A stochastic nonparametric approach for streamflow generation combining observational and paleoreconstructed data

James Prairie,[1] Kenneth Nowak,[2] Balaji Rajagopalan,[2,3] Upmanu Lall,[4] and Terrance Fulp[5]

Received 9 November 2007; revised 14 March 2008; accepted 17 April 2008; published 27 June 2008.

[1] The Colorado River basin experienced the worst drought on record during 2000–2004. Paleoreconstructions of streamflow for the preobservational period show droughts of greater magnitude and duration, indicating that the recent drought is not unusual. The rich information provided by paleoreconstructions should be incorporated in stochastic streamflow models, enabling the generation of realistic flow scenarios required for robust water resources planning and management. However, the magnitudes of reconstructed streamflow have a high degree of uncertainty. This apparent weakness of the paleodata has made their use in water resources planning contentious, despite their availability for many decades. However, few contest the accuracy of hydrologic state (i.e., dry and wet periods). A key question is how to combine the long paleoreconstructed streamflow information of lower reliability with the shorter observational data to develop a framework for streamflow simulation. We propose a unique stochastic streamflow simulation framework combining these two data sets. This has two components: (1) a nonhomogeneous Markov chain model, developed using the paleodata, which is used to simulate the hydrologic state, and (2) a nonparametric K-nearest neighbor (K-NN) time series bootstrap of observational flow magnitudes conditioned on the hydrologic state, thus combining the respective strengths of the two data sets. The framework is demonstrated for the Lees Ferry, Arizona, stream gauge on the Colorado River. The simulations show the ability to reproduce relevant statistics of the observational period and generate a rich variety of wet and dry sequences for use in sustainable management of water resources.

Citation: 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.

1. Introduction

[2] Effective long-term planning and management of water resources requires (1) a tool that can generate plausible streamflow scenarios and (2) a decision model to evaluate policy alternatives. Stochastic models are typically built on observed streamflow data and are then used to generate flow scenarios. There is a rich literature on models to simulate basin-wide flows in a linear [*Valencia and Schaake*, 1973; *Mejia and Rousselle*, 1976; *Tao and Delleur*, 1976; *Lane and Frevert*, 1990; *Salas et al.*, 1980; *Todini*, 1980; *Stedinger and Vogel*, 1984; *Stedinger et al.*, 1985; *Koutsoyiannis*, 1992; *Santos and Salas*, 1992; *Salas*, 1993; *Koutsoyiannis and Manetas*, 1996; *Koutsoyiannis*, 2001] or nonlinear [*Tarboton et al.*, 1998; *Kumar et al.*, 2000; *Sharma and O'Neill*, 2002; *Srinivas and Srinivasan*, 2005] framework. Observational data are usually limited in time, and thus the simulations have a limited range of interannual variability, especially for the magnitude and frequency of the extremes, which are crucial for robust long-term planning. This was underscored on the Colorado River basin during the recent severe and sustained drought. The basin experienced the worst drought on record from 2000 to 2004. Though this drought was unprecedented in the observed record (1906–2005), paleoreconstructions of streamflow from tree ring chronologies have shown droughts of greater magnitude and duration. A recent paleoreconstructed streamflow for the period 1490–1997, on the Colorado River at Lees Ferry, Arizona, a key gauge on the river [*Woodhouse et al.*, 2006], is shown in Figure 1 along with the observed flows. It is evident that the recent drought is unprecedented during the observed period, but the reconstructed streamflows prior to 1906 show severe droughts of 5 years in length at least four times over the approximately 500-year period, indicating that the recent drought is not unusual.

[3] Clearly, the rich information provided by paleoreconstructed streamflows should be incorporated in stochastic streamflow models to enable the generation of a realistic

[1] Bureau of Reclamation, University of Colorado, Boulder, Colorado, USA.

[2] Department of Civil, Environmental and Architectural Engineering, University of Colorado, Boulder, Colorado, USA.

[3] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA.

[4] Department of Earth and Environmental Engineering, Columbia University, New York, New York, USA.

[5] Bureau of Reclamation, Lower Colorado Region, Boulder City, Nevada, USA.

0043-1397/08/2007WR006684$09.00

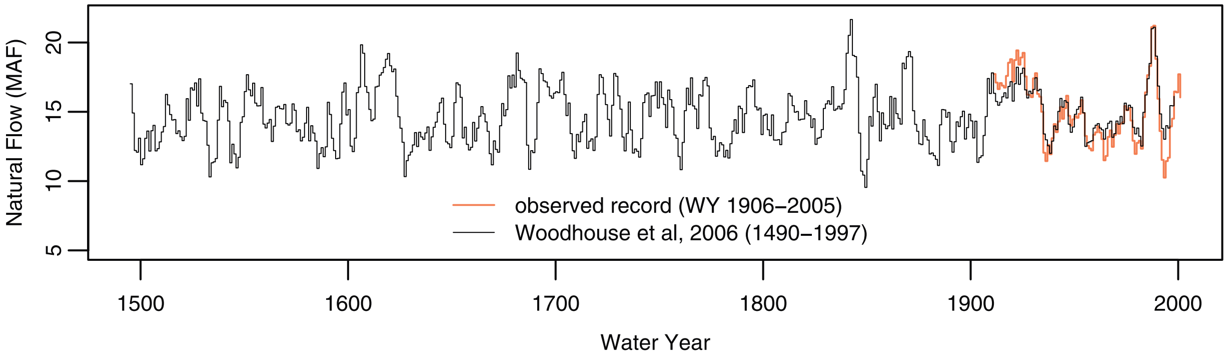


Figure 1. Five-year running means of historic and paleoreconstructed streamflow at Lees Ferry.

variety of plausible flow scenarios. However, the magnitudes of reconstructed streamflow have a high degree of uncertainty. Typically, a regression model is fit to the observed streamflow with a suite of tree ring observations as the predictors. This fitted model is then used to estimate streamflows in the preobservational period using the tree ring observations [*Meko et al.*, 1995]. The reconstructed streamflows can be sensitive to the choice of model as demonstrated by *Hidalgo et al.* [2000]. This apparent weakness of the paleoreconstructed flow data has made their use in a water resources planning context contentious, despite the availability of paleoreconstructed data for many decades. In spite of these apparent weaknesses, few argue about the duration and frequency of dry and wet (i.e., the hydrologic state) periods from the reconstructions [*Woodhouse et al.*, 2006]. The key question is how to combine the long paleoreconstructed streamflow information of lower reliability with the shorter but reliable observational data to develop a framework for simulation of streamflow scenarios.

[4] To address this question, we propose a new two-step process in which the hydrologic state (i.e., wet or dry) is modeled using the paleoreconstruction data and the flow magnitudes derived from the observational data. Specifically, a nonhomogeneous Markov chain model [*Rajagopalan et al.*, 1996, 1997] is built on the paleodata that is then used to simulate the hydrologic state. The flow magnitudes are then generated conditioned on the simulated hydrologic state using a K-nearest neighbor (K-NN) conditional time series bootstrap [*Lall and Sharma*, 1996], thereby using the strengths of both of these data sets. The data sets used, the proposed framework, and the application to the Lees Ferry, Arizona, stream gauge on the Colorado River are described in the following sections.

2. Data Sets

[5] As mentioned earlier, two data sets, paleoreconstructed streamflow and observed flows, are used in this study. These are described below.

2.1. Natural Streamflow

[6] The natural streamflow data for the Colorado River basin are developed by the Bureau of Reclamation (Reclamation) and updated regularly. Annual updates addressing data changes and additions are typical. Naturalized streamflows are computed by removing anthropogenic impacts (i.e., reservoir regulation, consumptive water use, etc.) from the recorded historic flows. *Prairie and Callejo* [2005] present a detailed description of methods and data used for the computation of natural flows in the Colorado River basin. This study uses the annual water year (September–October) natural streamflow at Lees Ferry, Arizona, for the period 1906–2005.

