



Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Hydrology 296 (2004) 135–163

Journal
of
Hydrology

www.elsevier.com/locate/jhydrol

Retrospective analysis and forecasting of streamflows using a shifting level model

V. Fortin^{a,*}, L. Perreault^a, J.D. Salas^b

^a*Hydro-Québec Research Institute, 1800, boul. Lionel-Boulet, Varennes, QC J3X 1S1, Canada*

^b*Department of Civil Engineering, Colorado State University, Fort Collins, CO 80523, USA*

Received 27 May 2003; revised 7 February 2004; accepted 19 March 2004

Abstract

Shifting level models have been suggested in the literature since the late 1970's for stochastic simulation of streamflow data. Parameter estimation for these models has been generally based on the method of moments. While this estimation approach has been useful for simulation studies, some limitations are apparent. One is the difficulty of evaluating the uncertainty of the model parameters, and another one is that the proposed model is not amenable to forecasting because the underlying mean of the process, which changes with time, is not estimated. In this paper, we reformulate the original shifting level model to conform to the so-called Hidden Markov Chain models (HMMs). These models are increasingly used in applied statistics and techniques such as Monte-Carlo Markov chain, and in particular Gibbs sampling, are well suited for estimating the parameters of HMMs. We use Gibbs sampling in a Bayesian framework for parameter estimation and show the applicability of the reformulated shifting level model for detection of abrupt regime changes and forecasting of annual streamflow series. The procedure is illustrated using annual flows of the Senegal River in Africa.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Shifting-level; Hidden markov chain; Forecasting; Bayesian analysis; Gibbs sampling; Stochastic hydrology

1. Introduction

Stochastic modeling of hydrologic time series has been widely used for planning and management of water resources systems such as for reservoir sizing and forecasting the occurrence of future hydrologic events (Salas et al., 1980; Loucks et al., 1981; Bras and Rodrigues-Iturbe, 1985; Hipel and McLeod, 1994). For example, stochastic models are used to generate synthetic series of water supply that may

occur in the future which are then utilized for estimating the probability distribution of key decision parameters such as reservoir storage size. Likewise, stochastic models may be utilized for forecasting water supplies and water demands days, weeks, months, and years in advance. In turn, the forecasts are used in planning and testing operating rules, in estimating future power output of hydroelectric systems, and during real-time systems' operations.

A number of stochastic models has been considered in the literature for synthetic generation and forecasting of hydrological processes (Bras and Rodrigues-Iturbe, 1985; Salas, 1993; Hipel and McLeod, 1994).

* Corresponding author.

E-mail addresses: vinfort@ireq.ca (V. Fortin), perreault.luc@ireq.ca (L. Perreault), jsalas@engr.colostate.edu (J.D. Salas).

Hydrologic processes such as annual streamflow and precipitation in some cases may be well represented by stationary linear models such as autoregressive (AR) and autoregressive moving average (ARMA) models. These models are usually capable of preserving the historical annual statistics, such as the mean, variance, skewness, and covariance. However, Hurst (1951) analyzed a large number of geophysical records and discovered that certain climatic and hydrologic records showed some evidence of long-term persistence. Even though ARMA models may preserve long-term related statistics such as storage and drought related statistics (O'Connell, 1971; Salas et al., 1979), alternative models such as the Fractional Gaussian Noise (Mandelbrot and Wallis, 1969), Broken Line (Rodriguez-Iturbe et al., 1972; Curry and Bras, 1978), Shifting Level (Klemes, 1974; Potter, 1976; Salas and Boes, 1980), and FARMA (Montanari et al., 1997) have been proposed in literature.

Many recent atmospheric and oceanic studies have shown that hydroclimatic processes exhibit abrupt shifting patterns (Yonetani and Gordon, 2001; Rial et al., 2004; Christiansen, submitted; Sveinsson et al., 2003; Schwing et al., 2003). The Shifting Level (SL) model is capable of simulating shifting streamflow patterns explicitly. It has been successfully applied for simulating net basin supplies of the Great Lakes System (Rassam et al., 1992) and annual streamflow series in the Quebec region in Canada (Salas, 2000). However, no forecasting procedure based on the SL model is available. In this paper, we develop the mathematical framework and estimation procedures for forecasting applications of the SL model.

Considerable research has been carried out in hydrology on developing mathematical tools and approaches for short- and long-term streamflow forecasting. Forecasting of hydrologic processes has been developed using similar approaches as for simulation, although many models and techniques are unique for either simulation or forecasting (Valdes et al., 2002). In any case, these models must take into account the uncertainty in the model parameters that arises because of insufficient historical data of the relevant processes under consideration. Bayesian analysis (Bernardo and Smith, 1994) offers a framework in which it is straightforward to issue probabilistic forecasts that take into account the uncertainty in the model parameters. Bayesian methods are often

regarded as more difficult to implement than classical approaches to parameter estimation, such as the method of moments and the method of maximum likelihood. However, for a large class of models, efficient numerical methods are now available for use in Bayesian analysis.

We show in this paper that the SL model belongs to the class of hidden Markov-chain models (HMMs). Monte-Carlo Markov chain (MCMC) methods, and in particular Gibbs sampling, are well suited for estimating the parameters of HMMs in a Bayesian framework. The Gibbs sampler was originally developed by Geman and Geman (1984) in the context of image restoration but Gelfand and Smith (1990) showed its applicability to general Bayesian statistical analysis. MCMC methods and HMMs models have been active areas of basic and applied research in statistics for many years, and have been gaining popularity in hydrology. Zucchini and Guttorp (1991) suggested HMMs for modeling rainfall while Kuczera and Parent (1998) introduced MCMC methods in the context of calibrating conceptual watershed models. Lu and Berliner (1999) used Gibbs sampling to estimate the parameters of a HMM developed for flow forecasting at a daily time step, Barreto and de Andrade (2000) applied MCMC methods to forecast monthly streamflows with an autoregressive model, Perreault et al. (2000a,b) used Gibbs sampling for retrospective analysis of annual flow series with change-point models, and Thyer and Kuczera (2000, 2003a,b) proposed a two-state HMM for modeling long-term persistence in annual precipitation series. This paper also builds upon ideas presented by Fortin et al. (2002) where Bayesian estimates were proposed for the parameters of the SL model when the long-term mean and variance are known.

By definition, HMMs include a hidden Markov process that is not observed directly. In comparison with other HMMs, which have been proposed to date in hydrology, the SL model has the particularity that the realization of its hidden Markov process (which corresponds to the local mean of the observed process) is a continuous variable that takes on values on the real line, instead of a finite number of values (typically two or three). As a by-product of the Bayesian estimation procedure presented in this paper, together with estimates of the parameters of the model, one can also estimate the hidden Markov

process. Thus, the method can also serve to characterize the local mean of the observed process. This analysis can be used to define dry and wet spells and to detect abrupt changes in the local mean of the process, in some sense generalizing the well-known segmentation method of Hubert et al. (1989), recently revisited by Kehagias (2004) using the HMM framework.

This paper is organized as follows. In Section 2, we propose a slightly different parameterization of the SL model of Salas and Boes (1980) to show that it belongs to the class of HMMs. In Section 3, we show how the parameters of the model can be estimated in a Bayesian framework using Gibbs sampling. In Section 4, we illustrate the estimation procedure using annual flows of the Senegal River and show how the Gibbs samples can be used for forecasting. Section 5 offers a general discussion on the results obtained and other potential applications of the suggested approach.

2. Representing the shifting level model as a hidden Markov model

Salas and Boes (1980) introduced a four-parameter SL model, which can be used to simulate time series showing sudden shifts in the mean. The observations $\mathbf{x} = (x_1, x_2, \dots, x_N)$ are considered to be realizations of the sum of two independent stochastic processes:

$$x_t = m_t + \varepsilon_t, m_t \perp \varepsilon_t, \\ p(\varepsilon_t | \sigma_\varepsilon^2) = \mathcal{N}(\varepsilon_t | 0, \sigma_\varepsilon^2), \varepsilon_t \text{ i.i.d.} \quad (1)$$

where \perp denotes independence between two random variables, $p(\cdot)$ denotes the probability density function (p.d.f.) of a random variable and $\mathcal{N}(\cdot | \mu, \sigma^2)$ denotes normal distribution with mean μ and variance σ^2 . While the term ε_t in (1) is a white noise with variance σ_ε^2 , the term m_t corresponds to the unobserved (or latent) mean level m_t corresponding to each observation x_t , i.e. $m_t = E[x_t | m_t]$. The mean level is also normally distributed with mean μ and variance σ_μ^2 , but it remains constant for epochs, which duration follows a geometric distribution. More precisely, we have:

$$p(m_t | \mu, \sigma_\mu^2) = \mathcal{N}(m_t | \mu, \sigma_\mu^2) \quad (2)$$

$$\Pr[m_t = m_{t+1} = \dots = m_{t+k} \\ \neq m_{t+k+1} | m_1, \dots, m_{t-1}, m_{t+k+2}, \dots, m_N, \eta] \\ = \eta \cdot (1 - \eta)^k$$

where η is the parameter of the geometric distribution, which corresponds to the probability that a transition to a new mean level occurs at any time point t . The SL model thus has four parameters: $\eta, \sigma_\varepsilon^2, \sigma_\mu^2$ and μ . It can be shown that the process $\{m_t\}$ is a Markov chain. Indeed, from the properties of the geometric distribution, the probability that an epoch ends at time t is constant, i.e. independent of t and of the value taken by the process at any other point in time. This can be seen by setting $k = 0$ in (2). Hence, if an epoch ends at time t then m_{t+1} is independent of the previous values $\mathbf{m}_t = (m_1, m_2, \dots, m_t)$, and if an epoch does not end at time t then $m_{t+1} = m_t$. It follows that the probability distribution of m_{t+1} given $\mathbf{m}_t = (m_1, m_2, \dots, m_t)$ depends only on the value of m_t (and on the parameters of the SL model), which makes the process $\{m_t\}$ Markovian.

It is readily seen that the expectation of x_t is equal to μ , and because of the independence between the two processes, the variance of x_t corresponds to the sum of the variance of the two processes:

$$E[x_t] = \mu, \quad \text{var}[x_t] = \sigma_\mu^2 + \sigma_\varepsilon^2 \quad (3)$$

Fig. 1 presents a sample of size $N = 100$ simulated using the SL process with parameters $\mu = 0, \sigma_\mu^2 = \sigma_\varepsilon^2 = 1/2$ (hence $\text{var}[x_t] = 1$) and $\eta = 0.2$. The latent variables m_1, m_2, \dots, m_N are plotted in bold, and the resulting observations $\mathbf{x} = (x_1, x_2, \dots, x_N)$ are represented by the finer line. The time series seems almost completely random before the 60th observation, though there are 12 shifts in the mean level during that period. Then a shift of considerable

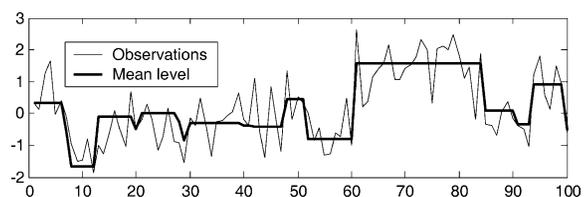


Fig. 1. $N = 100$ observations simulated from a SL process in which $\mu = 0, \sigma_\mu^2 = \sigma_\varepsilon^2 = 1/2$ and $\eta = 0.2$.

amplitude occurs, which lasts for 24 observations, followed by four more shifts during the last 16 years. Hence, the pseudo-periodicity caused by the shifting mean creates a times series, which appears more ‘predictable’ during certain epochs, and more ‘random’ at other times, a feature often associated with climatic and hydroclimatic time series.

2.1. A different parameterization of the SL model

The fact that two parameters of the SL model must add up to the total variance of the process $\{x_t\}$ leads to difficulties when estimating the parameters of the model in a Bayesian framework, essentially because the prior distribution of the two variance parameters σ_ε^2 and σ_μ^2 must then be modeled jointly. Indeed, if anything is known about the scale of the process, $\sigma_\mu^2 + \sigma_\varepsilon^2$, then prior information on σ_μ^2 sheds some light on the value of σ_ε^2 . Prior dependence between σ_ε^2 and σ_μ^2 then leads to difficulties both in the process of prior information elicitation and in the application of the numerical method proposed in this paper for approximating the posterior distribution. For this reason, it is useful to introduce a slightly different parameterization of the SL model. Let $\sigma^2 = \text{var}[x_t]$ and define ω as the ratio σ_μ^2/σ^2 . We will assume that $0 < \sigma_\mu^2 < \sigma^2$ so that $0 < \omega < 1$. We can then replace the parameter σ_μ^2 by $\omega \cdot \sigma^2$ and the parameter σ_ε^2 by $(1 - \omega) \cdot \sigma^2$, still leading to a four-parameter parameterization of the SL model, i.e. the mean μ and variance σ^2 of its stationary distribution, the parameter η , which corresponds to the probability of observing a transition from a mean level to another at any instant t , and the parameter ω , which can be interpreted as the proportion of the total variance of the process $\{x_t\}$, which is explained by the shifts in the mean level, represented by process $\{m_t\}$. Denote the vector of these four parameters by $\theta = (\mu, \sigma^2, \omega, \eta)$.

Again, for estimation purposes it is useful to associate with each time point t a Bernoulli variable z_t which takes the value 1 if a new epoch begins at time $t + 1$ and 0 otherwise. It follows from the properties of the geometric distribution that the random variables $\{z_t\}$ are independent and identically distributed (i.i.d.), with probability of success $\Pr[z_t = 1] = \Pr[m_t \neq m_{t+1}] = \eta$ (note that the probability that the mean level stays the same when a new epoch begins is zero since $\sigma_\mu^2 > 0$). The variables $\{z_t\}$ do not

introduce additional latent variables into the model, since each z_t is a deterministic function of m_t and m_{t+1} . Hence, knowledge of m_t and m_{t+1} implies perfect knowledge of z_t . However, when m_t and m_{t+1} are not both known, z_t remains a random variable.

2.2. Autocorrelation induced by a shifting mean

Salas and Boes (1980) have shown that the autocorrelation function for the SL model is similar to that of an ARMA(1,1) process. Indeed, even if observations are independent within each epoch, the random shifts in the mean level create dependence between observations. It can be shown that the autocorrelation ρ_k between x_t and x_{t+k} is a function of the parameters η and ω :

$$\rho_k = \omega \cdot (1 - \eta)^k \quad (4)$$

Proof: By definition, the autocorrelation of order k is the ratio of the covariance between x_t and x_{t+k} to the variance of the process: $\rho_k = \text{cov}(x_t, x_{t+k})/\sigma^2 = \text{cov}(m_t + \varepsilon_t, m_{t+k} + \varepsilon_{t+k})/\sigma^2$. Covariance being invariant with respect to white noise: $\text{cov}(m_t + \varepsilon_t, m_{t+k} + \varepsilon_{t+k}) = \text{cov}(m_t, m_{t+k}) = E[m_t \cdot m_{t+k}] - \mu^2$. Furthermore, m_t and m_{t+k} are equal if no transition have occurred between time t and time $t + k$, and are independent otherwise. From the properties of the geometric distribution, the probability that no shifts occurs between time t and $t + k$ is given by $\gamma = (1 - \eta)^k$. Hence, $E[m_t m_{t+k}] = E[m_t E[m_{t+k} | m_t]] = E[m_t (m_t \gamma + \mu \cdot (1 - \gamma))]$. It follows that $E[m_t \cdot m_{t+k}] = \gamma \cdot \sigma_\mu^2 + \mu^2 = \gamma \cdot \omega \cdot \sigma^2 + \mu^2$ and finally $\rho_k = \omega \cdot \gamma$. ■

Therefore, persistence in observations from a SL model decreases exponentially with time in the same manner as the linear ARMA(1,1) process.

