Modified K-NN Model for Stochastic Streamflow Simulation

James R. Prairie[1]; Balaji Rajagopalan[2]; Terry J. Fulp[3]; and Edith A. Zagona, M.ASCE[4]

Abstract: This paper presents a lag-1 modified K-nearest neighbor (K-NN) approach for stochastic streamflow simulation. The simulation at any time t given the value at the time $t-1$ involves two steps: (1) obtaining the conditional mean from a local polynomial fitted to the historical values of time t and $t-1$, and (2) then resampling (i.e., bootstrapping) a residual at one of the historical observations and adding it to the conditional mean. The residuals are resampled using a probability metric that gives more weight to the nearest neighbor and less to the farthest. The "residual resampling" step is the modification to the traditional K-NN time-series bootstrap approach, which enables the generation of values not seen in the historical record. This model is applied to monthly streamflow at the Lees Ferry stream gauge on the Colorado River and is compared to both a parametric periodic autoregressive and a nonparametric index sequential method for streamflow generation, each widely used in practice. The modified K-NN approach is found to exhibit better performance in terms of capturing the features present in the data.

DOI: 10.1061/(ASCE)1084-0699(2006)11:4(371)

CE Database subject headings: Colorado River; Nonlinear systems; Simulation; Streamflow.

Motivation

The Colorado River Simulation System (CRSS) (Bureau of Reclamation 1987) is widely used by the Bureau of Reclamation (BOR) to evaluate the effects of various policies (both water quantity and water quality related) that may be prescribed and implemented in the Colorado River basin. Capturing streamflow variability is key to such policy evaluations. To this end, one of the main components in the CRSS model is the stochastic streamflow simulation module to generate synthetic flow scenarios that reproduce the statistical properties of the observed data. The current technique for generating streamflows in the CRSS is the index sequential method (ISM). The ISM generates synthetic sequences by sequentially block resampling the historical time series. Consequently, it can only generate flow sequences observed in the historic record, which limits variability in the simulations. Policy analyses require a rich variety of statistically plausible sequences in order to better capture the variability in the flow. Clearly, ISM is limited in this requirement.

[1]Hydraulic Engineer, Bureau of Reclamation, Univ. of Colorado, UCB 421, Boulder, CO 80309-0421. E-mail: prairie@colorado.edu

[2]Assistant Professor, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado, UCB 426, Boulder, CO 80309-0426. E-mail: rajagopalan.balajir@colorado.edu

[3]Boulder Canyon Area Office Manager, Bureau of Reclamation, Lower Colorado Region, Boulder City, NV 89006. E-mail: tfulp@lc.usbr.gov

[4]Director, Center for Advanced Decision Support for Water and Environmental Systems (CADSWES), Univ. of Colorado, UCB 421, Boulder, CO 80309-0421. E-mail: zagona@cadswes.colorado.edu

Note. Discussion open until December 1, 2006. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on January 9, 2004; approved on September 14, 2005. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 11, No. 4, July 1, 2006. ©ASCE, ISSN 1084-0699/2006/4-371–378/$25.00.

The need to identify alternatives that will improve upon the ISM motivated the research presented in this paper. The ISM is a "nonparametric" method in that it makes no assumption of the functional form of the underlying model; instead, the method is data-driven. Keeping with the "nonparametric" spirit of ISM, we developed the proposed modified K-nearest neighbor (K-NN) method. The proposed approach retains all the aspects of the traditional K-NN time series bootstrap technique developed by Lall and Sharma (1996), but the "modification" enables simulating values not seen in the historical record. We evaluate the performance of our proposed approach by applying it to the monthly streamflow data from U.S. Geological Survey (USGS) stream gauge 09380000 located on the Colorado River at Lees Ferry, Arizona. We also compare the modified K-NN method with the ISM and a first-order periodic autoregressive model [PAR(1)], each widely used in practice.

The paper is organized as follows: a brief background on stochastic streamflow modeling including a description of the ISM and the PAR models is first presented, for the benefit of readers. Our proposed approach is then presented. A description of the results and summary conclude the presentation.

Background

Long-term operational and planning studies in a river basin require the ability to simulate realistic streamflow variability (McMahon et al. 1996). This ability involves developing a stochastic streamflow model to generate synthetic sequences of streamflow. These models work on the premise that the flow process is stationary. Hence the statistical characteristics of synthetic flows are similar to those observed in the historical record.

Streamflow simulation can be thought of as generating sequences from the conditional probability density function (PDF); e.g., for a first-order model, the conditional PDF is $f(y_t|y_{t-1})$, where y_t and y_{t-1}=streamflows at two consecutive time steps (e.g., seasons or months). Traditional stochastic streamflow models were developed within the linear autoregressive moving average (ARMA) and periodic autoregressive (PAR) framework and have extensive theoretical basis (see Yevjevich 1972; Stedinger and Taylor 1982; Bras and Iturbe 1985; Salas 1985).

One of the widely used parametric models for streamflow generation is the lag-1 PAR (Salas 1985), which linearly relates the streamflows in a season (or month) to the previous season, and has the form of

$$y_{\vartheta,\tau} = \mu_\tau + \Phi_{1,\tau}(y_{\vartheta,\tau-1} - \mu_{\tau-1}) + \varepsilon_{\vartheta,\tau} \qquad (1)$$

where ϑ=year, τ=season (or month); μ_τ=mean of the streamflow process in season τ, and $\Phi_{1,\tau}$=autoregressive parameter. The error $\varepsilon_{\vartheta,\tau}$ is assumed to be normally distributed with mean 0 and variance $\sigma^2(\varepsilon_\tau)$. The model parameters μ_τ, Φ_1, and $\sigma^2(\varepsilon)$ are estimated for each month from the data either by using method of moments or by approximating least-squares or Yule-Walker equations (Bras and Iturbe 1985; Salas 1985). A monthly PAR(1) model has 36 parameters. The model by construction preserves the mean, standard deviation, and lag(1) autocorrelation. By implication, y_t is also assumed to be normally distributed, and consequently, the joint $f(y_t, y_{t-1})$, and conditional $f(y_t|y_{t-1})$ probability density functions (PDF) are also normally distributed.

