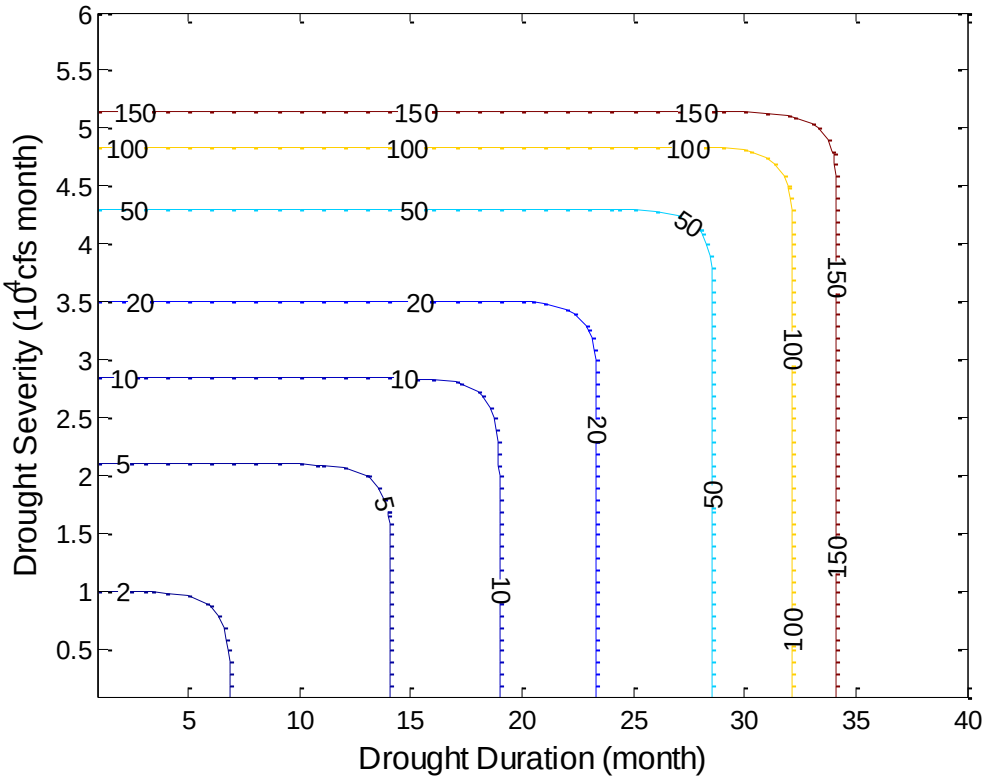

Figure 6. 5 Contours of joint return period (years) of drought duration and severity from entropy-based ME2 distribution.

The conditional return periods are shown in Figure 6. 6. For example, given drought severity $s \geq 1.5 \times 10^4$ cfs month, the conditional return period of the drought duration exceeding 25 months was around 100 years.

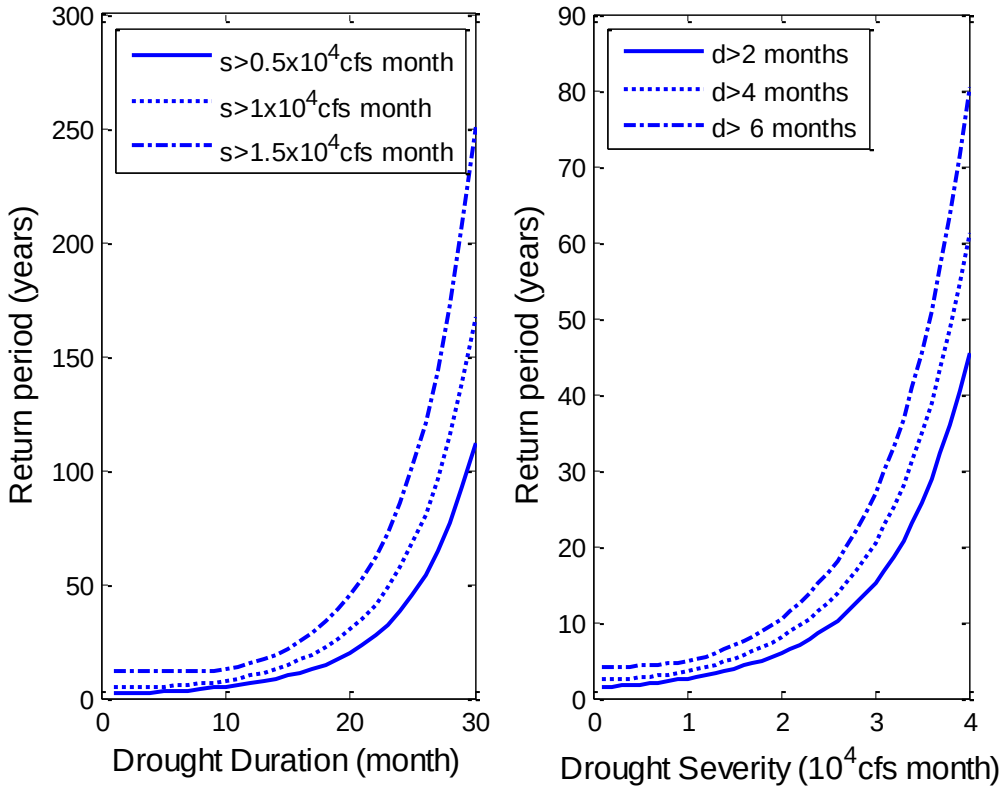

Figure 6. 6 Conditional return periods of drought duration and severity from entropy-based ME2 distribution.

6.4 Conclusion

An alternative method, based on entropy theory, is proposed for constructing the joint distribution of drought duration and severity. Separate and joint constraints of

drought duration and severity can be specified to derive the joint distribution using the principle of maximum entropy and marginal distributions can be derived from the joint distribution. The advantage of the proposed method is that it is flexible to incorporate different forms of marginal distributions of drought duration and severity. In this study, two entropy-based joint distributions are derived with different sets of constraints, both of which lead to different marginal distribution forms. One joint distribution is selected for the drought data defined by monthly streamflow of Brazos River at Waco, Texas. The theoretical joint probability from the entropy-based joint distribution fits the empirical probability well with an RMSE value of 0.06. The spread pattern of generated drought data from the entropy-based joint distribution also matches that of the observed drought data well. Generally the entropy-based joint distribution is capable of modeling drought duration and severity well. This entropy-based joint distribution is then used to derive the 2, 5, 10, 20, 50 and 100 year return periods for drought duration and severity. For a 100 year return period, the drought duration and drought severity are obtained as 32 months and 4.8×10^4 cfs month, respectively. Also it is found the conditional return period of the drought duration exceeding 25 months is around 100 years, given drought severity $s \geq 1.5\times10^4$ cfs month.

CHAPTER VII

ENTROPY-COPULA METHOD FOR DROUGHT ANALYSIS

7.1 Introduction

A number of drought indices, such as the Palmer drought severity index (PDSI) [*Palmer*, 1965] , standard precipitation index (SPI) [*McKee et al.*, 1993] and water deficit [*Dracup et al*., 1980], have been proposed to characterize droughts. A joint distribution is then required to characterize the correlation between drought duration and severity based on different drought indices.

The copula method has been extensively employed to construct joint distributions for drought analysis with different forms of marginal distributions. *Shiau* [2006] applied the copula method for bivariate drought analysis defined by the standardized precipitation index (SPI), and *Shiau et al.* [2007] investigated hydrologic droughts based on monthly streamflow in Yellow River, China. *Shiau and Modarres* [2009] used the copula method for the bivariate drought analysis in Iran with drought defined by the standardized precipitation index (SPI). *Song and Singh* [2010] used a trivariate copula to construct the joint distribution of drought duration, severity and inter-arrival time for drought analysis based on the streamflow data from Wei River basin, China.

The copula method has the ability to construct a joint distribution for drought analysis with different marginal distributions. However, when selecting marginal

distributions, generally commonly used distributions are used. This study proposes an entropy-copula method for deriving marginal distributions using entropy theory and then deriving a joint distribution for drought analysis using the copula method. The advantage of using entropy theory is that marginal distributions can be derived based on whatever information is available and one does not need to be restricted by the distribution forms that are commonly used. Furthermore, commonly used distributions can be derived as special cases of the entropy based distributions. In this study, an entropy-based marginal distribution with the first three moments is proposed and evaluated through comparison with other distributions. Results based on three types of datasets showed the entropy-based marginal distributions performed better in certain cases and can be used as candidate distributions for modeling drought variables. Application of the proposed entropy-copula method for constructing the joint distribution for drought analysis is illustrated with a case study based on Palmer drought severity index (PDSI) data of Climate Division 5 in Texas.