2.2. Paleoreconstructed Streamflow

[7] This study also uses the annual water year streamflow reconstructions from tree ring information at the Lees Ferry, Arizona, gauge, completed by *Woodhouse et al.* [2006] for the period 1490–1997. Tree ring widths are influenced by climate and available soil moisture and thus are good integrators of the weather fluctuations, just as streamflow is a watershed integration of hydrologic and climatologic processes. Consequently, the tree ring widths are well correlated with annual runoff. To gather ring width data, a series of trees are cored at multiple locations, chosen such that the tree species have annual rings sensitive to moisture availability. Selecting the species and the location is very important for this effort [*Meko et al.*, 1995]. Two core samples are taken from each tree for cross dating, and the ring widths are measured, obtaining the chronology of tree ring widths. The attractive aspect of tree-ring-based reconstructions, unlike other paleoproxy data, is that trees that put on annual rings have natural dating, with the outer ring corresponding to the current year and the subsequent inner rings corresponding to past years. A standard series of techniques [*Stokes and Smiley*, 1968; *Swetnam et al.*, 1985] are employed to process the ring width series. Typically, the series is first detrended to remove the effects of reduced ring width with aging. Next, the ring width series from various cores at a single location are combined to develop a "site chronology" [*Cook et al.*, 1990]. The site chronology is related to observed streamflow during the overlap period; typically, a multiple linear regression model is fit [*Weisberg*, 1985]. For the Colorado River at the Lees Ferry, Arizona, gauge, the regression model developed by *Woodhouse et al.* [2006], using all the available pool of chronologies (30 in total), explains approximately 84% of the annual variance of the observed streamflow. The fitted regression model is then used to estimate the streamflow during the preobservation period when tree ring information is available, thus obtaining the reconstructed streamflow series.

[8] Especially during high streamflow periods it is known that the tree ring widths are influenced by variables other

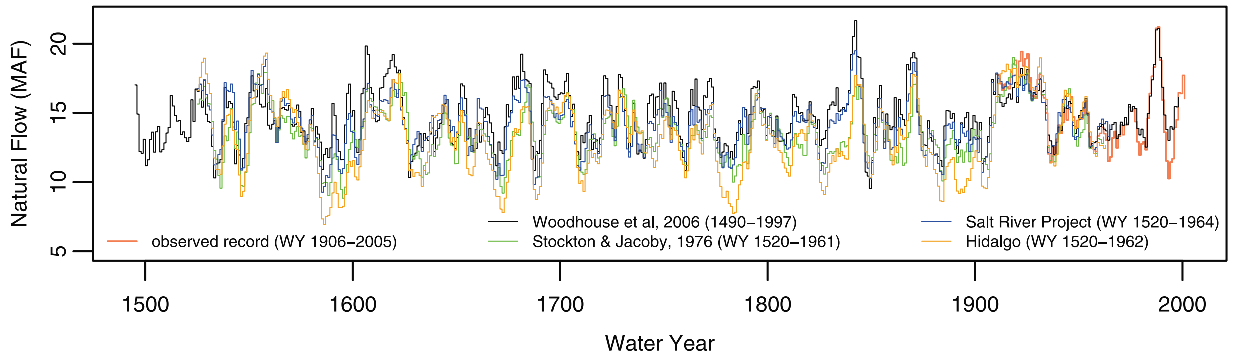


Figure 2. Five-year running means of recent and previous streamflow reconstructions at Lees Ferry.

than moisture availability, thus degrading their ability in accurately representing high flow years. Further, different data sets and techniques to process tree ring information can result in substantial differences in the reconstructed flows [*Hidalgo et al.*, 2000]. This can be seen in Figure 2, where four different streamflow reconstructions at the Lees Ferry, Arizona, gauge are shown, including the earliest reconstruction of *Stockton and Jacoby* [1976], later reconstructions by *Hidalgo et al.* [2000], that of *Hirschboeck and Meko* [2005] as part of the Salt River Project, and the most recent reconstruction by *Woodhouse et al.* [2006]. Each reconstruction used a different set of tree ring chronologies and different processing methods. Of particular interest is the increased severity of drought and reduced overall mean displayed by the Hidalgo reconstruction. Unfortunately, the variability across reconstructions has not helped instill confidence in use of these data by policy makers and water managers in the Colorado River basin, even with growing interest in wanting to use them. Despite their differences, reconstructions tend to agree quite well on “wet” and “dry” years [*Woodhouse et al.*, 2006], as seen in Figure 3. We found that three or more reconstructions agree on the

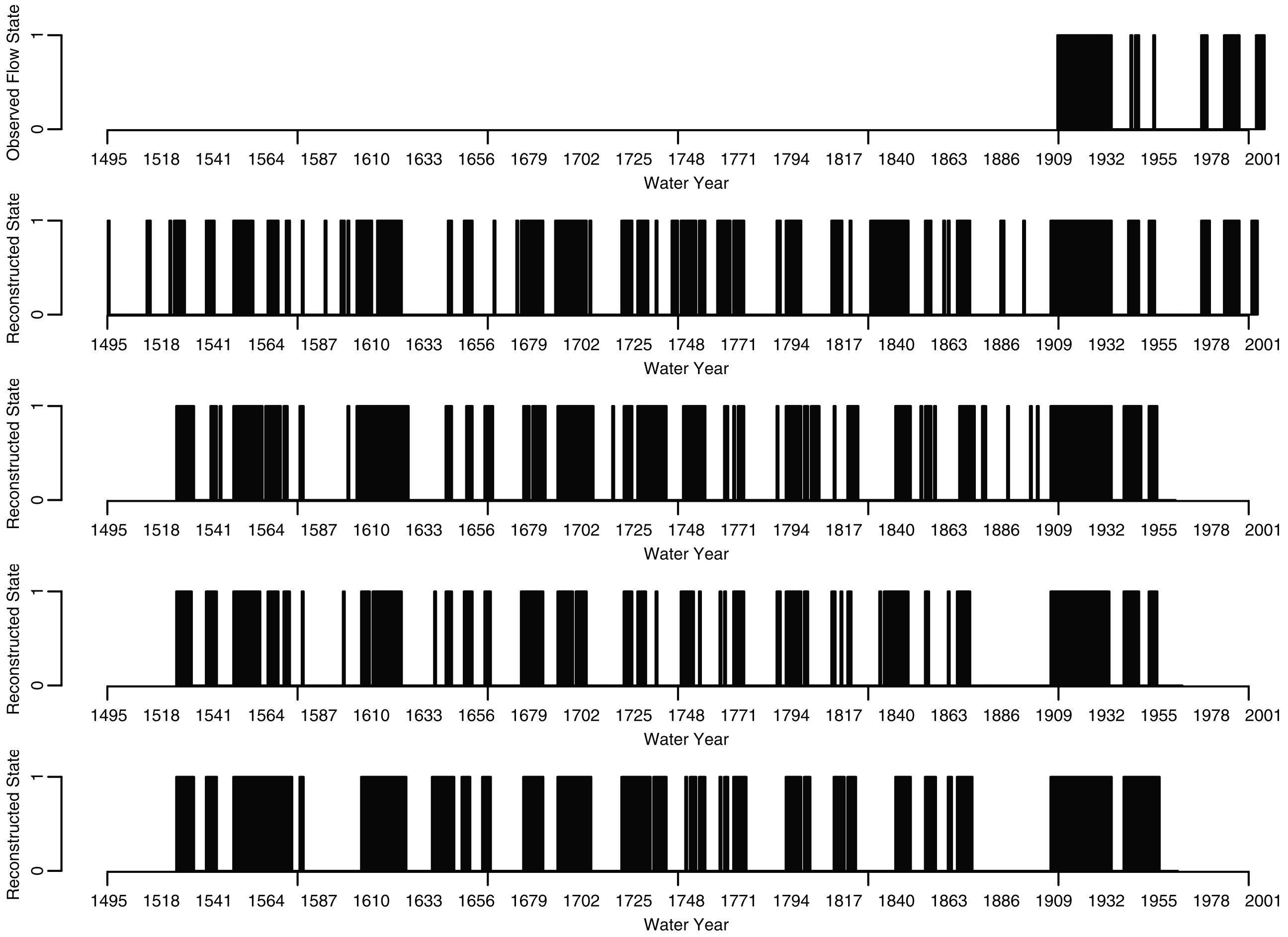


Figure 3. System state, i.e., wet (1) or dry (0), derived from 5-year running means for recent and previous streamflow reconstructions at Lees Ferry.

