ENVIRONMENTAL ENGINEERING SCIENCE
Volume 26, Number 8, 2009
© Mary Ann Liebert, Inc.
DOI: 10.1089/ees.2008.0341

ORIGINAL ARTICLE

Using Parametric and Nonparametric Methods to Model Total Organic Carbon, Alkalinity, and pH after Conventional Surface Water Treatment

Erin Towler,* Balaji Rajagopalan, and R. Scott Summers

Department of Civil, Environmental, and Architectural Engineering, University of Colorado, Boulder, Colorado.

Received: October 15, 2008 Accepted in revised form: May 26, 2009

Abstract

Predicting the behavior of natural organic matter (NOM), alkalinity, and pH during drinking water coagulation is difficult because of the heterogeneous chemical nature of NOM and the complexity of carbonate chemistry. Parametric and nonparametric statistical regression methods were implemented to model the removal of NOM, as measured by total organic carbon (TOC), from raw water by conventional surface water treatment and to track the behavior of pH and alkalinity. The United States Environmental Protection Agency (U.S. EPA) Information Collection Rule (ICR) database was sampled for raw water and postsedimentation data from conventional surface water plants. All models were evaluated in terms of their fit and predictive capability, and for all variables explored, the nonparametric local polynomial models outperformed their parametric linear least-squares counterparts. This was most pronounced with the pH model, and was attributed to the nonlinear relationship found between pH and one of the predictors. Between the sedimentation basin and the plant effluent, alkalinity was found to remain relatively constant, TOC decreased by 12% by filtration, and pH increased, consistent with chemical additions required to minimize corrosion in the distribution system. Modeling efforts in this article are meant to be complementary to previous chemical and process models of water treatment.

Key words: drinking water quality; water treatment; natural organic matter; statistical models; regression analysis

Introduction

Modeling drinking water treatment processes is challenging because of the complex chemical and physical nature of organic and inorganic constituents. The removal of natural organic matter (NOM) by coagulation, in particular, is difficult because of the heterogeneous chemical nature of NOM. Thus, utilities normally rely on historical records, operator experience, bench scale jar testing, and trial and error approaches to adjust process conditions. The lack of robustness in this approach compromises the ability of utilities to plan for future scenarios, such as estimating the economic impacts of proposed regulations or understanding the feasibility of inserting additional processes.

Improvements on this approach have come from theoretically motivated models, such as Edwards' (1997) Langmuir-based semiempirical model used to predict NOM removal during coagulation. However, use of Edwards' model is limited, as it requires the coagulation pH for input. The United States Environmental Protection Agency's (U.S. EPA) original Water Treatment Plant Model (WTPM) used an empirical equation that included initial total organic carbon (TOC), alum dose, and pH (Harrington *et al.*, 1992) to predict NOM removal, as measured by TOC. The WTPM was modified by Solarik *et al.* (2000) to include Edwards' model, which is coupled with a model of the carbonate system that allows the pH of coagulation to be predicted based on the initial pH, alkalinity, and the coagulant dose.

Another approach to modeling drinking water treatment processes is through statistical models. The biggest limitation to this approach can be obtaining a suitable dataset that includes relevant water quality and chemical treatment addition data. The availability of the U.S. EPA's Information Collection Rule (ICR) data (U.S. EPA, 2000), which represents the most comprehensive national drinking water-relevant dataset to date, provides a unique opportunity to examine field-scale data from treatment plants all over the United States (McGuire *et al.*, 2002). Multiple linear regression models have been used with this dataset to examine relationships between water quality, treatment, and disinfection byproduct (DBP) formation (Obolensky and Singer, 2008). In addition, nonparametric regression techniques have been used successfully in modeling NOM breakthrough in granular-activated carbon

Corresponding author: Department of Civil, Environmental, and Architectural Engineering, University of Colorado, UCB 428, Boulder, CO 80309-0428. *Phone:* 303-735-4147; *Fax:* 303-492-7317; *E-mail:* towlere@gmail.com

(GAC) adsorbers using the ICR database (Zachman *et al.*, 2007). Statistical models are straightforward to implement and provide valuable information to consultants and researchers in the water industry to aid in initial decision making, effective pilot-scale design, and regulatory planning.

In this article, two statistical regression methods—traditional linear regression and nonparametric regression—are compared and offered as tools for modeling the overall coagulation/flocculation/sedimentation process (from raw water through sedimentation basin) in conventional surface water treatment plants. The response variables modeled were TOC, a measure of NOM, and two water quality variables that impact NOM removal by coagulation: pH and alkalinity. The response of these water quality parameters from the sedimentation basin to finished water quality was also examined. The regression models were constructed using monthly water quality and chemical addition data recorded at conventional surface water plants monitored under the U.S. EPA's ICR (U.S. EPA, 2000). Objective criteria were used to evaluate the "best" predictors for each model, and they were validated in terms of their predictive skill. The nonparametric TOC model was coupled with an uncertainty simulator to examine output variability and the likelihood of exceeding a given limit. The proposed approach aims to gain insights from the extensive data-gathering effort of the ICR and to provide a predictive tool to help water utilities make better decisions. These approaches are offered as a complement to existing chemical and process models.

