AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2014WR015567

Key Points:

- Stochastic simulation of flow time series with regime-like behavior
- The method used a gamma HM and multinomial logistic regression model
- Model captures observed statistics and nonstationary spectral variability

Correspondence to:
C. Bracken,
cameron.bracken@colorado.edu

Citation:
Bracken, C., B. Rajagopalan, and E. Zagona (2014), A hidden Markov model combined with climate indices for multidecadal streamflow simulation, *Water Resour. Res.*, *50*, 7836–7846, doi:10.1002/2014WR015567.

Received 10 MAR 2014
Accepted 9 SEP 2014
Accepted article online 12 SEP 2014
Published online 8 OCT 2014

A hidden Markov model combined with climate indices for multidecadal streamflow simulation

C. Bracken[1,2], B. Rajagopalan[1,3], and E. Zagona[1,4]

[1]Department of Civil, Environmental, and Architectural Engineering, University of Colorado at Boulder, Boulder, Colorado, USA, [2]Bureau of Reclamation, Technical Service Center, Denver, Colorado, USA, [3]Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, Colorado, USA, [4]Center for Advanced Decision Support for Water and Environmental Systems, University of Colorado at Boulder, Boulder, Colorado, USA

Abstract Hydroclimate time series often exhibit very low year-to-year autocorrelation while showing prolonged wet and dry epochs reminiscent of regime-shifting behavior. Traditional stochastic time series models cannot capture the regime-shifting features thereby misrepresenting the risk of prolonged wet and dry periods, consequently impacting management and planning efforts. Upper Colorado River Basin (UCRB) annual flow series highlights this clearly. To address this, a simulation framework is developed using a hidden Markov (HM) model in combination with large-scale climate indices that drive multidecadal variability. We demonstrate this on the UCRB flows and show that the simulations are able to capture the regime features by reproducing the multidecadal spectral features present in the data where a basic HM model without climate information cannot.

1. Introduction and Background

We motivate, for clarity, the introduction and background of this research through the Upper Colorado River Basin (UCRB) streamflow variability. The annual Lees Ferry naturalized flow series, at the outlet of UCRB (Figure 1), exhibits a distinct regime-like behavior with sustained departures from the mean annual flow—1906–1930 represents a high flow period followed by nearly 50 years of lower than average flow with a sudden shift to higher flows in the mid-1980s and finally the recent prolonged drought—though still maintains a weak autocorrelation structure (lag 1 autocorrelation of 0.26, which is barely significant). A Hurst coefficient of 0.73 (using the regression of the spectral density function [*Taqqu et al*., 1995]) or 0.59 (using Robinson's method [*Robinson*, 1994]) also does not indicate a strong long-term persistence in terms of the Hurst effect [*Hurst*, 1951]. While we may never be certain if we are observing a stationary time series with a weak Hurst effect or a truly nonstationary series [*Koutsoyiannis and Montanari*, 2007], methods based on short-term persistence are clearly not adequate to explain the observations. Traditional time series models [*Salas et al*., 1980] are based on short-term memory and stationarity, thus, a weak autocorrelation leads to weak persistence and lower probability of long wet and dry spells. These long departures from the mean are important for multiyear reservoir planning in the UCRB since they stress the system far more than single wet or dry years. This behavior is similar to that suggested by *Akintuğ and Rasmussen* [2005] where persistence structure in the data (in terms of climate regimes) is not fully described by the autocorrelation function.

Links between large-scale climate fluctuations and UCRB hydrology are increasingly apparent [*Nowak et al*., 2012; *Timlsena et al*., 2009; *McCabe et al*., 2007; *Hunter et al*., 2006; *Grantz et al*., 2005; *Hidalgo*, 2003; *Piechota and Dracup*, 1996; *Nash and Gleick*, 1991]. Previous studies provide evidence of the varying influence of three oceanic climate phenomena, the El Niño Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), and the Atlantic Multidecadal Oscillation (AMO) on UCRB streamflow and precipitation. El Niño, PDO warm phase, and AMO cold phase are associated with increases in streamflow in the UCRB and La Niña, PDO cold phase, and AMO warm phase are associated with decreased streamflow [*Timlsena et al*., 2009]. In addition, coupling effects of the three phenomena are known to be important for streamflow generation. For example, a corresponding El Niño and PDO warm phase is associated with a 10–30% increase in UCRB seasonal streamflow volume as opposed to a 5–15% increase from either phenomenon alone [*Timlsena et al*., 2009]. Time-varying influence of these climate phenomena are thought to create the type of regime-switching behavior observed in the data.

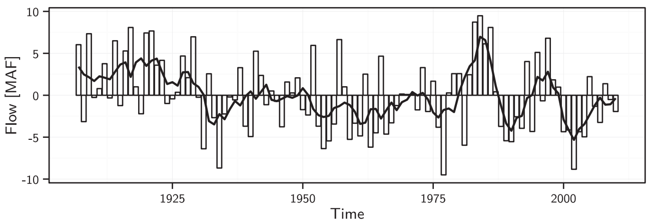

Figure 1. Lees Ferry annual naturalized flow time series. The black line represents a 5 year running mean.

Capturing the regime-switching behavior requires specialized methods capable of capturing the nonstationary spectrum. Fractional Gaussian models [*Koutsoyiannis*, 2002] that can account for Hurst effect, which are stationary models, can capture nonstationarity. However, if the Hurst effect is weak, such as the case here, these models might not be well suited. Traditional time series models based on autoregressive-moving average (ARMA) framework are not designed to capture nonstationary features in the spectrum and typically (but not necessarily) assume errors are normally distributed. The ARMA models reproduce a specific spectral characteristics that are smoothed and stationary [*Salas et al*., 1980]. Nonparametric methods based on block bootstrap [*Ouarda et al*., 1997], K-nearest neighbor bootstrap [*Lall and Sharma*, 1996; *Prairie et al*., 2008] are all attempts to alleviate the normality assumption, and they perform very well at capturing non-normal features such as bimodality as can be seen in the above references. They too cannot capture nonstationary spectral features because time domain models—traditional or nonparametric—map on to stationary spectra. Spectral-based methods decompose the time series into orthogonal frequency bands using wavelets and simulate each with an ARMA model and add them, known as WARM—this captures the spectral features very well [*Kwon et al*., 2007]. Improvements to WARM capture nonstationary spectrum have been proposed [*Nowak et al*., 2011]. While the spectral features are captured, the ARMA framework restricts the ability to simulate effectively the distributional properties.

