AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR018887

A Bayesian hierarchical nonhomogeneous hidden Markov model for multisite streamflow reconstructions

C. Bracken[1,2], B. Rajagopalan[1,3], and C. Woodhouse[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]School of Geography and Development, University of Arizona, Tucson, Arizona, USA

Key Points:
- A Bayesian hierarchical nonhomogeneous hidden Markov model (NHMM) for streamflow reconstruction is developed
- Flow is reconstructed from 1473 to 1997 at 20 sites in the Upper Colorado River Basin using tree-ring covariates
- Results are in agreement with published regression-based reconstructions

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

Citation:
Bracken, C., B. Rajagopalan, and C. Woodhouse (2016), A Bayesian hierarchical nonhomogeneous hidden Markov model for multisite streamflow reconstructions, *Water Resour. Res.*, *52*, 7837–7850, doi:10.1002/2016WR018887.

Received 4 MAR 2016
Accepted 15 SEP 2016
Accepted article online 20 SEP 2016
Published online 12 OCT 2016

Abstract In many complex water supply systems, the next generation of water resources planning models will require simultaneous probabilistic streamflow inputs at multiple locations on an interconnected network. To make use of the valuable multicentury records provided by tree-ring data, reconstruction models must be able to produce appropriate multisite inputs. Existing streamflow reconstruction models typically focus on one site at a time, not addressing intersite dependencies and potentially misrepresenting uncertainty. To this end, we develop a model for multisite streamflow reconstruction with the ability to capture intersite correlations. The proposed model is a hierarchical Bayesian nonhomogeneous hidden Markov model (NHMM). A NHMM is fit to contemporary streamflow at each location using lognormal component distributions. Leading principal components of tree rings are used as covariates to model nonstationary transition probabilities and the parameters of the lognormal component distributions. Spatial dependence between sites is captured with a Gaussian elliptical copula. Parameters of the model are estimated in a fully Bayesian framework, in that marginal posterior distributions of all the parameters are obtained. The model is applied to reconstruct flows at 20 sites in the Upper Colorado River Basin (UCRB) from 1473 to 1906. Many previous reconstructions are available for this basin, making it ideal for testing this new method. The results show some improvements over regression-based methods in terms of validation statistics. Key advantages of the Bayesian NHMM over traditional approaches are a dynamic representation of uncertainty and the ability to make long multisite simulations that capture at-site statistics and spatial correlations between sites.

1. Motivation

Tree-ring based streamflow reconstructions place our knowledge of present water resources systems in a centuries-long context. The rich variety of scenarios provided by reconstructions can be used in planning models to test the limits of existing systems. By utilizing simulated flow input at multiple locations on a river network, planning models allow water managers the opportunity to examine how systems will perform in scenarios which have not been observed in the historical record. As the next generation of planning models develops the ability to process probabilistic inputs, reconstruction models must be able to produce these inputs. Moreover, reconstruction models need to be able to produce spatially correlated inputs suitable for planning models operating on interconnected river networks. Multisite streamflow reconstructions from tree rings require the following properties:

1. Spatial correlation - Intersite correlation should be accurately captured in order to correctly portray the simultaneous behavior of locations on a network.
2. Realistic representation of uncertainty - Reconstructions should reflect the uncertainty inherent in any hydrologic system.
3. Rich variety of hydrologic sequences - Simulated sequences should represent the statistical properties of the historical record but should be able to produce unobserved scenarios.

Despite considerable methodological improvements in the field of streamflow reconstruction using tree-rings, previous studies have commonly focused on reconstructing a single site at a time. This does not explicitly account for the intersite correlations, although the correlations are sometimes captured as a

consequence of shared predictor tree-ring chronologies among the reconstructions. A standard approach to simulating flow on river networks is to simulate flows at a downstream gage (say the outlet to a basin) using the observed and paleo reconstructions [*Prairie et al*., 2008] and, subsequently, stochastically disaggregate them to all upstream sites [*Nowak et al*., 2010]. This approach would be obviated, however, if reconstructions were available at all the locations with the attendant uncertainties and intersite correlations––one motivation for this study. Applications of flow reconstruction models which explicitly capture intersite correlations have been limited. To our knowledge the only such model is that of *Devineni et al*. [2013] who used a Bayesian hierarchical regression approach to reconstruct flows in the Delaware River Basin, which is relatively small compared to the Upper Colorado River Basin (UCRB).

Bayesian models can naturally propagate uncertainty in parameter estimates to generate probabilistic reconstructions [*Devineni et al*., 2013]. If planning models are capable of incorporating such inputs, uncertainty can then be represented in terms of system variables that are familiar to water managers, enabling fully informed decision making. The Bayesian approach taken in this study provides a rich variety of streamflow reconstructions and subsequently can provide an improved representation of uncertainty for variables of interest for planning.

In this study, we propose a multisite streamflow reconstruction model which satisfies the necessary properties listed above. We take advantage of the set of high quality tree-ring chronologies in the UCRB to propose a Bayesian nonhomogeneous hidden Markov model (NHMM) capable of reconstructing streamflow at multiple sites with robust estimates of uncertainties while maintaining intersite correlations.

1.1. Reconstructions on the Upper Colorado River

The UCRB is a critically important surface water supply for the much of the southwestern US. In order to understand the long-term variability of the basin's flows, including periods of drought as well as pluvials, tree-ring data have been used to extend the observed gage values back in time. The earliest calibrated reconstruction of streamflow ever developed was for the UCRB at the Lees Ferry gauge along with 11 tributaries, in 1976 [*Jacoby and Stockton*, 1976], which provided the first opportunity to assess the gauge record in a long-term context. Stockton and Jacoby's Lees Ferry reconstruction showed that the period upon which the 1922 Colorado River Compact was based was the wettest period since at least 1520, and also indicated droughts had occurred in the past that were worse than those in the modern period. This reconstruction was followed by several efforts which used the same tree-ring data but different statistical and data treatment approaches [e.g., *Michaelsen et al*., 1990; *Hidalgo et al*., 2000] to reconstruct upper basin flows. A renewed interest in the nature of past droughts in the UCRB resulted in new data collections over the past several decades [*Woodhouse and Lukas*, 2006], which have enabled an updating of the earlier reconstructions, as well as extensions back in time with the use of remnant wood preserved on the landscape [e.g., *Woodhouse et al*., 2006; *Meko et al*., 2007]. Tree-ring data collections have also expanded across the UCRB, allowing additional reconstructions of tributary flows [e.g., *Woodhouse et al*., 2006; *Barnett et al*., 2010; *Gray et al*., 2011].