Data Set and Predictors

The data used to develop and validate the models in this investigation were from conventional surface water utility plants in the continental United States, and obtained from the U.S. EPA's ICR database (U.S. EPA, 2000). The ICR database covers 18 monthly intervals covering July 1997 through December 1998 (McGuire *et al.*, 2002). The predictors for the variables considered included raw water quality variables and selected chemical additions between raw water and the sedimentation basin.

For modeling the postsedimentation TOC (TOC_{sed}), the following predictors were considered: influent TOC (TOC_{in}), pH (pH_{in}), alkalinity (alk_{in}), turbidity ($turb_{in}$), temperature ($temp_{in}$), total hardness ($t\text{-}hard_{in}$), total specific ultraviolet absorbance ($TSUVA_{in}$), and coagulant dose. The modeling effort included all observations where the suite of paired predictors were complete (i.e., nonmissing). We constrained the dataset to influent TOC values of greatest interest for prediction, which we determined to be greater than 1 mg/L. In addition, an upper limit for influent TOC was determined to be 10 mg/L and turbidity values were constrained below 50 NTU, which removed large outliers, and resulted in a decrease in only 104 data points (about 4% of the available). For the resulting TOC modeling dataset, the sample size was 2,291, coming from 180 different utility water sources. For postsedimentation pH and alkalinity (pH_{sed} and alk_{sed}), the following predictors were considered: pH_{in}, alk_{in}, $turb_{in}$, $temp_{in}$, coagulant dose, lime dose, and chlorine dose. Turbidity values were again constrained to be less than 50 NTU, and the resulting sample sizes were 2,997 (186 utility water sources) and 3,019 (186 utility water sources), respectively, for modeling pH and alkalinity.

Model Development

In this investigation, both parametric linear regression and nonparametric local polynomial methods were employed. Because both of these techniques have been well documented in the literature, this article is limited to a brief overview of the main points of each technique. The reader is referred to the references throughout the following section for a detailed review.

Statistical prediction models can be represented as:

$$y = f(x) + e$$

where f is a function fit to a set of predictor variables (x), y is the dependent variable of interest, and e is the associated estimation error, generally assumed to be Normally distributed (with mean of 0 and variance σ^2) and independent. We present two approaches to estimating the function f, which are described below.

Linear regression

Traditionally, a linear relationship between the predictors and the dependent variable is assumed and is fit, of the form:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$$

where there are k predictor variables and the β coefficients are estimated from the data so as to minimize the mean squared errors (Helsel and Hirsch, 1995). The theory behind linear regression models is well developed with software packages to implement them readily available; hence, they are widely used (Helsel and Hirsch, 1995; Rao and Toutenburg, 1999). Higher orders of the predictor variables (e.g., squares and cubes) can be included in the above equation to fit nonlinear functional forms. The fitted equation is used to estimate the value of the dependent variable at future independent variable values.

However, this traditional approach has several drawbacks including (1) the assumption of a normal distribution of data and errors, (2) the assumption of a linear relationship between the predictors and the dependent variable, (3) higher order fits (e.g., quadratic or cubic) require large amounts of data for fitting, (4) the models may not be portable across data sets, and (5) estimates of model parameters are greatly influenced by outliers (Rajagopalan *et al.*, 2005).

Nonparametric regression

Nonparametric methods offer an attractive alternative to alleviating the drawbacks of the traditional linear regression approach. In this approach, the estimate of the function at any point, say, x^*, is influenced by the data points within a small neighborhood of x^*. Thus, no single equation is fit to the entire data, as in the case of traditional linear regression approach. This "local" fitting provides the capability to capture any nonlinear features that might be present locally in the data. As will be seen, the nonparametric methods are more computationally intensive than their linear counterpart, but with the enormous increase in computation power in recent years, this is no longer an issue. Lall (1995) provides an excellent review of nonparametric methods and their various applications to hydrologic applications.

There are several local functional estimation approaches, including kernel-based (Bowman and Azzalini, 1997), splines, *K*-nearest neighbors (*K*-NN) local polynomials (Owosina, 1992; Rajagopalan and Lall, 1999), and locally weighted polynomials (Loader, 1999). The locally weighted polynomials, henceforth referred to as LOCFIT, are used in this application, as they are computationally efficient, easy to implement, and robust. Furthermore, with the availability of the powerful LOCFIT library (Loader, 2004) in the statistical software R, the implementation is made easy. This has been successfully used for salinity and flow modeling (Prairie *et al.*, 2005, 2006), streamflow forecasting (Grantz *et al.,* 2005; Regonda *et al.,* 2006a) and in other hydrologic applications.