7.2 Entropy-copula method

7.2.1 Entropy-based marginal distribution

The entropy of a continuous random variable X on the interval [a, b] can be defined as [*Shannon*, 1948]:

$$H = -\int_a^b f(x) \ln f(x)\, dx \quad \text{i=0, 1,…, } m \tag{7.1}$$

where $f(x)$ is the probability density function (PDF) of the random variable X. According to the principle of maximum entropy developed by *Jaynes* [1957] , the probability

density function should be selected among all the distributions that maximize the entropy subject to given constraints (or known information). For a set of observations x_i (i=1, 2,…, n) of the random variable X, the known information from the observation is not the probability density function (PDF) $f(x)$ but the expectation of the function $f(x)$ (or constraints).

The general form of the constraints can be specified as:

$$\int_a^b g_r(x) f(x)\, dx = E(g_r) \tag{7.2}$$

where $g_r(x)$ is the known function with $g_0(x)$=1; $E(g_r)$ is the r-th expected value obtained from observations with g_0=1; and m is the number of constraints. When there are a variety of distributions that may be consistent with the specified constraints in equation (7.2), the principle of maximum entropy provides a way to select the one with maximum entropy. The maximum entropy-based probability density function can be obtained by maximizing the entropy in equation (7.1), subject to equations (7.2), using the method of Lagrange multipliers as [*Kesavan and Kapu*r, 1992]:

$$f(x) = \exp[-\lambda_0 - \lambda_1 g_1(x) - \lambda_2 g_2(x) \ldots - \lambda_m g_m(x)] \tag{7.3}$$

where λ_i, i=0, 1, 2, …, m are the Lagrange multipliers, which can be estimated with the Newton -Raphson algorithm [*Kapur*, 1989]. The cumulative distribution function, $G(x)$, can be derived accordingly as:

$$G(x) = \int_0^x \exp[-\lambda_0 - \lambda_1 g_1(t) - \lambda_2 g_2(t) \ldots - \lambda_m g_m(t)]dt \tag{7.4}$$

The commonly used distributions, such as exponential, gamma and normal, can be derived with different forms of constraints in equation (7.2) and the resulting maximum entropy distribution in equation (7.3) incorporates these distributions as special cases. For example, when the mean is used as a constraint, the entropy-based distribution is the exponential distribution. Similarly, when the mean and logarithm mean are used as constraints, the entropy-based distribution is the gamma distribution. Moreover, other forms of distributions that are not commonly used can also be derived using the entropy theory. Thus, the entropy based distribution in equation (7.3) provides flexible forms of marginal distributions.

The maximum entropy distribution in equation (7.3) is regarded as the least biased estimation of the PDF based on the given information or the maximally noncommittal to the missing information [*Jaynes*,1957]. Instead of selecting an empirical distribution by fitting to observations, the entropy-based distribution provides a way to make inferences of the underlying distribution with the use of constraints (or known information) derived from observations. For example, the selection of gamma distribution for modeling the datasets in the traditional approach can be interpreted as using the mean and logarithm mean as constraints to characterize the data.

Often observations of hydrologic variables are characterized by moment statistics, such as mean, standard deviation, and skewness. When the probability distribution of observations is inferred from these statistics of the underlying data, the maximum entropy based distribution can be constructed using these statistics as constraints. The maximum entropy-based distribution with certain moments as

constraints has been used to model the probability density function of data [*Mead and Papanicolaou*, 1984; *Wu*, 2003; *Gotovac et al.*, 2010]. A special case is that when the first two moments (or mean and standard deviation) are used as constraints, the maximum entropy-based distribution is the normal distribution. When the first three moments are employed as constraints expressed as:

$$\int_a^b f(x)\ dx = 1 \tag{7.5}$$

$$\int_a^b xf(x)\ dx = \bar{x} \tag{7.6}$$

$$\int_a^b x^2 f(x)\ dx = \overline{x^2} \tag{7.7}$$

$$\int_a^b x^3 f(x)\ dx = \overline{x^3} \tag{7.8}$$

the maximum entropy-based distribution can be expressed as:

$$f(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3) \tag{7.9}$$

From the statistical meaning of the first three moments, it is expected that the entropy-based distribution $f(x)$ in equation (7.9) characterizes the central tendency (mean), variability (standard deviation) and asymmetric property (skewness) of the data. The performance of this entropy based distribution (denoted as ENT) will be assessed in the following section.

7.2.2 Entropy-based marginal distribution

Sklar [1959] formulated the concept of copula for constructing multivariate distributions from univariate distributions that can be of different forms. For continuous random variables X and Y with their univariate distributions denoted $F_X(x)$ and $F_Y(y)$, the bivariate probability distribution can be expressed with copula C as:

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)) \tag{7.10}$$

The joint distribution relies on copula C which is unique if $F_X(x)$ and $F_Y(y)$ are continuous. Different copula families have been defined, such as the Archimedean and elliptical, which are discussed by *Nelsen* [2006] and *Joe* [1997]. A number of marginal distributions, such as exponential, gamma and weibull distribution, have been used to construct joint distributions with equation (7.10) for drought analysis. When $F_X(x)$ and $F_Y(y)$ are specified as commonly used distribution, equation (7.10) yields the joint distribution of random variables X and Y, which is the typical way the copula method is used.

With the entropy-based distribution as marginal distribution, the entropy-copula based joint distribution can be expressed from equations (7.4) and (7.10) as:

$$F_{X,Y}(x,y) = C(G_X(x), G_Y(y)) \tag{7.11}$$

Since the entropy-based distribution $G_X(x)$ or $G_Y(y)$ incorporates commonly used distributions as special cases, the entropy-copula based joint distribution in equation (7.11) can be regarded as a general framework for constructing a joint distribution. With

the flexible forms of $G_X(x)$ (or $G_Y(y)$) to model the marginal property and a number of copula families C to model the dependence structure, it is expected that the joint distribution in equation (7.11) is capable of modeling the underlying data separately and jointly. An extension of the proposed method to higher dimension is straightforward.

7.3 Method assessment

7.3.1 Data description

Three drought indices, Palmer drought severity index (PDSI), standard precipitation index (SPI) and streamflow deficits, are selected in this study for the evaluation of the proposed entropy based distribution in equation (7.9). The first and second types of datasets are the monthly PDSI and SPI data from January 1895 to December 2010 from 10 Climate Divisions in Texas, which can be obtained from the website: http://www.ncdc.noaa.gov/oa/climate/onlineprod/drought/xmgr.html. The third dataset is monthly streamflow from 14 stations on Colorado River in Texas.

7.3.2 Performance measures

Two measures are used to evaluate the performance of the proposed entropy-based (ENT) distribution. The quantiles corresponding to different return periods (or cumulative probability) estimated from the fitted distribution is generally needed in frequency analysis to assess the risk of drought occurrence. The root mean square error (RMSE) of quantiles was used here to assess the performance of ENT distribution defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - oq_i)^2} \tag{7.12}$$

where n is the number of observed values; y_i and oq_i are the quantiles estimated from the entropy-based distribution and observed quantiles corresponding to the empirical probabilities. In this study, the Gringorten plotting position formula was used to estimate the empirical cumulative (non-exceedance) probability which is expressed as [*Gringorten*, 1963]:

$$P_i = \frac{i - 0.44}{n + 0.12} \tag{7.13}$$

where P_i is the empirical probability for the i^{th} ordered observation from a record of length n.

The information-based measure, Akaike Information Criterion (AIC), developed by *Akaike* [1974] was also used for identifying the appropriate distribution defined as:

$$\mathrm{AIC} = -2\log(L) + 2\mathrm{N} \tag{7.14}$$

where L is the maximum likelihood of the model and N is the number of fitted parameters. The appropriate model is the one with the minimum AIC value.

7.3.3 Comparison

There are a variety of distributions that can be used as marginal distributions for modeling drought variables. In this study, the Gamma and Weibull distributions were selected as candidates as marginal distributions. The performance of the proposed entropy-based distribution was compared with these two distributions and a specific distribution was considered to perform best when it had minimum RMSE or AIC values. The number of cases that the ENT distribution performed best based on RMSE or AIC are shown in Table 7. 1 .

Table 7. 1 Number of cases of ENT distribution with the best performance for different types of datasets.