2.3. Joint probability distribution of the observations and the latent variables

To show that the SL model belongs to the class of HMMs and to perform the needed Bayesian analysis it is necessary to obtain the joint probability distribution of the observations and latent variables. From Eq. (1), it is straightforward to show that the probability distribution of x_t , given m_t is normal, centered at m_t

with variance $(1 - \omega) \cdot \sigma^2$:

$$p(x_t | m_t, \boldsymbol{\theta}) = p(x_t | m_t, \sigma^2, \omega) = \mathcal{N}(x_t | m_t, (1 - \omega) \cdot \sigma^2) \tag{5}$$

The variable z_t has a Bernoulli distribution defined by its probability of success: $\Pr[z_t = 0 | \eta] = 1 - \eta$ and $\Pr[z_t = 1 | \eta] = \eta$. As z_t is a discrete variable, there exists no continuous function $p(z_t | \eta)$ such that $\Pr[z_t \leq k | \eta] = \int_{-\infty}^k p(z_t | \eta) dz_t$. However, to simplify the notation in the remaining of the paper, it will be useful to model the uncertainty on z_t using an integrable p.d.f. $p(z_t | \eta)$. This can be done through the use of the Dirac delta function $\delta(\cdot)$, which verifies the following properties:

$$\delta(x - x_0) = 0 \text{ if } x \neq x_0, \int_a^b f(x) \delta(x - x_0) dx = \begin{cases} f(x_0) & \text{if } a < x_0 < b \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where f is a function continuous at $x = x_0$. It was introduced by Dirac (1958) as a mathematical artefact useful to represent singularities in a continuous space, and is hence very useful to deal with probability spaces comprised of both continuous and discrete variables. Notice that by letting $f(x) = 1$, we can show that the integral of a Dirac delta function is equal to one:

$$\int \delta(x - x_0) dx = 1 \tag{7}$$

An interesting property of the Dirac delta function is that it is the derivative of the Heaviside step function $H(x - x_0)$:

$$\delta(x - x_0) = H'(x - x_0) \tag{8}$$

$$H(x - x_0) = \begin{cases} 0 & \text{if } x < x_0 \\ 1 & \text{otherwise} \end{cases}$$

This is useful since the cumulative distribution of a discrete random variable z which takes real values k_1, k_2, \dots, k_J with probabilities p_1, p_2, \dots, p_J can be written as a weighted sum of Heaviside functions:

$$F(z) = \sum_{j=1}^J p_j \cdot H(z - k_j) \tag{9}$$

Hence, the p.d.f. of a discrete random variable can be expressed by the integrable

function $f(z) = \sum_{j=1}^J p_j \cdot \delta(z - k_j)$. It follows that the p.d.f of z_t can be written as:

$$p(z_t | \boldsymbol{\theta}) = p(z_t | \eta) = \eta \cdot \delta(z_t - 1) + (1 - \eta) \cdot \delta(z_t) \tag{10}$$

The value of the latent variable m_{t+1} depends both on the value of z_t and m_t : either $m_{t+1} = m_t$ if $z_t = 0$, or m_{t+1} is drawn from a normal distribution if $z_t = 1$. This can be written as:

$$p(m_{t+1} | m_1, \dots, m_t, z_1, \dots, z_N, \boldsymbol{\theta}) = p(m_{t+1} | m_t, z_t, \mu, \sigma^2, \omega) = (1 - z_t) \cdot \delta(m_{t+1} - m_t) + z_t \cdot \mathcal{N}(m_{t+1} | \mu, \omega \cdot \sigma^2) \tag{11}$$

where $\delta(\cdot)$ is once again the Dirac delta function, needed here to model the distribution of m_{t+1} as either a discrete distribution with a probability mass of one at m_t or a normal distribution centered on μ .

For mathematical convenience, the first latent variable m_1 can be considered to have a similar conditional probability distribution which would depend on an unknown initial condition m_0 as well as upon a Bernoulli variable z_0 which indicates whether or not $m_1 \neq m_0$. One can show by induction that the stationary distribution $p(m_t | \boldsymbol{\theta}) = \mathcal{N}(m_t | \mu, \omega \cdot \sigma^2)$ if:

$$p(m_0 | \boldsymbol{\theta}) = \mathcal{N}(m_0 | \mu, \omega \cdot \sigma^2) \tag{12}$$

Thus, the stationary distribution of m_t is a natural choice for the distribution of the initial condition m_0 . For estimation and forecasting purposes, it will also be useful to estimate the latent variable m_{N+1} . Let $\mathbf{m} = (m_0, m_1, \dots, m_N, m_{N+1})$ and $\mathbf{z} = (z_0, z_1, z_2, \dots, z_N)$. Combining (5) to (12), we can now obtain the joint distribution of the observations and of the latent variables, given the parameters of the model, which would correspond to the likelihood function of the model if we could observe the latent variables. First, note that since the error terms ε_t are i.i.d., the distribution of \mathbf{x} given the latent variables and the parameters is given by the product of their marginal distributions:

$$p(\mathbf{x} | \mathbf{m}, \boldsymbol{\theta}) = p(\mathbf{x} | \mathbf{m}, \sigma^2, \omega) = \prod_{t=1}^N p(x_t | m_t, \sigma^2, \omega) \tag{13}$$

Note also that since \mathbf{z} is a deterministic function of \mathbf{m} , $p(\mathbf{x} | \mathbf{m}, \boldsymbol{\theta}) = p(\mathbf{x} | \mathbf{m}, \mathbf{z}, \boldsymbol{\theta})$. Similarly, since the Bernoulli variables z_t are i.i.d., the distribution of \mathbf{z}

given the parameters is also simply the product of their marginal distributions:

$$p(\mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{z}|\boldsymbol{\eta}) = \prod_{t=0}^N p(z_t|\boldsymbol{\eta}) \quad (14)$$

By definition of conditional probability, the joint probability of the sequence $\mathbf{m} = (m_0, m_1, \dots, m_N, m_{N+1})$ given \mathbf{z} can be factorized as:

$$p(\mathbf{m}|\mathbf{z}, \boldsymbol{\theta}) = p(m_0|\mathbf{z}, \boldsymbol{\theta}) \cdot \prod_{t=0}^N p(m_{t+1}|m_0, m_1, \dots, m_t, \mathbf{z}, \boldsymbol{\theta}) \quad (15)$$

However, since $p(m_{t+1}|m_0, \dots, m_t, \mathbf{z}, \boldsymbol{\theta}) = p(m_{t+1}|m_t, z_t, \boldsymbol{\mu}, \sigma^2, \omega)$, we can write:

$$\begin{aligned} p(\mathbf{m}|\mathbf{z}, \boldsymbol{\theta}) &= p(\mathbf{m}|\mathbf{z}, \boldsymbol{\mu}, \sigma^2, \omega) \\ &= p(m_0|\boldsymbol{\mu}, \sigma^2, \omega) \cdot \prod_{t=0}^N p(m_{t+1}|m_t, z_t, \boldsymbol{\mu}, \sigma^2, \omega) \end{aligned} \quad (16)$$

Combining (13), (14) and (16), the joint distribution of all the observations and latent variables, given the parameters, is thus obtained as:

$$\begin{aligned} p(\mathbf{x}, \mathbf{m}, \mathbf{z}|\boldsymbol{\theta}) &= p(\mathbf{x}|\mathbf{m}, \sigma^2, \omega) \cdot p(\mathbf{m}|\mathbf{z}, \boldsymbol{\mu}, \sigma^2, \omega) \cdot p(\mathbf{z}|\boldsymbol{\eta}) \\ &= \prod_{t=1}^N p(x_t|m_t, \sigma^2, \omega) \cdot p(m_0) \\ &\quad \times \prod_{t=0}^N \{p(m_{t+1}|m_t, z_t, \boldsymbol{\mu}, \sigma^2, \omega) \cdot p(z_t|\boldsymbol{\eta})\} \end{aligned} \quad (17)$$

2.4. Hidden Markov models

Eq. (17) defines a Hidden Markov Model of order one (HMM, see Bengio, 1999). In an HMM of order

one, a single state variable, here $q_t = (m_t, z_t)$, summarizes all the relevant past information of the underlying process: it would enable forecasting the next state, $q_{t+1} = (m_{t+1}, z_{t+1})$ and the next observation, x_{t+1} . However, the state variable q_t is not observed directly. The joint distribution of an HMM is specified in terms of:

- the initial state probability distribution, here $p(q_0|\boldsymbol{\theta}) = p(m_0|\boldsymbol{\mu}, \sigma^2, \omega)$;
- the transition probability distribution, here $p(q_{t+1}|q_t, \boldsymbol{\theta}) = p(m_{t+1}|m_t, z_t, \boldsymbol{\mu}, \sigma^2, \omega) \cdot p(z_{t+1}|\boldsymbol{\eta})$;
- the emission probability distribution, here $p(x_t|q_t, \boldsymbol{\theta}) = p(x_t|m_t, \sigma^2, \omega)$.

HMMs are often represented graphically using a Bayesian network, also known as a *directed acyclic graph* (DAG), see Spiegelhalter et al. (1996), which illustrates the relationships between variables in the model (observations, parameters and latent variables). The nodes represent the variables and the edges between nodes represent the presence of a direct relationship between the corresponding variables. Nodes may be represented in two ways: either as a circle, denoting that the value of the corresponding variable is unknown and thus subject to estimation, or by a square in which case the value of that variable is known.

Fig. 2 presents a DAG corresponding to the SL model. It can be seen that the model has a hierarchical structure in which, for every time t , the first level represents the (possible) transition to a new mean; then, the second level represents the process which generates the new mean, given the previous one; and finally, the last level stands for the generation of the data, given the mean level. This model formalizes the belief that the mean level m_t depends on the value of the Bernoulli variable z_{t-1} and on the value of

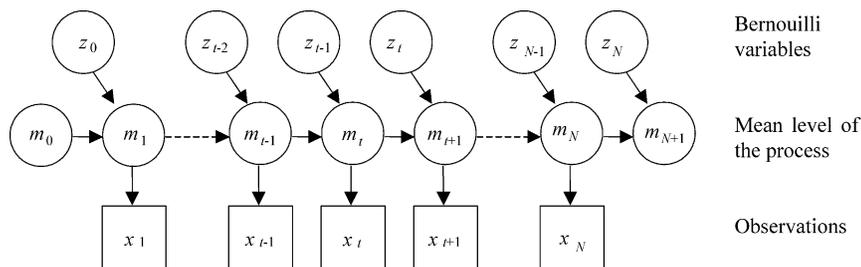


Fig. 2. Bayesian network representing the relationship between the latent variables and the observations in a SL model.

the previous mean level m_{t-1} . Similarly, the observation x_t depends on the value of m_t but is conditionally independent of the values of m_{t-1} and z_{t-1} . This conditional independence is represented in the graph by the absence of an edge between nodes m_{t-1} and x_t , and between z_{t-1} and x_t . Recognizing that the SL model is a HMM enables one to take advantage of estimation methods proposed for the parameters and latent variables of such models, and in particular Bayesian methods.

3. Bayesian estimation of the parameters

The first step in a Bayesian analysis is to set up a *full probability model*. That is, in addition to modeling the observable quantities \mathbf{x} , a prior distribution $p(\boldsymbol{\theta})$ is assumed for the parameters of the SL model. In the Bayesian perspective, specifying a prior distribution for the parameters is an integral part of the modeling task, with all hypotheses that modeling involves. Bayes' rule then provides the posterior distribution of the parameters given the data $p(\boldsymbol{\theta}|\mathbf{x})$, which is proportional to the joint distribution $p(\mathbf{x}, \boldsymbol{\theta})$, itself equal to the product of the likelihood function $p(\mathbf{x}|\boldsymbol{\theta})$ and of the prior distribution $p(\boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \propto p(\mathbf{x}, \boldsymbol{\theta}) \quad (18)$$

Just as the prior distribution reflects beliefs about the parameters prior to experimentation, the posterior distribution reflects the updated beliefs after observing the sample data. In the Bayesian framework, all statistical inference about the unknown parameters is based on the posterior distribution.

3.1. Likelihood function and prior distribution

Note that (17) does not correspond to the likelihood function, since the variables \mathbf{m} and \mathbf{z} are not observed. To obtain $p(\mathbf{x}|\boldsymbol{\theta})$, we need to integrate out the latent variables \mathbf{m} :

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{m}, \mathbf{z}(\mathbf{m})|\boldsymbol{\theta}) d\mathbf{m} \quad (19)$$

To establish the prior distribution of the parameters, we will assume that knowledge about the parameters

μ and σ^2 , which only characterize the location and scale of the stationary distribution of the observations, does not help to define the persistence of the phenomenon, which is characterized by the parameters ω and η . This implies that $p(\boldsymbol{\theta}) = p(\mu, \sigma^2) \cdot p(\omega, \eta)$. Furthermore, we will assume that ω and η are independent, i.e. $p(\omega, \eta) = p(\omega) \cdot p(\eta)$. Recall that η defines the mean length of epochs during which the local mean stays constant, and ω sets the level to which this shifting mean influences the observations. In a same region, different levels of observed persistence in annual streamflows could be explained by similar shifts in the regional climate, represented by the vector \mathbf{z} and characterized by the parameter η , which however affect basins to different degrees, characterized by the parameter ω . Thus, the parameter η defines the regional (or climatological) effect whereas the parameter ω defines the local effect of these regional variations. Assuming independence between these two parameters is equivalent to assuming that knowledge about the regional climate variations is not useful to assess to which degree a given basin will be influenced by these variations.

The parametric distributions of each parameter will also be chosen conveniently, mainly to facilitate the implementation of the estimation procedure. Given that the parameters η and ω are bounded between 0 and 1, we will assume that they are beta distributed, i.e.:

$$p(\omega, \eta) = p(\omega) \cdot p(\eta) = \mathcal{B}(\eta|s_\eta, \tau_\eta) \cdot \mathcal{B}(\omega|s_\omega, \tau_\omega) \quad (20)$$

where $\mathcal{B}(\theta|s, \tau) \propto \theta^{s\tau-1} \cdot (1-\theta)^{s-s\tau-1}$ corresponds to the p.d.f. of the beta distribution with hyperparameters $s > 0$ and $0 < \tau < 1$. We prefer this less common parameterization of the Beta distribution, since the hyperparameters s and τ are more easily interpreted. Indeed, it may be shown that τ is the expected value of θ whereas s can be interpreted as the weight given to the information modeled by the prior distribution. Smaller values of s lead to a more diffuse prior. Note that the uniform prior corresponds to the particular case $s = 2$ and $\tau = 1/2$, whereas Jeffrey's prior corresponds to the particular case $s = 1$ and $\tau = 1/2$. Both of these prior distributions are frequently used in situations where prior knowledge on the parameter θ is scarce.

For the location and scale parameters of the (normal) stationary distribution of the observations, we will assume a normal-inverse gamma prior,

with conditional dependence between μ and σ^2 , i.e.

$$p(\mu, \sigma^2) = p(\mu|\sigma^2) \cdot p(\sigma^2) \\ = \mathcal{N}(\mu|\nu, \kappa \cdot \sigma^2) \cdot \mathcal{G}^{-1}(\sigma^2|\alpha, \beta) \quad (21)$$

where $\mathcal{G}^{-1}(\sigma^2|\alpha, \beta) \propto (\sigma^2)^{-(\alpha+1)} \exp(-\beta/\sigma^2)$ corresponds to the p.d.f. of the inverse gamma distribution. Therefore, the joint prior distribution of the parameters is given by:

$$p(\theta) = p(\mu|\sigma^2) \cdot p(\sigma^2) \cdot p(\omega) \cdot p(\eta) \\ = \mathcal{N}(\mu|\nu, \kappa \cdot \sigma^2) \cdot \mathcal{G}^{-1}(\sigma^2|\alpha, \beta) \\ \times \mathcal{B}(\eta|s_\eta, \tau_\eta) \cdot \mathcal{B}(\omega|s_\omega, \tau_\omega) \quad (22)$$

Adding to Fig. 2, the nodes corresponding to these additional components and the edges that relate the parameters to the latent variables and observations, we obtain the Bayesian network

representation of the SL model for $t = 1, 2, \dots, N$ (Fig. 3). Note that observed data x_t and hyperparameters $(\nu, \kappa, \alpha, \beta, s_\omega, \tau_\omega, s_\eta, \tau_\eta)$ are represented by square nodes since they are known. The DAG of the SL model would be slightly different for $t = 0$ and $t = N + 1$. For $t = 0$, there is no predecessor state, and no observed value, while for $t = N + 1$, there is no successor state, and no observed value.

