Modified K-NN Model for Stochastic Streamflow Simulation

James Prairie¹, Balaji Rajagopalan², Terry Fulp³ and Edith Zagona⁴

Abstract

This paper presents a modified k-nearest neighbor approach for streamflow generation. 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 point ‘t+1’ given the value at the current time ‘t’ then involves two steps (i) obtaining the conditional mean from the local polynomial fit (ii) then resampling 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, less to the farthest. The local polynomial (an assumption free nonparametric approach) fit for the conditional mean, has the ability to capture any arbitrary (linear or nonlinear) dependence structure. The coupling of this with the residual resampling provides the capability to capture any arbitrary dependence and probability density functions exhibited by the data, unlike conventional methods that can capture only linear dependence and Gaussian probability density functions. This model is compared to both a parametric periodic auto regressive and a nonparametric index sequential method streamflow generator, each widely used in practice.

¹ U.S. Bureau of Reclamation, Univ. of Colorado, Boulder, Colorado
² Dept. Of Civil, Env., and Arch. Engg. Dept., Univ. of Colorado, Boulder, Colorado
³ U.S. Bureau of Reclamation, Lower Colorado Region, Boulder City, Nevada
⁴ Center for Advanced Decision Support for Water and Env. Systems, Univ. of Colorado, Boulder, Colorado
**Introduction**

The Colorado River Simulation System (CRSS) is widely used by the Bureau of Reclamation (USBR) 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. Streamflow variability is one of the key variables in policy analyses. One of the main components in the CRSS model is the stochastic streamflow generation module to generate synthetic flow scenarios that realistically reproduce the observed statistics and uncertainty [Prairie, 2002]. The current technique for generating streamflows used by the CRSS model is the Index Sequential Method (ISM). The ISM generates synthetic sequences by sequentially block resampling the historical time series. Consequently, there is no variety in the generated flow sequences. While the statistics of the observed data are reproduced in the simulations, lack of variety produces little uncertainty, which is important for policy analysis. The description of this method and these aspects will be discussed in the following sections and in the results section.

Thus, there is a need to identify alternatives that will improve upon the ISM. This motivated the development of the proposed modified k-nearest neighbor (K-NN) method for streamflow simulation. The paper is organized as follows: A brief background on stochastic streamflow modeling is first presented, followed by descriptions of ISM, Periodic Auto Regressive Model (PAR) and a nonparametric alternative. Our proposed model is next presented. We compare the models by applying them to monthly streamflow data from USGS stream gauge 09072500 (Colorado River near Glenwood Springs, CO). 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, 1996]. Typically, this ability involves developing a streamflow model (typically, stochastic) to generate synthetic sequences of streamflow. The generated sequences preserve some basic historic statistics (mean, standard deviation, lag(1) correlation for a first order model) and higher order statistics depending upon the model. These models work on the premise that the statistics of the historical flows are likely to occur in the future, i.e., the stationary assumption.

Stochastic streamflow models were traditionally developed in both Auto Regressive Moving Average (ARMA) and PAR models [see Yevjevich, 1972; Bras and Rodriguez-Iturbe, 1985; Salas, 1985; Stedinger and Taylor, 1982]. These are also referred to as parametric models because they involve selecting an appropriate model and fitting parameters to it. Parametric models, as they are based on regression (i.e., regression of the time series on its past values) framework, have all the assumptions of linear regression – most important of them being that the time series is normally (Gaussian) distributed [Salas, 1985]. More often than not, streamflows are not distributed normally, thereby violating this assumption. To address this, the data is transformed to a Gaussian distribution using a log or power transformation before fitting the model to the transformed data. The synthetic sequences generated from the model are 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 [Sharma et al., 1997; Salas, 1985; Bras and Rodriguez-Iturbe, 1985; Benjamin and Cornell, 1970]. Furthermore, transforming the data to a Gaussian distribution is often non-trivial and even involves subjectivity. Consequently, non-Gaussian features (heavy skew or bimodality) that may be present in the data will not be captured and reproduced effectively. Streamflows exhibit a variety of distributions including bimodality and strong skew, as can be seen in the flows from Colorado river basin [Sharma et al., 1997; Lall and
and as we show later in this paper when we compare the methods on the Glenwood Springs streamflow gauge.