Some time series models can explicitly capture regime-switching behavior without decomposition. The shifting level (SL) proposed by *Boes and Salas* [1978] can explicitly capture shifts in the mean of a series. In an SL model, the series is modeled as a sum of two independent stochastic processes, one for the mean component and one for the noise component [*Fortin et al*., 2004]. *Salas and Boes* [1980] describe the case of the Nile River Basin where the shifting means can produce spurious autocorrelation. The Lees Ferry data exhibits similar behavior, the serial correlation from 1906 to 1981 is 0.11 and from 1982 to 2010 it is 0.48 while for the whole period of record it is 0.26. This is indicative of a regime-switching behavior in terms of autocorrelation but the behavior is also apparent in the mean (Figure 1). The original SL model did not estimate the nonstationary mean and was therefore not useful for forecasting. A reformulated SL model was described by *Fortin et al.* [2004] and shown to be a special case of a class of models known as hidden Markov models.

Hidden Markov (HM) models (also known as Markov switching models or dependent mixture models) have wide applicability in hydrology for both simulation and forecasting. In HM models, a system switches between a fixed number of unobserved or "hidden" states via a Markov chain and corresponding transition probabilities. Each state corresponds to a probability distribution, called a component distribution, from which observed time series values are drawn. The original hydrologic applications of HM models were to rainfall data [*Jackson*, 1975; *Zucchini and Guttorp*, 1991; *Thyer and Kuczera*, 2000; *Mehrotra and Sharma*, 2005; *Greene et al*., 2008; *Kwon et al*., 2008; *Khalil et al*., 2010; *Yoo et al*., 2010; *Greene et al*., 2011]. In applications to streamflow, HM models [*Zucchini and Guttorp*, 1991; *Akintuğ and Rasmussen*, 2005; *Gelati et al*., 2010; *Evin et al*., 2011] are attractive because of their ability to simulate long persistence and regime-switching behavior in hydrologic time series. Regime changes, or switches between hidden states, especially on the annual scale have been attributed to regime shifts driven by large-scale climate features.

Akintuğ and Rasmussen [2005] discuss the correspondence of HM models with AR models. Though different in design, they suggest that an HM(m) model has a similar autocorrelation structure to an order AR$(m+1)$. This result is indeed important for data sets that exhibit significant autocorrelation at high lags but leaves open the question of applicability to data with very weak autocorrelation such as Lees Ferry. *Akintuğ and Rasmussen* [2005] use a homogeneous stationary hidden Markov model to simulate annual runoff for the Niagara River. The Niagara River exhibits strong autocorrelation and so is ideal to be simulated with HM models or higher-order ARMA models. Lees Ferry flow time series exhibits strong nonstationarity in the spectrum with decadal variability in recent years (Figure 7a)—the standard HM such as those used above would not capture this spectral feature, necessary for robust simulation of wet/dry epochs. Previous studies typically suggest that the HM states are driven by large-scale climate processes such as ENSO [*Gelati et al*., 2010] or PDO [*Akintuğ and Rasmussen*, 2005] but few studies actually include those indices directly into simulations or forecasts. *Gelati et al*. [2010] use a specialized HM model called the Markov modulated autoregressive model with exogenous input (MARX) with transition probabilities conditioned on sea surface temperatures. They make single step quarterly forecasts and longer-term simulations of runoff. Multifractal methods (e.g., Fractional Brownian Motion) [*Chi et al*., 1973; *Stedinger and Taylor*, 1982; *Hosking*, 1984; *Mesa and Poveda*, 1993; *Lohre et al*., 2003] are able to simulate long-memory processes (as described by the Hurst coefficient), which can reproduce nonstationary behaviors and may include exogenous inputs. However, including exogenous variables in these models is not straightforward and they are not guaranteed to capture nonstationary spectral features. *Henley et al*. [2011] developed a climate-informed multitime scale stochastic (CIMSS) framework that directly incorporates observed and paleo climate indices. The CMISS framework is able to capture observed wet/dry state run lengths better than the HM models they used but they do not discuss the ability of the model to capture local spectral features.

It is therefore clear, there is a need to develop a framework for time series simulation incorporating the regime-switching behavior of a hidden Markov model but that also captures observed nonstationary spectral features by incorporating large-scale climate information; this motivates the current study.

The paper is structured as follows: a brief overview of the model formulation and parameter estimation is given. The application of HM models to Lees's Ferry naturalized flow data is described. Results of the simulation and forecasting procedures are presented followed by the conclusions of this study. In Appendix A, we present the moments of a gamma HM model since they are not common in the literature.

2. Methodology

2.1. Model Formulation

As mentioned above, hidden Markov (HM) models are also known as Markov switching models, Markov mixture models, or dependent mixture models. An order m HM model transitions or switches between m "hidden" states according to a discrete Markov chain with transition probability matrix Γ. These states are typically described as climate regimes [*Thyer and Kuczera*, 2000; *Akintuğ and Rasmussen*, 2005; *Gelati et al*., 2010]. Each state prescribes a probability distribution known as a component distribution. The parameters of the component distributions are dependent on the state of the Markov process. Note that although the terms "wet" and "dry" are commonly used to describe states of an HM model, low flows can be generated in the "wet" state and vice versa.