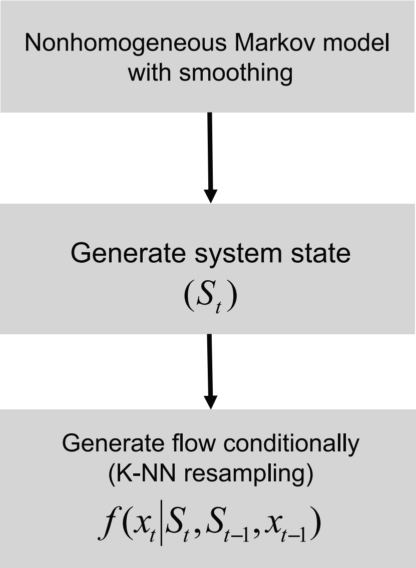

Nonhomogeneous Markov model
with smoothing
Generate system state
(S_t)
Generate flow conditionally
(K-NN resampling)
$f(x_t|S_t, S_{t-1}, x_{t-1})$

Figure 4. The nonparametric paleoconditioning (NPC) modeling framework description.

hydrologic state 88% of the time, while all four methods agree 65% of the time on an annual basis. This offers the potential to use the paleoreconstructed streamflows to model the hydrologic state (i.e., wet or dry) of the system and use the observational data for the flow magnitude. This forms the basis of our proposed framework.

3. Proposed Framework

[9] As mentioned above, the proposed framework combines the paleoreconstructed streamflows with the observational data in a framework for simulating robust streamflow scenarios for use in water resources management. The paleoreconstructed data are used to model the hydrologic state of the system. The median of the observed flows is used to define periods as wet if flow is greater than this threshold and dry if flow is less than this threshold. Epochs of wet and dry periods identified using this criterion are illustrated in Figures 2 and 3. They illustrate the persistence in wet/dry regimes that suggests a Markov chain based model. Because the state transition appears to be varying through time, a nonhomogeneous Markov chain modeling approach is appropriate. The streamflow magnitudes are then simulated from the conditional probability density function, given the wet or dry state using a nonparametric K-nearest neighbor bootstrap approach. The framework is shown in Figure 4. The description of these two components of the framework along with background information are provided below. Hereinafter we refer to this framework as nonparametric paleoconditioning (NPC).

3.1. Modeling the Hydrologic State

[10] Markov chains have been extensively used to model daily precipitation occurrence [*Gabriel and Neumann*, 1962; *Todorovic and Woolhiser*, 1975; *Smith and Schreiber*, 1974; *Salas*, 1993, and references within]. Typically, for a two-state (wet, dry) first-order model (i.e., state transition at the current time step depends on the previous state), the transition probabilities are directly estimated from the data by counting the proportion of transitions to a wet year from a dry year, P_{dw}, and the probability of a wet year followed by a dry year, P_{wd}. The probability of a dry year followed by a dry year can be obtained as $P_{dd} = 1 - P_{dw}$; likewise, the probability of a wet year followed by a wet year can be obtained as $P_{ww} = 1 - P_{wd}$. The transition probabilities can be readily used to simulate the hydrologic states and consequently, their frequencies. If these transition probabilities are assumed to be stationary and calculated from the entire data, then it is a "stationary" Markov chain. Here, though (Figures 2 and 3), the frequencies of wet and dry periods are varying (i.e., nonstationary) over time.

[11] The nonstationarity can be addressed in several ways. A moving window of some W time steps can be selected and the transition probabilities can be estimated for each time window and repeated by moving forward every time step. The transition probability estimates for each year are based on state observations present in the window length. An alternative, hidden Markov models, has been gaining popularity. In these, the underlying epochal (or regime) changes are modeled probabilistically and the transition probabilities are then conditionally estimated based on the epoch. These models have been applied to precipitation, climate, and streamflow data [see, e.g., *Zucchini and Guttorp*, 1991; *MacDonald and Zucchini*, 1997; *Lu and Berliner*, 1999; *Thyer and Kuczera*, 2000, 2003a, 2003b; *Akıntuğ and Rasmussen*, 2005]. Another approach to dealing with nonstationarity is the nonhomogeneous Markov models [*Hughes and Guttorp*, 1994; *Hughes et al.*, 1999; *Bellone et al.*, 2000; *Lambert et al.*, 2003]. For example, Fourier series were fit to model the changing transition probability with season for precipitation [*Woolhiser and Pegram*, 1979; *Roldan and Woolhiser*, 1982; *Feyerherm and Bark*, 1965].

[12] Nonparametric alternatives [e.g., *Rajagopalan et al.*, 1996, 1997; *Mehrotra et al.*, 2004; *Mehrotra and Sharma*, 2005] offer a more general and flexible approach. In particular, here we use the nonhomogeneous Markov model (NHM) developed by *Rajagopalan et al.* [1996], in which the transition probability at any time t is estimated as a weighted average of the transitions within a window of size H centered on t. The window size H is obtained from objective criteria. This was developed to model a daily precipitation process and subsequently applied for modeling the occurrence of El Niño–Southern Oscillation [*Rajagopalan et al.*, 1997]. We adapt the NHM framework for modeling the streamflow states described below.

[13] The transition probabilities, P_{dw} (t) and P_{wd} (t), for a given year are estimated by a discrete nonparametric kernel estimator given as

$$P_{dw}(t) = \frac{\sum_{i=2}^{n} K\left(\frac{t-t_i}{h_{dw}}\right) S_t[1-S_{t-1}]}{\sum_{i=2}^{n} K\left(\frac{t-t_i}{h_{dw}}\right) S_t} \tag{1}$$

$$P_{wd}(t) = \frac{\sum_{i=2}^{n} K\left(\frac{t-t_i}{h_{wd}}\right) [1-S_t]S_{t-1}}{\sum_{i=2}^{n} K\left(\frac{t-t_i}{h_{wd}}\right) [1-S_t]}, \tag{2}$$

where $K()$ = the kernel function, S_t = system hydrologic state (1 = wet, 0 = dry) at time t, S_{t-1} = system hydrologic state at time $t - 1$, $h_{()}$ = the kernel bandwidth, t = year of interest, and n = the number of values in the window $t - h_{()}$ to $t + h_{()}$. The discrete quadratic kernel function developed by *Rajagopalan and Lall* [1995] is used, which is given as

$$K(x) = \frac{3h}{(4h^2 - 1)}\left(1 - x^2\right) \quad \text{for} \quad |x| \leq 1, \tag{3}$$

where $x = (t - t_{()})/h_{()}$ measures the distance for event $t_{()}$ from the year of interest t within the bandwidth $h_{()}$, where $h_{()}$ is an integer. The weights from the kernel function are positive and sum to unity. It can be seen that the estimates of transition probabilities at any year t are based only on the transitions within a window $t - h_{()}$ to $t + h_{()}$.

[14] The transition probability estimators (1) and (2) are fully defined once the bandwidth $h_{()}$ is determined for each. An objective method based on a least squares cross-validation (LSCV) procedure [*Scott*, 1992] is used to select the optimal bandwidth that was developed by *Rajagopalan et al.* [1996] for the NHM case,

$$\text{LSCV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \hat{P}_{-t_i}(t_i)\right]^2, \tag{4}$$

where n = the number of observations (ndw or nwd), and $\hat{p}_{-t_i}(t_i)$ = the estimate of the transition probability ($\hat{p}_{WD}$ or $\hat{p}_{DW}$) at year t, based on data ranging from $t - h$ to $t + h$, with the exclusion of t (the transition at t should not be included when attempting to approximate that value). The 1 in equation (4) results from an assumption that the prior probability of transition is 1 for the years on which a transition has occurred. The value of h that minimizes the LSCV function is selected as the optimal bandwidth. The bandwidths h_{dw} and h_{wd} are objectively determined and subsequently used in the estimators (1) and (2) to estimate the transition probabilities for each year. The LSCV function does not always yield a clear minimum; therefore it is preferable to find an estimate for all available transitions to obtain a range of bandwidths. When a clear minimum is not found, it is recommended that a minimum delta h value (i.e., 0.0001) be determined to objectively find a minimum LSCV based on reaching the minimum delta h. We chose the clear minimum found within each complementary transition, but found little sensitivity over the range of possible bandwidths.

[15] Best Markov chain model orders are generally selected as the minimizers of the Akaike information criterion [*Gates and Tong*, 1976]. For the Lees Ferry paleoreconstructed data we found the two-state, first-order to be optimal.