In practice, more often than not, streamflows are not normally distributed, thereby violating the assumptions of the above model. To address this, a log or power transformation is applied to the data to transform the data to a normal distribution, and the model is fit to the transformed data. The synthetic sequences generated from the model are then back-transformed into the original space. This process of fitting the model on the transformed data and then back-transforming it often does not guarantee the preservation of statistics in the original space (Benjamin and Cornell 1970; Bras and Iturbe 1985; Salas 1985; Sharma et al. 1997). Furthermore, transforming the data to a normal distribution is often nontrivial and involves subjectivity. Consequently, non-Gaussian features such as heavy skew or bimodality that may be present in the data will not be captured and reproduced effectively in the simulations. The parametric models can simulate values and sequences not seen in the historical record past. However, the main disadvantages are (1) the data must be transformed to a Gaussian distribution to satisfy the assumptions of the model; (2) ε is generated from a normal distribution, and hence, any values from $-\infty$ to $+\infty$ can be simulated, potentially resulting in unrealistic values; and (3) only linear dependence between values for the order (lag) of the model are captured. Fernandez and Salas (1986) provide a periodic gamma autoregressive model to address the first two disadvantages, but any such improvements will involve additional parameter estimation and still cannot address bimodality and nonlinear features.

Nonparametric models, on the other hand, do not make any assumptions about the underlying form of the dependence (i.e., linear) or the PDFs. The simplest nonparametric approach is the ISM. It involves sequential block resampling of historic data as a synthetic trace. For example, if there are 90 years of historic data, the method extracts an M-year (e.g., 25-year) block from the start of the historic record, then shifts 1 year forward and extracts the second M-year block and so on, repeating the process 90 times. When the end of the historic record is reached, the record is continued from the beginning of the time series. A schematic of this technique is shown in Fig. 1. Since the historic data are resampled, all the distributional (i.e., PDF) properties and statistics present in the data are reproduced faithfully. The main disadvantage is that only historically observed sequences can be generated resulting in simulations that have limited variety. Ouarda et al. (1997) and Kendall and Dracup (1991) compared

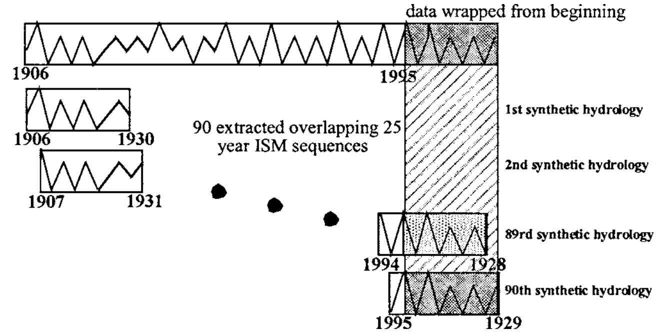

Fig. 1. Schematic of the ISM [adapted from Ouarda et al. (1997)]. The synthetic hydrologies, each 25 years in length, are shown below the original 90-year time series. The additional 24 years used for wraparound are shown in shading.

ISM and AR(1) models for projecting energy demand and reservoir storage capacities at Lake Mead and Powell, respectively. They found that the two methods were comparable in generating streamflow sequences that reproduced historic statistics. However, they found that AR(1) tended to underestimate dry and wet spells, while ISM, because of its constraints described earlier, could not reproduce the tails of the distributions (i.e., exceedance probabilities). The ISM is widely used by the BOR in the planning and management of the Colorado River basin.

Other nonparametric approaches that improve upon the ISM are (1) kernel methods, and (2) nearest-neighbor bootstrap methods. These are "local" estimators, in that, for a given value y_{t-1}, a small number of neighbors nearest to this value are obtained, and the conditional PDF $f(y_t|y_{t-1})$ is estimated based on these neighbors. If the neighbors include all the observed data points and if a normal distribution is fit, then this approach collapses to a linear parametric model. Kernel-based methods estimate the conditional PDF using a kernel function (or weight function) at each point of interest, y_{t-1}, which is then used in the simulation. A good general overview of the nonparametric techniques and their wide-ranging hydrologic applications can be found in Lall (1995). The kernel-based approach for streamflow generation was first developed by Sharma et al. (1997) and Tarboton et al. (1998). These approaches have also been successfully applied to rainfall modeling (Lall et al. 1996; Harrold et al. 2003b); flood frequency (Lall et al. l993; Moon and Lall 1994); groundwater applications (Adamoski and Feluch 1991); and streamflow forecasting (Smith 1991). The kernel-based methods can suffer from boundary problems (Lall 1995; Lall and Sharma 1996; Rajagopalan et al. 1997) that require appropriate kernel methods such as variable kernels (Sharma and O'Neil 2002). However, these methods are more difficult to use in higher dimensions.