The notation for HM models in the literature is somewhat nonstandard; we will adopt the notation of *Zucchini* [2009]. For an observed sequence X_t, $t=1, 2, \ldots, T$, the general form of an HM model is

$$\Pr(S_t|\mathbf{S}^{(t-1)})=\Pr(S_t|S_{t-1}), t=2, 3, \ldots, T \quad (1)$$

$$\Pr(X_t|\mathbf{X}^{(t-1)}, \mathbf{S}^{(t)})=\Pr(X_t|S_t), t \in \mathbb{N} \quad (2)$$

where S_t is the unobserved or "hidden" sequence that follows a simple first-order Markov process. $\mathbf{S}_t$ denotes the sequence $S_1, S_2, \ldots, S_T$. The transition probabilities, i.e., the conditional probabilities of transition from one hidden state to another, are defined as

$$\gamma_{jk}=\Pr(S_{i+1}=k|S_i=j) \quad (3)$$

or in matrix form

$$\Gamma=\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{bmatrix}. \quad (4)$$

In this formulation, the observed sequence X_t is dependent only on the current hidden state S_t. Note that in general X_t is not a Markov process [*Zucchini*, 2009].

The unobserved sequence S_t determines the state-dependent probability distribution

$$p_i(x) = \Pr(X_t = x | S_t = i). \tag{5}$$

For our purposes, p_i will represent a probability density function but it may similarly represent a probability mass function in the discrete case. Previous HM models of streamflow have used normal components distributions [*Jackson*, 1975; *Zucchini and Guttorp*, 1991; *Akintuğ and Rasmussen*, 2005; *Gelati et al*., 2010] but recent studies [*Wiper et al*., 2001; *Al-Saleh and Agarwal*, 2007; *Evin et al*., 2011] have explored the use of gamma component distributions. The gamma distribution is commonly used in the hydrologic modeling because of its lower bound of zero [*Salas et al*., 1980]. In this application, we use gamma component distributions, which are intuitive for strictly positive hydrologic applications such as streamflow and rainfall. In the case of a gamma component distribution

$$p_i(x) = g(x; k_i, \theta_i) = \frac{\theta_i^{k_i}}{\Gamma(k_i)} x^{k_i - 1} e^{-\theta_i x} \text{ for } x \geq 0 \tag{6}$$

where k_i is the state-dependent shape parameter, θ_i is the state-dependent rate parameter, and Γ is the gamma function. The result is analogous for normal component distributions.

We use a nonstationary version of the model described by *Akintuğ and Rasmussen* [2005]. Our model does not assume that the initial distribution is the stationary distribution and therefore allows the expected state to change in time. The stationary distribution, δ, can be computed conveniently from the identity [*Zucchini*, 2009]

$$\delta(\mathbf{I}_m - \Gamma + \mathbf{U}) = \mathbf{1}_m \tag{7}$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix, $\mathbf{U}$ is an $m \times m$ matrix of ones, and $\mathbf{1}_m$ is an m dimension row vector of ones.

2.2. Parameter Estimation and Model-Order Selection

Many methods exist to estimate the parameters of HM models. Commonly used techniques include direct maximization of the likelihood function [*Zucchini*, 2009; *Akintuğ and Rasmussen*, 2005] and Bayesian estimation procedures [*Thyer and Kuczera*, 2000, 2003]. Another common method is known as the Expectation Maximization (EM) algorithm [*Dempster et al*., 1977] for maximum likelihood estimation when some data are missing (in this case the hidden states). The EM algorithm provides a good compromise between the efficiency of direct maximization and the robustness of Bayesian techniques. The implementation of the EM algorithm in this context is known as the Baum-Welch algorithm [*Baum et al*., 1970]. The EM algorithm starts with an expectation step (E-step) to provide an estimate of the data likelihood given parameter estimates. The E-step is followed by the maximization step (M-step) where the data likelihood is maximized with respect to the parameters. The E-step and M-step are repeated until convergence of parameter values is achieved [*Zucchini*, 2009]. For gamma component distributions, the portion of the data likelihood that depends on the gamma parameters, no closed form equation exists so numerical maximization must be employed.

The EM algorithm requires initial parameter guesses, the transition probabilities (Γ), the parameters of the component distributions (Λ), and the initial distribution (δ). We use the following criteria for initial parameter guesses:

1. $\Gamma_0 = \mathbf{U}/m$, where $\mathbf{U}$ is an $m \times m$ matrix of ones.
2. The initial parameters are estimated by fitting a single component distribution to the entire data and then the same estimates are used for all the component distributions.
3. The initial distribution is first estimated as $\delta_0 = (1, \mathbf{0}_{m-1})$ where $\mathbf{0}_{m-1}$ is an m – 1 dimension row vector of zeros. If the choice for δ does not yield m distinct component distributions after employing the EM algorithm, then try $\mathbf{1}_m/m$ where $\mathbf{1}_m$ is an m dimension row vector of ones.

Both the Bayesian Information Criteria (BIC) [*Schwarz*, 1978] and the Akaike Information Criteria (AIC) [*Akaike*, 1974] can be used to determine the optimal model order for the HM model. Lower values of both AIC and BIC are favorable, where BIC more heavily penalizes higher numbers of parameters. In model selection, the model order is selected when increasing the order further would cause an increase in AIC or BIC.

2.3. Global Decoding

Global decoding of the hidden states was done using the Viterbi algorithm [*Forney*, 1973]. The Viterbi algorithm is a recursive procedure which maximizes the conditional probability of the sequence of states given the observed data values

$$\Pr(\mathbf{S}^{(T)}=\mathbf{s}^{(T)}|\mathbf{X}^{(T)}=\mathbf{x}^{(T)}). \tag{8}$$

The resulting sequence $s_1, s_2, \ldots, s_T$ is the most likely sequence of states, known as the global decoding. Complete details of the procedure are given in *Zucchini* [2009].

2.4. State Model—Incorporating Climate Information

In a traditional HM model, states are determined by the transition probability matrix of the underlying Markov process. Simulations from a traditional HM model are not conditioned on the climate and are therefore not expected to capture observed decadal variability. Assuming that decadal variability in streamflow is primarily the result of large-scale climate fluctuations, we propose an alternate method to simulate the state of the HM model that is directly informed by the climate system.