Data	Measure	Drought duraiton		Drought severity	
		RMSE	AIC	RMSE	AIC
PDSI (Total: 10)	ENT	10	4	9	3
	Gamma	0	6	0	0
	Weibull	0	0	1	7
SPI (Total: 10)	ENT	8	0	8	2
	Gamma	1	10	2	6
	Weibull	1	0	0	2
Streamflow (Total: 14)	ENT	14	9	12	7
	Gamma	0	5	0	1
	Weibull	0	0	2	6

Drought Variables from PDSI data

Palmer [1965] developed the PDSI as a measure of drought severity that incorporated precipitation, temperature and soil moisture. PDSI is among the most widely used drought indices for assessing a long term meteorological drought. It is a standardized measure with PDSI value larger than 4 representing extremely wet and less than -4 representing extreme drought. A drought event is defined when the PDSI is continuously negative.

For the PDSI data, the ENT distribution outperformed other two distributions for drought duration and severity based on the RMSE of quantiles. For example, for drought duration, the ENT distribution performed best for all 10 cases. Based on the AIC measure, the ENT distribution performed the best for 4 cases for drought duration and 3 for drought severity of the 10 datasets. These results show that the proposed entropy based distribution would be a good candidate in modeling drought properties of the PDSI data.

Drought Variables from SPI data

Standard Precipitation Index (SPI) is a probability drought index developed by *McKee et al.* [1993] based solely on precipitation information . The key feature of PSI is that drought can be measured at different time scales and thus both short-term and long-term droughts can be characterized. SPI is also a standardized measure with negative values for drought condition and positive values for wet condition. An SPI value larger than 2 represents extremely wet, whereas SPI value less than -2 represents extremely dry. A drought event is defined when the SPI is continuously negative.

The performance of the entropy based distribution for 1-month SPI data is also shown in Table 7. 1 . The ENT distribution performed well for modeling quantiles for SPI data. For example, ENT performed best for 8 out of 10 cases for drought severity. However, the ENT distribution did not perform as well based on the AIC values, since ENT distribution performed the best only for 0 and 2 cases for drought duration and severity, respectively. These results showed that the ENT distribution did not perform well for drought duration data based on SPI indices, while still was a good candidate for modeling drought severity of SPI data.

Drought variables from streamflow data

Streamflow deficit is one of the fundamental issues in water resources systems and is also used in this study. Streamflow drought was defined in terms of streamflow deficit from a certain truncation level and has been employed for drought analysis in a number of studies [*Zelenhasi and Salvai*, 1987; *Song and Singh*, 2010]. In this study,

monthly mean streamflow was used as the truncation level to define drought duration and severity.

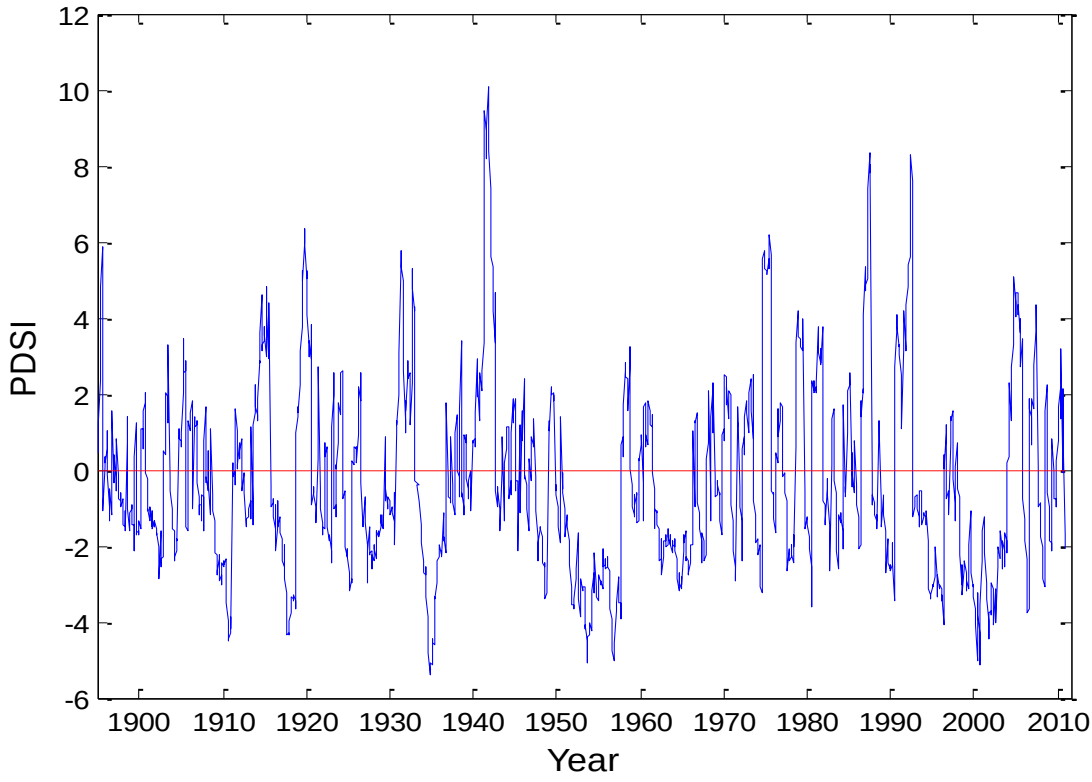

Figure 7. 1 Monthly PDSI data of Climate Division 5 in Texas.

From drought severity of streamflow datasets, ENT performed the best for 14 and 12 cases for drought duration and severity based on the RMSE measure. The ENT distribution performed best for 9 and 7 cases for drought duration and severity for all 14 cases based on the AIC values. Thus, generally the ENT distribution was found to be a

good candidate for modeling drought defined in terms of water deficit from streamflow data.

7.4 Case study

Monthly PDSI data from January 1895 to December 2009 of Climate Division 5, Texas, which is shown in Figure 7. 1, were used for illustrating the application of the proposed entropy-copula method for drought analysis. It was observed that PDSI was below zero for a relatively long time in the 1950s. This drought event occurred from August 1950 to September 1957 with drought duration of 86 months and severity of 262.6 from the PDSI data. The catastrophic drought in Texas in the 1950s was by far the worst in recorded history not only due to its intensity and coverage but also its persistence [*Lowry and Engineers*, 1959; *Riggio et al*., 1987]. From the PDSI data in this study, this drought event was also observed as the most severe one. Another drought is observed from January 1998 to December 2003 with drought duration of 72 months and severity of 196.8. These two drought events during the 1950s and 1998-2003 were selected for this study.

Table 7. 2 RMSE and AIC values of different distributions for the case study.

Drought variable	Measure	Distributions		
		ENT	Gamma	Weibull
Drought duration	RMSE	0.20	0.59	0.52
	AIC	171.06	175.94	175.31
Drought severity	RMSE	0.09	0.19	0.16
	AIC	-117.63	-102.78	-112.11

7.4.1 Marginal distribution

The proposed ENT distribution, Gamma distribution and Weibull distribution were selected as candidates for marginal distributions. The performance of these distributions in modeling the underlying data was compared based on RMSE or AIC. Results from these two measures for these distributions are shown in Table 7. 2

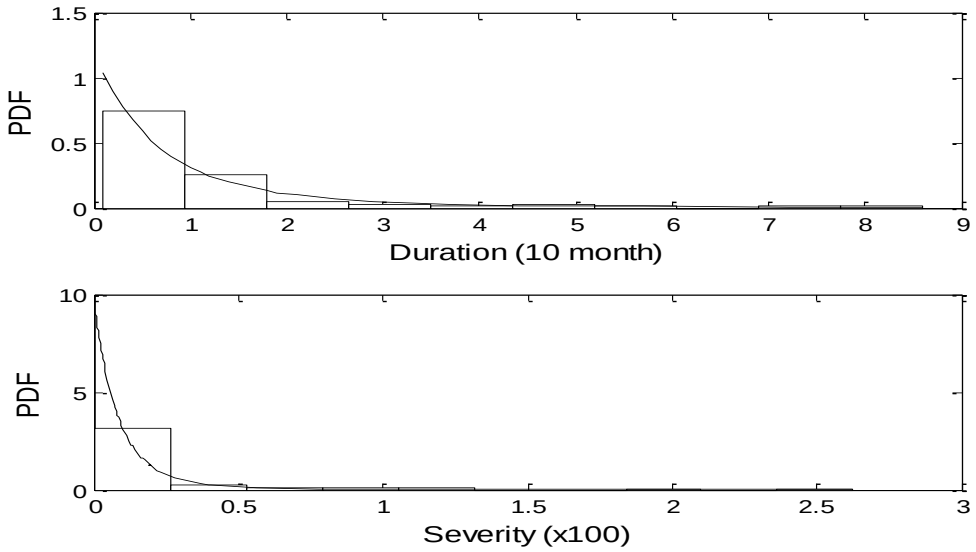

Figure 7. 2 Empirical histograms and entropy based-probability density function.