The parametric models generally preserve the mean, variance, and auto correlations (depending on the order of the model). As the time series is transformed to a Gaussian distribution (or near Gaussian) by appropriate transformation, the skewness is approximated. Last but not the least, these models require estimating several parameters, depending on the type of model, thereby, adding considerable uncertainty which can impact the simulations.

Nonparametric models unlike parametric, are assumption free and data driven. They strive to address the drawbacks of the parametric framework described above. The simplest nonparametric model is the ISM, which involves selecting blocks of historic data. For example, if we have 100 years of historical data, without wraparound we can generate 80 sequences of 20-year lengths, 70 sequences of 30-year lengths, etc. The advantages are that it is simple and easy to implement, assumption free, and can reproduce the entire distributional properties of the historic data: the mean, variance, auto-correlation, etc. The main disadvantage is that only historically observed sequences can be generated resulting in simulations with limited variety.

Kernel-based nonparametric models developed recently for streamflow simulation [Sharma et al., 1997], streamflow disaggregation [Tarboton et al., 1998] and for multivariate weather generation [Rajagopalan et al., 1997], have been shown to alleviate the drawbacks of parametric models and also ISM. In this approach, a kernel function (i.e., a weight function) is chosen with a limited extent (also known as bandwidth). At any point of estimate, all the historical observations are captured within the kernel function centered around the point of estimate. The captured points are weighted according to their distance to the point of estimate (i.e., more weights to nearest points and lesser to the farthest), a weighted sum is computed to estimate the
density function. A similar approach is used for regression estimation. Kernel methods have been successfully applied to a variety of problems – rainfall modeling [Lall et al., 1996]; flood frequency [Lall et al., 1993, Moon and Lall, 1994]; groundwater applications [Adamoski and Feluch, 1991], and streamflow forecasting [Smith, 1991]. Refer to Lall [1995] for an overview of the nonparametric techniques, in particular kernel-based methods and their hydrologic applications. However, the kernel methods suffer from severe boundary problems, more so in higher dimensions, that can bias the simulations. Hybrid methods using parametric and nonparametric methods have also been tried by Srinvas and Srinivasan, [2001]. In this, they fit the time series with a periodic autoregressive model that captures the dependence in the historic flow sequence. They then, do a moving block bootstrap (a nonparametric method) to resample the residuals. This hybrid improves upon the traditional parametric methods in preserving non-Gaussian features. The main drawback of this approach is that by fitting a periodic autoregressive model to the data first, nonlinear relationships cannot be captured. Furthermore, the hybrid approach requires several steps and parameters to be estimated.

Lall and Sharma [1996] developed a K-NN bootstrap method for time series resampling and applied it for streamflow simulation. This improves significantly over the kernel methods, and is easy to implement. In this method, K-NN from the historical data are found to the current feature vector. Once the neighbors are identified then there are several options (i) compute a weighted average to get a mean forecast [Yakowitz, 1985] or (ii) resample the neighbors with a probability metric that gives large weight to the nearest neighbors and less weight to the farthest, thereby generating ensembles [Rajagopalan and Lall, 1999; Yates et al., 2003; Lall and Sharma, 1996] or (iii) fit a polynomial to the k neighbors and use it to estimate the mean forecast and also resample the residuals to generate ensembles. As can be seen the approach provides a flexible framework and is easy to implement in higher
dimensions. Furthermore, the “feature vector” can be designed to include climate variables in addition to past values of the time series. In this paper we develop (iii) above which improves upon the traditional K-NN bootstrap developed by \[Lall and Sharma, 1996\] described in the following section.

In summary, the nonparametric models estimate the marginal and conditional probability density functions “locally” (e.g., the K-NN or the points within the kernel function) and simulate sequences from them. They are assumption free, and being “local” have the ability to capture any arbitrary probability density function or relationship, as can be seen from the application in this paper and in the above references.

**Index Sequential Method**

The ISM is a nonparametric stochastic technique that generates synthetic streamflows by sequentially block bootstrapping the historic time series. Block bootstrapping is a technique in which a block of the historic time series is resampled as a synthetic trace. For example, our historic time series for stream gauge 09072500 is 90 years in length, from water years 1906 to 1995. To model 25 years into the future, the method extracts a 25-year block from the historic streamflow record then shifts one year forward and extracts 25 years again, 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 Figure 1. The intent is that every year to be simulated sees all the streamflows of the historic record. Consequently, the simulated sequences have the same distributional properties as the historic data.