Ameliorating the boundary problems of the kernel-based approach, Lall and Sharma (1996) developed a K-NN bootstrap method for time series resampling and applied it for streamflow simulation. In this method, first the K-NN of y_{t-1} from the historic data are found, and then the neighbors are resampled via a weight function that assigns large weight to the nearest neighbors and small to the farthest (Lall and Sharma 1996). This method is akin to estimating the conditional PDF, $f(y_t|y_{t-1})$, and simulating from it. Furthermore, while obtaining the K-NN relevant external information, such as large-scale climate features [e.g., El Nino Southern Oscillation (ENSO)], can be easily included. This approach has also been used in multivariate stochastic daily weather simulation (Rajagopalan and Lall 1999; Yates et al. 2003) and

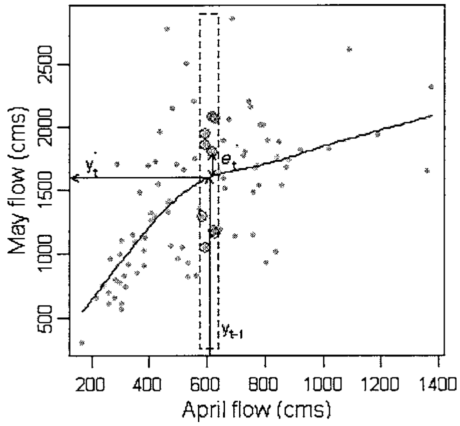

Fig. 2. Nonlinear local regression fit to May flows dependent on April flows are depicted by the solid line

placed in a Markovian framework to improve representation of longer-term variability in a daily model (Harrold et al. 2003a). The only slight drawback is that at any given time point values not seen in the historic record cannot be simulated; however, unlike ISM, a rich variety in the sequences can be obtained.

Hybrid models combining parametric and nonparametric methods have also been developed (Srinivas and Srinivasan 2001). In these models, the streamflow time series is partially prewhitened with a periodic autoregressive model (i.e., parametric model) to remove the dependence in the historic flow sequence; then, a moving block bootstrap (i.e., nonparametric model) is used to resample the residuals from the partially prewhitened streamflow. The hybrid models perform well but still involve extensive steps in fitting and generating the streamflows.

We propose a modification to the traditional K-NN time series bootstrap that addresses this drawback, as described in the following section.

Modified K-NN Method

As mentioned earlier, one of the drawbacks of the K-NN time series bootstrap technique is that values not seen in the historic record cannot be simulated. To address this, Lall and Sharma (1996) briefly mentioned a modification in their conclusion section. Here we formally develop and demonstrate the modification's utility.

The steps involved in the modification can be described using Fig. 2. This figure shows the scatter plot of April and May streamflows from the Lees Ferry stream gauge. The solid line shows a local (or nonparametric) fit through the scatter. The nonparametric fit is a locally weighted regression scheme (Loader 1999; Rajagopalan and Lall 1998) in that, at any point y^*_{t-1}, a local polynomial is fit to the K-NN. The size of the neighborhood (i.e., K) and the order of the polynomial (p) are obtained using an objective criteria called *generalized cross validation* [see the above references and Prairie (2002); Grantz et al. (2006); and Prairie et al. (2005)]. This estimation is repeated at all data points to obtain the solid line contained in Fig. 2. We used the package LOCFIT ⟨http://cm.belllabs.com/cm/ms/departments/sia/project/locfit/⟩ developed by Loader (1999) for fitting the local polynomials.

The modified K-NN algorithm proceeds as follows:

1. A local polynomial is fit for each month dependent on the previous month (as in Fig. 2):

$$y_t = g(y_{t-1}) + e_t \tag{2}$$

   where $g(y_{t-1})$=local polynomial fitted as described above.
2. The residuals (e_t) from the fit are saved.
3. Once we have the simulated value of the flow for the current month y^*_{t-1}, we estimate the mean flow of the next month $\hat{y}^*_t$ from Eq. (2) not including the residual.
4. K-NN of y^*_{t-1} (these are highlighted in Fig. 2) are obtained.
5. The neighbors are weighted using the weight function

$$W(i) = \frac{1/i}{(\sum_{i=1}^{K} 1/i)} \tag{3}$$

   This weight function gives more weight to the nearest neighbor and less weight to the farthest neighbor. The weights are normalized to create a probability mass function or "weight metric." Other weight functions with the same philosophy (i.e., more weights to nearest neighbors and lesser weights to farther neighbors) can be used as well. We found little or no sensitivity to the choice of the weight function.
6. One of the neighbors is resampled using the "weight metric" obtained from Eq. (3), above. Consequently, its residual (e^*_t) is added to the mean estimate $\hat{y}^*_t$. Thus the simulated value for the next time step becomes $y^*_t = \hat{y}^*_t + e^*_t$.
7. Steps 1–6 are repeated for other months to obtain an ensemble of simulations. The output from steps 1 and 2 can be saved for each month and used for successive years.

Lall and Sharma (1996) suggested both an objective criteria based on generalized cross validation and a heuristic scheme to select a K, the number of nearest neighbors. They stated that the heuristic scheme of ($K = \sqrt{N}$) does not appreciably change K, compared with the GCV. We adopted the same scheme here, where N=number of sample data points.

This approach has three clear features distinct from traditional K-NN:

1. Values not seen in the historic record can be simulated;
2. Residual resampling captures the local uncertainty more effectively; and
3. The local regression fit has the ability to capture any arbitrary (linear or nonlinear) relationship among the streamflows and extrapolate beyond the range of the observations.

Model Evaluation

We compared the three models [ISM, PAR(1), and modified K-NN] by applying them to the natural streamflows at USGS stream gauge 09380000 (Colorado River at Lees Ferry, Ariz.). The monthly natural streamflow data calculated by the BOR from historic gauge records by removing anthropogenic effects such as consumptive use, reservoir regulation, etc., were available for the 90-year period from 1906 to 1995 (James Prairie, personal data compilation, March 14, 2005).