The suitability of the tree-ring data in this region for reconstructions of water year streamflow has made skillful reconstructions possible. Lower to midelevation conifer species here––*Pinus edulis*, *Pinus pondersa*, and *Pseudotsuga menziseii*––are highly sensitive to variations in moisture [*Woodhouse and Lukas*, 2006]. The association between annual growth of these conifers and winter precipitation is among the strongest in the world [*St George and Ault*, 2014]. Growth is highly dependent on winter precipitation, which influences the spring and early summer soil moisture important for the onset of growth [*St George and Ault*, 2014], while winter precipitation is also the primary driver of water year runoff in the UCRB [*Nowak et al*., 2012; *Vano et al*., 2012]. As a consequence, annual tree growth is well correlated with water year streamflow, providing a suitable paleohydrologic proxy for past streamflow [*Woodhouse and Lukas*, 2006]. Because of the close correspondence between annual tree growth and water year flows, the UCRB has become a test bed for the development of innovative reconstruction techniques. Traditional reconstruction approaches [*Woodhouse et al*., 2006; *Meko et al*., 2007] have been regression-based, but new approaches [*Gangopadhyay et al*., 2009, 2015] have also produced skillful reconstructions, with additional benefits. Regression methods, though widely used, do not have a dynamic representation of uncertainty, i.e., uncertainty bounds are constant in every year. It seems reasonable to expect that certain flow regimes are more difficult to reconstruct than others, particularly high flow years, and a static uncertainty representation cannot capture this. Planning

models require long ensemble sequences of streamflow representing a wide variety of system behaviors while capturing observed statistics. From the regression models, ensembles are generated for each year independently based on the uncertainty bounds, as a result they may not capture the year-to-year correlation of the sequence without an additional step in the reconstruction process. *Gangopadhyay et al*. [2009] used a nonparametric K-nearest neighbor (KNN) resampling approach, gaining the ability to simulate long flow series with a dynamic uncertainty representation. A potential drawback in the simulations from a KNN approach are that values higher (lower) than the maximum (minimum) observed flow will never be generated. More recently, *Gangopadhyay et al.* [2015] built on this nonparametric approach to reconstruct inputs for a simple water balance model [*McCabe and Wolock*, 2011], enabling the reconstruction of suite of hydroclimatic variables for the UCRB.

This paper is organized as follows. Section 2 describes the model structure including procedures for implementation, section 3 describes the data for the UCRB, while section 4 describes the the calibration and reconstruction steps from the fitted best model, along with model validation. Section 5 presents the results and model diagnostics and section 6 contains summary and discussion.

2. Model Structure

The structure of the multisite Bayesian hierarchical NHMM consists of the following components: (i) a *marginal site model* in the form of a NHMM fitted to contemporary data at each site with temporal variability of model parameters modeled as a function of a set of covariates, (ii) a *spatial model* in the form of a Gaussian elliptical copula for capturing the multisite joint probability distribution and consequently the intersite correlations, (iii) the *likelihood function* and its computation for the hierarchical model and (iv) description of *prior distributions* for the model parameters. These components are described below. The best model selection procedure and model implementation are described after the model description. This procedure is applied to identify the optimal number of states in the NHMM and the optimal set of covariates.

2.1. Marginal Site Model––NHMM

The marginal site model is a nonhomogeneous hidden Markov model (NHMM) fitted to observed flow time series at each site. Hidden Markov models have a long history in hydrology [*Hughes and Guttorp*, 1994; *Hughes et al*., 1999; *Robertson et al*., 2004; *Mehrotra and Sharma*, 2005; *Ak intuğ and Rasmussen*, 2005; *Bracken et al*., 2014]. A hidden Markov model is composed of a small number of unobserved (latent) states representing the state of the climate. Each state is represented by a component probability distribution. In each year, the system switches randomly between the hidden states sampling from the corresponding component distribution to generate observed values. The so-called transition probability matrix (tpm) determines the chance of switching to each state given the current state. In a homogeneous hidden Markov model, the tpm is constant in time as opposed to a nonhomogeneous version in which the transition probabilities are allowed to vary in time according to some covariates.

Consider a random process Y_{st} for $s=1,\ldots,S$ sites and $t=1,\ldots,T$ years. The observed values y_{st} are assumed to depend on a latent nonhomogeneous discrete time Markov process Z_{st} with state space $\mathbf{K}=\{1,\ldots,K\}$ where K is the number of latent states. At each site, the hidden state switches independently of other sites according to a first order Markov process. Let $\mathbf{x}_t$ be a vector of $P+1$ covariate values at time t, the first order Markov assumption implies

$$P(Z_{st}|Z_{s1},\ldots,Z_{s(t-1)},\mathbf{x}_1,\ldots,\mathbf{x}_T)=P(Z_{st}|Z_{s(t-1)},\mathbf{x}_t)$$

The marginal data model for K states and $p=1,\ldots,P$ covariates is

$$P(Y_{st}=y_{st}|Z_{st}=z_{st},\mathbf{x}_t=\mathbf{x}_t)=\text{LogNormal}(\mu_{sk}+\mathbf{w}_{sk}\mathbf{x}_t,\sigma_{sk})$$

where Y_{st} is the random variable representing annual flow in year t at site s, $Z_{st}\in\mathbf{K}$ is the latent (hidden) state, μ_{sk} is the state dependent intercept term at site s for state k, $\mathbf{w}_{sk}$ is a $1\times q+1$ vector of coefficients (including an intercept), $\mathbf{x}_t$ is a $1\times q+1$ vector of covariates and σ_{sk} is the state dependent scale parameter at site s for state k.

The lognormal distribution is a natural distributional choice for two main reasons: (1) It is defined only for positive values making it a natural fit for modeling streamflow data, and (2) the location parameter has no

theoretical bounds, a convenient property for modeling purposes. Another reasonable choice would be the gamma distribution, though it lacks the convenience of property 2.

The transition probabilities are modeled using a multinomial logistic link function

$$P(Z_{st}=i|Z_{s(t-1)}=j,\mathbf{x}_t=\mathbf{x}_t)=p_{ij}^{(st)}=\frac{\exp\left(\boldsymbol{\beta}_{ij}^{(s)}\mathbf{x}_{st}\right)}{\sum_{\ell=1}^{K}\exp\left(\boldsymbol{\beta}_{i\ell}^{(s)}\mathbf{x}_{st}\right)} \tag{1}$$

where $p_{ij}^{(st)}$ is the probability of transitioning from state i to state j at time t for site s, $\boldsymbol{\beta}_{ij}$ is a $1\times Q+1$ vector of coefficients. In order for the parameters to be identified, one $\boldsymbol{\beta}_{ij}$ per row of the transition probability matrix needs to be set to a $1\times Q+1$ vector of zeros. In the case of 2 states, the multinomial logistic link function simplifies to a logistic link function. For the special case of $t=1$

$$P(Z_{s1}=i)=1/K$$

which implies we do not fit the initial state distribution as an extra parameter. For a Kth order model for S sites, the NHMM has $S[K^2(Q+1)+2K]$ parameters. Note that the covariates used in the data layer and the transition layer can be different but we use the same covariates for both.

2.2. Spatial Model

Let $\mathbf{Y}_t=(Y_{1t},\ldots,Y_{St})$ be a vector of flow random variables at time t with corresponding realizations $\mathbf{y}_t=(y_{1t},\ldots,y_{St})$. The joint distribution is assumed to be represented by a Gaussian elliptical copula which allows for any marginal distribution to be specified while maintaining joint normality after transformation. The Gaussian copula constructs the joint cumulative distribution function (cdf) of $\mathbf{Y}_t$ as

$$F_C(\mathbf{y}_t)=\Phi_\Sigma(\mathbf{u}_t) \tag{2}$$

where $\Phi_\Sigma(\mathbf{u}_t)$ is the joint cdf of an S-dimensional multivariate normal distribution with covariance matrix Σ, $\mathbf{u}_t=(u_{1t},\ldots.u_{St})$, $u_{it}=\phi^{-1}(F_{it}[y_{it}])$ where ϕ is the cdf of the standard normal distribution and F_{it} is the marginal cdf at site i and time t. The corresponding joint probability distribution function (pdf) is

$$f_C(\mathbf{y}_t)=\frac{\prod_{i=1}^{S}f_{it}[y_{it}]}{\prod_{i=1}^{S}\psi[u_{it}]}\Psi_\Sigma(\mathbf{u}_t) \tag{3}$$

where f_{it} is the marginal lognormal mixture pdf at site i and time t, ψ is the standard normal pdf and Ψ_Σ is the joint pdf of an S-dimensional multivariate normal distribution.