The ENT distribution outperformed other two distributions for drought duration and likewise for drought severity. For example, the AIC value for drought severity for the ENT distribution was -117.63, while that for gamma and Weibull distribution was -102.78 and -112.11, respectively. Thus the ENT distribution was selected for modeling drought duration and severity from the PDSI data. Empirical histograms and probability density functions (PDF) for drought duration and severity are shown in Figure 7. 2 .

Generally theoretical PDFs fitted the empirical histograms well. The cumulative probability estimated from the entropy based distribution and that from the empirical plotting position formula are shown in Figure 7. 3 . These results showed that the entropy-based distribution was satisfactory for modeling drought duration and severity.

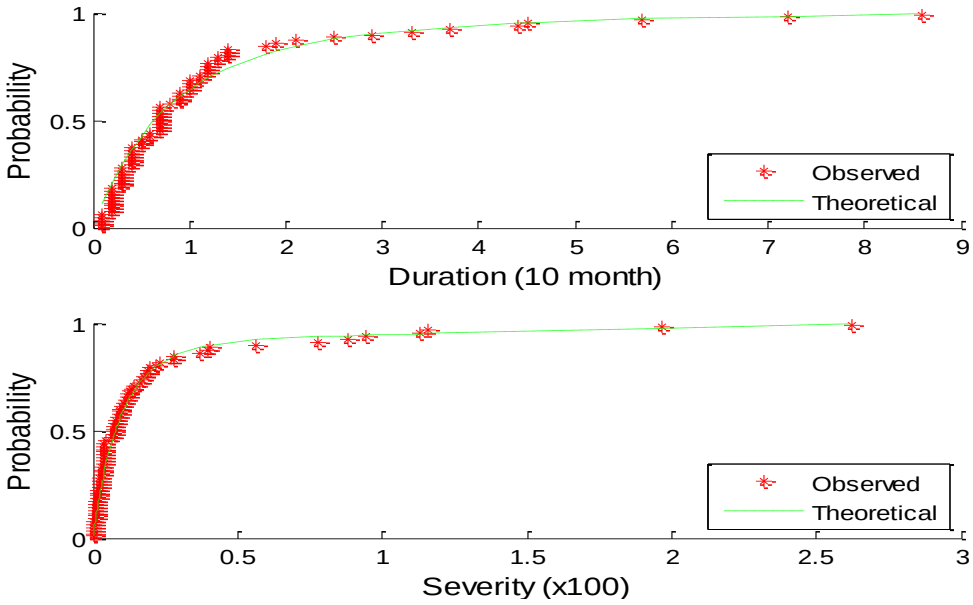

Figure 7. 3 Empirical and entropy-based cumulative distribution function.

7.4.2 Entropy-copula based joint distribution

The entropy-based marginal distributions in equation (7.9) were used for constructing the entropy-copula based joint distribution in equation (7.11). Three copulas from the Archimedian family, namely Clayton, Frank, and Gumbel, were selected to construct the joint distribution for comparison. The inference function for marginal (IFM) method was used to estimate the copula parameter in which parameters

of the marginal distribution and joint distribution were split [*Joe*, 1997]. The parameter of the copula were estimated with maximum likelihood method whereas the copula is selected based on the AIC values. In this study, the Gumbel copula had minimum AIC values and thus was selected to construct the joint distribution.

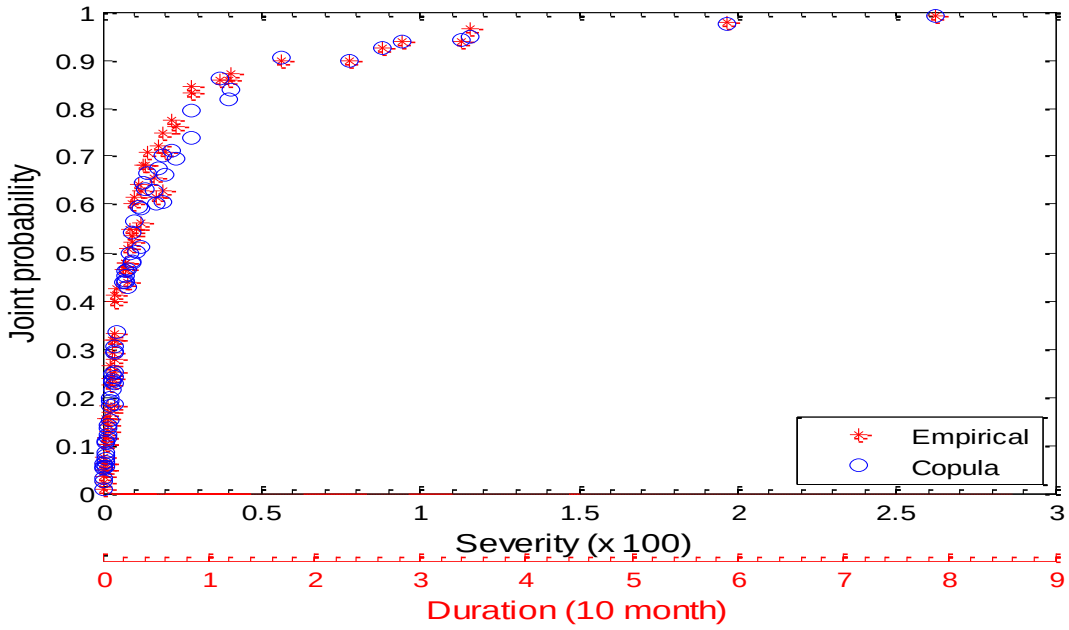

Figure 7. 4 Comparison of empirical and theoretical joint probability distributions.

For the drought duration and severity pairs (d_1, s_1), (d_2, s_2),.., (d_m, s_m), the empirical joint cumulative (non-exceedance) probability can be expressed as [*Yue et al.*, 1999]:

$$P(D \le d_k, S \le s_k) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} - 0.44}{m + 0.12} \tag{7.15}$$

where m is the length of the observation; and n_{ij} is the number of occurrences of the pair (d_i, s_i) for $d_i \le d_k$, $s_i \le s_k$ for $1 \le i \le k$. For the observed drought duration and severity pairs, the theoretical joint probability can be computed through equation (7.11).

Comparison of empirical and theoretical joint probability distributions is shown in Figure 7. 4. Generally the theoretical probability distribution fitted the empirical distribution well and thus the entropy-copula based joint distribution was capable of modeling drought duration and severity jointly well.

Table 7. 3 Univariate return period for drought duration and severity.

Return period (years)	Drought duration (months)	Drought severity
2	2.2	2.6
5	11.5	14.3
10	21.0	27.5
20	33.6	57.0
50	54.8	175.9
100	70.8	213.6

7.4.3 Drought analysis

A common approach to drought analysis is based on fitting distributions to drought variables and then analyzing return periods corresponding to some occurrence levels of drought events. The univariate return period for drought duration *D* and for drought severity *S* can be defined as [*Shiau et al*., 2007]:

$$T_D = \frac{E(L)}{1 - G_D(d)} \qquad T_S = \frac{E(L)}{1 - G_S(s)} \tag{7.16}$$

where T_D and T_S are the univariate return periods for drought duration D and drought severity S, respectively; $E(L)$ is the mean drought interval time; $G_D(d)$ and $G_S(s)$ are the cumulative distributions of drought duration and drought severity, respectively. The univariate return periods for drought duration and severity for the return period 2, 5, 10, 20, 50 and 100 years are shown in Table 7. 3.

The empirical return period can also be obtained with equation (7.16) with the cumulative probability estimated from the plotting position formula in equation (7.13). For the drought of the 1950s, the theoretical return period was 223 years from drought duration and 765 years from drought severity. From observations, the empirical return period was 206 years, which is close to the theoretical one estimated from drought duration while it differed greatly from that estimated from drought severity.

The empirical return period for the drought 1950s was obtained from equation (7.13) with the rank of the observed values. Though the accuracy of the empirical return period with the highest-rank was low [*Stedinger*, 1993; *Beckers and Alila*, 2004], it is given hereafter for reference. For the drought during the year 1998-2003, the return period was 106 years from drought duration and 69 years from drought severity. The empirical return period for this drought event was 74 years, which is close to the theoretical value estimated from drought severity.