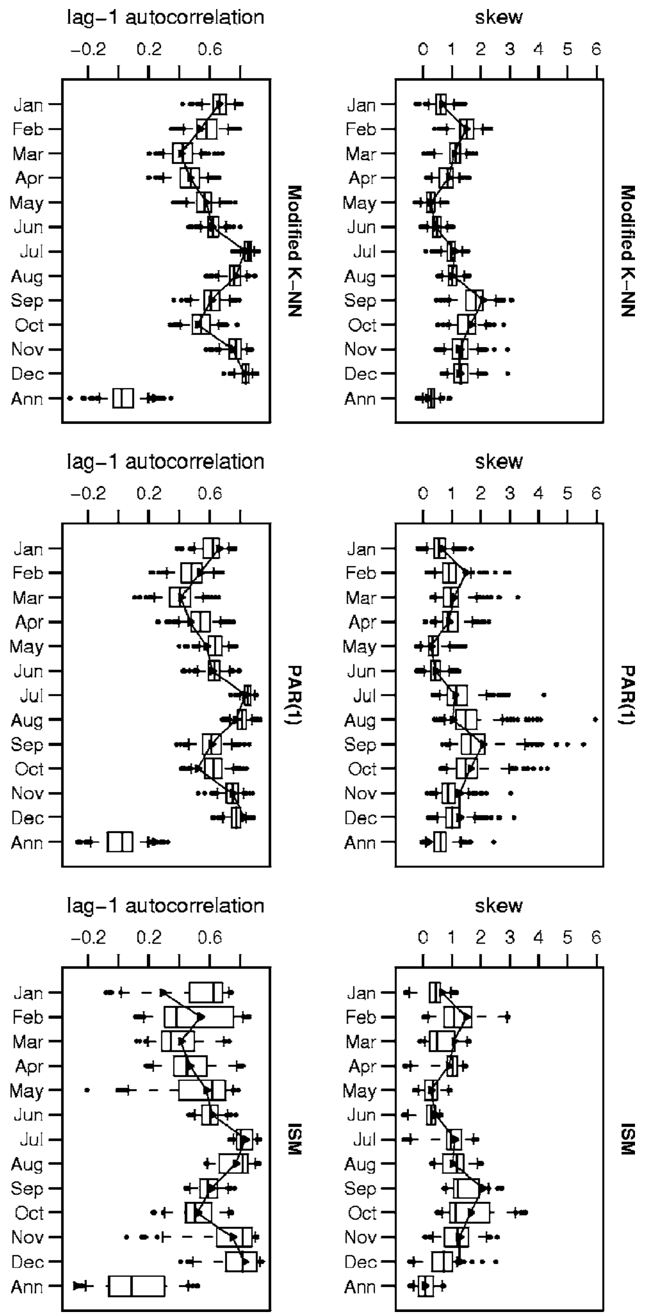

Fig. 3. Boxplots of coefficient of skew (upper plots) and lag-1 autocorrelation (lower plots) from simulations from the three models. The historic values are shown as solid circles and joined by the solid line.

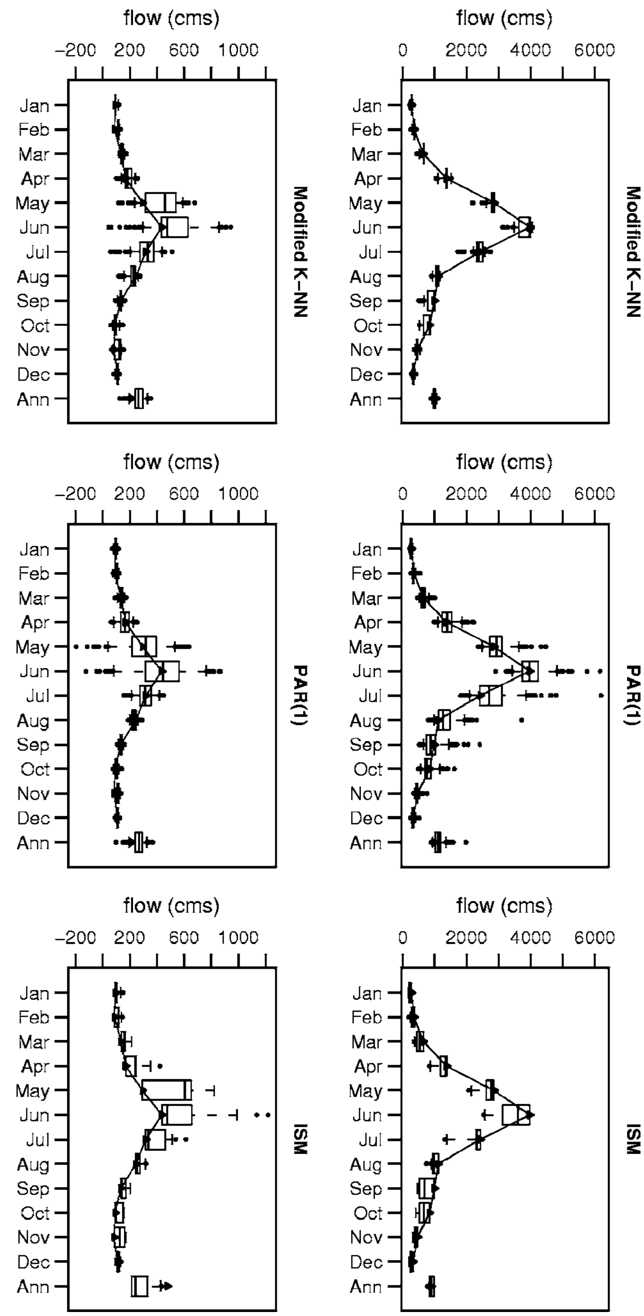

Fig. 4. Boxplots of maximum (upper plots) and minimum (lower plots) flow values

We generated 250 simulations from these models each of the same length as the historic data; for the ISM, we could generate only 90 simulations, each of 25 years in length. For the PAR(1) model we used the semiautomated software package SAMS (stochastic analysis, modeling, and simulation) developed by Salas et al. (2000) which performs transformations, model fitting, and simulation. This package allows the user to perform transformation to normal distribution with the assistance of log probability plots and a skewness test.