The dependence between sites is assumed to be a function of distance [*Renard*, 2011]. The dependence matrix is constructed with a simple exponential model

$$\Sigma(i,j)=\exp\left(-||\mathbf{s}_i-\mathbf{s}_j||/a_0\right) \tag{4}$$

where a_0 is the copula range parameter. The copula range parameter controls the distance over which sites are correlated, a very large value of a_0 indicates stations far apart are highly correlated and a very small value of a_0 indicates sites are independent. Note that the values in this dependence matrix are not covariances, so by analogy with the variogram, the dependence model is termed the dependogram [*Renard*, 2011].

2.3. Likelihood Computation

The marginal data likelihood is computed with the so called forward algorithm for HMMs [*Zucchini*, 2009]. Let $\boldsymbol{\delta}$ be a K-vector whose elements are all equal to $1/K$, let $\mathbf{P}_{st}$ be a $K\times K$ diagonal matrix whose diagonal elements are $P(y_{st}|z_{st}=k,\mathbf{x}_t),k=1,..,K$ and let $\Gamma_{st}=[p_{ij}^{(st)}]_{i,j=1}^{K}$. In matrix form, the likelihood function for the NHMM is the product of the likelihood at each of the S sites

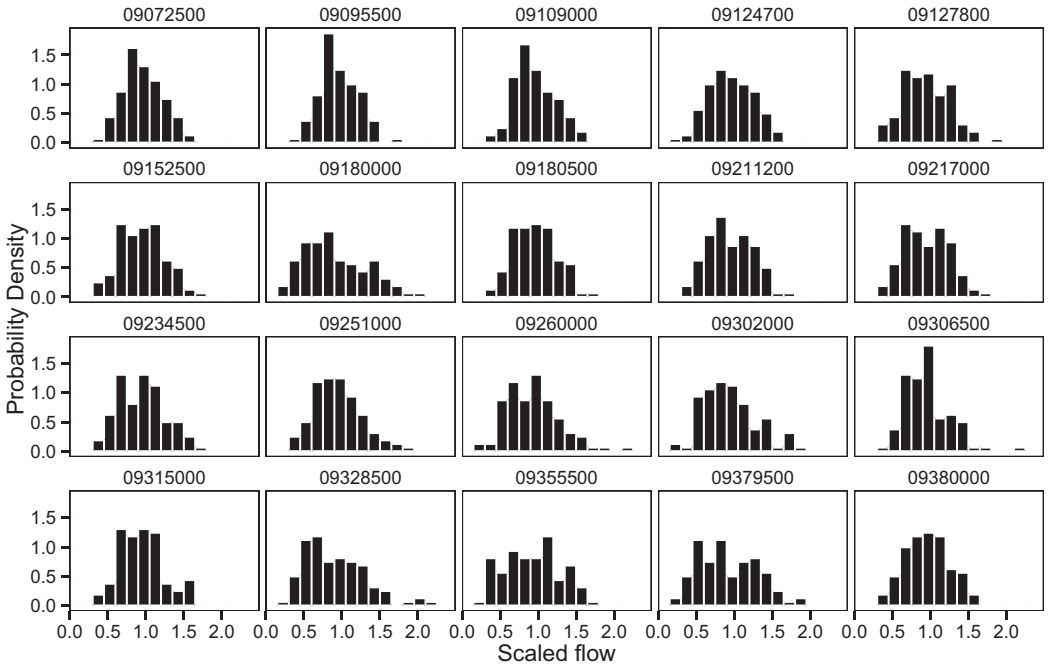

Figure 1. Histograms of scaled naturalized flow data at 20 sites in the upper Colorado river basin.

$$L_{\text{NHMM}} = \boldsymbol{\delta}\left(\prod_{s=1}^{S} \mathbf{P}_{s1}\Gamma_{s2}\mathbf{P}_{s2}\cdots\Gamma_{sT}\mathbf{P}_{sT}\right)\mathbf{1}'. \tag{5}$$

The copula likelihood is a product of independent copula densities in each year using equation (3)

$$L_{Cop} = \prod_{t=1}^{T} f_C(\mathbf{y}_t). \tag{6}$$

The full likelihood is a product of the copula likelihood and the NHMM likelihood

$$L = L_{\text{NHMM}} L_{Cop}. \tag{7}$$

For comparison purposes, it is useful to fit the NHMM without the copula. In this case the full likelihood is simply L_{NHMM}. The model in this case is equivalent to fitting a single-site NHMM individually for each station. The model behavior with and without the copula is discussed in the results section.

2.4. Priors

We use weakly informative $N(0, 1)$ priors for $\mathbf{w}_{sk}$, μ_{sk} and σ_{sk} knowing that the scale of the parameters will all be approximately between -1 and 1 (because of the scaled flow data). For a_0 we use $N(d, d/3)$ where d is the maximum distance between sites.

The multinomial logistic function has the property $f(\mathbf{x}) = f(a\mathbf{x})$ where a is any scalar. As such, $\boldsymbol{\beta}_{ij}$ can be multiplied by any constant and still produce the same transition probabilities. In practice if very diffuse priors are set, the values of $\boldsymbol{\beta}_{ij}$ can become arbitrarily large, potentially introducing numerical errors. It becomes necessary to control the scale of $\boldsymbol{\beta}_{ij}$ via priors. In this case, we set the priors to be $N(0, 0.1)$ to keep these parameters on the same scale as all other model parameters which are approximately in the range $[-1, 1]$ due to scaling of the data.

2.5. Implementation

The model was implemented in the Stan modeling language [*Stan Development Team*, 2015a] using the RStan interface [*Stan Development Team*, 2015b]. Stan uses the No-U-Turn Sampler (NUTS), an implementation of Hamiltonian Monte Carlo (HMC) [*Hoffman and Gelman*, 2014]. The NUTS sampler deals well with highly correlated parameters, tends to need very few warmup iterations and typically produces nearly

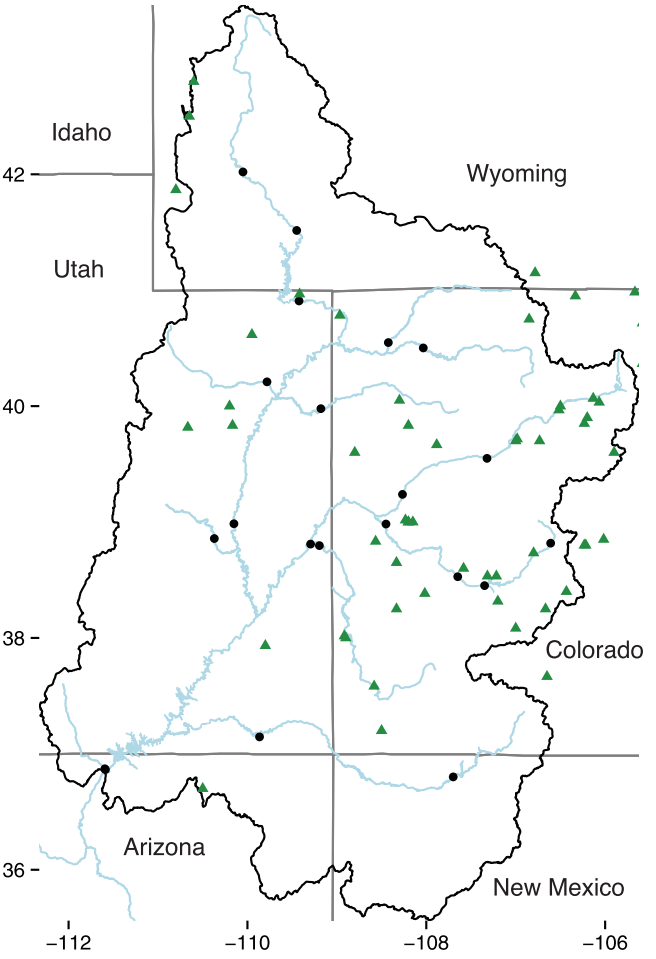

Figure 2. Map of the Upper Colorado River basin, with flow sites (black dot) and tree ring sites (green triangles).

uncorrelated samples. For these reasons, very long chains are usually not needed, nor is thinning. The tradeoff in using the NUTS sampler in this application was much longer computation time per sample compared to a traditional Metropolis-Hastings or Gibbs sampler.

