An enhanced nonparametric streamflow disaggregation model with genetic algorithm

T. Lee,[1] J. D. Salas,[2] and J. Prairie[3]

Received 21 January 2009; revised 1 April 2010; accepted 20 April 2010; published 31 August 2010.

[1] Stochastic streamflow generation is generally utilized for planning and management of water resources systems. For this purpose, a number of parametric and nonparametric models have been suggested in literature. Among them, temporal and spatial disaggregation approaches play an important role particularly to make sure that historical variance-covariance properties are preserved at various temporal and spatial scales. In this paper, we review the underlying features of existing nonparametric disaggregation methods, identify some of their pros and cons, and propose a disaggregation algorithm that is capable of surmounting some of the shortcomings of the current models. The proposed models hinge on k-nearest neighbor resampling, the accurate adjusting procedure, and a genetic algorithm. The models have been tested and compared to an existing nonparametric disaggregation approach using data of the Colorado River system. It has been shown that the model is capable of (1) reproducing the season-to-season correlations including the correlation between the last season of the previous year and the first season of the current year, (2) minimizing or avoiding the generation of flow patterns across the year that are literally the same as those of the historical records, and (3) minimizing or avoiding the generation of negative flows. In addition, it is applicable to intermittent river regimes.

Citation: Lee, T., J. D. Salas, and J. Prairie (2010), An enhanced nonparametric streamflow disaggregation model with genetic algorithm, *Water Resour. Res.*, *46*, W08545, doi:10.1029/2009WR007761.

1. Introduction

[2] Stochastic generation is generally required for planning and management of water resources systems. For river systems involving several sites, the generation model must be capable of reproducing the relationships among the sites in addition to the statistical properties at the individual sites. For this purpose, multivariate models have been proposed in literature such as multivariate autoregressive moving average (ARMA) and multivariate periodic ARMA models [e.g., *Salas et al.*, 1980; *Loucks et al.*, 1981]. These models have been called parametric models in literature. In addition, disaggregation models have been developed such as $\mathbf{Y} = AX + B\mathbf{V}$ of *Valencia and Schaake* [1973], where $\mathbf{Y}$ is a vector representing seasonal data, X is a vector of the corresponding annual data, $\mathbf{V}$ is a vector of independent standard normal noises, and A and B are parameter matrices. A key requirement of this model is that the sum of the disaggregated values adds up to X (for all sites). The foregoing disaggregation model, however, requires estimating a large number of parameters. For this reason, some parsimonious models have been proposed [e.g., *Stedinger and Vogel*, 1984; *Santos and Salas*, 1992]. More recently, *Koutsoyiannis and Manetas* [1996] developed the accurate adjusting procedure (AAP) that combines a model for the lower scale (e.g., monthly) and a model for the higher scale (e.g., yearly) in such a way as to match the generated sequences at each time scale.

[3] The literature of stochastic streamflow simulation has been enriched by the emergence of nonparametric modeling alternatives such as block bootstrapping [*Vogel and Shallcross*, 1996] and k-nearest neighbors resampling (KNNR) [e.g., *Lall and Sharma*, 1996; *Buishand and Brandsma*, 2001]. Alternative nonparametric disaggregation methods have also been proposed such as the method of fragments (MF) [e.g., *Porter and Pink*, 1991; *Srikanthan and McMahon*, 1982; *Svanidze*, 1980], the nonparametric disaggregation (NPD) of *Tarboton et al.* [1998], and the NPD with KNNR (called NPDK) of *Prairie et al.* [2007].

[4] There are pros and cons of both parametric and nonparametric models, and some of these have been reviewed by *Rajagopalan et al.* [2009]. Among them, three issues that are relevant emerge: (1) the preservation of cross-boundary correlations (e.g., in temporal disaggregation, the correlation of the last season of the previous year with the first season of this year may not be preserved), (2) the generation of negative quantities, and (3) the generation of flow patterns that may be repetitive or too close to the historical sequences. The first two shortcomings may occur in both parametric and nonparametric approaches, while the third one generally occurs in nonparametric models.

[5] The issue of cross-boundary correlation was first realized by *Mejia and Rousselle* [1976], who discovered that the Valencia and Schaake (VS) model as originally pro-

[1] INRS-ETE, Quebec, Quebec, Canada.

[2] Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, Colorado, USA.

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

Copyright 2010 by the American Geophysical Union.
0043-1397/10/2009WR007761

posed did not preserve the correlation of the first season of the current year with the seasons of the previous year. Mejia and Rousselle (MR) suggested the additional term needed to fix this shortcoming in the VS model. An alternative disaggregation model developed by *Lane* [1978] and incorporated in the software called LAST essentially hinges on the MR framework. In addition, a practical way to get around the referred shortcoming (in both parametric and nonparametric disaggregation approaches) has been to start the generation at a season where the correlation with the previous season is small or not significant. This remedy, however, is not always sufficient where the correlations are high every season.

[6] Though measured streamflows are positive quantities, parametric and nonparametric models may generate negative values. For example, in parametric models negative values may occur because of the Gaussian assumption of the underlying variable in most typical models such as ARMA. Since hydrologic data are generally skewed, transforming the data into normal is used to comply with the Gaussian assumption, but data transformation does not eliminate the possibility of generating negatives (with some exceptions). Non-Gaussian models are also available, e.g., the periodic gamma autoregressive (PGAR) model [*Fernandez and Salas*, 1986], and *Todini* [1980] suggested using a skewed noise **V** in the VS disaggregation model. However, the PGAR model is only available for single site generation and Todini's suggestion is quite limited in the applicable range of skewness. In nonparametric kernel density-based models such as nonparametric model with long-term dependence (NPL) [*Sharma and O'Neill*, 2002] and NPD [*Tarboton et al.*, 1998], negative values may be generated because Gaussian kernels are employed. Likewise, in NPDK [*Prairie et al.*, 2007], negatives may occur because of the constant linear adjustment utilized. A way to circumvent this problem has been to drop the negatives or regenerate, but this procedure may not be adequate if too many negatives are generated. Another alternative in nonparametric models may be to employ a boundary kernel [*Simonoff*, 1996], but such an alternative has not been applied or implemented in existing nonparametric disaggregation models.

[7] Resemblance of historical patterns in generated data generally occurs in nonparametric techniques such as block bootstrapping [*Vogel and Shallcross*, 1996], KNNR [*Lall and Sharma*, 1996], and NPDK [*Prairie et al.*, 2007]. These approaches generate temporal and spatial flow patterns that are repetitive and close or identical to the historical ones. Generating the same or similar values and patterns is not ideal as argued by *Maheepala and Perera* [1996]. From a practical standpoint, no major problem arises in generating the same pattern(s) since the historical characteristics are reproduced. However, from a statistical (simulation) standpoint, it is just natural to expect that the generated sequences vary enough to be different from those observed in the historical records. In fact, if we had one more year of seasonal streamflows (beyond the historical record), that sequence of seasonal values will be different from any other year of the historical record. To the best knowledge of the authors, there are no ways to get around the referred repetition problem in nonparametric models.

[8] The review of literature reveals the need for making modifications to existing nonparametric disaggregation models so as to avoid the drawbacks identified above, namely, (1) the lack of preservation of cross-boundary correlations, (2) the generation of negatives, and (3) the repetition of the same historical patterns across the year for temporal disaggregation or the same patterns for multiple sites in spatial disaggregation. Thus, we propose appropriate modifications to existing approaches that will eliminate the referred shortcomings. In addition, the proposed approach is applicable for intermittent streamflows, which the NPD and NPDK models cannot adequately simulate. Specifically, we investigate the NPD and NPDK models, identify their limitations, and reveal their similarity to AAP. To surmount their drawbacks, modifications of the NPDK model are suggested and verified. In addition, a genetic algorithm is proposed to ameliorate the alluded repetition problem.

2. Review of Two Existing Disaggregation Approaches

[9] The accurate adjusting procedure (AAP) of *Koutsoyiannis and Manetas* [1996] and the nonparametric disaggregation with KNNR (denoted as NPDK for short) developed by *Prairie et al.* [2007] are reviewed herein. First, we define some notation and useful terms. In stochastic disaggregation, a higher-level value is split into multiple lower-level values in such a way as to preserve the statistical characteristics at both levels. For example, in temporal disaggregation yearly data are disaggregated into seasonal data, and in spatial disaggregation mainstream station data are disaggregated into data at substations. Lower-level variables (e.g., seasons) are denoted as $\mathbf{Y} = (Y_1, \ldots, Y_d)^T$ where d is the number of seasons and X denotes the upper-level (e.g., annual) variable. An important feature in disaggregation lexicon is the additivity property, i.e.,

$$Y_1 + Y_2 + \ldots + Y_d = X. \tag{1}$$

The disaggregation approaches suggested in this paper require initially choosing a candidate lower-level variable set. Then these variables are adjusted to meet additivity. The candidate lower-level variables are denoted as $\tilde{\boldsymbol{Y}} = [\tilde{Y}_1, \ldots, \tilde{Y}_d]^T$ and their sum denoted as $\tilde{X}$. Generally, the notations here are applicable for temporal and spatial disaggregation. In some cases, we will also use $Y_{\nu,\tau}^{(s)}$, where $s = 1, \ldots, S$ (S is the number of sites), $\nu = 1, \ldots, N$ (N is the number of years of record), and $\tau = 1, \ldots, d$. Furthermore, μ_Z and σ_Z represent the mean and standard deviation of random variable Z and $\sigma_{Z_1Z_2}$ represents the covariance between Z_1 and Z_2. Also the lower-case letter y is employed to denote the observed data of the variable Y.

2.1. Accurate Adjusting Procedure

[10] *Koutsoyiannis and Manetas* [1996] developed a scheme for coupling two different models for the lower-level and higher-level variables. Both models are fitted independently, and the generation of data may be summarized as follows:

[11] 1. The higher-level data X (e.g., yearly) is generated from a given model (e.g., ARMA).

[12] 2. The d-dimensional lower-level data $\tilde{\mathbf{Y}} = [\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_d]^T$ are generated from a lower-level model (e.g., periodic ARMA (PARMA)) independently from the generation of the higher-level data.

