WATER RESOURCES RESEARCH, VOL. 49, 767–789, doi:10.1002/wrcr.20063, 2013

A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation

Chao Li,[1] Vijay P. Singh,[1,2] and Ashok K. Mishra[1,3]

Received 22 December 2011; revised 25 October 2012; accepted 14 December 2012; published 6 February 2013.

[1] This paper presents an improved bivariate mixed distribution, which is capable of modeling the dependence of daily rainfall from two distinct sources (e.g., rainfall from two stations, two consecutive days, or two instruments such as satellite and rain gauge). The distribution couples an existing framework for building a bivariate mixed distribution, the theory of copulae and a hybrid marginal distribution. Contributions of the improved distribution are twofold. One is the appropriate selection of the bivariate dependence structure from a wider admissible choice (10 candidate copula families). The other is the introduction of a marginal distribution capable of better representing low to moderate values as well as extremes of daily rainfall. Among several applications of the improved distribution, particularly presented here is its utility for single-site daily rainfall simulation. Rather than simulating rainfall occurrences and amounts separately, the developed generator unifies the two processes by generalizing daily rainfall as a Markov process with autocorrelation described by the improved bivariate mixed distribution. The generator is first tested on a sample station in Texas. Results reveal that the simulated and observed sequences are in good agreement with respect to essential characteristics. Then, extensive simulation experiments are carried out to compare the developed generator with three other alternative models: the conventional two-state Markov chain generator, the transition probability matrix model, and the semiparametric Markov chain model with kernel density estimation for rainfall amounts. Analyses establish that overall the developed generator is capable of reproducing characteristics of historical extreme rainfall events and is apt at extrapolating rare values beyond the upper range of available observed data. Moreover, it automatically captures the persistence of rainfall amounts on consecutive wet days in a relatively natural and easy way. Another interesting observation is that the recognized "overdispersion" problem in daily rainfall simulation ascribes more to the loss of rainfall extremes than the under-representation of first-order persistence. The developed generator appears to be a sound option for daily rainfall simulation, especially in particular hydrologic planning situations when rare rainfall events are of great importance.

Citation: Li, C., V. P. Singh, and A. K. Mishra (2013), A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation, *Water Resour. Res.*, *49*, doi:10.1002/wrcr.20063.

1. Introduction

[2] Daily rainfall is a major input to drive many models of hydrologic, agricultural, ecological, and other environmental systems [*Mehrotra et al.*, 2012; *Kleiber et al.*, 2011]. A great deal of attention has therefore been devoted to daily rainfall modeling. Considering the fact that daily rainfall is non-negative with point mass at zero, a discrete-continuous mixed distribution with a probability density function (PDF) of the following form is obtained:

$$\phi(x) = (1 - p_1)\delta(x) + p_1 f(x), \tag{1}$$

[3] This form is usually used to represent the at-site distribution of daily rainfall X, where p_1 is the probability of rainfall occurrence; $\delta(x)$ is the one-dimensional Dirac delta function, which becomes ∞ if and only if x is 0, and becomes 0 otherwise; and $f(x)$ is a skewed density for rainfall amounts. Discrete-continuous mixed distributions of this form have been used in the literature for daily rainfall downscaling [*Cannon*, 2008; *Carreau and Vrac*, 2011].

[4] To simultaneously model multiple rainfall series (e.g., rainfall at multiple sites or of several successive days), it is logical to extend the univariate distribution in

Additional Supporting Information may be found in the online version of this article.

[1]Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA.
[2]Department of Civil and Environmental Engineering, Texas A&M University, College Station, Texas, USA.
[3]Pacific Northwest National Laboratory, Richland, Washington, USA.

Corresponding author: C. Li, Department of Biological and Agricultural Engineering, Texas A&M University, College Station, TX 77843-2117, USA. (lichsunny@gmail.com)

0043-1397/13/10.1002/wrcr.20063

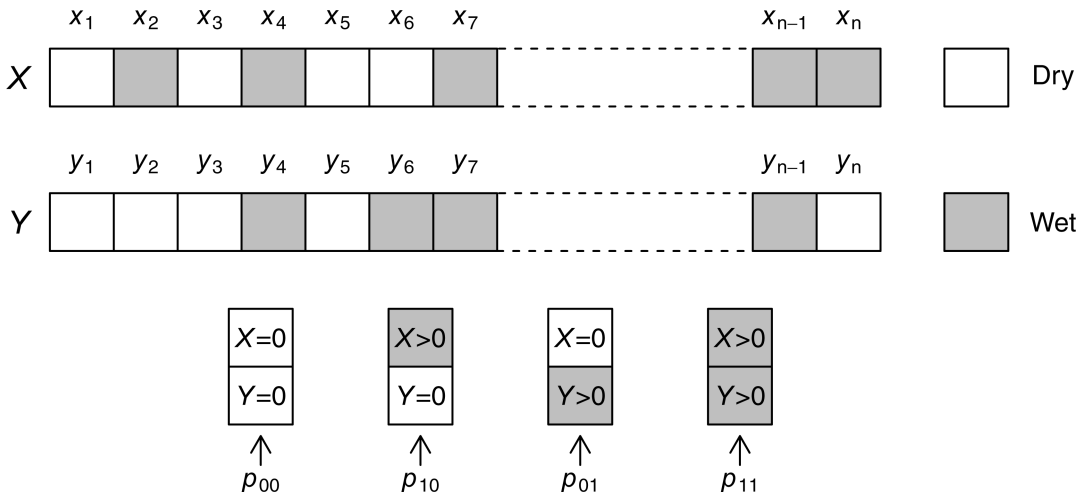

Figure 1. Schematic showing the notation of two discrete-continuous rainfall series.

equation (1) to its multivariate analog. In theory, there is nothing to limit building a joint discrete-continuous mixed distribution for fully multivariate analysis. In practice, however, this is hardly achievable, as the model complexity increases for higher powers (2^d for d-dimension model). One simple extension is to the bivariate level, which models the pairwise dependence of daily rainfall X and Y. The usefulness of a bivariate discrete-continuous mixed distribution can be recognized in several aspects. For example,

[5] 1. If X and Y denote daily rainfall of two consecutive days, then from the bivariate distribution one may derive the conditional distribution of rainfall of current day given that of previous day, which serves as the "engine" for single-site daily rainfall simulation.

[6] 2. If X and Y are spatially averaged rainfall of two neighboring watersheds or rainfall of two rain gauges, then one may use the bivariate distribution for simultaneous simulation of rainfall series while preserving their dependence structure.

[7] 3. If X represents satellite or radar rainfall estimates and Y denotes ground observations, then a best guess (regression) or a conditional distribution (ensemble regression) of actual rainfall given that of satellite or radar estimate may be yielded from the bivariate distribution.

[8] Thus, a well-formulated bivariate discrete-continuous mixed distribution would have much practical appeal.

[9] There are some models, in the context of rainfall simulation, that do allow multisite modeling of daily rainfall. In addition to the multivariate autoregressive model of *Bárdossy and Plate* [1992], the nonparametric hidden Markov chain model developed by *Hughes and Guttorp* [1994], the nearest neighbor bootstrap technique of *Rajagopalan and Lall* [1999], and the regionalized daily rainfall generation approach of *Mehrotra et al*. [2012], another notable multivariate modeling framework is the one proposed by *Wilks* [1998, 2009]. In this framework, each site follows its own model, while the dependence among sites is maintained by driving individual models with spatially correlated random variates. Owing to its advantage of being simple in extending from single-site to multisite simulation, this framework has been frequently used and improved. For instance, *Mehrotra and Sharma* developed semiparametric and nonparametric multisite models for daily rainfall simulation [2007a, 2007b] and downscaling [2005, 2010]; *Thompson et al*. [2007], *Brissette et al*. [2007], and *Tarpanelli et al*. [2012] improved it such that the correlated random variates can be efficiently generated. It must, however, be realized that the aforementioned multisite models are designed specifically for rainfall simulation rather than formulating a joint distribution for multiple rainfall series. They might be unsuitable for the application in scenario 3 as listed above, unless additional efforts are made to reformulate the models. A multivariate or bivariate mixed distribution might be used not only for simulation but also for statistical inference (regression and ensemble regression), which is applicable for situations similar to scenario 3.

[10] We return to the problem of formulating a bivariate discrete-continuous mixed distribution. The first work on this type of distribution was introduced by *Shimizu* [1993]. Given that both X and Y are zero-inflated random variables, there are four possible mutually exclusive classes, as illustrated in Figure 1,

$$S := \{[X = 0, Y = 0], [X > 0, Y = 0], [X = 0, Y > 0], [X > 0, Y > 0]\}.$$

[11] By the rule of total probability, a bivariate PDF analogous to equation (1) is structured as

$$\phi(x,y) = p_{00}\delta(x,y) + p_{10}h_X(x)\delta(y) + p_{01}h_Y(y)\delta(x) + p_{11}h(x,y), \tag{2}$$

and the corresponding cumulative distribution function (CDF) is

$$\Phi(x,y) = p_{00} + p_{10}H_X(x) + p_{01}H_Y(y) + p_{11}H(x,y), \tag{3}$$

where

$$p_{00} = P(X = 0, Y = 0),$$
$$p_{10} = P(X > 0, Y = 0),$$
$$p_{01} = P(X = 0, Y > 0),$$
$$\text{and} \quad p_{11} = P(X > 0, Y > 0)$$

represent the occurrence probabilities of the four classes, respectively; $\delta(x, y)$ is the two-dimensional Dirac delta

function which yields ∞ if and only if both x and y are 0, and yields 0 otherwise; $\delta(x)$ and $\delta(y)$ hold the same meaning as in equation (1); $h_X(x)$, $h_Y(y)$, and $h(x, y)$ are the PDFs of rainfall amounts within relevant classes, respectively; $H_X(x)$, $H_Y(y)$, and $H(x, y)$ are the corresponding CDFs.

[12] After the pioneering work of *Shimizu* [1993], the bivariate mixed distribution has been applied to investigate the properties of the Pearson's correlation coefficient between rainfall gauges [*Habib et al*., 2001; *Ha and Yoo* 2007; *Yoo and Ha*, 2007] and has been improved such that the joint behavior of contemporaneous rainfall amounts can be properly modeled [*Herr and Krzysztofowicz*, 2005]. The most recent treatment of this distribution was given by *Serinaldi* [2008, 2009a, 2009b], in which the copula theory was used to construct the joint density $h(x, y)$. The copula theory does circumvent restrictions to the marginal distributions and can model different dependence structures with different copulae. This model is not yet without limitation. First, a limited number of copula families (sometimes even one comprehensive family) were used, which may not suffice to describe various autocorrelation structures of rainfall amounts and may thus be of limited use to simulate rainfall of different climate areas. Moreover, the significance of marginal distributions was overlooked. The Weibull, gamma, and Pearson type III distributions were used for the marginal distributions of rainfall amounts, i.e., for $h_X(x)$, $h_Y(y)$ and the margins of $h(x, y)$ [*Serinaldi*, 2009b]. These distributions perform well in describing the usual behavior of rainfall. However, they might not necessarily perform well in capturing unusual behavior or rare events [*Vrac and Naveau*, 2007; *Furrer and Katz*, 2008; *Hundecha et al*., 2009; *Carreau et al*., 2009; *Carreau and Vrac*, 2011; *Hundecha and Merz*, 2012], as can be seen from Figure 2, which shows observed against modeled rainfall quantiles by the Weibull, gamma, and Pearson type III distributions, respectively, at a sample station in Texas. Till now, we are not yet aware of research on this bivariate mixed distribution specifically accommodating the heavy-tailed property of rainfall amounts, which is much common for rainfall at finer time scales.

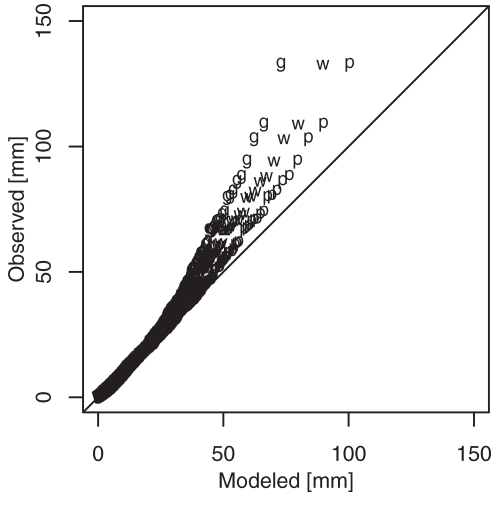

Figure 2. QQ plots of observed versus gamma (g), Weibull (w), and Pearson III (p) modeled quantiles of rainfall amounts of a sample station in Texas.

[13] In view of the above-mentioned limitations, the goal of this research is to further improve the bivariate mixed distribution based on the work of *Shimizu* [1993], *Herr and Krzysztofowicz* [2005], and *Serinaldi* [2009b]. Innovations in the improvements are twofold. One is the appropriate selection of an optimal copula family from a wider choice of admissible candidates such that the joint behavior of rainfall amounts can be realistically modeled. The other one is the introduction of a hybrid distribution for marginal rainfall, which improves the characterization of extremes while retaining a decent fit for low to moderate values. Although the hybrid distribution was reported in our previous work [*Li et al*., 2012], therein only the distribution of single-site rainfall amount was of interest and no mechanism was designed for generating synthetic rainfall series. Here, we extend the hybrid distribution capable of bivariate inference and simulation of daily rainfall, with both occurrence and amount simultaneously taken into account. In addition, by generalizing daily rainfall as a Markov process with autocorrelation described by the improved distribution, a stochastic rainfall generator is developed and analyzed in this research. Although presented here is a single-site model, it may be used as building blocks for multisite simulation following the approach of *Wilks* [1998]. Attributing to the hybrid marginal distributions, characteristics of historical extreme rainfall events can be preserved in the synthetic series and rare rainfall events beyond the upper range of available observed data may be reasonably extrapolated. An implementational merit of the generator is that it unifies rainfall occurrence and amount processes into a single one. As a consequence, the lag-1 autocorrelation of daily rainfall may be automatically captured in a relatively natural and simple way without much extra work if any.

[14] Besides the aforementioned research on multisite rainfall simulation, it is better to mention some other representative single-site models such that one can get an overall picture about the differences between the suggested generator and other alternatives. A typical approach for single-site daily rainfall breaks down the simulation into two stages. The first stage simulates rainfall occurrence process. Among others, the two-state Markov chain model introduced by *Gabriel and Neumann* [1962] has been extensively used. Once the occurrence series is simulated, the second stage simulates rainfall amounts on wet days. To that end, independent random numbers are generated from a fitted parametric distribution, such as exponential distribution [*Todorovic and Woolhiser*, 1975], gamma distribution [*Richardson*, 1981], and mixed exponential distribution [*Wilks*, 1998]. Rather than focusing on rainfall simulation, *Katz* [1974, 1977] derived some important inferential statistics of this model, for instance, probability distributions for the number of wet days, maximum daily rainfall, and rainfall totals over a given period. It is apparent that the suggested generator bears similarity to this model. The differences, also the merits, of the suggested generator are the following: on one hand, instead of breaking down the occurrence and amount processes, it unifies them into a single one; and on the other hand, instead of assuming independence of rainfall amounts of two consecutive wet days, it properly accounts for the dependence.