3.2. Modeling the Flow Magnitudes

[16] The streamflow magnitudes, as mentioned earlier, are modeled based on the observed data and conditioned upon the hydrologic state simulated using the paleodata. This model can be described as the conditional probability density function (PDF),

$$f(x_t|S_t, S_{t-1}, x_{t-1}), \tag{5}$$

where the flow at the current time $t = x_t$ conditioned on the current system state = S_t, previous system state = S_{t-1}, and previous flow = x_{t-1}.

[17] Simulation from this conditional PDF is achieved by a K-NN bootstrap method [*Lall and Sharma*, 1996; *Rajagopalan and Lall*, 1999]. Typically, K-NN are identified in the observational data of the current feature vector [S_t, S_{t-1}, x_{t-1}]. One of the neighbors is selected, based on a metric that gives the highest probability to the nearest neighbor and the lowest to the farthest. The corresponding streamflow of the year that sequentially follows the selected neighbor is the simulated value for the current time.

[18] This case is unique in that the feature vector includes discrete and continuous variables. Further, the discrete variables indicate system state as 0 or 1, i.e., dry or wet, while the continuous variable is a considerably larger value. If this disparity in magnitude is not considered in the neighbor choice, the state information will not influence the neighbor choice. The neighbor would be chosen based solely on x_{t-1}. Therefore determination from the feature vector [S_t, S_{t-1}, x_{t-1}] is split into two steps. First the discrete variables are identified as members in one of the four categories (*ww*, *wd*, *dw*, *dd*) identified from the state vector [S_t, S_{t-1}]. In the second step, the K-nearest neighbors of x_{t-1} that lie within the appropriate category are identified. The flow for the following year, x_t, corresponding to the neighbor selected for x_{t-1}, is then sampled.

[19] In this work, $K_j = n_j$, where $j = 1,..,4$ represent the four state categories and n is the number of values in each category. With a larger observational data set the number of nearest neighbors can also be based on the heuristic scheme $K = \sqrt{n}$ [*Lall and Sharma*, 1996], following the asymptotic arguments of *Fukunaga* [1990]. Objective criteria such as generalized cross validation (GCV) can also be used [*Lall and Sharma*, 1996, *Prairie et al.*, 2005] The K_j neighbors were weighted with the function

$$W(i) = \left(\frac{1}{i}\right)\bigg/\left(\sum_{i=1}^{K}\frac{1}{i}\right).$$

3.3. Implementation Algorithm

[20] The complete framework combines the two models. The simulation proceeds as follows. First, a simulation horizon is identified, which is application dependent. Suppose a T-year horizon is chosen.

[21] 1. Randomly resample a block of T years from the paleoreconstructed streamflows, say 1651–1680.

[22] 2. Generate flow states $S(t)$ where $t = 1, 2,\ldots,T$ using the transition probabilities of the resampled years from step 1 above.

[23] 3. Generate flow magnitudes $x(t)$ for each $t = 1,2,\ldots,T$ from the conditional PDF $f(x_t|S_t, S_{t-1}, x_{t-1})$ using the K-NN bootstrap approach described in the previous section.

[24] 4. Repeat steps 2 and 3 to obtain as many simulations as required.

4. Model Evaluation

[25] The proposed framework (NPC) is applied to the paleoreconstructed streamflows (1490–1997) and observed

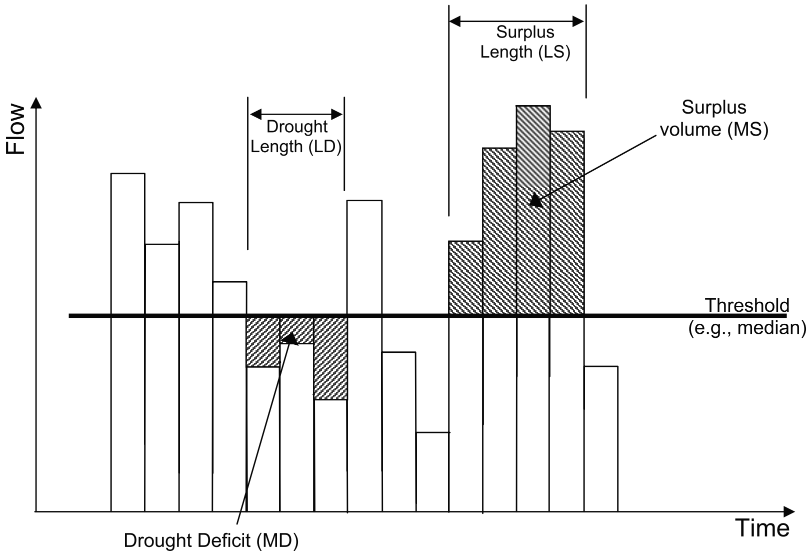


Figure 5. Definition of surplus and drought statistics.

natural flows (1906–2005) at Lees Ferry, Arizona, on the Colorado River. For this work, 500 simulations, each 100 years in length (same as the length of the observed flows), were generated.

[26] A suite of basic distributional statistics are computed including the annual (1) mean, (2) standard deviation, (3) coefficient of skew, (4) maximum, (5) minimum, and (6) lag-1 autocorrelation. Surplus and drought statistics include the average length surplus (avgLS), average length drought (avgLD), average surplus (avgS), and deficit (avgD) volume. Surplus (drought) is defined as values above (below) a threshold, here the median of the observed record. Figure 5 describes the computation of these surplus and drought statistics based on the threshold.

[27] The results are displayed as box plots where the box represents the interquartile range (IQR) and whiskers extend to the 5th and 95th percentiles of the simulations and outliers are shown as points beyond the whiskers. The statistics of the observed record are represented as a triangle, and the statistics of the paleoreconstructed record are represented as a circle. Performance on a given statistic is judged as good when the observed or paleostatistic, depending on the statistic of interest, falls within the interquartile range of the box plots, while increased variability is indicated by a wider box plot.

5. Results

[28] First the four sets of time-varying transition probabilities estimated from the NHM estimator (equations, (2), (3), and (4)) over the paleoperiod are shown in Figure 6. The optimal bandwidth minimizing the LSCV was found to be 37 years for the wet-wet transition and 19 for the dry-dry transition. The other two transition probabilities are complements of these. The epochal behavior in the transition probabilities is quite apparent. We draw attention to two epochs, (1) the early 1900s when the probability of transition to a wet state is higher than 0.5 and the transition to dry state is much lower than 0.5, which is also the epoch when the water sharing compact agreements on the Colorado River basin were developed, the wettest epoch in the past 500 years. In contrast, (2) the early 1600s is when the probability of transition to a dry state is much higher than 0.5, which is one of the driest periods in the paleorecord. There is also a steady decline in the probability of transition to a wet state in recent decades and a corresponding increase to dry states. Thus using these varied transition probabilities will provide a richer variety of wet and dry sequences, as seen in the results that follow.

[29] The simulations capture the basic distributional statistics of the observed streamflow within the IQR (Figure 7). This is consistent with the methodology in that the K-NN bootstrap approach resamples the observed data. Since the generated sequences are of the same length as the observed, the basic statistics of the observed streamflows are well captured, as to be expected. These distributional statistics of the paleorecord are not expected to be captured.

[30] Box plots of surplus and drought statistics are shown in Figure 8, along with the corresponding values from the observed record represented as a triangle and those from the paleorecord represented as a circle. The simulations from NPC generate longer drought and surplus sequences relative to observed, which can be seen by the observed statistics

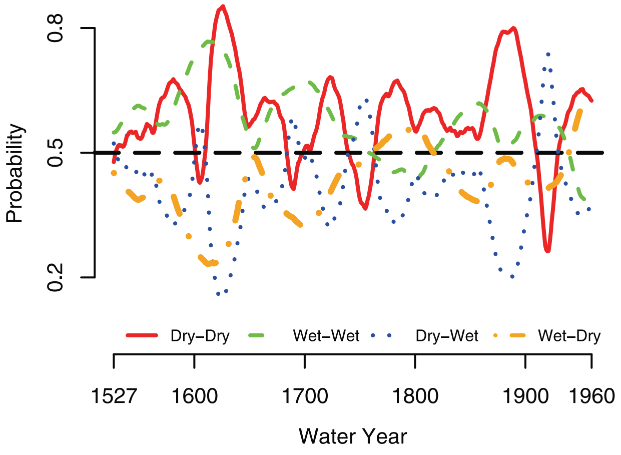


Figure 6. Transition probabilities from the paleostreamflows using the nonhomogeneous Markov estimator.