The implementation steps of LOCFIT are as follows: for any point of interest, say, x^*

(1) $K=(\alpha \times N)$ *K*-NN from the observational data are identified, where α is the fraction of the observational data (ranging between 0 and 1) and N is the sample size.
(2) A polynomial of order P is fit to the identified *K*-NN. In this study, first-order ($P=1$) and second-order ($P=2$) polynomials were considered.
(3) The fitted polynomial (from steps 1 and 2) is used to estimate the value of the dependent variable, $Y(x^*)$, at x^*.
(4) The residuals from the polynomial fitted to the *K*-NN are used to obtain the standard error variance (σ_{le}^2) of the estimate (Loader, 1999, pp. 29–30).
(5) Repeat (1) through (4) for all points of interest.

The polynomial coefficients are estimated by minimizing the weighted mean squared errors—as opposed to the mean squared errors in the traditional linear regression. The *K*-NNs are weighted based on their proximity to x^* with highest weights to the nearest neighbors and zero weights to the farthest. Any weight function can be used to provide the weights, and the approach is insensitive to the choice of the weight function. The neighbors are selected based on the Euclidean distance in the variable space, and the variables are scaled before the distance is computed. Notice that if K is set to N (i.e., all the available observation data), P is set to 1, and all the neighbors are given equal weights, this approach collapses to the traditional linear regression. Thus, the local polynomial approach offers a general framework with the traditional linear regression model being a subset.

The two parameters of the approach, K and P, have to be identified for a given observation data. This is obtained using the generalized crossvalidation (GCV) function. The combination of K and P that minimizes the GCV function is chosen as the best set of parameters for the LOCFIT. The GCV function is defined as,

$$GCV(K,P)=\frac{\sum_{i=1}^{N}\frac{(y_i-\hat{y}_i)^2}{N}}{\left(1-\frac{m}{N}\right)^2}$$

where $y_i-\hat{y}_i$ is the residual (error) between the observed and predicted values, N is the number of data points, and m is the degrees of freedom of the fitted polynomial (Loader, 1999, p. 31). If all of the points are used (i.e., $\alpha=1$ so $K=N$) and weighted equally and, the polynomial order is 1, then the GCV for the parametric linear regression is calculated. The GCV has been found to be a good estimate of the predictive risk of the model, unlike other functions, which are goodness of fit measures (Craven and Whaba, 1979). A step-by-step overview of this process can be seen in Prairie *et al.* (2005). In this application, the LOCFIT package of the statistical software R, developed by Loader (2004), was employed.

The GCV measure can also be used to identify the best subset of predictors (Regonda *et al.*, 2005, 2006b). This entails finding the combination of predictors (and the associated parameters K and P) that result in a minimum GCV value. This has been used in modeling water quality variables and shown to improve upon the traditional stepwise regression methods (Zachman *et al.*, 2007).

Alternative regression frameworks

This investigation compares two modeling approaches, but we note that other variations in the multivariate regression framework can be explored. A more generalized linear modeling framework, known as generalized linear modeling (GLM) (McCullagh and Nelder, 1989) is more flexible than the traditional linear regression model presented here. In addition, the so-called "curse of dimensionality," or the fact that the density of data become sparser in high dimensional space, is a limitation of multivariate regression. Principle component (PC) regression is one alternative, whereby the dominant modes of variability are determined by a PC analysis and then regressed against the response variable (Hidalgo *et al.*, 2000). This framework eliminates the problem of multicollinearity, allows for the regression to be performed in a lower dimensional space, and has been successfully demonstrated in water resources applications (Regonda *et al.*, 2006b; Gangopadhyay *et al.*, in press). Additive models are another appealing, nonparametric alternative that can alleviate problems associated with high-dimensional data analysis and improve the interpretability of the results (Hastie and Tibshirani, 1990; Wood, 2006). Another modeling approach that has been explored in a host of water management applications is artificial neural networks (Zhang and Stanley, 1999; Yu *et al.*, 2000; Dawson and Wilby, 2001; Shariff *et al.*, 2004; Schulze *et al.*, 2005).

We also note that the dataset modeled in this study is an example of longitudinal data, because the measurements were collected repeatedly over time (i.e., 18 months), for different subjects (i.e., individual utility water sources). Longitudinal data often results in mean values being clustered within subjects, as well as serial correlation within subjects. We note that these are potential limitations to both modeling frameworks presented here. As such, procedures for modeling and analyzing this type of data have surfaced as a field in their own right (e.g., Weiss, 2005). Linear mixed models are an example of a longitudinal data modeling framework that can provide for both of these types of limitations (Diggle *et al.*, 1994).

