Environmental Modelling & Software 63 (2015) 230–239

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A programming tool to generate multi-site daily rainfall using a two-stage semi parametric model

Rajeshwar Mehrotra [a,*], Jingwan Li [a], Seth Westra [b], Ashish Sharma [a]

[a] Water Research Centre, School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia
[b] School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, Australia

A R T I C L E I N F O

Article history:
Received 13 May 2014
Received in revised form
13 October 2014
Accepted 13 October 2014
Available online 7 November 2014

Keywords:
Markov model
Multi-site rainfall
Weather generator
Spatial correlations
Kernel density estimation
Nonparametric
Low-frequency variability
Longer-term aggregated variables
MRS

A B S T R A C T

Many problems in hydrology and agricultural science require extensive records of rainfall from multiple locations. Temporal and/or spatial coverage of rainfall data is often limited, so that stochastic models may be required to generate long synthetic rainfall records. This study describes a multi-site rainfall simulator (MRS) to stochastically generate daily rainfall at multiple locations. The MRS is available as an open-source software package in the R statistical computing environment. The software includes statistical analysis and graphics functions, and can display statistics and graphs at multiple time scales, including from individual sites and areal averages. The MRS thus provides a detailed set of modelling functions to simulate and analyse daily rainfall. The capabilities of the package are demonstrated using 30 gauges located in Sydney, Australia, and the results show that the model preserves observed year-to-year variability, interannual persistence and various daily distributional and space–time dependence attributes.

1. Introduction

Many problems in hydrology and agricultural science require extensive records of rainfall from multiple locations (Mehrotra and Sharma, 2007a,b; Wilks, 1998). However, observational records are often short and only represent a single realization of the possible future patterns of rainfall, so that stochastic models are often used to augment the observational records. The synthetic rainfall sequences should accurately reflect the statistical characteristics of the historical rainfall record. Depending on the problem, these statistics can include annual total rainfall, seasonality, interannual persistence, occurrence probability and intermittency, intensity, extremes, and the dependence between sites.

The development of a multi-site stochastic rainfall model that preserves all of these statistics has proved to be extremely challenging. Most models either fail to account for, or poorly simulate, the observed low-frequency variability (e.g. the year-to-year dependence) in daily rainfall (Katz and Parlange, 1993, 1998; Wilks, 1999a,b) as well as the spatial dependence across multiple point locations (Mehrotra and Sharma, 2007a,b). The focus on Markovian dependence at the daily scale also limits the accurate reproduction of extended drought frequencies because only a few days' memory is retained at a time (Buishand, 1978; Guttorp, 1995; Racsko et al., 1991; Semenov and Porter, 1995). Failure to address these challenges can lead to poor outcomes for hydrological or agricultural activities in a region. The simulated streamflow, for example, might misrepresent drought risk, leading to suboptimal policies that will result in suboptimal catchment management outcomes (Mehrotra and Sharma, 2007a,b).

Several alternative methods for multi-site rainfall simulation have been proposed, including nonparametric nearest-neighbour resampling (Beersma and Buishand, 2003; Mehrotra et al., 2004), and parametric alternatives, such as multi-site chain-dependent processes (Wilks, 1998; Qian et al., 2002; Brissette et al., 2007; Mehrotra and Sharma, 2007b; Burton et al., 2013). Bearing in mind that the large-scale rainfall generating mechanisms might lead to similar rainfall patterns over a region (e.g. see Leonard et al., 2014), weather state based models (also referred to as hidden Markov chain models; see Hughes and Guttorp, 1994; Thyer and Kuczera, 2003a,b; Mehrotra et al., 2004) have also been

* Corresponding author. Tel.: +61 2 93851540; fax: +61 2 9385 6139.
E-mail address: raj.mehrotra@unsw.edu.au (R. Mehrotra).

http://dx.doi.org/10.1016/j.envsoft.2014.10.016
1364-8152/

developed. Multisite space–time dependence has been modelled using a censored power-transformed multivariate Gaussian distribution (Bardosy and Plate, 1992; Ailliot et al., 2009). Models based on a Neyman–Scott point process coupled with a spatial Poisson process of cell centers (Cowpertwait, 1995; Leonard et al., 2008) and fractal cascade model (Jothityangkoon et al., 2000) have been used to generate rain cell events in space and time.

Several methods have also been proposed to preserve low-frequency rainfall variability, including those that allow variations in the stochastic model parameters by conditioning on one or several atmospheric variables (Hughes and Guttorp, 1994; Hughes et al., 1999; Mehrotra et al., 2004; Katz and Parlange, 1993; Katz and Zheng, 1999; Wilks, 1989), and those that conditionally modify the model parameters on the basis of the values of some aggregated time scale predictor variables, such as antecedent rainfall state (wet or dry) at various levels of aggregation at gauged (Harrold et al., 2003a,b) and ungauged (Mehrotra et al., 2012) locations. The rainfall occurrence (Harrold et al., 2003a) was resampled from the historical record of rainfall occurrence, conditional to the current values of antecedent rainfall state, while the amount (Harrold et al., 2003b) was conditioned on the rainfall amount on the previous day and a 365 days wetness state index variable to closely reproduce the historical longer-term variability. A novel method for reconstructing both the observed spatial (inter-site) as well as temporal correlation statistics has been proposed by Clark et al. (2004a,b). The approach explains a large part of the observed interannual variability in the generated series without involving major changes in the basic structure of the rainfall generation models. Sharma and Mehrotra (2010) provide a comprehensive review of various rainfall generation approaches.