The globally decoded states are a time series, $s_1, s_2, \ldots, s_T$, taking values $1, \ldots, m$. Using the globally decoded state series, we train a multinomial logistic regression model to obtain estimates of the probability the system was in a particular state. Multinomial logistic regression is a generalization of binary logistic regression where the response variable can take on K possible outcomes instead of two [*Hastie et al*., 2002]. Predictor variables can be continuous or discrete. If the order of the HM model is 2, the state model simplifies to binary logistic regression. The reader is referred to *Hastie et al*. [2002] for the multinomial extension in the case of 3 or more states. The form of the state model follows from an appropriate generalized linear model [*McCullagh and Nelder*, 1989]

$$\eta_t=\ln\frac{\pi_t}{1-\pi_t}=\beta_0+\beta\cdot\mathbf{x}_t \tag{9}$$

where η_t is known as the logit link function, π_t is the probability that the system was in state 1 in year t, β_0 is the intercept term, β is a vector of the logistic regression parameters, and $\mathbf{x}_t$ is a vector of climate indices at year t. Discrete states are related to climate indices by the logistic function

$$\pi_t=\frac{\exp(\beta_0+\beta\cdot\mathbf{x}_t)}{1+\exp(\beta_0+\beta\cdot\mathbf{x}_t)} \tag{10}$$

The logistic function can only take values between 0 and 1, interpreted as probabilities. The parameters can be estimated with iteratively reweighted least squares (IRLS) using Newton-Raphson/Fisher-scoring iterations [*Hastie et al*., 2002].

2.5. Simulation Procedure

Given an HM model and a state model, together referred to as the climate-informed hidden Markov model (HMC), we can generate simulations using the state model predictions instead of the traditional simulation approach of using the HM transition probabilities. By doing so, we capture the decadal variability present in the climate indices. This is akin to using nonhomogeneous transition probabilities [*Robertson et al*., 2004].

1. Draw a uniform random number $r=U(0,1)$.

2. For each year t:

   2.1. Determine the state estimate, $\hat{s}_t$, by finding the sample rank of the first element of the vector $[r, C_1, C_2, \ldots, C_m]$ where $C_j=\sum_{i=0}^{j}\pi_{i,t}$, the sequence of partial sums of the state probabilities at time t.

   2.2. For each year t, draw a random number from the state-dependent component distribution determined by $\hat{s}_t$.

3. Repeat steps 1–3 to generate the desired number of simulations.

3. Results

Ensembles of Lees Ferry naturalized streamflow each 105 years long, same as the historical data, are simulated using the two HM models—the basic HM model with gamma component distributions, referred to as HMG and a climate-informed version of the same referred to as HMC, as described above. Both methods sample from identical gamma component distributions but differ in their simulation of the hidden state.

First, the best HMG model is fitted—for this the AIC and BIC were used and both resulted in a second-order model, i.e., two states each with a component gamma distribution. The model parameters—the state transition probabilities from 1 year to next and the parameters of the two gamma distributions—are shown in Table 1. The two gamma distributions along with the histogram and the combined distribution of the data are shown in Figure 2. The two distributions of the two hidden states, based on the location of their peaks characterize two "regimes"—normal and wet.

From the global decoding, the most likely sequence of hidden states that evolved the historical flow time series is shown in Figure 3b. The persistence of the system in terms of prolonged stretches of these states—i.e., regimes—is readily apparent. This is further highlighted in Figure 3a which shows the historic flow time series overlaid with the mean of the period over which the model was in a particular state (Figure 3a). We can also see that the states correspond to wet and dry epochs in the data. The period in the early 1900s corresponds the time in which the Colorado River Compact [*Colorado River Commission and Coauthors*, 1923] was signed which is known to be a period of above average flow and the same during the 1980s. The hidden states and the corresponding gamma distribution reveal the regime-like behavior of the flow series.

Table 1. HM Model Parameters

Parameter	Model
k	$\begin{bmatrix} 12.11480 \\ 22.02217 \end{bmatrix}$
θ	$\begin{bmatrix} 0.889508 \\ 1.221267 \end{bmatrix}$
Γ	$\begin{bmatrix} 0.98 & 0.02 \\ 0.08 & 0.92 \end{bmatrix}$
δ	$[0, 1]$

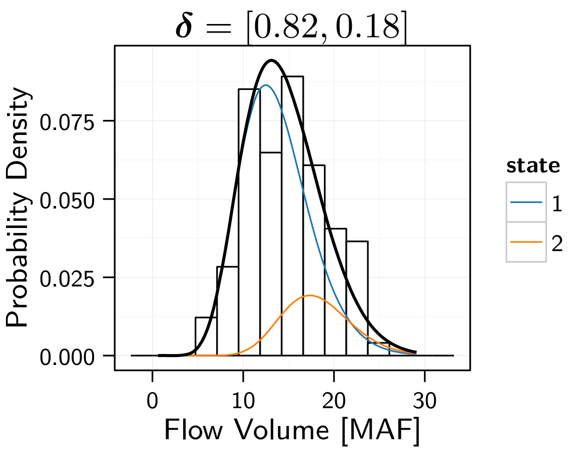

Figure 2. Stationary distributions of gamma component HM model.

To fit the HMC model, a state generation model is developed using climate indices. For predictors of state, we chose time series of climate indices that are known to influence streamflow in the UCRB as potential predictors: AMO, PDO, and ENSO (specifically the Niño3 index) [*Nowak et al.*, 2012; *Timlsena et al.*, 2009; *McCabe et al.*, 2007; *Hunter et al.*, 2006; *Grantz et al.*, 2005; *Hidalgo*, 2003; *Piechota and Dracup*, 1996; *Nash and Gleick*, 1991]. To capture the temporal dependence, we also included previous year's state, S_{t-1}, as a predictor. By including climate indices along with S_{t-1}, we expected to capture both decadal variability and persistence of state. Using the globally decoding states in Figure 3b, the dependent variable is this two category time series. Given that the optimal HM model order is 2, we fit a two category multinomial logistic regression model, which in this case

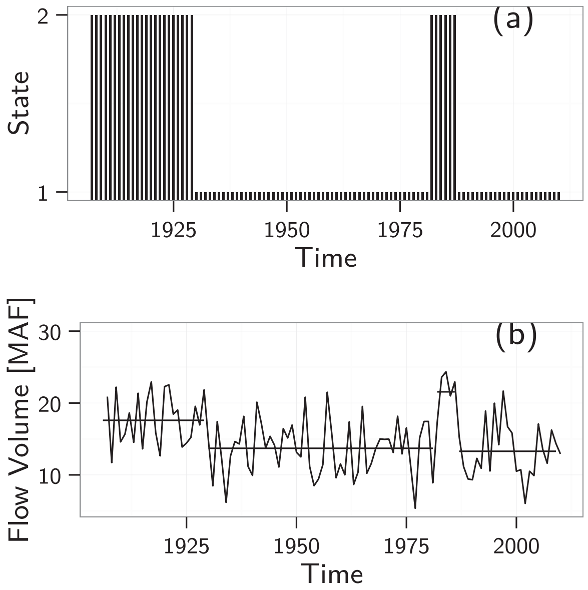

Figure 3. (a) Global decoding of the HM model using the Viterbi algorithm. (b) Plotted below is the Lees Ferry annual time series. The horizontal lines indicate the mean of the periods above where the model was in a particular state.