Model Evaluation and Validation

For each dependent variable explored in this investigation, a suite of linear and local polynomial regression models was fit with different variable combinations ranging from 2 to 6. The "best" variable subset for each of the two methods was chosen based on the GCV score. *R*-squared and root-mean-squared error (RMSE) were calculated on the model errors (observed minus predicted) as a way to quantify and compare model performances with other studies. These measures, in addition to the GCV score, provided a comprehensive quantification of internal validation. In practice, the skill of these models in a predictive setting is desired. For this purpose, it is common to fit the model on a portion of the data and predict the withheld data and compute the skills. Although this is an acceptable approach, the predictive skill depends on the data withheld. To address this the following approach was used: (1) 10% of the data are randomly selected and withheld, (2) the models are fit to the remaining data set (using the same K and P as the original model), (3) the fitted models are used to estimate the values at the withheld points (see Step 3 of the LOCFIT implementation) and skill measures computed, (4) Steps (1) through (4) are repeated a number of times (100 in this case). The skill scores from the simulations from the two methods are compared as box plots.

Results and Discussion

TOC model results

The modeling of postsedimentation TOC concentration was first considered. Table 1 shows the variable combinations, GCV scores, R^2, and RMSE values for the top five local polynomial models (referred henceforth as, NP-1 through NP-5) and linear models (referred henceforth as, L-1 through L-5). In the table, the regression coefficients for each independent variable were reported for the linear model, but were designated as Xs for the local polynomial models because the coefficients change depending on where the prediction is being estimated. The table also reports the α and P parameters for the NP models. In terms of variable selection, all 10 "best" models chose alk$_{in}$, TOC$_{in}$, and coagulant dose to be included. pH$_{in}$ was chosen the least, with only one of the linear models including it. Temp$_{in}$ was chosen by all of the linear models, but by only two of the local polynomial models. Turb$_{in}$, TSUVA$_{in}$, and t-hard$_{in}$ were chosen in only some of the models for both methods. In terms of GCV score, R^2, and RMSE, the top five local polynomial models performed better than the top five linear models. However, within each respective method (NP or L), there was little difference in the GCV, R^2, and RMSE values among the top five models. This is common in real data sets and often ignored in traditional stepwise regression approaches. In such cases, multimodel approaches that combine estimations from all the top models have been advocated and shown to improve the predictions (e.g., Regonda *et al.*, 2006b). A multimodel approach was investigated in this study, but did not effectively improve the model results, thus, only the top model within each category (i.e., NP-1 and L-1) was used.

Figure 1 shows the comparison of the observed and estimated values for the TOC$_{sed}$ concentrations from L-1 and NP-1 models. It can be seen that estimates from NP-1 showed a tighter scatter around the one-to-one line (straight lines in the figure) compared to the linear model. The residuals from linear regression model showed much more heteroscedasticity and skewness than the NP-1 model (figures not shown). This invalidates the assumption of residual normality and furthermore, a constant error variance assumption in the face of heteroscedasticity can pose a problem in estimating valid prediction intervals. Transformation of the data, identifying additional predictor variables, and iterated parameter estimation methods are potential modifications to contend with this problem (Carroll and Ruppert 1988; Helsel and Hirsch, 1995). The NP-1 model exhibited mild heteroscedasticity as mentioned above, and the error variance is estimated "locally" (Loader, 1999) at each point; hence, the impact on the estimates and predictions are not widespread. It was interesting to note that for NP-1, the best α was found to be 0.5 (i.e., half of the data points were used in each local estimation) and that a second degree polynomial ($P = 2$)

Table 1. Selected Predictor Variables and Goodness-of-Fit Statistics for Five Best-Fit Nonparametric and Linear Models for Predicting Postsedimentation Total Organic Carbon Concentrations

	Nonparametric TOC$_{sed}$ models					*Linear TOC$_{sed}$ models*				
Variable	*NP-1*	*NP-2*	*NP-3*	*NP-4*	*NP-5*	*L-1*	*L-2*	*L-3*	*L-4*	*L-5*
Intercept						0.211	0.285	0.202	0.128	0.266
pH$_{in}$									0.0102	
alk$_{in}$ (mg/L[a])	X	X	X	X	X	0.00299	0.00271	0.00270	0.00262	0.00477
turb$_{in}$ (NTU)	X	X		X	X	−0.00531				−0.00489
temp$_{in}$ (C)	X	X				0.00954	0.00948	0.00980	0.00980	0.00975
TOC$_{in}$ (mg/L)	X	X	X	X	X	0.546	0.542	0.539	0.540	0.546
TSUVA$_{in}$ (L/m-mg)	X		X		X		−0.0320			−0.0224
coag[b] (mmol/L)	X	X	X	X	X	−0.865	−0.835	−0.887	−0.882	−0.825
t-hard$_{in}$ (mg/L)				X	X	0.00130	0.00133	0.00141	0.00141	
GCV	0.178	0.182	0.184	0.184	0.185	0.251	0.252	0.252	0.253	0.253
R^2	0.829	0.819	0.810	0.813	0.821	0.715	0.714	0.712	0.712	0.712
RMSE	0.388	0.398	0.407	0.405	0.397	0.499	0.500	0.501	0.501	0.501
α	0.50	0.35	0.35	0.50	0.60					
P	2	2	2	2	2					

Predictor variables are designated with Xs for the nonparametric (NP) fits and with regression coefficient values for the linear (L) fits.