[13] 3. The sum $\tilde{X} = \sum_{\tau=1}^{d} \tilde{Y}_\tau$ is determined and its distance from the generated data X is calculated as $\Delta = \left|X - \tilde{X}\right|/\sigma(X)$ where $\sigma(X)$ is the standard deviation of X.

[14] 4. If $\Delta > \varepsilon$, where ε is the tolerance level (0.1–1), regenerate the data set $\tilde{\mathbf{Y}}$. Otherwise, the lower-level data are adjusted to meet additivity. The three adjustments are as follows:

$$(\text{linear})\, Y_\tau = \tilde{Y}_\tau + \lambda_\tau \left(X - \tilde{X}\right), \tau = 1, \ldots, d \tag{2}$$

$$(\text{proportional})\, Y_\tau = \tilde{Y}_\tau \left(X/\tilde{X}\right), \tau = 1, \ldots, d, \tag{3}$$

$$(\text{power})\, Y_\tau = \tilde{Y}_\tau \left(X/\tilde{X}\right)^{\lambda_\tau/\eta_\tau}, \tau = 1, \ldots, d, \tag{4}$$

where $\lambda_\tau = \sigma(Y_\tau, X)/\sigma^2(X)$, $\eta_\tau = \mu(Y_\tau)/\mu(X)$, $\sigma(Y_\tau, X) = \mathrm{Cov}(Y_\tau, X)$, and $\mu(Z)$ is the mean of the variable Z.

[15] 5. The steps 1–4 are repeated until all the higher-level data are disaggregated.

[16] The linear adjustment (equation (2)) preserves the mean and the variance-covariance matrix of the lower-level variables [*Koutsoyiannis and Manetas*, 1996]. However, negatives may be generated and higher-order statistics such as skewness may be biased. Therefore, the linear adjustment is applicable where the lower-level variables exhibit small skewness. In addition, *Koutsoyiannis* [1994] showed that the proportional adjustment (equation (3)) is appropriate for lower-level variables that are independent with gamma marginals having the same scale parameter and different shape parameters. Simulation experiments showed that the assumption of independence may be relaxed. Furthermore, proportional adjustment is useful for intermittent data. If a lower-level value is zero, the proportional adjustment does not change the zero unlike the linear adjustment. The power adjustment (equation (4)) is a generalization of the proportional adjustment, but many trials are required to meet additivity [*Koutsoyiannis and Manetas*, 1996]. The linear and proportional adjustments are employed here in our paper using a different modeling framework.

2.2. Nonparametric Disaggregation Model

[17] *Tarboton et al.* [1998] developed a nonparametric disaggregation (NPD) approach. It employs the nonparametric conditional density estimate $f(\mathbf{Y}|X)$. The coordinates of the lower-level variable vector $\mathbf{Y}$ are rotated into a new vector space $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$ by

$$\mathbf{Z} = \mathbf{RY}, \tag{5}$$

where $\mathbf{R}$ is a $d \times d$ rotation matrix, which is obtained from the Gram Schmidt Orthonormalization [*Lay*, 1997]. The rotation estimation procedure guarantees that the last coordinate of Z is aligned perpendicular to the hyperplane in equation (1). Thus, the last element of the rotated variable Z_d is

$$Z_d = X/\sqrt{d}. \tag{6}$$

The essence of the NPD model for generating the $\mathbf{Y}$ values is first generating the Zs and then backrotating from equation (5) so that $\mathbf{Y} = \mathbf{R}^{-1}\mathbf{Z} = \mathbf{R}^T\mathbf{Z}$ where $\mathbf{R}^{-1} = \mathbf{R}^T$ for standard basis.

[18] *Tarboton et al.* [1998] used the multivariate kernel density estimate to generate $Z_1, \ldots, Z_{d-1}$ and Z_d is obtained from equation (6) where X has been previously generated (separately) from a given model. Since using the d-dimensional multivariate density estimate is cumbersome, *Prairie et al.* [2007] employed KNNR to generate the $\mathbf{Z}$ variables instead of generating from a multivariate density estimate. Their disaggregation generation procedure (NPDK) is summarized as follows:

[19] 1. Estimate the $\mathbf{R}$ matrix and obtain $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,d})$, $i = 1, \ldots, N$ (N is the record length, e.g., number of years for temporal disaggregation) from the historical data using equation (5).

[20] 2. The higher-level data X is generated from a specified model. Then, set $Z_d = X/\sqrt{d}$ (equation (6)).

[21] 3. The k-nearest neighbors are obtained from the distances between Z_d and $z_{i,d}$, $i = 1, \ldots, N$, and the k-closest values are chosen. The k-neighbors are assigned weights as,

$$w_m = \frac{1/m}{\sum_{j=1}^{k} 1/j}, m = 1, 2, \ldots, k, \tag{7}$$

where $k = \sqrt{N}$ [*Prairie et al.*, 2007]. Subsequently, one of the k-neighbors is selected by random generation from the discrete weighted distribution obtained from equation (7). Assume that the ℓth value, i.e., the ℓth row (from $i = 1, \ldots, N$) is selected.

[22] 4. Then the d - 1 elements of $\mathbf{Z}$ (i.e., $Z_1, Z_2, \ldots, Z_{d-1}$) are taken from $\mathbf{z}_\ell$ of step 1, i.e., $Z_1 = z_{\ell,1}$, $Z_2 = z_{\ell,2}$, $\ldots$, $Z_{d-1} = z_{\ell,d-1}$. Therefore, $\mathbf{Z} = (z_{\ell,1}, z_{\ell,2}, \ldots, z_{\ell,d-1}, Z_d)$.

[23] 5. The $\mathbf{Z}$ vector is back-rotated into the original space by $\mathbf{Y} = \mathbf{R}^T\mathbf{Z}$.

[24] 6. Steps 2–5 are repeated until the generation length is met.

2.3. Further Examination of the NPD

[25] Without loss of generality, the NPD procedure is investigated for a two-dimensional case. Following *Lay* [1997], it may be shown that

$$\mathbf{R} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}. \tag{8}$$

Then $\mathbf{Z}$ becomes

$$\mathbf{Z} = \mathbf{RY} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_1/\sqrt{2} - y_2/\sqrt{2} \\ y_1/\sqrt{2} + y_2/\sqrt{2} \end{pmatrix}. \tag{9}$$

Then back rotation is performed considering that $Z_2 = X/\sqrt{2}$, which gives

$$\mathbf{Y} = \mathbf{R}^T\mathbf{Z} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} y_1/\sqrt{2} - y_2/\sqrt{2} \\ X/\sqrt{2} \end{pmatrix} = \begin{pmatrix} y_1 + (X - x)/2 \\ y_2 + (X - x)/2 \end{pmatrix}, \tag{10}$$

where $x = y_1 + y_2$ is the historical higher-level value. Equation (10) shows that the NPD procedure distributes the difference between X (generated) and x (historical) equally among the lower-level values. This analysis has been performed for higher values of d with the same results (not included here).

[26] The foregoing analysis of NPD reveals the strong similarity between NPD and the linear adjustment of AAP, which suggests that a simpler NPD procedure may be devised as (1) given the generated higher-level value X, find from the historical lower-level data the set $\mathbf{y} = (y_1, \ldots, y_d)$ whose sum is close to X employing KNNR, and (2) the selected lower-level data are adjusted as

$$\mathbf{Y}^* = \begin{pmatrix} y_1 + (X - x)/d \\ y_2 + (X - x)/d \\ \vdots \\ y_d + (X - x)/d \end{pmatrix}. \tag{11}$$

This simpler procedure gives the same result as that of *Prairie et al.* [2007]. However, additional modifications are needed to avoid (1) the lack of preservation of cross-boundary correlations, (2) the generation of negatives, and (3) the repetition of temporal and spatial historical flow patterns.

3. Model Description

[27] The analysis presented in section 2 showed that the NPDK disaggregation procedure of *Prairie et al.* [2007] performs a linear adjustment with a constant scaling factor $\lambda_\tau = 1/d$ (equation (11)) for all lower levels ($\tau = 1, \ldots, d$) in a similar fashion as the linear adjustment (2) of AAP [*Koutsoyiannis and Manetas*, 1996]. The difference is that NPDK uses KNNR to find a close set of lower-level generated data whose sum is close to the higher-level value while AAP employs a repetition process. Another difference is the scaling factors utilized. From the investigation of the two disaggregation models above, we propose a disaggregation algorithm that will be able to surmount the shortcomings of both. In addition, the proposed approach will be applicable to intermittent river regimes. The procedure proposed in this paper hinges on using parts of AAP and NPD approaches with substantial modifications as summarized below. First, KNNR is employed to find the candidate lower-level values, whose sum is close to the (previously) generated higher-level value. Then the adjusting procedure is applied to meet additivity. One modification is to include the value of the last season of the previous year in selecting the lower-level sequence. Also to avoid generating the same historical pattern in a year, we employ a genetic algorithm (GA) mixture for the lower-level variables as in the work of *Lee* [2008]. Furthermore, the appropriate adjustment will be utilized so that the method will be applicable to intermittent flows.

3.1. Proposed Nonparametric Disaggregation: The KLA and KPA Approaches

[28] The proposed procedure starts by generating the higher-level variable X, then separately (independently) employs KNNR to generate the lower-level sequence (e.g., seasonal data) so that their sum is close to X. The final step is to adjust the disaggregated values to meet additivity. For easy reference, the suggested nonparametric disaggregation based on KNNR with linear adjustment is called KLA and that with proportional adjustment is called KPA. We will describe the procedure with focus on temporal disaggregation (e.g., annual to seasonal). The procedure is also applicable to spatial disaggregation, which is described in section 3.3.

[29] The specific steps of the proposed temporal disaggregation procedure are as follows:

[30] 1. Fit a model to the historical annual data x_i such as ARMA [e.g., *Salas et al.*, 1980], shifting mean [*Sveinsson et al.*, 2003], KNN bootstrapping [*Lall and Sharma*, 1996], the modified KNN [*Prairie et al.*, 2006], and KNN with gamma kernel (KGK) [*Salas and Lee*, 2010]. Then generate an annual series X_ν, $\nu = 1, \ldots, N^G$, where N^G is the generation length.

