Environ. Sci. Technol. **2009,** *43,* 1407–1411

Simulating Ensembles of Source Water Quality Using a K-Nearest Neighbor Resampling Approach

ERIN TOWLER,*,† BALAJI RAJAGOPALAN,†,‡ CHAD SEIDEL,§ AND R. SCOTT SUMMERS†

Civil, Environmental, and Architectural Engineering Department, University of Colorado, 428 UCB, Boulder, Colorado 80309; Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309; Damon S. Williams Associates, 1624 Market Street, Suite 475, Denver, Colorado 80202

Received July 29, 2008. Revised manuscript received December 17, 2008. Accepted December 19, 2008.

Climatological, geological, and water management factors can cause significant variability in surface water quality. As drinking water quality standards become more stringent, the ability to quantify the variability of source water quality becomes more important for decision-making and planning in water treatment for regulatory compliance. However, paucity of long-term water quality data makes it challenging to apply traditional simulation techniques. To overcome this limitation, we have developed and applied a robust nonparametric *K*-nearest neighbor (*K*-nn) bootstrap approach utilizing the United States Environmental Protection Agency's Information Collection Rule (ICR) data. In this technique, first an appropriate "feature vector" is formed from the best available explanatory variables. The nearest neighbors to the feature vector are identified from the ICR data and are resampled using a weight function. Repetition of this results in water quality ensembles, and consequently the distribution and the quantification of the variability. The main strengths of the approach are its flexibility, simplicity, and the ability to use a large amount of spatial data with limited temporal extent to provide water quality ensembles for any given location. We demonstrate this approach by applying it to simulate monthly ensembles of total organic carbon for two utilities in the U.S. with very different watersheds and to alkalinity and bromide at two other U.S. utilities.

Introduction

Surface waters often exhibit significant spatial and temporal variability in water quality that can be attributed to climatological, geological, and water management factors. Many of these surface waters serve as the raw or "influent" water for drinking water treatment plants, and the influent water quality can exert a major impact on the water quality of the treatment plant effluent, or "finished-water" quality. Utilities that operate water treatment plants face complex treatment decisions when balancing changing influent water quality with finished water quality objectives. These decisions become even more complicated in light of new and changing regulatory requirements. To aid in decision-making, tools are needed to help utilities better characterize and understand influent water quality in order to assess treatment plant performance in light of changing regulations. To this end, a technique that can generate realistic influent water quality ensembles, and consequently quantify the variability, is very important. In some cases, the quantification of the influent variability can provide sufficient information to evaluate the impacts of regulatory changes and treatment options. In other cases, the ensembles can be used as input for an assessment model, from which the uncertainty and risk surrounding the decision variable(s) of interest can be robustly quantified.

Traditional methods for modeling uncertainty involve fitting a probability density function (pdf) to the observed data and using it to simulate "scenarios" that capture the variability, i.e., the Monte Carlo approach. There is a very rich history of this approach, especially in hydrology and water resources management (*1, 2* and references therein). However, this traditional approach has several drawbacks including (i) limited choice of the pdf that can be fit to the data, especially if the data exhibits skewness (*3*) or bimodal distribution, (ii) no choice other than normal distribution for more than one variable, (iii) not portable across sites—i.e., a single pdf cannot be prescribed for all the locations, and (iv) greatly influenced by outliers. Adding to these drawbacks is the fact that most national water quality information databases are substantially limited in some way, e.g., historical time frame, sample location geography, sample parameters, etc. (*4*). For example, few utilities have monitored total organic carbon (TOC), a parameter of concern, for more than 10 years. The United States Environmental Protection Agency's (USEPA) Information Collection Rule (ICR) database is the most comprehensive national drinking water-relevant data set, yet it is limited to an 18 month sample period (*5*). Thus, it is not possible to realistically apply the pdf-based traditional simulation approach to 18 observations at each location. Clearly, alternate approaches are needed. The objective of this paper is to develop a simple, robust, and flexible framework to generate influent water quality values and associated variability at locations that may have limited or no observed data.