The joint return period of drought duration and severity can be defined for the case when drought duration or severity exceeds specific values ($D \geq d$ or $S \geq s$) (denoted

as Type I joint return period) or the case when both drought duration and severity exceed specific values ($D \geq d$ and $S \geq s$) (denoted as Type II joint return period). The two return periods are expressed as:

$$T_{DS} = \frac{E(L)}{P(D \geq d \quad or \quad S \geq s)} = \frac{E(L)}{1 - F_{DS}(d,s)} \tag{7.17}$$

$$T'_{DS} = \frac{E(L)}{P(D \geq d \quad and \quad S \geq s)} = \frac{E(L)}{1 - G_D(d) - G_S(s) + F_{DS}(d,s)} \tag{7.18}$$

where $F_{DS}(d, s)$ is the joint cumulative probability that can be obtained from equation (7.11).

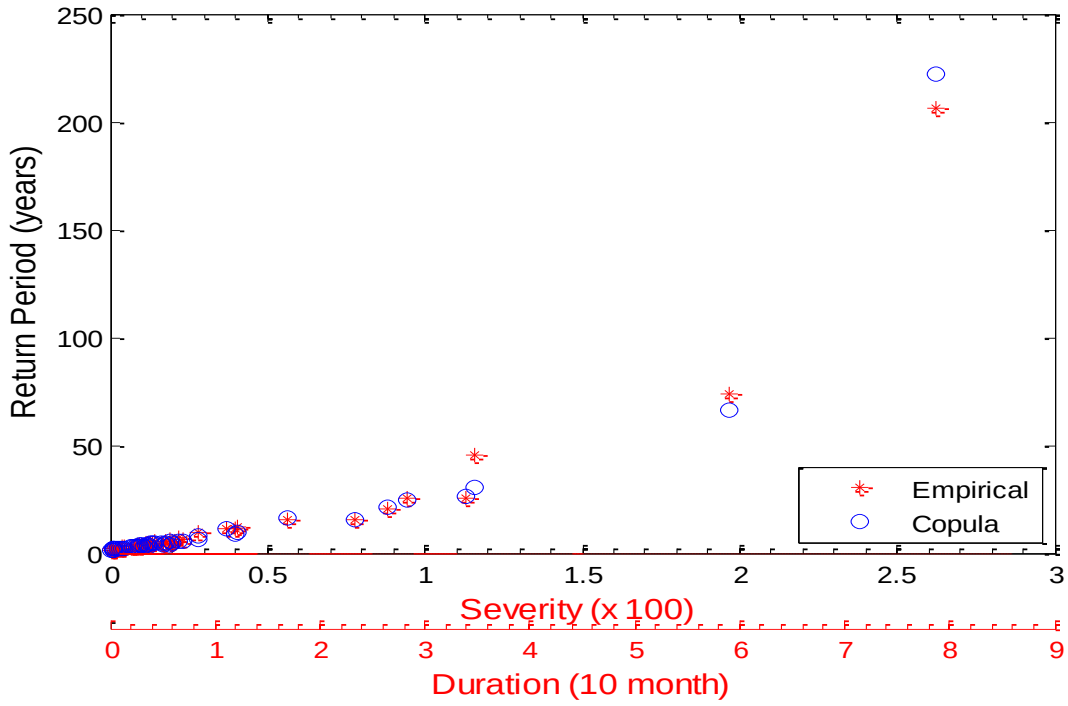

Figure 7. 5 Comparison of theoretical and empirical type I joint return period.

For the observed drought duration and severity pairs, the theoretical joint return period can be obtained from equation (7.17) with the corresponding joint probability estimated from equation (7.11), while the empirical joint return period can also be obtained with the empirical probability estimated from equation (7.15), as shown in Figure 7. 5. Generally the theoretical joint return period for the observed drought duration and severity pairs fitted the empirical one well. For example, the theoretical joint return period during the year 1998-2003 was 66.3 years, which is close to the theoretical return period of 74.0 years.

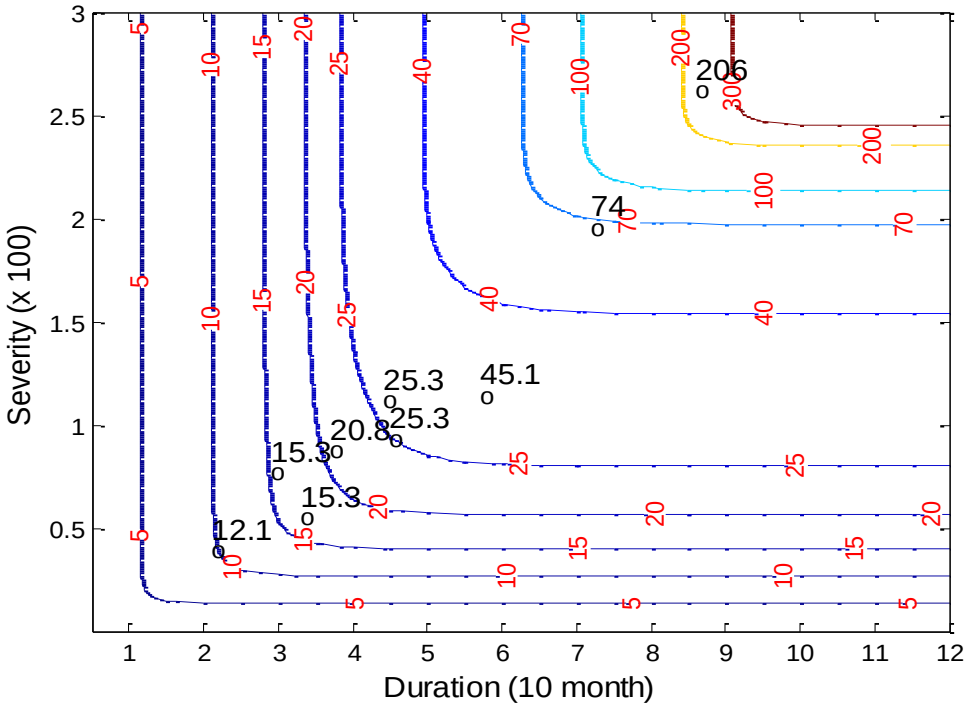

Figure 7. 6 Type I joint return period of drought duration and severity.

The theoretical joint return period for different combinations of drought duration and severity estimated from equation (7.17) is shown in Figure 7. 6. Certain empirical return periods (> 12 years) are also superimposed on Figure 7. 6 for comparison with contour of the return period. It can be seen that the empirical return period values are close to the theoretical values on the nearest contour lines. For the extreme drought during the year 1950s, the corresponding theoretical return period was 222 years, which is close to the empirical return period 206 years. Similarly, the empirical return period 74.0 years corresponding to observed drought during the year 1998-2003 is near the contour line with return period 70 years.

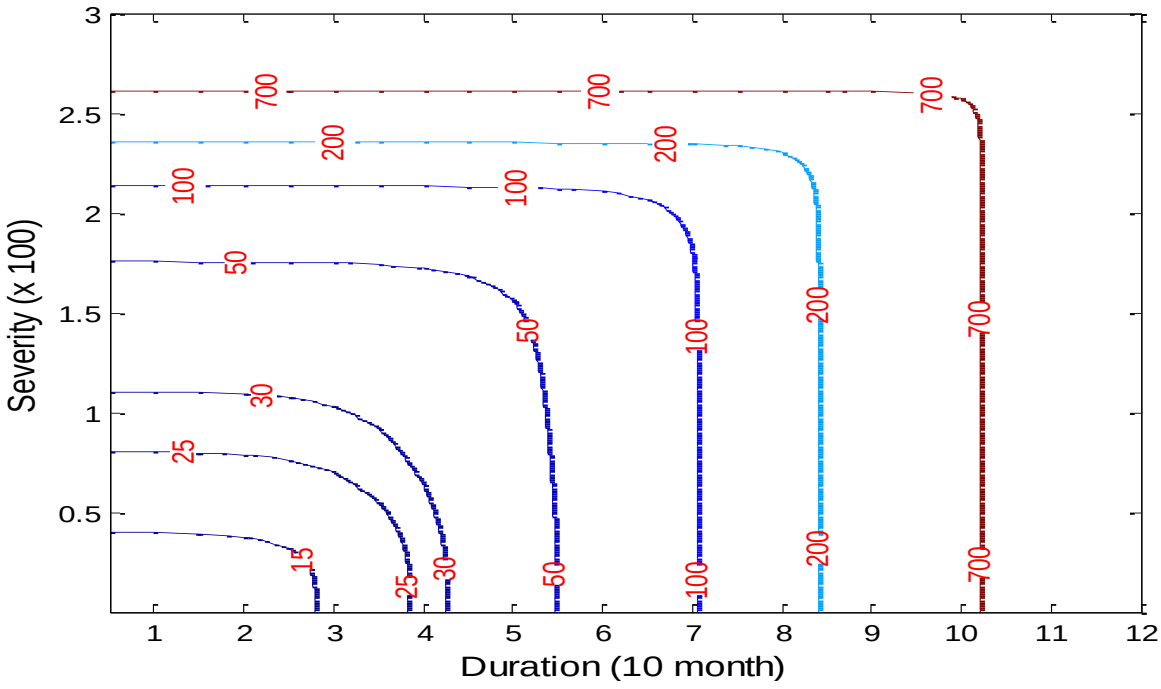

Figure 7. 7 Type II joint return period of drought duration and severity.