*Ouarda et al. [1997]* compared the nonparametric ISM to a traditional parametric autoregressive method to determine how well the two modeling techniques allowed project dependable capacity to be estimated. Their study found
using ISM, synthetically generated sequences statistically corresponded acceptably with sequences generated from an AR(1) model. Because power generation was of interest, the study looked at the cumulative density function for total energy out and found that the two techniques compared well at the lower and middle sections, which are critical to energy production modeling.

*Kendall and Dracup* [1991] also compared ISM and an AR(1) model. Again, an annual model was developed to compare the two techniques. Their study compared reservoir storage capacities at Lake Mead and Lake Powell. They also found that the two techniques generated sequences that are not significantly different. Generally, they found that flows generated by the AR(1) model result in slightly higher storage levels than the ISM. They also stated that the AR(1) model has a tendency to underestimate the occurrence of severe droughts. Further, the ISM did not perform well when the tails of the distribution were of interest. They found at 90 percent exceedance and greater, the AR(1) model produced lower flows. They recommended considering AR(1) models if tails of the distribution are important.

Neither of these studies explored higher order statistics, such as the PDF of the generated synthetic flows or the bivariate probability of the current month’s flow dependent on the previous month’s flow and also extreme statistics. For application in the CRSS model, which is the motivation of this study, reproducing the extreme events (wet and dry periods) are crucial in effective policy decision-making. What we find is that the ISM is limited in its ability to reproduce a variety of wet and dry sequences since it resamples chunks of historical record.

Advantages of this technique, which have led to its widespread use, are that it is simple to implement and that it does not require the modeler to estimate parameters, and it reproduces all of the historical statistics.
Periodic Auto Regressive Method (PAR)

The periodic auto regressive (PAR) model is one of the traditional parametric methods to model and simulate synthetic streamflow. The PAR model is also termed a seasonal auto regressive model and is distinguished from other auto regressive methods in that it explicitly models seasonality, such as the case with streamflow.

The general equation for a PAR model of order $\rho$ is given by

$$y_{\vartheta, \tau} = \mu_{\tau} + \sum_{j} \Phi_{j, \tau} (y_{\vartheta, \tau-j} - \mu_{\tau-j}) + \epsilon_{\vartheta, \tau}$$  \hspace{1cm} (1)

where: $y$ is the streamflow process,

$\vartheta$ is the year, $\tau$ is the season,

$\mu_{\tau}$ is the mean of the process in season $\tau$ streamflow,

$\Phi$ is the auto regressive parameter,

$\epsilon_{\vartheta, \tau}$ is the uncorrelated normally distributed noise term with mean 0 and variance $\sigma^{2}(\epsilon)$.

The season could represent months or another subset of a year. In our investigation, we developed a lag(1) monthly model or a model of order $\rho = 1$ with $\tau = 12$. This model can be written as

$$y_{\vartheta, \tau} = \mu_{\tau} + \Phi_{1, \tau} (y_{\vartheta, \tau-1} - \mu_{\tau-1}) + \epsilon_{\vartheta, \tau}$$  \hspace{1cm} (2)

which means each season represents a month, and the current month flow is dependent linearly on the previous month’s flow. The model parameters $\Phi$ and $\sigma^{2}(\epsilon)$ are estimated for each month from the data. With $\rho = 1$ and $\tau = 12$ (representing 12 seasons), 12 estimates of $\Phi_{\rho, \tau}, \mu$, and $\sigma^{2}(\epsilon)$ have to be computed, requiring estimation of 36 parameters. Method of Moments, approximating Least
Squares, or Yule-Walker equations are used to estimate the auto regressive parameters [Salas, 1985; Bras and Rodriguez-Iturbe, 1985].

The PAR(1) model by its construction, preserves the mean, standard deviation, and lag(1) correlation. Furthermore, being in the realm of linear regression, the data is assumed to be normally distributed and as such, the simulations generally confirm to a Gaussian probability function. If the data is non-Gaussian, it has to be first transformed to a Gaussian shape via box-cox transformations before the model is fit. We transformed the Colorado River data with a log transformation using the equation $y = \ln(x + a)$ where: $y$ is the transformed process, $x$ is the original process, and $a$ is estimated to bring the skewness coefficient towards zero. This adds additional steps in the modeling process. Furthermore, many times it is hard to obtain a “best” transformation (e.g., if the data has a bimodal distribution). This imposes a serious limitation on the model.