[a] As mg/L CaCO$_3$.

[b] Coagulant dose.

TOC, total organic carbon (mg/L); GCV, generalized crossvalidation; RMSE, root-mean-square error. See Data Set and Predictors for variable definitions.

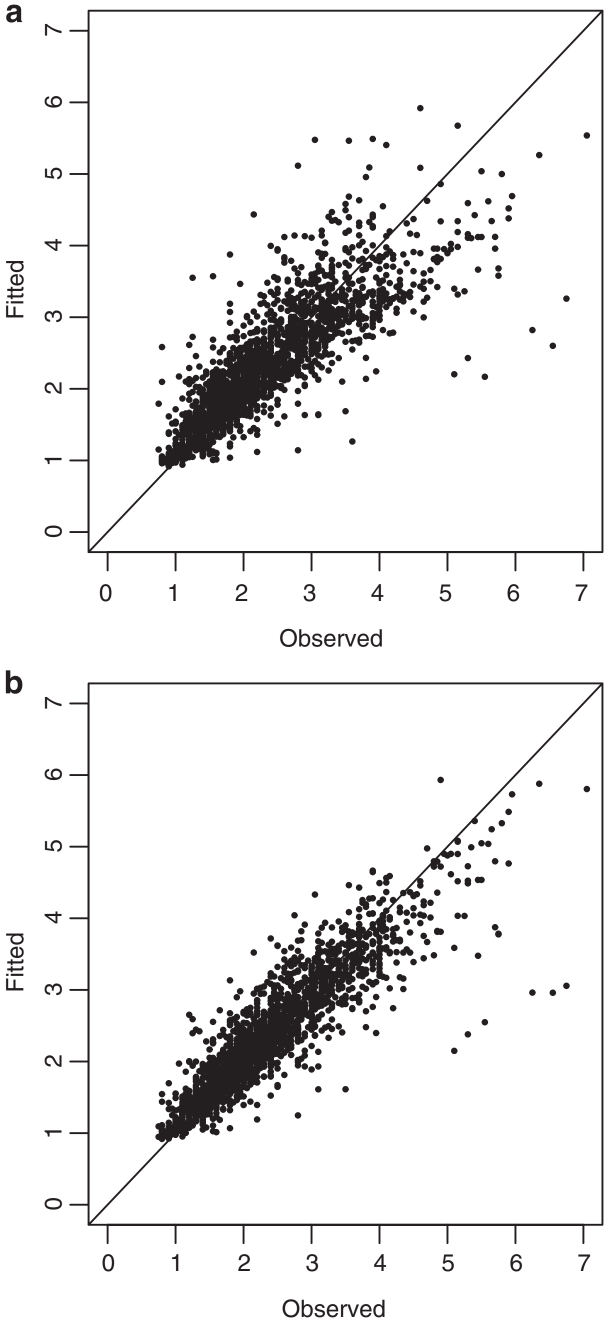

FIG. 1. Observed vs. fitted TOC concentrations in the sedimentation basin (TOC_{sed}) from the best linear model (**a**) and the best local polynomial model (**b**).

provided the best fit, indicating that there were underlying nonlinearities in the data.

The direction of each coefficient in the L-1 TOC model was examined for their consistency with prior knowledge. The chemistry and application of coagulation for NOM removal is summarized in Letterman *et al.* (1999). The alk_{in} and $t\text{-}hard_{in}$ coefficients were positive, consistent with the fact that higher alkalinity (and higher total hardness) waters are associated with characteristics that make them less amenable to coagulation (Archer and Singer, 2006). The $turb_{in}$ coefficient was negative, which was consistent with findings that coagulated turbidity forms flocs that can serve as adsorption sites for NOM, which would decrease TOC_{sed} (Letterman *et al.*, 1999). The coefficient of $temp_{in}$ being positive was counterintuitive, as higher rates of reaction, and therefore more TOC removal, would be expected in association with higher temperatures. The TOC_{in} had a positive coefficient, because higher raw water TOC values will likely yield relatively higher TOC_{sed} values. The coefficient associated with the coagulant dose was negative, underscoring the fact that TOC was removed during coagulation. It should be noted that each regression coefficient represents a "partial" effect, conditioned on all other predictors being held constant. When predictors are correlated, the resulting regression model and the associated parameter coefficients can be different from a model based on any single variable. Regression based on principal component analysis obviates this problem, and can therefore be an attractive alternative (Hidalgo *et al.*, 2000).