The joint return period for different combinations of drought duration and severity estimated from equation (7.18) was also computed, as shown in Figure 7. 7. For the drought in the 1950s, the return period obtained from the proposed method was 770 years, while for the drought during the year 1998-2003, the joint return period was 113 years. Note these return periods are larger than those estimated from drought duration and severity separately.

The conditional return period $T_{D|S\geq s}$ for drought duration given drought severity exceeding a certain threshold *s* can be defined as [*Shiau*, 2003; *Shiau*, 2006]:

$$\mathrm{T}_{\mathrm{D|S}\geq \mathrm{s}} = \frac{T_S}{P(D \geq d, S \geq s)} = \frac{E(L)}{[1-Gs(s)][1-G_D(d)-G_S(d)+C(G_D(d),G_S(s))]} \qquad (7.19)$$

Similarly, the conditional return period for drought severity given drought duration exceeding a certain threshold *d* can be defined as:

$$\mathrm{T}_{\mathrm{S|D}\geq \mathrm{d}} = \frac{T_D}{P(D \geq d, S \geq s)} = \frac{E(L)}{[1-G_D(d)][1-G_D(d)-G_S(d)+C(G_D(d),G_S(s))]} \qquad (7.20)$$

The conditional return period of drought duration given drought severity and drought severity given drought duration is shown in Figure 7. 8. For example, the drought duration corresponding to the conditional return period 100 years was 30 months given $s > 20$. These results can help assess the risk of drought occurrence for water resources planning and management.

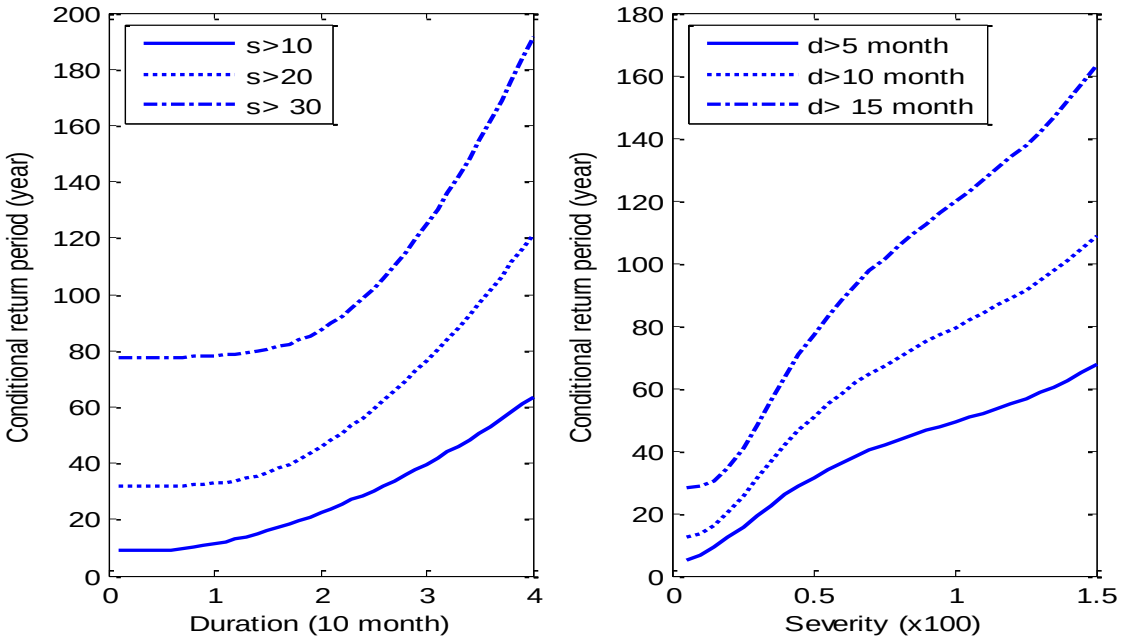

Figure 7. 8 Conditional return period for drought duration given drought severity and drought severity given drought duration.

7.5 Conclusion

The proposed entropy-copula method can be regarded as a general framework for constructing joint distributions. The advantage of the proposed method is that it provides more flexible marginal distributions which can be derived from the principle of maximum entropy with different constraints. The following conclusions can be drawn from this study:

(1) An entropy-based distribution based on the first three moments is found to be a good candidate for modeling drought variables through comparison with other distributions for three different types of datasets.

(2) Based on the PDSI data in Climate Division 5 in Texas, the proposed entropy-copula method is shown to model drought duration and severity separately and jointly well.

(3) For the drought during the 1950s in Texas, the return period is estimated as 223 years from drought duration and 765 from drought severity. The joint return periods obtained from equation (7.17) and (7.18) are around 222 and 770 years, respectively.

(4) For the drought during 1998-2003 in Texas, the return period is estimated as 106 years from drought duration and 69.0 from drought severity. The joint return periods obtained from equation (7.17) and (7.18) are 66 and 113 years, respectively.

CHAPTER VIII

CONCLUSION

The research presented in this study focuses on the application of entropy theory in hydrologic analysis and simulation consisting of rainfall analysis, streamflow simulation and drought analysis. The effect of time duration, climate zone and the distance from the Gulf of Mexico on the frequency distribution of annual rainfall maxima is analyzed and an entropy based distribution is proposed to model extreme rainfall values for rainfall analysis. The entropy and entropy-copula methods are proposed for monthly streamflow simulation that is capable of preserving the statistics of historical streamflow. These two methods are also extended for multi-site annual streamflow simulation. For drought analysis, the entropy method and entropy-copula method are also developed for constructing the joint distribution for drought variables. The following conclusions are drawn from this study.

8.1 Rainfall analysis

(1) The frequency distributions of annual rainfall maxima are highly skewed for short durations, like 15 min, but tend to be smoothed when the duration is relatively long.

(2) The rainfall distributions show different patterns across different regions. In northern and western parts, like the CS and SA climate zones, distributions are sharp; however, they are relatively smooth in the southeast, like the SH climate zone.

(3) The frequency distribution of rainfall near the Gulf of Mexico is smoother than that far away from the Gulf.

(4) Using the Monte Carlo simulation, the entropy based distribution with the first four moments as constraints (ENT4) is shown to be comparable with the commonly used generalized extreme value (GEV) distribution and is preferable for the datasets with high skewness.

(5) The ENT4 distribution is shown to be a good candidate to model annual rainfall maxima of different time duration, climate zones, and distances from the Gulf of Mexico across Texas.

8.2 Streamflow simulation

(1) The entropy based method for single-site monthly streamflow simulation is shown to be capable of satisfactorily preserving the basic statistics (including mean, standard deviation, skewness, maximum values, minimum values) and lag-one correlation of historical streamflow, based on monthly streamflow in the Colorado River basin.

(2) The advantage of the entropy based method is that no assumption is made about the marginal distribution of historical data and data transformation is not needed. The proposed method can be extended to preserve more statistical characteristics (e.g., kurtosis and more lag correlations) while it will be computationally cumbersome when more statistics are to be preserved.

(3) The entropy-copula method for the single-site monthly streamflow simulation is shown to preserve the basic statistics and lag-one correlation well. Furthermore, the nonlinear dependence can also be preserved due to the copula component.

(4) The extended entropy-copula method is shown to improve the preservation of the lag-one correlation at the annual scale, higher-order correlation and inter-annual statistics.

(5) The entropy method and entropy-copula method are extended for the multi-site annual streamflow simulation and shown to preserve the mean, standard deviation and skewness well based on annual streamflow simulation at four sites in the Colorado River basin.