[31] 2. Consider the first generated annual value X_1 and determine the distances Δ_i between X_1 and the historical annual data x_i, $i = 1, \ldots, N$ (N is the historical record length) as

$$\Delta_i = |X_1 - x_i|, i = 1, \ldots, N \tag{12}$$

and arrange the distances from the smallest to largest one.

[32] 3. Determine the number of nearest neighbors k as $k = \sqrt{N}$, the corresponding weights $w_1, w_2, \ldots, w_k$ from equation (7) and the cumulative weights $cw_m = \sum_{r=1}^{m} w_r$, $m = 1, \ldots, k$. Then take one among the k values of Δ_i by random generation using the cumulative weight distribution cw_m, $m = 1, \ldots, k$. Assume the selected one corresponds to the ℓth year (in the array of the historical seasonal data $y_{i,\tau}$), then the values $y_{\ell,\tau}$, $\tau = 1, \ldots, d$ are the candidate generated data, i.e., $\tilde{\boldsymbol{Y}}_1 = \{\tilde{Y}_{1,1}, \ldots, \tilde{Y}_{1,d}\} = \{y_{\ell,1}, \ldots, y_{\ell,d}\}$ and $\tilde{X}_1 = \sum_{\tau=1}^{d} \tilde{Y}_{1,\tau} = \sum_{\tau=1}^{d} y_{\ell,\tau}$ is the corresponding annual value. The logic is that the seasonal sequences whose sums are closer to X_1 have a higher probability to be chosen according to the weights from equation (7). The GA mixture may be applied to mix the candidate data $\tilde{\mathbf{Y}}_1$ with another data set whose sum is close to X_1. However, for sake of clarity, this additional step is explained separately in section 3.2.

[33] 4. The selected seasonal data set $\tilde{\boldsymbol{Y}}_1 = \{\tilde{Y}_{1,1}, \ldots, \tilde{Y}_{1,d}\}$ are adjusted with a linear or proportional adjusting procedure to obtain the generated lower-level data set $\mathbf{Y}_1 = \{Y_{1,1}, \ldots, Y_{1,d}\}$ so that their sum is equal to X_1 of step 2. For example, for linear adjustment equation (2) gives $Y_{1,\tau} = \tilde{Y}_{1,\tau} + \lambda_\tau(X_1 - \tilde{X}_1)$, where $\lambda_\tau = \sigma(y_{1,\tau} x_1)/\sigma^2(x_1)$. Likewise, for proportional adjustment, equation (3) gives $Y_{1,\tau} = \tilde{Y}_{1,\tau}(X_1/\tilde{X}_1)$.

[34] 5. The next year X_ν generated in step 1 is now considered (e.g., $v = 2$), and we want to generate the corresponding seasonal values. In order to take into account the effect of the last season of the previous year, we use the weighted distances

$$\Delta_i = \sqrt{\varphi_1 (X_\nu - x_i)^2 + \varphi_2 \left(Y_{\nu-1,d} - y_{i-1,d}\right)^2}, \; i = 2, \ldots, N, \tag{13}$$

where $Y_{\nu-1,d}$ is the generated value of the last season of the previous year and $y_{i-1,d}$ is the historical value of the last season of the previous year (respect to year i). The scaling factors φ_1 and φ_2 are determined by $\varphi_1 = 1/\sigma^2(x_i)$ and $\varphi_2 = 1/\sigma^2(y_{i,d})$, respectively. Including the additional term to obtain Δ_i in (13) preserves the relation between the last season of the

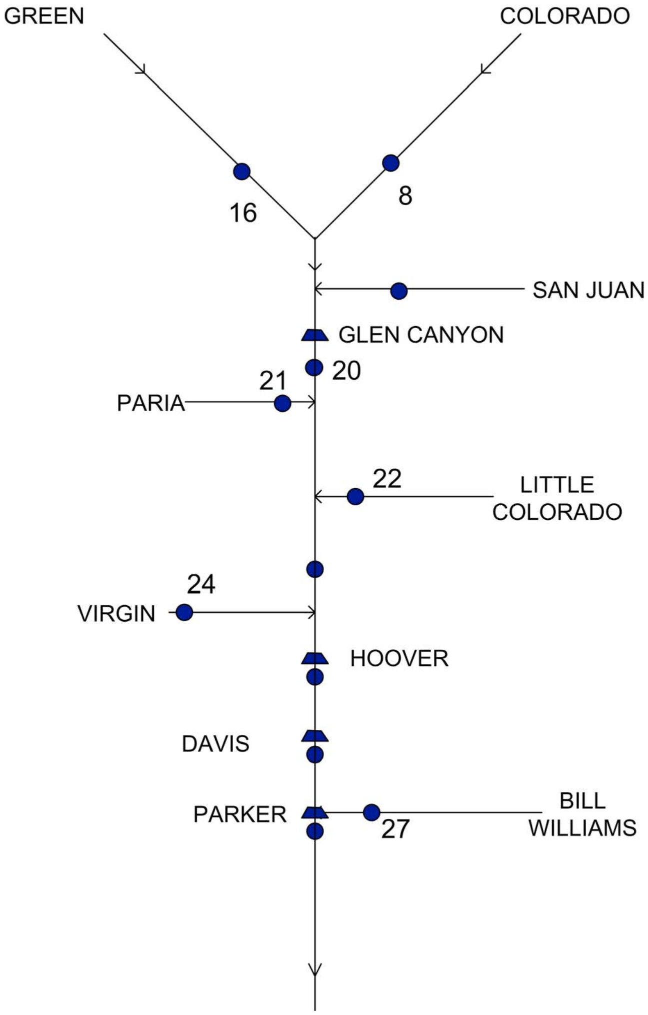


Figure 1. Schematic of part of the Colorado River system where sites numbered 20, 21, 22, 24, and 27 are shown.

previous year and the first season of the current year. Then, from the k smallest values of Δ_i, one is selected at random using the weights as in step 3. This selection will lead to the candidate generated seasonal data $\tilde{\boldsymbol{Y}}_\nu = \{\tilde{Y}_{\nu,1}, \ldots, \tilde{Y}_{\nu,d}\} = \{y_{\nu,1}, \ldots, y_{\nu,d}\}$. This sequence may be mixed using the GA (section 3.2) and then adjusted to obtain the generated seasonal data $\mathbf{Y}_\nu = \{Y_{\nu,1}, \ldots, Y_{\nu,d}\}$. This step is repeated until the generation length N_G is met.

[35] A few remarks are in order. Referring to equation (13) alternatively, one may include other lagged seasons beyond the last season of the previous year. However, such an alternative may degrade the important relation between X_ν and x_i in equation (13). Also the scaling factors φ_1 and φ_2 may be determined by the adaptive Metropolis algorithm suggested by *Mehrotra and Sharma* [2006]. The foregoing alternatives have not been tested in the current paper. In addition, we used the heuristic method for determining the number of nearest neighbors, i.e., $k = \sqrt{N}$, as suggested by *Lall and Sharma* [1996]. It has performed well in previous applications [e.g., *Yates et al.*, 2003]. Alternatively, one may also use the generalized cross validation method for determining k. The variability of the resampled sequences is related to k. If k is too small, the generated values will be too similar to the historical values, which is undesirable because one would expect the generated sequences, while maintaining the historical statistics, to be different from the historical values.

3.2. Mixing With Genetic Algorithm

[36] The disaggregation model suggested above produces repetitive patterns of generated data across the year. This occurs because in the KNNR selection procedure (steps 3 and 5, section 3.1), the entire seasonal sequence for the year is selected as a block. This shortcoming was previously discussed by *Porter and Pink* [1991] and *Lee and Salas* [2008]. Seasonal repetition is not desirable because it contradicts the stochastic nature of hydrological processes, i.e., a variety of seasonal patterns occurs throughout the year and one would expect the generated seasonal sequence of the hydrologic variable at hand to differ from those observed in the historical record. *Lee* [2008] suggested a mixing procedure based on GA in the context of multisite simulation to overcome this problem. Here we apply the concept of mixing in the context of the proposed disaggregation approach to avoid generating patterns identical to the historical ones. Using GA one may apply three processes: reproduction, crossover, and mutation. In our disaggregation procedure, we will only use the crossover process to avoid further changes in the generated data because GA may have some effect on the season-to-season correlations. First, the underlying concepts are explained, and then a summarized procedure is given.

[37] Recall that in steps 3 and 5 above (section 3.1) we obtained the generated seasonal data denoted by $\tilde{\boldsymbol{Y}}_\nu = \{\tilde{Y}_{\nu,1}, \ldots, \tilde{Y}_{\nu,d}\}$ and its corresponding annual data $\tilde{X}_\nu = \sum_{\tau=1}^{d} \tilde{Y}_{\nu,\tau}$. We will rename these variables as $\tilde{\mathbf{Y}}^1_\nu = \{\tilde{Y}^1_{\nu,1}, \ldots, \tilde{Y}^1_{\nu,d}\}$ and $\tilde{X}^1_\nu$ because for purposes of mixing we need to generate another seasonal data set as in step 3 or 5, whose annual value is similar to X^1_ν. Let us denote the second candidate by $\tilde{\boldsymbol{Y}}^2_\nu = \{\tilde{Y}^2_{\nu,1}, \ldots, \tilde{Y}^2_{\nu,d}\}$. The crossover mixing process of GA is performed with either random or competition selection.

[38] Random selection chooses one of two values for the lower-level data (i.e., $\tilde{Y}^1_{\nu,\tau}$ or $\tilde{Y}^2_{\nu,\tau}$) as

$$\tilde{Y}_{\nu,\tau} = \tilde{Y}^1_{\nu,\tau} \qquad \text{if} \qquad u_\tau < p, \tag{14}$$

otherwise $\tilde{Y}_{\nu,\tau} = \tilde{Y}^2_{\nu,\tau}$, where p is a given probability and u_τ is a uniform (0,1) random number. Choosing $p > 0.5$ will give preference to selecting $\tilde{Y}^1_{\nu,\tau}$. One may also vary p with the season τ.