Table 1. Copula Families Used in this Research[a]

Copula	$C(u, v)$	$\tau=g(\theta)$	Ω^θ	Ω^τ
Clayton	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$1-\frac{2}{2+\theta}$	$[0,\infty)$	$(0, 1]$
Frank	$-\frac{1}{\theta}\ln\left(1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right)$	$1-\frac{4}{\theta}\left(1-\frac{1}{\theta}\int_0^\theta t/(e^t-1)dt\right)$	$R\backslash\{0\}$	$[-1, 1]\backslash\{0\}$
Gumbel	$\exp\left(-\left((-\ln u)^\theta+(-\ln v)^\theta\right)^{1/\theta}\right)$	$1-\frac{1}{\theta}$	$[1,\infty)$	$[0, 1]$
Survival Clayton	$u+v-1+\left((1-u)^{-\theta}+(1-v)^{-\theta}-1\right)^{-1/\theta}$	$1-\frac{2}{2+\theta}$	$[0,\infty)$	$(0, 1]$
A12[b]	$\left(1+\left(\left(u^{-1}-1\right)^\theta+\left(v^{-1}-1\right)^\theta\right)^{1/\theta}\right)^{-1}$	$1-\frac{2}{3\theta}$	$[1,\infty)$	$\left[\frac{1}{3}, 1\right]$
A14[b]	$\left(1+\left(\left(u^{-1/\theta}-1\right)^\theta+\left(v^{-1/\theta}-1\right)^\theta\right)^{1/\theta}\right)^{-\theta}$	$1-\frac{2}{1+2\theta}$	$[1,\infty)$	
FGM	$uv+\theta uv(1-u)(1-v)$	$\frac{2\theta}{9}$	$[-1, 1]$	$\left[-\frac{2}{9}, \frac{2}{9}\right]$
Joe	$1-\left((1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta\right)^{1/\theta}$	No closed form[c]	$[1,\infty)$	$(0, 1]$
Gaussian	$\int_{-\infty}^{\Phi^{-1}(u)}\int_{-\infty}^{\Phi^{-1}(v)}\frac{1}{2\pi\sqrt{1-\theta^2}}\exp\left(\frac{2\theta xy-x^2-y^2}{2\left(1-\theta^2\right)}\right)dxdy$	$\frac{2}{\pi}\arcsin\theta$	$[-1, 1]$	$[-1, 1]$
Student	$\int_{-\infty}^{t_\vartheta^{-1}(u)}\int_{-\infty}^{t_\vartheta^{-1}(v)}\frac{\Gamma((\vartheta+2)/2)}{\Gamma(\vartheta/2)\pi\vartheta\sqrt{1-\theta^2}}\left(1+\frac{x^2-2\theta xy+y^2}{\vartheta}\right)^{(\vartheta+2)/2}dxdy$	$\frac{2}{\pi}\arcsin\theta$	$[-1, 1]$	$[-1, 1]$

$\Phi(\cdot)$, standard Gaussian distribution function; $t_\vartheta(\cdot)$, Student distribution function with ϑ degrees of freedom.
[a]Representative contour plots for each copula family can be found in the Supporting Information.
[b]Numbers denote Archimedean copulas as listed in *Nelson* [2006].
[c]$\tau=g(\theta)$ can be approximated by $\tau=\alpha$ arctan (θ/β) using Monte Carlo simulation.

Reproduction of the structure of daily autocorrelation is recognized as a crucial test for a stochastic rainfall generator [*Gregory et al*., 1993]. There exist alternative models that do not assume independence of rainfall amounts. One is the multistate Markov chain model also known as transition probability matrix (TPM) model [*Haan et al*., 1976; *Srikanthan and McMahon*, 1985; *Srikanthan et al*., 2005]. A second alternative is the nonparametric model developed by *Harrold et al*. [2003a, 2003b]. This model was then adjusted and incorporated into other multisite rainfall simulation and downscaling models [*Mehrotra and Sharma*, 2005, 2007a, 2007b, 2010; *Mehrotra et al*., 2012]. For elaborate reviews of stochastic rainfall simulation studies done in the past and those done more recently, one can refer to the work by *Srikanthan and McMahon* [2001] and by *Sharma and Mehrotra* [2010], respectively. To better understand the merits and demerits of the suggested generator, we shall compare it with three alternate models: the conventional Markov chain generator [*Richardson*, 1981], the TPM model [*Srikanthan and McMahon*, 1985], and the modified nonparametric model of *Harrold et al*. [2003a, 2003b] with parametric Markov chain for rainfall occurrences and nonparametric kernel density estimation (KDE) for rainfall amounts.

[15] The rest of this paper is organized as follows. Section 2 introduces the improved bivariate mixed distribution. Section 3 describes algorithms for the simulation of random numbers and for the estimation of distribution parameters. Based on the improved distribution, section 4 presents a single-site daily rainfall generator and tests it on a sample station in Texas. Section 5 continues with extensive simulation experiments to compare it with other advanced alternatives. Finally, conclusions are presented in section 6.

2. Improved Bivariate Mixed Distribution

2.1. Constructing $h(x, y)$ With Copula

[16] Our first improvement to the bivariate mixed distribution is reflected in introducing more and various copula families as admissible candidates to construct the joint density $h(x, y)$. The objective is to reduce the risk of obtaining a misrepresented relationship between X and Y when $X > 0$ and $Y > 0$. With the use of copula theory [*Joe*, 1997; *Nelson*, 2006], the bivariate CDF can be written as follows:

$$H(x,y) = C(F(x), G(y)), \tag{4}$$

where $C(\cdot)$ is the copula function; the marginal CDFs $F(x)$ and $G(y)$ are given as:

$$F(x) = P(X \le x | X > 0, Y > 0),$$

$$G(y) = P(Y \le y | X > 0, Y > 0).$$

[17] From equation (4), the corresponding PDF can be decomposed into:

$$h(x,y) = f(x) \cdot g(y) \cdot c(F(x), G(y)), \tag{5}$$

where $c(\cdot)$ is the copula density, and $f(x)$ and $g(y)$ are the PDFs of $F(x)$ and $G(y)$, respectively. Ten different copula families are considered as admissible candidates for this work (Table 1). They can model a wide variety of

dependence structures, including the lower and upper tail dependencies, and cover most bivariate analyses found in the hydrological literature.

2.2. Modeling Rainfall Amounts With Hybrid Distribution

[18] Our second improvement is to introduce a hybrid exponential and generalized (HEG) Pareto distribution for rainfall amounts. The objective is to improve the characterization of extremes. The HEG distribution has the PDF of

$$f_{\mathrm{HEG}}(x) = \frac{1}{Z}\left(f_{\mathrm{E}}(x;\mu)I_{(0,\theta]}(x) + f_{\mathrm{GP}}(x;\kappa,\sigma,\theta)I_{(\theta,\infty)}(x)\right), \quad \mu,\kappa,\sigma,\theta > 0, \tag{6}$$

the CDF of

$$F_{\mathrm{HEG}}(x) = \frac{1}{Z}\left(F_{\mathrm{E}}(x;\mu)I_{(0,\theta]}(x) + (F_{\mathrm{E}}(\theta;\mu) + F_{\mathrm{GP}}(x;\kappa,\sigma,\theta))I_{(\theta,\infty)}(x)\right), \tag{7}$$

and the p-quantile function of

$$x_p = (-\mu\ln(1-pZ))I_{[0,F_E(\theta)]}(p) + \left(\theta + \sigma/\kappa\left(\left(pZ - 2 + e^{-x/\mu}\right)^{-\kappa} - 1\right)\right)I_{(F_E(\theta),1]}(p), \tag{8}$$

where

$$Z = F_{\mathrm{E}}(\theta;\mu) + 1,$$

$$f_{\mathrm{E}}(x;\mu) = 1/\mu e^{-x/\mu}I_{(0,\theta)}(x),$$

$$F_{\mathrm{E}}(x;\mu) = \left(1 - e^{-x/\mu}\right)I_{(0,\theta)}(x),$$

$$f_{\mathrm{GP}}(x;\kappa,\sigma,\theta) = 1/\sigma(1+\kappa(x-\theta)/\sigma)^{-1/\kappa-1}I_{(\theta,\infty)}(x),$$

$$F_{\mathrm{GP}}(x;\kappa,\sigma,\theta) = \left(1 - (1+\kappa(x-\theta)/\sigma)^{-1/\kappa}\right)I_{(\theta,\infty)}(x),$$

and $I_A(\cdot)$ is the indicator function. To ensure the continuity of the PDF at the junction point, θ can be expressed as a function of μ and σ as follows:

$$\theta = -\mu\ln(\mu/\sigma). \tag{9}$$

[19] Therefore, the HEG distribution can be fully parameterized by $\mathbf{P}=[\mu, \kappa, \sigma]$. In the improved distribution, $h_X(x)$, $h_Y(y)$, $f(x)$, and $g(y)$ all belong to the HEG family. Note that for simplicity, subscript "HEG" will be dropped from the following relevant equations.

2.3. Conditional Distributions

[20] Two types of conditional distributions, derived from the improved bivariate mixed distribution, constitute the cornerstones for its applications:

I) $P(Y \le y|X = x, X > 0, Y > 0)$ (or $P(X \le x|Y = y, X > 0, Y > 0)$)

II) $P(Y \le y|X = x, X \ge 0)$ (or $P(X \le x|Y = y, Y \ge 0)$)

2.3.1. Type I Conditional Distribution

[21] Consider the following situation. Given contemporaneous occurrence of rainfall at both sites, one wants to know the conditional distribution of rainfall amount at one site given that at the other site, i.e., the conditional distribution of Y given $X = x$, $X > 0$, and $Y > 0$ (or of X given $Y = y$, $X > 0$, and $Y > 0$). This conditional CDF is given as follows:

$$H_{Y|X=x}(y|X = x, X > 0, Y > 0) = c_1(F(x), G(y)), \tag{10}$$

where $c_1(\cdot)$ is the partial derivative of copula $C(\cdot)$ with respect to its first argument [*Zhang* and *Singh*, 2007]. The p-quantile function of the distribution is given as follows:

$$y_{p|X=x,X>0,Y>0} = G^{-1}\left(c_{1_2}^{-1}(F(x), p)\right), \tag{11}$$

where $G^{-1}(\cdot)$ is the inverse of $G(\cdot)$, which can be directly computed by equation (8); $c_{1_2}^{-1}(\cdot)$ is the quasi-inverse of $c_1(\cdot)$ with respect to its second argument.

2.3.2. Type II Conditional Distribution

[22] Now consider another situation where one is interested in the conditional distribution of Y given $X = x$ and $X \ge 0$ (or of X given $Y = y$ and $Y \ge 0$). In this case, no prior knowledge about the wet or dry state of Y (or X) is available. This conditional CDF can be split into two parts [*Herr*, 1999; *Herr and Krzysztofowicz*, 2005]. The first part is the distribution of Y given $X = x$ and $X = 0$:

$$\Phi_{Y|X=0}(y|X = 0) = \frac{p_{00} + p_{01}H_Y(y)}{p_{00} + p_{01}}. \tag{12}$$

[23] The second part is of Y given $X = x$ and $X > 0$:

$$\Phi_{Y|X=x}(y|X = x, X > 0) = \frac{p_{10}h_X(x) + p_{11}f(x)c_1(F(x), G(y))}{p_{10}h_X(x) + p_{11}f(x)}. \tag{13}$$

[24] Correspondingly, the p-quantile functions are as follows:

$$y_{p|X=0} = H_Y^{-1}\left(\frac{(p_{00} + p_{01})p - p_{00}}{p_{01}}\right), \tag{14}$$

$$y_{p|X=x,X>0} = G^{-1}\left(c_{1_2}^{-1}\left(F(x), \frac{(p_{10}h_X(x) + p_{11}f(x))p - p_{10}h_X(x)}{p_{11}f(x)}\right)\right), \tag{15}$$

where $H_Y^{-1}(\cdot)$ and $G^{-1}(\cdot)$ are the inverses of $H_Y(\cdot)$ and $G(\cdot)$, respectively. Derivation for the CDFs of the type II conditional distribution is presented in Appendix A.

3. Simulation and Estimation

3.1. Random Number Simulation

[25] Random vectors can be simulated from the bivariate mixed distribution by an algorithm in what follows:

[26] *Algorithm 1*:

1. Draw p uniformly distributed over [0, 1].
2. If $p < p_{00}$, set $x = 0$ and $y = 0$.

3. If $p_{00} \le p < p_{00}+p_{10}$, then draw a random value from $H_X(x)$ for x and set $y = 0$.

4. If $p_{00}+p_{10} \le p < p_{00}+p_{10}+p_{01}$, then set $x = 0$ and draw a random value from $H_Y(y)$ for y.

5. If $p \ge p_{00}+p_{10}+p_{01}$, then draw a bivariate random vector from the joint distribution $C(F(x), G(y))$.

[27] In applications, oftentimes it is needed to generate random samples Y (or X) from the conditional distribution $Y|X$ (or $X|Y$). Suppose $X = x$, the generation algorithm can be summarized by the following steps:

[28] *Algorithm 2*:

1. Draw a random number p which is uniformly distributed over [0, 1].

2. If $x = 0$, determine if $(p_{00}+p_{01})p-p_{00}$ is positive or negative. If it is positive, then set $y = y_{p|X=0}$, and set $y = 0$ otherwise.

3. If $x > 0$, determine if $(p_{10}h_X(x) + p_{11}f(x))p - p_{10}h_X(x)$ is positive or negative. If it is positive, then set $y = y_{p|X=x,X>0}$, and set $y = 0$ otherwise.

4. Repeat the above steps with new x, if necessary.

3.2. Parameter Estimation

[29] The discrete probabilities, the marginal HEG distributions, and the copula function are unknowns to be estimated. Although theoretically one can perform direct maximum likelihood (ML) estimation, for simplicity and flexibility we break down the estimation into three parts: (1) estimate the discrete probabilities; (2) estimate the marginal HEG distributions; and (3) select and estimate the copula function. A stepwise estimation procedure is presented as follows. The forthcoming Monte Carlo simulation in section 3.3 will provide empirical justification for this separate estimation strategy.

Step 1. Estimate the Discrete Probabilities

[30] The discrete probabilities can be estimated by the ML method as follows:

$$\hat{p}_{10} = \sum_{i=1}^{n} \frac{I_{(0,\infty)}(x_i)I_0(y_i)}{n}, \tag{16}$$

$$\hat{p}_{01} = \sum_{i=1}^{n} \frac{I_0(x_i)I_{(0,\infty)}(y_i)}{n}, \tag{17}$$

$$\hat{p}_{11} = \sum_{i=1}^{n} \frac{I_{(0,\infty)}(x_i)I_{(0,\infty)}(y_i)}{n}, \tag{18}$$

$$\hat{p}_{00} = 1 - \hat{p}_{10} - \hat{p}_{01} - \hat{p}_{11}, \tag{19}$$

where n is the number of data pairs in the sample set.

Step 2. Estimate the Marginal HEG Distributions

[31] The ML method can be used to estimate parameters of HEG distribution. In cases when it has problems, a decent alternative is the maximum goodness-of-fit (MGF) method with the right-tail Anderson-Darling statistic [*Li et al.*, 2012]. In practice, one may apply the ML or the MGF method for estimation of parameters in the HEG distribution, whichever provides better results.

Step 3. Select and Estimate the Copula Function

[32] A two-stage algorithm is used to determine the copula function. The first stage identifies the most suitable copula family from the 10 candidates (Table 1), and the second stage estimates parameters of the identified family.

[33] For the first stage, there are several criteria such as Akaike information criterion (AIC) [*Akaike*, 1974], Bayesian information criterion (BIC) [*Schwarz*, 1978], Bayesian copula selection (BCS) [*Huard et al.*, 2006], and Genest-Rémillard goodness-of-fit test [*Genest et al.*, 2009]. The Genest-Rémillard goodness-of-fit test tends to fail to discriminate among different families when the association between variables is weak, which is usually the case for rainfall amounts of two consecutive wet days [*Serinaldi*, 2009b]. For this reason, we use democratic voting among families elicited from the first three criteria. If three different families are elicited, we arbitrarily follow the AIC criterion.

[34] The AIC and BIC criteria are calculated as AIC=$-2\mathrm{LL}_C+2p$ and BIC=$-2\mathrm{LL}_C+p\log n_4$, respectively, where LL_C is the log-likelihood of the sample vectors $[u_i, v_i]$ and $i = 1, 2, \ldots, n_4$; u_i and v_i are computed by $u_i = F(x_i)$ and $v_i = G(y_i)$; p is the number of parameters of the copula family; and $n_4 = \sum_{i=1}^{n} I_{(0,\infty)}(x_i)I_{(0,\infty)}(y_i)$. The smaller the AIC or BIC is, the better is the family.

[35] The BCS weight for a candidate copula family is computed as follows:

$$W = \int_{\Omega^\tau} \prod_{i=1}^{n_4} c\big(u_i, v_i | g^{-1}(\tau)\big) \mathrm{d}\tau, \tag{20}$$

where $g^{-1}(\tau)$ is the inverse of $g(\theta)$ and Ω^τ is the domain of τ, both of which have been included in Table1; the other components hold the same meaning as defined before. In the case of Student copula, the degree of freedom is estimated first and then the weight is computed. The best family is identified as the one with maximum BCS weight.