Thus the model is fitted in the transformed space and the simulations are “back-transformed” to the original space. The back-transformation is not guaranteed to preserve the basic and higher order statistics of the transformed data [Sharma et al., 1997; Salas, 1985; Bras and Rodriguez-Iturbe, 1985].

In our application here, we used the software package developed at Colorado State University, “Stochastic Analysis, Modeling, and Simulation” (SAMS) [Salas et al., 2000]. SAMS transforms the flow data to approximate a Gaussian distribution by the transformation process described earlier.

The advantages to using a parametric model like PAR are that (1) it can simulate values and sequences not seen in the past and (2) a fully developed theoretical background exists. The main disadvantages are that the data must be transformed to a Gaussian distribution to satisfy the assumption of the model and that $\varepsilon$ is generated from a normal distribution; hence, any values from $-\infty$ to $+\infty$ can be simulated, resulting in unrealistic values.
Traditional K-NN Method

Recently developed nonparametric techniques try to address the drawbacks of ISM and parametric models. These include, kernel based [Sharma et al., 1997; Tarboton et al., 1998], and K-NN based time series resampling [Lall and Sharma, 1996]. As mentioned earlier, the nonparametric techniques approximate the conditional and marginal densities of a time series and simulate from these. For example, a lag(1) model can be expressed as a simulation from the conditional PDF

\[ f(y_t|y_{t-1}) \]  

(3)

where \( y \) is the time series of flows.

The parametric models, such as PAR, essentially simulate from a Gaussian conditional distribution because they assume the data is normally distributed, as described in the previous section. The nonparametric models approximate the conditional PDF shown in (3), from the data, either through kernel estimators or by K-NN time series resampling – this forms the crux of the nonparametric approach. Thereby, they have the ability to reproduce any arbitrary PDF structure that might be present in the data, without any prior assumptions, unlike the parametric counterpart.

Lall and Sharma [1996] introduced the K-NN time series resampling model, distinguishing it from the ISM. Unlike ISM, the K-NN model does not resample a block of the historic time series for each generated time series. Instead, the conditional PDF is approximated using K-NN of the current value and one of the neighbors is selected as the value for the next time step. The method is briefly described below (for additional details see Lall and Sharma [1996] or Rajagopalan and Lall [1999]):

1. Find the k-nearest neighbors to the current month’s flow. (Basically, compute
the distance from the current month’s flows to all the flows of the current month in the historical record. The distances are sorted and the nearest k neighbors are found)

2. The neighbors are weighted based on their distance to the current month’s flow such that the nearest neighbor gets the largest weight and the farthest neighbor the least weight.

3. The weights are normalized to create a probability mass function, also known as “weight metric”.

4. Using this “weight metric” one of the neighbors is resampled. The successor of the resampled neighbor becomes the simulated value for the next month.

The steps are repeated to generate several simulations.

The number of neighbors k can be thought of as a smoothing parameter – in that, for smaller k the PDF approximation is based on a few points and therefore has the ability to capture local features, while a larger k can smooth out local features. Of course, getting an appropriate k is the key. There are several methods for obtaining k, heuristic and objective methods such as Generalized Cross Validation (GCV).

This approach has been extended for daily weather generation by Rajagopalan and Lall [1999] and for spatial weather generation by Yates et al. [2003].

One of the drawbacks of the K-NN technique is that the values not seen in the historical record cannot be simulated; therefore, there is reduced variety in the simulations. Nonparametric models that use kernel density estimators alleviate this problem [Sharma et al., 1997]. However, as mentioned earlier, the kernel based methods have severe problems at the boundaries of the variables (i.e., 0, for streamflows) resulting in bias [Lall and Sharma, 1996]. Furthermore, they can simulate negative values but to a lesser degree than the parametric models.

To address the drawbacks of the K-NN model we developed a modification to
Modified K-NN Method

We modified the traditional K-NN method described above, to alleviate its main drawback – i.e. not being able to generate values not seen in the historical record. The modification we developed here is briefly mentioned in the conclusion of Lall and Sharma [1996] and Rajagopalan and Lall [1999].