A suite of basic and higher-order statistics are computed from the simulations and compared with those of the historic data. These are

1. Monthly mean flows, standard deviation, lag-1 correlation (i.e., month-to-month correlation), coefficient of skewness, and maximum and minimum flows;
2. Marginal, bivariate, and conditional PDFs. The probability density functions are estimated using the nonparametric kernel density estimators (Bowman and Azzalini 1997); and
3. Drought (longest and maximum drought) and surplus (longest and maximum surplus) statistics. The longest drought statistic is the maximum number of consecutive years of flows below a threshold value; the maximum drought statistic is the maximum volume of water during a drought. The computation is vice versa for surplus statistics.

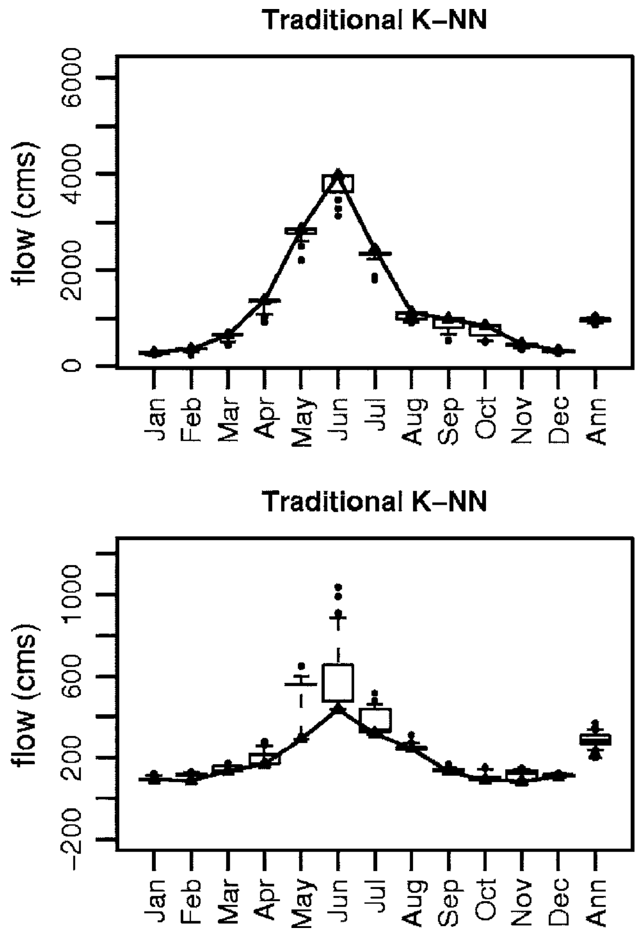

Fig. 5. Boxplots of maximum (upper plot) and minimum (lower plot) flow values from traditional K-NN

A sampling of these statistics is presented and described in the following section.

Results

The results are presented as boxplots, with the box representing the interquartile range and the whiskers extending to the 5th and 95th percentile of the simulations; the statistics from the historic record are linked by a solid line. Wider boxes are indicative of increased variability; the historic values falling within the box suggest a good reproduction of that statistic.

As expected, all the models preserved the mean and standard deviation (figures not shown). Boxplots of skew coefficient (Fig. 3, right side) and lag-1 autocorrelation (Fig. 3, left side) indicate that the modified K-NN captures these two statistics very well. The PAR(1) model captures the skew in all months except in February, August, and November. This suggests that the transformations in these months were not fully effective. The lag-1 autocorrelations are very well captured, as to be expected given that it is a lag-1 model. The ISM simulations have a larger boxplot, indicative of increased variability. This is due to the fact that each simulation's length is 25 years, while the simulations from modified K-NN and PAR(1) are of the same length as the historic record (i.e., 90 years) and consequently, tighter boxes. If the ISM simulations are also made to be 90 years in length, then each simulation's statistics will be exactly the same.

Boxplots of maximum (Fig. 4, right side) and minimum (Fig. 4, left side) values show that the ISM model does not simulate values not seen in historic record, so the observed values and the top of the boxplot are the same. The PAR(1) model captures these two statistics well, but notice the negative values that are generated in the minimum values, especially in the high-flow months of May and June. These values result because the error term in Eq. (1) is simulated from a normal distribution.