[36] Once the most suitable copula family is identified, the ML method is then used to estimate the parameters. As inferred from equation (5), the log-likelihood function can be decomposed as follows:

$$\mathrm{LL}_{XY} = \mathrm{LL}_X + \mathrm{LL}_Y + \mathrm{LL}_C, \tag{21}$$

where

$$\mathrm{LL}_X = \sum_{i=1}^{n_4} \log(f(x_i)),$$

$$\mathrm{LL}_Y = \sum_{i=1}^{n_4} \log(g(y_i)), \quad \text{and}$$

$$\mathrm{LL}_C = \sum_{i=1}^{n_4} \log(c(F(x_i), G(y_i))).$$

[37] Equation (21) suggests that the marginal distributions and the copula function can be estimated separately. After fitting the marginal distributions, copula parameters are determined by numerically maximizing LL_c.

3.3. Preliminary Monte Carlo Simulation

3.3.1. Simulation Design

[38] To roughly test the asymptotic properties of the estimators for the bivariate mixed distribution, Monte Carlo simulation was carried out. Random sample sets with

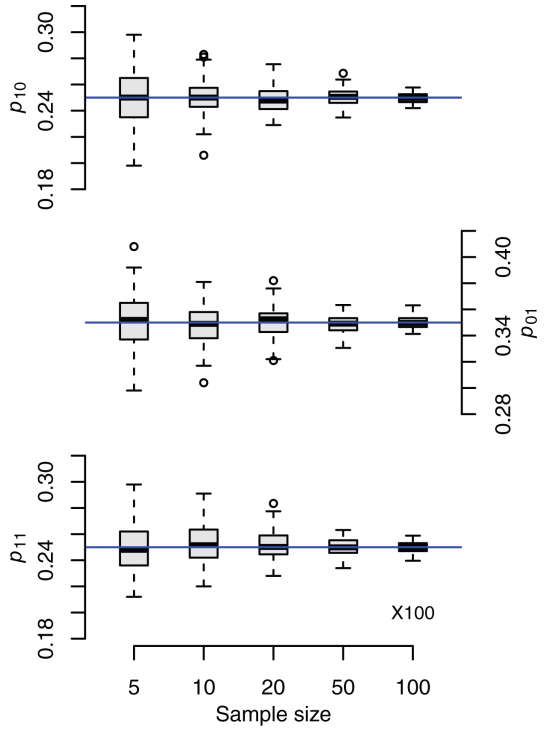

Figure 3. Behavior of discrete probability estimates as sample size increases. True values are marked by horizontal lines.

varying sizes were generated from the distribution using Algorithm 1. Model parameters were estimated following the stepwise procedure in section 3.2. The sample size was increased from 500 to 10,000 with varying factors. For each sample size, random sampling and parameter estimation were repeated with 100 trials.

[39] The parent distribution was parameterized as follows. Arbitrarily, the discrete probabilities were set as p_{00}=0.15, p_{10}=0.25, p_{01}=0.35, and p_{11}=0.25. The marginal distributions $h_X(x)$, $h_Y(y)$, $f(x)$, and $g(y)$ were assumed to be identical with parameters **P**=[5.22, 0.18, 16.30]. Two copula families were used to mimic different dependence structures. One was Gumbel copula with $\tau = 0.75$ ($\theta = 4.0$), which can simulate the upper tail dependence. The other was Clayton copula also with $\tau = 0.75$ ($\theta = 6.0$), which is capable of modeling the lower tail dependence.

3.3.2. Simulation Results

[40] Estimates for the discrete probabilities and parameters of $h_X(x)$ are shown by box plots in Figures 3 and 4, respectively, as the sample size increases. The true value of each parameter is marked by a horizontal line. The results for the other marginal distributions are not presented as they demonstrated similar patterns. Two major points can be inferred from the figures: (1) the discrepancies between the means of the estimates and true values were very small, even negligible when the sample size was sufficiently large; and (2) the spread of the estimates notably decreased as the sample size increased. Thus, in general, for both discrete probabilities and marginal distributions, the estimators described in section 3.2 behave asymptotically consistently and efficiently.

[41] With the aid of democratic voting, the number of successful identifications for the true copula family was summarized in Table 2. It may be seen that (1) the democratic voting converged to the right family as the sample size increased; (2) AIC and BIC outperformed BCS in identifying the true model; and (3) the democratic voting was safer than any single criterion.

[42] After identifying the most suitable copula family, the corresponding parameter was estimated by the ML method. Box plots for the estimates of Kendal's τ for the Gumbel and Clayton copulae are shown in Figure 5. As a note, this figure contains the estimates in trials when the true family was successively identified only. It appears that the ML estimator for the copula parameter behaves as expected.

[43] The above Monte Carlo simulation indicates three major points. First, the separate estimation strategy described in section 3.2 seems to work adequately respecting asymptotic consistency and efficiency. Second, the simulation Algorithm 1 appears to perform well in the sense that parameters estimated from the samples generated by Algorithm 1 statistically reproduce the true values. Third, the bivariate mixed distribution is expected to model different dependence structures with the use of different copulae. As a final point, it is realized that a more comprehensive study to properties of the improved bivariate mixed distribution and its estimation are required in the future considering the limited scope of this small simulation experiment.

4. Application to Daily Rainfall Simulation

[44] Among different recognized applications of the improved bivariate mixed distribution, how to use it for

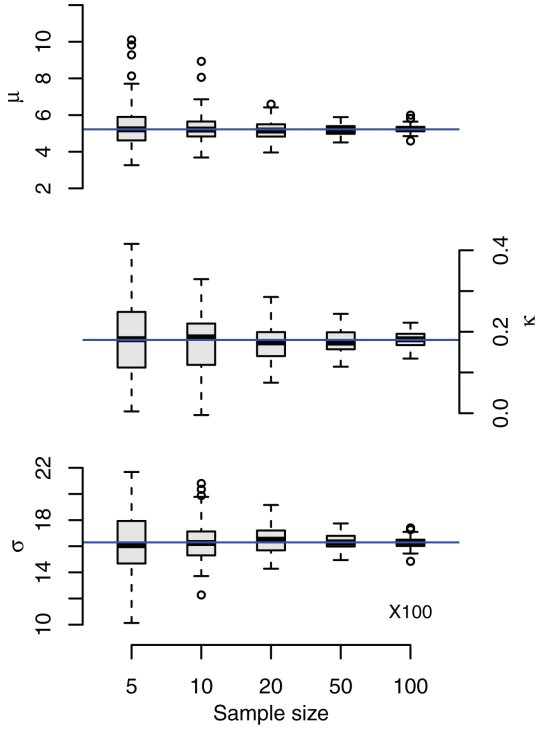

Figure 4. Behavior of HEG distribution parameters estimates computed from the subset $X > 0$ and $Y = 0$ as sample size increases. True values are marked by horizontal lines.

Table 2. Number of Successful Identifications of the True Copula Family Over 100 Trials

	Gumbel Family ($\tau = 0.75$)				Clayton Family ($\tau = 0.75$)			
Sample Size ($\times$100)	AIC	BIC	BCS	Democratic Voting	AIC	BIC	BCS	Democratic Voting
5	73	73	65	73	100	100	100	100
10	84	84	79	84	100	100	100	100
20	96	96	94	98	100	100	100	100
50	98	98	98	98	100	100	100	100
100	100	100	100	100	100	100	100	100

daily rainfall simulation is of particular interest in this research.

4.1. Bivariate Mixed Distribution-Based Markov Chain Generator

[45] Daily rainfall can be generalized as a Markov process with autocorrelation described by the bivariate mixed distribution. Let X and Y in equation (2) denote rainfall of days $t-1$ and t, respectively, then the conditional distribution of rainfall of day t given that of day $t-1$ can be modeled by equations (12) and (13). Rainfall simulation may proceed through sequentially sampling from the conditional distribution following the steps in Algorithm 2. For simplicity, we shall hereinafter refer to it as a bivariate mixed distribution-based Markov chain generator (BMC). In particular, BMCs with HEG and gamma distributions for rainfall amounts are denoted by BMC-H and BMC-G, respectively.

[46] However, caution has to be exercised while using BMC-H for rainfall simulation, as it is recognized that on occasion extremely large values may be generated. A similar problem is faced by other rainfall generators that simulate rainfall amounts in terms of distributions with generalized Pareto tails, such as dynamic mixture of gamma and generalized Pareto distribution [*Vrac and Naveau*, 2007; *Hundecha et al*., 2009; *Hundecha and Merz*, 2012] and hybrid gamma and generalized Pareto distribution [*Furrer and Katz*, 2008]. In these compound distributions, tail index (κ) of the Pareto component is usually forced to be positive in order to meet the heavy-tailed nature of daily rainfall amounts. The PDF of a generalized Pareto distribution with positive tail index is slowly varying at infinity [*Feller*, 1968], which means that high quantiles (e.g., 0.9999+) would be very large. Even though these high quantiles are less likely to be generated, the possibility does exist. Very small amounts of extremely high values may significantly change the nature of simulated sequences, especially with respect to rainfall frequency analysis that usually involves block maxima or peaks over threshold only. Considering the fact that rainfall at a given site is bounded below and above by 0 and a finite value, corrections to BMC-H are needed such that infeasible large values can be screened. At the same time, it is necessary to emphasize that as a stochastic rainfall generator, it should be able to reasonably extrapolate unseen rare rainfall events significantly beyond the upper range of available observed data.

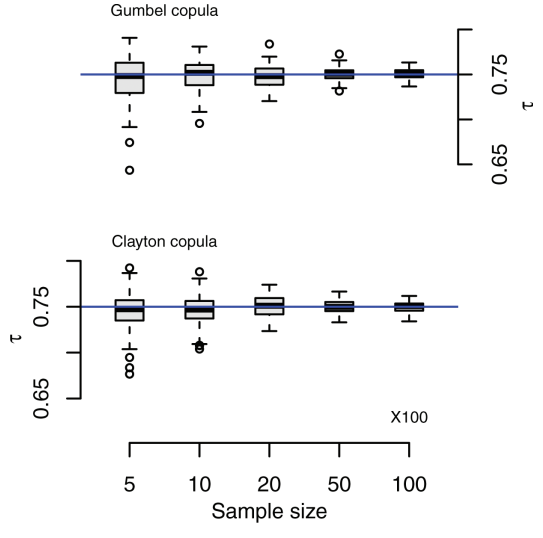

Figure 5. Behavior of the Kendall's τ estimates for Gumbel family and the Clayton family computed from the subset $X > 0$ and $Y > 0$ as sample size increases. True values are marked by horizontal lines.

[47] Keeping the above two accounts in mind, a modification proposal for algorithm 2 is presented in the following. Suppose in the current month, the observed maximum daily rainfall amount is $a_{\max}$. We assume that simulated rainfall amounts in this month should be no greater than 200% of $a_{\max}$. From the fitted $H_Y(y)$, a upper-bound percentile (p_{up1}) can be obtained by evaluating $H_Y(y)$ at $2a_{\max}$. Similarly, another upper-bound percentile (p_{up2}) can be obtained from $G(y)$. In step 2, if $(p_{00}+p_{01})p-p_{00}$ is positive, which means that the day to be simulated is wet, then we repeatedly generate a uniform random variate p_1 over [0, 1] and evaluate

$$\frac{(p_{00}+p_{01})p_1-p_{00}}{p_{01}},$$

until this quantity is exactly between 0 and p_{up1}; the corresponding value is denoted as p_r. The rainfall amount is then simulated as $H_Y^{-1}(p_r)$. Similar modification is made in step 3 as follows. If $(p_{10}h_X(x)+p_{11}f(x))p-p_{10}h_X(x)$ is positive, then quantity

$$c_{1_2}^{-1}\left(F(x),\frac{(p_{10}h_X(x)+p_{11}f(x))p_1-p_{10}h_X(x)}{p_{11}f(x)}\right),$$

is repeatedly evaluated on randomly generated uniform variates p_1s until it falls between 0 and p_{up2}. Denote the

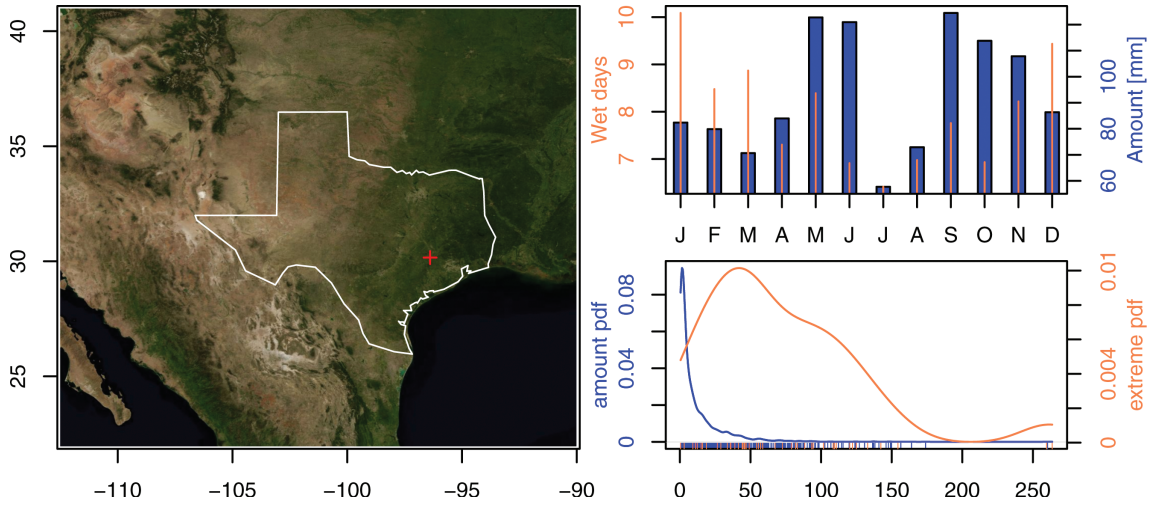

Figure 6. Location map of the selected rainfall station in Texas (map); the distributions of the (top right) number of wet days and rainfall totals over months of year and the (bottom right) kernel density estimations of nonzero and annual extremes of daily rainfall at the selected station.

resulting value as p_r. Then the simulated amount is $G^{-1}(p_r)$.

[48] The above procedure screens unreasonable large values, retains certain extrapolation ability, and ensures autocorrelation of rainfall amounts to be maintained. We tested this procedure at several stations and found that the upper-bound percentile is almost always greater than 0.999, indicating that the tail behavior learned from observed data is slightly intervened only.

4.2. Data

[49] Rainfall records spanning over a period from 1960 to 2005 at station TX411048 in Texas were used to test the BMC-H generator. This station was selected mainly for the account that no missing values exist in the time window. The location of the station (map), distributions of the number of wet days and the rainfall totals over months of year (top right), and an overall picture about the empirical PDFs of nonzero and annual extremes of daily rainfall (bottom right) are shown in Figure 6. To avoid identifying dew and other noise as rainfall, a value of 0.3 mm was adopted as the significant rainfall threshold, which means that only days with amounts greater than 0.3 mm were considered as wet.

[50] Historical observations were first stratified into months of year. Then, BMC-H was fitted for each calendar month. An implicit assumption involved in the stratification is that rainfall process is stationary within a given month but nonstationary across different months. On one hand, this assumption insures a large enough sample size such that the model can be estimated with reasonable accuracy, and on the other hand, it properly takes rainfall seasonality into account. It is noted that the same data stratification was used for fitting other alternative models in the forthcoming comparison analyses.

4.3. Preliminary Evaluation of the BMC-H Generator

[51] Copula characterizes the joint behavior of rainfall amounts of two consecutive wet days. The fundamental objective of copula selection is to adequately represent the dependence structure of the data under consideration. To demonstrate the necessity of considering more copula families as candidates, we analyzed the copula selection for each month based on different criteria. For the ease of comparison, the 10 candidates were ranked by ascending AIC and BIC values or by descending BCS weights. After that, families with smaller AIC or BIC values or greater BCS weights would gain front ranks. The results are presented in Figure 7. As can be observed, both AIC and BIC resulted in the same ranks, which were much different from those elicited from the BCS criterion. Other interesting findings include the following: (1) of the 10 candidates, the most often selected families were Clayton and survival Clayton; (2) no matter which criterion was followed, A12 and A14 were the two least suitable families as both of them admits Kendall's τ no less than 0.333, whereas the sample estimates from the rainfall records under consideration were up to 0.118 only; (3) and besides the commonly used families, other families like FGM and Joe might be required for a more realistic simulation. The above analyses suggest that it is always preferable to select a suitable copula family from various candidates rather than to adopt a unique one for each month and for stations from different climate areas; otherwise, a suboptimal model might be obtained, which would in turn misrepresent the autocorrelation behavior of rainfall amounts.