In this modification, a local (or nonparametric) regression is fitted to the successive monthly flows – e.g., a regression between March and February flows. Given the flow in the current month the fitted regression is used to obtain the mean flow of the next month. Now, k neighbors are computed from the data for the current month’s streamflow and residuals from the regression at these k neighbors are resampled using the same approach described in the previous section and added to the mean flow. Thus, instead of resampling the historical values, residuals are resampled from the neighborhood. This has three clear advantages:

(i) Values not seen in the historical record can be simulated
(ii) Residual resampling better captures local variability.
(iii) The local regression fit has the ability to capture any arbitrary (linear or nonlinear) relationship between the streamflows.

The algorithm is outlined below using Figure 2. This figure shows the scatter plot of February and March streamflows at the Glenwood Springs location. 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]. This is similar to the K-NN approach in that, at any point, a local polynomial is fit to the k nearest neighbors, The size of the neighborhood (i.e., k) and the order of the polynomial are obtained using an objective criteria called Generalized Cross Validation (see the above references and Prairie [2002]). This estimation is repeated
at several points to obtain the solid line in Figure 2. We used the package LOCFIT
developed by Loader (http://cm.belllabs.com/cm/ms/departments/sia/project/locfit/)
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:

   \[ y_t = f(y_{t-1}) + e_t \]  

2. The residuals \((e_t)\) from the fit are saved.

3. Once we have the value of the flow for the current month \(y_{t-1}\), we estimate the
   mean flow of the next month \(y_t^\ast\) from (4).

4. We then estimate the k-nearest neighbors to \(y_{t-1}\) (these are shown in big
   circles in Figure 2).

5. The neighbors are weighted using the weight function:

   \[ \frac{y_j}{\sum_{j=1}^{k} y_j} \]  

   (5)

   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 5,
   above. Consequently, its residual \((e_t^\ast)\) is resampled and added to the mean
   estimate \(y_t^\ast\). Thus, the simulated value for the next time step becomes,
\[ y_i^* + e_i^*. \]

7. Repeat 6 to obtain as many simulations as required. Repeat steps 1 through 6 for other months.

*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 mentioned that the heuristic scheme of \( k = \sqrt{N} \) works well in almost all the cases for \( 1 \leq \rho \leq 6 \) and \( N \geq 100 \), and we adopted the same scheme here, where \( \rho \) is the dimension of the model and \( N \) is the number of data points. In our monthly lag(1) \( (\rho = 1) \) model, and \( N \) is the number of years of data.

As can be seen, the modified K-NN model generates values not seen in the historic record and also has the ability to generate extreme values not seen in the history. Furthermore, it retains the basic capability of the traditional K-NN methodology of reproducing all the basic and higher order (i.e., marginal and conditional PDF) statistics.

**Model Evaluation**

We compared the three models (ISM, PAR and modified K-NN) by applying them to the natural streamflow at USGS stream gauge 09072500 (Colorado River near Glenwood Springs, CO). The monthly natural flow data was available for the 90-year period from 1906 to 1995. Natural flows were calculated by the USBR, from historic gauge records by removing anthropogenic effects such as consumptive use, reservoir regulation, imports, and exports. High variability in both annual and monthly flows can be seen in Figure 3.

We generated 100 simulations from these models of the same length as the historical data – for the ISM it will only be 90 simulations. A suite of statistics are computed from the simulations and compared with those of the historical data.
Model Evaluation Criteria

The following statistics are computed for comparison:

**Basic Statistics**

Monthly mean flows; standard deviation, lag-1 correlation (i.e., month to month correlation), coefficient of skewness, maximum and minimum flows.

**Higher order Statistics**

Marginal, bivariate and conditional PDFs. The density functions are estimated using nonparametric kernel density estimates [Bowman and Azzalini, 1997].