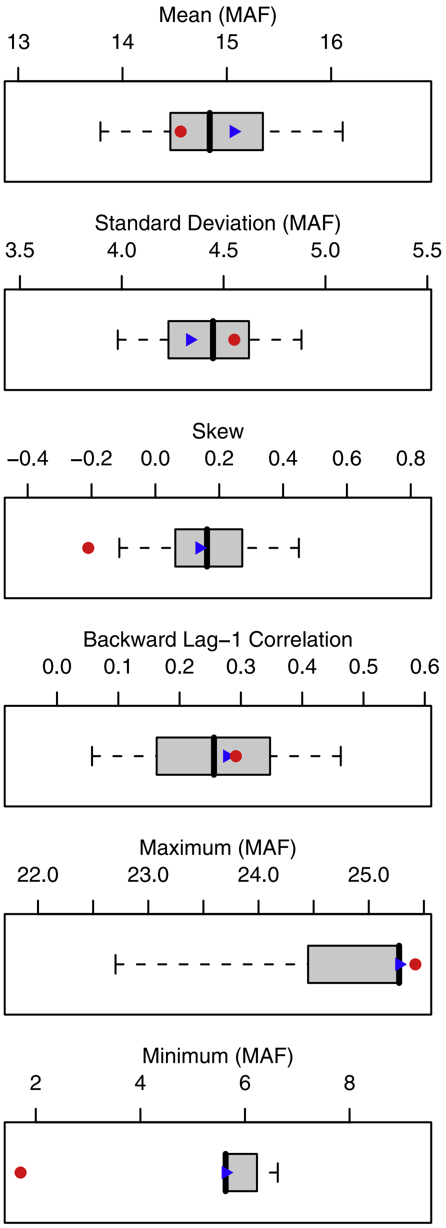


Figure 7. Box plots of basic statistics from NPC simulations. Statistics of the observed period are shown as blue triangles, and those of the paleoflows are shown as red circles.

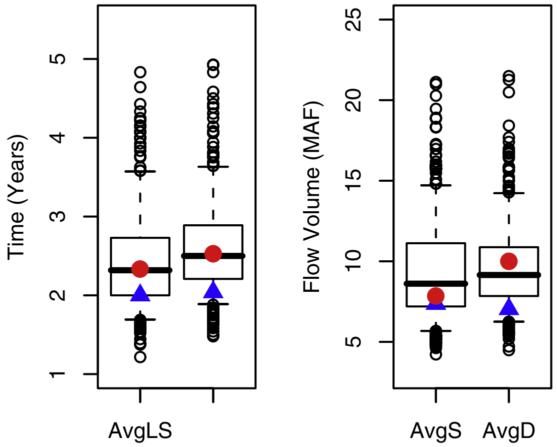


Figure 8. Box plots of drought and surplus statistics from NPC simulations. Statistics of the observed period are shown as blue triangles, and those of the paleo are shown as red circles.

falling low within or below the IQR in Figure 8. The avgLS and avgLD of the paleodata are well reproduced in the NPC simulations. The avgS and avgD are influenced by both the magnitudes of flow, which are resampled from the observed record, and the state sequences from the paleorecord; therefore these statistics represent a blend of both these records. For comparison, a simple K-NN lag-1 model as described by *Lall and Sharma* [1996] was used to resample the observed natural flow record with no influence from the paleorecord. Figure 9 shows the drought and surplus statistics from this simple model. The avgLS and avgLD as well as the avgS and avgD of the observed record are captured well within the IQR from this simulation, but the corresponding statistics from the paleorecord are not, as to be expected. Also, these simulations display reduced variability, as a tightened IQR, compared with Figure 8. The NPC framework is able to produce more varied drought and surplus sequences than what can be obtained from resampling only the observed data.

[31] The distribution of surplus and drought lengths is displayed in Figures 10 and 11 as histograms, respectively, for the observed, paleo, and NPC simulations. The histogram from the NPC simulations appear to be a smoothed version of that from the paleorecord, and also, the observed record has limited longest wet and dry spell lengths. Visually, the tail behavior of the histograms from the paleorecord and NPC simulations can be seen to be different from the observed record. The risk of a 6-year or longer dry spell (i.e., probability of exceedance) is 0% from the observed, 10.1% from the paleo, and 8.6% from the NPC simulations. The NPC provides a better sense of this risk, while the observed data show no risk of this. Also, the tails of the NPC drought and surplus plots extend to include event durations not seen in either the paleo or the observed. This is a new contribution to the field and is valuable in

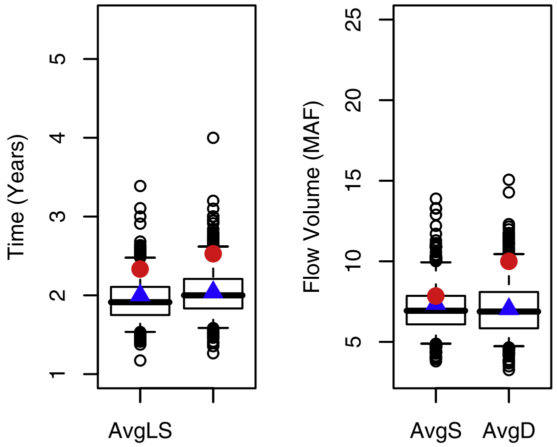
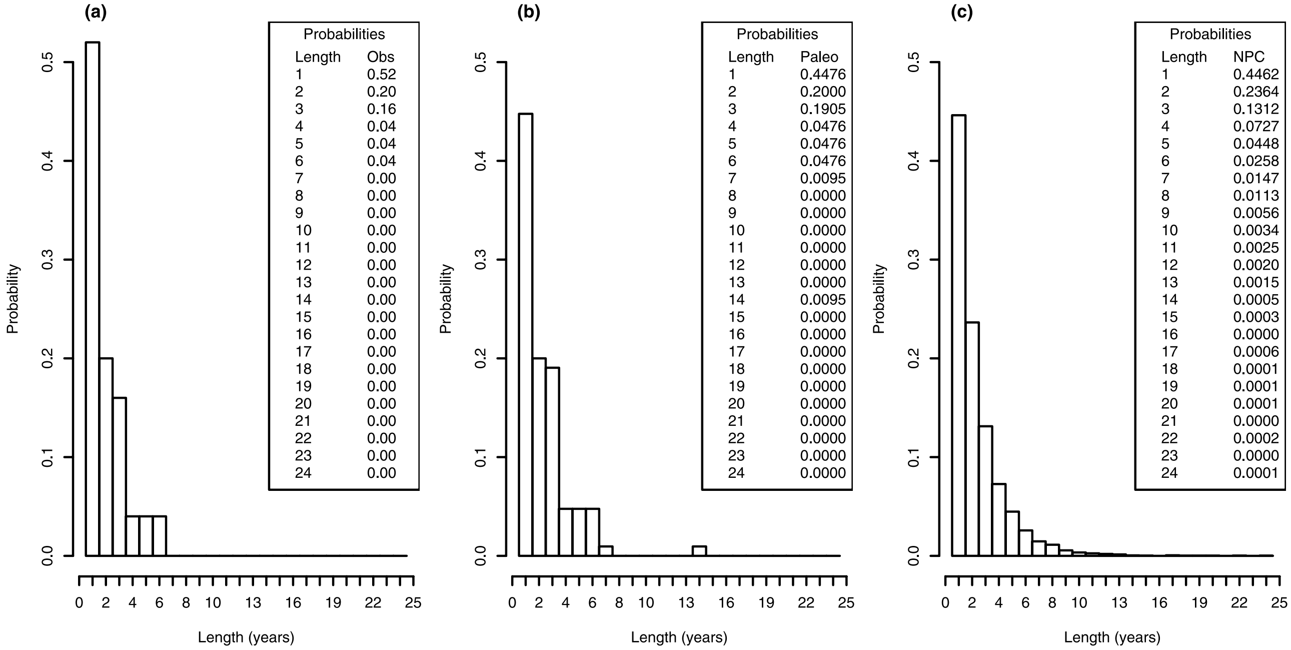


Figure 9. Box plots of drought and surplus statistics from K-NN lag-1 resampling of the observed data. Statistics of the observed period are shown as blue triangles, and those of the paleo are shown as red circles.