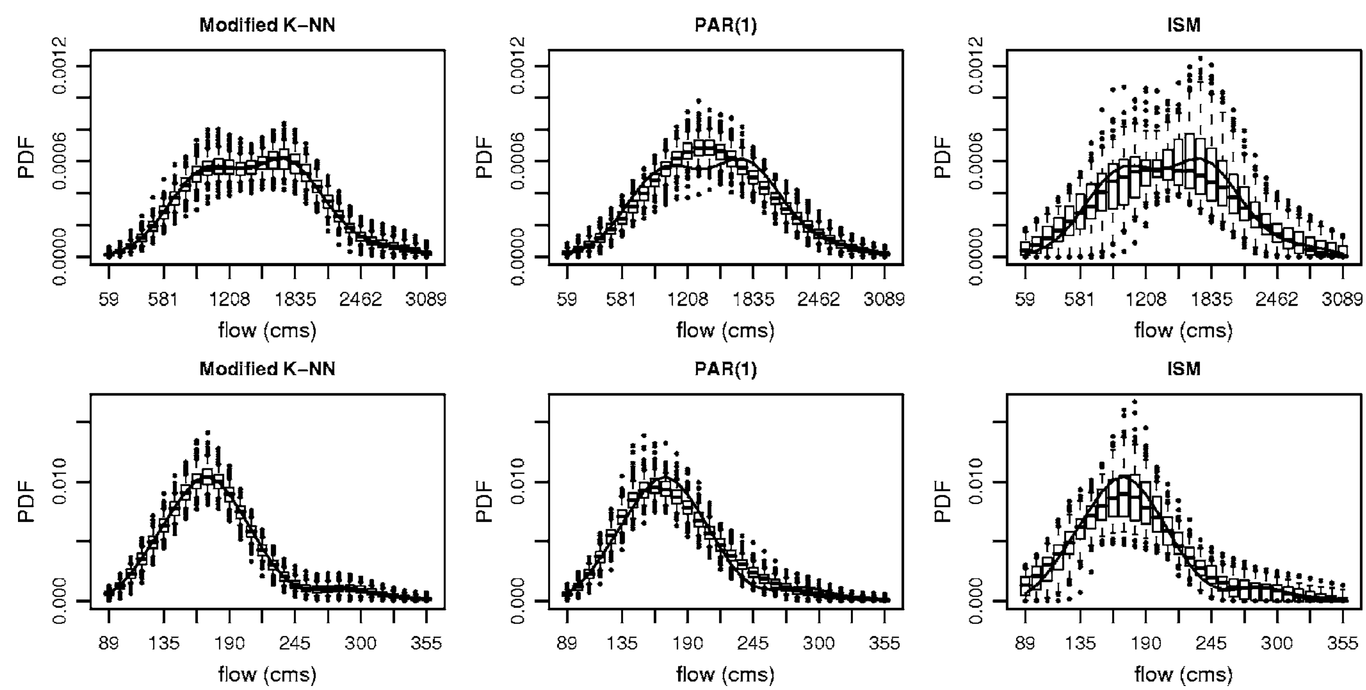

Fig. 6. Boxplots of PDFs of May (upper plots) and December (lower plots) flows

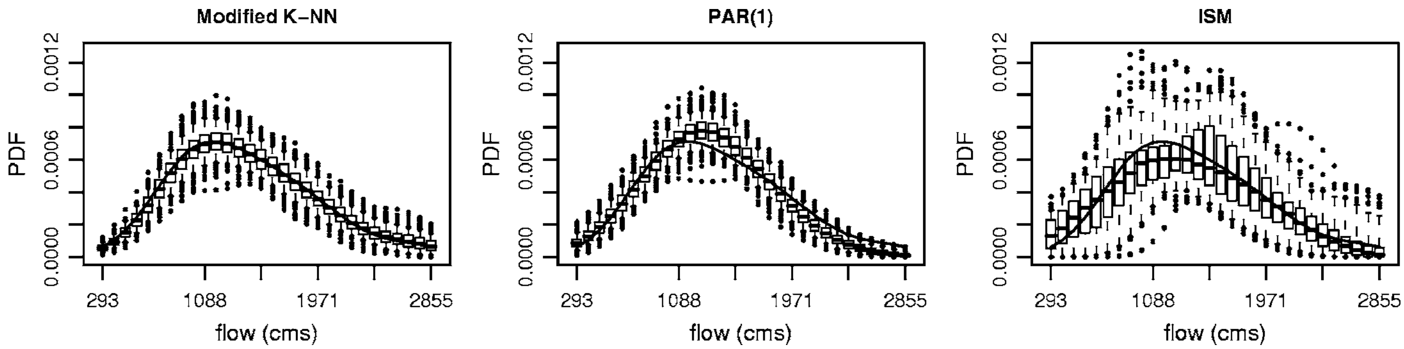

Fig. 7. Boxplots of conditional PDF of June flows, conditioned on May flow of 467 cm

The modified K-NN preserves both the maximum and minimum values, and it generates values not seen in the historic record, which is a result of the proposed modification (versus the traditional K-NN time series bootstrap). The range of extrapolation depends on the nature of the scatterplot (e.g., Fig. 2) and the magnitude of the residuals in the neighborhood. For example, simulation of a value in May outside the range of observation will depend on the simulated value in April.

As stated in the motivation section, the modified K-NN approach retains all aspects of the traditional K-NN approach but provides the capability to generate values not seen in the historical record and also limited extrapolation to generate extreme values. The traditional K-NN is constrained to the range of the observed streamflows in simulating the extremes, as can be seen in Fig. 5. Except for the maximum and minimum statistic, the traditional K-NN and modified K-NN reproduce all the statistics in a similar manner (figures not shown).

One of the key statistics that indicates the performance of these models is the PDF. Boxplots of PDFs for May and December streamflows are shown in Fig. 6, top and bottom rows, respectively. A clear bimodality can be seen in the PDF of the May streamflows that is well captured by the Modified K-NN and ISM but not by PAR(1); in fact, the PAR(1) simulations resemble a Gaussian distribution. The boxes are wider for ISM simulations, as expected and described earlier. The modified K-NN reproduces the true shape of the historical PDF very well in both months. Reproduction of the PDF has significant implications in terms of the threshold exceedance probabilities and, consequently, in policy analyses.

We computed the conditional PDF of June flows given a May flow of 467 cubic meters per second (cms), and the boxplots of the same are shown in Fig. 7. Here too, the modified K-NN method does well in capturing the true form of the underlying PDF, which is slightly non-Gaussian. The PAR(1) tends to simulate a normal-looking PDF since it assumes normal distributions for the joint and conditional probabilities. The ISM captures the feature but with large variability.