[52] The upper-bound percentiles for screening irrational over large values of rainfall amounts are listed in Table 3. It is seen that all these percentile values are very close to 1 (greater than 0.999), which means that the tail behavior of rainfall amounts of each month learned from available observed data was slightly altered only.

[53] To evaluate the performance of BMC-H (as well as other alternate models), 200 sequences, each with the same length as historical records (46 years), were generated. We evaluated the model by descriptive statistics from the following five aspects: (1) to reproduce basic occurrence statistics; (2) to amount statistics; (3) to reproduce the historical distribution of rainfall amounts; (4) to reproduce

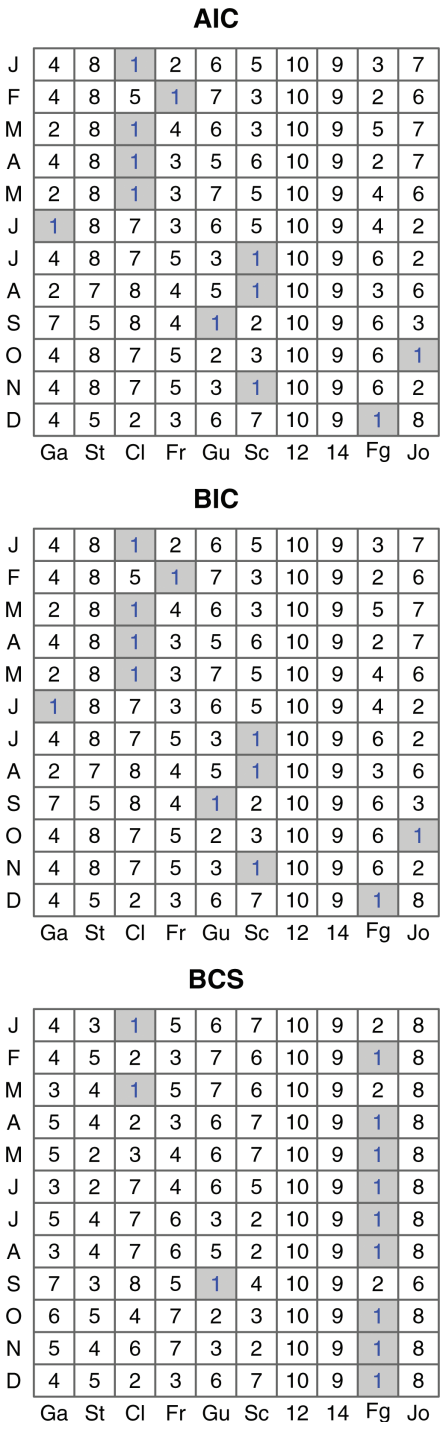

Figure 7. Copula ranks elicited by different criteria. Ga: Gaussian copula; St: Student copula; Cl: Clayton copula; Fr: Frank copula; Gu: Gumbel copula; Sc: survival Clayton copula; 12: A12 copula; 14: A14 copula; Fg: FGM copula; and Jo: Joe copula.

characteristics of extreme rainfall events; and (5) to reproduce autocorrelation of rainfall amounts.

[54] Basic occurrence statistics analyzed in this research are the number of wet days (NM), the number of dry days (ND), the number of wet spells (NWS), the number of dry spells (NDS), maximum wet spell length (MWSL), and maximum dry spell length (MDSL). Monthly patterns of these statistics are visually summarized by box plots in Figure 8. Apparently, all these statistics were reasonably well reproduced, indicating that BMC-H can satisfactorily simulate the persistence of rainfall occurrences. The root-mean-square errors (RMSEs) provided a quantitative confirmation of this observation (Table 4).

[55] Basic amount statistics are the yearly mean and standard deviation of monthly rainfall totals, and the results are presented in Figure 9. Inspection of Figure 9 reveals that BMC-H seems to be doing a quite good job of reproducing these statistics. Except for September for which a slight trend toward underestimating the mean was detected, no significant overestimation or underestimation was found in other months.

[56] To check how well BMC-H reproduces the historical distribution of rainfall amounts, we looked at the empirical QQ plots of simulated values against observations on natural and logarithmic (base 2) scales, respectively, as shown in Figure 10. It appears that the distribution of simulated amounts was in fair agreement with that of the observations.

[57] Realistic simulation of the entire distribution of rainfall amounts would help BMC-H to reproduce characteristics of extreme rainfall events. To verify this expectation, extreme weather indices were used for further evaluation. Maximum 1-day rainfall amount measures 1-day block extreme rainfall events. Monthly pattern of this statistics is shown in Figure 11 (top). To circumvent misleading, it is better to explain how this graph was plotted. First, for a given month and year, we picked out the maximum daily rainfall value from observations. Thereby we had 46 values, one for each year. Then, we averaged these values such that we obtained one smoothed value, as marked by the rectangular. Similarly, we obtained another 200 values, one for each simulated sequence, as displayed by the box plot. Figure 11 (top) was accomplished by repeating the above steps for each month. For simplicity, this smoothed quantity is referred to henceforth as SM1A. As expected, BMC-H reproduced SM1A with reasonably good accuracy. One disturbing instance was found in September, for which SM1A was underestimated.

[58] Moreover, rainfall fractions due to wet days with amounts greater than large rainfall quantiles (e.g., 0.90, 0.95, and 0.99) over a period measure extreme rainfall events from a view point close to frequency analysis [*Lennartsson et al.*, 2008]. Take the computation of rainfall fraction corresponding to the 0.95 quantile as an example. Assume over a given period that there are 100 wet days

Table 3. Upper-Bound Percentiles for Screening Irrational Over Large Values of Rainfall Amounts to be Simulated

Month	1	2	3	4	5	6	7	8	9	10	11	12
$F_Y(y)$	0.9999	0.9992	0.9991	0.9997	1.0000	0.9995	0.9995	0.9993	0.9997	0.9996	0.9999	0.9994
$G(y)$	0.9998	1.0000	1.0000	0.9999	0.9999	0.9994	1.0000	0.9999	0.9992	0.9994	0.9995	0.9998

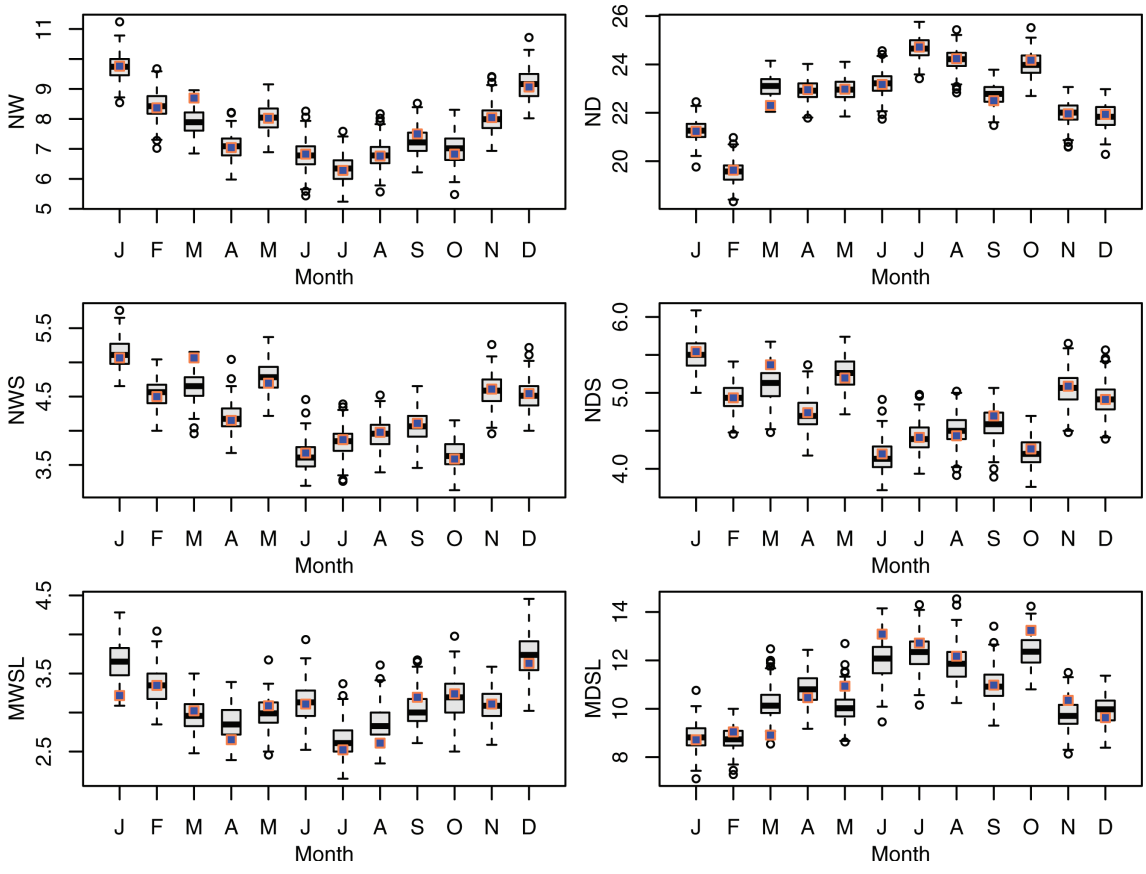

Figure 8. Box plots of basic raifnall occurrence statistics of observed and BMC-H-simulated sequences. Orange rectangulars with blue filled denote observed values.

with amount of a_1, $a_2, \ldots,$ a_{100}, respectively; the 0.95 quantile of these 100 values is denoted as a; suppose there are 4 days whose amounts are greater than a, e.g., a_7, a_{12}, a_{25}, and a_{70}, then the fraction corresponding to the 0.95 quantile can be computed as $(a_7+a_{12}+a_{25}+a_{70})/(a_1+a_2+\ldots+a_{100})$. To obtain reliable estimates for the quantiles, all the 46 years of data were pooled together. The results are presented in Figure 11 (bottom). It is apparent that the three statistics were reasonably well preserved.

[59] For the evaluation of BMC-H in capturing autocorrelation of daily rainfall amounts, simulated and observed lag-1 autocorrelations measured by Kendall's τ correlation coefficient are presented in Figure 12 (top). As can be observed, no significant evidence for misrepresentation was detected except for a slight overestimation in January and March. It is worthwhile to note that we used Kendall's τ rather than the Pearson's correlation coefficient as the dependence measure, as the former is more suitable for non-Gaussian distributed rainfall data.

[60] Autocorrelation of rainfall amounts has a direct influence on rainfall event volumes. We therefore continued to investigate whether or not the 2-, 3-, and 4-day rainfall event volumes were reasonably preserved. As consecutive wet days may last over months and over years, the volume statistics were computed again by pooling the 46 years of data. The results are shown in Figure 12 (middle), from which it is inferred that BMC-H did a good job of simulating short-term rainfall event volumes.

[61] In addition to Kendall's τ correlation coefficient which mainly measures the central dependence, it is interesting to look at the upper tail dependence as well. Intuitively, the upper tail dependence can be understood as how likely extreme rainfall events to occur together [*Cherubini et al.*, 2004]. A nonparametric estimator recommended by *Frahm et al.* [2005] was used to compute the upper tail dependence coefficient as follows:

$$\hat{\lambda} = 2 - 2\exp\left(1/n\sum_{i=1}^{n}\log\left(\frac{\sqrt{\log(1/u_i)\log(1/v_i)}}{\log\left(1/\max(u_i,v_i)^2\right)}\right)\right), \tag{22}$$

where n is the number of paired consecutive wet days; $u_i = \hat{F}_{t-1}\left(x_{t-1}^i\right)$ and $v_i = \hat{G}_t\left(x_t^i\right)$; $\hat{F}_{t-1}(\cdot)$ and $\hat{G}_t(\cdot)$ are the empirical CDFs of rainfall amounts of days $t-1$ and t, respectively. The results are presented in Figure 12 (bottom). It can be seen that BMC-H exhibited a decent performance in simulating the dependence of rainfall extremes.

Table 4. Root-Mean-Square Error of Basic Rainfall Occurrence Statistics

	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
BMC-H												
NW	0.61	0.46	1.13	0.61	0.58	0.53	0.43	0.48	0.71	0.51	0.49	0.55
ND	0.61	0.46	1.13	0.61	0.58	0.53	0.43	0.48	0.71	0.51	0.49	0.55
NWS	0.22	0.20	0.53	0.21	0.22	0.23	0.25	0.28	0.19	0.21	0.24	0.21
NDS	0.27	0.18	0.37	0.21	0.19	0.20	0.22	0.21	0.19	0.19	0.21	0.19
MWSL	0.50	0.25	0.21	0.20	0.25	0.25	0.24	0.34	0.43	0.34	0.21	0.26
MDSL	0.69	0.54	1.75	0.75	0.89	1.10	0.81	0.73	0.62	0.97	0.77	0.79
BMC-G												
NW	0.67	0.49	0.42	0.50	0.60	0.57	0.41	0.48	0.59	0.50	0.49	0.55
ND	0.67	0.49	0.42	0.50	0.60	0.49	0.41	0.48	0.59	0.50	0.49	0.55
NWS	0.23	0.20	0.21	0.23	0.22	0.22	0.24	0.30	0.21	0.23	0.24	0.21
NDS	0.25	0.20	0.26	0.22	0.21	0.21	0.21	0.21	0.24	0.22	0.24	0.21
MWSL	0.41	0.25	0.19	0.24	0.26	0.28	0.22	0.33	0.35	0.28	0.22	0.24
MDSL	0.67	0.58	0.77	0.77	0.85	1.16	0.74	0.73	0.69	1.05	0.76	0.79
CMC-H[a]												
NW	0.69	0.46	1.13	0.46	0.57	0.47	0.42	0.48	0.61	0.47	0.49	0.57
ND	0.69	0.46	1.13	0.46	0.57	0.47	0.42	0.48	0.61	0.47	0.49	0.57
NWS	0.23	0.21	0.31	0.23	0.24	0.23	0.23	0.30	0.22	0.22	0.24	0.21
NDS	0.26	0.20	0.21	0.22	0.22	0.22	0.21	0.22	0.23	0.23	0.23	0.21
MWSL	0.41	0.23	0.30	0.21	0.24	0.23	0.22	0.34	0.35	0.28	0.22	0.26
MDSL	0.68	0.58	1.36	0.78	0.94	1.22	0.85	0.70	0.64	0.96	0.76	0.78
TPM												
NW	0.46	0.43	0.47	0.41	0.46	0.43	0.36	0.46	0.48	0.49	0.50	0.50
ND	0.47	0.43	0.47	0.41	0.46	0.43	0.36	0.46	0.48	0.49	0.50	0.50
NWS	0.22	0.21	0.24	0.24	0.22	0.21	0.20	0.24	0.22	0.22	0.24	0.21
NDS	0.24	0.21	0.30	0.21	0.21	0.19	0.20	0.20	0.19	0.21	0.23	0.22
MWSL	0.52	0.21	0.23	0.27	0.21	0.22	0.18	0.34	0.25	0.28	0.22	0.27
MDSL	0.56	0.60	0.61	0.63	1.05	1.39	0.78	0.73	0.79	1.11	0.86	0.64

[a]SMC-K model had statistically the same results as CMC-H because they both use conventional two-state Markov chain model for the simulation of rainfall occurrence.

4.4. Benefits From Using HEG Distribution for Rainfall Amounts

[62] A major innovation of the improved bivariate mixed distribution is one of using a rather sophisticated marginal distribution to characterize the entire distribution of rainfall amounts. One might raise a question as to whether the synthetic rainfall sequences are really sensitive to the choice of marginal distribution. If not, then there is no need to spend extra time and effort on a complex model. In this respect, a comparison between BMC-H and BMC-G was carried out.

[63] The unique difference between BMC-H and BMC-G lies in the distribution used for rainfall amounts. Thus, there should be no overwhelming difference between them in simulating rainfall occurrences, as verified by RMSEs in Table 4. Either no significant difference was observed in reproducing basic amount statistics. We therefore pay special attention to look if BMC-H outperforms BMC-G in capturing the entire distribution of rainfall amounts and in reproducing characteristics of extreme rainfall events. Figure 13 shows QQ plots of BMC-G simulated versus observed rainfall amounts. It can be seen that BMC-G resulted in a slight overestimation of the central part and a serious underestimation of the tail part of the distribution. As a consequence, extreme rainfall characteristics, for instance, SM1A and large rainfall fractions, were consistently underestimated (Figure 14). When compared with BMC-G, BMC-H provides a notable gain in capturing the entire distribution of rainfall amounts and preserving characteristics associated with extreme rainfall events.