3.2. Gibbs sampling

The integration operation plays a fundamental role in Bayesian analysis, whether it is for calculating the normalizing constant in (18) or for evaluating the marginal posterior distribution of each parameters to make inference about these unknown quantities. For example, to evaluate the posterior marginal density $p(\eta|\mathbf{x})$, the parameters μ, σ^2 and ω must be integrated

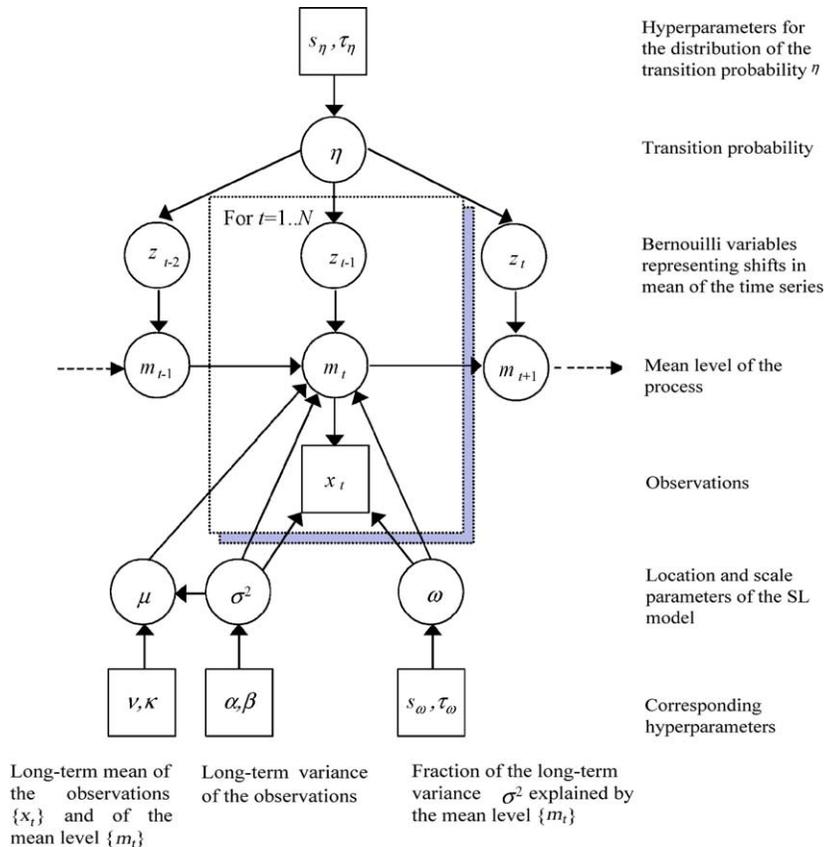


Fig. 3. Bayesian network representing the relationship between the hyperparameters, parameters, latent variables and the observations in a SL model.

out of (18), which is already an integral with respect to \mathbf{m} (cf. Eq.(19)). Explicit evaluation of such integrals is generally not possible. However, the Gibbs sampler, a tool particularly suitable for HMMs models, offers a straightforward solution to such problems. The idea of Gibbs sampling is to simulate from the so-called complete conditional distribution (CCD) of each unknown quantity (parameters and latent variables), which is obtained by conditioning on the data but also on the values of the remaining unknown quantities, which in turn are obtained from the previous iteration. Under mild conditions, a Gibbs sampler generates a Markov chain whose stationary distribution is precisely the posterior distribution of interest (Gelfand and Smith, 1990). The output of this Markov chain, once it has passed its transient stage, can be used to estimate parameters and functions of these parameters, such as moments and marginal densities.

For example, given a parameter space $\theta = (\theta_1, \dots, \theta_j)$ and the CCD $p(\theta_j | \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_J)$ for every parameter of the model, Table 1 presents a general algorithm for the Gibbs sampler.

Different implementations of the Gibbs sampler are possible, for example one can choose to draw the parameters in a certain order, or regroup certain parameters and draw from their joint CCD. For the SL model, it was essential to sample from the joint distribution of $q_t = (m_t, z_t)$ as the distribution of \mathbf{z} given \mathbf{m} is a Dirac distribution.

After a sufficiently large number of iterations R , the sampler has converged and the following simulated values can then be used for estimating different features of the posterior joint distribution, for example moments and marginal densities. Full details on theoretical properties of the algorithm and of the Gibbs samples obtained can be found in Gelfand and Smith (1990).

Note that Gibbs sampling is only appropriate when it is reasonably simple to compute and sample from

the CCDs. Otherwise, other sampling techniques such as the Metropolis-Hasting algorithm may be much simpler to implement (Robert and Casella, 1999).

3.3. The complete conditional distributions

The conditional independence structure of the SL model (Fig. 3) allows an easy evaluation of the CCD. Denoting the CCD of a parameter θ_j by $p(\theta_j | -)$, Bayes' rule implies that $p(\theta_j | -)$ is proportional to the joint distribution of the observations, latent variables and parameters:

$$p(\theta_j | -) = p(\theta_j | \mathbf{x}, \mathbf{m}, \mathbf{z}, \theta_{(j)}) = \frac{p(\mathbf{x}, \mathbf{m}, \mathbf{z}, \theta)}{\int p(\mathbf{x}, \mathbf{m}, \mathbf{z}, \theta) d\theta_j} = \frac{p(\mathbf{x}, \mathbf{m}, \mathbf{z} | \theta) p(\theta)}{\int p(\mathbf{x}, \mathbf{m}, \mathbf{z} | \theta) p(\theta) d\theta_j} \quad (23)$$

where $\theta_{(j)}$ stands for a vector containing all element of θ except θ_j . The normalizing constant in (23) is much easier to compute than in (18), in part because it involves a single integral, but also because all the probability distributions in (23), which do not involve θ_j can be taken out of the integral and cancelled out, being constant with respect to θ_j . We can now obtain the CCDs for the parameters $\mu, \sigma^2, \eta, \omega$ and in the same manner the CCDs for the latent variables m_{N+1} and q_t . Let \mathbf{M} be the set of time steps at which new epochs begin, including the initial mean level, i.e. $\mathbf{M} = \{0\} \cup \{t : z_{t-1} = 1, 1 \leq t \leq N + 1\}$. Let also $r = \sum_{t=0}^N z_t$ and $\bar{m} = 1/(r + 1) \sum_{t \in \mathbf{M}} m_t$. Proof of the following results is presented in Appendix A:

$$p(\mu | -) = p(\mu | r, \bar{m}, \sigma^2, \omega) = \mathcal{N}\left(\mu \left| \frac{\kappa \cdot (r + 1) \cdot \bar{m} + \omega \cdot v}{\kappa \cdot (r + 1) + \omega}, \frac{\kappa \cdot \omega \cdot \sigma^2}{\kappa \cdot (r + 1) + \omega} \right.\right) \quad (24)$$

$$p(\sigma^2 | -) = p(\sigma^2 | \mathbf{x}, \mathbf{m}, \mathbf{z}, \omega) = \mathcal{G}^{-1}\left(\sigma^2 \left| \frac{N+r}{2} + 1 + \alpha, \frac{\sum_{t=1}^N (x_t - m_t)^2}{2(1 - \omega)} + \frac{\sum_{t \in \mathbf{M}} (m_t - \mu)^2}{2\omega} + \frac{(m - v)^2}{2\kappa} + \beta \right.\right) \quad (25)$$

Table 1
General Gibbs sampling algorithm

Specify arbitrary starting values $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_J^{(0)})$
 For $i = 1$ to R , where R is the number of Gibbs samples to be drawn
 For $j = 1$ to J , where J is the number of parameters of the model
 Draw $\theta_j^{(i)}$ from $p(\theta_j | \theta_1^{(i-1)}, \dots, \theta_{j-1}^{(i-1)}, \theta_{j+1}^{(i-1)}, \dots, \theta_J^{(i-1)})$
 Next j
 Next i

$$p(\eta|-) = p(\eta|r, N) = \mathcal{B}(\eta|s_\eta\tau_\eta + r, s_\eta - s_\eta\tau_\eta + N - r + 1) \quad (26)$$

$$p(\omega|-) = p(\omega|\mathbf{x}, \mathbf{m}, \mathbf{z}, \mu, \sigma^2) \propto \prod_{t=1}^N \mathcal{N}(x_t|m_t, (1-\omega)\cdot\sigma^2) \times \prod_{t \in \mathbf{M}} \mathcal{N}(m_t|\mu, \omega\cdot\sigma^2) \cdot \mathcal{B}(\omega|s_\omega, \tau_\omega) \quad (27)$$

$$p(m_{N+1}|-) = p(m_{N+1}|q_N, \mu, \sigma^2) = (1 - z_N) \cdot \delta(m_{N+1} - m_N) + z_N \cdot \mathcal{N}(m_{N+1}|\mu, \sigma^2) \quad (28)$$

$$p(q_t|-) = p(z_t|-) \cdot p(m_t|\mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}) \quad (29)$$

where $\mathbf{a}_{(t)}$ denotes a vector \mathbf{a} from which element a_t has been removed. The CCD $p(m_t|\mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta})$ corresponds to the distribution of m_t conditional on all other parameters and latent variables, with the exception of z_t .

$$p(z_t|-) = p(z_t|m_t, m_{t+1}) = \begin{cases} \delta(0) & \text{if } m_t = m_{t+1} \\ \delta(1) & \text{if } m_t \neq m_{t+1} \end{cases} \quad (30)$$

$$p(m_0|\mathbf{x}, \mathbf{m}_{(0)}, \mathbf{z}_{(0)}, \boldsymbol{\theta}) = p(m_0|m_1, \eta, \mu, \sigma^2) = (1 - \eta) \cdot \delta(m_1 - m_0) + \eta \cdot \mathcal{N}(m_0|\mu, \sigma^2) \quad (31)$$

$$p(m_t|\mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}) = p(m_t|x_t, q_{t-1}, m_{t+1}, \mu, \sigma^2, \eta, \omega), \quad 1 \leq t \leq N = (1 - z_{t-1}) \cdot \delta(m_t - m_{t-1}) + z_{t-1} \cdot [(1 - \eta^*) \cdot \delta(m_{t+1} - m_t) + \eta^* \cdot \mathcal{N}(m_t|(1-\omega)\cdot\mu + \omega\cdot x_t, \omega(1-\omega)\cdot\sigma^2)] \quad (32)$$

where

$$\eta^* = \left[1 + \frac{1-\eta}{\eta} \times \frac{\mathcal{N}(x_t|m_{t+1}, (1-\omega)\cdot\sigma^2)}{\mathcal{N}(x_t|\mu, \sigma^2)} \right]^{-1} \quad (33)$$

3.4. Sampling from the complete conditional distributions

Standard simulation methods are available to sample from the CCD of μ, σ^2 and η . Since the CCD of m_{N+1} (Eq. (28)) is either a Dirac or a normal distribution (depending on the value of z_N), sampling from this distribution simply means setting m_{N+1} equal to m_N if $z_N = 0$, and drawing from a normal distribution if $z_N = 1$. Sampling from the CCD of q_t can be done by first sampling m_t from $p(m_t|\mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta})$ (Eq. (32)) and then by setting $z_t = 0$ if $m_t = m_{t+1}$ and $z_t = 1$ otherwise. Sampling from $p(m_t|\mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta})$ is relatively simple. First, if $z_{t-1} = 0$, then m_t is set to the value of m_{t-1} . If on the contrary $z_{t-1} = 1$, then $p(m_t|\mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta})$ is simply a mixture of a Dirac and a normal distribution. Sampling from this mixture implies first drawing a random number u from a uniform distribution on $[0,1]$ to choose between the Dirac and the normal distribution depending on whether or not $u > \eta^*$, and then drawing a random number from the selected distribution.

Sampling from the CCD of ω is a bit more difficult since the normalizing constant cannot be computed analytically. However, a simple inversion method works nicely for this univariate distribution: a random number u uniformly distributed between 0 and 1 is drawn and the value ω for which $F(\omega|-) = \int_{-\infty}^{\omega} p(\omega|-)d\omega = u$ is sought. To estimate the cumulative distribution we use a very simple but robust approach: $p(\omega|-)$ is evaluated for values of ω between 0 and 1 by steps of $\Delta\omega = 10^{-3}$, $F(\omega|-)$ is approximated by $\hat{F}(\omega|-) = \sum_{i=1}^{\omega/\Delta\omega} p(i\cdot\Delta\omega|-) / \sum_{i=1}^{1/\Delta\omega} p(i\cdot\Delta\omega|-)$ for values of ω which are multiples of $\Delta\omega$, and is linearly interpolated in between. This works quite well because the parameter ω is bounded between 0 and 1, but speed and accuracy gains could be obtained by using a more efficient algorithm. However, even for moderate sample sizes, simulation of the latent variables takes most of the computing time.

Table 2 details a first implementation of the Gibbs sampler for the SL model. In practice, we found that reversing the order in which the parameters are sampled at each iteration seems to speed up convergence to the stationary distribution. This approach is known as the reversible Gibbs sampler

Table 2
Gibbs sampling algorithm for the SL model

A. Specify arbitrary starting values $\mu^{(0)}, \sigma^{2(0)}, \omega^{(0)}, \eta^{(0)}, m_{N+1}^{(0)}, q_t^{(0)} = (m_t^{(0)}, z_t^{(0)}), t = 0, \dots, N + 1$

B. For $i = 1$ to R , where R is the number of Gibbs samples to be drawn

1. Let $M^{(i-1)} = \{0\} \cup \{t : z_{t-1}^{(i-1)} = 1, 1 \leq t \leq N + 1\}, r^{(i-1)} = \sum_{t=0}^N z_t^{(i-1)}$ and $\bar{m}^{(i-1)} = \frac{1}{r^{(i-1)} + 1} \sum_{t \in M^{(i-1)}} m_t^{(i-1)}$
2. Draw $u^{(i)}$ from a uniform distribution on $(0, 1)$
3. Find $\omega^{(i)}$ such that $\tilde{F}(\omega^{(i)} | \mathbf{x}, \mathbf{m}^{(i-1)}, \mathbf{z}^{(i-1)}, \mu^{(i-1)}, \sigma^{2(i-1)}) = u^{(i)}$
4. Draw $\eta^{(i)}$ from $p(\eta | r^{(i-1)}, N)$
5. Draw $\sigma^{2(i)}$ from $p(\sigma^2 | \mathbf{x}, \mathbf{m}^{(i-1)}, \mathbf{z}^{(i-1)}, \omega^{(i)})$
6. Draw $\mu^{(i)}$ from $p(\mu | r^{(i-1)}, \bar{m}^{(i-1)}, \sigma^{2(i)}, \omega^{(i)})$
7. Draw $m_0^{(i)}$ from $p(m_0 | m_1^{(i-1)}, \eta^{(i)}, \mu^{(i)}, \sigma^{2(i)})$
8. Let draw $z_0^{(i)} = 0$ if $m_0^{(i)} = m_1^{(i-1)}$ and $z_0^{(i)} = 1$ otherwise
9. For $t = 1$ to N
 - (a) Draw $m_t^{(i)}$ from $p(m_t | x_t, q_{t-1}^{(i)}, m_{t+1}^{(i-1)}, \mu^{(i)}, \sigma^{2(i)}, \eta^{(i)}, \omega^{(i)})$
 - (b) Let $z_t^{(i)} = 0$ if $m_t^{(i)} = m_{t+1}^{(i-1)}$ and $z_t^{(i)} = 1$ otherwise
10. Next t
11. Draw $m_{N+1}^{(i)}$ from $p(m_{N+1} | q_N^{(i)}, \mu^{(i)}, \sigma^{2(i)})$

C. Next i

(Robert and Casella, 1999). To initialize the Gibbs sampler, we generate the initial parameter vector $\theta^{(0)}$ by sampling from the prior distribution $p(\theta), z$ from $p(\mathbf{z} | \eta^{(0)}), m_0$ from $\mathcal{N}(m_0 | \mu^{(0)}, \omega^{(0)}, \sigma^{2(0)})$, and then each m_t in turn from $p(m_t | m_{t-1}^{(0)}, z_{t-1}^{(0)}, \mu^{(0)}, \sigma^{2(0)}, \omega^{(0)})$.

3.5. Checking for convergence

One of the main difficulties with Gibbs sampling is to evaluate how many iterations must be performed to obtain accurate results. Different numerical procedures are available to check for convergence (Robert, 1998). However, it is always a good idea to make sure that results obtained from a number of runs having different starting points are similar. To ensure this, we apply two different procedures. We first compare the posterior expected values of the parameters obtained by Gibbs sampling for different runs. For each run and each parameter θ , we compute the arithmetic mean over the second half of the Gibbs sample:

$$\bar{\theta}^{(R)} = \frac{1}{\lfloor R/2 \rfloor} \sum_{i=\lfloor R/2 \rfloor + 1}^R \theta^{(i)} \tag{34}$$

If the Gibbs sampler has converged and if a sufficient number of iterations have been performed afterwards, then this estimate of the posterior

expectation of the parameter θ should be approximately equal for independent runs. The coefficient of variation $CV(\bar{\theta}^{(R)})$ gives a useful indication of the accuracy of the posterior expectation.