**Reservoir Statistics**

Drought (longest and maximum drought) and surplus (longest and maximum surplus) statistics. These are important in river basins for reservoir operation. The longest drought statistic is the maximum number of consecutive years of flows below the median flow, the maximum drought statistic is the maximum volume of water during a drought - vice-versa for surplus statistics

**Results**

The statistics from simulation ensembles and the historic data are shown as boxplots. The boxplots display the interquantile range (IQR) and whiskers extending to 1.5 * IQR for the PDFs of the 100 synthetic natural flow traces. The interquantile range indicates the range for 50 percent of the data around the mean. The horizontal line inside the IQR depicts the median of the data. The whiskers approximate the 5 percent and 95 percent confidence for the traces. Data beyond the whiskers (1.5 * IQR) are termed outliers and indicated by a solid circle. An example boxplot is given:
Historic data is shown as a solid circle with a solid line connecting each month.

As expected, all the models preserved well the mean and standard deviation for ISM (Figure 4), PAR (Figure 5), and K-NN (Figure 6). Skew coefficients and lag(1) correlations are shown in Figures 7, 8, 9, respectively. Skew coefficients are not well reproduced by PAR(1) because this depends on how well the transformation is able to approximate a Gaussian distribution, while modified K-NN model and ISM preserved the skews very well. The lag(1) correlations are preserved by all the models, as they are constructed to do so. Note that the ISM exactly reproduced the statistics with no variation around the historic values, because it uses the historic record exactly to develop the simulations. All the statistics from ISM simulations exhibit this trait, hence, we only discuss the comparisons with PAR(1) and modified K-NN model below.

Boxplots of maximum and minimum values are shown in Figures 10, 11 and 12, for ISM, PAR and K-NN, respectively. The modified K-NN preserves both of them, while PAR seems to overestimate the maximum and minimum values. This could be the result of transformation, wherein small values in the transformed space can lead to large values in the original space. The modified K-NN generated maximum and minimum values that are not seen in the historical data; which is an improvement over the traditional K-NN.
Figures 13-15 show the boxplots of PDFs of January and February streamflows, for the three models. For the month of January, the historical PDF seems to be close to Gaussian and this is reproduced by PAR (Figure 14) and the modified K-NN (Figure 15). The historical PDF of February flows exhibited a skewed distribution while, the PAR model (Figure 14) tended to reproduce a Gaussian equivalent of this and the nonparametric K-NN (Figure 15) captured the non-Gaussian feature very well. Boxplots from ISM are shown in Figure 13. Similar observations can be seen for the PDFs of September flows and annual flows, shown in Figure 16, 17, and 18, for the three models, ISM, PAR and K-NN, respectively.

The bivariate PDF was computed for the months of May and June from the historic data (Figure 19(a)) and from one of the simulations from PAR(1) (Figure 19(b)) and modified K-NN (Figure 19(c)) models. The historic bivariate PDF shows a significant non-Gaussian feature in the bivariate PDF. The PAR(1) model generated a Gaussian distribution, as expected, while the K-NN model preserved the non-Gaussian feature exhibited in the historic data.

We now computed the conditional PDF (shown in Figure 20) of June flows given a May flow of 187.6 cms. The simulations for PAR(1), as expected generated a Gaussian feature while the modified K-NN captures the slight non-Gaussian feature seen in the historical data.

The key point to note from these figures is that the PAR model is constrained to reproduce a Gaussian PDF, consistent with its theory, while the modified K-NN has the ability to reproduce any arbitrary PDF structure.

Lastly, we compared these models simulations on the drought and surplus statistics shown in Figures 21, 22, and 23 for the three models. The boxplots are shown as the ratio of the values from the simulations to the historical values. Both PAR(1) and K-NN are seen to preserve the drought and surplus statistics quite well. However, modified K-NN seems to underestimate the maximum drought.
Summary and Discussion

We presented a modified K-NN model that improves upon the traditional K-NN time series bootstrap. The proposed method fits a local polynomial between the variables. Next, the simulations involve obtaining the mean value from the local polynomial fit bootstrapping the historical residuals (i.e., residuals from the local fit at the historical values) that are added to the mean value to obtain ensembles. Consequently, it has the following three advantages (i) values not seen in the history can be generated, including extreme values, (ii) as the modeling philosophy is local, this provides the ability to capture any arbitrary (linear or non-linear) dependence structure and any arbitrary PDFs (Gaussian or non-Gaussian) exhibited by the data and (iii) it provides an assumption free and parsimonious framework that is easy to implement.