4.5. Benefits From Accounting for Autocorrelation of Rainfall Amounts

[64] As was previously mentioned, an implementational benefit from applying the bivariate mixed distribution for

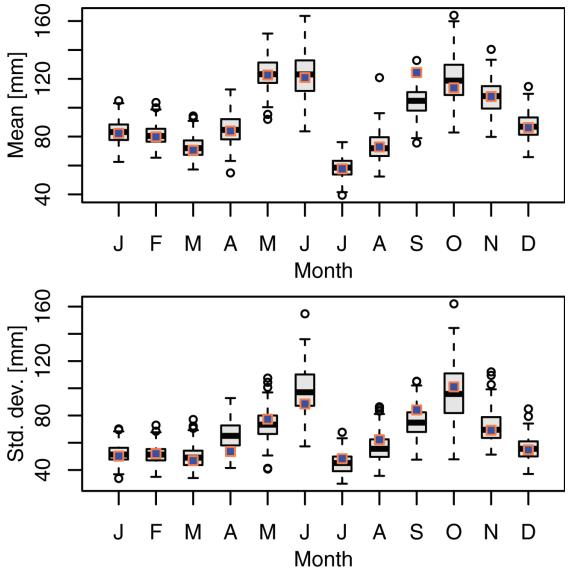

Figure 9. Box plots of basic rainfall amount statistics of observed and BMC-H-simulated sequences. Orange rectangulars with blue filled denote observed values.

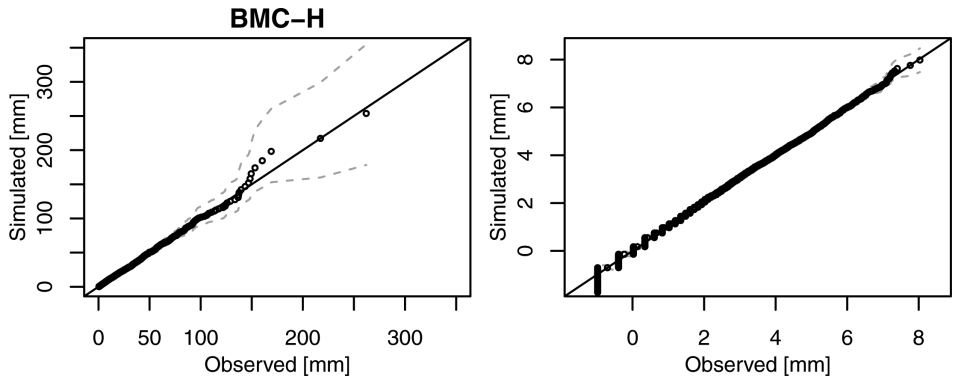

Figure 10. Empirical QQ plots on (right) natural and (left) logarithmic scales of rainfall observations versus simulations from BMC-H. Dashed gray lines represent the 95% confidence bounds.

daily rainfall simulation is that autocorrelation of rainfall amounts can be properly taken into account in a relatively natural and easy way. This is an important advantage of BMC over the conventional two-state Markov chain generator (CMC) [*Gabriel and Neumann*, 1962; *Richardson*, 1981].

[65] Both BMC and CMC have the same assumption for rainfall occurrence. They would not have significant distinction in preserving basic occurrence statistics (Table 4). A major difference between BMC and CMC is that the latter assumes independence of rainfall amounts, whereas the former does not. Hence, we focus on the performance in reproducing autocorrelation of rainfall amounts. For the sake of fair comparison, the HEG distribution was used for both models. Figure 15 is the same as Figure 12 but for the CMC model. As observed from Figure 15, the simulated Kendall's τ correlation coefficients were almost symmetrically distributed about 0 throughout the year without clear seasonal cycles and regardless of observed values. Nearly the same observations hold for the upper tail dependence coefficients. The underestimated autocorrelation of rainfall amounts will in turn lead to misrepresentation of rainfall event volumes, as signified by overestimated or underestimated 2- and 3-day event volumes. In summary, CMC fails to transfer the autocorrelation information inherent in observed rainfall records into simulated sequences, whereas BMC successfully does that.

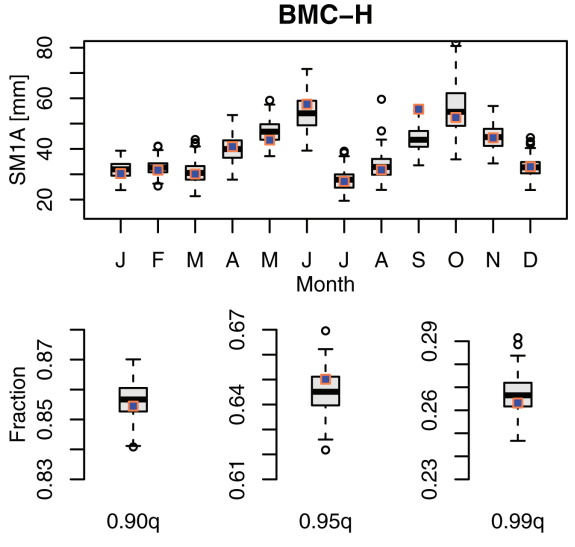

Figure 11. Box plots of (top) SM1A and (bottom) rainfall fractions due to wet days with amounts greater than 0.90, 0.95, and 0.99 quantiles of observations and simulations from BMC-H. Orange rectangulars with blue filled denote observed values.

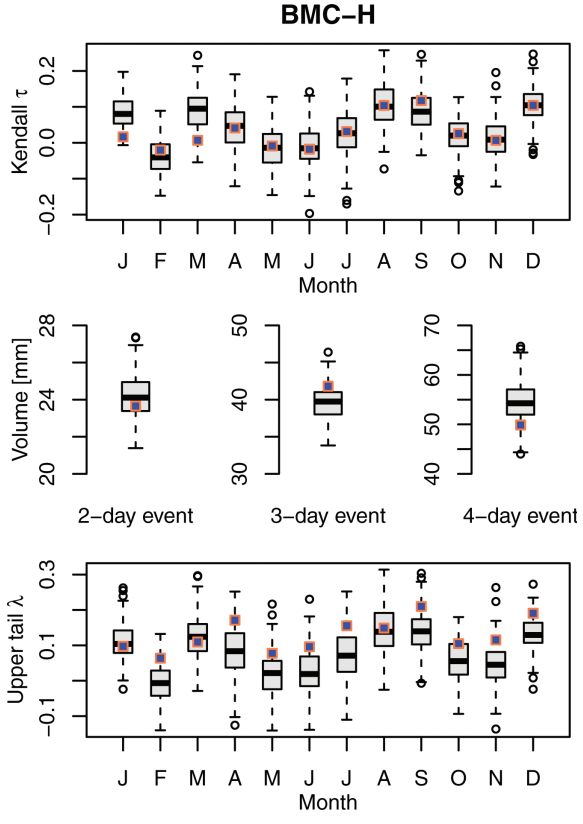

Figure 12. Box plots of (top) Kendall's τ correlation coefficient, (middle) 2-, 3-, and 4-day rainfall event volumes, and (bottom) upper tail dependence coefficient of observations and simulation from BMC-H. Orange rectangulars with blue filled denote observed values.

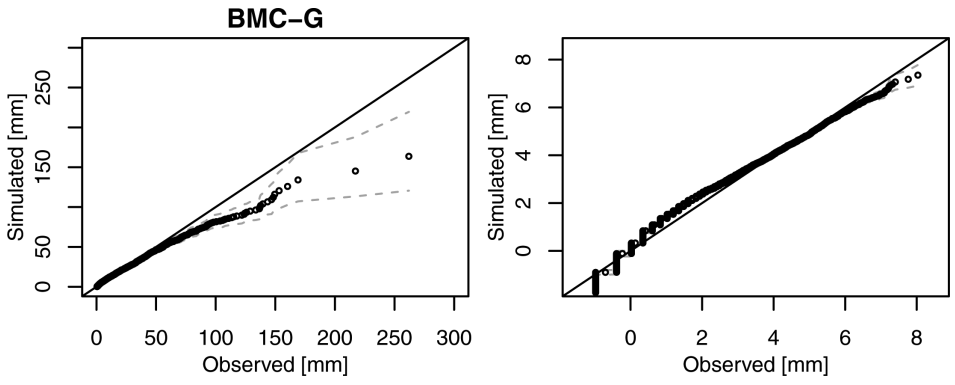

Figure 13. Empirical QQ plots on (right) natural and (left) logarithmic scales of rainfall observations versus simulations from BMC-G. Dashed gray lines represent the 95% confidence bounds.

[66] From Figure 15, one might have also noticed that the 4-day rainfall event volume was somewhat better reproduced by CMC than by BMC, even though the former assumes that rainfall amounts are independent and identically distributed. It means that long-term rainfall event volume is independent of lag-1 autocorrelation of rainfall amounts. In view of this observation, we can remark that it seems not necessary to take a more advanced model for high-order autocorrelation of the rainfall records under consideration and that first-order Markovian dependence (as is used in BMC) seems adequate.

4.6. Benefit for Reducing Effects of "Overdispersion"

[67] A typical challenge in daily rainfall simulation is the effect of overdispersion, which generally refers to the case where simulated rainfall only represents a smoothed long-term variance [*Katz and Zheng*, 1999; *Wilks*, 1999; *Mehrotra and Sharma*, 2007a, 2007b; *Kim et al*., 2012]. There are two types of overdispersion. One is related to the rainfall occurrence process, as indicated by underestimated long-term dry and wet spells, and the other one to the rainfall amount process, as signified by deflated variance of seasonal and annual rainfall totals. It is opined that the amount of overdispersion partly results from the independence assumption for rainfall amounts. The BMC generator considers autocorrelation of rainfall amounts and thus should be able to reduce amount of overdispersion. To gain

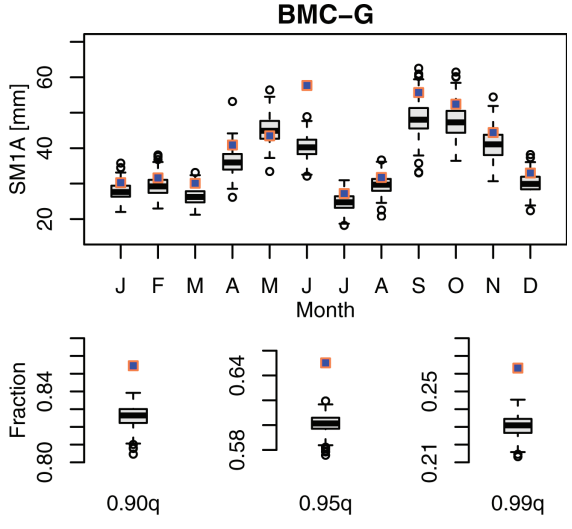

Figure 14. Box plots of (top) SM1A and (bottom) rainfall fractions due to wet days with amounts greater than 0.90, 0.95, and 0.99 quantiles of observations and simulations from BMC-G. Orange rectangulars with blue filled denote observed values.

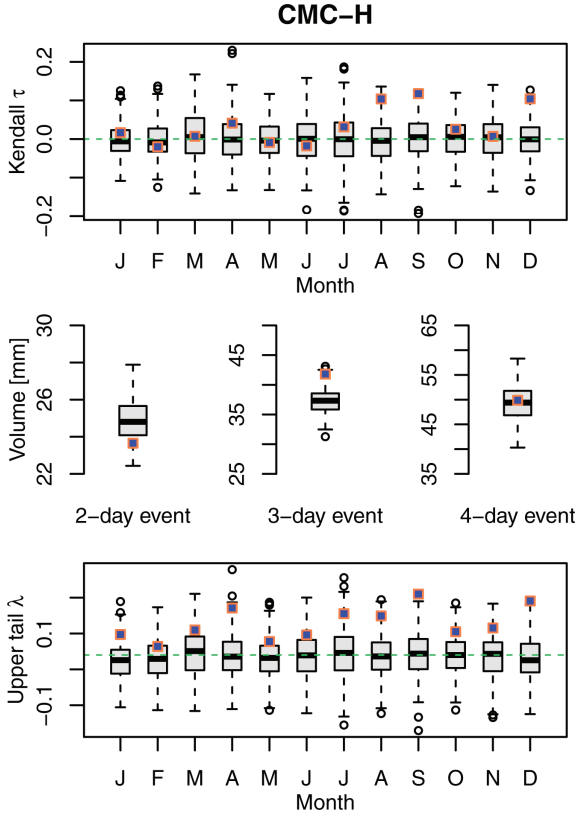

Figure 15. Box plots of (top) Kendall's τ correlation coefficient, (middle) 2-, 3-, and 4-day rainfall event volumes, and (bottom) upper tail dependence coefficient of observations and simulations from CMC-H. Orange rectangulars with blue filled denote observed values. Dashed green lines are zero reference lines.

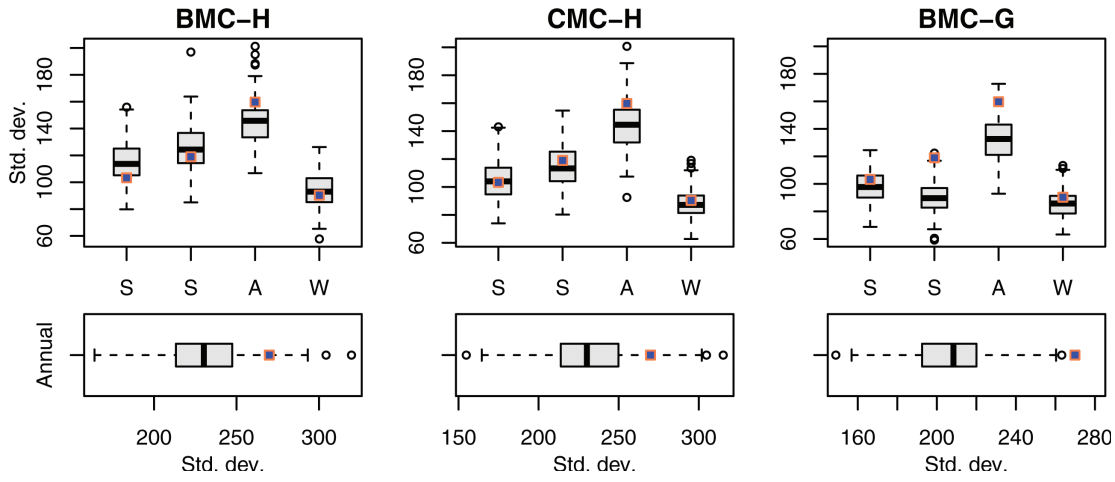

Figure 16. Box plots of (top) standard deviations of seasonal and (bottom) annual rainfall totals of observations and simulations from (left) BMC-H, (middle) CMC-H, and (right) BMC-G, respectively. Orange rectangulars with blue filled denote observed values.

insights into this point, we calculated standard deviations of seasonal and annual rainfall totals (Figure 16). Yet, there was no evident difference between BMC-H and CMC-H (CMC with HEG distribution for rainfall amounts), which implies that properly modeling lag-1 autocorrelation seems to contribute little if any to the reduction of overdispersion here. We then proceeded to look for the standard deviations corresponding to BMC-G, as presented in Figure 16 (right). At this time, the effect of overdispersion became apparent, especially for rainfall totals of dry seasons and at annual scale. Considering the difference between BMC-H and BMC-G, the above analyses suggest that to reduce overdispersion, preserving lag-1 autocorrelation is relatively less important than preserving extreme rainfall characteristics. The improved bivariate distribution provides a gain in reducing the effect of overdispersion.

5. Comparison With Other Advanced Daily Rainfall Generators

[68] In terms of comparing BMC-H with simple benchmark models (BMC-G and CMCs), our purpose above was to efficiently appreciate the advantage of BMC-H in reproducing characteristics related to extreme rainfall and lag-1 autocorrelation of rainfall amounts. One might be more willing to see comparison between BMC-H and other relatively advanced models. In this section, we compare BMC-H with two alternate models. Both of them are among the most frequently used stochastic generators for daily rainfall. One is the TPM model [*Haan et al*., 1976; *Srikanthan and McMahon*, 1985], and the other one is a semiparametric model with parametric Markov chain for rainfall occurrences and nonparametric KDE for rainfall amounts (SMC-K). The SMC-K model is rooted in the one more recently developed by *Harrold et al*. [2003b], wherein a somewhat complex algorithm is used for the generation of rainfall occurrences [*Harrold et al*., 2003a]. Nevertheless, its application requires large sample size [*Srikanthan et al*., 2005]. Considering the relatively short records available for this research, we replace it by a simple Markov chain model. For the benefit of the reader, TPM and SMC-K are introduced in Appendices B and C, respectively.