Mehrotra and Sharma (2007b) proposed a semi-parametric multi-site rainfall model referred to as the 'multi-site Markov model – kernel density estimator' (MMM-KDE), which is designed to preserve both the time and space dependent structures in rainfall simulations. Similar to other multi-site models, the MMM-KDE requires extensive pre-processing of observed information to estimate the model parameters. The model is effective, but is potentially challenging to implement due to the large number of steps in the algorithm, as well as the complexity of some of the mathematical concepts.

To represent many of these spatial and temporal features of rainfall variability, we develop and describe a software tool called the multi-site rainfall simulator (MRS), which runs the MMM-KDE algorithm and analyses daily rainfall at multiple locations. Once the MRS is installed, an interactive screen asks the user to enter some basic information required to run the model. The model-generated results can be visualised in the form of tables and graphs and can be saved for use in other applications at a later stage. A help file is also provided with the package. The capabilities and utilities of the tool have been demonstrated by applying it to data from 30 rain gauges around Sydney, Australia, and the results are presented in the form of plots and tables of observed and simulated at-site statistics and several spatio-temporal dependence attributes.

The paper is organised as follows. The model, its features and use are described in Section 2. In Section 3 the software architecture is described. Details of the application of the model and a comparison of the various results are presented in Section 4. We conclude the paper by presenting the summary in Section 5.

2. The multi-site rainfall simulator (MRS)

MRS is a multi-site Markov model that is designed to account for low-order dependence, by making use of (a) aggregated time scale predictor variables to preserve the longer-time scale dependence (i.e., low-frequency variability) and (b) spatially correlated random numbers to maintain the desired spatial correlations in the generated rainfall sequences. These are features unique to the MRS.

The rainfall simulation proceeds in two stages. In the first stage, rainfall occurrences are generated using a two-state, first order Markov model. At each time step, Markovian transition probabilities are modified on the basis of the aggregated wetness over pre-specified period(s) of time in the recent past. In the second stage, rainfall amounts on simulated wet days are generated using a nonparametric kernel density estimation approach assuming first order Markovian dependence. To begin with, models for occurrence and amounts are applied independently at each point location.

The spatial dependence in the rainfall occurrences and amount series are then induced by using spatially correlated random numbers in the generation process. This significantly simplifies the model structure. The seasonal transition is maintained by estimating the daily Markovian probabilities and correlations using a moving window (Rajagopalan and Lall, 1999; Mehrotra et al., 2006; Mehrotra and Sharma, 2007a,b), thereby avoiding the sharp transitions from one month to another. A brief description of the temporal and spatial aspects of the model is provided in the following sections; for more details refer to Mehrotra and Sharma (2007b, 2010).

2.1. Representation of temporal dependence

A single-site first order Markov model is defined as $P(O_t|O_{t-1})$ where O_t refers to the rainfall occurrences at time step t. Inclusion of additional continuous predictors $\mathbf{X}_t$ as conditioning variables modifies the first order conditional dependence to $P(O_t|O_{t-1},\mathbf{X}_t)$. Examples of predictors that can be used include large-scale atmospheric variables and/or antecedent rainfall at various levels of aggregation. Expanding the conditional expression and rearranging the terms leads to:

$$P(O_t = 1|O_{t-1} = i, \mathbf{X}_t) = \frac{P(O_t = 1, O_{t-1} = i)}{P(O_{t-1} = i)} \times \frac{f(\mathbf{X}_t|O_t = 1, O_{t-1} = i)}{f(\mathbf{X}_t|O_{t-1} = i)} \tag{1}$$

The first expression on the right of (1) defines the transition probabilities $P(O_t|O_{t-1}=i)$ of a first order Markov model (representing order-one dependence,p_1), whereas the second expression signifies the effect of including the additional predictor set $\mathbf{X}_t$ in the model. The second expression is approximated as a multivariate normal, which is likely to be a reasonable assumption when $\mathbf{X}_t$ represents the aggregated predictor variables (e.g. rainfall occurrences) over a period such as a year, or smoothly varying large-scale atmospheric variables. Under specific instances where the assumption of a multivariate normal may not be valid, it is possible to transform the data or use alternative probability distributions (e.g. see Mehrotra and Sharma, 2010).

The second expression of (1), when expanded as a multivariate normal, leads to the following simplification for $P(O_t|O_{t-1}=i,\mathbf{X}_t)$:

$$P(O_t=1|O_{t-1}=i,\mathbf{X}_t) = p_{1,i} \times \frac{f(\mathbf{X}_t|O_t=1,O_{t-1}=i)}{\left[f(\mathbf{X}_t|O_t=1,O_{t-1}=i)p_{1,i}\right] + \left[f(\mathbf{X}_t|O_t=0,O_{t-1}=i)p_{0,i}\right]} \tag{2a}$$

where $p_{1,i}$ is the baseline transition probability of the first order Markov model defined by $p_{1,i}=P(O_t=1|O_{t-1}=i)$, with p_{0i} being equal to 1 − p_{1i}; $\mu_{1,i}$ represents the mean vector