We also apply the non-parametric method proposed by Brooks and Gelman (1998). For each parameter and each run $j = 1, \dots, J$, we compute the interval length Δ_j of the empirical $100(1 - \alpha)\%$ interval, say 80%, from the second half of the Gibbs sample. We then compute the interval length Δ for the same coverage probability from the super-sample obtained by mixing the results from all runs. If the Gibbs sampler has converged, the values Δ_j should be only slightly smaller than Δ on average. In practice, convergence is assumed when the ratio $\hat{R}_{BG} = \Delta / \frac{1}{J} \sum_{j=1}^J \Delta_j$ is smaller than 1.2 for every parameter.

It is also possible to take advantage of the Gibbs samples drawn from each independent Markov chain to improve the accuracy. Indeed, the posterior distribution of any function of the parameters can be better estimated from the super-sample obtained by mixing the Gibbs samples from each independent run. For example, since the results obtained for J different chains are independent, the coefficient of variation of the posterior expectation for this super-sample can be estimated by $(CV(\bar{\theta}^{(R)})) / \sqrt{J - 1}$.

3.6. Elicitation of prior information

Before using the proposed Bayesian estimation procedure, the values of the hyperparameters must be set by the hydrologist on the basis of his expertise prior to analyzing the data set. This can be a complex and lengthy process. An interesting method consists in obtaining prior guesses for quantiles of the stationary distribution $p(x_t)$ and of the conditional distribution $p(x_t | x_{t-1})$, which can then be translated into values for the hyperparameters. Another approach, which we use in this paper, consists in using observations at a neighboring site to estimate the hyperparameters. However, we will only use regionalization techniques to estimate the parameters of the normal-inverse gamma prior distribution, which characterize the mean and variance of the stationary distribution of the process, and assume uniform prior distributions for the parameters controlling the persistence of the process, η and ω .

3.7. Illustration of the estimation procedure using synthetic data

To better understand the Bayesian estimation procedure we applied the procedure to the simulated sample obtained in Section 2 (Fig. 1). Recall that this sample was simulated using as parameter values $\eta = 0.2$, $\omega = 1/2$, $\mu = 0$ and $\sigma^2 = 1$. We will suppose that the mean and variance of the process are known (respectively, 0 and 1) and assume uniform prior distributions for η and ω . It is straightforward to adapt the Gibbs sampling algorithm to cases where some parameters are known a priori: these parameters are simply not drawn randomly from their CCD but simply kept equal to their known value for all iterations.

Based on the last 10,000 iterations of a 20,000 iterations run, Fig. 4 shows the posterior distributions of the parameters η and ω . While there is significant uncertainty left on the parameter values, the posterior expectation of each parameter (respectively $E[\eta|\mathbf{x}] = 0.18$ and $E[\omega|\mathbf{x}] = 0.54$) is close to the true parameter value.

Fig. 5a shows the expected value of the latent variables $\{m_t, t = 1, 2, \dots, N\}, E[m_t|\mathbf{x}]$, together with the observations and the exact values of m_t . It can be seen that the mean level is relatively well estimated on average, despite the uncertainty remaining on the parameters of the model. Fig. 5b presents the expected value of z_{t-1} for each time step $E[z_{t-1}|\mathbf{x}]$. This corresponds to the posterior probability that a new epoch begins at time t . It can be seen that only a few transitions are detected with high probability: out of the 17 transitions present in the time series, only at

four time steps does the probability of a transition reach the 0.5 mark. It is not surprising to observe that these correspond to shifts of larger magnitude. Hence, while the estimation procedure is capable of estimating the transition probability of the process and its mean level as a function of time, it is incapable of detecting all shifts present in a noisy data set. However, the proposed estimation procedure still serves as a coherent method for detecting sudden shifts in a time series when multiple shifts are present, and it suffices for forecasting purposes to be able to estimate correctly the parameters of the model and the current mean level of the process.

4. Application

The foregoing modeling approach will be illustrated using annual streamflows of the Senegal River. This river is located in Western Africa and has a drainage area of approximately $S = 218,000 \text{ km}^2$ at the Bakel station. Fig. 6 shows the time series of the annual flow data for the period 1903–1986, which has been downloaded from <http://www.cig.ensmp.fr/~hubert>. This time series shows a long-term persistence. When fitting an autoregressive AP(p) model to this time series, the Akaike Information Criterion (Akaike, 1974) suggests an AR(2) model, which has the same number of parameters as the SL model. Using the segmentation procedure introduced by Hubert et al. (1989), the persistence in this time series could be explained by sudden shifts in the local mean and variance of the process (Hubert, 2000).

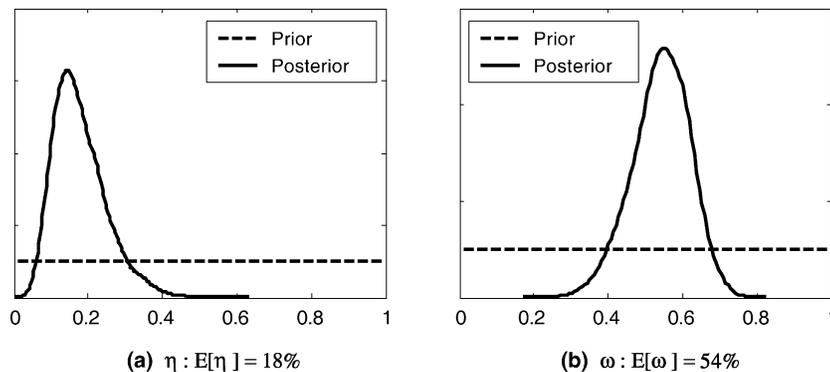


Fig. 4. Posterior distribution for the parameters of the SL model applied to synthetic data: (a) for η in which $E[\eta|\mathbf{x}] = 0.18$ (the true parameter value being 0.2), (b) for ω in which $E[\omega|\mathbf{x}] = 0.54$ (the true parameter value being 0.5).

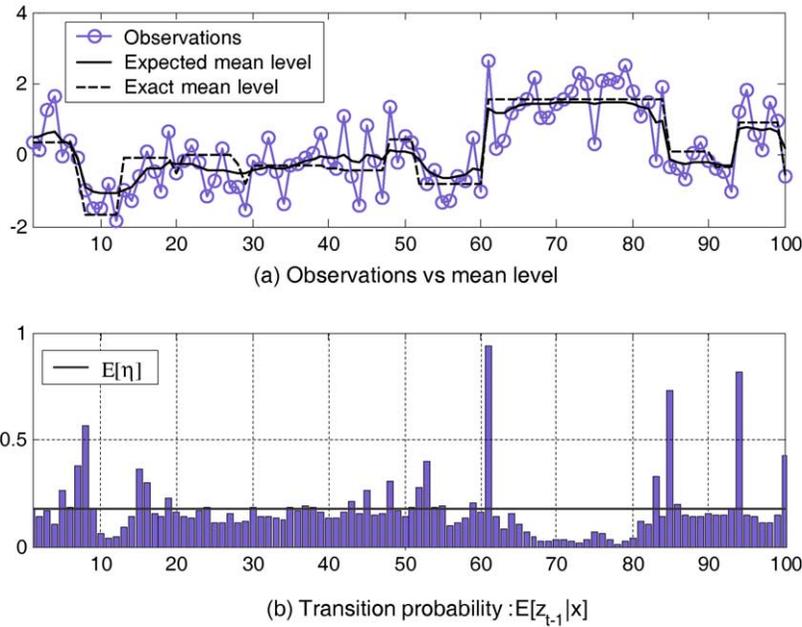


Fig. 5. (a) Expected value of the mean level and (b) transition probability estimated from synthetic data.

At the 95% confidence level, this segmentation procedure identified four change-points in the times series, with new epochs beginning in 1922, 1937, 1950 and 1968.

4.1. Prior information modelling

In addition to the Senegal River data we use the annual streamflows of the Niger River for the period

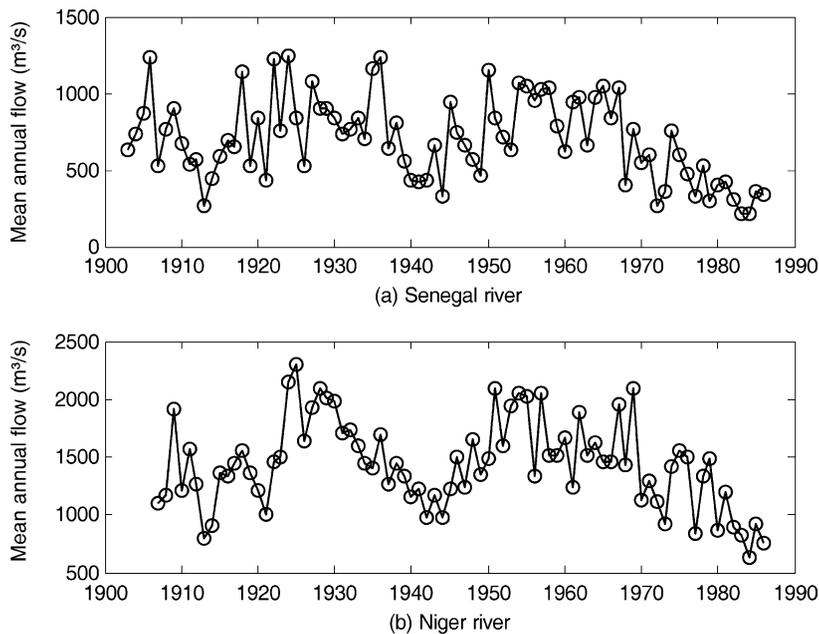


Fig. 6. Time series of annual flows for the Senegal River at Bakel (1903–1986) and the Niger River (1907–1986).

1907–1986 to obtain a prior distribution for the parameters of the stationary distribution, assuming that the coefficient of variation $CV[x] = \sigma[x]/E[x]$ and the specific discharge $Q_s = E[x]/A$ are similar for both sites (where A denotes the area of the watershed). The values of these statistics for the Niger River are $CV[x] \cong 0.3$ and $Q_s \cong 3 \text{ l/s/km}^2$. Hence, our prior estimates of the mean and variance of the annual flows of the Senegal River are $E(\mu) = 700 \text{ m}^3/\text{s}$ and $E(\sigma^2) = (200 \text{ m}^3/\text{s})^2$, respectively. Furthermore, from experience we estimate that $E(\mu)$ is within 30% of the population value, two times out of three, and that $E(\sigma^2)$ is also within 100% of the population value, two times out of three (recall that for a normal distribution, the probability associated with an interval of plus or minus one standard deviation about the mean is approximately 2/3). Hence, the standard deviation of the prior distribution should be about 30% of the prior expectation for μ , and 100% for σ^2 , which means that we would like to let $CV[\mu] = 0.3$ and $CV[\sigma^2] = 1$, approximately.

From the properties of the normal-inverse gamma prior (Bernardo and Smith, 1994), it is known that $E[\sigma^2] = \beta/(\alpha - 1)$, $E[\mu] = E[\mu|\sigma^2] = \nu$, $\text{var}[\sigma^2] = \beta^2/[(\alpha - 1)^2(\alpha - 2)]$ and $\text{var}[\mu|\sigma^2] = \kappa\sigma^2$, so that $CV[\mu] = E[CV[\mu|\sigma^2]] = E[\sqrt{\kappa}\cdot\sigma/\nu] \cong \sqrt{\kappa}\cdot CV[x]$ and $CV[\sigma^2] = 1/\sqrt{\alpha - 2}$. Hence, reasonable values for the parameters of the normal-inverse gamma distribution are: $\alpha = 2 + 1/CV[\sigma^2]^2 = 3$, $\beta = (\alpha - 1) \cdot E[\sigma^2] = 80,000$, $\nu = E[\mu] = 700$ and $\kappa = (CV[\mu]/CV[x])^2 = 1$. The parameters of the prior distributions of η and ω are chosen to obtain uniform prior distributions: $s_\eta = s_\omega = 2$, and $\tau_\eta = \tau_\omega = 1/2$.

4.2. Inference for the parameters

Using Gibbs sampling, we ran 10 Markov chain in parallels, drawing from each chain 20,000 Gibbs samples, and keeping in each case the results from the last 10,000 iterations. Fig. 7a–d show the prior and posterior densities for each parameter of the SL model based on 100,000 values. The posterior density were

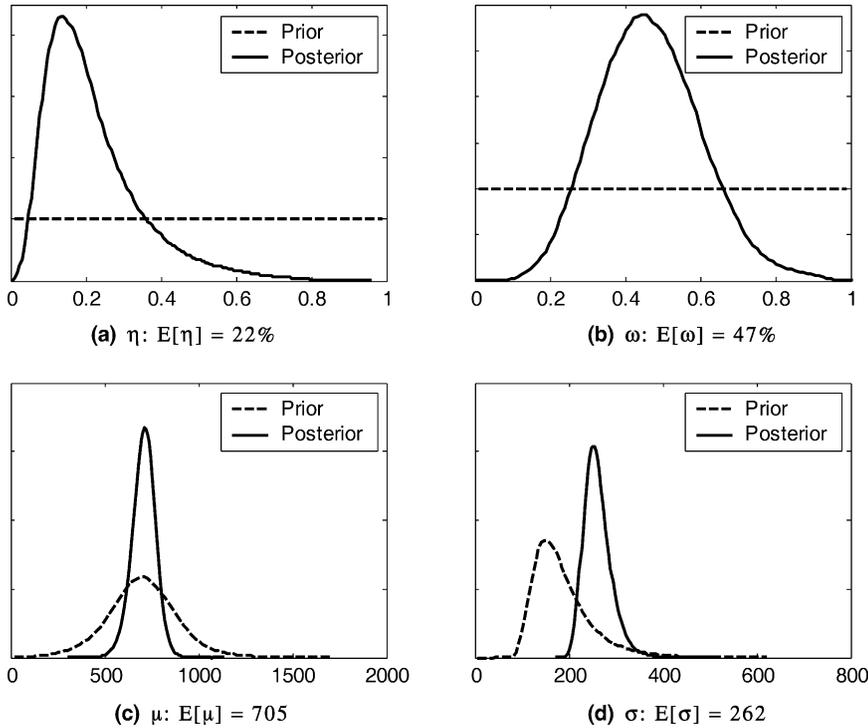


Fig. 7. Posterior distribution for the parameters of the SL model applied to the Senegal River annual flows: (a) for η in which $E[\eta|x] = 0.22$, (b) for ω in which $E[\omega|x] = 0.47$, (c) for μ in which $E[\mu|x] = 705$, and (d) for σ in which $E[\sigma|x] = 262$.

Table 3
Correlation between the parameters of the SL model for the Senegal River

	η	ω	μ	σ^2
η	1	0.44	0.20	-0.10
ω	0.44	1	0.10	0.48
μ	0.20	0.10	1	-0.05
σ^2	-0.10	0.48	-0.05	1

estimated by kernel smoothing with a normal kernel function. The expected values of the parameters are indicated in the diagrams. The expected value of the transition probability η is about 0.2 and the variance explained by the shifting mean process is about 50%, but there remains a large amount of uncertainty on these parameters. For example, a 95% credibility interval for η gives [0.06;0.56] and a 95% credibility interval for ω gives [0.22;0.76]. Note that the posterior distribution of the standard deviation σ is presented instead of the variance σ^2 as its interpretation is generally easier.

The relatively large uncertainty on the posterior marginal distributions of η and ω can be explained in part by the fact that the parameters are not

independent. The correlation between each pair of parameters can be easily estimated from the Gibbs samples (Table 3). It is seen that the largest correlations are between η and ω , which account for the persistence of the time series, and between ω and σ^2 , which are the variance parameters. These large correlations may delay convergence of the Gibbs sampler, but are correctly taken into account by the Gibbs sampling approach to retrospective and predictive analysis.