Alkalinity and pH model results

Table 2 shows the variable combinations, GCV scores, R^2, RMSE, α, and P values for the best-fit models for post sedimentation alkalinity (alk_{sed}) and pH (pH_{sed}). The NP-1 model for alk_{sed} performed better than its L-1 counterpart in terms of GCV and RMSE, but was quite similar in terms of the fitted R^2. In addition, plots of the NP-1 and L-1 estimates with the observed values for alk_{sed} were very similar (not shown), with the L-1 model exhibiting only slightly more spread, but no systematic pattern of over- or underestimation.

The NP-1 model for pH_{sed} performed much better than its L-1 counterpart in terms of GCV, R^2, and RMSE and with a tighter scatter (Fig. 2) in all pH ranges, but especially for observed pH_{sed} values below 7 and above 8. For pH_{sed} values of about 9.5, there was substantial underestimation by the NP-1 model, but below this value the model fit was better. The L-1

Table 2. Selected Predictor Variables and Goodness-of-Fit Statistics for Modeling pH and Alkalinity in the Sedimentation Basin

	Model			
	alk_{sed}		*pH_{sed}*	
Variable	*NP-1*	*L-1*	*NP-1*	*L-1*
Intercept		17.5		4.90
pH_{in}		−2.48	X	0.274
alk_{in} (as mg/L $CaCO_3$)	X	0.972	X	0.00476
$turb_{in}$ (NTU)	X	−0.135	X	−0.00892
$temp_{in}$ (C)		−0.133		−0.00875
Coagulant dose (mmol/L)	X	−23.4	X	
Lime dose (mg/L)	X	0.229	X	0.0466
Chlorine dose (mg/L)	X		X	−0.0306
GCV	70.9	139	0.172	0.479
R^2	0.978	0.950	0.819	0.351
RMSE	7.80	11.8	0.366	0.691
α	0.20	—	0.20	—
P	2	—	2	—

Predictor variables are designated with Xs for the best nonparametric (NP-1) and with regression coefficient values for the best linear (L-1) fits.

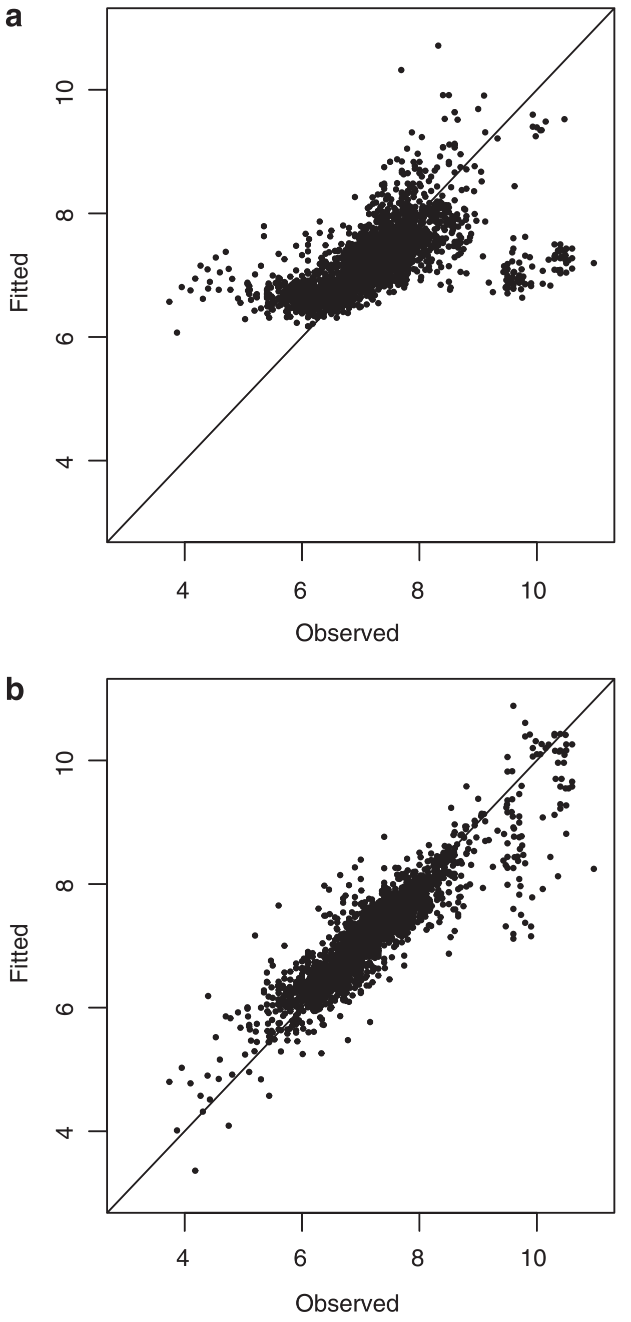

FIG. 2. Observed vs. fitted pH in the sedimentation basin (pH_{sed}) data for the best linear model (**a**) and the best local polynomial model (**b**).