[39] Competition selection of the generated data may be employed for improving the reproduction of the historical season-to-season correlations. For example, if the intent is to increase the lag-1 correlations, one must choose $\tilde{Y}^1_{\nu,\tau}$ if

$$\left[\frac{\tilde{Y}^1_{\nu,\tau}-\hat{\mu}_\tau}{\hat{\sigma}_\tau}\times\frac{y_{\nu,\tau-1}-\hat{\mu}_{\tau-1}}{\hat{\sigma}_{\tau-1}}\right] > \left[\frac{\tilde{Y}^2_{\nu,\tau}-\hat{\mu}_\tau}{\hat{\sigma}_\tau}\times\frac{y_{\nu,\tau-1}-\hat{\mu}_{\tau-1}}{\hat{\sigma}_{\tau-1}}\right], \tag{15}$$

otherwise $\tilde{Y}^2_{\nu,\tau}$ is chosen where $\hat{\mu}_\tau$ and $\hat{\sigma}_\tau$ are the sample mean and standard deviation, respectively, for season τ. One may also combine the criteria (14) and (15) such that $\tilde{Y}^1_{\nu,\tau}$ is selected if either (14) or (15) is met, otherwise $\tilde{Y}^2_{\nu,\tau}$ is chosen. Through testing these two alternatives, i.e., random

Table 1. Basic Monthly Statistics for the Colorado River at Lees Ferry, Site 20[a]

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Yearly
Mean	580.89	480.82	382.53	356.61	393.78	645.20	1199.95	3037.20	4054.34	2190.44	1083.17	671.37	15076.31
Standard deviation	272.01	141.53	95.86	78.63	97.58	211.39	512.46	1146.76	1572.35	1012.25	423.97	309.70	4365.30
Skewness	1.641	1.215	1.223	0.590	1.419	1.081	0.961	0.2713	0.4266	1.1327	0.9464	1.9532	0.1402
Lag-1 correlation	0.558	0.758	0.826	0.703	0.552	0.482	0.470	0.5923	0.6251	0.8311	0.7815	0.6373	0.283
$(\sigma_y/\sigma_x)^{2}$[b] (%)	0.388	0.105	0.048	0.032	0.050	0.234	1.378	6.901	12.974	5.377	0.943	0.503	100

[a] Units for the mean and the standard deviation are in 1000 acre-feet (1 acre-foot = 1.2335×10^3 m^3).

[b] The ratio of the standard deviation of monthly data over that of the yearly data is represented by σ_y/σ_x.

selection based on (14) and competition selection based on (15), it is clear that both have some effect on the seasonal correlations. The general pattern is that using random selection with constant p underestimates the seasonal correlations for some seasons as p increases. Conversely, competition selection has the opposite effect. Therefore, our experiments suggest that a proper combination of the two approaches can achieve quite good results in reproducing the seasonal correlations. One may also obtain good results based on random selection only by varying p (across the year). Either way, one must perform extensive trial and error to find acceptable correlations. This issue is discussed further in section 4.

[40] The summarized GA procedure is described here assuming that in step 5 above, we obtained the generated seasonal data $\tilde{\boldsymbol{Y}}_\nu = \{\tilde{Y}_{\nu,1}, \ldots, \tilde{Y}_{\nu,d}\}$ and the corresponding annual data $\tilde{X}_\nu$. The specific steps are (1) redefine the generated data sets as $\tilde{\mathbf{Y}}_\nu^1$ and $\tilde{X}_\nu^1$, respectively. (2) A second seasonal data set is generated using KNNR that is close to X_ν^1. For this purpose, we find the distances $\Delta_i = |X_\nu^1 - x_i|, i = 1, \ldots, N$ and they are ordered from the smallest to largest one. (3) We use k and the cumulative weight probabilities of equation (7) as in step 3 above. Among the k smallest distances, one is selected at random using the referred weight probabilities. Thus, the year that corresponds to the selected distance defines the seasonal data that is taken from the historical data array. Then the second candidate seasonal sequence is $\tilde{\boldsymbol{Y}}_\nu^2 = \{\tilde{Y}_{\nu,1}^2, \ldots, \tilde{Y}_{\nu,d}^2\}$ whose annual total is close to X_ν^1. (4) The two data sets $\tilde{\mathbf{Y}}_\nu^1$ and $\tilde{\mathbf{Y}}_\nu^2$ are mixed with GA to create the new seasonal data set, say $\tilde{\mathbf{Y}}_\nu^{GA}$. For this purpose, we use the random selection criterion (14), the competition selection criterion (15), or a combination of the two. As noted above, the proposed disaggregation models with linear and proportional adjustments have been denoted as KLA and KPA, respectively. To distinguish where the addition of the genetic algorithm is made, they are identified as KLAG and KPAG disaggregation approaches, respectively.

3.3. Nonparametric Procedure for Spatial Disaggregation

[41] The procedure for spatial disaggregation is similar to that for temporal disaggregation, but for ease of the reader, we summarize it assuming that we wish to disaggregate the yearly streamflows at a key station (say downstream) into yearly streamflows at d substations (upstream). Let the annual variable at the key station be denoted by X_ν and the corresponding variables at substations by $Y_\nu^{(s)}$, $s = 1, \ldots, d$, where s represents the station and d is the total number of stations. Thus, under the foregoing assumptions equation (1) applies, i.e., $Y_\nu^{(1)} + Y_\nu^{(2)} + \ldots + Y_\nu^{(d)} = X_\nu$.

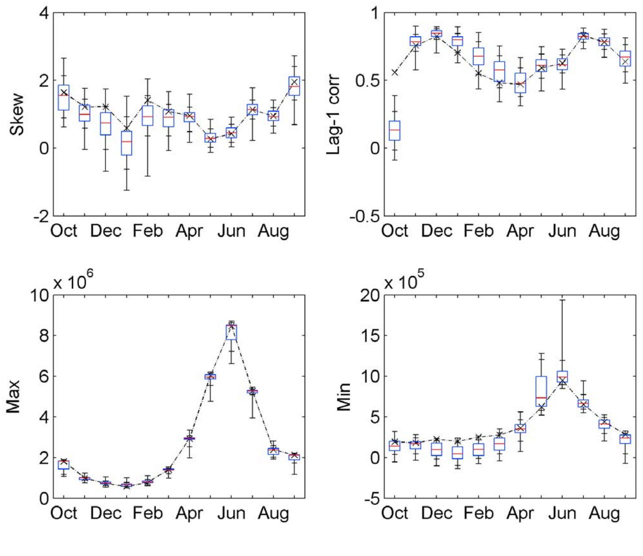


Figure 2. Box plots of the monthly statistics of the Colorado River at site 20 estimated from the generated (disaggregated) data based on the NPDK model and corresponding historical statistics (cross marks and dash-dotted line). The units for max and min are in acre-feet (1 acre-foot = 1.2335×10^3 m^3).

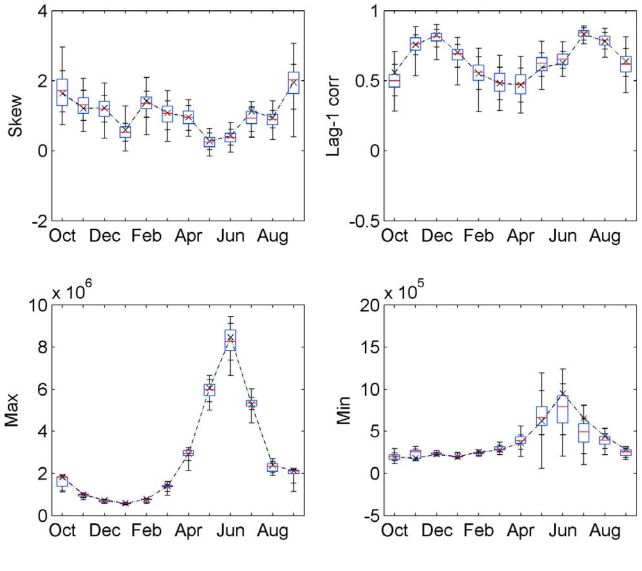


Figure 3. Box plots of the monthly statistics of the Colorado River at site 20 estimated from the generated (disaggregated) data based on the KLA model and corresponding historical statistics (cross marks and dash-dotted line). The units for max and min are in acre-feet.

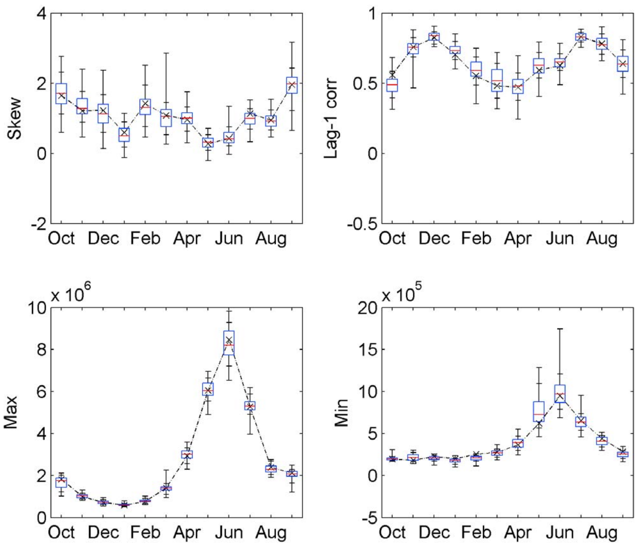


Figure 4. Box plots of the monthly statistics of the Colorado River at site 20 estimated from the generated (disaggregated) data based on the KPA model and corresponding historical statistics (cross marks and dash-dotted line). The units for max and min are in acre-feet.

[42] The specific steps of the proposed spatial disaggregation procedure are as follows:

[43] 1. Fit a model to the historical key station data x_i, $i = 1, \ldots, N$ (N is historical record length). Then generate the series X_ν, $\nu = 1, \ldots, N^G$, where N^G is generation length.

[44] 2. Consider X_ν and determine the distances $\Delta_i = |X_\nu - x_i|$, $i = 1, \ldots, N$ and arrange them from the smallest to the largest one. Determine the number of nearest neighbors $k = \sqrt{N}$ and take one among the k values of Δ_i by random generation using the cumulative weight distribution as in equation (7). Assuming the selected one corresponds to the ℓth year, the values of the historical data of the substations for year ℓ are the candidate generated data, i.e., $\tilde{\mathbf{Y}}_\nu = \{\tilde{Y}_\nu^{(1)}, \ldots, \tilde{Y}_\nu^{(d)}\} = \{y_\ell^{(1)}, \ldots, y_\ell^{(d)}\}$ and $\tilde{X}_\nu = \sum_{s=1}^{d} \tilde{Y}_\nu^{(s)}$. If GA mixture is desired, follow steps 1–4 as in section 3.2, otherwise skip to step 3.