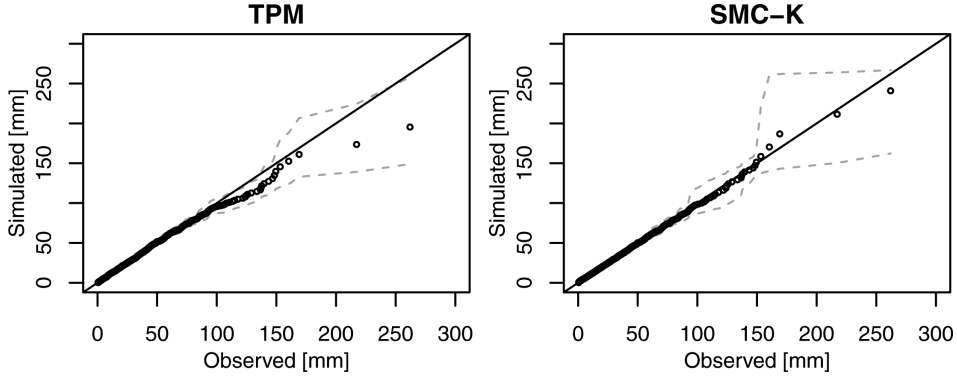

Figure 17. Empirical QQ plots of rainfall observations versus simulations from (left) TPM and (right) SMC-K, respectively. Dashed gray lines represent the 95% confidence bounds.

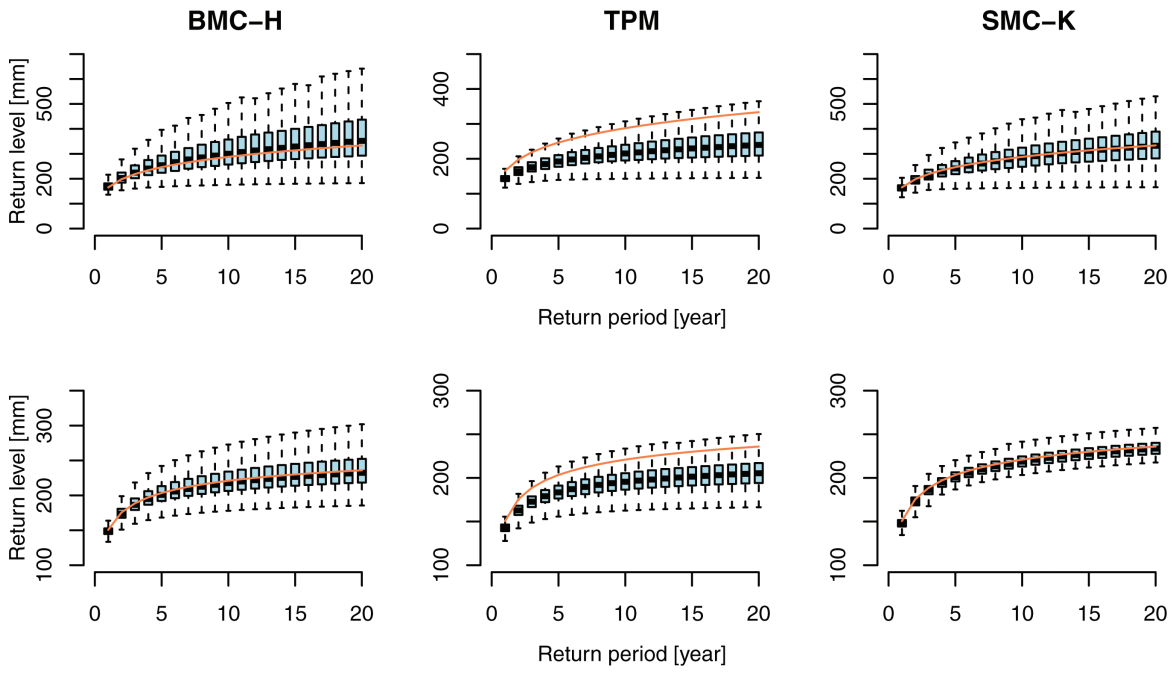

Figure 18. The return period and return level relationships derived from frequency analysis performed on (top) annual block maximum daily rainfall and (bottom) rainfall exceedances over 0.90 quantiles corresponding to observations (solid lines) and simulations (box plots) from (left) BMC-H, (middle) TPM, and (right) SMC-K, respectively.

5.1. Difference in Reproducing Basic Occurrence and Amount Statistics

[69] A quantitative assessment of the performance of BMC-H, TPM, and SMC-K in reproducing basic occurrence statistics was made by comparing RMSEs of each statistics. The results for both BMC-H and TPM are given in Table 3. In general, BMC-H performed a little bit inferior to TPM. A multistate Markov chain model seems more suitable for the rainfall occurrence process here. Note that SMC-K and CMC-H apply exactly the same algorithm for the simulation of rainfall occurrences. Comparing BMC-H with SMC-K in reproducing basic occurrence statistics is therefore equivalent to comparing it with CMC-H. As was already discussed in section 4.4, BMC-H presented similar performance to CMC-H regarding the simulation of rainfall occurrences. It should perform similarly to SMC-K as well. With regard to basic amount statistics (mean and standard deviation), the three models performed nearly the same. No one was completely convincingly better than the other.

5.2. Difference in Reproducing Overall Distributional Properties of Rainfall Amounts

[70] Figure 17 presents QQ plots of the observed against generated rainfall amounts from the TPM and SMC-K models, respectively. Generally, BMC-H outperformed TPM but slightly inferior to SMC-K, which exhibited surprising correspondences between observed and simulated quantiles, in both tail and central parts. The nice correspondence arises from the basic machinery of SMC-K. As illustrated in Appendix C, simulating rainfall amounts from a target density built by KDE is a kind of conditional bootstrapping smoothed by Gaussian kernels [*Sharma and O'Neill*, 2002]. Through smoothing, values different from observations can be simulated; however, the large-sample behavior of the smoothed values is statistically similar to the observations. It may, however, be noted that such sampling machinery of SMC-K might fail to bridge the gap between the "bulk" and the tail of observed rainfall amounts and the gap between extreme values sparsely scattered in the tail domain if these gaps are too wide. In the TPM model, rainfall amounts are divided into a number of wet states according to their magnitudes. The uniform distribution is used for rainfall amounts within each of the wet states except for the last, for which a shifted gamma distribution is assumed. This approach can essentially be seen as a piecewise linear approximation to the distribution function below a threshold, with a gamma-shaped tail above. The piecewise modeling did offer somewhat an improvement in characterizing the tail behavior of rainfall but still not as well as BMC-H and SMC-K.

[71] As mentioned above, simulating rainfall amounts by SMC-K is a conditional reshuffle-perturbation procedure performed on available observed data. Therefore, it is expected to preserve most observed distributional properties of rainfall amounts, including the overall distribution of rainfall amounts and distributions of extremes, such as block maxima and peaks over threshold. To compare the three models in reproducing distributions of rainfall extremes, two types of frequency analysis were performed. One is on annual maxima of daily rainfall. The other one is on rainfall exceedances over a threshold (0.90 quantile). The results of daily rainfall return period

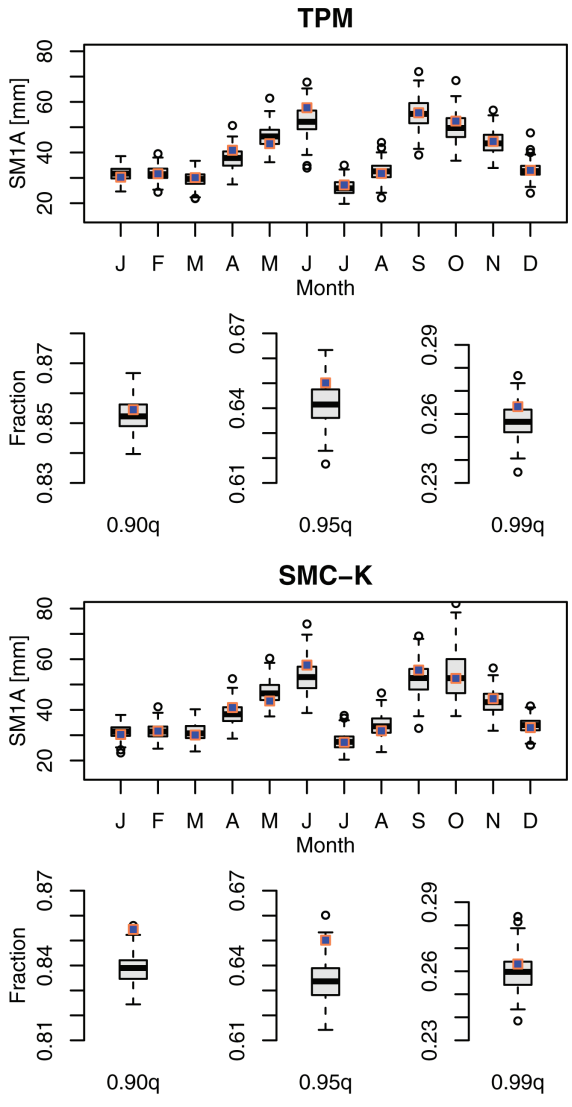

Figure 19. Box plots of SM1A and rainfall fractions due to wet days with amounts greater than 0.90, 0.95, and 0.99 quantiles of observations and simulations from (top) TPM and (bottom) SMC-K, respectively. Orange rectangulars with blue filled denote observed values.

and return level (P-L) relationships derived from the two types of frequency analysis, respectively, are presented in Figure 18. Among the three models, SMC-K and BMC-H performed comparably, whereas TPM seriously underestimated both types of P-L relationship. It was recognized that owing to the screening procedure, the P-L relationships can be reasonably preserved by BMC-H. Without screening, a small number of extremely large values might be simulated, which would raise these relationships, especially the first type. When compared with BMC-H, SMC-K tends to generate rainfall realizations bearing a somewhat too close resemblance to available observed data. BMC-H provides more diverse rainfall realization scenarios especially with respect to extremes which will in turn offer diverse risk scenarios. In situations, if rare rainfall events are of particular interest, BMC-H is preferable.

5.3. Difference in Reproducing Characteristics of Extreme Rainfall Events

[72] The comparison of the observed and simulated SM1A and large rainfall fractions by TPM and SMC-K, respectively, is shown in Figure 19. Together with Figure 11, Figure 19 indicates the following: (1) the three models performed comparably in reproducing SM1A, and (2) BMC-H performed similarly to TPM and better than SMC-K in reproducing large rainfall fractions. SMC-K demonstrated a trend toward underestimating rainfall fractions, implicating that the nonparametric KDE is less apt at capturing unusual or rare rainfall events. One point worth noting is with respect to the computation of SM1A. As was explained in section 4.2, SM1A represents a smoothed value over years. Figure 20 demonstrates the highest rather than the smoothed maximum 1-day rainfall amount of each month. To distinguish between these two quantities, the shorthand notation for this highest order statistics will be M1A. As observed from this figure, BMC-H emerged as the best choice, whereas neither TPM nor SMC-K performed as well as MBC-H. In particular, both TPM and SMC-K underestimated M1A. The former performed slightly better than the latter. The reasoning for the difference lies in the treatment of rainfall extremes. BMC-H

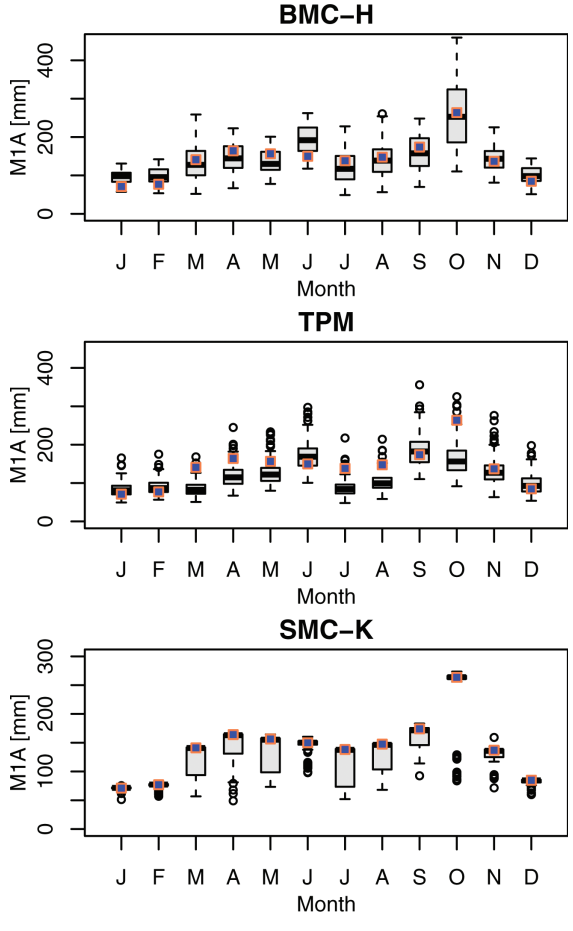

Figure 20. Box plots of M1A of observations and simulations from (top) BMC-H, (middle) TPM, and (bottom) SMC-K. Orange rectangulars with blue filled denote observed values.

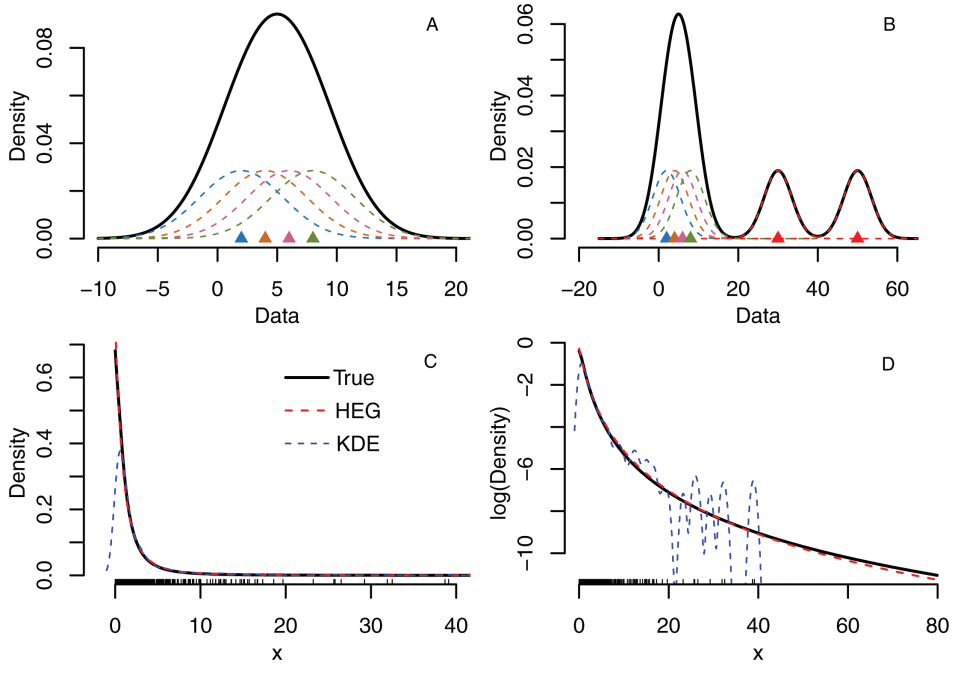

Figure 21. (a) Demonstration for kernel density estimation based on sample from a light-tailed distribution. (b) The same as in Figure 21(a) but based on sample from a heavy-tailed distribution. (c) Kernel density estimation and parametric HEG probability density function fitted to data sampled from a dynamic mixture of gamma and generalized Pareto distribution. (d) The same as in Figure 21(c) but potted on a semilogarithmatic scale.

characterizes rainfall exceedances over a threshold by a generalized Pareto distribution, which is motivated from the extreme value theory specifically for describing unusual behavior or rare events, and thus can properly reproduce historical extremes and reasonably extrapolate unseen values beyond the upper range of available observed data. TPM characterizes large rainfall in terms of a gamma distribution, which can simulate values significantly greater than the observed extremes [*Srikanthan et al*., 2005]. However, the gamma distribution is not heavy enough to adequately capture the unusual tail behavior. This explains why TPM underestimated M1As for several months and why many outliers appeared in the box plots. SMC-K uses a nonparametric KDE to approximate the distribution of rainfall amounts. An implicit assumption of KDE is that the support of the underlying distribution is the same as the range of the available sample, or at least approximately so. This assumption implies that values much greater than the observed extremes cannot be simulated unless larger bandwidth is exerted on kernels in the tail domain. This interprets why SMC-K-simulated M1A had a J-shaped distribution.