Recent developments in nonparametric methods (see ref *6* for an overview of these methods and their applications to hydrologic problems) offer attractive alternatives to alleviate these drawbacks to a large extent. Within this, the *K*-nearest neighbor (*K*-nn) bootstrap technique (*7*) and its variations have been developed and applied successfully to generate scenarios of daily weather (*1, 8, 9*) and streamflows (*7, 10, 11*) for water management and salinity (*12*). This bootstrap method is data driven, easy to implement, and portable. Here, we adapt and develop a *K*-nn bootstrap technique to simulate influent water quality scenarios using the ICR database (*13*). While the ICR database has only 18 monthly values for the period of July 1997 through December 1998, it covers 500 treatment plants within 296 water systems across the U.S. (*5*). A strength of this approach is its ability to use the extensive spatial information to simulate temporal variability at a location, which can be challenging for conventional methods. In the following sections, we describe the methodology and demonstrate its applicability by applying it to two water utilities with very different watersheds for TOC and to two other water utilities for alkalinity and bromide.

* Corresponding author phone: 303-735-4147; fax: 303-492-7317; e-mail: towlere@colorado.edu.
† Civil, Environmental, and Architectural Engineering Department, University of Colorado.
‡ Cooperative Institute for Research in Environmental Sciences, University of Colorado.
§ Damon S. Williams Associates.

10.1021/es8021182 CCC: $40.75
Published on Web 01/30/2009

Methodology. Any attempt to generate scenarios is a conditional pdf simulation problem. For example, a utility might be interested in generating an ensemble of a water quality variable (x) for a given month conditioned on a suite of variables (referred to as the "feature vector", $\vec{y}$). Simulation involves fitting the conditional pdf, $f(x|\vec{y})$, and performing a Monte Carlo simulation. As mentioned above, traditional methods for computing the conditional pdf are extremely difficult. This is especially true in higher dimensions, due to limited choice of distribution functions (only normal distribution is possible in higher dimensions) and paucity of data. The K-nn bootstrap technique is akin to fitting the conditional pdf and simulating from it in a data driven manner.

To illustrate this technique, let us suppose that monthly ensembles of a water quality variable, x, are required at a given surface source water location. Also suppose that there are m surface source water data available (from the ICR database). The "feature vector", $\vec{y}$, consists of p explanatory variables. In this application, three variables are included (i.e., $p = 3$), which are: WQ = annual average water quality concentration, Lat = latitude, and Lon = longitude. In this application, the annual average value is calculated from the monthly 1998 data. As can be seen, the aim is to generate monthly ensembles of the water quality variable by conditioning on the annual average along with a suite of influencing variables. The explanatory variables chosen here represent the best variables available from the ICR data set. Each explanatory value helps to constrain the simulations, with the goal of making the ensembles realistic and adhere to their historical statistical properties. However, it is important to note that if this technique were transferred to another data set, other explanatory variables might be more appropriate. This is discussed further in the concluding remarks of this paper.

The steps of the algorithm are as follows:

(1) The user specified feature vector of p variables is defined. Here, as mentioned above, the feature vector is

$$y_{\text{user}} = \begin{bmatrix} \text{WQ}_{\text{user}} \\ \text{Lat}_{\text{user}} \\ \text{Lon}_{\text{user}} \end{bmatrix} \tag{1}$$

(2) The feature vector of all m source waters from the ICR database is constructed as

$$y_{\text{ICR}} = \begin{bmatrix} \text{WQ}_1 & \text{Lat}_1 & \text{Lon}_1 \\ \dots & \dots & \dots \\ \text{WQ}_i & \text{Lat}_i & \text{Lon}_i \\ \dots & \dots & \dots \\ \text{WQ}_m & \text{Lat}_m & \text{Lon}_m \end{bmatrix} \tag{2}$$

The user's feature vector will be one of the entries in the y_{ICR} matrix if monthly data at the user location is available.