4.3. Convergence analysis

To evaluate the accuracy of the results obtained, we compared the posterior expectation of the parameters as the number of iterations increases for each of the 10 independent runs, using Eq. (34). Fig. 8 presents the coefficient of variation of the posterior expectation estimated from these ten independent Markov chains for each parameter, as the number of iterations is increased from 1 to 20,000 (of course, the total number of iterations performed is ten times higher). It can be seen that the coefficient of variation is about one order of magnitude larger for the transition probability than for the long-term mean. Still, after

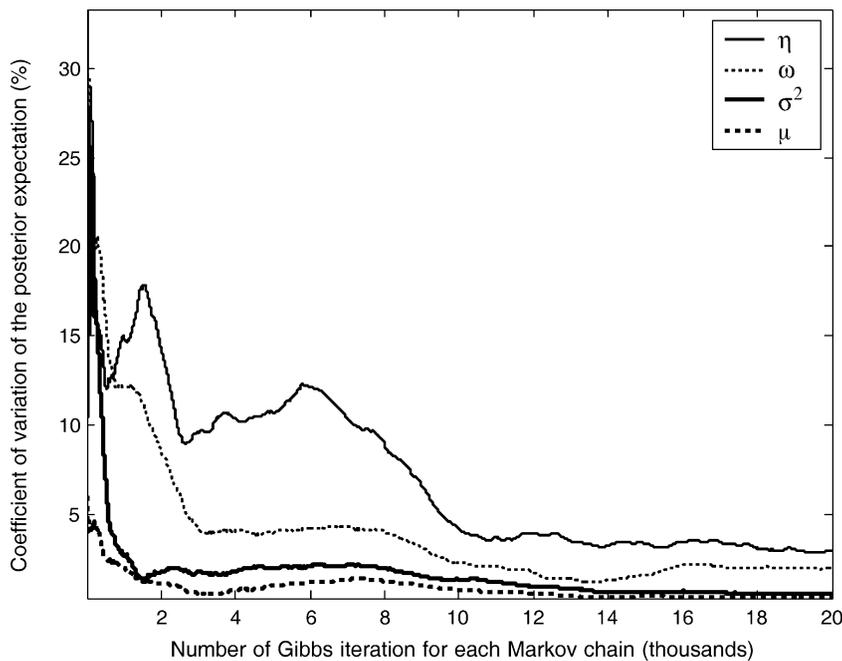


Fig. 8. Coefficient of variation of the posterior expectation of each parameter, estimated from ten independent Markov chains.

10,000 iterations for each run, the CV is less than 5% for all parameters (relative to its expected value, the most variable parameter is η , but with an expected value of 0.22, a 5% coefficient of variation corresponds to a standard deviation of 0.01). Therefore, we consider the results to be sufficiently accurate to support the analyses presented in this paper.

The Brooks and Gelman (1998) procedure leads to the same conclusion: the value of the statistic \hat{R}_{BG} computed for a 80% credible interval is smaller than 1.02 for all parameters after 20,000 iterations (the maximum acceptable value being 1.2). In fact, even for the most variable parameter, η , \hat{R}_{BG} is smaller than 1.2 after only 3500 iterations.

4.4. Retrospective analysis

Fig. 9a shows the expected value of the latent variables $\{m_t, t = 1, 2, \dots, N\}$, $E[m_t|\mathbf{x}]$, together with the observations, while Fig. 9b presents the expected value of z_{t-1} for each year, $E[z_{t-1}|\mathbf{x}]$. This corresponds to the posterior probability that a new epoch begins on year t . These two graphs can be very useful for identifying sudden shifts in the time series and corresponding periods of low or high flows. For example, it is likely that a period of low flows began in 1968 (with a probability of about 0.7). In fact, according to these results, the four years likely to correspond to the beginning of a new epoch are 1922,

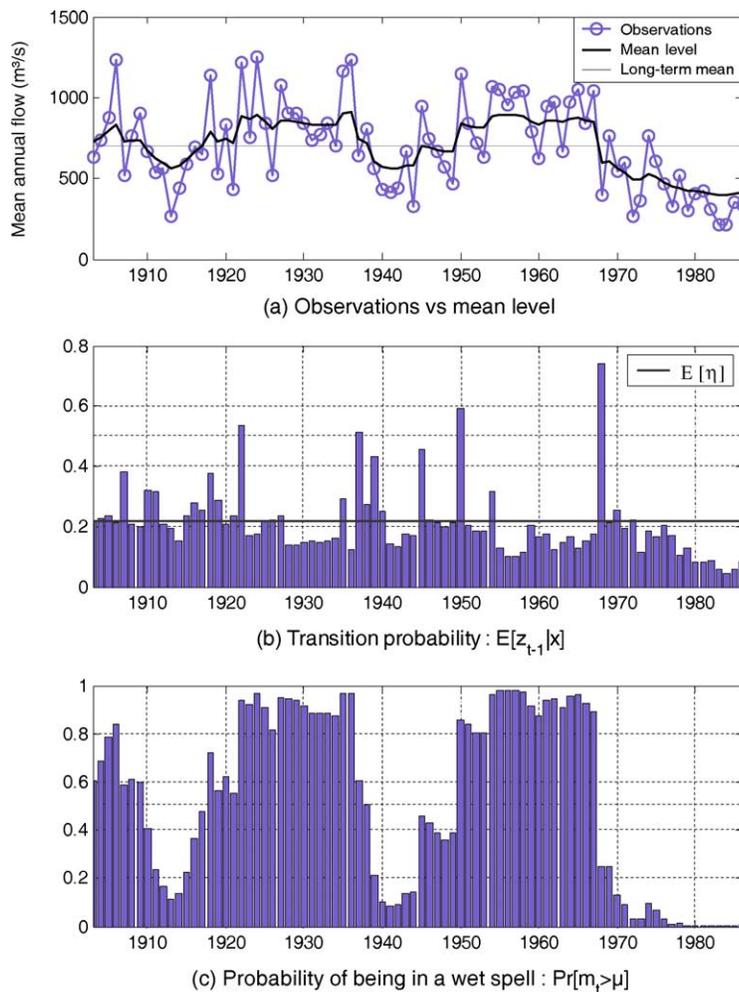


Fig. 9. (a) Expected value of the mean level, (b) transition probability and (c) probability of wet spell.

1937, 1950 and 1968. Indeed, only for these four years is $E[z_{t-1}|\mathbf{x}]$ higher than 0.5. Note that these epochs correspond exactly to those identified at the 95% confidence level by the segmentation procedure of Hubert (2000). Hence, the proposed procedure generalizes Hubert's segmentation procedure insofar as it gives similar results while providing an assessment of the uncertainty on the parameters and latent variables of the model. However, the SL model is more constrained since it imposes a geometric distribution for the duration of epochs and assumes the variance to be constant across epochs.

From the properties of the geometric distribution, the expected duration of epochs is $E[1/\eta|\mathbf{x}] = 6.4$ years, whereas the segmentation procedure of Hubert (2000) identifies only five segments in an 84 year time series, with an average length of 17 years. However, as was shown in Section 3.7 with synthetic data, many shifts of small magnitude can go undetected, which can explain the discrepancy between the expected duration of epochs and the number of segments identified.

The SL model also offers a new approach for defining dry and wet spells. Indeed, if we look at the plot of the expected mean level $E[m_t|\mathbf{x}]$ over time and compare it with the long-term mean (Fig. 9a), we can define dry and wet spells quite naturally by classifying years as wet or dry depending if $E[m_t|\mathbf{x}] > E[\mu]$ or $E[m_t|\mathbf{x}] < E[\mu]$. This classification defines six periods: 1903–1909 (wet), 1910–1917 (dry), 1918–1938 (wet), 1939–1949 (dry), 1950–1967 (wet) and 1968–1986 (dry). Of course, there is also uncertainty associated with this classification, which can be measured for example by the probability of a given year t being wet, $\Pr[m_t > \mu]$ (Fig. 9c). For example, it can be observed that the probability of being in a wet spell is close to zero after 1980.

4.5. Adequacy of the SL model to represent the observed persistence

While the SL model may be appropriate to represent sudden shifts in the regional climate, the shifts cause autocorrelation as shown in Section 2.2. Thus, it should be used with caution when one suspects that a significant part of the autocorrelation of the process under consideration is caused by storage effects in the watershed. When both abrupt shifts and storage affect the autocorrelation of

the underlying process, the shifting mean autoregressive (SMAR) model proposed by Sveinsson and Salas (2001) in which observations follow an autoregressive AR(1) process between shifts in the mean, may be a better model although no Bayesian estimation and forecasting methodology has been developed for this model at the moment.

In the case of the Senegal River, Hubert (2000) suggested that it is in fact the regional climate which is changing, and that the observed autocorrelation in the time series is induced by the sudden shifts. This is supported by the fact that the observed lag-1 autocorrelation between any two shifts identified by Hubert (2000) is lower than the lag-1 autocorrelation computed on the complete time series ($\hat{\rho}_1 = 0.4$). In fact, during the periods 1903–1921, 1922–1936, 1937–1949 and 1950–1967 the observed lag-1 autocorrelation is even negative. It is only for the last period (1968–1986) that the observed lag-1 autocorrelation is close to 0.4. But the strongest evidence in favor of a climatic explanation for the persistence in this data set is the spatial coherence of the abrupt shifts, the neighboring Niger River showing similar shifts at similar times.

Furthermore, note that from the Gibbs samples the posterior expectation for the lag-1 autocorrelation, $E[\rho_1] = E[\omega \cdot (1 - \eta)]$, is estimated at 0.4, the standard deviation of the posterior distribution of ρ_1 being about 0.1. Hence, the SL model is capable of representing the observed persistence as measured by $\hat{\rho}_1$.

Finally, a classical residual analysis was performed on $\hat{\varepsilon}_t = E[m_t] - x_t$ to ensure that no signal remained in the residuals. This time series passes Kendall's trend test at the 5% level with a p -value of 0.7, Wald-Wolfowitz's independence test at the 5% level with a p -value of 0.1, the Chi-square goodness-of-fit test for normality (with 11 classes) at the 5% level with a p -value of 0.5, and the empirical moments test for normality (based on the sample skewness and kurtosis), with p -values of 0.5 for skewness and 0.98 for kurtosis.

4.6. Probabilistic forecasting

From a Bayesian point of view, a forecast of the next observation y_{N+1} should be based on the posterior distribution of these future observations given \mathbf{x} ,

which is also known as the predictive distribution $p(y_{N+1}|\mathbf{x})$. While this distribution is unknown, from Eq. (5) the distribution of y_{N+1} given the parameters and latent variables is:

$$p(y_{N+1}|m_{N+1}, \sigma^2, \omega) = \mathcal{N}(y_{N+1}|m_{N+1}, (1 - \omega) \cdot \sigma^2) \tag{35}$$

Since we obtained through Gibbs sampling a sample from the posterior distribution $p(m_{N+1}, \omega, \sigma^2|\mathbf{x})$, we can generate a sample from $p(y_{N+1}|\mathbf{x})$ by simulating y_{N+1} from $\mathcal{N}(y_{N+1}|m_{N+1}, (1 - \omega) \cdot \sigma^2)$ for each Gibbs sample $m_{N+1}^{(i)}, \omega^{(i)}, \sigma^{2(i)}$. This sample can then be used to produce a probabilistic forecast of y_{N+1} . Fig. 10 presents this probabilistic forecast, together with the predictive distribution of m_{N+1} (the local mean) and the posterior distribution of μ (the long-term mean). It can be seen that the expected value of y_{N+1} (480 with a standard deviation of 260) is well below the long-term mean, since the local mean (480 ± 170) is much lower than the long-term mean (710 ± 60). From Hubert (2000), the observed streamflow for that year was around 200 m³/s, breaking the record for the lowest observed flow.

It is also possible, using the SL model, to produce a long-term probabilistic forecast for as many years ahead as needed. Of course, the hypotheses of the SL model must still hold in the future. In particular, the stationarity hypothesis is of concern in the case of the Senegal River, since the streamflow has remained under the long-term mean for the last 12 years of the record.

Conditional on the value of the parameters and on the latent variable $q_N = (m_N, z_N)$, a probabilistic forecast of the observations $\mathbf{y} = (y_{N+1}, \dots, y_{N+k})$ and latent variables $\mathbf{m}_y = (m_{N+1}, \dots, m_{N+k}), \mathbf{z}_y = (z_{N+1}, \dots, z_{N+k})$ can be obtained by sampling from their joint probability distribution, which can be derived in the same manner as (17):

$$p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y|q_N, \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{m}_y, \sigma^2, \omega) \cdot p(\mathbf{m}_y|\mathbf{z}_y, \mu, \sigma^2, \omega) \cdot p(\mathbf{z}_y|\eta) = \prod_{t=N+1}^{N+k} p(y_t|m_t, \sigma^2, \omega) \times p(m_t|m_{t-1}, z_{t-1}, \mu, \sigma^2, \omega) \cdot p(z_t|\eta) = \prod_{t=N+1}^{N+k} p(y_t, q_t|q_{t-1}, \boldsymbol{\theta}) \tag{36}$$

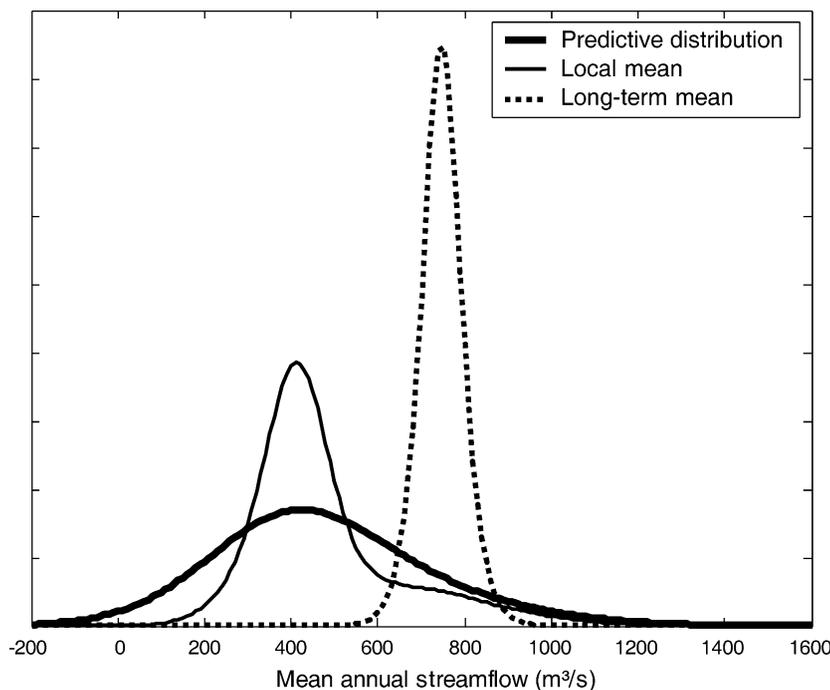


Fig. 10. Probabilistic forecast of the 1987 mean annual streamflow of the Senegal River.

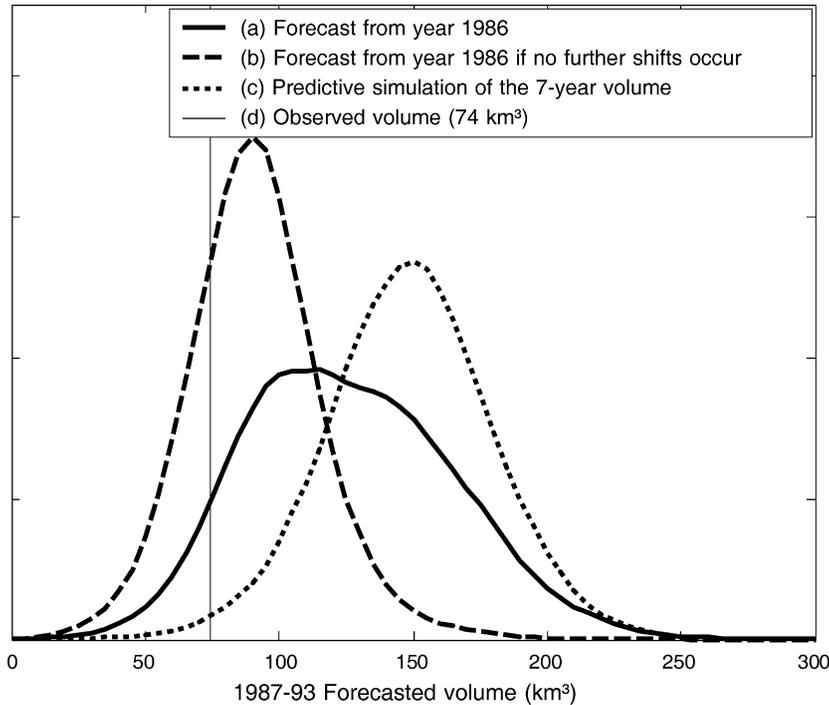


Fig. 11. Probabilistic forecast of the 1987–1993 volume of the Senegal River under three hypotheses: (a) forecast assuming stationarity of streamflows; (b) forecast assuming that the mean of the process stays indefinitely at the 1986 level; (c) simulation of the 7-year volume (no knowledge of initial conditions assumed).