Three chains of length 3,000 were run, with the first 1,000 iterations discarded as warmup, resulting in 6,000 samples for each parameter. To assess convergence, we compute the $\hat{R}$ statistic [*Gelman and Rubin*, 1992] and visually inspect trace plots. All chains for which the maximum $\hat{R}$ value was above 1.1 were discarded. Depending on the model order, chains took from a few hours to several days to run on a 8-core 2.6 GHz Intel Core i7 processor.

3. Data

Naturalized water year (October–September) flow data at 20 sites on the Upper Colorado River for the period 1906–1997 were provided by the Bureau of Reclamation. The flow data at each site were scaled by dividing by the respective standard deviations. Figure 1 shows histograms of the scaled data.

We obtained tree-ring chronologies, previously used in UCRB flow reconstructions, from *Meko et al.* [2007] and *Woodhouse et al.* [2006]. Moisture-sensitive chronologies, from *Pinus edulis*, *Pinus pondersa*, *Pinus flexilis*, and *Pseudotsuga menziseii*, are composed of dated and measured ring widths from multiple trees at each chronology site. We used the residual version of the chronologies, with persistence presumed due to biological factors removed. Fifty chronologies with complete records from 1473 to 1997 were included. Figure 2 shows the Upper Colorado River Basin with locations of the tree-ring chronologies and the natural flow sites.

Principal Component Analysis was performed on the 50 tree-ring chronologies and the first three principal components (PCs) were found to explain 58% of the variance and thus were considered as potential predictors.

For a comparison of the reconstructions resulting from the Bayesian NHMM approach with published regression-based reconstructions, we collected all of the available existing reconstructions corresponding to USBR natural flow nodes in the Upper Colorado River Basin above Lees Ferry. In total there are 15 available reconstructions; 10 from *Woodhouse et al.* [2006]; 2 from *Barnett et al.* [2010]; 3 from *Gray et al.* [2011]. Multiple reconstructions of Lees Ferry flow have been generated; we use the most recent, from *Meko et al.* [2007]. Table 1 shows a list of natural flow gage names and citations for previous reconstructions, the data from these reconstructions can be obtained from http://treeflow.info.

4. Reconstruction, Calibration, and Validation

In order to reconstruct past streamflow based on covariate information, a calibration period is required in order to train the model. The full calibration period is 1906–1997, but for validation, models were calibrated

Table 1. Natural Flow Gage Names and IDs and Citations for Previous Reconstructions

Gage Name	Gage ID	Previous Reconstruction
Colorado River At Glenwood Springs, CO	09072500	*Woodhouse* [2006]
Colorado River Near Cameo, CO	09095500	
Taylor River Below Taylor Park Res., CO	09109000	
Gunnison River Above Blue Mesa Res., CO	09124700	
Gunnison River At Crystal Reservoir, CO	09127800	*Woodhouse* [2006]
Gunnison River Near Grand Junction, CO	09152500	*Woodhouse* [2006]
Dolores River Near Cisco, UT	09180000	*Woodhouse* [2006]
Colorado River Near Cisco, UT	09180500	*Woodhouse* [2006]
Green River Below Fontenelle Res., WY	09211200	*Barnett* [2010]
Green River Near Green River, WY	09217000	*Woodhouse* [2006]
Green River Near Greendale, UT	09234500	*Barnett* [2010]
Yampa River Near Maybell, CO	09251000	*Gray* [2011]
Little Snake River Near Lily, CO	09260000	*Gray* [2011]
Duchesne River Near Randlett, UT	09302000	
White River Near Watson, UT	09306500	*Gray* [2011]
Green River At Green River, UT	09315000	*Woodhouse* [2006]
San Rafael River Near Green River, UT	09328500	
San Juan River Near Archuleta, NM	09355500	*Woodhouse* [2006]
San Juan River Near Bluff, UT	09379500	*Woodhouse* [2006]
Colorado River At Lees Ferry, AZ	09380000	*Hidalgo* [2000] *Woodhouse* [2006] *Meko* [2007] *Gangopadhyay* [2009]

on the years 1952–1997 and validated on the years 1906–1951. Training the model results in a collection of samples from the posterior distribution. To reconstruct, a NHMM simulation is conducted using the parameter values from a single posterior sample and the tree-ring covariate data. Each simulation from the NHMM is known as a realization or trace. This process is repeated for each posterior sample, propagating the uncertainty present in the parameter estimates to the reconstructions.

The full reconstruction procedure is as follows:

1. Select a posterior sample for each parameter;
2. With equal probability, choose a random initial state for each site;
3. Draw one sample from the Gaussian copula with a dependence matrix based on the current sample of the copula range parameter;
4. Based on the current state for each site, transform the copula sample to the current component distribution for each site using the inverse lognormal function;
5. Transition to a new state for each site based on the transition probability matrix (computed from equation (1)) for the current year;
6. Repeat steps 3–5 for all reconstruction years;
7. Repeat steps 1–6 for each posterior sample;
8. Repeat steps 1–7 for each posterior chain.

4.1. Model Selection

Model selection was conducted using the DIC (deviance information criterion) [*Spiegelhalter et al.*, 2002] and the BIC (Bayesian information criterion) [*Schwarz*, 1978]. DIC is more widely used for Bayesian models but BIC tends to favor a more parsimonious model so we computed both metrics. BIC and DIC were computed for model orders from 2 to 11 including 1–3 PCs. Model selection is conducted using models fit on the entire historical record.

4.2. Validation

A validation run was conducted (using the optimal NHMM model) in which the model was fit to the half data (1952–1997) and tested on the other half of the data, not used in the model calibration (1906–1951). The reduction of error (RE) statistic was used to assess goodness of fit over the validation period [*Fritts et al.*, 1990]. Positive values of RE indicate skill above the climatology of the calibration period with higher values of RE indicating

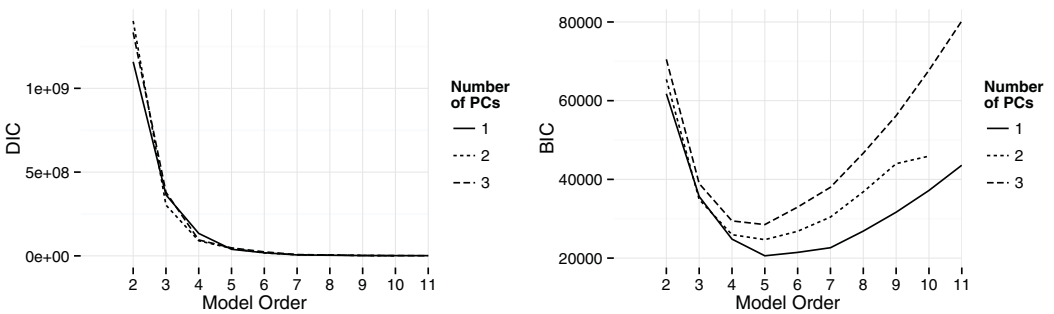

Figure 3. Plots of (left) DIC and (right) BIC for model orders 2–11 and with 1–3 PCs.