(3) The covariance matrix S of the y_{ICR} matrix is computed.

(4) Weights for the three variables of the feature vector are assigned as

$$W = [W_{\text{WQ}} \quad W_{\text{Lat}} \quad W_{\text{Lon}}] \tag{3}$$

(5) Weighted Mahalanobis distances d_i are computed between the y_{user} vector and the vector of the ith eligible source water in the y_{ICR} matrix as

$$d_i = \sqrt{(W \times (y_{\text{user}} - y_i)^T) S^{-1} (W^T \times (y_{\text{user}} - y_i))} \tag{4}$$

for all $i = 1$ to m. T is the transpose of the vector. The Mahalanobis distance has an advantage over other distance metrics in that the components of the feature vector do not need to be scaled (*8, 14*).

(6) The distances d_i are sorted and the K-nearest neighbors are chosen. The neighbor with the smallest d_i value is considered the first or "nearest" neighbor, and so on until the Kth neighbor. There are several methods for selecting K, but the heuristic rule, $K = \sqrt{m}$, with its theoretical justifications (*7, 15*) seems to work well and has been used by all the earlier applications in the aforementioned references. However, this can be chosen using objective methods as well (*7*). For this application, there are 323 surface water treatment plants that are potential neighbors in the ICR database, so $K = \sqrt{m} = \sqrt{323} \approx 18$.

(7) A probability metric is used to assign weights to each of the K-nearest neighbors. To this end, a weight function given

$$p_j = \frac{1/j}{\sum_{i=1}^{K} 1/i} \tag{5}$$

for all $j = 1$ to K is created. This results in the closest neighbor receiving the highest weight and the furthest (i.e., the Kth neighbor) receiving the lowest weight, directly corresponding to how frequently each neighbor is resampled. For example, with 18 nearest neighbors, the nearest neighbor gets selected about 29% of the time ($p_1 = 0.29$), while the 18th/Kth neighbor gets picked about 1.6% of the time ($p_{18} = 0.016$). By taking the cumulative sum of the weights, a cumulative distribution function that enables the resampling of neighbors is created as

$$cp_i = \sum_{j=1}^{i} p_j \tag{6}$$

for all $i = 1$ to K. Other weight functions using the distances to the neighbors (e.g., bisquare function) can also be devised. However, it has been found from a variety of applications that the simulations are robust to the choice of the weight function — this is demonstrated in Lall and Sharma (*7*).

(8) To simulate a value for the first month (e.g., January), one of the K neighbors is resampled using the cumulative weight function described in step 7, and the January value corresponding to the identified neighbor is selected. This is repeated to generate simulations for all the twelve months and for as many years as desired.

In this application, all the variables in the feature vector are assigned equal weights in the computation of the distance and consequently the selection of the nearest neighbors (i.e., $W_{\text{WQ}} = W_{\text{Lon}} = W_{\text{Lat}}$). However, these weights can be optimized (*16, 17*) or prescribed by the user as relevant to the situation. For instance, if the variability of the water quality parameter is known to be similar longitudinally, then the weights on the latitude and the other member(s) of the feature vector can be reduced or eliminated, thus constraining the neighborhood to the longitudinal direction. The flexibility of the weights in the neighbor selection is illustrated in Figure 1.

Results

For each utility considered, the K-nn technique was used to generate 500 influent water quality values for each month, January (J) through December (D). In Figures 2–5, the ensembles are shown as box plots in which the box represents the 25th and 75th percentile, the whiskers show the 5th and 95th percentiles, points are values outside this range, and the horizontal line represents the median. In addition, each figure includes a box plot representing the annual average value (WQ from the feature vector) for each of the K neighbors. The statistics of the simulations were compared to that of the observed ICR data at this location to evaluate the performance of the technique. The observed ICR data