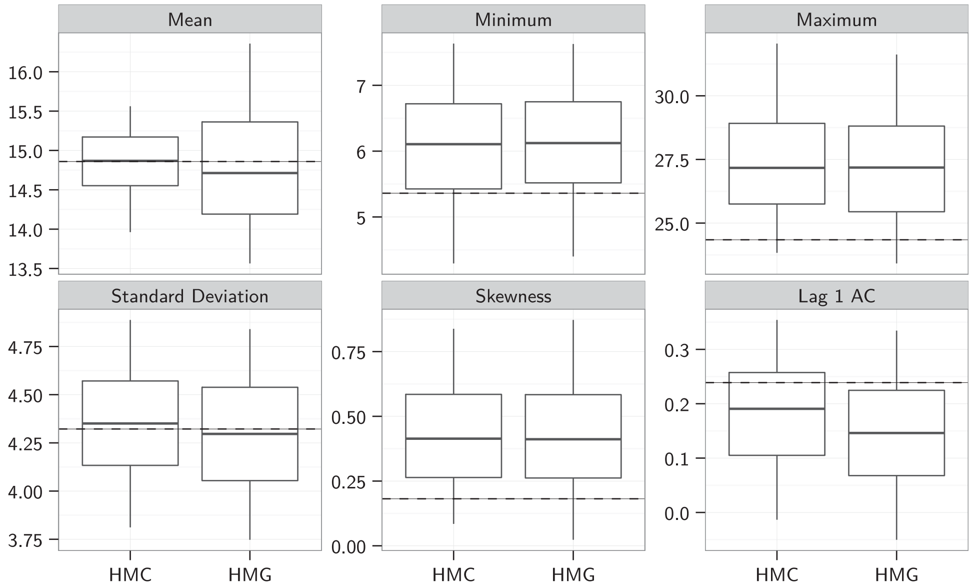

Figure 4. Basic simulation statistics of HMC framework and basic HMG. Boxplots represent the spread of the simulations and the horizontal line represents the observed value.

is simply a binary logistic regression model described in the previous sections. The best model based on AIC was obtained as:

$$\eta_t = \beta_0 + \beta_1 \mathrm{AMO}_t + \beta_2 \mathrm{PDO}_t + \beta_3 S_{t-1}. \tag{11}$$

We present parameter estimates and results from a basic HM model using gamma component distributions (HMG), as well as the same HM model (HMC), using a climate-based state model, referred to as the HMC. As a control case, we also include results where we impose a perfect knowledge of the HM states.

Following *Guimarães and Santos* [2011], we generated 1200 ensembles from both models, HMG and HMC, each of length 105 (the same length of the data). Figure 4 shows the boxplots of basic distributional statistics—mean, standard deviation, skewness, lag-1 autocorrelation, minimum, and maximum—along with observed value as horizontal line. The mean and standard deviations are captured precisely while values of maximum and minimum are simulated outside of the range of available data, a feature important for planning studies which can identify limitations of existing systems and management strategies under conditions not seen within a historic record. Skewness is oversimulated—though the magnitude of the oversimulation is small it is likely an artifact of the underlying gamma distributions—a trade off for including a lower bound of zero. Figure 5 shows the boxplot of the probability density function (PDF) of simulations from HMC model along with the observed. This too is captured well—nearly identical simulation of the PDF was seen from the basic HMG model (figure not shown). The ability of HM models to simulate prolonged spells (wet and dry) is of interest here as it is one indicator of regime-switching behavior. A view of the run lengths is provided in Figure 6. The frequency of runs of various lengths from the simulations of HMC model is shown as boxplots along with the observed frequency. As can be seen, the simulations capture the distribution of prolonged runs very well. The ability to capture the observed run length distribution is important for water resources system reliability. Simulations from HMG

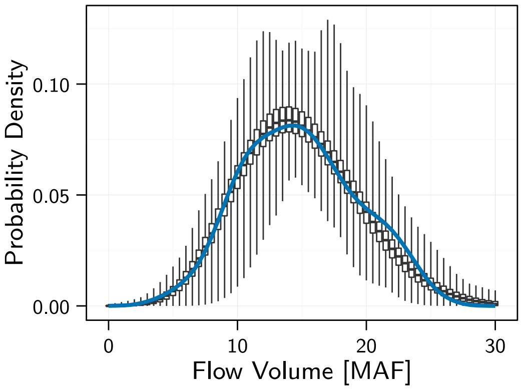

Figure 5. Observed (solid line) and boxplots of simulated PDF with the HMC model.

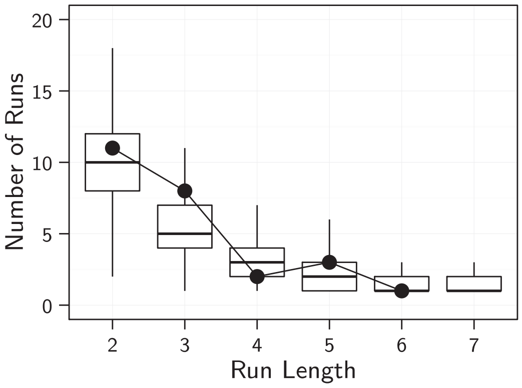

Figure 6. Observed (black points) and simulated (boxplots) run lengths from the HMC model.

also showed similar performance (figure not shown).