Last, we compared these models on the ability to reproduce the drought and surplus statistics, as shown in Fig. 8. The boxplots are shown as the ratio of the values from the simulations to the historical values. The modified K-NN preserves the surplus statistics but the PAR(1) tends to overestimate the maximum surplus. Both PAR(1) and modified K-NN tend to underestimate the drought statistics. The drought statistics are long memory variables that are not guaranteed to be captured with a lag(1) model.

Summary and Discussion

This work was motivated by the need to develop a robust alternative to ISM in the CRSS model. To this end, we developed a modified K-NN method of lag-1 for stochastic streamflow simulation. Our proposed approach retains all the features (i.e., ability to capture any arbitrary PDF and dependence structure present in the data) of the traditional K-NN time series bootstrap technique developed by Lall and Sharma (1996), but the "modification" enables simulating values not seen in the historic record. The modification was discussed briefly in Lall and Sharma (1996) in their summary section, which we developed and implemented in this paper. In this model, first, a local polynomial (a nonparametric function) is fitted to estimate the mean of the conditional probability density function. The simulation at any time t given the value at the time $t-1$, then, involves two steps: (1) obtaining the conditional mean from the local polynomial fit, and (2) then resampling (i.e., bootstrapping) a residual at one of the historic observations and adding it to the conditional mean. The residuals are resampled using a probability metric that gives more weight to the nearest neighbor, less to the farthest. This model was applied to monthly natural streamflows at the USGS Lees Ferry stream gauge on the Colorado River and was compared to PAR(1) and ISM models, each widely used in practice. We found the modified K-NN approach to exhibit better performance in terms of faithfully capturing all the features present in the data, especially the non-Gaussian PDFs.

One of the drawbacks of this model is that it does not capture the interannual variability very well (as seen from the simulation of drought statistics). This requires incorporation of long-term dependence structure. Sharma and O'Neil (2002) included the sum of the previous 12 monthly flows as a conditioning variable in addition to the previous month's flows in a kernel-based nonparametric streamflow generation model and showed that such an inclusion better captures the interannual variability. We plan to adopt a similar strategy in the modified K-NN model, in future research.

Another disadvantage of the modified K-NN approach is when the sample size is small. In such situations, the number of neighbors to bootstrap the residuals will be small and consequently will limit variety in the ensembles. To address this, there are two possible approaches: (1) subjectively change (increase) the number of neighbors to obtain a good variety in the ensembles, or (2) use the local standard error from the local polynomial (Loader 1999) to generate random normal deviates and then add these to the mean estimates to generate the ensembles. In an application for

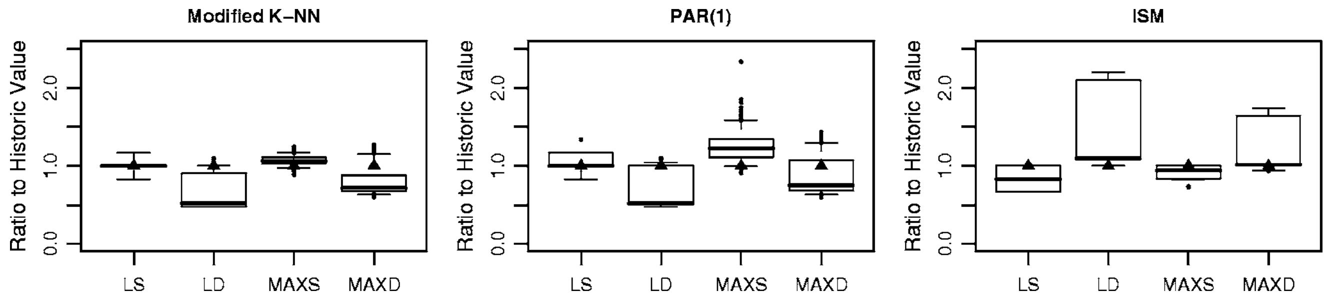

Fig. 8. Boxplots of drought and surplus statistics for simulations from the three models. The statistics longest drought (LD), maximum drought (MD), longest surplus (LS), and maximum surplus (MS) are described in the model evaluation section.

forecasting Thailand summer rainfall (Singhrattna et al. 2005) with a small sample size ($N=25$), the latter approach was implemented with good results.

One of the significant advantages of the K-NN (or modified K-NN) framework is that variables can be easily added to the conditioning vector. For instance, if large-scale climate features such as ENSO are known to modulate the variability of streamflows at a location, an index of ENSO can be included into the conditioning vector, thus incorporating relevant climate information in the simulation. This would be computationally intensive and unwieldy in a parametric approach such as PAR or similar frameworks. We applied this modified K-NN approach with good success to Truckee-Carson River basin streamflow forecasting (Grantz et al. 2006), Thailand summer rainfall forecasting (Singhrattna et al. 2005), and salinity modeling on the Colorado River basin (Prairie et al. 2005). A variant of this approach has been applied for improved streamflow forecasts in north-eastern Brazil (DeSouza and Lall 2003).

The K-NN framework with the proposed modification provides an attractive and robust alternative to the parametric approaches. Extensions to include large-scale climate information or space-time disaggregation of streamflows are relatively easy and parsimonious.

Acknowledgments

Funding for this work by the Bureau of Reclamation is gratefully acknowledged. The writers thank Dave Trueman for his strong support and encouragement. The writers would also like to thank three anonymous reviewers whose comments significantly improved the paper.

Notation

The following symbols are used in this paper:

e_t = error term at time t;
i = index term;
K = number of neighbors;
N = sample size;
p = order of the polynomial;
t = time index;
W = weight function;
y = dependent variable;
y^* = dependent variable plus an error term;
ε = error term;
ϑ = year index;
μ = estimate of the mean;
τ = season index; and
Φ = autoregressive parameter.