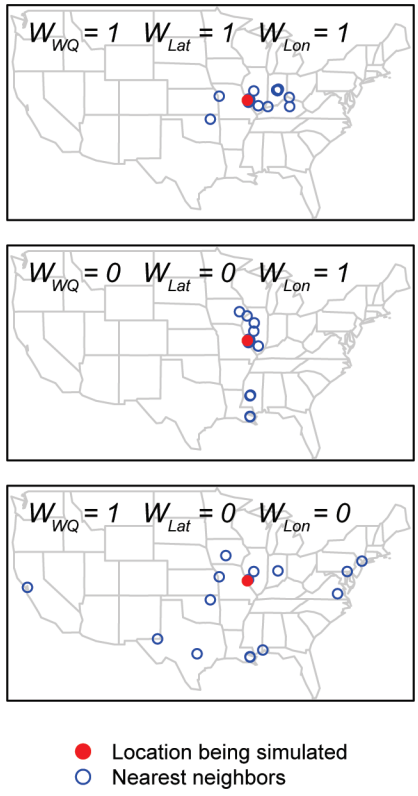

FIGURE 1. The flexibility of the weighting scheme is shown by the following cases: (a) all weights are equal, (b) neighbors are chosen by similarity in longitude, and (c) neighbors are only chosen by their similarity in terms of annual average concentration.

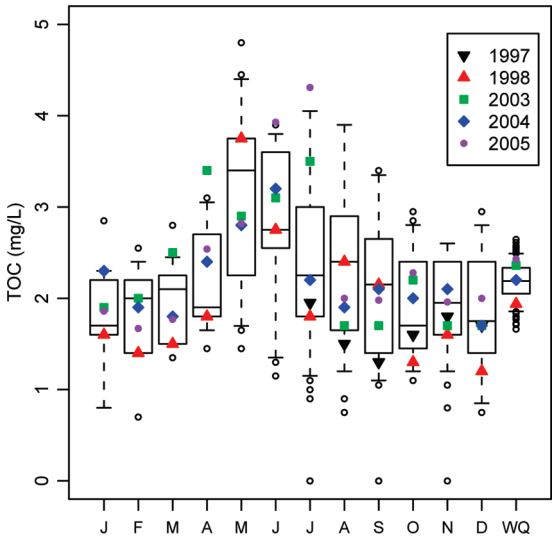

FIGURE 2. Monthly TOC concentration simulations (J through D), annual average TOC concentrations for the *K*-nearest neighbors (WQ), and measured values for the City of Boulder's Betasso Water Treatment Plant.

included 6 months in 1997 and 12 months in 1998. If the observed value falls within the range of the ensemble it suggests that the simulations well-capture the historical properties.

Simulation of monthly influent TOC concentrations for the City of Boulder's Betasso Water Treatment Plant (CO) was first considered. The influent water quality to this plant is impacted by snowmelt during spring runoff. Simulations for monthly TOC at the Boulder utility are shown in Figure 2. In addition to the ICR data from 1997 and 1998, data

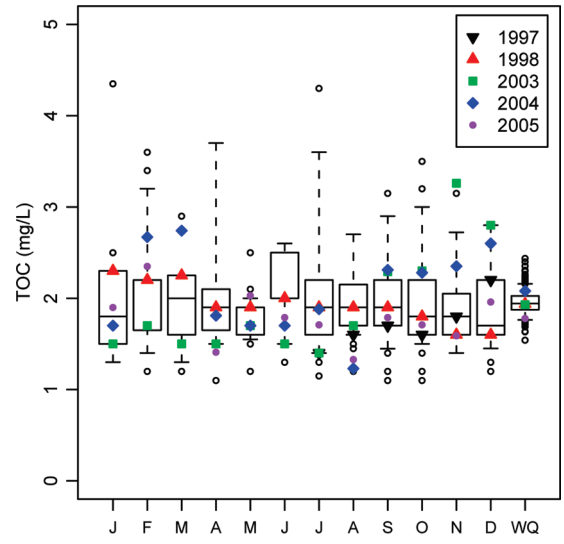

FIGURE 3. Monthly TOC concentration simulations (J through D), annual average TOC concentrations for the *K*-nearest neighbors (WQ), and measured values for the for the City of Birmingham's Carson Filter Plant.