5.4. Further Explanation for the Difference in Reproducing Extreme Rainfall Event Characteristics

[73] Analyses in the above section suggest that when compared with BMC-H and TPM, SMC-K is less apt at reproducing characteristics of extreme rainfall events. This is consistent with the statement made by *Srikanthan et al*. [2005]. It is also in agreement with the point of view of existing researches that KDE might provide misleading tail behavior for heavy-tailed data, for instance, *Markovich* [2007] and *Carreau and Bengio* [2009] amongst many others. At the first glance, however, it seems against the previous observations in section 5.2 that SMC-K performs rather well in reproducing distributions of overall rainfall amounts and extremes. To see why they make sense, we first enumerate typical characteristics of samples from heavy-tailed and light-tailed distributions. If a random sample is from a heavy-tailed distribution, then there are sparse observations or "outliers" isolated from other values or the "bulk" of the sample, as demonstrated by the rug plot in Figure 21(b). Although if a random sample is from a light-tailed distribution, there is no such outliers and all the sample data are compactly distributed, as demonstrated by the rug plot in Figure 21(a). For a light-tailed distribution, the nonparametric KDE is a good estimator for the corresponding PDF, as illustrated by the solid line in Figure 21(a). However, if the distribution is heavy-tailed, KDE provides a misleading approximation to the tail domain. In general, it has sharp bumps centered at the outliers and does not provide correct rate of the PDF decay to 0 [*Silverman*, 1986], as illustrated here by the solid line in Figure 21(b). As a result, high quantiles of the underlying distribution will be underestimated. Moreover, a PDF with bumps in the tail domain is hard to interpret, representing an unrealistic feature in rainfall.

[74] The distorted tail estimation for heavy-tailed data by KDE can be more realistically demonstrated in terms of a simple Monte Carlo experiment. First, we generated a random sample of size 1000 from a dynamic mixture of gamma and generalized Pareto distribution, which has been extensively used for daily rainfall simulation and

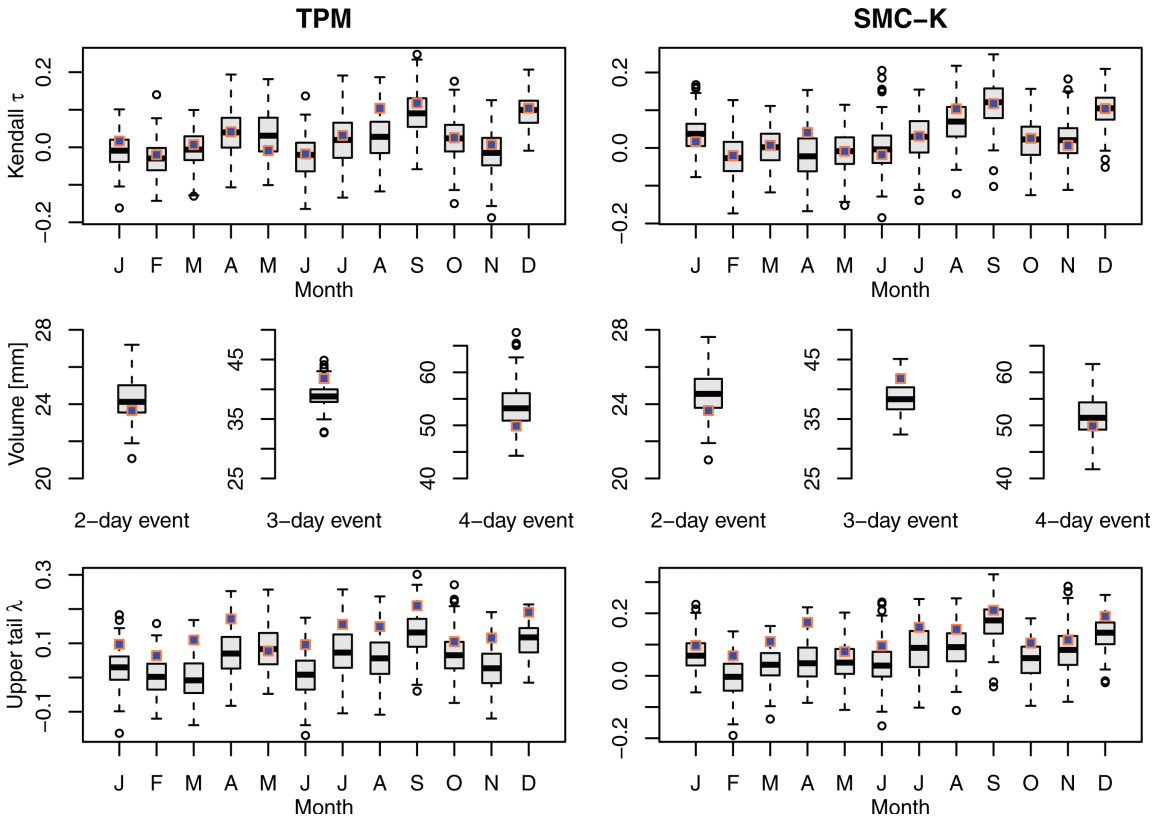

Figure 22. Box plots of (top) Kendall's τ correlation coefficient, (middle) 2-, 3-, and 4-day rainfall event volumes, and (bottom) upper tail dependence coefficient of observations and simulations from (left) TPM and (right) SMC-K, respectively. Orange rectangulars with blue filled denote observed values.

downscaling [*Vrac and Naveau*, 2007; *Hundecha et al.*, 2009; *Hundecha and Merz*, 2012]. The parent distribution was parameterized such that it honors typical distributional characteristics of daily rainfall. We fitted the random sample using nonparametric KDE with Gaussian kernel (Appendix C) and parametric HEG distribution, respectively, and the results are presented in Figures 21(c) and 21(d). The rug plots indicate existing sparse observations toward the tail end of the distribution. Figure 21(c) implies that both KDE and HEG provided a reasonable fit for low to moderate values. However, the tail part was distorted by the former, which is efficiently illustrated by Figure 21(d). The distorted tail will cause the underestimation of high quantiles. For instance, the real 0.9995 quantile is 64.77, whereas the KDE-estimated quantile was 39.28, which is slightly different from the largest sample value (39.11) only.

[75] It must, however, be emphasized that the above discussion should not be interpreted as suggesting either parametric distributions or nonparametric estimators are preferred to the other type, but properly appreciating the different working machineries of BMC-H and SMC-K for simulating rainfall amounts. In situations where values much higher than the observed extremes are not as prominent, nonparametric rainfall generators (such as SMC-K) are preferred. Here, we want to point out again that as a stochastic rainfall generator, it should be able to reasonably extrapolate unseen rare rainfall events significantly beyond the upper range of historical records. In the context of climate change, growing extreme weather events (e.g., severe storms and snowfalls) have been witnessed in different places over the world. Most of the extreme events are unseen in recent history and may lead to large losses. Extrapolating certain high values becomes more imperative to meet climate change-induced risks faced by hydrologic design and planning.

5.5. Difference in Reproducing Lag-1 Autocorrelation of Rainfall Amounts

[76] The comparison of historical and simulated lag-1 autocorrelation of rainfall amounts, the corresponding upper tail dependence, and rainfall event volumes by TPM and SMC-K is shown in Figure 22. Recalling what have been observed in Figure 12 and comparing them with Figure 22, three major points can be inferred: (1) all the three models closely reproduced lag-1 autocorrelation of rainfall amounts; however, BMC-H preserved seasonal cycles of autocorrelation slightly better than the other two models; (2) the 2- and 3-day rainfall event volumes were slightly better reproduced by BMC-H, whereas the 4-day rainfall event volume was preserved reasonably well by the three models; and (3) the upper tail dependence of rainfall extremes was somewhat more likely to be underestimated throughout the year by TPM than by BMC-H and SMC-K.

The first point seems to be different from the conclusion obtained by *Srikanthan et al*. [2005], stating that TPM does not perform as well as SMC-K in preserving autocorrelation of rainfall amounts. This may be due to the relatively higher autocorrelations of their rainfall records (approximately ranged from 0.2 to 0.4). The TPM model is not explicitly structured for reproducing autocorrelation of rainfall amounts. In situations where the autocorrelation is small, TPM provides a decent performance. When the autocorrelation is high, the performance of TPM will deteriorate and the advantage of BMC-H and SMC-K will be apparent. Point 2 supports our previous suspicion that the assumed Markovian first-order dependence is adequate for the rainfall amount process under consideration. It is interesting to note that although the averaged behavior of upper tail dependence simulated by BMC-H and SMC-K were similar, the former presented much clearer seasonal cycles than the latter. The above three points highlight the value of BMC-H in representing central as well as upper tail dependencies of daily rainfall amounts.

6. Conclusions

[77] Based on the work by *Shimizu* [1993], *Herr and Krzysztofowicz* [2005], and *Serinaldi* [2008, 2009a, 2009b], we present an improved bivariate mixed distribution, which is useful for pairwise daily rainfall analysis. The improved distribution is more flexible in modeling various dependence structures and is capable of modeling the entire range of rainfall when it is heavy-tail distributed. Several problems involved in hydrology can be addressed with the aid of this bivariate distribution, for example, daily rainfall simulation, radar rainfall bias correction [*Smith et al*., 2012], uncertainty estimation of satellite rainfall [*Gebremichael et al*., 2011], and so forth.

[78] Among several recognized applications of the improved distribution, particularly presented here is its utility for single-site daily rainfall simulation. A stochastic daily rainfall generator is developed, which generalizes daily rainfall as a Markov process with autocorrelation described by the improved bivariate mixed distribution. Instead of breaking down rainfall occurrence simulation and amount simulation separately, the developed generator unifies them and thus autocorrelation of daily rainfall is automatically accounted for. The developed generator is first tested on a sample station in Texas. The results reveal that the simulated and observed sequences are in good agreement with respect to essential characteristics. To efficiently appreciate the advantage of the developed generator in reproducing characteristics related to extreme rainfall and the lag-1 autocorrelation of rainfall amounts, we compared it with two benchmark models (BMC-G and CMC-H).

[79] In addition, extensive simulation experiments are carried out to compare it with two other relatively advanced alternate models: the TPM model and the semiparametric Markov chain model with parametric Markov chain for rainfall occurrences and nonparametric KDE for rainfall amounts (SMC-K). The results show that (1) when compared with TPM and SMC-K, the developed generator is apt at reproducing the central and upper tail dependencies of rainfall amounts of two consecutive wet days; (2) the developed generator performs best in simulating maximum daily rainfall amount in the sense that it can reasonably extrapolate unseen rare rainfall values significantly beyond the upper range of available observed data; (3) with regard to reproducing the entire distribution of rainfall amounts and the distributions of rainfall extremes such as annual maxima and peaks over threshold, there is no clear-cut difference between the developed generator and SMC-K except that the former is more apt at providing diverse rainfall realization scenarios and risk scenarios. Another interesting observation found in this research is that to reduce the amount of overdispersion, preserving lag-1 autocorrelation is relatively less important than preserving extreme rainfall characteristics.

[80] It is realized that the improved bivariate mixed distribution bears similarities with the error model for satellite rainfall of *Gebremichael et al*. [2011]. With the given satellite rainfall estimates, they developed two separate conditional distributions of ground observations: one for when satellite estimates are zero and another one for positive values. This bivariate mixed distribution combines both instances. It is therefore not hard to modify the presented generator for obtaining the ensemble estimates (or distribution) of ground rainfall observations with the given satellite estimates. Nevertheless, it is important to note that in this situation, the model should be calibrated and validated before putting it into application. Actually, the developed generator is applicable not only for rainfall simulation but also for streamflow simulation, especially it is suitable for stations located in arid or semiarid regions, where zero observations are not uncommon. When all observations are nonzero, the generator developed here reduces to a model similar to the one proposed by *Lee and Salas* [2011] and *Hao and Singh* [2012].

Appendix A: Derivation of the Type II Conditional Distribution

[81] As a note, derivations in this appendix make substantial use of the work of *Herr* [1999] and *Zhang* and *Singh* [2007].

A1. Case 1: $X = x$ and $X = 0$

[82] Applying the definition of conditional probability yields the following equation:

$$\Phi_{Y|X=0}(y|X=0) = P(Y \le y, X=0)/P(X=0). \tag{A1}$$

[83] The denominator of equation (A1) is

$$P(X=0) = P(X=0, Y=0) + P(X=0, Y>0) = p_{00} + p_{01}, \tag{A2}$$

and the numerator is

$$P(Y \le y, X=0) = P(Y=0, X=0)P(Y \le y | Y=0, X=0) + P(Y>0, X=0)P(Y \le y | Y>0, X=0) = p_{00} + p_{01}H_Y(y). \tag{A3}$$

[84] Substituting the denominator and numerator back into equation (A1) yields the following equation:

$$\Phi_{Y|X=0}(y|X=0) = (p_{00} + p_{01}H_Y(y))/(p_{00} + p_{01}). \quad \text{(A4)}$$

A2. Case 2: $X = x$ and $X > 0$

[85] Similarly, applying the definition of conditional probability leads to the following equation:

$$\phi_{Y|X=x}(y|X=x, X>0) = \phi(x,y|X>0, Y\geq 0)/\phi_X(x|X>0). \quad \text{(A5)}$$

[86] Marginalizing Y out from equation (2) results in the following equation:

$$\begin{aligned}\phi_X(x|X\geq 0) &= \int \phi(x,y)dy \\ &= p_{00}\delta(x) + p_{10}h_X(x) + p_{01}\delta(x) + p_{11}f(x). \end{aligned} \quad \text{(A6)}$$

[87] From equation (A6), it is apparent to obtain the denominator of equation (A5) as follows:

$$\phi_X(x|X>0) = p_{10}h_X(x) + p_{11}f(x). \quad \text{(A7)}$$

[88] Rewrite the numerator of equation (A5) as follows:

$$\phi(x,y|X>0, Y\geq 0) = p_{10}\delta(y)h_X(x) + p_{11}h(x,y). \quad \text{(A8)}$$

[89] Rewrite $h(x, y)$ in equation (A8) as follows:

$$h(x,y) = f(x)h_{Y|X}(y|X=x, X>0, Y>0). \quad \text{(A9)}$$

[90] Substituting equations (A7), (A8) and (A9) into equation (A5) yields the PDF of Y given $X = x$ and $X > 0$ as follows:

$$\phi_{Y|X=x}(y|X=x, X>0) = \frac{p_{10}\delta(y)h_X(x) + p_{11}f(x)h_{Y|X}(y|X=x, X>0, Y>0)}{p_{10}h_X(x) + p_{11}f(x)}. \quad \text{(A10)}$$

[91] Integrating equation (A10) yields the corresponding CDF as follows:

$$\Phi_{Y|X=x}(y|X=x, X>0) = \frac{p_{10}h_X(x) + p_{11}f(x)H_{Y|X}(y|X=x, X>0, Y>0)}{p_{10}h_X(x) + p_{11}f(x)}. \quad \text{(A11)}$$

[92] After applying the copula theory, $H_{Y|X}(y|X=x, X>0, Y>0)$ in equation (A11) can be expressed by equation (10). Substituting equation (10) into equation (A11) yields the following equation:

$$\Phi_{Y|X=x}(y|X=x, X>0) = \frac{p_{10}h_X(x) + p_{11}f(x)c_1(F(x), G(y))}{p_{10}h_X(x) + p_{11}f(x)}. \quad \text{(A12)}$$

Appendix B: TPM Generator

[93] The TPM model used herein follows the work of *Haan et al*. [1976] and *Srikanthan* and *McMahon* [1985], with slight modifications: (1) use 0.3 mm rather than 0.1 mm as the significant rainfall threshold; (2) adjust the number of states for each month such that there are enough data for the last state; and (3) accommodate the boundary problem caused by the changing number of states.