(a)

Probabilities

Length	Obs
1	0.52
2	0.20
3	0.16
4	0.04
5	0.04
6	0.04
7	0.00
8	0.00
9	0.00
10	0.00
11	0.00
12	0.00
13	0.00
14	0.00
15	0.00
16	0.00
17	0.00
18	0.00
19	0.00
20	0.00
21	0.00
22	0.00
23	0.00
24	0.00

(b)

Probabilities

Length	Paleo
1	0.4476
2	0.2000
3	0.1905
4	0.0476
5	0.0476
6	0.0476
7	0.0095
8	0.0000
9	0.0000
10	0.0000
11	0.0000
12	0.0000
13	0.0000
14	0.0095
15	0.0000
16	0.0000
17	0.0000
18	0.0000
19	0.0000
20	0.0000
21	0.0000
22	0.0000
23	0.0000
24	0.0000

(c)

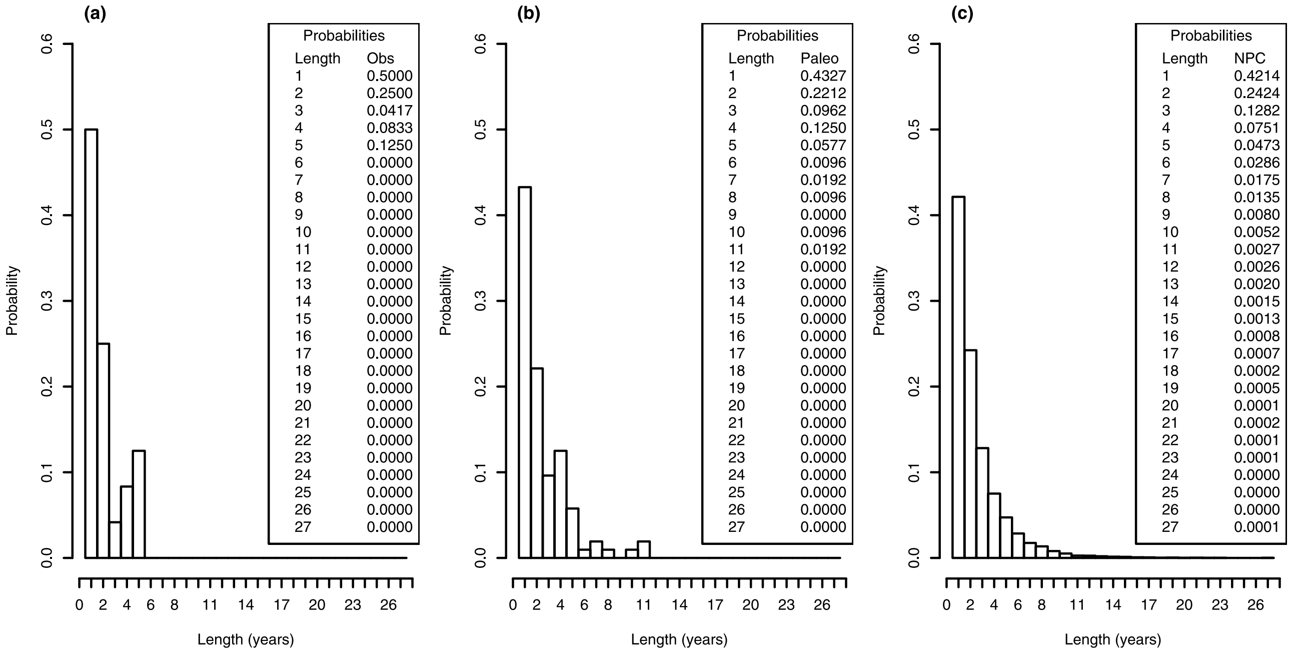
Probabilities

Length	NPC
1	0.4462
2	0.2364
3	0.1312
4	0.0727
5	0.0448
6	0.0258
7	0.0147
8	0.0113
9	0.0056
10	0.0034
11	0.0025
12	0.0020
13	0.0015
14	0.0005
15	0.0003
16	0.0000
17	0.0006
18	0.0001
19	0.0001
20	0.0001
21	0.0000
22	0.0002
23	0.0000
24	0.0001

Figure 10. Histogram of surplus lengths from (a) observed, (b) paleo, and (c) NPC simulations.

quantifying risk and planning for extreme events. The impacts on results from a decision support system that incorporated alternate hydrologic simulations including NPC simulations were published by *Bureau of Reclamation* [2007] for the Colorado River basin operations. This study found that use of NPC simulations indicated greater risk of lower reservoir conditions than when only using simulation based on the observed or paleorecord alone. These findings indicated the importance of developing sequences of flows not seen in the observed period but probable based on state information from the paleorecord.

[32] In the Colorado River basin the critical sequence of concern is a series of droughts connected over 12 years with surplus years interspersed. Such sequences are not represented in the drought statistics described above; *Timilsena et al.* [2007] address this through the use of a 5-year moving average to determine the hydrologic variable and thus periods of drought. Furthermore, the drought and surplus statistics estimated above are based on a preselected threshold (here it is the median streamflow of the observed period). Thus the results are sensitive to this selected threshold. To avoid this, a better approach is to determine the required storage for a given streamflow sequence to meet various demand levels. This incorporates the effect of multiple linked droughts and thus is more realistic in representing critical droughts. The algorithm, termed the



(a)

Probabilities

Length	Obs
1	0.5000
2	0.2500
3	0.0417
4	0.0833
5	0.1250
6	0.0000
7	0.0000
8	0.0000
9	0.0000
10	0.0000
11	0.0000
12	0.0000
13	0.0000
14	0.0000
15	0.0000
16	0.0000
17	0.0000
18	0.0000
19	0.0000
20	0.0000
21	0.0000
22	0.0000
23	0.0000
24	0.0000
25	0.0000
26	0.0000
27	0.0000

(b)

Probabilities

Length	Paleo
1	0.4327
2	0.2212
3	0.0962
4	0.1250
5	0.0577
6	0.0096
7	0.0192
8	0.0096
9	0.0000
10	0.0096
11	0.0192
12	0.0000
13	0.0000
14	0.0000
15	0.0000
16	0.0000
17	0.0000
18	0.0000
19	0.0000
20	0.0000
21	0.0000
22	0.0000
23	0.0000
24	0.0000
25	0.0000
26	0.0000
27	0.0000

(c)

Probabilities

Length	NPC
1	0.4214
2	0.2424
3	0.1282
4	0.0751
5	0.0473
6	0.0286
7	0.0175
8	0.0135
9	0.0080
10	0.0052
11	0.0027
12	0.0026
13	0.0020
14	0.0015
15	0.0013
16	0.0008
17	0.0007
18	0.0002
19	0.0005
20	0.0001
21	0.0002
22	0.0001
23	0.0001
24	0.0000
25	0.0000
26	0.0000
27	0.0001

Figure 11. Histogram of drought lengths from (a) observed, (b) paleo, and (c) NPC simulations.

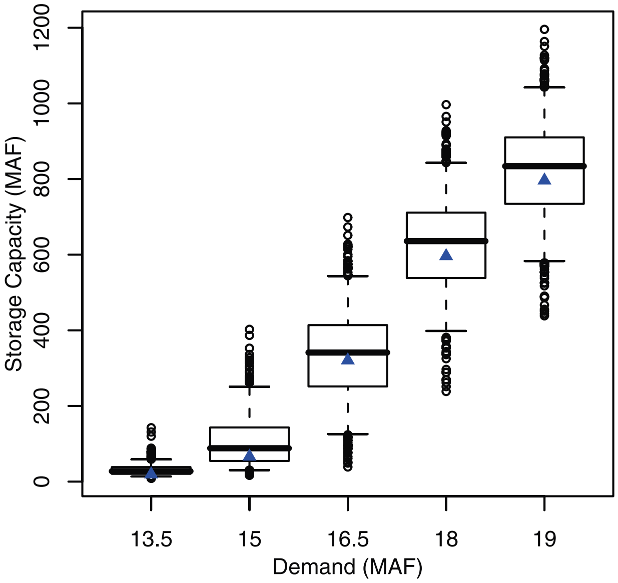


Figure 12. Box plots of demand-storage from the sequent peak algorithm on the NPC simulations. The corresponding values from the observed data are shown as triangles.

sequent peak algorithm [*Loucks et al.*, 1981], used for this purpose is given as

$$S'_i = \begin{cases} S'_{i-1} + d - y_i \\ 0 \end{cases} \tag{6}$$

$$S_c = \max\left[S'_i, \ldots, S'_N\right], \tag{7}$$

where S_i' is the storage at time step i, d is the demand or yield, y_i is the streamflow from a sequence of N values at time i, and S_c is the storage capacity. This is also widely used for designing reservoir capacities.