(6) The entropy method preserves the (linear) Pearson correlation of streamflow between different sites well for all cases, while the entropy-copula method does not perform as well for certain sites. However, the entropy-copula method outperforms the entropy method in preserving the Kendall and Spearman correlation.

8.3 Drought analysis

(1) The joint distribution constructed from the entropy method is shown to be capable of modeling drought variables based on the drought data of monthly streamflow of Brazos River at Waco, Texas. Different forms of marginal distributions can be obtained depending on the constraints.

(2) An entropy based distribution with the first three moments as constraints is shown to be a good candidate for modeling drought variables based on three different datasets.

(3) The entropy-copula method for constructing the joint distribution can be regarded as a general framework for drought analysis and is shown to be suitable for modeling the drought properties based on the Parmer drought severity index (PDSI) data in Climate Division 5 in Texas.

(4) For two drought events during the 1950s and during 1998-2003 in Texas, the return periods estimated from drought duration and severity separately and jointly are obtained.

REFERENCES

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control, 19*(6), 716-723.

Alila, Y. (1999), A hierarchical approach for the regionalization of precipitation annual maxima in Canada, *J. Geophys. Res.*, *104*(D24), 31,645-31,655, doi:610.1029/1999JD900764.

Asquith, W. (1998), Depth-duration frequency of precipitation for Texas, *USGS Water–Resources Investigations Report*, *98–4044*, Austin, TX.

Balakrishnan, N., and C. Lai (2009), *Continuous Bivariate Distributions*, Springer, New York.

Beckers, J., and Y. Alila (2004), A model of rapid preferential hillslope runoff contributions to peak flow generation in a temperate rain forest watershed, *Water Resour. Res.*, *40*, W03501, doi:03510.01029/02003WR002582.

Botero, B., and F. Francés (2010), Estimation of high return period flood quantiles using additional non-systematic information with upper bounded statistical models, *Hydrology and Earth System Sciences*, *14*(12), 2617-2628.

Carr, J. (1967), The climate and physiography of Texas, *Report 53*, Texas Water Development Board, Austin, TX.

Clemen, R. T., and T. Reilly (1999), Correlations and copulas for decision and risk analysis, *Management Science*, *45*(2), 208-224.

Dracup, J., K. S. Lee, and E. G. Paulson Jr (1980), On the definition of droughts, *Water Resour. Res.*, *16*(2), 297-302, doi:10.1029/WR016i002p00297.

Fernandez, B., and J. D. Salas (1990), Gamma autoregressive models for stream-flow simulation, *Journal of Hydraulic Engineering*, *116*(11), 1403-1414.

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

Finzi, G., E. Todini, and J. R. Wallis (1975), Comment upon multivariate synthetic hydrology, *Water Resour. Res.*, *11*(6), 844-850, doi:10.1029/WR011i006p00844.

Genest, C., and A.-C. Favre (2007), Everything you always wanted to know about copula modeling but were afraid to ask, *J. Hydrol. Eng.*, *12*(4), 347-368.

Genest, C., A.-C. Favre, J. Béliveau, and C. Jacques (2007), Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data, *Water Resour. Res.*, *43*, W09401, doi:09410.01029/02006WR005275.

Genest, C., J. F. Quessy, and B. Rémillard (2006), Goodness of fit procedures for copula models based on the probability integral transformation, *Scandinavian Journal of Statistics*, *33*(2), 337-366.

González, J., and J. Valdés (2003), Bivariate drought recurrence analysis using tree ring reconstructions, *J. Hydrol. Eng.*, *8*(5), 247-258.

Gotovac, H., V. Cvetkovic, and R. Andricevic (2010), Significance of higher moments for complete characterization of the travel time probability density function in heterogeneous porous media using the maximum entropy principle, *Water Resour. Res.*, *46*, W05502, doi:05510.01029/02009WR008220.

Gringorten, I. I. (1963), A plotting rule for extreme probability paper, *J. Geophys. Res.*, *68*(3), 813-814.

Grygier, J. C., and J. R. Stedinger (1988), Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, *24*(10), 1574-1584, doi:1510.1029/WR1024i1010p01574.

Hao, Z., and V. Singh (2011), Single-site monthly streamflow simulation using entropy theory, *Water Resour. Res.*, *47*, W09528, doi:09510.01029/02010WR010208.

Hao, Z., and V. Singh (2012), Entropy-copula method for single-site monthly streamflow simulation, *Water Resour. Res.*, in revision.

Hipel, K., and A. McLeod (1978), Preservation of the rescaled adjusted range, 2: Simulation studies using Box-Jenkins models, *Water Resour. Res.*, *14*(3), 509-516, doi:10.1029/WR014i003p00509.

Hipel, K., A. McLeod, and E. McBean (1979), Hydrologic generating model selection, *J. Water Resour. Plann. Manage.*, *105*(2), 223-242.

Huff, F. A., and J. R. Angel (1992), Rainfall frequency atlas of the Midwest, *Bulletin 71*, Illinois State Water Survey, Champaign, IL.

Jaynes, E. (1957), Information theory and statistical mechanics, *Physical Review*, *106*(4), 620-630.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Johnson, M. (1987), *Multivariate Statistical Simulation*, John Wiley & Sons, New York.

Kapur, J. (1989), *Maximum-Entropy Models in Science and Engineering*, John Wiley & Sons, New York.

Kendall, D., and J. Dracup (1992), On the generation of drought events using an alternating renewal-reward model, *Stochastic Hydrology and Hydraulics*, *6*(1), 55-68.

Kesavan, H., and J. Kapur (1992), *Entropy Optimization Principles with Applications*, Academic Press, New York.

Kim, T., J. Valdés, and C. Yoo (2003), Nonparametric approach for estimating return periods of droughts in arid regions, *J. Hydrol. Eng.*, *8*(5), 237-246.

Kim, T., J. Valdes, and C. Yoo (2006), Nonparametric approach for bivariate drought characterization using Palmer drought index, *J. Hydrol. Eng.*, *11*(2), 134-143.

Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, *32*(7), 2105-2117, doi:2110.1029/2196WR00488.

Lall, U. (1995), Recent advance in nonparametric function estimation, *U.S. Natl. Rep. Int. Union Geod. Geophys. 1991–1994*, *Rev. Geophys.*, *33*, 1093-1102.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, *32*(3), 679-693, doi:10.1029/95WR02966.

Larkin, T., and G. Bomar (1983), Climatic atlas of Texas, *Report LP-192*, Texas Dept. of Water Resources, Austin, TX.

Lee, T., and J. D. Salas (2006), Record extension of monthly flows for the Colorado River system, Bureau of Reclamation, U.S. Dept. of the Interior, Denver.

Lee, T., J. D. Salas, and J. Prairie (2010), An enhanced nonparametric streamflow disaggregation model with genetic algorithm, *Water Resour. Res.*, *46*, W08545, doi:10.1029/2009WR007761.

Lee, T., and J. D. Salas (2011), Copula-based stochastic simulation of hydrological data applied to Nile River flows, *Hydrology Research*, *42*(4), 318-330.

Lettenmaier, D. P., and S. J. Burges (1977), An operational approach to preserving skew in hydrologic models of long-term persistence, *Water Resour. Res.*, *13*(2), 281-290, doi:210.1029/WR1013i1002p00281.

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

Lowry, R. L. (1959), A study of droughts in Texas, *Bulletin 5914*, Texas Board of Water Engineers, Austin, TX.

Matalas, N. C. (1967), Mathematical assessment of synthetic hydrology, *Water Resour. Res.*, *3*(4), 937-945.

Mathier, L., L. Perreault, B. Bobée, and F. Ashkar (1992), The use of geometric and gamma-related distributions for frequency analysis of water deficit, *Stochastic Hydrology and Hydraulics*, *6*(4), 239-254.

Matz, A. (1978), Maximum likelihood parameter estimation for the quartic exponential distribution, *Technometrics*, *20*(4), 475-484.

McKee, T. B., N. J. Doesken, and J. Kleist (1993). The relationship of drought frequency and duration to time scales, in *Eighth Conference on Applied Climatology*, Am. Meteorol. Soc., Anaheim, CA.

Mead, L., and N. Papanicolaou (1984), Maximum entropy in the problem of moments, *J. Math. Phys*, *25*(8), 2404-2417.

Mejia, J., and J. Rousselle (1976), Disaggregation models in hydrology revisited, *Water Resour. Res.*, *12*(2), 185-186, doi:110.1029/WR1012i1002p00185.

Nadarajah, S. (2007), A bivariate gamma model for drought, *Water Resour. Res.*, *43*, W08501, doi:08510.01029/02006WR005641.