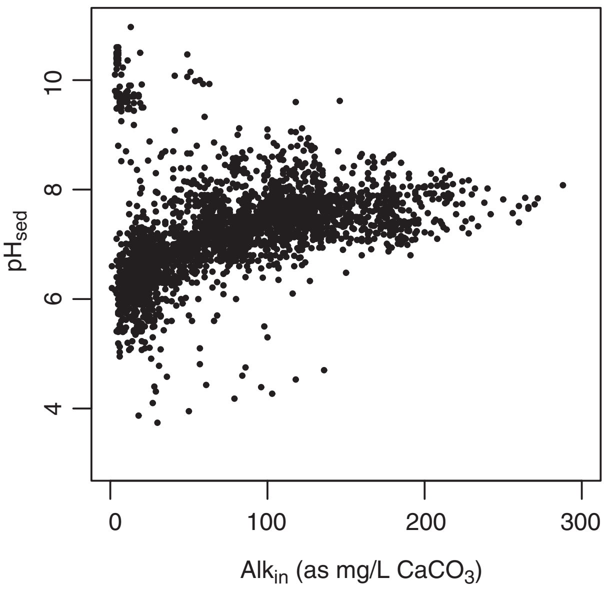

FIG. 3. Influent alkalinity (alk_{in}) vs. pH in the sedimentation basin (pH_{sed}).

model did a relatively good job at estimating observed values between 6 and 8, but showed a systematic bias with observed values below 6 being overestimated and values above 8 underestimated. This was investigated further by examining the relationship of each of the predictor variables with the pH_{sed}. In terms of the relationship between alk_{in} and pH_{sed}, a strong nonlinearity can be observed (Fig. 3). This nonlinearity could be contributing to the ineffectiveness of the linear model in this case. It was interesting and consistent to note that for both pH_{sed} and alk_{sed} models the best α for the NP-1 model was found to be 0.2 (i.e., 20% of the points were used in each local estimation) and a second degree polynomial ($P=2$) was utilized.

Similar signs in the coefficients of L-1 pH and alkalinity models were expected, because pH and alkalinity are linked through carbonate equilibrium. Indeed, this was the case with alk_{in}, $turb_{in}$, $temp_{in}$, and lime. However, in the pH model, the coefficient of pH_{in} was found to be positive as to be expected, unlike what was found for the alkalinity model. The coagulant dose coefficient was negative for post sedimentation alkalinity, although coagulant dose was found not to be a contributor to postsedimentation pH, somewhat surprising because coagulants consume alkalinity and can lead to lower pH values. The chlorine dose was found to contribute to the pH model with a negative coefficient, as would be expected when chlorine is added as an acid.

Validation: TOC, alkalinity, and pH models

The results of validating these models by witholding 10% of the data 100 times are shown for R^2 in Fig. 4 as box plots. The box represents the 25th and 75th percentile (inner quartile range), the whiskers show the 5th and 95th percentiles, points are values outside this range, and the horizontal line represents the median. The box plots show the range of uncertainty, with a wider box indicating larger uncertainty. The overlaid gray triangle shows the values from the fit based on the entire data. Table 3 shows the results for RMSE. In general, the NP-1 model exhibited higher R^2 and lower RMSE values relative to the L-1 model, especially for pH_{sed}. We note that this validation method is robust in that it provides an estimation of the predictive skill uncertainty. Because the goal of these models is to provide skillful prediction, these additional metrics are more informative than any single measure that only tests the goodness of fit.

Up to this point, the model validation was conducted using data from the original ICR database. In practice, one might be interested in evaluating the models using more recent data