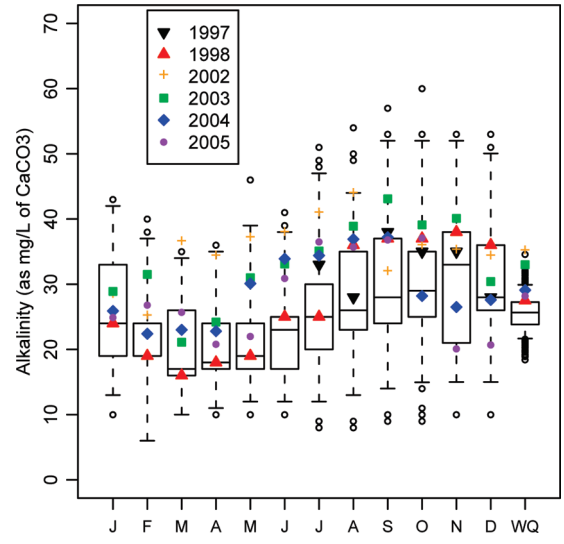

FIGURE 4. Monthly alkalinity concentration simulations (J through D), annual average alkalinity concentrations for the *K*-nearest neighbors (WQ), and measured values for the New Jersey American Water Swimming River Treatment Plant.

representing monthly and annual average observed values from 2003, 2004, and 2005 are overlaid on the figure. The seasonality of the water quality is very well captured and most of the observed values fall within the range of the ensembles.

Figure 3 shows the simulation of plant influent TOC concentration values for the City of Birmingham's Carson Filter Plant (AL) using the utility's ICR data. The raw water source, Inland Lake, is a large reservoir that attenuates the impact of changes in runoff water quality. While the observed annual TOC values, with a mean of 1.93 mg/L, are not that dissimilar from those of Boulder, with a mean of 2.23 mg/L, there is much less seasonal variation in the monthly values and the range of the monthly inner quartile values is less than that for the Boulder data. Again the simulations well capture the distribution of the recently observed data.

The technique can be extended to other water quality variables as well. Figure 4 shows simulations of the source water alkalinity data for the New Jersey American Water Swimming River Treatment Plant (NJ). The alkalinity varies

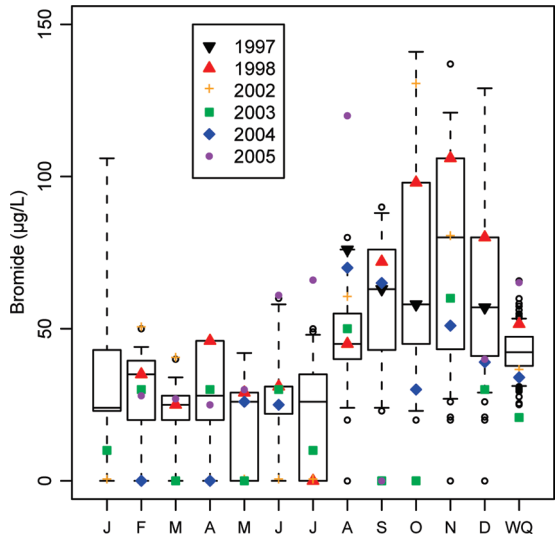

FIGURE 5. Monthly bromide concentration simulations (J through D), annual average bromide concentrations for the *K*-nearest neighbors (WQ), and measured values for the Greater Cincinnati Water Works Richard Miller Water Treatment Plant.

seasonally with lower values occurring in the spring and higher values in the fall. The simulation captures the seasonality of the alkalinity data. Observed monthly data from 2002 to 2005 tends to fall on the top whiskers, but largely within the range of the simulations.