Nadarajah, S. (2009), A bivariate pareto model for drought, *Stochastic Environmental Research and Risk Assessment*, *23*(6), 811-822.

Narasimhan, B., R. Srinivasan, S. Quiring, and J. Nielsen-Gammon (2008), Digital climatic atlas of Texas, *Report 2005-483-5591*, Texas Water Development Board Contract, Texas A&M University, TX.

National Fibers Information Center (1987), *The Climate of Texas Counties*, University of Texas, Austin and Texas A&M University, College Station, TX.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer, New York.

North, G., J. Schmandt, and J. Clarkson (1995), *The Impact of Global Warming on Texas*, University of Texas Press, Austin, TX.

Nowak, K., J. Prairie, B. Rajagopalan, and U. Lall (2010), A non-parametric stochastic approach for multisite disaggregation of annual to daily streamflow, *Water Resour. Res.*, *46*, W08529, doi:10.1029/2009WR008530.

Nowak, K., B. Rajagopalan, and E. Zagona (2011), Wavelet Auto-Regressive Method (WARM) for multi-site streamflow simulation of data with non-stationary spectra, *Journal of Hydrology*, *410*(1-2), 1-12.

Palmer, W. (1965), Meteorological drought, *Weather Bur. Res. Pap. 45*, U.S. Dep. of Commer., Washington, D. C.

Parrett, C. (1997), Regional analysis of annual precipitation maxima in Montana, *USGS Water-Resources Investigations Report, 97-4004*, Helena, Mont.

Prairie, J., B. Rajagopalan, T. Fulp, and E. Zagona (2006), Modified K-NN model for stochastic streamflow simulation, *J. Hydrol. Eng.*, *11*(4), 371-378.

Prairie, J., B. Rajagopalan, U. Lall, and T. Fulp (2007), A stochastic nonparametric technique for space-time disaggregation of streamflows, *Water Resour. Res.*, *43*, W03432, doi:03410.01029/02005WR004721.

Riggio, R. F., G. W. Bomar, and T. J. Larkin (1987), Texas drought: Its recent history (1931-1985), *Report LP 87-04*, Texas Water Commission, Austin, TX.

Salas, J., and J. Delleur (1980), *Applied Modeling of Hydrologic Time Series*, Water Resources Publication, Littleton, Colo.

Salas, J. D., F. Fu, A. Cancelliere, D. Dustin, D. Bode, A. Pineda, and E. Vincent (2005), Characterizing the severity and risk of drought in the Poudre River, Colorado, *J. Water Resour. Plann. Manage.*, *131*(5), 383-393.

Salas, J., and T. Lee (2010), Nonparametric simulation of single-site seasonal streamflows, *J. Hydrol. Eng.*, *15*(4), 284-296.

Salvadori, G. (2007), *Extremes in Nature: An Approach Using Copulas*, Springer, New York.

Santos, E., and J. Salas (1992), Stepwise disaggregation scheme for synthetic hydrology, *J. Hydrol. Eng.*, *118*(5), 765-784.

Savic, D., D. Burn, and Z. Zrinji (1989), Comparison of streamflow generation models for reservoir capacity-yield analysis, *Water Resour. Bull.*, *25*(5), 977-983.

Schaefer, M. (1990), Regional analyses of precipitation annual maxima in Washington State, *Water Resour. Res.*, *26*(1), 119-131, doi:10.1029/WR026i001p00119.

Schölzel, C., and P. Friederichs (2008), Multivariate non-normally distributed random variables in climate research–introduction to the copula approach, *Nonlinear Processes in Geophysics*, *15*(5), 761–772.

Shannon, C. E. (1948), A mathematical theory of communications, *Bell Syst. Tech. J.* , *27*(7), 379-423.

Shannon, C., and W. Weaver (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.

Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res.*, *38*(7), 1100, doi:1110.1029/2001WR000953.

Sharma, A., D. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*(2), 291-308, doi:210.1029/1096WR02839.

Shiau, J. (2003), Return period of bivariate distributed extreme hydrological events, *Stochastic Environmental Research and Risk Assessment*, *17*(1), 42-57.

Shiau, J. (2006), Fitting drought duration and severity with two-dimensional copulas, *Water resources management*, *20*(5), 795-815.

Shiau, J., S. Feng, and S. Nadarajah (2007), Assessment of hydrological droughts for the Yellow River, China, using copulas, *Hydrological Processes*, *21*(16), 2157-2163.

Shiau, J., and H. Shen (2001), Recurrence analysis of hydrologic droughts of differing severity, *J. Water Resour. Plann. Manage.*, *127*(1), 30-40.

Shiau, J. T., and R. Modarres (2009), Copula based drought severity duration frequency analysis in Iran, *Meteorological Applications*, *16*(4), 481-489.

Singh, V. P. (1992), *Elementary Hydrology*, Prentice Hall, Upper Saddle River, N.J.

Singh, V. P. (1998), *Entropy-Based Parameter Estimation in Hydrology*, Springer, New York.

Sivakumar, B., and R. Berndtsson (2010), *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*, World Scientific, Hackensack, N. J.

Sklar, A. W. (1959), Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, *8*, 229–231.

Smith, J. (1993), Moment methods for decision analysis, *Management Science*, *39*(3), 340-358.

Smith, O., S. Adelfang, and J. Tubbs (1982), A bivariate gamma probability distribution with application to gust modeling, *NASA Tech. Mem. 82483*, Marshall Space Flight Center, Huntsville, Alabama.

Song, S., and V. Singh (2010), Frequency analysis of droughts using the Plackett copula and parameter estimation by genetic algorithm, *Stochastic Environmental Research and Risk Assessment*, *24*(5), 783-805.

Srinivas, V., and K. Srinivasan (2005), Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows, *Journal of Hydrology*, *302*(1-4), 307-330.

Stedinger, J., D. Pei, and T. Cohn (1985), A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res.*, *21*(5), 665-675, doi:610.1029/WR1021i1005p00665.

Stedinger, J. R., R. M. Vogel, and E. Foufoula-Georgiou (1993), Frequency analysis of extreme events, in *Handbook of Hydrology*, edited by D. R. Maidment, pp. 18.1–18.66, McGraw-Hill, New York.

Tarboton, D., A. Sharma, and U. Lall (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*(1), 107-119, doi:110.1029/1097WR02429.

Trivedi, P. K., and D. M. Zimmer (2005), Copula modeling: an introduction for practitioners, *Foundations and Trends in Econometrics*, *1*(1), 1-111.

Valencia, R. D., and J. J. Schaake (1973), Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, *9*(3), 580-585, doi:510.1029/WR1009i1003p00580.

Vogel, R., and A. Shallcross (1996), The moving blocks bootstrap versus parametric time series models, *Water Resour. Res.*, *32*(6), 1875-1882, doi:1810.1029/1896WR00928.

Vogel, R., and J. Stedinger (1988), The value of stochastic streamflow models in overyear reservoir design applications, *Water Resour. Res.*, *24*(9), 1483–1490, doi:1410.1029/WR1024i1009p01483.

Wang, W., and M. T. Wells (2000), Model selection and semiparametric inference for bivariate failure-time data, *Journal of the American Statistical Association, 95*(449), 62-72.

Wu, X. (2003), Calculation of maximum entropy densities with application to income distribution, *Journal of Econometrics, 115*(2), 347-354.

Yevjevich, V. (1967), An objective approach to definitions and investigations of continental hydrologi droughts, *Hydrology Paper 23*, Colorado State Univ., Fort Collins.

Yue, S., T. Ouarda, B. Bobée, P. Legendre, and P. Bruneau (1999), The Gumbel mixed model for flood frequency analysis, *Journal of Hydrology*, *226*(1-2), 88-100.

Zelenhasi , E., and A. Salvai (1987), A method of streamflow drought analysis, *Water Resour. Res.*, *23*(1), 156-168, doi:10.1029/WR023i001p00156.

Zellner, A., and R. Highfield (1988), Calculation of maximum entropy distributions and approximation of marginal posterior distributions, *Journal of Econometrics*, *37*(2), 195-209.

VITA

Name:	Zengchao Hao
Address:	321E Scoates Hall, 2117, TAMU,
	College Station, TX 77843-2117
Email Address:	hzc07@tamu.edu
Education:	B.S., Agricultural Engineering,
	China Agricultural University, China, 2005
	M.S., Civil Engineering,
	Tsinghua University, China, 2007
	Ph.D., Biological and Agricultural Engineering,
	Texas A&M University, U.S., 2012