The spectral signature of a time series is an important attribute to consider in simulation [*Kwon et al.*, 2007]—as it captures the low-frequency variability and nonstationarity. The wavelet spectrum of Lees Ferry flow series (Figure 7a) exhibits strong variability in the decadal period band in recent years. This nonstationarity is important in characterizing the wet period of 1980s and the ongoing prolonged drought. A nonclimate-informed hidden Markov model such as HMG does not capture the nonstationary behavior of the evolution of the states which imparts spectral nonstationarity. This can be seen in Figure 7b which shows the median wavelet spectrum from the simulations and the power in the decadal band is smeared over the entire time period. We imposed the state sequence from global decoding in HMG—in that, the state sequence from Figure 3a is taken and flow magnitudes are generated from the corresponding gamma distribution. The median wavelet spectrum from these simulations is shown in Figure 7c, which is closer to the spectrum of the historic flow series (Figure 7a). This indicates that the nonstationarity in the spectrum is a result of the sequence of the hidden states, which too is nonstationary and not captured by a stationary Markov transition. Figure 7d shows the median spectrum from HMC simulations, in which the logistic regression model is used to estimate the state probabilities for each year based on the climate indices and,

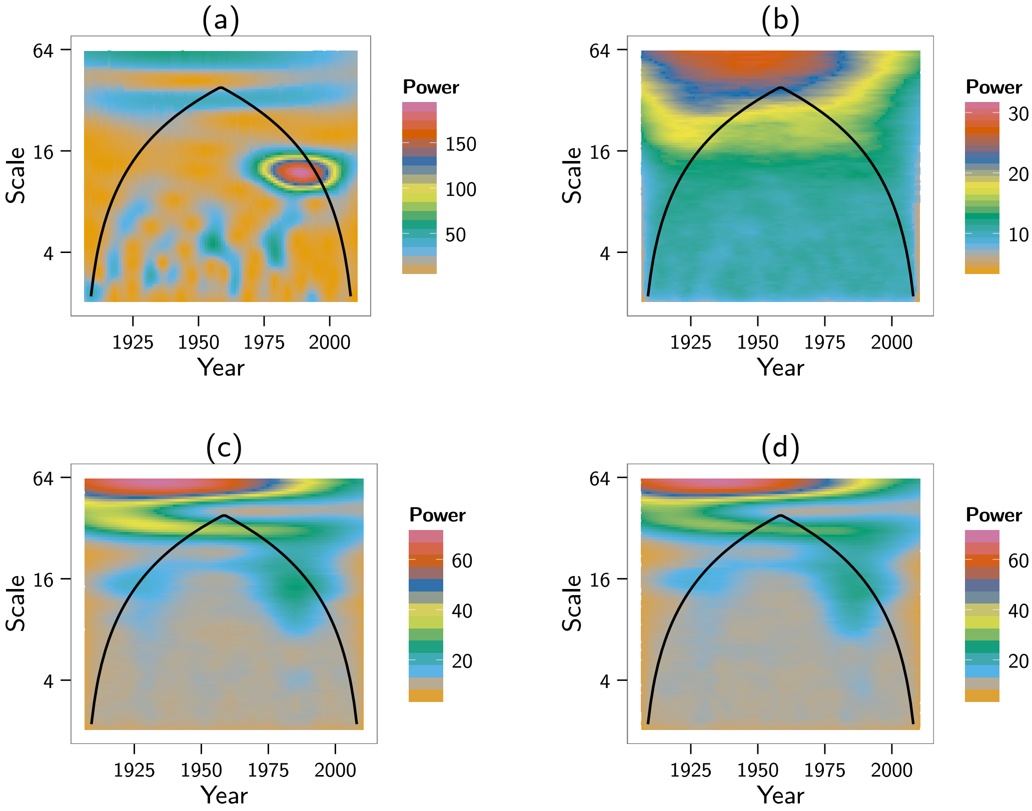

Figure 7. (a) Lees Ferry observed wavelet spectrum, (b) median wavelet spectrum from HMG simulations, (c) median wavelet spectrum when the globally decoded HM states are imposed on the simulations, and (d) median wavelet spectrum from the HMC simulations, incorporating climate information.

consequently the state sequences are generated. The median spectrum from these simulations captures the spectral features of the historic flow (Figure 7a), especially in the decadal period band, very well. The HMC framework, by incorporating climate information, is able to capture the observed spectral nonstationary.

4. Conclusions

We developed a flexible framework for stochastic simulation of flow time series that exhibited regime-like behavior. The method used two components—a gamma component distribution for each hidden state (or regime) from the hidden Markov model and a multinomial logistic regression that incorporates large-scale climate information—to model the nonstationarity in state transitions. The framework was applied to the naturalized flow from the outlet of the Upper Colorado River Basin—a series exhibiting low autocorrelation, regime-switching behavior, and nonstationary decadal variability, attributes not possible to capture using traditional time series methods. The HMC was shown to capture all of the distributional statistics of the data as well as the observed nonstationary decadal variability in the spectrum.

The HMC framework is complementary to spectral-based simulation methods [*Nowak et al*., 2011; *Kwon et al*., 2007] in that it enables the incorporation of large-scale climate information [*Henley et al*., 2011]. The nonstationarity in the state transitions enabled by HMC is akin to simulating from a nonstationary Markov chain [e.g., *Prairie et al*., 2008]. This framework could be readily extended to multiple sites on a river network using either disaggregation or a multivariate HM model [*Hughes and Guttorp*, 1994; *Hughes et al*., 1999; *Mehrotra and Sharma*, 2005; *Mares et al*., 2014]. Similarly, multivariate HM models can be used to model the climate indices and flow jointly. The modeling approach is hierarchical in nature which provides opportunities for Bayesian modeling [*Berliner et al*., 2000; *Thyer et al*., 2009; *Cooley and Sain*, 2010; *Schliep et al*., 2010; *Cooley et al*., 2012]. Extensions of this approach to additional basins and seasonal ensemble forecasting will be helpful for water resources management.

