Journal of Hydrology 529 (2015) 1442–1450

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Modeling of daily rainfall sequence and extremes based on a semiparametric Pareto tail approach at multiple locations

Byung-Jin So [a], Hyun-Han Kwon [a,*], Dongkyun Kim [b], Seung Oh Lee [b]

[a] Department of Civil Engineering, Chonbuk National University, Jeonju-si 561-756, Republic of Korea
[b] Department of Civil Engineering, Hongik University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 7 April 2015
Received in revised form 8 August 2015
Accepted 17 August 2015
Available online 28 August 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Uwe Haberlandt, Associate Editor

Keywords:
Daily rainfall
Piecewise Kernel-Pareto distribution
Multi-site
Markov chain
Extreme rainfall

SUMMARY

A stochastic generation framework for simulation of daily rainfall at multiple sites is presented in this study. The limitations of a Gamma distribution-based Markov chain model for reproducing high-order moments are well-known, and the problems have increased the uncertainties when the models are used in establishing water resource plans. In this regard, this study attempted to develop a semiparametric model based on a piecewise Kernel-Pareto distribution for simulation of daily rainfall in order to further improve the existing model in terms of reproducing extremes, and in addition, the algorithm to reproduce the spatial correlation was combined. The proposed model can essentially be seen as a piecewise distribution approach constructed by parametrically modeling the tails of the distribution using a generalized Pareto and the interior by kernel density estimation methods. As a result, a Kernel-Pareto distribution-based Markov chain model has been shown to perform well at reproducing most statistics, such as mean, standard deviation, skewness and kurtosis. The proposed model provided a significantly improved estimate of design rainfalls for all the stations.

1. Introduction

A long-term water resource plan is driven by the need to establish strategies for mitigating the adverse impact of uncertain climate risks and their effects. For this purpose, stochastic generation models are widely used to generate plausible scenarios of weather variables that have statistically similar characteristics as that of observed variables. The stochastic model is only able to reproduce statistical properties of weather variables when the observed record is long enough to understand the underlying structure in a statistical way. However, these observed records, providing sufficient temporal and spatial coverage, are not readily available. A way to alleviate this is to generate artificial sequences of weather variables with finite length and use these as inputs to hydrologic models for risk and reliability assessment in the design and operation of water resource systems, given rare occurrences of these variables. Rainfall, temperature, and solar radiation are the main weather variables modeled through the stochastic model. In particular, rainfall at a daily time scale is one of the key variables in the hydrologic applications.

Stochastic models generating rainfall sequences at a single site are easy to develop and implement. A common approach to modeling at single sites has been developed by assuming a low-order Markovian process (Gabriel and Neumann, 1962; Katz and Zheng, 1999; Salas, 1993; Wilks, 1999a; Wilks and Wilby, 1999) describing the temporal dependence of rainfall occurrence patterns, and employing the probability density function (pdf) such as gamma distribution, exponential distribution, and some mixture distribution representing rainfall amounts given its occurrences (e.g. wet days) independently (Kim et al., 2012; Richardson, 1981; Stern and Coe, 1984; Woolhiser and Pegram, 1979). Wilks and Wilby (1999) provide a comprehensive introduction and history of stochastic weather generators. Recent reviews of stochastic rainfall generators have been covered by Maraun et al. (2010) and Wilks (2010).

There has been a major concern about the overdispersion in daily rainfall simulation which commonly refers to the case where simulated rainfall underestimates rainfall extremes and first-order persistence (Buishand, 1978; Katz and Parlange, 1998; Katz and Zheng, 1999; Kim et al., 2012; Mehrotra and Sharma, 2007a, 2007b), and suggesting their use can lead to an unrealistic scenario for impact studies of water resources associated with the risk of climate variability. The overdispersion is due to either an inappropriate model structure for high frequency variability in daily rain-

* Corresponding author.
E-mail address: hkwon@jbnu.ac.kr (H.-H. Kwon).

http://dx.doi.org/10.1016/j.jhydrol.2015.08.037
0022-1694/

fall series or low frequency variability in the climate system. The gamma and exponential distribution are commonly used to model the amounts of daily rainfall, but it often fails to characterize the underlying properties of the extreme daily rainfall (Wilks, 1999a). Wilks (1999a) indicated that an exponential mixture model showed better performances in cases that the extremes are only less than 100 mm. Qian et al. (2008) and Semenov (2008) applied semiparametric approaches using empirical distribution to improve simulations of extreme rainfall, and results indicated that the extremes are reproduced reasonably well. Other approaches have been implemented using mixture distributions (e.g. gamma and generalized Pareto distribution) and showed a substantial improvement in terms of reproducing extreme rainfall (Furrer and Katz, 2008; Vrac and Naveau, 2007).

Other recent interests include spatio-temporal rainfall modeling via multi-site weather generators (Kleiber et al., 2012) that allow us to simultaneously simulate spatially and temporally correlated rainfall sequences. It is decisive to preserve the spatial correlation of the behavior of precipitation intensity and occurrences in watershed scales, and these play important roles in water resource planning and management. Wilks (1998) laid the foundation for spatio-temporal modeling of rainfall based on multivariate Gaussian distribution to simulate correlated occurrence and amount in daily rainfall sequences. In recent years, many methods have been proposed for the spatial and temporal modeling of rainfall at multiple stations, including k-nearest resampling models (Apipattanavis et al., 2007; Rajagopalan and Lall, 1999), hidden Markov models (Ailliot et al., 2009; Charles et al., 1999; Hughes et al., 1999; Khalil et al., 2010; Kwon et al., 2009), generalized chain-dependent process models (Zheng and Katz, 2008a, 2008b; Zheng et al., 2010), and copula-based models (Bardossy and Pegram, 2009; Li et al., 2013).