We compared this method to ISM and PAR(1), two widely used nonparametric and parametric models, respectively, in practice. The USBR uses ISM in their operational purpose. The models were compared on their ability to reproduce basic and higher order statistics (in particular, the joint, marginal and conditional PDFs) on monthly streamflow data from USGS stream gauge 09072500 (Colorado River near Glenwood Springs, CO). The modified K-NN performed much better in reproducing all the statistics. The PAR(1) model, as per its background theory is constrained to simulate Gaussian features. Non-Gaussian and non-linear features will be simulated to the extent that the transformations are adequate in transforming the data to a Gaussian distribution before applying the model. Furthermore, the PAR modeling framework involves estimating several parameters which can grow exponentially with the order of the model. The ISM on the other hand reproduces historical traces too closely, as a result, there is very little or no variety in the simulations and hence, its utility in capturing the uncertainty is limited.
The monthly model proposed in this paper does not capture the interannual variability very well (as seen in the drought statistics). This requires modification such as including the sum of previous twelve and more monthly flows to the conditioning vector in addition to the previous month’s flows [Sharma and O’Neil, 2002]. Sharma and O’Neil [2002] show that this modification in a kernel-based streamflow simulation model better captures the interannual variability. We plan to adopt a similar strategy in the modified K-NN model.

One of the significant advantages of the K-NN or (modified K-NN) framework is that variables can be easily added in the feature vector - e.g., if large scale climate information such as the state of El Nino Southern Oscillation (ENSO) index is important for streamflow simulations, this can easily be included into the conditioning vector. Preliminary results from including climate information in the modified K-NN streamflow simulation model to forecast spring streamflows of Truckee-Carson basin in, NV, USA are very encouraging [Grantz et al., 2002]. A variant of this approach has been applied for improved streamflow forecasts in North East Brazil [DeSouza and Lall, 2003].

In closing, the proposed modified K-NN model is an attractive and flexible alternative to other nonparametric methods. Extensions to include climate information, or space-time disaggregation of streamflows are relatively easy and parsimonious.

**Acknowledgments** This work was funded by the Bureau of Reclamation. The authors would like to thank Dave Trueman for his strong support and encouragement. Logistical support from CADSWES to conduct this research is thankfully acknowledged.
References


Figure Captions

Figure 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.

Figure 2. Nonlinear local regression fit to March natural flows dependent on February natural flows are depicted by the solid line ($k = 10$). A least square fit is shown with the dotted line.

Figure 3. A time series of annual water year natural flow and (b) monthly natural flow from water year 1906 to 1995 for USGS stream gauge 09072500 (Colorado River near Glenwood Springs, CO). The time series exhibit a high rate of variability both annually and monthly.

Figure 4. Boxplots of mean (upper plot) and standard deviation (lower plot) from ISM simulations and the historic values are shown as solid circles and joined with solid line.

Figure 5. Boxplots of mean (upper plot) and standard deviation (lower plot) from PAR(1) simulations and the historic values are shown as solid circles and joined with solid line.

Figure 6. Same as Figure 5, but for modified K-NN simulations.

Figure 7. Boxplots of lag-1 correlation (upper plot) and coefficient of skew (lower plot) from ISM simulations and the historic values are shown as solid circles and joined with solid line.

Figure 8. Boxplots of lag-1 correlation (upper plot) and coefficient of skew (lower plot) from PAR(1) simulations and the historic values are shown as solid circles and joined with solid line.

Figure 9. Same as Figure 8, but from modified K-NN simulations.

Figure 10. Boxplots of maximum (upper plot) and minimum (lower plot) from ISM simulations and the historic values are shown as solid circles and joined with solid line.

Figure 11. Boxplots of lag-1 maximum (upper plot) and minimum (lower plot) from PAR(1) simulations and the historic values are shown as solid circles and joined with solid line.

Figure 12. Same as Figure 11, but from modified K-NN simulations.
Figure 13. Boxplots of PDFs of January flows (upper plot) and February flows (lower plot) from ISM simulations. The historical PDF is shown as solid line.

Figure 14. Boxplots of PDFs of January flows (upper plot) and February flows (lower plot) from PAR(1) simulations. The historical PDF is shown as solid line.

Figure 15. Same as Figure 14, but for simulations from modified K-NN.

Figure 16. Boxplots of PDFs of September flows (upper plot) and Annual flows (lower plot) from ISM simulations. The historical PDF is shown as solid line.