Figure 5 shows the simulation of the influent bromide concentration in the Ohio River at the Greater Cincinnati (OH) Water Works Richard Miller Water Treatment Plant. The bromide varies seasonally with high values occurring during the late summer and fall, which is when the river has a lower streamflow. The simulation captures this seasonality of the ICR bromide data and predicts high monthly variability, especially in the fall. Some (28%) of the observed monthly data from 2002 to 2005 were below the detection limit (which varied between 10 and 20 μg/L) and were therefore set to 0. It is noted that this biases the 25th percentile values for May and July, as well as the 5th percentile values for the first part of the year (January through July), to appear artificially low. However, it is generally the higher bromide values that are of concern, and the higher range percentiles ($\geq$50th) are not affected by this treatment of the left-censored data. The ability of the ensemble to capture the observed data was good, however, some of the more recent observations fall outside the range of the simulations, suggesting that longer records of bromide data might be useful in fully capturing the range of its variability.

Discussion

In this paper, we develop a *K*-nn based bootstrap technique for simulating the variability of influent water quality at a location by conditioning on a feature vector that is formed from the best available explanatory variables. The technique is demonstrated by its application to locations with varied watersheds and to the different variables of TOC, alkalinity, and bromide. In this method, nearest neighbors of a feature vector are identified and one of the neighbors is selected to generate an ensemble. This is akin to "locally" (in the vicinity of the feature vector) estimating the conditional pdf based on the *K*-nearest neighbors and simulating from it. This "local" aspect is the main difference from the traditional methods, which provides the ability to capture any variability structure present in the data. Traditional methods, which involve fitting a parametric pdf for each month, would be flawed due to severe data limitations. The *K*-nn approach on the other hand, takes advantage of the extensive data in space to overcome the limited temporal data, as is the case with the ICR database. As can be seen, the technique is simple, robust, portable, and theoretically sound. Another advantage of this methodology is that with an appropriate feature vector, realistic variability of the influent water quality can be generated for locations that may have limited or no observed data.

The framework has flexibility in selecting a wide neighborhood range depending on the site specific features. This can include varying the weights associated with each component of the feature vector in order to get the optimal neighbors for a location. We found the spatial variables (latitude and longitude) and average water quality value to be indispensable, but other explanatory variables when available need to be considered. For instance, if a database that included water quality, hydrologic factors, topographic information, land-use type, etc., became available, this resampling scheme could be modified to include the relevant explanatory variables in the feature vector. In addition, the availability of a longer monthly time series could change the time scale of the feature vector: instead of conditioning on average annual concentration, each month could be conditioned on its own average monthly concentration. As such, this resampling method that was crafted specifically for the ICR database could be useful in many environmental engineering decision contexts.

The *K*-nearest neighbors selected remain the same for all the months but each monthly simulation comes from any of the neighbors, thus providing a rich variety in the monthly sequences. Since the simulations are based on resampling of the historical observations, "new" values not present in the database are not generated, but at any location a rich variety of plausible values not seen at that location are simulated, as can be seen by the box plots. New values can be obtained by fitting a local polynomial to the nearest neighbors (e.g., ref (*18*),) and resampling from the local error structure (*12*). Other variations in this multivariate regression framework can also be explored, such as generalized linear modeling (*19*). The regression framework can be computationally intensive and also may require several models for each month to capture the seasonality. The *K*-nn resampling method presented here provides a simpler and robust alternative. However, we are exploring these complementary alternatives.

Ensemble simulation and forecasting approaches are gaining favor in water resource management (*20*−*22*) and have enormous potential for water utility decision-making and planning. By using the open source ICR database, we have developed a tool that can be used to simulate variability on a large-scale with limited data, useful for regulatory, treatment, and risk assessments.