[45] 3. The disaggregated data at the substations $\tilde{\mathbf{Y}}_\nu = \{\tilde{Y}_\nu^{(1)}, \ldots, \tilde{Y}_\nu^{(d)}\}$ are adjusted with linear or proportional relation as in equation (2) or (3), respectively, to obtain the generated data at substations $\mathbf{Y}_\nu = \{Y_\nu^{(1)}, \ldots, Y_\nu^{(d)}\}$ so that their sum is equal to X_ν of step 2.

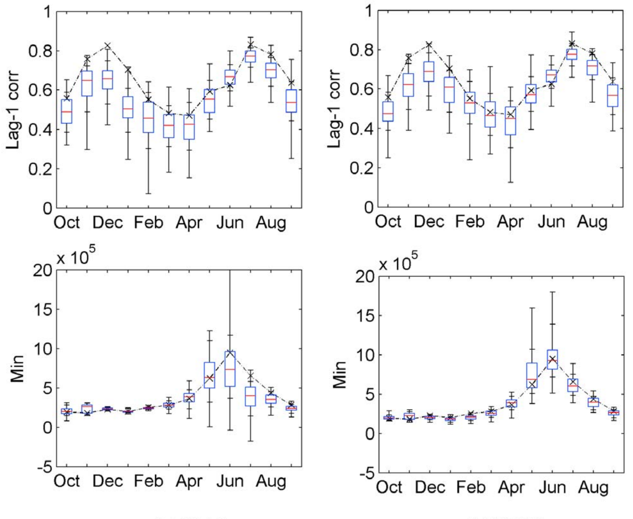
[46] 4. Repeat steps 2–4 for all $\nu = 1, \ldots, N_G$.

[47] It must be noted that the foregoing procedure assumes that the sum of the flows of the substations must be equal to the flow at the key station. Otherwise, two options can be followed. Either create an artificial substation so that the sum of the flows at substations plus that of the artificial station is equal to the key station flows (thus the algorithm above applies) or modify the procedure above by using some type of spatial adjustment (e.g., refer to the work of *Sveinsson et al.* [2009]).

4. Data Description, Model Assessment, and Applications

[48] The proposed models are verified using data of the Colorado River system. The Colorado is a major river system in the western United States, and the Bureau of Reclamation uses 29 gaging sites within the system for long-term planning studies. Relevant information on the Colorado River system can be found in the work of *Lee and Salas* [2006]. We illustrate the proposed nonparametric disaggregation models based on site 20 of the Colorado River at Lees Ferry station (referred to as site 4 in the work of *Prairie et al.* [2007]) for application of temporal disaggregation and sites 21 (Paria River at Lees Ferry, AZ), 22 (Little Colorado River near Cameron, AZ), 24 (Virgin River at Littlefield, AZ), and 27 (Bill Williams River below Alamo Dam, AZ) for application of spatial disaggregation. The locations of the sites used in this study are shown in Figure 1. The historical gaged data have been naturalized [*Prairie and Callejo*, 2005], and part of the data has been extended using linear regression and nonparametric bootstrapping [*Lee and Salas*, 2006] so that all sites have data for the period 1906–2003. The complete data set may be found at the Web site http://www.usbr.gov/lc/region/g4000/NaturalFlow/index.html. Because the Colorado River management model runs with data in English units, the streamflow data employed in our study is based on acre-feet (1 acre-foot = 1.2335 × 10^3 m^3).

[49] One hundred samples are generated from each model with the same length as the historical data (98 years). Various basic statistics are estimated from the historical and generated data to verify the model performance such as mean, standard deviation, skewness, maximum, minimum, and lag-1 correlations at the seasonal and yearly time scales. Box plots are employed to display the variability of the generated statistics. The end line of the box (interquartile range) indicates the 25 and 75 percentiles, while the cross lines above the box on the whisker denotes the 90 percentile and the maximum and the cross lines below the box denotes the 10 percentile and the minimum. The dash-dotted line connecting the cross marks represents historical statistics. The cross or serial relations in the generated data are



(a) KLAG (b) KPAG

Figure 5. Box plots of the monthly lag-1 correlations and minimums of the Colorado River at site 20 derived from the generated (disaggregated) data based on the models (left) KLAG and (right) KPAG. The historical values are also shown (cross marks and dash-dotted line). The units for max and min are in acre-feet.

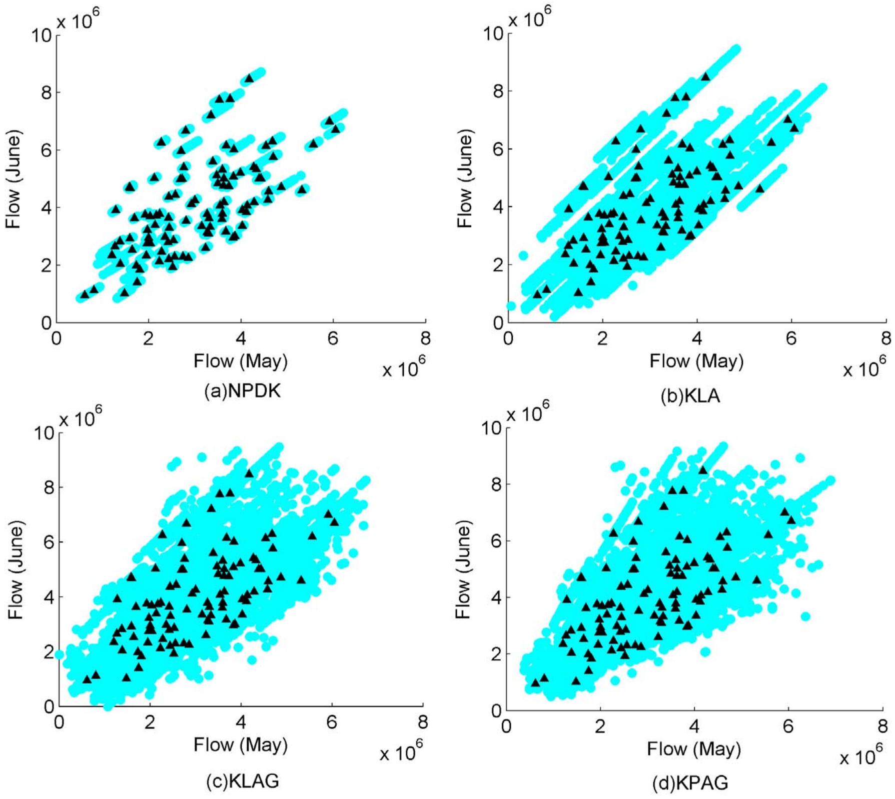


Figure 6. Scatterplots of the flows of the Colorado River at site 20 for the months of May (horizontal) and June (vertical) for the historical data (triangles) and generated data (gray circles) obtained from the models (a) NPDK, (b) KLA, (c) KPLG, and (d) KPAG. The units are in acre-feet.

checked using scatterplots. Furthermore, operational statistics such as maximum drought and surplus amounts and lengths and storage capacity based on water demand levels that are fractions of the historical mean (0.7–1.0) were also compared.

4.1. Temporal Disaggregation

[50] For temporal disaggregation, yearly and monthly data of the Colorado River at site 20 have been used to validate the performance of the proposed model and compared to that of the NPDK model of *Prairie et al.* [2007]. The basic monthly and yearly statistics of the historical data are shown in Table 1. The last row shows the ratios of the monthly variances divided by the variance of the yearly data. They indicate that the wet months (MJJ) contribute most of the yearly variance while the contribution of the dry months is much smaller. The KGK model developed by *Salas and Lee* [2010] was employed for the yearly data generation. In turn these data were disaggregated using the models suggested in this paper and the existing NPDK model. Overall five types of models were tested: (1) nonparametric disaggregation model with KNN (NPDK); (2) disaggregation with KNN and linear adjustment, KLA; (3) KLA with genetic algorithm, KLAG; (4) disaggregation with KNN and proportional adjustment, KPA; and (5) KPA with genetic algorithm, KPAG. From the five models, various test statistics were estimated and compared. The comparison of the historical and generated annual statistics (not shown) indicated that the KGK model reproduced the historical basic statistics, surplus, and storage capacity statics quite well (indicated by the historical statistic falling within the interquartile range), although some overestimation was noted in the drought statistics.

[51] First, the monthly statistics of the historical and generated data from models NPDK and KLA were compared. The monthly means and standard deviations are well reproduced by both models (not shown). The monthly skewnesses, lag-1 correlations, maximums, and minimums are shown in Figures 2 (NPDK) and 3 (KLA). Figure 2 for the NPDK model shows some underestimation of the monthly skewness for several months, which is also reflected in the minimum flows for the same months. Likewise, Figure 2 shows a significant underestimation of the correlation of October of the current year with September of the previous year since NPDK has no structure linking this year with the previous year. This shortcoming has been fixed by adding one more term in the value selection (13). The improvement of the lag-1 correlations is clearly shown in Figures 3 and 4 for the KLA and KPA models, respectively. Also Figure 5 shows the results of the monthly lag-1 correlations and minimum for the KLAG and KPAG models (models with GA mixing where $p = 0.5$ in (14)). In these cases one may note that while the lag-1 correlation for the first month has been greatly improved with respect to that of NPDK (Figure 2), some underestimation of the correlations occurs for other months. The effect of the parameter p in (14) for the GA mixture was further examined. We varied the value of p and compared the effects on the lag-1 correlations and found that a small value of p such as 0.05 or 0.10 yield

