Available online at www.sciencedirect.com
ScienceDirect
Environmental Modelling & Software 23 (2008) 412–421
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

A Bayesian method for multi-site stochastic data generation: Dealing with non-concurrent and missing data, variable transformation and parameter uncertainty

Q.J. Wang*

CSIRO Land and Water, PO Box 56, Highett, Victoria 3190, Australia

Received 22 November 2006; accepted 26 April 2007
Available online 24 October 2007

Abstract

Stochastically generated stream flow and climatic data may be used as input to water resources simulation models for planning purposes. Dealing with non-concurrent and missing data, variable transformation and parameter uncertainty presents a significant challenge in the development of methods for stochastic data generation. In this paper, a Bayesian method is introduced for multi-site stochastic generation of annual stream flow and climatic data. A contemporaneous autoregressive lag-one model CAR(1) with the Box−Cox transformation is used to capture key statistical structure of multiple annual stream flow and climatic time series while keeping the number of model parameters to a minimum. The posterior joint distribution of the model parameters is formulated, allowing for inputs of historical data series that are not continuous or concurrent, thus avoiding the need to infill or truncate data records and maximising the value of available data. Parameter and uncertainty inference are solved numerically by using Markov Chain Monte Carlo simulations. Subsequent stochastic generation of data fully accounts for parameter uncertainty. In addition, a re-parameterization scheme is used to handle the problem of strong inter-parameter dependence from the Box−Cox transformation. The method was applied to the Melbourne Water supply system to demonstrate its computational feasibility.

Keywords: Stochastic data generation; Time series; Stream flow; Climatic variables; Parameter uncertainty; Bayesian method; Variable transformation; Non-concurrent and missing data

1. Introduction

To aid water resource system planning, simulation models may be used to evaluate system performance. The usefulness of simulation models is often limited by the lack of flow and climatic data series long enough to represent possible future scenarios. To overcome this limitation, stochastic models have been used to artificially generate sequences of flows and climatic variables. These stochastic models are to mimic observed statistical behaviours considered important to water resource systems (e.g. Thomas and Fiering, 1962; Matalas, 1967).

A thorough review of hydrological time series models can be found in Salas (1993). The most popular stochastic model for multiple annual flows and climatic variables is the Markov autoregressive lag-one model AR(1) (Matalas, 1967; McMahon and Srikanthan, 1982; Kuczera, 1986). One special family of the AR(1) model is the contemporaneous autoregressive lag-one model CAR(1) (Hipel and McLeod, 1994). Conceptually, the CAR(1) model assumes that different time series variables are statistically related to one another only at the same time. A variable at one time can be statistically related to the variable at a different time only within one time series. The CAR(1) model preserves the important statistical properties of cross-series correlation and within-series auto-correlation, while having far fewer parameters than the AR(1) model.

The Box−Cox transformation is often used to transform flows and climatic variables so that they become multinormal

* Tel.: +61 3 92526139.
E-mail address: QJ.Wang@csiro.au

1364-8152/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2007.04.013

and amenable to an AR(1) model (e.g. McMahon and Srikanthan, 1982; Kuczera, 1986; Thyer et al., 2002). In addition, if the variables are subject to some underlying changes, they need to be adjusted before a stationary AR(1) model can be assumed. For example, for forested catchments in the Melbourne Water supply system, Langford (1975) and Kuczera (1985) found that stream flow was reduced following major bushfires due to the conversion of largely mature mountain ash forest to an establishing regrowth forest. Kuczera (1986) introduced a bushfire and stream flow impact function to adjust the stream flow.

Historical data are most valuable in establishing a stochastic model and should be fully used. Nearly all the available methods for the calibration of multiple time series stochastic model require concurrent and consecutive data sets. To satisfy this requirement, two approaches are often employed. The first approach is to infill missing data, usually by some kind of regression. However, such a practice does not necessarily increase, sometimes may even corrupt, the information content of the original data (Salas, 1993). In addition, infilled data sets may render mathematical operations inoperable (Kuczera, 1987). The second approach is to truncate the data series so that only concurrent and consecutive data are used. Such an approach is a terrible waste of valuable data.