Figure 17. Boxplots of PDFs of September flows (upper plot) and Annual flows (lower plot) from PAR(1) simulations. The historical PDF is shown as solid line.

Figure 18. Same as Figure 17, but for simulations from modified K-NN.

Figure 19. Bivariate probability density function for May and June flows of (a) historical data (b) one simulation from PAR(1) and (c) one simulation from modified K-NN.

Figure 20. Conditional PDF of June flows, conditioned on May flow of 1876 cms from PAR simulations (upper plot) and modified K-NN (lower plot).

Figure 21. Boxplot of drought and surplus statistics from ISM simulations. The boxplots are shown as the ratio of values from simulations to the historical values.

Figure 22. Boxplot of drought and surplus statistics from PAR(1) simulations. The boxplots are shown as the ratio of values from simulations to the historical values.

Figure 23. Same as Figure 22 but for simulations from modified K-NN.
Figure 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.

Figure 2: Nonlinear local regression fit to March natural flows dependent on February natural flows is depicted by the solid line (k = 10). A least square fit is shown with the dotted line.
Figure 3: (a) A time series of annual water year natural flow and (b) monthly natural flow from water year 1906 to 1995 for USGS stream gauge 09072500 (Colorado River near Glenwood Springs, CO). The time series exhibit a high rate of variability both annually and monthly.
Figure 4: Boxplots of mean (upper plot) and standard deviation (lower plot) from ISM simulations and the historic values are shown as solid circles and joined with solid line.
Figure 5: Boxplots of mean (upper plot) and standard deviation (lower plot) from PAR(1) simulations and the historic values are shown as solid circles and joined with solid line.
Figure 6: Same as Figure 5, but for modified K-NN simulations.
Figure 7: Boxplots of lag-1 correlation (upper plot) and coefficient of skew (lower plot) from ISM simulations and the historic values are shown as solid circles and joined with solid line.
Figure 8: Boxplots of lag-1 correlation (upper plot) and coefficient of skew (lower plot) from PAR(1) simulations and the historic values are shown as solid circles and joined with solid line.
Boxplots of 'new' KNN Simulated Lag(1) Correlation of Flows

Boxplots of 'new' KNN Simulated Skew of Flows

Figure 9: Same as Figure 8, but from modified K-NN simulations.
Figure 10: Boxplots of maximum (upper plot) and minimum (lower plot) from ISM simulations and the historic values are shown as solid circles and joined with solid line.
Figure 11: Boxplots of lag-1 maximum (upper plot) and minimum (lower plot) from PAR(1) simulations and the historic values are shown as solid circles and joined with solid line.
Figure 12: Same as Figure 11, but from modified K-NN simulations.
Figure 13: Boxplots of PDFs of January flows (upper plot) and February flows (lower plot) from ISM simulations. The historical PDF is shown as solid line.
PDFs from PAR(1) method for January

![Probability Distribution](image1)

PDFs from PAR(1) method for February

![Probability Distribution](image2)

Figure 14: Boxplots of PDFs of January flows (upper plot) and February flows (lower plot) from PAR(1) simulations. The historical PDF is shown as solid line.
Figure 15: Same as Figure 14, but for simulations from modified K-NN.
Figure 16: Boxplots of PDFs of September flows (upper plot) and Annual flows (lower plot) from ISM simulations. The historical PDF is shown as solid line.
Figure 17: Boxplots of PDFs of September flows (upper plot) and Annual flows (lower plot) from PAR(1) simulations. The historical PDF is shown as solid line.
PDFs from 'new' KNN method for September

PDFs from 'new' KNN method for Annual

Figure 18: Same as Figure 17, but for simulations from modified K-NN.
Figure 19: Bivariate probability density function for May and June flows of (a) historical data (b) one simulation from PAR(1) and (c) one simulation from modified K-NN.
Figure 20: Conditional PDF of June flows, conditioned on May flow of 187.6 cms from PAR simulations (upper plot) and modified K-NN (lower plot).
Figure 21: Boxplot of drought and surplus statistics from ISM simulations. The boxplots are shown as the ratio of values from simulations to the historical values.

Figure 22: Boxplot of drought and surplus statistics from PAR(1) simulations. The boxplots are shown as the ratio of values from simulations to the historical values.
Figure 23: Same as Figure 22 but for simulations from modified K-NN.