Table B1. State Partition of TPM Model Used in This Research

State Number	Intermediate State (mm)	Last State (mm)
1	[0, 0.3)	
2	[0.3, 0.9)	[0.3, ∞)
3	[0.9, 2.9)	[0.9, ∞)
4	[2.3, 6.9)	[2.9, ∞)
5	[6.9, 14.9)	[6.9, ∞)
6	[14.9, 30.9)	[14.9, ∞)
7	[30.9, ∞)	

[94] The states partition of the TPM model is given in Table B1. Alternations from state i to j are determined by the estimated TPM P whose elements are p_{ij}, where i, $j = 1, 2, \ldots, c$, and c is the maximum number of states. The uniform distribution is used for rainfall amounts of wet states except the last for which a shifted gamma distribution is used.

[95] Due to the relatively short historical records available for this study, parameter estimates with reliable accuracy, when assuming seven states for all months, might not be obtained. We therefore adjusted the number of states such that for each month there were at least 80 observations for the last state. However, this may cause a boundary problem. For instance, suppose there are five states in June and three states in July and suppose the simulated state for last day of June happens to be 5, then for the first day of July, the simulation algorithm will crash. To avoid this problem, we arbitrarily assume a dummy state as the last state of the current month to aid the simulation of state for the first day of the month.

[96] Software for the TPM model described in *Srikanthan et al*. [2005] can be freely available at http://toolkit.ewater.com.au/Tools/SCL. It must, however, be noted that the TPM model implemented in the present research is somewhat different from the one in *Srikanthan et al*. [2005]. To avoid misleading due to different results from different models, we provide the computational procedure in MATLAB for interested readers.

Appendix C: Semiparametric Markov Chain Generator With KDE for Rainfall Amounts

[97] The simulation model described in this appendix mainly follows the work of *Harrold et al*. [2003b]. Major differences are as follows: (1) instead of the nonparametric model of *Harrold et al*. [2003a], we used the conventional two-state Markov chain model for the simulation of rainfall occurrences; (2) instead of dividing wet days into four classes, we divided them into two classes conditionally on the previous day is wet or dry; (3) instead of using 31-day moving window to accommodate rainfall seasonality, we assumed that rainfall is stationary within months; (4) we applied the least-squares cross-validation (LSCV) method suggested by *Sharma et al*. [1997] to obtain the optimal kernel smoothing parameter rather than the adjusted bandwidth derived from a trial and error procedure; (5) to avoid

generating values smaller than 0.3 mm, we repeated the random sampling until a reasonable value was obtained rather than using "variable kernel," which will lead to bias in the density estimate [*Silmonoff*, 1996; *Salas and Lee*, 2010].

[98] Suppose a synthetic rainfall occurrence sequence has been generated from a two-state Markov chain model, for the ease of explanation, we assume that the day whose amount to be simulated is in month m, then the simulation can be split into two cases.

C1. Class 1: The Previous Day Is Dry

[99] A univariate kernel density is used for rainfall amounts of class 1 as follows:

$$\hat{f}(x) = \sum_{i=1}^{n_1} N((x - x(i))/h)/n_1 h, \tag{C1}$$

where n_1 is the number of observations $x(i)$ in class 1 within month m; h is the kernel bandwidth, which is determined by LSCV; and $N(\cdot)$ is the PDF of the standard Gaussian distribution. To simulate rainfall amounts of this class, (1) first pick an x' value from $x(i)$ ($i = 1, 2, \ldots, n_1$) with equal probability; (2) then perturb this value by drawing a random variate from a Gaussian distribution with mean x' and variance h^2; and (3) repeat the perturbation until a value no less than 0.3 mm is obtained and assign it as the simulated amount.

C2. Class 2: The Previous Day Is Wet

[100] Suppose the simulated amount of previous day is x_{t-1}^g, then a conditional kernel density is used to approximate the distribution of rainfall amount x_t of current day conditioning on x_{t-1}^g

$$\hat{f}\left(x_t | x_{t-1}^g\right) = \sum_{i=1}^{n_2} w(i) N\left(x_t; b(i), \lambda^2 S'\right), \tag{C2}$$

where

$$w(i) = N\left(x_{t-1}^g; x_{t-1}(i), \lambda^2 S_{t-1}\right) / \sum_{j=1}^{n_2} N\left(x_{t-1}^g; x_{t-1}(j), \lambda^2 S_{t-1}\right),$$

$$b(i) = x_t(i) + \left(x_{t-1}^g - x_{t-1}(i)\right) S_{t-1,t} / S_{t-1},$$

$$S' = S_t - S_{t-1,t}^2 / S_{t-1},$$

n_2 is the number of observed pairs $[x_{t-1}(i), x_t(i)]$ in class 2 within month m; λ is the optimal kernel smoothing factor determined by LSCV; $N(\cdot; \alpha, \beta)$ is the PDF of a Gaussian distribution with mean α and variance β; S_{t-1} and S_t are the sample variance of $x_{t-1}(i)$ and $x_t(i)$ ($i = 1, 2, \ldots, n_2$), respectively; and $S_{t-1,t}$ is the sample covariance between $x_{t-1}(i)$ and $x_t(i)$. To simulate rainfall amounts of this class, (1) first pick an $[x_{t-1}', x_t']$ vector from $[x_{t-1}(i), x_{t-1}(i)]$ ($i = 1, 2, \ldots, n_2$) with probability $w(i)$; (2) then compute b' with $x_{t-1}(i)$ and $x_t(i)$ in $b(i)$ replaced by x_{t-1}' and x_t', respectively; (3) sample a random variate from a Gaussian distribution with mean b' and variance $\lambda^2 S'$; and (4) repeat sampling until a value no less than 0.3 mm is obtained and assign it as the simulated amount.

[101] **Acknowledgments.** This work was financially supported in part by the United States Geological Survey (USGS, Project ID: 2009TX334G) and TWRI through the project "Hydrological Drought Characterization for Texas under Climate Change, with Implications for Water Resources Planning and Management."

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control*, *19*(6), 716–722.

Bárdossy, A., and E. Plate (1992), Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resour. Res.*, *28*(5), 1247–1259.

Brissette, F. P., M. Khalili, and R. Leconte (2007), Efficient stochastic generation of multi-site synthetic precipitation data, *J. Hydrol.*, *345*, 121–133.

Cannon, A. J. (2008), Probabilistic multisite precipitation downscaling by an expanded Bernouilli-Gamma density network, *J. Hydrometeorol.*, *9*, 1284–1300.

Carreau, J., and Y. Bengio (2009), A hybrid Pareto model for asymmetric fat-tailed date: The univariate case, *Extremes*, *12*, 53–76.

Carreau, J., and M. Vrac (2011), Stochastic downscaling of precipitation with neural network conditional mixture models, *Water Resour. Res.*, *47*, W10502, doi:10.1029/2010WR010128.

Carreau, J., P. Naveau, and E. Sauquet (2009), A statistical rainfall-runoff mixture model with heavy-tailed components, *Water Resour. Res.*, *45*, W10437, doi:10.1029/2009WR007880.

Cherubini, U., E. Luciano, and W. Vecchiato (2004), *Copula Methods in Finance*, Wiley, Chichester, West Sussex, U. K.

Feller, W. (1968), *An Introduction to Probability Theory and Its Application I, II*, 3rd ed., Wiley, New York.

Frahm, G., M. Junker, and R. Schmidt (2005), Estimating the tail dependence coefficient, *Insur. Math. Econ.*, *37*, 80–100.

Furrer, E. M., and R. W. Katz (2008), Improving the simulation of extreme precipitation events by stochastic weather generators, *Water Resour. Res.*, *44*, W12439, doi:10.1029/2008WR007316.

Gabriel, K. R., and J. Neumann (1962), A Markov chain model for daily rainfall occurrence at Tel Aviv, *Q. J. R. Meteorol. Soc.*, *88*, 90–95.

Gebremichael, M., G.-Y. Liao, and J. Yan (2011), Nonparametric error model for a high resolution satellite rainfall product, *Water Resour. Res.*, *47*, W07504, doi:10.1029/2010WR009667.

Genest, C., B. Remillard, and D. Beaudoin (2009), Goodness-of-fit tests for copulas: A review and a power study, *Insur. Math. Econ.*, *44*, 199–213.

Gregory, J. M., T. M. Wigley, and P. D. Jones (1993), Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals, *Clim. Dyn.*, *8*, 299–310.

Ha, E., and C. Yoo (2007), Use of mixed bivariate distributions for deriving inter-station correlation coefficients of rain rate, *Hydrol. Processes*, *21*(22), 3078–3086.

Haan, C. T., D. M. Allen, and J. D. Street (1976), A Markov chain model of daily rainfall, *Water Resour. Res. 12*(3), 443–449.

Habib, F., W. F., Krajewski, and G. J. Ciach (2001), Estimation of rainfall interstation correlation, *J. Hydrometeorol.*, *2*, 621–629.

Hao, Z., and V. P. Singh (2012), Entropy-copula method for single-site monthly streamflow simulation, *Water Resour. Res.*, *48*, W06604, doi:10.1029/2011WR011419.

Harrold, T., A. Sharma, and S. J. Sheather (2003a), A nonparametric model for stochastic generation of daily rainfall occurrence, *Water Resour. Res.*, *39*(10), 1300, doi:10.1029/2003WR002182.

Harrold, T., A. Sharma, and S. J. Sheather (2003b), A nonparametric model for stochastic generation of daily rainfall amounts, *Water Resour. Res.*, *39*(12), 1343, doi:10.1029/2003WR002570.

Herr, H. D. (1999), A bivariate precipitation uncertainty processor for probabilistic river forecasting, M.S. thesis, p. 118, School of Eng. and Appl. Sci., Univ. of Va., Charlottesville, Va.

Herr, H. D., and R. Krzysztofowicz (2005), Generic probability distribution of rainfall in space: The bivariate mode, *J. Hydrol.*, *306*, 234–236.

Huard, D., G. Evin, and A.-C. Favre (2006), Bayesian copula selection, *Comput. Stat. Data Anal.*, *51*, 809–822.

Hughes, J. P., and P. Guttorp (1994), A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena, *Water Resour. Res.*, *30*(5), 1535–1546, doi:10.1029/93WR02983.

Hundecha, Y., and B. Merz (2012), Exploring the relationship between changes in climate and floods using a model-based analysis, *Water Resour. Res.*, *48*, W04512, doi:10.1029/2011WR010527.

Hundecha, Y., M. Pahlow, and A. Schumann (2009), Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes, *Water Resour. Res.*, *45*, W12412, doi:10.1029/2008WR007453.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, New York.

Katz, R. W. (1974), Computing probabilities associated with the Markov chain model for precipitation, *J. Appl. Meteor.*, *13*(8), 953–954.

Katz, R. W. (1977), Precipitation as a chain-dependent process, *J. Appl. Meteorol.*, *16*(7), 671–676.

Katz, R. W., and X. Zheng (1999), Mixture model for overdispersion of precipitation, *J. Clim.*, *12*, 2528–2537.

Kim, Y., R. W. Katz, B. Rajagopalan, G. P. Podesta, and E. M. Furrer (2012), Reducing overdispersion in stochastic weather generators using a generalized linear modeling approach, *Clim. Res.*, *53*, 13–24.

Kleiber, W., R. W. Katz, and B. Rajagopalan (2011), Daily spatio-temporal precipitation simulation using latent and transformed Gaussian processes, *Water Resour. Res.*, *48*, W01523, doi:10.1029/2011WR011105.

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

Lennartsson, J., A. Baxevani, and D. Chen (2008), Modeling precipitation in Sweden using multiple step Markov chains and a composite model, *J. Hydrol.*, *363*, 42–59.

Li, C., V. P. Singh, and A. K. Mishra (2012), Simulation of the entire range of daily precipitation using a hybrid probability distribution, *Water Resour. Res.*, *48*, W03521, doi:10.1029/2011WR011446.

Markovich, N. (2007), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, Wiley, Chichester, U. K.

Mehrotra, R., and A. Sharma (2005), A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences, *J. Geophys. Res.*, *110*, D16108, doi:10.1029/2004JD005677.

Mehrotra, R., and A. Sharma (2007a), A semi-parametric model for stochastic generation of multi-site daily rainfall exhibiting low-frequency variability, *J. Hydrol.*, *335*, 180–193.

Mehrotra, R., and A. Sharma (2007b), Preserving low-frequency variability in generated daily rainfall sequences, *J. Hydrol.*, *345*, 102–120.

Mehrotra, R., and A. Sharma (2010), Development and application of a multisite rainfall stochastic downscaling framework for climate change impact assessment, *Water Resour. Res.*, *46*, W07526, doi:10.1029/2009WR008423.

Mehrotra, R., S. Westra, A. Sharma, and R. Srikanthan (2012), Continuous rainfall simulation: 2. A regionalized daily rainfall generation approach, *Water Resour. Res.*, *48*, W01536, doi:10.1029/2011WR010490.

Nelsen, R. B. (2006), *An Introduction to Copulas*, 2nd ed., Springer, New York.

Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other variables, *Water Resour. Res.*, *35*(10), 3089–3101.

Richardson, C. W. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resour. Res.*, *17*(1), 182–190.

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

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Serinaldi, F. (2008), Analysis of inter-gauge dependence by Kendall's τ_K, upper tail dependence coefficient, and 2-copulas with application to rainfall fields, *Stochastic Environ. Res. Risk Assess.*, *22*, 671–688.

Serinaldi, F. (2009a), Copula-based mixed models for bivariate rainfall data: An empirical study in regression perspective, *Stochastic Environ. Res. Risk Assess.*, *23*, 677–693.

Serinaldi, F. (2009b), A multisite daily rainfall generator driven by bivariate copula-based mixed distributions, *J. Geophys. Res.*, *114*, D10103, doi:10.1029/2008JD011258.

Sharma, A., D. G. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*(2), 291–308, doi:10.1029/96WR02839.

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:10.1029/2001WR000953.

Sharma, A., and R. Mehrotra (2010), Rainfall generation, in *Rainfall: State of the Science*, edited by M. Gebremicheal, pp. 215–246, American Geophysical Union, Washington, D. C.

Shimizu, K. (1993), A bivariate mixed lognormal distribution with an analysis of rainfall data, *J. Appl. Meteorol.*, *32*, 161–171.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Silmonoff, J. S. (1996), *Smoothing Methods in Statistics*, Springer, New York.

Smith, J. A., M. L. Baeck, G. Villarini, C. Welty, A. J. Miller, and W. F. Krajewski (2012), Analyses of long-term, high-resolution radar rainfall data set for the Baltimore metropolitan region, *Water Resour. Res.*, *48*, W04504, doi:10.1029/2011WR010641.

Srikanthan, R., and T. A. McMahon (1985), Stochastic generation of rainfall and evaporation data, Tech. Pap. 84, Aust. Resour. Counc., Canberra.

Srikanthan, R., and T. A. McMahon (2001), Stochastic generation of annual, monthly, and daily climate data: A review, *Hydrol. Earth Syst. Sci.*, *5*, 653–670, doi:10.5194/hess-5-653-2001.

Srikanthan, R., T. I. Harrold, A. Sharma, and T. A. McMahon (2005), Comparison of two approaches for generation of daily rainfall data, *Stochastic Environ Res. Risk Assess.*, *19*, 215–226, doi:10.1007/s00477-004-0226-0.

Tarpanelli, A., M. Franchini, L. Brocca, S. Camici, F. Melone, and T. Moramarco, (2012), A simple approach for stochastic generation of spatial rainfall patterns, *J. Hydrol.*, *472*, 63–76.

Thompson, C. S., P. J. Thomson, and X. Zheng (2007), Fitting a multisite daily rainfall model to New Zealand data, *J. Hydrol.*, *340*, 25–39, doi:10.1016/j/jhydrol.2007.03.020.

Todorovic, P., and D. A. Woolhiser (1975), A stochastic model for n-day precipitation, *J. Appl. Meteorol.*, *14*(1), 17–24.

Vrac, M., and P. Naveau (2007), Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resour. Res.*, *43*, W07402, doi:10.1029/2006WR005308.

Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol.*, *210*, 178–191.

Wilks, D. S. (1999), Interannual variability and extreme-value characteristics of several stochastic daily precipitation models, *Agric. For. Meteorol.*, *93*, 153–169.

Wilks, D. S. (2009), A gridded multisite weather generator and synchronization to observed weather data, *Water Resour. Res.*, *45*, W10419, doi:10.1029/2009WR007902.

Yoo, C., and E. Ha (2007), Effect of zero measurements on the spatial correlation structure of rainfall, *Stochastic Environ. Res. Risk Assess.*, *21*, 287–297.

Zhang, L., and V. P. Singh (2007), Bivariate rainfall frequency distributions using Archimedean copulas, *J. Hydrol.*, *332*, 93–109.