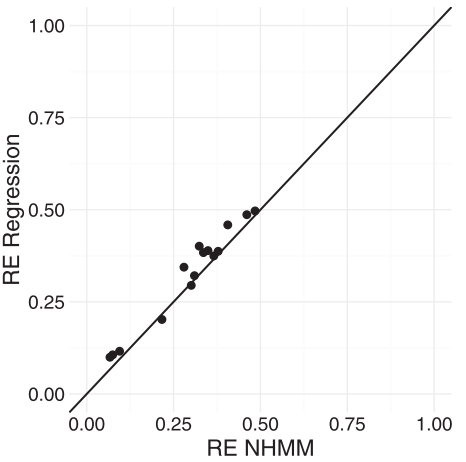

Figure 4. Reduction of Error (RE) statistic for the NHMM versus regression during the validation period (1906–1951). Both models are trained on data from 1952 to 1997 using identical predictors.

better correspondence with observations. Statistics (max, min, mean and standard deviaiton) were also computed for the validation period and the reconstruction period using the validation models (trained on half the data) and compared to that of the observed.

Published UCRB reconstructions use single-site regression models which select the best set of predictors appropriate for each site, while the NHMM uses the same set of predictors at all the sites. In order to enable a consistent comparison with regression-based estimates over this period, linear regression models were fitted at each site using the same suite of predictors as the NHMM and over the same period of time (1952–1997). The RE was computed at each site for the 1906–1951 period from the individual regression models and compared with that from NHMM. Furthermore, statistics from the best fit regression models over the reconstruction period 1473–1905 were compared with those from NHMM.

4.3. Label Switching

Label switching occurs in mixture models and HMMs when state labels (1, 2, etc.) randomly switch during the fitting procedure [*Jasra et al*., 2005]. For example in the single site NHMM in section 2.1, if 2 states are used, the system switches between 2 probability distributions at each time step with some predetermined transition probability. The label switching issue occurs because assigning labels of 1 or 2 to each distribution is arbitrary. In practice, the labels may randomly switch in two subsequent model fits. In a Bayesian setting, this switching occurs when identical priors are used for state specific parameters such that the posterior is unchanged when state labels are permuted. There is no general way to alleviate this issue and typically this is remedied by problem specific postprocessing or priors. This issue is especially prevalent when using Bayesian fitting procedures since the state labels can switch within and between chains.

In our application label switching did occur. Label switching was overcome by reconstructing (simulating) from each chain individually, only combining the results after simulation.

5. Results

5.1. Model Selection and Validation Results

DIC values (Figure 3, left) converge at around model order 5 or 6 and BIC values show a clear minimum at model order 5 (Figure 3, right). Given this, we choose the best model as one with 5 hidden states and 1 PC.

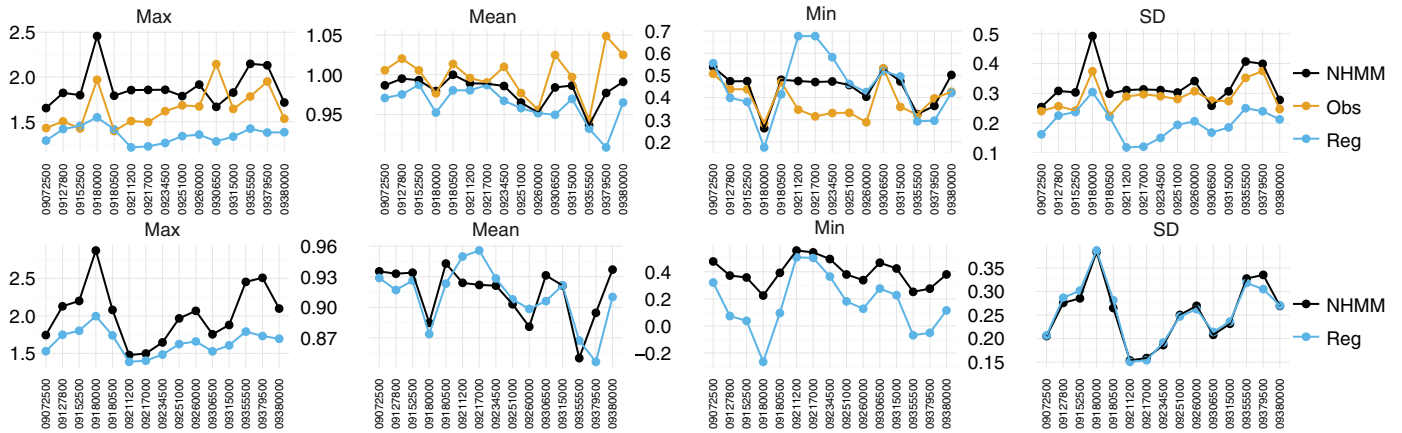

Figure 5. Statistics for each station during (top) the validation period (four plots) and (bottom) the reconstruction period (1473–1905) (four plots) for the median NHMM (Black), fitted Regression (Blue) and observations (Orange). In both cases, models are trained on data from 1952 to 1997. Note the lines are for visual clarity and do not imply a relationship between stations.

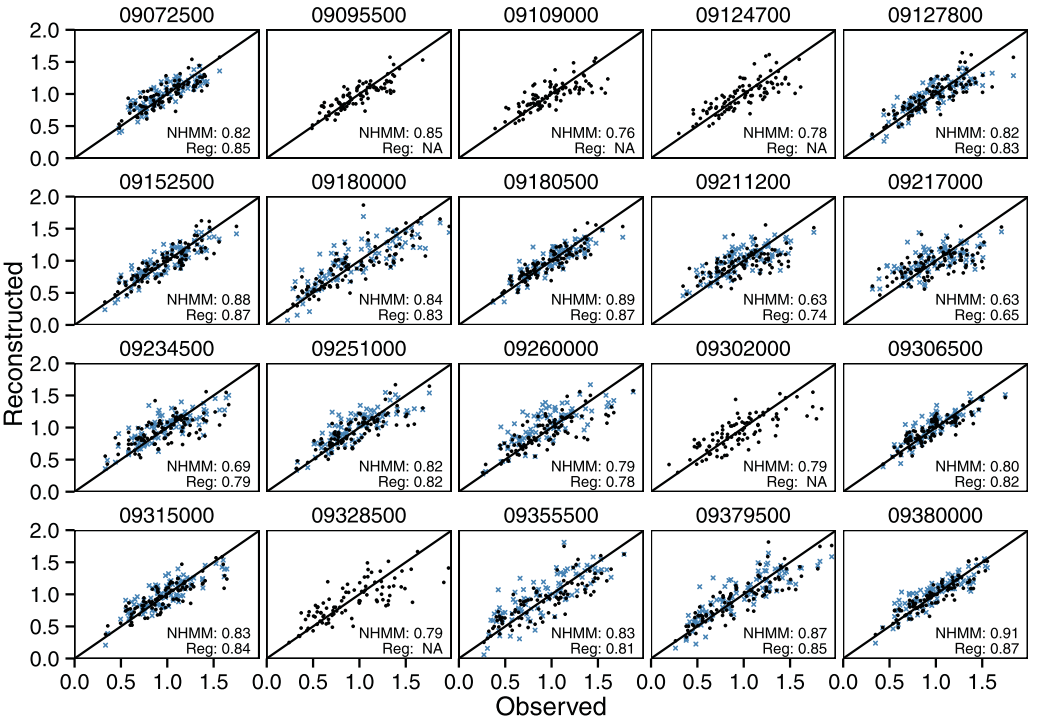

Figure 6. Observed flow plotted against published regression-based reconstructions (blue crosses) and median NHMM reconstructions (black points) for each station. Correlation coefficients for NHMM versus observed, and published regression estimates versus observed, can be found on the bottom right of each panel.