References

Adamoski, K., and Feluch, W. (1991). "Application of nonparametric regression to groundwater level prediction." *Can. J. Civ. Eng.*, 18, 600–606.

Benjamin, J. R., and Cornell, C. A. (1970). *Probability, statistics, and decision for civil engineers*, McGraw-Hill, New York.

Bowman, A. W., and Azzalini, A. (1997). *Applied smoothing techniques for data analysis*, Clarendon, Oxford, U.K.

Bras, R. L., and Iturbe, I. (1985). *Random functions and hydrology*, Addison-Wesley, Reading, Mass.

Bureau of Reclamation. (1987). *Colorado River simulation system, system overview*, U.S. Dept. of the Interior, Denver.

DeSouza, F. A., and Lall, U. (2003). "Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a mutlivariate, semi-parametric algorithm." *Water Resour. Res.*, 39, 1307–1320.

Fernandez, B., and Salas, J. D. (1986). "Periodic gamma-autoregressive process for operational hydrology." *Water Resour. Res.*, 22(10), 1385–1396.

Grantz, K., Rajagopalan, B., Clark, M., and Zagona, E. (2006). "A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts." *Water Resour. Res.*, 41, W10410.

Harrold, T. I., Sharma, A., and Sheather, S. J. (2003a). "A nonparametric model for stochastic generation of daily rainfall occurence." *Water Resour. Res.*, 39(10), 1300.

Harrold, T. I., Sharma, A., and Sheather, S. J. (2003b). "A nonparametric model for stochastic generation of daily rainfall amounts." *Water Resour. Res.*, 39(12), 1343.

Kendall, D. B., and Dracup, J. A. (1991). "A comparison of index-sequential and AR(1) generated hydrologic sequences." *J. Hydrol.*, 122, 335–352.

Lall, U. (1995). "Recent advances in nonparametric function estimation: Hydraulic applications." *Rev. Geophys.*, 33, 1093–1102.

Lall, U., Moon, Y.-I., and Bosworth, K. (1993). "Kernel flood frequency estimators: Bandwidth selection and kernel choice." *Water Resour. Res.*, 29(4), 1003–1015.

Lall, U., Rajagopalan, B., and Tarboton, D. G. (1996). "A nonparametric wet/dry spell model for resampling daily precipitation." *Water Resour. Res.*, 32(9), 2803–2823.

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

Loader, C. (1999). *Local regression and likelihood*, Springer, New York.

McMahon, T. A., Pretto, P. B., and Chew, F. H. S. (1996). "Synthetic hydrology—Where's the hydrology." *Stochastic Hydraulics '96*, K. S. Tickle, I. C. Goulter, C. Xii, S. A. Wasimi, and F. Douchart, eds., Balkema, Rotterdam, The Netherlands, 3–14.

Moon, Y.-I., and Lall, U. (1994). "Kernel function estimator for flood frequency analysis." *Water Resour. Res.*, 30(11), 3095–3103.

Ouarda, T., Labadie, J. W., and Fontane, D. G. (1997). "Index sequential hydrologic modeling for hydropower capacity estimation." *J. Am. Water Resour. Assoc.*, 33(6), 1337–1349.

Prairie, J. R. (2002). "Long-term salinity prediction with uncertainty analysis: Application for Colorado River above Glenwood Springs, CO." MS thesis, Univ. of Colorado, Boulder, Colo.

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

Rajagopalan, B., and Lall, U. (1998). "Locally weighted polynomial estimation of spatial precipitation." *J. Geogr. Inf. Decis. Anal.*, 2(3), 48–57.

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

Rajagopalan, B., Lall, U., Tarboton, D. G., and Bowles, D. S. (1997). "Multivariate nonparametric resampling scheme for simulation of daily weather variables." *J. Stochastic Hydrol. Hydrauli.*, 11(1), 65–93.

Salas, J. D. (1985). "Analysis and modeling of hydrologic time series." *Handbook of hydrology*, D. R. Maidment, ed., McGraw-Hill, New York, 19.1–19.72.

Salas, J. D., Chung, C., Lane, W. L., and Frevert, D. K. (2000). *Stochastic analysis, modeling, and simulation (SAMS), Version 2000—User's manual*, Department of Civil Engineering, Hydrologic Science and Engineering Program, Colorado State Univ., Fort Collins, Colo.

Sharma, A., and O'Neill, R. (2002). "A nonparametric approach for representing interannual dependence in monthly streamflow sequences." *Water Resour. Res.*, 38(7), 5.1–5.10.

Sharma, A., Tarboton, D. G., and Lall, U. (1997). "Streamflow simulation: A nonparametric approach." *Water Resour. Res.*, 33(2), 291–308.

Singhrattna, N., Rajagopalan, B., Clark, M., and Krishna Kumar, K. (2005). "Forecasting Thailand summer monsoon rainfall." *Int. J. Climatol.*, 25(5), 649–664.

Smith, J. A. (1991). "Long-range streamflow forecasting using nonparametric regression." *Water Resour. Bull.*, 27(1), 39–46.

Srinivas, V. V., and Srinivasan K. (2001). "A hybrid stochastic model for multiseason streamflow simulation." *Water Resour. Res.*, 37(10), 2537–2549.

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

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

Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. (2003). "A technique for generating regional climate scenarios using a nearest neighbor bootstrap." *Water Resour. Res.*, 39(7), 1199.

Yevjevich, V. M. (1972). *Stochastic processes in hydrology*, Water Resources Publication, Colorado State Univ., Fort Collins, Colo.