Although various approaches have been proposed for spatio-temporal modeling, the underestimation of extremes in multi-site daily rainfall simulation models has not been extensively studied. In view of the above mentioned limitations, this study proposes a semiparametric multi-site daily rainfall simulation model to better reproduce extremes. For the multi-site model, this study employs a parametric model for spatially correlated rainfall occurrence and amount introduced by Wilks (1998). The semiparametric model in this study can essentially be seen as a piecewise distribution approach constructed by parametrically modeling the tails (i.e. above a threshold) of the distribution using a generalized Pareto, and the interior (i.e. below) by kernel density estimation methods. The proposed model is compared to the well-known daily rainfall simulator based on a two-state, the first order Markov chain with two-parameter Gamma distribution (hereafter referred to as "Gamma model"). The paper is arranged as follows. In Section 2, we first introduce the proposed semiparametric modeling approach for the multi-site Markov chain model for daily rainfall series in South Korea. The study area and the data used are then described briefly in Section 3. We illustrate the proposed model on a network of 15 stations in South Korea that are taken from the database of the Korea Meteorological Administration (KMA), and the results are then discussed in the order of the key research questions indicated above in Section 4. Conclusions and recommendations are then presented.

2. Methodology

The proposed model for generating spatially correlated rainfall sequences is to first generate rainfall occurrences, and then simulate rainfall amounts at multiple locations. Especially, this study much focuses on a semiparametric multi-site daily rainfall simulation model to better reproduce extremes. This section begins describing the first-order Markov chain approach with the occurrence model.

2.1. Single-site Markov Chain Model (SSMC)

The daily rainfall simulation model proposed in this study is a chain-dependent process which consists of a first-order two-state Markov process representing rainfall occurrence, with serially independent rainfall amounts on wet days. The use of the Markov chain for the modeling of rainfall occurrences has been suggested by many studies (Gabriel and Neumann, 1962; Haan et al., 1976; Mehrotra and Sharma, 2007a; Todorovic and Woolhiser, 1975; Wilks and Wilby, 1999). The observed rainfall sequences are regarded simply as a series of two states (i.e. rain or no rain day) that can be treated as binary sequences (i.e. 0 or 1) with the first-order Markov chain model representing the temporal dependence between rain and no rain days on successive days.

2.2. Multi-site Markov chain model

Floods associated with heavy convective rainfall may be connected with a relatively localized area of the watershed so that rainfall modeling using a simple SSMC for an entire watershed may result in an underestimation or overestimation of areal rainfall (or design rainfall), which leads to an unreliable estimation of design floods. For these reasons, the multi-site rainfall simulation model for weather networks would require preserving spatial dependence. This study employed Wilks's approach in representing the spatial variability of rainfall. Many studies have been based on Wilks's approach (Khalili et al., 2009; Mehrotra and Sharma, 2007a; Mehrotra et al., 2006; Qian et al., 2002; Wilks, 1999b), and the Wilks's approach has been further improved in multi-site rainfall simulators (Brissette et al., 2007; Khalil et al., 2010, 2009; Srikanthan and Pegram, 2009). In this section, the Wilks's approach is first briefly outlined.

Wilks (1998) extended the SSMC for rainfall to a multi-site model by joining individual SSMCs through a sequence of random numbers that is a spatially correlated, but temporally independent, and the Wilks method works well with parametric weather generators such as WGEN (Richardson, 1981). The spatial correlation structure associated with the sequence of random numbers is derived from an empirical process in order to guarantee proper representation of observed spatial correlation in the rain-gauge network.

Given a network of m weather stations, there are $m(m-1)/2$ pairwise correlations which need to be reproduced in the uniform random vector ($\mathbf{u}$), used for a rainfall occurrence model in the Markov chain process. Basically, the vector $\mathbf{u}$ is temporally independent (i.e. $Corr(u_n(a), u_{n+1}(b)) = 0$) while spatially correlated (i.e. $Corr(u_n(a), u_n(b)) \neq 0$) across weather stations. If the observed rainfall occurrence is denoted as $O_n^o(a)$ and $O_n^o(b)$ for locations a and b, respectively, a sample correlation coefficient between $O_n^o(a)$ and $O_n^o(b)$ is represented as:

$$c^o(a,b) = Corr(O_n^o(a), O_n^o(b)) \tag{1}$$

The uniform random variates, $u_n(a)$ and $u_n(b)$, are first transformed to Gaussian random variates, $g_n(s)$, through a Gaussian cumulative density function ($\Phi(\cdot)$) in order to specify the spatial dependence at location s, and the correlation coefficient between the Gaussian random variates $g_n(a)$ and $g_n(b)$ can be written as Eq. (3).

$$u_n(s) = \Phi(g_n(s)) \tag{2}$$

$$cg(a,b) = Corr(g_n(a), g_n(b)) \tag{3}$$

We then need to find the correlation $cg(a,b)$ so that the random variates $u_n(a)$ and $u_n(b)$ result in simulated rainfall occurrences $O_n(s)$ that show a correlation coefficient $c(a,b)$ that is similar to that of observed $c^o(a,b)$. There is a problem that spatial correla-