Simulating from $p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | q_N, \boldsymbol{\theta})$ is done iteratively by sampling from $p(y_t, q_t | q_{t-1}, \boldsymbol{\theta})$ for $t = N + 1, \dots, N + k$. As for the 1-year ahead forecast, uncertainty on the parameters of the SL model can be taken into account by sampling from $p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | q_N, \boldsymbol{\theta})$ for each Gibbs sample $(q_N^{(i)}, \boldsymbol{\theta}^{(i)})$. Using this approach, we obtain a sample from the predictive distribution $p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | \mathbf{x}) \propto p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | q_N, \boldsymbol{\theta}) \cdot p(q_N, \boldsymbol{\theta} | \mathbf{x})$. This sample can be used to issue probabilistic forecasts of any function $g(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y)$, for example the total volume $V(\mathbf{y}) = \sum_{t=N+1}^{N+k} y_t$.

Furthermore, conditional forecasts can also be issued, for example to take into account the possibility that the process is not stationary and that the streamflow regime has changed due to human influence on the watershed and on the climate, and will not go back to previous levels. This could be done by letting $z_{N+1} = \dots = z_{N+k} = 0$ in the simulation process (which we will denote simply by $\mathbf{z}_y = 0$), instead of simulating \mathbf{z}_y from $p(\mathbf{z}_y | \eta)$. We would then obtain a sample from the distribution $p(\mathbf{y}, \mathbf{m}_y | \mathbf{x}, \mathbf{z}_y = 0)$.

Fig. 11 presents three different forecasts of the 1987–1993 total streamflow volume for the Senegal River: the predictive distribution, $p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | \mathbf{x})$, the predictive distribution conditional on the additional information that no further shifts occur in these 7 years, $p(\mathbf{y}, \mathbf{m}_y | \mathbf{x}, \mathbf{z}_y = 0)$ and finally, for comparison purposes, a probabilistic forecast of the 7-year volume conditional on the hypothesis that a shift occurs at the end of the observed record, i.e. $p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | \mathbf{x}, z_N = 1)$. Note from (11) and (36) that random samples from $p(\mathbf{y}, \mathbf{m}_y | \mathbf{x}, z_N = 1)$ are independent of the latent variable q_N since $p(m_{N+1} | m_N, z_N = 1, \mu, \sigma^2, \omega) = \mathcal{N}(m_{N+1} | \mu, \omega \sigma^2)$. Since m_{N+1} is drawn from the same distribution as m_0 (Eq. (12)), $p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | \mathbf{x}, z_N = 1)$ corresponds to our best forecast of the 7-year volume if we had only the information on the values of the parameters, and not on the value of the latent variable q_N . This would happen for example if we were asked to make a prediction of the 7-year volume without knowing the years for which the prediction was valid, such as is the case in a typical stochastic simulation framework. For this reason, we

refer to this distribution as the predictive simulation of the 7-year volume.

The observed volume for 1987–1993 was obtained from Fig. 1 of Hubert (2000). At 74 km^3 , it is well below the expectation of the simulated 7-year volume (148 km^3 with a standard deviation of 32 km^3), and much closer to the forecasted 7-year volume ($126 \pm 39 \text{ km}^3$). It is however quite close to the prediction made under the hypothesis that the mean level of 1986 persists throughout the forecast period ($91 \pm 25 \text{ km}^3$). While this result weighs in favor of non-stationarity, the non-exceedance probability of the observed volume computed from $p(\mathbf{y}, \mathbf{m}_y, \mathbf{z}_y | \mathbf{x})$ is still over 8%.

Notice on Fig. 10 that the SL model predicts with a small but non-negligible probability that the flow might be negative, which is obviously not possible. There is of course nothing in the model that prevents negative flows from being simulated or forecasted. If the probability of such an event is significant, one should condition on the event $y_{N+1} \geq 0$, which is straightforward with Gibbs sampling: it only implies dropping in the analysis all Gibbs iteration for which $y_{N+1} < 0$. Another solution to eliminate the possibility for negative forecasts would be to perform a logarithmic transformation on the flows prior to the analysis. As the stationary distribution of x_t is a normal distribution for the SL model (being the sum of two independent processes with normal stationary distributions), the decision of whether or not to transform the data prior to the analysis could be based on an analysis of the linearity of the observations on normal probability paper. In the case of the Senegal River annual flows, it is interesting to note that the observations pass the empirical moments test for normality, but not their logarithm. Hence, no transformation was performed on the data. Given the dependence in the observations, one must however be cautious not to reject the SL model on the basis of a standard test for normality.

4.7. Point forecasting—a comparison with linear ARMA(p, q) models

As was mentioned before, persistence is such in the Senegal River annual flow series that the AIC criterion suggests an AR(2) model, amongst autoregressive AR(p) models. It is interesting to evaluate

how the SL model fares when compared to the AR(2) model selected by the AIC criterion. Another interesting model to consider is the ARMA(1,1), since its autocorrelation structure is similar to that of the SL model. Recall that an ARMA(p, q) model can be written as:

$$(x_t - \mu) = \varepsilon_t + \sum_{k=1}^p a_k \cdot (x_{t-k} - \mu) + \sum_{k=1}^q b_k \cdot \varepsilon_{t-k} \quad (37)$$

where ε_t is a white noise with variance σ_ε^2 , μ is the long-term mean of the process, a_1, \dots, a_p are the autoregressive parameters and b_1, \dots, b_q are the moving average parameters. Hence, an ARMA(p, q) model has $p + q + 2$ parameters. Hence, the SL, AR(2) and ARMA(1,1) all have four parameters.

Only a comparison of point forecasts will be presented in this section, as Bayesian estimation and forecasting for ARMA(p, q) models is outside the scope of this paper. We propose to compare the forecasts which would have been issued by the SL, AR(2) and ARMA(1,1) models for each year, by fitting each model on observations (x_1, x_2, \dots, x_k) and then forecasting the observation x_{k+1} by a point forecast \hat{x}_{k+1} . We showed in Section 4.6 that the SL model could also be used to issue a forecast under the hypothesis that no further shifts occur after the last observed shift. We also want in this section to compare this conditional forecast with the other models.

The predictive distribution of the SL model being generally asymmetric, the posterior expectation may not make a good point forecast. Indeed, it is not the value most likely to occur: it is the mode of the posterior distribution, which corresponds to the most likely value. We shall therefore use the mode of the posterior distribution as our point forecast for the SL model. Of course, it would be preferable to choose a point forecast by maximizing the utility function of the forecast, but this information is generally not available. Estimating the mode of the posterior distribution is not straightforward from Gibbs samples. We obtain it through kernel smoothing, with a normal kernel function optimized for a normal distribution, by finding the value which maximizes the kernel estimation of the p.d.f.

For each time step, we performed only 10,000 iterations for the SL model, keeping the last 5000 for forecasting. To fit the AR(2) and ARMA(1,1) models to the data, we applied the ARMAX function from The MathWorks MATLAB System identification toolbox version 5.0.2. The function is able to fit the models to our dataset for $k \geq 8$. We can therefore compare the performance of the two models from 1910 through 1986. For each year and each model M , we thus obtain a point forecast \hat{x}_{k+1}^M , which we would like to aggregate to obtain a measure of the accuracy of each model. It is tempting to only apply usual criteria such as bias (BIAS) mean absolute error (MAE) or root mean square error (RMSE), but this gives the same weight to each year, whereas the number of observations used to fit the models is increasing, and thus the accuracy of the forecasts is changing with time. As a surrogate for the (hopefully increasing) accuracy of the forecasts, we propose to give a weight to each forecast \hat{x}_{k+1}^M , which is proportional to the number k of observations used to fit the model, and then compute a weighted bias (WBIAS), a weighted MAE (WMAE) and a weighted RMSE (WRMSE) for each model, as well as bias, MAE and RMSE:

$$\begin{aligned}
 \text{BIAS} &= \frac{1}{N - N_0} \sum_{k=N_0}^{N-1} \hat{x}_{k+1}^M - x_{k+1} \\
 \text{MAE} &= \frac{1}{N - N_0} \sum_{k=N_0}^{N-1} |\hat{x}_{k+1}^M - x_{k+1}| \\
 \text{RMSE} &= \sqrt{\frac{1}{N - N_0} \sum_{k=N_0}^{N-1} (\hat{x}_{k+1}^M - x_{k+1})^2} \\
 \text{WBIAS} &= \frac{2}{N^2 - N - N_0^2 + N_0} \sum_{k=N_0}^{N-1} k(\hat{x}_{k+1}^M - x_{k+1}) \\
 \text{WMAE} &= \frac{2}{N^2 - N - N_0^2 + N_0} \sum_{k=N_0}^{N-1} k|\hat{x}_{k+1}^M - x_{k+1}| \\
 \text{WRMSE} &= \sqrt{\frac{2}{N^2 - N - N_0^2 + N_0} \sum_{k=N_0}^{N-1} k(\hat{x}_{k+1}^M - x_{k+1})^2}
 \end{aligned}
 \tag{38}$$

where N_0 is the minimum number of observations kept for model fitting (here $N_0 = 8$), and N is the total number of observations available (here $N = 84$).

Table 4 shows the value of each statistic for the SL, AR(2) and ARMA(1,1) models, as well as for two other simple approaches to forecasting, the AR(0) and naïve forecast. The AR(0) model suggests that the time series is a white noise, and thus its point forecast is simply the past observed average annual flow: $\hat{x}_{k+1}^{\text{AR}(0)} = \sum_{i=1}^k x_i/k$. The naïve forecast is even simpler: it simply corresponds to the previous observation: $\hat{x}_{k+1}^{\text{NAIVE}} = x_k$. In practice, such simple approaches to forecasting can prove hard to beat. The conditional forecast issued with the SL model under the hypothesis that no more shifts in the mean occur is denoted SL-c. The standard deviation of each statistic is also indicated in Table 4, estimated using the Jackknife technique (Quenouille, 1956).

The results obtained show that the naïve forecast is the less biased, whereas the SL-c forecast leads to the smallest error, albeit only by 2–4% depending on the criterion. The SL and ARMA(1,1) models have similar performance, followed by the AR(2) and AR(0) models. Furthermore, the naïve forecast beats the AR(0) and AR(2) models in terms of WMAE. This is in part because all models perform relatively poorly in the latter part of the time series, starting in 1968, as shown by Fig. 12, as they systematically overestimate the annual flow: the amplitude of the bias is six to nine times larger than its value up to 1968. This behavior suggests a structural change in the process underlying the time series, which cannot be accounted for by a stationary model, such as the SL and ARMA(p, q) models. The SL model forecasts can be issued under the hypothesis that the last shift in the mean is a permanent feature, but in that case change-point models such as proposed by Perreault et al. (2000a,b) would be more coherent with this hypothesis.

5. Discussion and conclusion

The SL model has proven to be useful for modeling and generating time series of annual streamflow, but had never been used to our knowledge in forecasting mode or for retrospective analysis and segmentation of time series. This model assumes that the observations are normally distributed with a variance that is constant in time, but with a mean that stays constant

Table 4
Comparison of point forecasts of the Senegal River annual flows

	MAE	RMSE	BIAS	WMAE	WRMSE	WBIAS
SL	198 ± 2	239 ± 2	27 ± 3	196 ± 2	231 ± 2	53 ± 3
SL-c	193 ± 2	237 ± 2	19 ± 3	185 ± 2	225 ± 2	38 ± 3
ARMA(1,1)	197 ± 3	241 ± 3	23 ± 4	192 ± 2	231 ± 2	44 ± 2
AR(2)	213 ± 2	251 ± 2	33 ± 2	207 ± 2	240 ± 4	52 ± 4
AR(0)	232 ± 2	277 ± 3	54 ± 2	249 ± 2	291 ± 4	96 ± 5
Naive	222 ± 3	290 ± 3	7 ± 4	201 ± 3	265 ± 3	9 ± 4

for epochs of length distributed according to a geometric distribution.

5.1. Advantages and difficulties of a Bayesian approach

In this paper, we estimate the parameters of a SL model using a Bayesian approach, along with Gibbs sampling. This estimation method takes into account the whole sample for estimating the parameters, not

only the first moments and the autocovariance function as in Salas and Boes (1980). Furthermore, the Bayesian framework provides an assessment of the uncertainty on the parameters, as well as estimates of the mean level of the process and of years at which shifts in the mean are more likely to have occurred. As a by-product, Gibbs sampling also provides all the information needed to make probabilistic forecasts. Gibbs sampling also provides the probability distribution of the residuals, which is useful for model

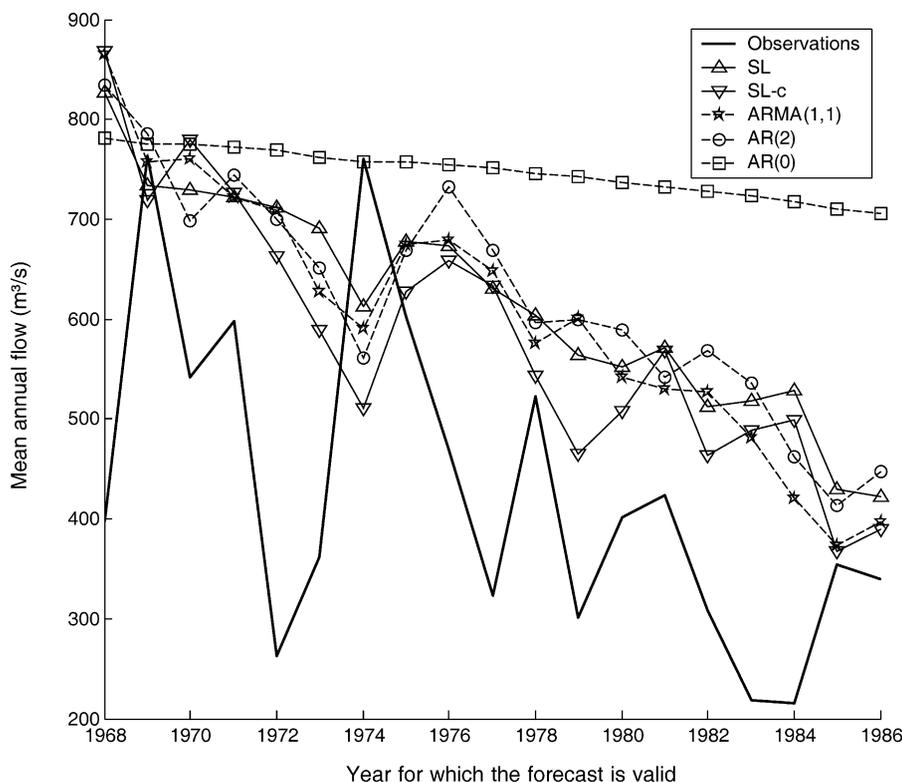


Fig. 12. Point forecast of the mean annual flow of the Senegal River from 1968 to 1986.

verification since the residuals can then be tested for randomness. However, all Bayesian methods require a careful assessment of the prior distribution. Furthermore, with Gibbs sampling there is the additional challenge of making sure that the Gibbs sampler has converged to the stationary distribution.

5.2. Forecasting with regime-switching models

In an application of the model to point forecasting of the Senegal River annual flows, the SL model has not performed better than a linear autoregressive model with moving average ARMA(1,1), which has the same number of parameters as the SL model. However, it did so when we considered a conditional forecast where we supposed that the last shift in the mean to have occurred in the time series was a permanent feature, and that no more shifts would occur in the future. The same phenomenon was observed for probabilistic forecasting: the conditional forecast performed better than the unconditional one. The surprising performance of this conditional forecast points to the fact that the Senegal River annual flows time series may not be stationary, and that a more permanent change in either land use or climate has occurred since the last shift, identified in 1968.