Acknowledgments

We acknowledge AwwaRF project 3115, "Decision Tool to Help Utilities Develop Simultaneous Compliance Strategies" for partial financial support on this research effort. In addition, we thank the staff of the City of Boulder's Betasso Water Treatment Plant (CO), the City of Birmingham's Carson Filter Plant (AL), the New Jersey American Water Swimming River Treatment Plant (NJ), and the Greater Cincinnati (OH) Water Works Richard Miller Water Treatment Plant for providing data. We are grateful to two anonymous reviewers whose comments greatly improved the manuscript.

Literature Cited

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

(2) Rajagopalan, B.; Lall, U.; Tarboton, D. G.; Bowles, D. S. Multivariate nonparametric resampling scheme for generation of daily weather variables. *Stochastic Hydrol. Hydraulics* **1997**, *11*, 65–93.
(3) Helsel, D. R.; Hirsch, R. M. *Statistical Methods in Water Resources*; Elsevier: Amsterdam; New York, 1995.
(4) Frey, M. F.; Seidel, C.; Edwards, M.; Parks, J. L. *Occurrence Survey of Boron and Hexavalent Chromium*; Awwa Research Foundation: Denver, CO, 2004.
(5) McGuire, M. J.; McLain, J. L.; Obolensky, A. *Information Collection Rule Data Analysis*; AWWA Research Foundation: Denver, CO, 2002.
(6) Lall, U. Recent advance in nonparametric function estimation—Hydrologic application. *Rev. Geophys.* **1995**, *33*, 1093–1102.
(7) Lall, U.; Sharma, A. A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.* **1996**, *32*, 679–693.
(8) Yates, D.; Gangopadhyay, S.; Rajagopalan, B.; Strzepek, K. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resour. Res.* **2003**, *39*, 1199.
(9) Buishand, T. A.; Brandsma, T. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resour. Res.* **2001**, *37*, 2761–2776.
(10) Prairie, J. R.; Rajagopalan, B.; Fulp, T. J.; Zagona, E. A. Modified K-NN model for stochastic streamflow simulation. *J. Hydrol. Eng.* **2006**, *11*, 371–378.
(11) Grantz, K.; Rajagopalan, B.; Clark, M.; Zagona, E. A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resour. Res.* **2005**, *41*, W10410.
(12) Prairie, J. R.; Rajagopalan, B.; Fulp, T. J.; Zagona, E. A. Statistical nonparametric model for natural salt estimation. *J. Environ. Eng.-Asce* **2005**, *131*, 130–138.
(13) United States Environmental Protection Agency ICR Auxiliary 1 Database Version 5.0. Query Tool Version 2.0 (CD-ROM) 2000.
(14) Davis, J. C. *Statistics and Data Analysis in Geology*; Wiley: New York, 1986.
(15) Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: New York, 1972.
(16) Young, K. C. A multivariate chain model for simulating climatic parameters from daily data. *J. Appl. Meteorol.* **1994**, *33*, 661–671.
(17) Yakowitz, S.; Karlsson, M. Nearest neighbor methods with application to rainfall/runoff prediction, In *Stochastic Hydrology*; Macneil, J. B.; Humphries, G. J., Eds.; D. Reidel: Norwell, MA, 1987.
(18) Towler, E. L. *Characterizing and Incorporating Uncertainty in Water Quality and Treatment*; Master's Thesis, University of Colorado: Boulder, CO, 2006.
(19) McCullagh, P.; Nelder, J. A. *Generalized Linear Models*; Chapman and Hall: London; New York, 1989.
(20) Souza, F. A.; Lall, U. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. *Water Resour. Res.* **2003**, *39*, 1307.
(21) Grantz, K.; Rajagopalan, B.; Zagona, E.; Clark, M. Water management applications of climate-based hydrologic forecasts: Case study of the Truckee-Carson River Basin. *J. Water Resour. Plann. Manage.-Asce* **2007**, *133*, 339–350.
(22) Regonda, S. K.; Rajagopalan, B.; Clark, M.; Zagona, E. A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin. *Water Resour. Res.* **2006**, *42*, W09404.

ES8021182