[33] The algorithm is run for various demand (yield) levels with the historic flow (triangle), and each trace of the 500 simulations (box plots) shown in Figure 12. The sequences generated from the NPC framework introduce significant and realistic flow variability and as a result, reduced system reliability. For example, consider a demand of 16.5 million acre feet (MAF; 1.233 × 10^9 m^3). To reliably meet this demand, based on the historic inflow sequence (triangle), a storage capacity of 325 MAF is required. The box plot shows considerable variability in the required storage capacity based on the 500 traces simulated from the combined framework. Furthermore, the box plot, shown as a PDF (Figure 13a) or a cumulative distribution function (CDF) (Figure 13b), can easily be used to find the reliability. It is clear that a demand of 16.5 MAF cannot reliably be met 98.9% of the time for a storage capacity of 60 MAF (the approximate current storage capacity of the Colorado River basin). The reliability is the area under the PDF curve below 60 MAF which is 1 minus the area of the hatched region in Figure 13a, or (1– 0.989 = 0.011) as read from the CDF. The reliability of alternate storage capacities can be found from Figure 13a or 13b in a similar manner.

[34] The sequent peak method assumes that the demand level is constant through time and must be met in all years. In real operations, however, this is not the case. As a result, the reliability estimates obtained above tend to be too simplistic and conservative and provide only a coarse representation of the actual system reliability. Therefore we urge caution in using these results to read policy implications. To fully appreciate the actual operations of the water resources in a river basin, a decision support system that incorporates variable demand schedules, proper topographic layout for river system reservoir, diversion points, and operating policies must be used. This will help provide realistic estimates of reliability for the various decision components

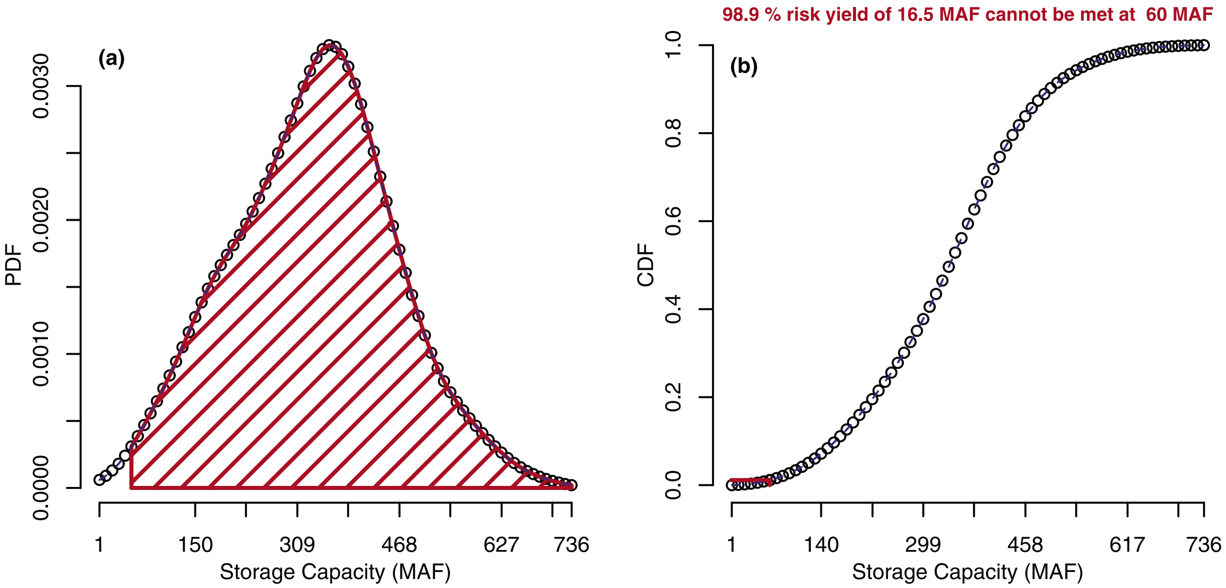


Figure 13. (a) Probability density function and (b) cumulative distribution function for 16.5 million acre feet demand box plot from Figure 12.

of the system, as demonstrated by *Prairie* [2006] and *Bureau of Reclamation* [2007, Appendix N].

[35] As seen from the results, the utility of the proposed framework of combining information from paleoreconstructions and observations is to produce a rich variety of wet and dry spells, which are crucial for robust water resources planning. Investigation of the spell distribution shows that the combination approach generates a higher risk of extended wet and dry spells. This risk will have a significant impact on the water resources management in the basin, especially when the current modeling framework does not model events greater than 5 years in length. The spell variability generated here is much richer than what can be obtained from traditional time series modeling of the observed data [*Prairie*, 2006].

6. Summary and Discussion

[36] A novel framework for combining information from multiple sources in generating scenarios was developed. The methodology is data driven, flexible, and easy to implement. Other variations of the framework are possible, especially for generating state information, such as (1) fitting a stationary Markov chain separately on different epochs specified by the user, or (2) bootstrapping blocks of paleodata and using the state information.

[37] The presented framework combines the long paleoreconstructed streamflow information of lesser reliability with the shorter but reliable observational data. The framework has two components: (1) a nonhomogeneous Markov chain model developed on the paleodata that is then used to simulate the hydrologic state, and (2) a K-nearest neighbor (K-NN) time series bootstrap to simulate the streamflow magnitude from the observational data conditioned on the hydrologic state and the previous flow magnitude. This framework combines the respective strengths of the two data sets. Furthermore, it is robust and parsimonious. The framework was applied to paleoreconstructed streamflow and observational data for the Lees Ferry, Arizona, streamflow gauge on the Colorado River. The simulations showed the ability to capture all the distributional statistics of the observational period and also generate a rich variety of wet and dry sequences that will benefit the sustainable management of water resources in the basin.

[38] It is difficult to quantify the significance of combining the paleodata with the observational data in comparison with solely using the observational period data when generating simulations. A Kolmogorov-Smirnov test was used to compare the distributions from both methods, and a significant difference in the overall distributions was not found. However, the distributions do present different tail probabilities that cannot be demonstrated with the Kolmogorov-Smirnov test. These differing tail probabilities present a revised picture of risk that is typically determined with a decision support system. *Prairie* [2006, chapter 5] presents results from a decision support system that demonstrate using the combine data set versus the observational data alone influences decision variables sensitive to tail probabilities such as those affected by extreme events, for instance, protracted drought or surplus.

[39] In the presented results, currently the threshold used to determine system state as well as drought and surplus statistics is based on the median of the observed flow. This threshold can be modified or more states could be included as required on a case by case basis.

[40] A slightly modified version of this technique can be used to generate streamflow sequences based on climate change projections. In this modification the state sequence would be generated using the paleo and observed data and the streamflow magnitudes would be resampled from the PDF of flows from climate change projections.

[41] The annual streamflow generated at Lees Ferry, Arizona, from this approach can be spatially and temporally disaggregated [*Prairie et al.*, 2007] obtaining monthly flow scenarios at all the gauges in the basin. These scenarios are used in a basin-wide decision model [*Prairie*, 2006] and help determine realistic estimations for risk and reliability of various decision components in the water resources system, facilitating effective long-term planning.

[42] **Acknowledgments.** Funding for this research by the Bureau of Reclamation's Lower Colorado regional office via grant 04PG303326 is gratefully acknowledged. Continued support by Kib Jacobson of the Bureau of Reclamation's Upper Colorado regional office is appreciated. Thanks are also due the Center for Advanced Decision Support in Water and Environmental Systems (CADSWES) at the University of Colorado, Boulder, for use of its facilities and computational support.

References

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

Bellone, E., J. P. Hughes, and P. Guttorp (2000), A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts, *Clim. Res.*, *15*, 1–12.

Bureau of Reclamation (2007), Final environmental impact statement: Colorado River interim guidelines for lower basin shortages and coordinated operations for Lakes Powell and Lake Mead, Bur. of Reclam., U.S. Dep. of the Interior, Boulder City, Nev.