The fact that a regime-switching non-linear model like the SL model does not clearly outperform linear models for point forecasting, despite its apparent success at describing the time series, was to be expected, since linear models can provide good approximation of non-linear time series when the noise level is high, see for example Ramsey (1996) and Dacco and Satchell (1999). However, van Dijk and Franses (2003) suggest that non-linear models can prove superior for forecasting extreme events, and that these events being generally of more importance, this should be reflected in the criterion used for model comparison. Finally, it is possible that a non-linear model improves probabilistic forecasts while not improving upon point forecasts (Clements and Hendry, 1999).

It should also be noted that other models could be devised to forecast annual flows of the Senegal River, for example a non-parametric approach, such as an adaptation of the model proposed by Sharma and O'Neill (2002) for monthly streamflow sequences. But the advantage of the SL model is that it can be used both for retrospective analysis, including for

identification of multiple shifts in a time series, and forecasting. It can then be easier to explain the forecast to the decision-maker by showing how the basin responds to the regional climate, and how the induced persistence can be used to make useful forecasts. Experience in retrospective analysis, stochastic simulation and forecasting of annual flows has taught us that a coherent approach to these problems is of some importance.

5.3. Research perspectives

To be able to model both the persistence induced by climate variability and the persistence induced by storage effects in the watershed, it would be quite interesting to extend this Bayesian SL model to include autocorrelation in the residuals, as proposed by Sveinsson and Salas (2001) and to develop a multivariate model. Since the SL model is in part justified by the suggestion that apparent shifts in the annual mean of streamflows could be induced by sudden shifts in climate, it would make sense, in a given region, to assume that the shifts occur at the same time $\mathbf{z} = (z_0, z_1, \dots, z_N)$ at all sites, along with parameter η which governs these shifts, and let the proportion ω of variance explained by these sudden shifts be allowed to vary from site to site, along with the long-term mean and variance at each site. Forecasting could furthermore be improved by taking into account explanatory variables, such as a seasonal forecast of precipitations or climatic indices.

To better evaluate the probabilistic forecasts produced by a SL model, we also plan to retrospectively compare probabilistic forecasts obtained with a SL model with probabilistic forecasts issued by ARMA processes for different annual flow time series. Scoring probabilistic forecasts is, however, more challenging than scoring point forecasts. We suggest the use of a pseudo Bayes factor (Gelfand and Dey, 1994): for any two probabilistic forecasts of a decision variable y expressed by two p.d.f.'s $p_1(y|\mathbf{x})$ and $p_2(y|\mathbf{x})$, the ratio $B(y_0) = p_1(y_0|\mathbf{x})/p_2(y_0|\mathbf{x})$ evaluated at the observed value y_0 is an interesting relative measure of both accuracy and precision of a probabilistic forecast, and a geometric average of $B(y_0)$ yields a score which is known as a pseudo-Bayes factor.

Finally, some efforts should be devoted to optimizing the estimation procedure in order to reduce

computing time, in particular the generation of parameter ω and a stopping criterion for assessing from parallel runs that sufficient iterations have been performed for inference and forecasting purposes. Indeed, on an entry-level PC computer, we are able to perform about 2000 iterations per minute for the 1903–1986 Senegal River dataset. Hence, an accurate estimation of the posterior distribution takes 1 h 40 min in this case, which can be a limitation for some applications, for example if the procedure were to be used as a screening test for sudden changes in the mean of a process. Reducing computing time would open the possibility for the development of a generic software for retrospective analysis and forecasting using the SL model. For the time being, a MATLAB routine is available from the corresponding author.

5.4. Climate change

We are aware that the approach proposed in this paper for streamflow forecasting does not address one of the most challenging questions at the heart of long-term streamflow forecasting in the 21st century: given the possibility of a human-induced climate change, will streamflows in this century be anything like we have known in the 20th century? Indeed, the SL model is stationary, and in forecast mode will inevitably produce probabilistic forecasts, which converge towards the stationary distribution, once the effects of the initial conditions are gone. If the effects of a human-induced climate change are already visible in the observed record of streamflows, it might be possible from a Bayesian perspective to test the hypothesis that observed streamflows are stationary by comparing two models which allow for shifts in the mean, one being stationary (M_1 , the SL model), and one being non-stationary (M_2). A good candidate for M_2 could be the change-point model proposed by Perreault et al. (2000a). This model assumes that there is a single shift in the mean in the period of record, and implicitly assumes that this shift is a permanent feature, so that future streamflows will have a mean level equal to the mean observed after this shift. Since the models are not nested, we propose to test for stationarity by performing a Bayes factor analysis for these two models. As a by-product of this analysis, we would get the posterior probability of each model given the data, which could then even be used to

forecast from a mixture of the two models, to account for possible non-stationarity in streamflows.

Acknowledgements

The research project that led to this paper was funded by Hydro-Québec Production. Discussions with Pr Jacques Bernier, Pr Michel Slivitzky and Pr Éric Parent were helpful in developing the estimation procedure and analyzing the results. Pr Anne-Catherine Favre and Dr Salaheddine El-Adlouni performed useful convergence analyses on the Gibbs samples, and Rose-Carline Evra also contributed to a previous version of this manuscript as part of her undergraduate co-op program in Statistics. Finally, the comments made by the reviewers resulted in a significant improvement of the manuscript.

Appendix A

Proofs of Eqs. (24)–(33), i.e. the CCDs are shown in this section. For each parameter, the CCD is obtained from (23) by canceling the terms that do not involve this parameter and integrating the numerator to obtain the normalizing constant.

We start by presenting a certain number of intermediate results which will be needed to complete the proofs. First, let us observe that if \mathbf{z} is known, the conditional distribution of \mathbf{m} given \mathbf{z} and $\theta, p(\mathbf{m}|\mathbf{z}, \mu, \sigma^2, \omega)$, is a product of normal distributions. Indeed:

$$\begin{aligned} p(\mathbf{m}|\mathbf{z}, \mu, \sigma^2, \omega) &\propto \mathcal{N}(m_0|\mu, \omega\sigma^2) \\ &\cdot \prod_{t=0}^N \{(1 - z_t) \cdot \delta(m_{t+1} - m_t) + z_t \cdot \mathcal{N}(m_{t+1}|\mu, \omega\sigma^2)\} \\ &\propto \mathcal{N}(m_0|\mu, \omega\sigma^2) \cdot \prod_{\{t: z_t=1\}} \mathcal{N}(m_{t+1}|\mu, \omega\sigma^2) \\ &\propto \prod_{t \in \mathbf{M}} \mathcal{N}(m_t|\mu, \omega\sigma^2) \end{aligned} \quad (\text{A1})$$

where $\mathbf{M} = \{0\} \cup \{t : z_{t-1} = 1, 1 \leq t \leq N + 1\}$.

Two properties of the normal distribution will also be useful

$$\begin{aligned} \int \mathcal{N}(x_t|m_t, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_t|\mu, \omega\sigma^2) dm_t \\ = \mathcal{N}(x_t|\mu, \sigma^2) \end{aligned} \quad (\text{A2})$$

$$\begin{aligned} & \mathcal{N}(x_t|m_t, (1-\omega)\sigma^2) \cdot \mathcal{N}(m_t|\mu, \omega\sigma^2) \\ &= \mathcal{N}(x_t|\mu, \sigma^2) \cdot \mathcal{N}(m_t|(1-\omega)\mu + \omega x_t, \omega(1-\omega)\sigma^2) \end{aligned} \tag{A3}$$

Proof of (A2):

$$\begin{aligned} & \int \mathcal{N}(x_t|m_t, (1-\omega)\sigma^2) \cdot \mathcal{N}(m_t|\mu, \omega\sigma^2) dm_t \\ &= \frac{1}{2\pi\sigma^2\sqrt{\omega(1-\omega)}} \int \exp\left(-\frac{1}{2\sigma^2}\left[\frac{(x_t-m_t)^2}{1-\omega} + \frac{(m_t-\mu)^2}{\omega}\right]\right) dm_t \\ &= \frac{1}{2\pi\sigma^2\sqrt{\omega(1-\omega)}} \exp\left(-\frac{(x_t-\mu)^2}{2\sigma^2}\right) \\ & \quad \times \int \exp\left(-\frac{[m_t-(\mu+\omega x_t-\omega\mu)]^2}{2\omega(1-\omega)\sigma^2}\right) dm_t \\ &= \mathcal{N}(x_t|\mu, \sigma^2) \cdot \int \mathcal{N}(m_t|\mu+\omega(x_t-\mu), \omega(1-\omega)\sigma^2) dm_t \\ &= \mathcal{N}(x_t|\mu, \sigma^2) \end{aligned}$$

Proof of (A3):

$$\begin{aligned} & \mathcal{N}(x_t|\mu, \sigma^2) \cdot \mathcal{N}(m_t|(1-\omega)\mu + \omega x_t, \omega(1-\omega)\sigma^2) \\ &= \frac{1}{2\pi\sigma^2\sqrt{\omega(1-\omega)}} \\ & \quad \times \exp\left(-\frac{\omega(1-\omega)(x_t-\mu)^2 + [m_t-(1-\omega)\mu - \omega x_t]^2}{2\omega(1-\omega)\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\omega\sigma^2}} \exp\left(-\frac{(x_t-m_t)^2}{2(1-\omega)\sigma^2}\right) \frac{1}{\sqrt{2\pi(1-\omega)\sigma^2}} \\ & \quad \times \exp\left(-\frac{(m_t-\mu)^2}{2\omega\sigma^2}\right) \\ &= \mathcal{N}(x_t|m_t, (1-\omega)\sigma^2) \cdot \mathcal{N}(m_t|\mu, \omega\sigma^2) \end{aligned}$$

A.1. Complete conditional distribution for μ

$$p(\mu|-) \propto p(\mathbf{m}|\mathbf{z}, \mu, \sigma^2, \omega) \cdot p(\mu|\sigma^2) \tag{A4}$$

Hence, from (A1):

$$\begin{aligned} p(\mu|-) &\propto \prod_{t \in \mathbf{M}} \mathcal{N}(m_t|\mu, \omega\sigma^2) \cdot \mathcal{N}(\mu|v, \kappa\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\omega^{-1} \sum_{t \in \mathbf{M}} (m_t - \mu)^2 + \kappa^{-1}(\mu - v)^2\right]\right\} \\ &\propto \exp\left\{-\frac{\kappa(r+1) + \omega}{2\kappa\omega\sigma^2}\right. \\ & \quad \left. \times \left[\mu^2 - 2\mu \frac{\kappa(r+1)\bar{m} + \omega v}{\kappa(r+1) + \omega}\right]\right\} \\ &= \mathcal{N}\left(\mu \mid \frac{\kappa(r+1)\bar{m} + \omega v}{\kappa(r+1) + \omega}, \frac{\kappa\omega\sigma^2}{\kappa(r+1) + \omega}\right) \end{aligned} \tag{A5}$$

A.2. Complete conditional distribution for σ^2

$$\begin{aligned} p(\sigma^2|-) &\propto p(\mathbf{x}|\mathbf{m}, \sigma^2, \omega) \cdot p(\mathbf{m}|\mathbf{z}, \mu, \sigma^2, \omega) \\ & \quad \times p(\mu|\sigma^2) \cdot p(\sigma^2) \end{aligned} \tag{A6}$$

Hence, from (A1):

$$\begin{aligned} p(\sigma^2|-) &\propto \prod_{t=1}^N \mathcal{N}(x_t|m_t, (1-\omega)\sigma^2) \\ & \quad \times \prod_{t \in \mathbf{M}} \mathcal{N}(m_t|\mu, \omega\sigma^2) \cdot \mathcal{N}(\mu|v, \kappa\sigma^2) \cdot \mathcal{G}^{-1}(\sigma^2|\alpha, \beta) \\ &\propto (\sigma^2)^{-\frac{N}{2} + \frac{r+1}{2} + \frac{1}{2} + (\alpha+1)} \\ & \quad \times \exp\left(-\sigma^2 \left[\frac{\sum_{t=1}^N (x_t - m_t)^2}{2(1-\omega)} + \frac{\sum_{t \in \mathbf{M}} (m_t - \mu)^2}{2\omega} + \frac{(m-v)^2}{2\kappa} + \beta\right]\right) \\ &= \mathcal{G}^{-1}\left(\sigma^2 \mid \frac{N+r}{2} + 1 + \alpha, \frac{\sum_{t=1}^N (x_t - m_t)^2}{2(1-\omega)} + \frac{\sum_{t \in \mathbf{M}} (m_t - \mu)^2}{2\omega} + \frac{(m-v)^2}{2\kappa} + \beta\right) \end{aligned} \tag{A7}$$

A.3. Complete conditional distribution for η

$$\begin{aligned}
 p(\eta|-) &\propto p(\mathbf{z}|\eta) \cdot p(\eta) \propto \prod_{t=0}^N \{ \eta \cdot \delta(z_t - 1) \\
 &\quad + (1 - \eta) \cdot \delta(z_t) \} \cdot \eta^{s_\eta \tau_\eta - 1} \cdot (1 - \eta)^{s_\eta - s_\eta \tau_\eta - 1} \\
 &\propto \eta^{s_\eta \tau_\eta + r - 1} \cdot (1 - \eta)^{s_\eta - s_\eta \tau_\eta + N - r} \\
 &= \mathcal{B}(\eta | s_\eta \tau_\eta + r, s_\eta - s_\eta \tau_\eta + N - r + 1)
 \end{aligned}
 \tag{A8}$$

A.4. Complete conditional distribution for ω

The CCD for ω is only known to a constant:

$$\begin{aligned}
 p(\omega|-) &\propto p(\mathbf{x}|\mathbf{m}, \sigma^2, \omega) \cdot p(\mathbf{m}|\mathbf{z}, \mu, \sigma^2, \omega) \cdot p(\omega) \\
 &\propto \prod_{t=1}^N \mathcal{N}(x_t | m_t, (1 - \omega) \cdot \sigma^2) \\
 &\quad \prod_{t \in \mathbf{M}} \mathcal{N}(m_t | \mu, \omega \cdot \sigma^2) \cdot \mathcal{B}(\omega | s_\omega, \tau_\omega)
 \end{aligned}
 \tag{A9}$$

A.5. Complete conditional distribution for m_{N+1}

$$\begin{aligned}
 p(m_{N+1}|-) &= \frac{(1 - z_N) \cdot \delta(m_{N+1} - m_N) + z_N \cdot \mathcal{N}(m_{N+1} | \mu, \omega \cdot \sigma^2)}{(1 - z_N) \int \delta(m_{N+1} - m_N) dm_{N+1} + z_N \int \mathcal{N}(m_{N+1} | \mu, \omega \cdot \sigma^2) dm_{N+1}} \\
 &= (1 - z_N) \cdot \delta(m_{N+1} - m_N) + z_N \cdot \mathcal{N}(m_{N+1} | \mu, \sigma^2)
 \end{aligned}
 \tag{A10}$$

A.6. Complete conditional distribution for $q_t = (m_t, z_t), t = 0, 1, \dots, N$

The CCD of q_t is slightly different for $t = 0$ and $t > 0$:

$$\begin{aligned}
 p(q_0|-) &\propto p(m_1 | m_0, z_0, \mu, \sigma^2, \omega) \cdot p(z_0 | \eta) \\
 &= p(m_0, z_0 | m_1, \boldsymbol{\theta}) = p(m_0, z_0 | -)
 \end{aligned}
 \tag{A11}$$

$$\begin{aligned}
 p(q_t|-) &\propto p(x_t | m_t, \sigma^2, \omega) \cdot p(m_t | m_{t-1}, z_{t-1}, \mu, \sigma^2, \omega) \\
 &\quad \cdot p(m_{t+1} | m_t, z_t, \mu, \sigma^2, \omega) \cdot p(z_t | \eta) \\
 &= p(m_t, z_t | x_t, m_{t-1}, m_{t+1}, z_{t-1}, \boldsymbol{\theta}) \\
 &= p(m_t, z_t | -)
 \end{aligned}
 \tag{A12}$$

By definition of conditional probability, the CCD of q_t can be written as the product of the CCD of z_t and the distribution of m_t conditional on all observations, parameters and latent variables with the exception of z_t , which is obtained using the law of total probability:

$$\begin{aligned}
 p(q_t|-) &= p(z_t|-) \cdot p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}) \\
 &= p(z_t|-) \cdot \int p(m_t, z_t | -) dz_t
 \end{aligned}
 \tag{A13}$$