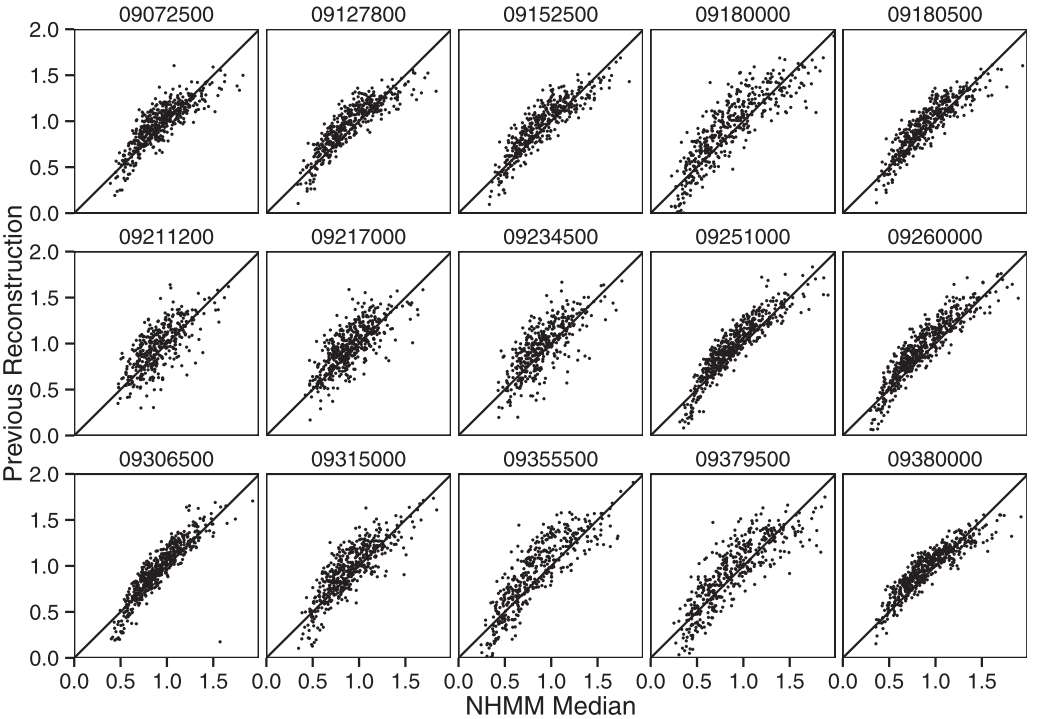

Figure 7. NHMM median plotted against published regression-based reconstructions where available.

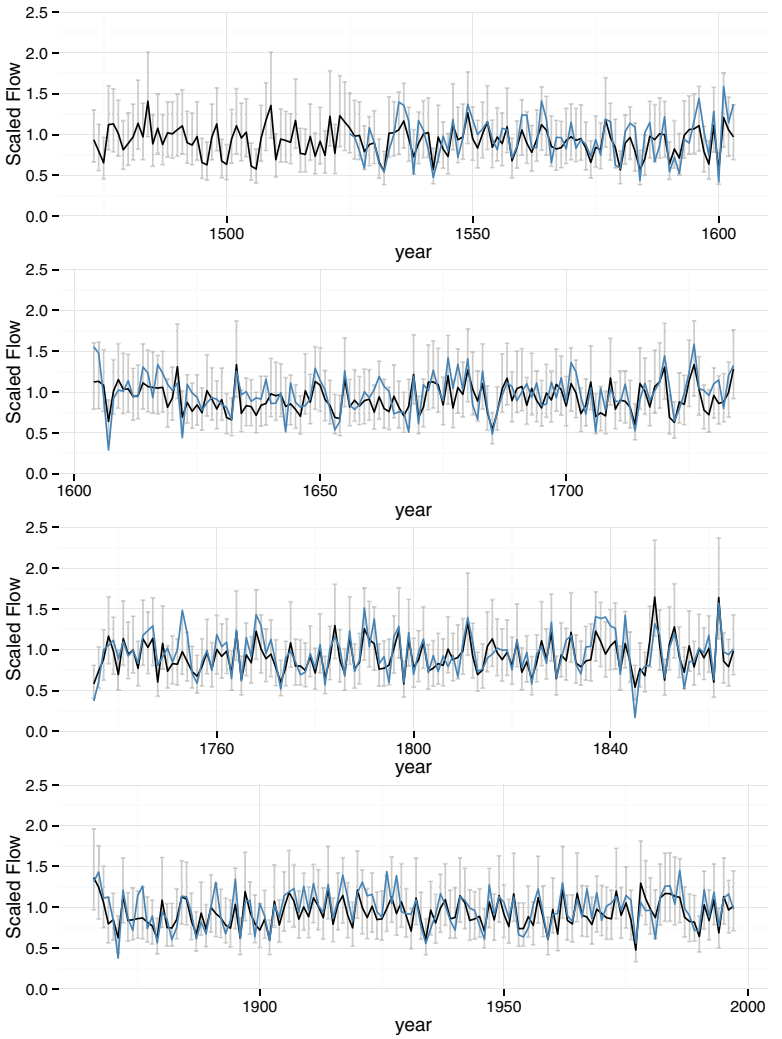

Figure 8. Reconstructions at Lees Ferry, Arizona, gage number 09380000, for the period 1473 - 1997, the NHMM median (black line) with 80% CI in gray and, *Meko et al.* [2007] reconstruction (blue lines).

The RE statistics for the NHMM median versus the fitted regression models are quite similar (Figure 4), indicating both match observations similarly when trained on the same data. The regression models were fitted over the same training period (1952–1997) with the same predictors as the NHMM, in this case the first PC.

Figure 5 shows statistics (max, min, mean and standard deviation) for the validation period (Top 4 boxes) and for the reconstruction period (bottom 4 boxes). The NHMM estimates are shown in black and those from fitted regression in blue along with the observed statistics in orange, during the validation period.

During the validation period, the fitted regression estimates of maximum and standard deviation underestimate the observed while NHMM tends to overestimate. The mean and minimum are comparably captured by both models during in the validation period. In the reconstruction period, 1473–1905 both models perform similarly for the standard deviation and mean. The NHMM tends to estimate higher minimum and maximum values, a reflection of skewed and fat tailed nature of the lognormal distribution used in the NHMM compared to the normal distribution assumed in regression.