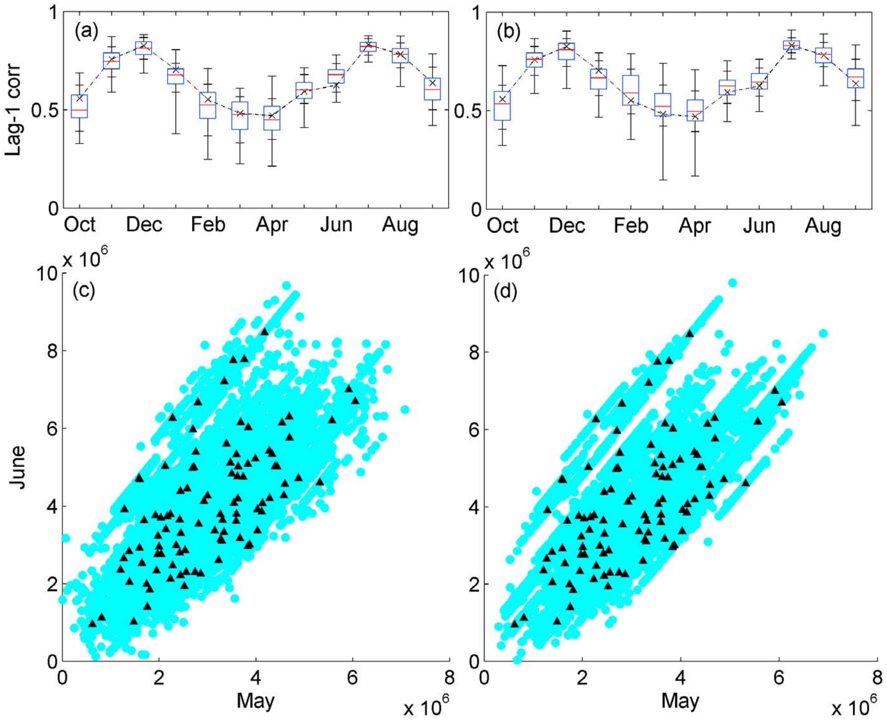


Figure 7. (a and b) Lag-1 correlations and (c and d) scatterplots between flows of May and June of KLAG for Colorado River station 20 for (left) method 1 random selection as in (16) with varying p and (right) method 2 combination of random selection (14) and competition (15). The crossover probabilities used for each month are method 1, p = [0, 0, 0, 0.05, 0.1, 0.1, 0.1, 0.3, 0.5, 0.1, 0.1, 0.1] and method 2, p = [0.8, 0.8, 0.8, 0.4, 0.5, 0.5, 0.05, 0.05, 0.05, 0.5, 0.5, 0.5]. The units for the scatterplots are in acre-feet.

reasonable reproduction of the lag-1 monthly correlations and still provide a good mixing of the data. Alternatively, as pointed out in section 3.2 above, one could obtain very good reproduction of seasonal correlations either by varying the parameter p with the season or combining the two methods of GA mixing, i.e., random selection and competition (this issue is further discussed below).

[52] Negatives may be generated by the NPDK model as shown in Figure 2. The low flow months (DJFM) are highly affected resulting in negative values. This occurs because NPDK uses the linear adjustment (11) where the difference between the generated annual value and the summation of the selected monthly values is distributed equally among the months without considering the variability of the individual months (Table 1.) Thus, the high flow months are not affected much from the adjustments while the low flow months are highly affected and result in higher variability. This is the main reason for the bias produced by the NPDK model. In contrast the KLA model does not produce any negative values as shown in Figure 3 since the difference of the historical and generated yearly values are distributed considering the contribution of monthly covariances (as in equation (2)). Our experience shows that linear adjustments are not convenient where the data are highly skewed because negative values may occur even with KLA. In addition, Figure 4 shows that the minimum values are better preserved with the KPA model. This model guarantees that no negative values are generated (unless negative values occur in the historical data). Note that some underestimations of the minimum flows may still occur with the KLA and KLAG models (Figures 3 and 5) especially for June and July although with less chance of generating negative values. On the other hand, the KPA and KPAG models show better preservation of the minimum values (Figures 4 and 5).

[53] The scatterplot in Figure 6 displays the relationship between the flows for May (horizontal) and June (vertical) obtained for the historical and generated data based on models NPDK, KLA, KLAG, and KPAG. Figure 6a shows that the generated data are always extremely close to the historical values for the NPDK model, which indicates the repetitious data pattern generated by NPDK. Figure 6b also reveals the directional patterns of the data generated based on KLA model and occurs for data generated based on KPA model (not shown). For both KLA and KPA models, the variations of generated flows are more desirable (wider) than those obtained by NPDK model but still the pattern is directional. On the other hand, the introduction of GA mixture produces generated data with wider spread as shown in Figure 6c for model KLAG and Figure 6d for KPAG. As mentioned in section 3.2, the underestimation of the lag-1 cross correlation in applying GA can be remedied by either (1) varying each month the crossover probability p of the random selection (14) or (2) employing the combination of the random and competition selections. On the basis of both methods 1 and 2, we searched the optimal p value (i.e., as close as possible to the historical lag-1 cross correlation) by trial and error. The results of methods 1 and 2 are shown in Figures 7 (left) and 7 (right), respectively. The lag-1 corre-

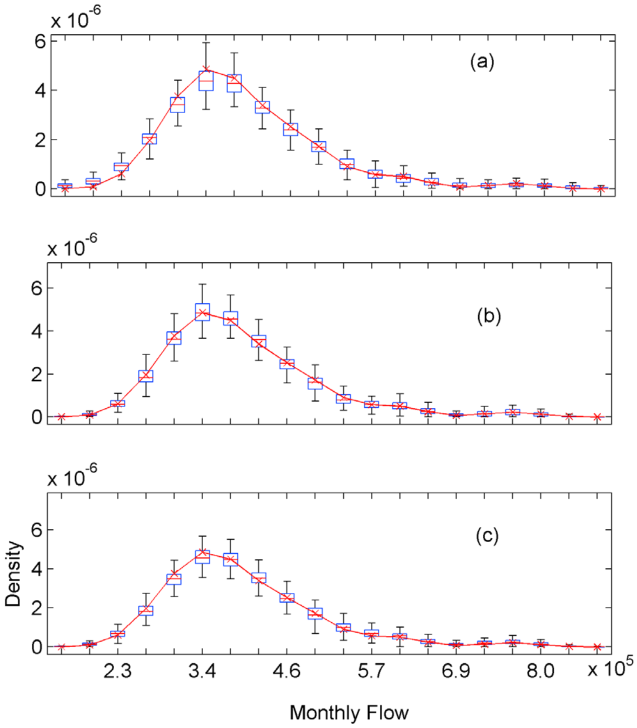


Figure 8. Kernel density estimates using normal kernel for the month of February of the Colorado River at site 20 obtained from the generated data based on models (a) NPDK, (b) KLA, and (c) KLAG and from the historical data (cross marks and dash-dotted line). The units for the density are in 1/acre-feet.

lations using KLAG are shown in Figures 7a and 7b and a scatterplot for May (horizontal) and June (vertical) flows are shown in Figures 7c and 7d. The lag-1 correlations of all months are well preserved in both cases. The generated values are properly distributed (Figures 7c and 7d). Less spread is observed in Figure 7d than in Figure 7c because for the months in question the *p* values for method 2 were smaller than those used for method 1.

[54] The densities of the historical and generated monthly data estimated with a normal kernel and the asymptotic optimal bandwidth [*Simonoff*, 1996] are shown in Figure 8 for February for models NPDK (a), KLA (b), and KLAG (c). For the NPDK model, the density around the mode is somewhat underestimated while it is overestimated in the low flow range. Slight underestimation is also observed for the KLAG model near the mode caused by the GA mixture. The density estimates for the other months are reasonably well preserved for all models.

[55] The ratios of drought, surplus, and storage statistics obtained from the monthly historical and generated data for the various models were also compared. To some extent, the behavior of those statistics depends on the generated yearly data. As mentioned above the alternative disaggregation models are based on the same yearly generated data (from the KGK model). For illustration, Figure 9 shows the KPA model box plots obtained for the various statistics for water demand levels that are fractions of the historical mean (0.7–1.0). It shows that generally the historical statistics are well reproduced. Some underestimation of the drought statistics for threshold levels 0.8 and 0.9 occurs but the results are reasonable for thresholds 0.7 and 1.0. Likewise, for the storage capacity some overestimation occurs for levels 0.8 and 0.9, but the results are reasonable for 0.7 and 1.0. There are not many differences in the results obtained for the other models (not shown).

4.2. Spatial Disaggregation

[56] To demonstrate spatial disaggregation, we use the tributary sites of the Lower Colorado River system (sites 21, 22, 24, and 27). Two of these sites (22 and 27) are intermittent so the group selected is a combination of nonintermittent and intermittent flow sites. These data are highly skewed, not only at the monthly but also at the yearly scales (Table 2). For this application, we created an index station that is the sum of the flows at the referred four sites. Table 2 shows that site 21 has the lowest contribution of variance (only about 0.07%) with respect to the variance of the total flow at the index station. The yearly data of the index station was generated using the KGK model [*Salas and Lee*, 2010]. From the yearly data, the monthly data of the index station were obtained applying the nonparametric temporal disaggregation with proportional adjustment and GA mixing (KPAG). Then the monthly flows at the index station were spatially disaggregated to obtain the monthly flows at the referred four sites. Both the NPDK and KPA models were compared for the spatial disaggregation step.

[57] We compared the monthly statistics of the four sites after spatial disaggregation of the monthly flows at the index station using both models. Figure 10 shows the box plots obtained for the monthly means and minimums for site 21. Significant biases are observed in the referred statistics for the NPDK model. Such biases especially occur for the smaller streams (e.g., site 21) where the means and standard deviations are much smaller than for the other sites 22, 24, and 27 (Table 2). Also, a significant number of negatives are generated as shown in the plot for minimum flows in Figure 10. On the other hand, the monthly means for site 21 are well preserved by the KPA model and no negatives are generated although during the dry months some negative biases are observed but the results are realistic. Also the other basic statistics obtained from NPDK are significantly biased for site 21. However, for sites 22, 24, and 27, the key basic statistics obtained from both NPDK and KPA models are comparable except that a significant number of negatives are generated by the NPDK model. Likewise, the monthly cross correlations between flows at sites 21 and the other sites are not well reproduced for many months by the NPDK model. For example, the monthly correlations between sites 21–22 and 21–24 are shown in Figure 11 for both NPDK and KPA models. In general, the correlations are well reproduced by the KPA model but for many months the correlations are either significantly underestimated or overestimated by the NPDK model. However, the cross correlations between sites 22–24, 22–27, and 24–27 are reasonably well reproduced by both models. We also compared the yearly basic statistics derived from the monthly flows for the four sites. Again for site 21 the annual statistics are not reproduced by NPDK model. The results for the other sites are comparable and reasonably well reproduced, although for site 27 significant underestimation of the standard deviation, skewness, and maximum occurs using both models.