Appendix A: Moments of the Gamma HM

We provide moments and the autocorrelation function for a hidden Markov model of order *m* with gamma component distributions since it is not widely used. Corresponding formulas for normal HM models are given in *Akintuğ and Rasmussen* [2005] and *Frühwirth-Schnatter* [2006]. Let δ denote the stationary distribution of the HM model then:

$$\mathrm{E}(X_t)=\sum_{i=1}^{m}\frac{\delta_i k_i}{\beta_i} \tag{A1}$$

$$\mathrm{Var}(X_t)=\sum_{i=1}^{m}\delta_i\left[a_i^2+\frac{k_i}{\beta_i^2}\right] \tag{A2}$$

where

$$a_i=\frac{k_i}{\beta_i}-\mathrm{E}(X_t) \tag{A3}$$

and

$$\mathrm{Skew}(X_t)=\mathrm{Var}(X_t)^{-3/2}\sum_{i=1}^{m}\delta_i a_i\left[a_i^2+3\left(\frac{k_i}{\beta_i^2}\right)^2\right]. \tag{A4}$$

The autocorrelation function is

$$\rho(k)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\delta_i k_i k_j \gamma_{ij}(k)}{\beta_i\beta_j}-[\mathrm{E}(X_t)]^2}{\mathrm{Var}(X_t)} \tag{A5}$$

where $\gamma_{ij}(k)$ is the *i*, *j* entry of Γ^k.

Acknowledgments
Funding for this research by Bureau of Reclamation and Western Water Assessment RISA program at the University of Colorado is gratefully acknowledged. Data available at: http://bechtel.colorado.edu/bracken/hmm-sim-src/data/.

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control*, *19*(6), 716–723.

Akintuğ, B., and P. F. Rasmussen (2005), A Markov switching model for annual hydrologic time series, *Water Resour. Res.*, *41*, W09424, doi:10.1029/2004WR003605.

Al-Saleh, J. A., and S. K. Agarwal (2007), Finite mixture of gamma distributions: A conjugate prior, *Comput. Stat. Data Anal.*, *51*(9), 4369–4378.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.*, *41*(1), 164–171.

Berliner, L. M., C. K. Wikle, and N. Cressie (2000), Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling: EBSCOhost, *J. Clim.*, *13*(22), 3953.

Boes, D. C., and J. D. Salas (1978), Nonstationarity of the mean and the Hurst phenomenon, *Water Resour. Res.*, *14*(1), 135–143.

Chi, M., E. Neal, and G. K. Young (1973), Practical application of fractional Brownian Motion and noise to synthetic hydrology, *Water Resour. Res.*, *9*(6), 1523–1533.

Colorado River Commission and Coauthors (1923), Colorado River Compact. Government Printing Office Rep. II, 5 pp.

Cooley, D., and S. R. Sain (2010), Spatial hierarchical modeling of precipitation extremes from a regional climate model, *J. Agric. Biol. Environ. Stat.*, *15*(3), 381–402.

Cooley, D., D. Nychka, and P. Naveau (2012), Bayesian spatial modeling of extreme precipitation return levels.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc., Ser. B*, *39*(1), 1–38.

Evin, G., J. Merleau, and L. Perreault (2011), Two-component mixtures of normal, gamma, and Gumbel distributions for hydrological applications, *Water Resour. Res.*, *47*, W08525, doi:10.1029/2010WR010266.

Forney, G. D. (1973), The viterbi algorithm, *Proc. IEEE*, *61*(3), 268–278.

Fortin, V., L. Perrault, and J. Salas (2004), Retrospective analysis and forecasting of streamflows using a shifting level model, *J. Hydrol.*, *296*(1–4), 135–163.

Frühwirth-Schnatter, S. (2006), Finite Markov mixture modeling, in *Finite Mixture and Markov Switching Models*, pp. 301–318, Springer, N. Y.

Gelati, E., H. Madsen, and D. Rosberg (2010), Markov-switching model for nonstationary runoff conditioned on El Niño information, *Water Resour. Res.*, *46*, W02517, doi:10.1029/2009WR007736.

Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona (2005), A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts, *Water Resour. Res.*, *41*, W10410, doi:10.1029/2004WR003467.

Greene, A. M., A. W. Robertson, and S. Kirshner (2008), Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time-scales using a hidden Markov model, *Q. J. R. Meteorol. Soc.*, *134*(633), 875–887.

Greene, A. M., A. W. Robertson, P. Smyth, and S. Triglia (2011), Downscaling projections of Indian monsoon rainfall using a non-homogeneous hidden Markov model, *Q. J. R. Meteorol. Soc.*, *137*(655), 347–359.

Guimarães, R., and E. G. Santos (2011), Principles of stochastic generation of hydrologic time series for reservoir planning and design: A case study, *J. Hydrol. Eng.*, *16*(11), 891–898.

Hastie, T., R. Tibshirani, and Friedman (2002), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics*, 2nd ed., Springer, N. Y.

Henley, B. J., M. A. Thyer, G. Kuczera, and S. W. Franks (2011), Climate-informed stochastic hydrological modeling: Incorporating decadal-scale variability using paleo data, *Water Resour. Res.*, *47*, W11509, doi:10.1029/2010WR010034.

Hidalgo, H. (2003), ENSO and PDO effects on hydroclimatic variations of the Upper Colorado River Basin, *J. Hydrometeorol.*, *4*(1), 5–23.

Hosking, J. R. M. (1984), Modeling persistence in hydrological time series using fractional differencing, *Water Resour. Res.*, *20*(12), 1898–1908.

Hughes, J. P., and P. Guttorp (1994), Incorporating spatial dependence and atmospheric data in a model of precipitation, *J. Appl. Meteorol.*, *33*(12), 1503–1515.

Hughes, J. P., P. Guttorp, and S. P. Charles (1999), A non-homogeneous hidden Markov model for precipitation occurrence, *J. R. Stat. Soc., Ser. C*, *48*(1), 15–30.

Hunter, T., G. Tootle, and T. Piechota (2006), Oceanic-atmospheric variability and western U.S. snowfall, *Geophys. Res. Lett.*, *33*, L13706, doi:10.1029/2006GL026600.