5.2. Comparisons With Published Reconstructions

For an additional comparison with observations and published reconstructions, Figure 6 shows the NHMM median for the observation period (1906–1997) versus observations (black points) as well as the mean estimates from the published regressions (see references in Table 1) for the same period (blue crosses). These results are produced from models fitted to the entire observation period. The NHMM and published estimates

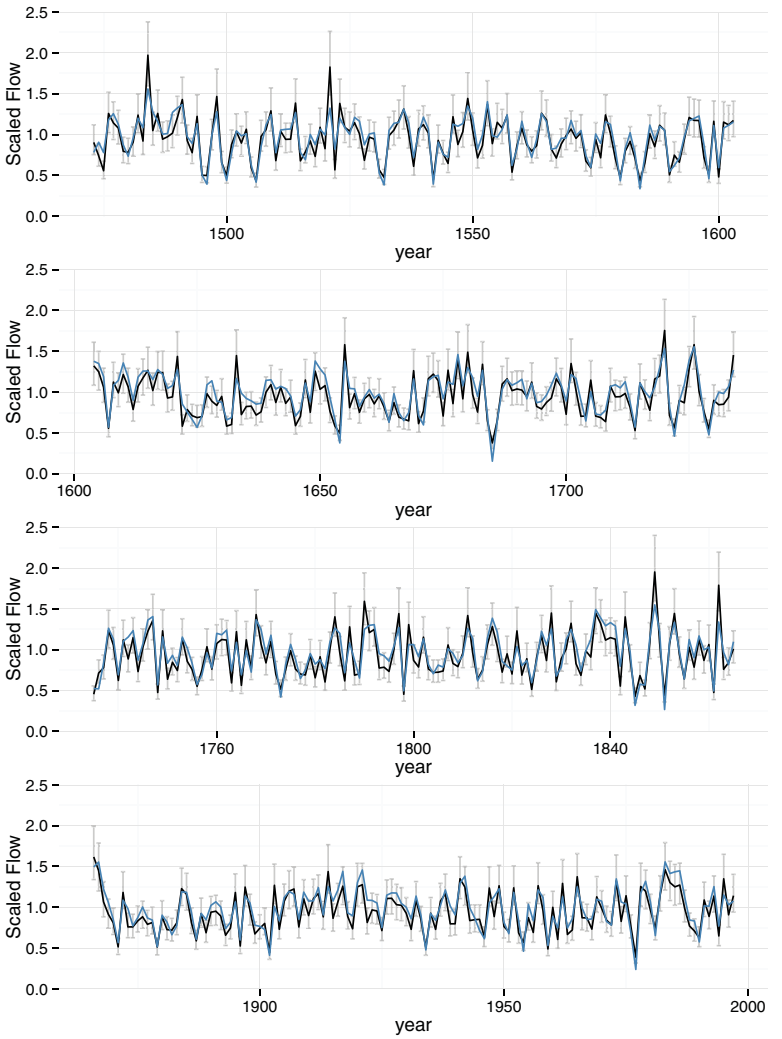

Figure 9. Same as Figure 8 but reconstructions at Green River Near Green River, Wyoming, gage number 09315000.

both compare well to observed flow with correlation coefficients higher than 0.7 (see values in the plots in Figure 6), which is not unexpected given the common predictor pool. Note that the NHMM correlation values are computed from the median of all NHMM simulations, which represent the average behavior of the NHMM."

Figure 7 shows the median of the NHMM reconstructions plotted against 15 published regression-based reconstructions. It can be seen that overall the median reconstructions compare quite well. The NHMM tends to produce higher low flow values, lower middle flow values and higher high flows compared to regression-based reconstructions (see the crescent shaped scatter in Figure 7). We argue that this is due to the log normal component distributions, which have a lower bound of zero and fatter tail than the Normal distribution used in regression-based methods. This property of producing larger high values than regression is beneficial for representing wet extremes in that the NHMM can produce higher unobserved flows. This feature may lead to more accurate detection of pluvials, potentially of interest to water managers. The tendency of the model to produce higher low flow values than regression is a potential drawback, suggesting that it may be useful to use a multimodel approach to capture the strengths in both methods.

To additionally compare the NHMM reconstructions with previous efforts, we show the median of the NHMM reconstructions (Figures 8 and 9) versus regression-based reconstructions from *Meko et al*. [2007] and *Woodhouse et al*. [2006] at two key location on the UCRB – Lees Ferry and Green River Near Green River, WY. The black lines show the median of the NHMM reconstructions, with the 80% credible interval (CI) in grey. Blue lines show the previous regression-based reconstructions respectively.

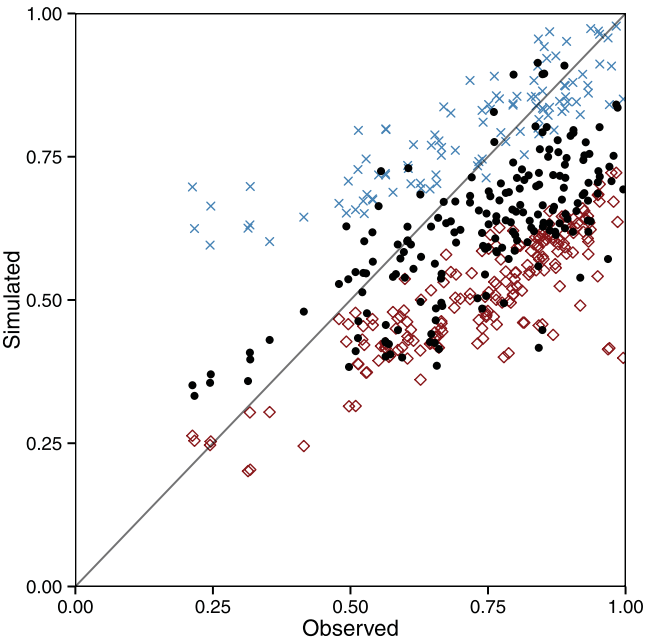

Figure 10. Observed versus simulated intersite correlations from published regression-based estimates (blue crosses), the NHMM with the copula (black dots) and the NHMM without the copula (red diamonds), for the period 1906–1997.

5.3. Intersite Correlations

Correlation between sites is an important feature to capture in multisite reconstructions and simulations. As we described in the method section, the copula is instrumental in capturing the intersite correlations. Figure 10 shows the intersite correlations of median estimates from the NHMM with the copula (black points), median estimates from the NHMM without the copula (red diamonds) and from the published regression-based estimates (blue crosses). As expected, using a copula with the NHMM improves the intersite correlations compared to a version of the model without a copula. Regression-based approaches tend to capture high correlations well but consistently overestimate moderate correlations. This is likely due to the fact that many of the previous reconstructions use identical tree-ring covariates, propagating the high correlations to the reconstructions. The NHMM without the copula does not have an explicit inclusion of intersite correlation hence it tends to underestimate correlations, especially higher values.

Although some correlations of the NHMM with copula are underestimated, they provide a favorable overall representation of correlation compared to regression models. Note, too, that the NHMM correlations shown in Figure 10 are the median of many simulations, but true correlations would be represented by a distribution of values (not shown) from each simulation.

6. Summary and Conclusions

A multisite streamflow reconstruction method is presented using a Bayesian hierarchical nonhomogeneous hidden Markov model. The model consists of single-site NHMMs that are bound together with an elliptical copula and fit together in a fully Bayesian framework. The model is applied to reconstruct annual flows at 20 sites on the Upper Colorado River Basin from 1473 to 1906 and the results are in good agreement with previous (regression-based) reconstructions for this region. The model has several advantages over traditional regression methods including the ability to simulate multiple sites simultaneously while preserving intersite correlations and a dynamic representation of uncertainty in each reconstructed year.

A data set generated by the NHMM consists of many annual time series of flow at each specified location. The range of values seen in all simulated time series represent the uncertainty present in the reconstructed values. Water resources planning models which can utilize probabilistic inputs require data in just such a format making the NHMM a valuable tool for producing water resources planning model input data.

Model selection uncovered that the optimal model included 5 hidden states and remarkably only required a single principal component from the tree-ring data. The NHMM performed similarly to regression in terms of matching observations during the validation period in which half the data was dropped.

Some key differences exist between reconstructions from the NHMM and regression:

1. Neither method perfectly captures the intersite correlations of the observed data. Regression tends to overestimate correlations lower than about 0.75, a consequence of using shared tree-ring predictors. The NHMM provides a more consistent representation of intersite correlations but tends to underestimate higher correlations by about 10–20%.