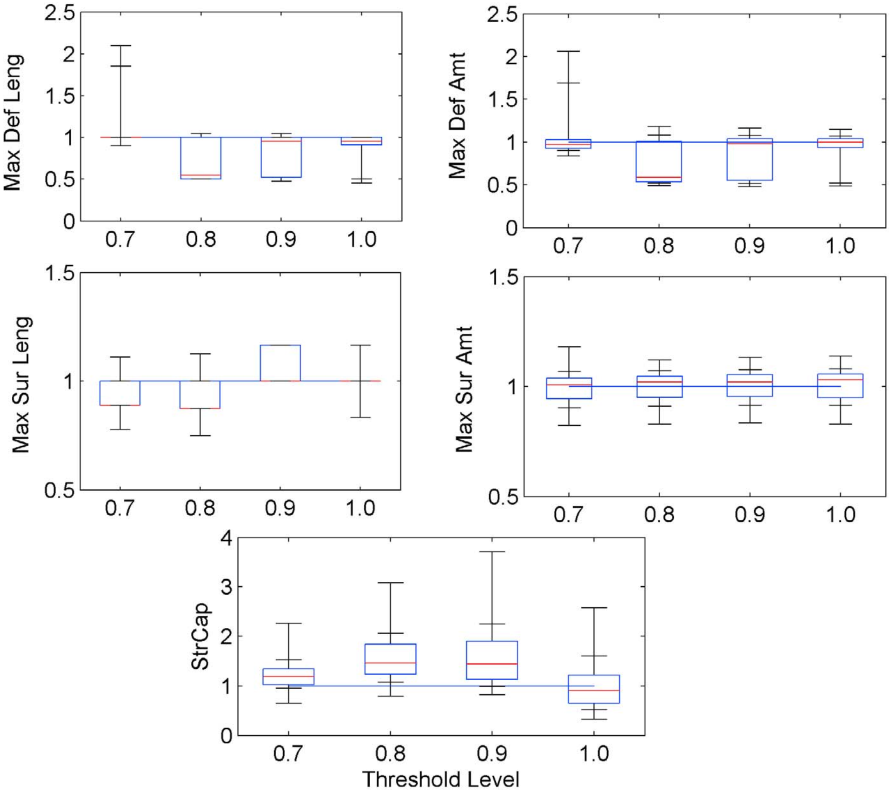


Figure 9. Box plots of ratios of drought, surplus, and storage statistics (for demand threshold levels 0.7–1.0) derived from the generated monthly data over the corresponding historical statistic for the Colorado River at site 20. The generated data were based on the KPA model. Departures from the horizontal line at ratio equal to 1 suggest some degree of bias. Max Def Leng, maximum deficit length; Max Def Amt, maximum deficit amount; Max Sur Leng, maximum surplus length; Max Sur Amt, maximum surplus amount; StrCap, storage capacity.

[58] Furthermore, drought, surplus, and storage capacity statistics estimated based on monthly and annual flows were also compared for water demand levels in the range 0.7–1.0. Box plots of the ratios of the statistics obtained from generated data divided by the corresponding historical statistics were used for comparison. For example, Figures 12 and 13 show for the yearly data the box plots obtained for site 21 based on the NPDK and KPA models, respectively. As occurred with other monthly and yearly statistics for site 21, the results from the NPDK model are poor especially for the maximum surplus amount, which is significantly overestimated. The results for the other sites are quite reasonable and comparable for both models except for site 27 where significant underestimation for the drought and surplus amounts and storage capacity were obtained.

5. Summary and Conclusions

[59] For several decades, stochastic disaggregation has been a valuable tool for generating hydrologic data. It has been useful for temporal and spatial disaggregation. For this purpose, parametric and nonparametric disaggregation approaches have been suggested in literature. From reviewing the existing nonparametric disaggregation models and uncovering their pros and cons, we suggest improved disaggregation models that overcome some of the shortcomings of the existing models. The proposed disaggregation models hinge on generating the higher-level variable *X* (e.g., annual data) based on any parametric or nonparametric model, then independently applying KNNR for generating the lower-level sequence *Y* (e.g., seasonal data) in such a way that their sum is close to the generated data *X*. Then the lower-level values are adjusted to meet additivity. For this purpose, linear and proportional adjustments may be applied. In addition, genetic algorithm mixing was suggested to avoid generating the same historical pattern in a year or in space as the case may be. We recommend the proportional adjustment models (KPA or KPAG) in the case where data without negative values are highly skewed, such as the tributary streams of the Lower Colorado River system. On the other hand, we recommend the linear adjustment models (KLA or KLAG) in the case of

Table 2. Basic Yearly Statistics for the Colorado River Basin at Sites 21, 22, 24, and 27[a]

	Site 21	Site 22	Site 24	Site 27	Index Site
Mean	21.12	180.41	169.97	98.19	469.69
Standard deviation	8.31	140.40	88.27	125.02	314.34
Skewness	0.839	2.008	1.677	2.673	1.98
Lag-1 correlation	0.146	−0.038	0.061	0.061	0.01
$(\sigma_y/\sigma_x)^2$[b] (%)	0.070	19.950	7.885	15.818	100

[a] Units for the mean and the standard deviation are in 1000 acre-feet (1 acre-foot = 1.2335 × 10^3 m^3).

[b] The ratio of the standard deviation of monthly data over that of the yearly data is represented by σ_y/σ_x.

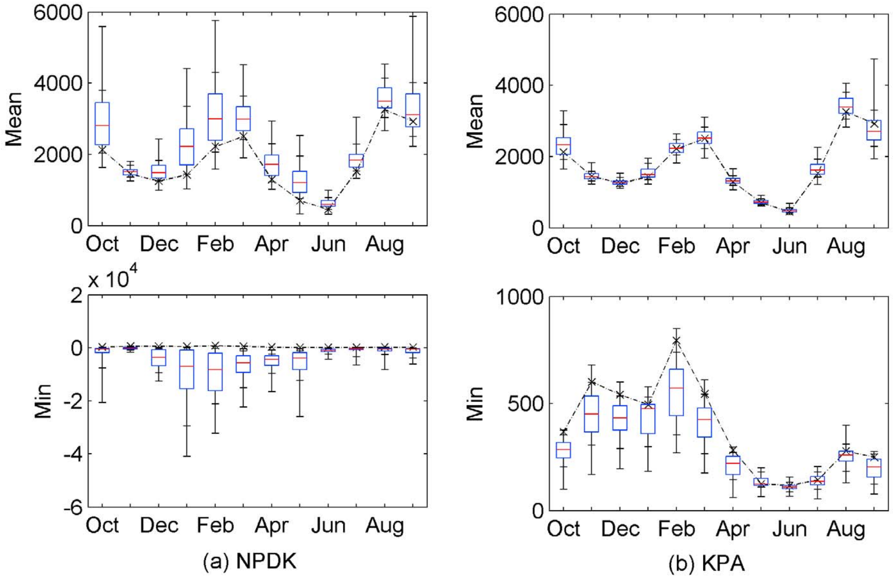


Figure 10. Box plots of monthly means and minimums obtained from the generated (monthly) data for site 21 of the Colorado River based on models (a) NPDK and (b) KPA and from the historical data (cross marks and dash-dotted line). The units are in acre-feet.

data with small skewness and/or negative values, such as for intervening flows of the Colorado River system.

[60] The proposed models for temporal and spatial disaggregation have been tested and applied using data of the Colorado River system. The applications indicate that the suggested modeling procedure gives reasonable and improved results compared to existing nonparametric disaggregation approaches. In particular the proposed models overcome the drawbacks mentioned in the paper by *Prairie et al.* [2007], such as the inability to capture the correlation between the first month of the current year and the last month of the previous year (for temporal disaggregation) and the proper preservation of extrema (minimum and maximum). The former is overcome by including the variable of the last month of the previous year in the KNNR selection and the latter is accomplished by using the concept of accurate adjusting. In addition, the proposed disaggregation models have the ability to model jointly intermittent and nonintermittent variables using proportional adjustment. Furthermore, more variable flow patterns can be obtained using the genetic algorithm mixture. A drawback employing this algorithm is the possible underestimation or

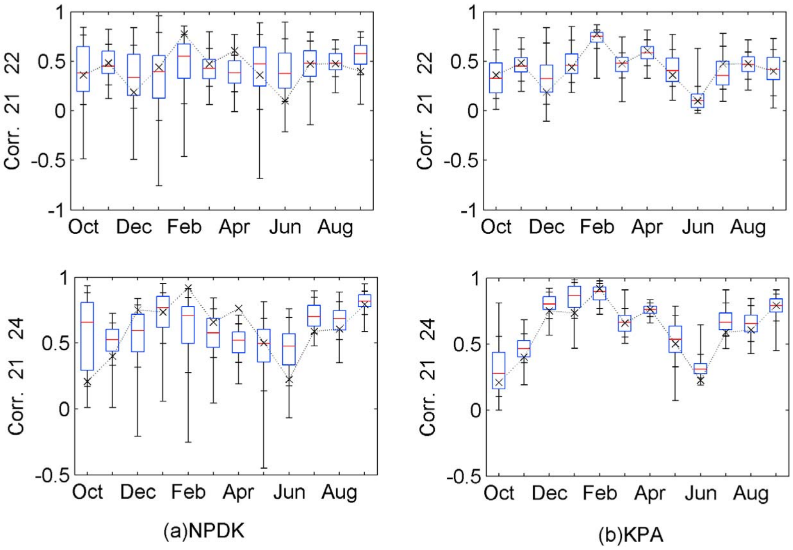


Figure 11. Box plots of the monthly lag-0 cross-correlations between sites 21–22 and 21–24 of the Colorado River obtained from the generated data based on models NPDK and KPA and from the historical data (cross marks and dash-dotted line).