Given \mathbf{m}, z_t is a deterministic, binary function of m_t and m_{t+1} , being equal to zero if and only if $m_t = m_{t+1}$

and to one otherwise. Hence, the CCD of z_t can be written as:

$$p(z_t|-) = p(z_t | m_t, m_{t+1}) = \begin{cases} \delta(0) & \text{if } m_t = m_{t+1} \\ \delta(1) & \text{if } m_t \neq m_{t+1} \end{cases}
 \tag{A14}$$

The conditional distribution $p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta})$ is easily obtained for $t = 0$:

$$\begin{aligned}
 p(m_0 | \mathbf{x}, \mathbf{m}_{(0)}, \mathbf{z}_{(0)}, \boldsymbol{\theta}) &= \int p(m_0, z_0 | -) dz_0 = \int p(m_0, z_0 | m_1, \boldsymbol{\theta}) dz_0 = \frac{\int p(m_1 | m_0, z_0, \mu, \sigma^2, \omega) \cdot p(z_0 | \eta) dz_0}{\int \int p(m_1 | m_0, z_0, \mu, \sigma^2, \omega) \cdot p(z_0 | \eta) dm_0 dz_0} \\
 &= \frac{(1 - \eta) \cdot \delta(m_1 - m_0) + \eta \cdot \mathcal{N}(m_0 | \mu, \sigma^2)}{(1 - \eta) \cdot \int \delta(m_1 - m_0) dm_0 + \eta \cdot \int \mathcal{N}(m_0 | \mu, \sigma^2) dm_0} = (1 - \eta) \cdot \delta(m_1 - m_0) + \eta \cdot \mathcal{N}(m_0 | \mu, \sigma^2)
 \end{aligned}
 \tag{A15}$$

For $t > 0$, it is simpler to consider separately the two cases $z_{t-1} = 0$ and $z_{t-1} = 1$. The case $z_{t-1} = 0$ is simple, since it means that $m_{t-1} = m_t$. Since m_{t-1} is also known, the conditional distribution $p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta})$ is necessarily a Dirac distribution centered on m_{t-1} :

$$p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}, z_{t-1} = 0) = \delta(m_{t-1} - m_t) \quad (\text{A16})$$

In the case $z_{t-1} = 1$, $p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta})$ can be written as:

$$\begin{aligned} p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}, z_{t-1} = 1) &= \int p(m_t, z_t | -) dz_t \\ &= \frac{\int p(x_t | m_t, \sigma^2, \omega) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2) \cdot p(m_{t+1} | m_t, z_t, \mu, \sigma^2, \omega) \cdot p(z_t | \eta) dz_t}{\int \int p(x_t | m_t, \sigma^2, \omega) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2) \cdot p(m_{t+1} | m_t, z_t, \mu, \sigma^2, \omega) dm_t p(z_t | \eta) dz_t} \\ &= \frac{\mathcal{N}(x_t | m_t, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2) [(1 - \eta) \cdot \delta(m_{t+1} - m_t) + \eta \cdot \mathcal{N}(m_{t+1} | \mu, \omega \sigma^2)]}{\int \mathcal{N}(x_t | m_t, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2) [(1 - \eta) \cdot \delta(m_{t+1} - m_t) + \eta \cdot \mathcal{N}(m_{t+1} | \mu, \omega \sigma^2)] dm_t} \end{aligned} \quad (\text{A17})$$

From the properties of the Dirac distribution, the following proposition is immediate:

$$\begin{aligned} &\int \mathcal{N}(x_t | m_t, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2) \\ &\times \delta(m_{t+1} - m_t) dm_t \\ &= \mathcal{N}(x_t | m_{t+1}, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_{t+1} | \mu, \omega \sigma^2) \end{aligned} \quad (\text{A18})$$

Combining (A2) and (A18), we can simplify the denominator of (A17) and obtain:

$$p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}, z_{t-1} = 1) = \frac{\mathcal{N}(x_t | m_t, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2) [(1 - \eta) \cdot \delta(m_{t+1} - m_t) + \eta \cdot \mathcal{N}(m_{t+1} | \mu, \omega \sigma^2)]}{\mathcal{N}(m_{t+1} | \mu, \omega \sigma^2) [(1 - \eta) \cdot \mathcal{N}(x_t | m_{t+1}, (1 - \omega)\sigma^2) + \eta \cdot \mathcal{N}(x_t | \mu, \sigma^2)]} \quad (\text{A19})$$

Once again, using the properties of the Dirac distribution:

$$\begin{aligned} &\mathcal{N}(x_t | m_t, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2) \cdot \delta(m_{t+1} - m_t) \\ &= \mathcal{N}(x_t | m_{t+1}, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_{t+1} | \mu, \omega \sigma^2) \\ &\times \delta(m_{t+1} - m_t) \end{aligned} \quad (\text{A20})$$

This allows us to divide both numerator and denominator by $\mathcal{N}(m_{t+1} | \mu, \omega \sigma^2)$ and obtain:

$$\begin{aligned} p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}, z_{t-1} = 1) &= \frac{(1 - \eta) \cdot \mathcal{N}(x_t | m_{t+1}, (1 - \omega)\sigma^2) \cdot \delta(m_{t+1} - m_t)}{(1 - \eta) \cdot \mathcal{N}(x_t | m_{t+1}, (1 - \omega)\sigma^2) + \eta \cdot \mathcal{N}(x_t | \mu, \sigma^2)} \\ &+ \frac{\eta \cdot \mathcal{N}(x_t | m_t, (1 - \omega)\sigma^2) \cdot \mathcal{N}(m_t | \mu, \omega \sigma^2)}{(1 - \eta) \cdot \mathcal{N}(x_t | m_{t+1}, (1 - \omega)\sigma^2) + \eta \cdot \mathcal{N}(x_t | \mu, \sigma^2)} \end{aligned} \quad (\text{A21})$$

We can then use (A3) to express (A21) as a mixture of two distributions:

$$\begin{aligned} p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}, z_{t-1} = 1) &= (1 - \eta^*) \cdot \delta(m_{t+1} - m_t) + \eta^* \\ &\times \mathcal{N}(m_t | (1 - \omega)\mu + \omega x_t, \omega(1 - \omega)\sigma^2) \end{aligned} \quad (\text{A22})$$

where

$$\eta^* = \left[1 + \frac{1 - \eta}{\eta} \times \frac{\mathcal{N}(x_t | m_{t+1}, (1 - \omega)\sigma^2)}{\mathcal{N}(x_t | \mu, \sigma^2)} \right]^{-1} \quad (\text{A23})$$

Combining (A16) and (A22) we obtain:

$$\begin{aligned}
 p(m_t | \mathbf{x}, \mathbf{m}_{(t)}, \mathbf{z}_{(t)}, \boldsymbol{\theta}) &= (1 - z_{t-1}) \cdot \delta(m_t - m_{t-1}) \\
 &+ z_{t-1} \cdot [(1 - \eta^*) \cdot \delta(m_{t+1} - m_t) \\
 &+ \eta^* \cdot \mathcal{N}(m_t | (1 - \omega) \cdot \mu \\
 &+ \omega \cdot x_t, \omega \cdot (1 - \omega) \cdot \sigma^2)] \quad (\text{A24})
 \end{aligned}$$

References

- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AP-19, 716–723.
- Barreto, G., de Andrade, M., 2000. Bayesian Inference and Markov Chain Monte Carlo Methods Applied to Streamflow Forecasting. *Proceedings Sixth International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Funchal, vol. 2., FOR-034.
- Bengio, Y., 1999. Markovian Models for Sequential Data. *Neural Computing Surveys* 2, 129–162.
- Bernardo, J.M., Smith, M.F.A., 1994. *Bayesian Theory*, Wiley, Chichester.
- Bras, R.L., Rodrigues-Iturbe, I., 1985. *Random Functions and Hydrology*, Addison-Wesley, Menlo Park, CA, 575 pp.
- Brooks, S.P., Gelman, A., 1998. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical statistics*, 7, 434–455.
- Christiansen B., 2003. Evidence for non-linear climate change: two stratospheric regimes and a regime shift. *J. Climate*, 16, 3681–3689.
- Clements, M.P., Hendry, D.H., 1999. *Forecasting Non-stationary Economic Time Series*, MIT Press, Cambridge, MA.
- Curry, K., Bras, R.L., 1978. Theory and application of the multivariate broken line, disaggregation and monthly autoregressive streamflow generators for the Nile River, Technical Report 78-5, Ralph M. Parsons Lab., Mass. Inst. Of Technol, Cambridge.
- Dacco, R., Satchell, S., 1999. Why do Regime-switching models forecast so badly? *Journal of Forecasting* 18, 1–16.
- van Dijk, D., Franses, P.H., 2003. Selecting a nonlinear time series model using weighted tests of equal forecast accuracy, *Econometric Institute Report EI 2003-10*, Econometric Institute, Erasmus University, Rotterdam.
- Dirac, P.A.M., 1958. *The Principles of Quantum Mechanics* (Forth Edition), Oxford U. Press.
- Fortin, V., Perreault, L., Ondo, J.-C., Evra, R.-C., 2002. Bayesian long-term forecasting of annual flows with a shifting-level model, *Proceedings of the Symposium on Managing the Extremes—Floods and Droughts*, Environmental and Water Resources Institute of ASCE, Roanoke, VA.
- Gelfand, A.E., Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.
- Gelfand, A.E., Dey, D.K., 1994. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society B* 56, 501–514.
- Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741.
- Hipel, K.W., McLeod, A.I., 1994. *Time Series Modelling of Water Resources and Environmental Systems*, Elsevier, Amsterdam.
- Hubert, P., 2000. The segmentation procedure as a tool for discrete modeling of hydrometeorological regimes. *Stochastic Environmental Research and Risk Assessment* 14, 297–304.
- Hubert, P., Carbonnel, J.P., Chauuche, A., 1989. Segmentation des séries hydrométéorologiques. Application à des séries de précipitations et de débits de l’Afrique de l’Ouest. *Journal of Hydrology* 110, 349–367.
- Hurst, H.E., 1951. Long-term storage capacity of reservoirs. *Transactions of ASCE* 116, 770–779.
- Kehagias, A., 2004. A hidden Markov Model Segmentation procedure for hydrological and environmental times series, *Stoch. Env. Res. and Risk Ass.*, 18(2), 117–130.
- Klemes, V., 1974. The Hurst phenomena: a puzzle? *Water Resources Research* 10(4), 675–688.
- Kuczera, G., Parent, É., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology* 211(1–4), 69–85.
- Loucks, D.P., Stedinger, J.R., Haith, D.A., 1981. *Water Resource Systems Planning and Analysis*, Prentice Hall, Englewood Cliffs.
- Lu, Z.-Q., Berliner, L.M., 1999. Markov switching time series models with application to a daily runoff series. *Water Resources Research* 35(2), 523–534.
- Mandelbrot, B.B., Wallis, J.R., 1969. Computer experiments with fractional Gaussian noises, part 1, averages and variances. *Water Resources Research* 5(1), 228–241.
- Montanari, A., Rosso, R., Taquq, M.S., 1997. Fractionally differenced ARIMA models applied to hydrologic time series: identification, estimation and simulation. *Water Resources Research* 33, 1035–1044.
- O’Connell R.T., 1971. A simple stochastic modeling of Hurst’s law, *International Symposium On Mathematic Models in Hydrology*, Warsaw.
- Perreault, L., Bernier, J., Bobée, B., Parent, É., 2000a. Bayesian change-point analysis in hydrometeorological time series, part 1, the normal model revisited. *Journal of Hydrology* 235, 221–241.
- Perreault, L., Bernier, J., Bobée, B., Parent, É., 2000b. Bayesian change-point analysis in hydrometeorological time series, part 2, comparison of change-point models and forecasting. *Journal of Hydrology* 235, 242–263.
- Potter K.W., A Stochastic model of the Hurst phenomenon: nonstationarity in hydrologic processes. PhD Dissertation, John Hopkins University, 1976.
- Quenouille, M.H., 1956. Notes on bias in estimation. *Biometrika* 43, 353–360.
- Ramsey, J.B., 1996. If nonlinear models cannot forecast, what are they? *Studies in Nonlinear Dynamics and Econometrics* 1, 65–86.
- Rassam, J.C., Fagherazzi, L.D., Bobée, B., Mathier, L., Roy, R., Carballada, L., 1992. Beauharnois-Les Cèdres spillway: design

- flood study with stochastic approach, Final report to the Experts Committee, Hydro-Québec, Montreal, Canada, 105 p.
- Rial, J.A., Pielke, R.A. Sr, Beniston, M., Claussen, M., Canadell, J., Cox, P., Held, H., de Noblet-Ducoudré, N., Prinn, R., Reynolds, J., Salas, J.D., 2004. Nonlinearities, Feedbacks, and Critical Thresholds Within the Earth's Climate System. *Climatic Change* in press.
- Robert, C., 1998. Discretization and MCMC convergence assessment, Lecture Notes 135, Springer, New York.
- Robert, C., Casella, G., 1999. Monte Carlo statistical methods, Springer, New York.
- Rodriguez-Iturbe, I., Mejia, J.M., Dawdy, D.R., 1972. Streamflow simulation. 1, A new look at Markovian models, fractional Gaussian noise and crossing theory; 2, the broken line process as a potential model for hydrologic simulation. *Water Resources Research* 8, 921–941.
- Salas, J.D., 1993. Analysis and modeling of hydrologic time series. In: Maidment, D., (Ed.), *Handbook of Hydrology*, McGraw Hill, New York, (Chapter 19).
- Salas, J.D., 2000. Stochastic Analysis and Modeling for Simulation and Forecasting, Consulting report for Hydro-Québec, Colorado State University.
- Salas, J.D., Boes, D.C., 1980. Shifting level modeling of hydrologic series. *Advances in Water Resources* 3, 59–63.
- Salas, J.D., Boes, D.C., Pegram, G.G.S., Yevjevich, V., 1979. The Hurst phenomenon as a preasymptotic behavior. *Journal of Hydrology* 44, 1–15.
- Salas, J.D., Delleur, J.W., Yevjevich, V., Lane, W.L., 1980. Applied modeling of hydrologic time series, *Water Resources Publications*, Littleton, CO.
- Sharma, A., O'Neill, R., 2002. A nonparametric approach for representing interannual dependence in monthly streamflow sequences. *Water Resources Research* 38(7), 2105–2117.
- Spiegelhalter, D.J., Thomas, A., Best, N.G., 1996. Computation on Bayesian graphical models. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics 5*, Oxford University Press, New York, pp. 407–425.
- Sveinsson, O.G.B., Salas, J.D., 2001. Stochastic modeling and simulation of the great lakes net basin supplies based on univariate and multivariate shifting mean (with persistence), Consulting Report for Hydro-Québec, Colorado State University.
- Sveinsson, O.G.B., Salas, J.D., Boes, D.C., Pielke, R.A. Sr, 2003. Modeling of long term variability of climatic and hydrologic processes. *Journal of Hydrometeorology* 4, 489–496.
- Schwing, F.B., Jiang, J., Mendelsohn, R., 2003. Coherency of multi-scale abrupt changes between the NAO, NPI, and PDO. *Geophysical Research Letters* 30(7), 1406–1409.
- Thyer, M., Kuczera, G., 2000. Modelling long term persistence in hydro-climatic time series using a hidden state Markov model. *Water Resources Research* 36(11), 3301–3310.
- Thyer, M., Kuczera, G., 2003a. A hidden Markov model for modeling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach. *Journal of Hydrology* 275, 12–26.
- Thyer, M., Kuczera, G., 2003b. A hidden Markov model for modeling long-term persistence in multi-site rainfall time series 1. Real data analysis. *Journal of Hydrology* 275, 27–48.
- Valdes, J.B., Burlando, P., Salas, J.D., 2002. Stochastic Forecasting of Precipitation and Streamflow Processes. In: Potter, T.D., Colman, B. (Eds.), *Handbook of Weather, Climate, and Water*, Wiley, New York, (Chapter 34).
- Yonetani, T., Gordon, H.B., 2001. Abrupt changes as indicators of decadal climate variability. *Climate Dynamics* 17, 249–258.
- Zucchini, W., Guttorp, P., 1991. A hidden Markov model for space-time precipitation. *Water Resources Research* 27(8), 1917–1923.