2. By design the NHMM has the ability to simulate at multiple sites simultaneously, generating sequences of flows at each site known as traces. While it is possible with the at-site regression models to generate traces, the ability to generate robust traces without extra postprocessing is limited because the regression models do not have any information on year-to-year transitions.
3. The NHMM can be scaled to accommodate any number of sites on a river network, or even sites across river networks without the need to build separate models for each site. We chose the UCRB for our application because of the availability of high quality data and a multitude of existing reconstructions from which to compare and validate. This method is general enough to apply to any river network with available tree-ring data. One approach that may improve the NHMM even further would be to select predictors (from among the PCs) individually for each site on a network.
4. The NHMM and regression make fundamentally different assumptions about the underlying probability distribution for flow. Regression assumes flow values are drawn from a normal distribution with a mean that changes based on tree-ring covariates and a fixed variance. The NHMM assumes the underlying distribution of flow is a mixture of lognormal distributions in that the mean changes based on the hidden state and tree-ring covariates and the variance changes with hidden state. One consequence of using the lognormal component distributions is minimum and maximum flows from the NHMM tend to be higher than those reconstructed via regression though there is obviously no data during this period to verify, which method is more accurate. During the validation period (1906–1951) the NHMM tended to better reproduce observed mean and variance which were consistently underestimated by regression. In the validation period the NHMM captured the maxima and minima better at some sites and regression performed better at others.

One potential drawback of the NHMM is a slight under-simulation of observed intersite correlations on average. This may be able to be improved upon through the use of different component distributions, a different copula or including more physical information about the river network, though these were not investigated in this study. Regression on the other hand tends to overestimate intersite correlations, a detail which may be relevant depending on the application.

Some future work using the NHMM includes incorporating other hydrometeorological variables such as soil moisture or temperature as in *Gangopadhyay et al.* [2015] and exploring other hierarchical model structures, which could improve computation time. In addition, this approach could be applied to other basins to assess its performance under different hydroclimatic conditions and with a difference set of tree-ring chronologies.

Acknowledgments
Funding for this research by a Science and Technology grant from Bureau of Reclamation is gratefully acknowledged. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research. The authors are thankful for support from the Janus supercomputer staff at the University of Colorado. Finally, the authors greatly appreciate the constructive comments of three anonymous reviewers. Data are available at http://civil.colorado.edu/bracken/uc_treering_recon/. Pre and postprocessing analysis was conducted using the R language [*R Core Team*, 2014].

References

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.

Barnett, F. A., S. T. Gray, and G. A. Tootle (2010), Upper Green River Basin (United States) streamflow reconstructions, *J. Hydrol. Eng.*, *15*(7), 567–579.

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.

Devineni, N., U. Lall, N. Pederson, and E. Cook (2013), A tree-ring-based reconstruction of Delaware River basin streamflow using hierarchical Bayesian regression, *J. Clim.*, *26*, 4357–4374, doi:10.1175/JCLI-D-11-00675.1.

Fritts, H. C., J. Guiot, and G. A. Gordon (1990), Verification, in *Methods of Dendrochronology Applications in the Environmental Sciences*, edited by E. R. Cook and L. A. Kairiukstis, pp. 178–185, Springer, N. Y.

Gangopadhyay, S., B. Harding, B. Rajagopalan, J. Lukas, and T. Fulp (2009), A nonparametric approach for paleohydrologic reconstruction of annual streamflow ensembles, *Water Resour. Res.*, *45*, W06417, doi:10.1029/2008WR007201.

Gangopadhyay, S., G. J. McCabe, and C. A. Woodhouse (2015), Beyond annual streamflow reconstructions for the Upper Colorado River Basin: A paleo-water-balance approach, *Water Resour. Res.*, *51*, 9763–9774, doi:10.1002/2015WR017283.

Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*(4), 457–472.

Gray, S. T., J. J. Lukas, and C. A. Woodhouse (2011), Millennial-length records of streamflow from three major Upper Colorado River Tributaries 1, *J. Am. Water Resour. Assoc.*, *47*(4), 702–712.

Hidalgo, H. G., T. C. Piechota, and J. A. Dracup (2000), Alternative principal components regression procedures for dendrohydrologic reconstructions, *Water Resour. Res.*, *36*(11), 3241–3249.

Hoffman, M. D., and A. Gelman (2014), The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo, *J. Mach. Learning Res.*, *15*, 1593–1623.

Hughes, J. P., and P. Guttorp (1994), A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena, *Water Resour. Res.*, *30*(5), 1535–1546.

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.

Jacoby, G. C., and C. W. Stockton (1976), Long-term surface-water supply and streamflow trends in the Upper Colorado River basin based on tree-ring analyses, *Natl. Sci. Found. Lake Powell Res. Proj. Bull.*, *18*, 70.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005), Markov Chain Monte Carlo methods and the label switching problem in bayesian mixture modeling, *Stat. Sci.*, *20*(1), 50–67.

McCabe, G. J., and D. M. Wolock (2011), Independent effects of temperature and precipitation on modeled runoff in the conterminous United States, *Water Resour. Res.*, *47*, W11522, doi:10.1029/2011WR010630.
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.
Meko, D. M., C. A. Woodhouse, C. A. Baisan, T. Knight, J. J. Lukas, M. K. Hughes, and M. W. Salzer (2007), Medieval drought in the upper Colorado River Basin, *Geophys. Res. Lett.*, *34*, L10705, doi:10.1029/2007GL029988.
Michaelsen, J., H. A. Loaiciga, L. Haston, and S. Garver (1990), Estimating drought probabilities in California using tree rings, *Tech. Rep. B-57105*, Calif. Dep. of Water Resour., Univ. of Calif., Santa Barbara.
Nowak, K., J. Prairie, B. Rajagopalan, and U. Lall (2010), A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow, *Water Resour. Res.*, *46*, W08529, doi:10.1029/2009WR008530.
Nowak, K., M. Hoerling, B. Rajagopalan, and E. Zagona (2012), Colorado River Basin hydroclimatic variability, *J. Clim.*, *25*(12), 4389–4403.
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.
R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna.
Renard, B. (2011), A Bayesian hierarchical approach to regional frequency analysis, *Water Resour. Res.*, *47*, W11513, doi:10.1029/2010WR010089.
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.
Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*, 461–464.
Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002), Bayesian measures of model complexity and fit, *J. R. Stat. Soc., Ser. B*, *64*(4), 583–639.
St George, S., and T. R. Ault (2014), The imprint of climate within Northern Hemisphere trees, *Quat. Sci. Rev.*, *89*, 1–4.
Stan Development Team (2015a), *Stan: A C++ Library for Probability and Sampling, Version 2.7.0*.
Stan Development Team (2015b), *RStan: The R interface to Stan, Version 2.7.0*.
Vano, J. A., T. Das, and D. P. Lettenmaier (2012), Hydrologic sensitivities of Colorado River runoff to changes in precipitation and temperature*, *J. Hydrometeorol.*, *13*(3), 932–949.
Woodhouse, C. A., and J. J. Lukas (2006), Drought, tree rings and water resource management in Colorado, *Can. Water Resour. J.*, *31*(4), 297–310.
Woodhouse, C. A., S. T. Gray, and D. M. Meko (2006), Updated streamflow reconstructions for the Upper Colorado River Basin, *Water Resour. Res.*, *42*, W05415, doi:10.1029/2005WR004455.
Zucchini, W. (2009), Hidden Markov models for time series: An introduction using R. Chapman & Hall/CRC, Boca Raton, Fla.