Hurst, H. E. (1951), Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng.*, *116*, 770–808.

Jackson, B. B. (1975), Markov mixture models for drought lengths, *Water Resour. Res.*, *11*(1), 64–74.

Khalil, A. F., H.-H. Kwon, U. Lall, and Y. H. Kaheil (2010), Predictive downscaling based on non-homogeneous hidden Markov models, *Hydrol. Sci. J.*, *55*(3), 333–350.

Koutsoyiannis, D. (2002), The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrol. Sci. J.*, *47*(4), 573–595.

Koutsoyiannis, D., and A. Montanari (2007), Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resour. Res.*, *43*, W05429, doi:10.1029/2006WR005592.

Kwon, H.-H., U. Lall, and A. F. Khalil (2007), Stochastic simulation model for nonstationary time series using an autoregressive wavelet decomposition: Applications to rainfall and temperature, *Water Resour. Res.*, *43*, W05407, doi:10.1029/2006WR005258.

Kwon, H.-H., U. Lall, and J. Obeysekera (2008), Simulation of daily rainfall scenarios with interannual and multidecadal climate cycles for South Florida, *Stochastic Environ. Res. Risk Assess.*, *23*(7), 879–896.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, *32*(3), 679–693.

Lohre, M., P. Sibbertsen, and T. Könning (2003), Modeling water flow of the Rhine River using seasonal long memory, *Water Resour. Res.*, *39*(5), 1132, doi:10.1029/2002WR001697.

Mares, C., I. Mares, H. Huebener, M. Mihailescu, U. Cubasch, and P. Stanciu (2014), A hidden Markov model applied to the daily spring precipitation over the Danube basin, *Adv. Meteorol.*, *0142*(1), 1–11.

McCabe, G., J. Betancourt, and H. Hidalgo (2007), Associations of decadal to multidecadal sea-surface temperature variability with upper Colorado River Flow, *J. Am. Water Resour. Assoc.*, *43*, 183–192, doi:10.1111/j.1752-1688.2007.00015.x.

McCullagh, J., and P. Nelder (1989), *Generalized Linear Models*, 2nd ed., Chapman and Hall, Boca Raton, Fla.

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

Mesa, O. J., and G. Poveda (1993), The Hurst effect: The scale of fluctuation approach, *Water Resour. Res.*, *29*(12), 3995–4002.

Nash, L., and P. Gleick (1991), Sensitivity of streamflow in the Colorado basin to climatic changes, *J. Hydrol.*, *125*(3–4), 221–241.

Nowak, K., M. Hoerling, B. Rajagopalan, and E. Zagona (2012), Colorado River Basin hydroclimatic variability, *J. Clim.*, *25*(12), 4389–4403.

Nowak, K. C., B. Rajagopalan, and E. Zagona (2011), Wavelet auto-regressive method (WARM) for multi-site streamflow simulation of data with non-stationary spectra, *J. Hydrol.*, *410*(1), 1–12.

Ouarda, T., J. Labadie, and D. Fontane (1997), Indexed sequential hydrologic modeling for hydropower capacity estimation, *J. Am. Water Resour. Assoc.*, *33*, 1337–1349.

Piechota, T. C., and J. A. Dracup (1996), Drought and regional hydrologic variation in the United States: Associations with the El Niño-Southern Oscillation, *Water Resour. Res.*, *32*(5), 1359–1373.

Prairie, J., K. Nowak, B. Rajagopalan, U. Lall, and T. Fulp (2008), A stochastic nonparametric approach for streamflow generation combining observational and paleoreconstructed data, *Water Resour. Res.*, *44*, W06423, doi:10.1029/2007WR006684.

Robertson, A. W., S. Kirshner, and P. Smyth (2004), Downscaling of daily rainfall occurrence over Northeast Brazil using a hidden Markov model, *J. Clim.*, *17*(22), 4407–4424.

Robinson, P. M. (1994), Semiparametric analysis of long-memory time series, *Ann. Stat.*, *22*(01), 515–539.

Salas, J., and D. Boes (1980), Shifting level modelling of hydrologic series, *Adv. Water Resour.*, *3*(2), 59–63.

Salas, J. D., J. D. A. Yevjevich, and W. L. Lane (1980), *Applied Modeling of Hydrologic Time Series*, Water Resour. Publ, Highlands Ranch, Colo.

Schliep, E. M., D. Cooley, S. R. Sain, and J. A. Hoeting (2010), A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling, *Extremes*, *13*(2), 219–239.

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

Stedinger, J. R., and M. R. Taylor (1982), Synthetic streamflow generation: 1. Model verification and validation, *Water Resour. Res.*, *18*(4), 909–918.

Taqqu, M. S., V. Teverovsky, and W. Willinger (1995), Estimators for long-range dependence: An empirical study, *Fractals*, *03*(04), 785–798.

Thyer, M., and G. Kuczera (2000), Modeling long-term persistence in hydroclimatic time series using a hidden state Markov model, *Water Resour. Res.*, *36*(11), 3301–3310.

Thyer, M., and G. Kuczera (2003), A hidden Markov model for modelling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach, *J. Hydrol.*, *275*(1–2), 12–26.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.

Timlsena, J., T. Piechota, G. Tootle, and A. Singh (2009), Associations of interdecadal/interannual climate variability and long-term Colorado River Basin streamflow, *J. Hydrol.*, *365*(3–4), 289–301.

Wiper, M., D. R. Insua, and F. Ruggeri (2001), Mixtures of gamma distributions with applications, *J. Comput. Graphical Stat.*, *10*(3), 440–454.

Yoo, J. H., A. W. Robertson, and I.-S. Kang (2010), Analysis of intraseasonal and interannual variability of the Asian summer monsoon using a hidden Markov model, *J. Clim.*, *23*(20), 5498–5516.

Zucchini, W. (2009), *Hidden Markov Models for Time Series: An Introduction Using R*, CRC Press, Boca Raton, Fla.

Zucchini, W., and P. Guttorp (1991), A hidden Markov model for space-time precipitation, *Water Resour. Res.*, *27*(8), 1917–1923.