Kuczera (1987) introduced to the hydrology literature a useful method to estimate the AR(1) model from an incomplete data set. It involves the use of an iterative procedure to obtain a maximum likelihood solution. The method was formulated for the standard AR(1) model that does not require any prior normalisation of the model variables. Although not in the context of the AR(1) model, Wang (2001) explicitly acknowledged missing data in his formulation of a posterior distribution of model parameters, and allowed information transfer between correlated data series that had both concurrent and non-concurrent records by a joint inference of the model parameters. This Bayesian method can be applied to the AR(1) model, using the Kuczera (1987) formulation of the likelihood function. When the Box–Cox transformation is required to normalise the model variables to an AR(1) model, only minor extension to the Kuczera formulation of the likelihood function is necessary to include the effect of the transformation.

Another Bayesian approach to dealing with missing data is to treat missing data as unknown parameters, which are jointly inferred with model parameters (Gelman et al., 1995; Thyer and Kuczera, 2003a,b). This approach is straightforward in mathematical formulation but highly demanding in numerical computation during implementation. For instance, in the example of application to be described later in the paper, the use of this approach would mean that a total of 449 new parameters had to be introduced to represent missing data, to add to an already large number (116) of model parameters. This would place dramatically increased demand on statistical inference of the joint distribution of the parameters. For this reason, whenever mathematically possible the formulation of the posterior distribution of the model parameters should explicitly account for the effect of missing data, rather than treating missing data as unknown parameters.

A Bayesian approach enables the quantification of parameter uncertainty (e.g. Wang, 2001; Thyer and Kuczera, 2003a,b). In practice, there is considerable uncertainty with model parameters, even when all available historical data are efficiently used in the model establishment. Parameter uncertainty should be taken into account when a model is used to generate stochastic data series (Stedinger et al., 1985). The quantification of parameter uncertainty for multiple time series models presents a significant challenge. However, with recent advances in Bayesian computation through the use of Markov Chain Monte Carlo (MCMC) simulations, the Bayesian method has become a powerful technique for inferring complicated statistical models and dealing with parameter uncertainties (Gelman et al., 1995; Wang, 2001; Thyer and Kuczera, 2003a,b).

In this paper, a Bayesian formulation and solution of the statistical inference of a stochastic model for multi-site annual stream flow and climatic variables is introduced. As the method was developed for the Melbourne Water supply system, the Kuczera (1986) bushfire model is used to describe the impact of bushfires on flows. Annual flows are adjusted to conditions of mature mountain ash. The Box–Cox transformation is applied to the adjusted annual flows and to annual climatic variables. The transformed flow and climatic variables are assumed to follow a contemporaneous autoregressive lag-one model CAR(1). A Bayesian formulation of the posterior distribution of the model parameters allows all available historical data to be used without the need to infill missing data or truncate non-concurrent periods of data. Parameter and uncertainty inference are solved numerically by using MCMC simulations. Subsequent stochastic generation of data fully accounts for parameter uncertainty. The paper includes a brief demonstration of the method for the Melbourne Water supply system.

2. Model formulation

2.1. Impact of bushfires on flows

The Kuczera (1986) relationship between reduction in annual flow and age of mountain ash stand is used to account for the impact of bushfires. The reduction is with reference to annual flow if all mountain ash was mature. For a catchment that has a number of (age) stands of mountain ash, reduction in annual flow is given by:

$$f(\text{year}) = \sum_{s=1}^{S(\text{year})} A_s L_{\max} \delta_s \exp(1-\delta_s)$$
$$\delta_s = \max[0, K(\text{year} - \text{YEAR}_s - 2)] \quad (1)$$

where f(year) – reduction in annual flow, expressed in depth of water, in year 'year', s – the sth mountain ash stand, the total number of stands being S, A_s – fraction of catchment covered by the sth stand of mountain ash, YEAR_s – origin year of the sth ash stand, $L_{\max}$ – maximum reduction in annual flow following a fire, expressed in depth of water; $L_{\max} \geq 0$, K – a parameter defining that the maximum reduction in annual flow occurs $(2 + 1/K)$ years after the fire; K in years^{-1} and $K > 0$.