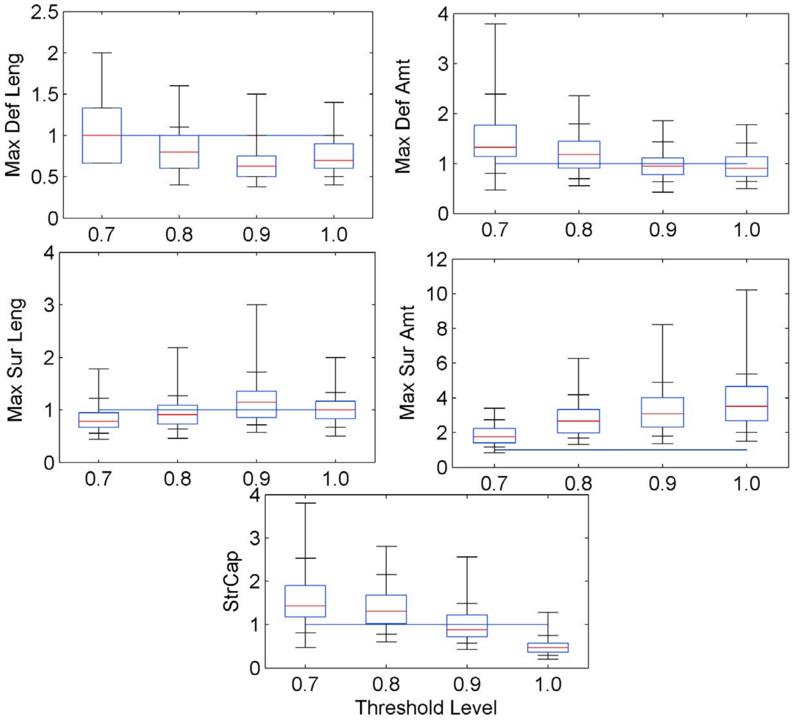


Figure 12. As in Figure 9, except that site 21 is employed and the applied model is NPDK to produce the figure.

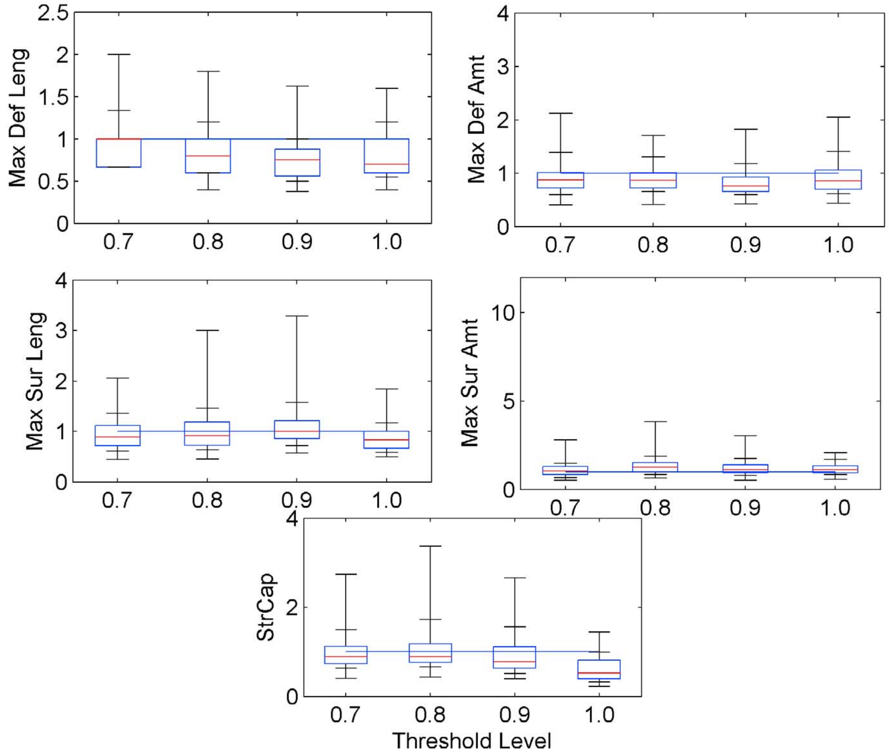


Figure 13. As in Figure 12, except that the applied model is KPA to produce the figure.

overestimation of temporal or spatial correlations (as the case may be). However, we showed that by trial and error one can reproduce such correlations quite well.

[61] **Acknowledgments.** The authors wish to acknowledge the financial support of the Bureau of Reclamation Lower Colorado Region, Contract "Development of Stochastic Hydrology for the Colorado River System." Special thanks to D. Frevert, B. Lane, and T. Fulp for their technical assistance and support throughout the study. The comments of three anonymous reviewers and the editor and associate editor improved an earlier version of the manuscript extensively.

References

Buishand, T. A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resour. Res.*, *37*, 2761–2776, doi:10.1029/2001WR000291.

Fernandez, B., and J. D. Salas (1986), Periodic gamma autoregressive processes for operational hydrology, *Water Resour. Res.*, *22*, 1385–1396, doi:10.1029/WR022i010p01385.

Koutsoyiannis, D. (1994), A stochastic disaggregation method for design storm and flood synthesis, *J. Hydrol.*, *156*, 193–225.

Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, *32*, 2105–2117, doi:10.1029/96WR00488.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, *32*, 679–693, doi:10.1029/95WR02966.

Lane, W. L. (1978), *Applied Stochastic Techniques (LAST Computer Package): User Manual*, Bureau of Reclam., Denver, Colo.

Lay, D. C. (1997), *Linear Algebra and Its Applications*, 2nd ed., 486 pp., Addison-Wesley, Reading, Mass.

Lee, T., and J. D. Salas (2006), Record extension of monthly flows for the Colorado River system, 155 pp., Bur. of Reclam., U.S. Dept. of the Inter. (Available at http://www.usbr.gov/lc/region/g4000/NaturalFlow/Final.RecordExtensionReport.2006.pdf.)

Lee, T., and J. D. Salas (2008), Periodic stochastic model for simulating intermittent monthly streamflows of the Colorado River system, paper presented at World Environmental & Water Resources Congress 2008, Honolulu, Hawaii, 12–16 May.

Lee, T. S. (2008), Stochastic simulation of hydrologic data based on nonparametric approaches, Ph.D. dissertation, 346 pp., Colo. State Univ., Fort Collins, Colo.

Loucks, D. P., J. R. Stedinger, and D. A. Haith (1981), *Water Resources Systems Planning and Analysis*, 559 pp., Prentice-Hall, Englewood Cliffs, N. J.

Maheepala, S., and B. J. C. Perera (1996), Monthly hydrologic data generation by disaggregation, *J. Hydrol.*, *178*, 277–291.

Mehrotra, R., and A. Sharma (2006), Conditional resampling of hydrologic time series using multiple predictor variables: A K-nearest neighbour approach, *Adv. Water Resour.*, *29*, 987–999.

Mejia, J. M., and J. Rousselle (1976), Disaggregation models in hydrology revisited, *Water Resour. Res.*, *12*, 185–186, doi:10.1029/WR012i002p00185.

Porter, J. W., and B. J. Pink (1991), A method of synthetic fragments for disaggregation in stochastic data generation, paper presented at International Hydrology and Water Resources Symposium 1991, Inst. of Eng., Perth, West. Aust., Australia, 2–4 Oct.

Prairie, J. R., B. Rajagopalan, T. J. Fulp, and E. A. Zagona (2006), Modified K-NN model for stochastic streamflow simulation, *J. Hydrol. Eng.*, *11*, 371–378.

Prairie, J., and R. Callejo (2005), Natural flow and salt computation methods, 122 pp., Bur. of Reclam., U.S. Dept. of the Inter., Salt Lake City, Utah.

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

Rajagopalan, B., J. Salas, and U. Lall (2009), Stochastic methods for modeling precipitation and streamflow, in *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*, edited by B. Sivakumar and R. Berndtsson, World Sci., Singapore.

Salas, J. D., and T. Lee (2010), Non-parametric simulation of single site seasonal streamflows, *J. Hydrol. Eng.*, *15*, 284–296.

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

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

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.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, 353 pp., Springer, New York.

Srikanthan, R., and T. A. McMahon (1982), Stochastic generation of monthly streamflows, *J. Hydraul. Div. Am. Soc. Civ. Eng.*, *108*, 419–441.

Stedinger, J. R., and R. M. Vogel (1984), Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, *20*, 47–56, doi:10.1029/WR020i001p00047.

Svanidze, G. G. (1980), *Mathematical Modeling of Hydrologic Systems*, Water Resour. Publ., Fort Collins, Colo.

Sveinsson, O. G. B., J. D. Salas, D. C. Boes, and R. A. Pielke (2003), Modeling the dynamics of long-term variability of hydroclimatic processes, *J. Hydrometeorol.*, *4*(3), 489–505.

Sveinsson, O. G. B., T. S. Lee, J. D. Salas, W. L. Lane, and D. K. Frevert (2009), Developments on Stochastic Analysis, Modeling, and Simulation (SAMS 2009), in *Great Rivers*, *Proc. World Environ. Water Resour. Congr.*, edited by S. Starrett, 1–10.

Tarboton, D. G., A. Sharma, and U. Lall (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*, 107–119, doi:10.1029/97WR02429.

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

Valencia, D., and J. C. Schaake (1973), Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, *9*, 580–585, doi:10.1029/WR009i003p00580.

Vogel, R. M., and A. L. Shallcross (1996), The moving blocks bootstrap versus parametric time series models, *Water Resour. Res.*, *32*, 1875–1882, doi:10.1029/96WR00928.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek (2003), A technique for generating regional climate scenarios using a nearest-neighbor algorithm, *Water Resour. Res.*, *39*(7), 1199, doi:10.1029/2002WR001769.

T. Lee, INRS-ETE, 490 de la Couronne, Quebec, QC G1K 9A9, Canada. (tae_sam.lee@ete.inrs.ca)

J. Prairie, Bureau of Reclamation, 421 UCB, University of Colorado, Boulder, CO 80309, USA.

J. D. Salas, Department of Civil and Environmental Engineering, B208 Engineering Bldg., Colorado State University, Fort Collins, CO 80523-1372, USA.