Cook, E. R., K. Briffa, S. Shiyatov, and V. Mazepa (1990), Tree-ring standardization and growth-trend estimation, in *Methods of Dendrochronology: Applications in the Environmental Sciences*, edited by E. R. Cook and L. A. Kairiukstis, pp. 153–162, Springer, New York.

Feyerherm, A. M., and L. D. Bark (1965), Statistical methods for persistent precipitation patterns, *J. Appl. Meteorol.*, *4*, 320–328.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Academic, San Diego, Calif.

Gabriel, K. R., and J. Neumann (1962), A Markov chain model for daily rainfall occurrence at Tel Aviv, *Q. J. R. Meteorol. Soc.*, *88*, 90–95.

Gates, P., and H. Tong (1976), On Markov chain modeling to some weather data, *J. Appl. Meteorol.*, *15*, 1145–1151.

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.

Hirschboeck, K. K., and D. M. Meko (2005), A tree-ring based assessment of synchronous extreme streamflow episodes in the Upper Colorado and Salt-Verde-Tonto River Basins: Final report, Lab. of Tree-Ring Res., Univ. of Ariz., Tucson.

Hughes, J. P., and P. Guttorp (1994), A class of stochastic models for relating synoptic-scale 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, *Appl. Stat.*, *48*(1), 15–30.

Koutsoyiannis, D. (1992), A nonlinear disaggregation method with a reduced parameter set for simulation of hydrologic series, *Water Resour. Res.*, *28*(12), 3175–3191.

Koutsoyiannis, D. (2001), Coupling stochastic models of different timescales, *Water Resour. Res.*, *37*(2), 379–391.

Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, *32*(7), 2105–2117.

Kumar, D. N., U. Lall, and M. R. Peterson (2000), Multisite disaggregation of monthly to daily streamflow, *Water Resour. Res.*, *36*(7), 1823–1833.

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

Lambert, M. F., J. P. Whiting, and A. V. Metcalfe (2003), A non-parametric hidden Markov model for climate state identification, *Hydrol. Earth Syst. Sci.*, *7*(5), 652–667.

Lane, W. L., and D. K. Frevert (1990), *Applied Stochastic Techniques Users Manual*, Bur. of Reclam., Denver, Colo.

Loucks, D. P., J. R. Stedinger, and D. A. Haith (1981), *Water Resource System Planning and Analysis*, Prentice-Hall, Upper Saddle River, N. J.

Lu, Z. Q., and L. M. Berliner (1999), Markov switching time series models with application to daily runoff series, *Water Resour. Res.*, *35*, 523–534.

MacDonald, I. L., and W. Zucchini (1997), *Hidden Markov and Other Models for Discrete-Valued Time Series*, CRC Press, 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.

Mehrotra, R., A. Sharma, and I. Cordery (2004), Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrences, *J. Geophys. Res.*, *109*, D14107, doi:10.1029/2004JD004823.

Mejia, J. M., and J. Rousselle (1976), Disaggregation models in hydrology revisited, *Water Resour. Res.*, *12*(2), 185–186.

Meko, D., C. W. Stockton, and W. R. Boggess (1995), The tree-ring record of severe sustained drought, *Water Resour. Bull.*, *31*(5), 789–801.

Prairie, J. (2006), Stochastic nonparametric framework for basin wide streamflow and salinity modeling: Application for the Colorado River basin, Ph.D. dissertation, Univ. of Colo., Boulder.

Prairie, J., and R. Callejo (2005), Natural flow and salt computation methods, U.S. Dep. of Interior, Salt Lake City, Utah. (Available at http://www.usbr.gov/lc/region/g4000/NaturalFlow/NaturalFlowAndSaltComptMethodsNov05.pdf)

Prairie, J., B. Rajagopalan, T. Fulp, and E. Zagona (2005), Statistical nonparametric model for natural salt estimation, *J. Environ. Eng.*, *131*(1), 130–138.

Prairie, J. R., B. Rajagopalan, U. Lall, and T. J. Fulp (2007), A stochastic nonparametric technique for space-time disaggregation of streamflows, *Water Resour. Res.*, *43*, W03432, doi:10.1029/2005WR004721.

Rajagopalan, B., and U. Lall (1995), A kernel estimator for discrete distributions, *J. Nonparametric Stat.*, *4*, 409–426.

Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res.*, *35*(10), 3089–3101.

Rajagopalan, B., U. Lall, and D. G. Tarboton (1996), Nonhomogeneous Markov model for daily precipitation, *J. Hydrol. Eng.*, *1*(1), 33–39.

Rajagopalan, B., U. Lall, and M. A. Cane (1997), Anomalous ENSO occurrences: An alternate view, *J. Clim.*, *10*, 2351–2357.

Roldan, J., and D. A. Woolhiser (1982), Stochastic daily precipitation models: 1. A comparison of occurrence processes, *Water Resour. Res.*, *18*(5), 1451–1459.

Salas, J. (1993), Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by D. Maidment, pp. 19.1–19.72, McGraw-Hill, New York.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane (1980), *Applied Modeling of Hydrologic Time Series*, 484 pp., Water Resour. Publ., Highlands Ranch, Colo.

Santos, E. G., and J. D. Salas (1992), Stepwise disaggregation scheme for synthetic hydrology, *J. Hydraul. Eng.*, *118*(5), 765–784.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley, Hoboken, N. J.

Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res.*, *38*(7), 1100, doi:10.1029/2001WR000953.

Smith, J. A., and H. A. Schreiber (1974), Point processes of seasonal thunderstorm rainfall: 1. Distribution of rainfall event, *Water Resour. Res.*, *10*(3), 418–423.

Srinivas, V. V., and K. Srinivasan (2005), Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows, *J. Hydrol.*, *302*, 307–330.

Stedinger, J. R., and R. M. Vogel (1984), Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, *20*(1), 47–56.

Stedinger, J. R., D. Pei, and T. A. Cohn (1985), A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res.*, *21*(5), 665–675.

Stockton, C. W., and G. C. Jacoby (1976), Long-term surface-water supply and streamflow trends in the Upper Colorado River Basin, *Lake Powell Res. Proj. Bull. 18,* Natl. Sci. Found., Arlington, Va.

Stokes, M. A., and T. L. Smiley (1968), *An Introduction to Tree-Ring Dating*, Univ. of Ariz. Press, Tucson.

Swetnam, T. W., M. A. Thompson, and E. K. Sutherland (1985), *Using Dendrochronology to Measure Radial Growth of Defoliated Trees, Agric. Handb.*, vol. 639, For. Serv., U.S. Dep. of Agric., Washington, D. C.

Tao, P. C., and J. W. Delleur (1976), Multistation, multiyear synthesis of hydrologic time series by disaggregation, *Water Resour. Res.*, *12*(6), 1303–1312.

Tarboton, D. G., A. Sharma, and U. Lall (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*(1), 107–119.

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

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

Thyer, M., and G. Kuczera (2003b), A hidden Markov model for modeling long-term persistence in multi-site rainfall time series: 2. Real data analysis, *J. Hydrol.*, *275*, 27–48.

Timilsena, J., T. C. Piechota, H. Hidalgo, and G. Tootle (2007), Five hundred years of hydrological drought in the Upper Colorado River Basin, *J. Am. Water Resour. Assoc.*, *43*(3), 798–812.

Todini, E. (1980), The preservation of skewness in linear disaggregation schemes, *J. Hydrol.*, *47*, 199–214.

Todorovic, P., and D. A. Woolhiser (1975), Stochastic model of n-day precipitation, *J. Appl. Meteorol.*, *14*(1), 17–24.

Valencia, D. R., and J. C. Schaake (1973), Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, *9*(3), 580–585.

Weisberg, S. (1985), *Applied Linear Regression*, 2nd ed., John Wiley, Hoboken, N. J.

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.

Woolhiser, D. A., and G. G. S. Pegram (1979), Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models, *J. Appl. Meteorol.*, *18*, 34–42.

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

T. Fulp, Bureau of Reclamation, Lower Colorado Region, Boulder City, NV 89006, USA.

U. Lall, Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA.

K. Nowak and B. Rajagopalan, Department of Civil, Environmental and Architectural Engineering, University of Colorado, Boulder, CO 80309, USA.

J. Prairie, Bureau of Reclamation, University of Colorado, 421-UCB, Boulder, CO 80309, USA. (jprairie@uc.usbr.gov)