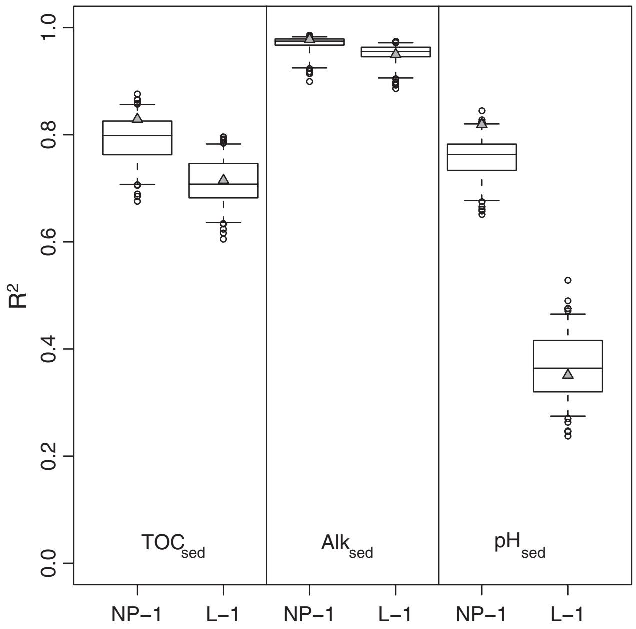

FIG. 4. R^2 values of simulations when 10% of the data is dropped for modeling postsedimentation TOC (TOC_{sed}), alkalinity (alk_{sed}), and pH (pH_{sed}). NP-1 and L-1 are the best models for the nonparametric and linear cases, respectively, for each dependent variable being considered. Gray triangles are R^2 values when all of the data is used in the model.

Table 3. Percentile Values from the Simulations of RMSE Values When 10% of the Data Is Dropped for Each Dependent Variable Modeled in the Sedimentation Basin

	Model					
	TOC_{sed} (mg/L)		*Alk_{sed} (as mg/L $CaCO_3$)*		*pH_{sed}*	
Percentile	*NP-1*	*L-1*	*NP-1*	*L-1*	*NP-1*	*L-1*
5%	0.36	0.42	68	97	0.35	0.60
25%	0.38	0.46	76	100	0.38	0.65
50%	0.41	0.49	85	110	0.42	0.69
75%	0.45	0.52	99	130	0.45	0.73
95%	0.51	0.57	150	160	0.48	0.79

NP-1 and L-1 are the best models for the nonparametric and linear cases, respectively.

(e.g., post-ICR). Under ideal circumstances, such an approach would benefit from recent data from all of the utilities that participated in the original ICR. Not surprisingly, this type of data is not available, but we were able to obtain post-ICR data from five U.S. utilities that were part of the original ICR who were interested in the TOC model. The data set contained 352 data points from the five U.S. utilities, with dates ranging from 2001 to 2007. All of the data were from utilities that practice conventional surface water treatment (at least from influent to the sedimentation basin). The validation was demonstrated on the nonparametric TOC model because it outperformed its parametric counterparts and was of greatest interest to the utilities involved. However, two of the utilities were unable to provide raw water observations of UV absorbance that could be used in the analysis. Thus, the external validation was done on the second best nonlinear model (see Table 1), which did not include TSUVA, which is calculated from UV absorbance, as an input variable. The RMSE TOC_{sed} value for this independent data set was 0.86 mg/L, which was higher than the 0.40 mg/L that was calculated for NP-2 (see Table 1).

The fact that the nonparametric model did not perform as well with the post-ICR data set as might be expected brings up a few points. First, the empirical models were built on older data sets than the post-ICR data used to validate it. This should not be a problem, as long as there have not been changes to the process being modeled. However, some process changes have occurred in the last 10 years since the ICR data was collected. For one, the ICR database was constructed prior to the disinfection and disinfection byproducts (D/DBP) rules. Thus, some utilities may have begun to practice enhanced coagulation. To account for these changes, the model needs to be updated by incorporating data as it becomes available. This can be accomplished by simply inputting the new data and rerunning the model. As an extension to this analysis, we appended a random quarter of the new data to the ICR data and created an updated nonparametric model, with an RMSE of 0.41 mg/L. When all of the independent data was run through the updated model, the RMSE was 0.49 mg/L, which is closer to the value from the original model.

Incorporating input variability: TOC model

A useful application of this work is that the models can be coupled with an uncertainty simulator to evaluate various output scenarios. Because postsedimentation TOC is mainly affected by influent TOC, the best TOC_{sed} model (NP-1) was run with 500 input scenarios per month, which were generated for influent TOC (Towler *et al.*, 2009). The other predictor variables were held constant at their average monthly values. Essentially, the input scenarios are generated from a probability density function (pdf) of the possible influent values, which can be run through the model to gain an output pdf. This output can be used to represent the potential variability of the TOC in the sedimentation basin.

To demonstrate, we examine one utility in the ICR database whose TOC concentrations varied seasonally. The ICR data was collected before the D/DBP Rule, but this provides an opportunity to retrospectively examine how this information could have been used for initial planning purposes. For instance, in the Stage 2 D/DBP Rule, one method of compliance is achieved when the TOC concentration after the sedimentation basin is less than 2 mg/L. By running the influent TOC simulations through the TOC NP-1 model, we can examine the output scenarios, which allow for the calculation of the probability of exceeding this limit. These results are shown for 2 months—May and November—in Fig. 5. Here it can be seen that for this utility, November only has a 12% chance of exceeding that value, but May has a 73% chance of exceeding that value. This information is useful for utility planning. The operational changes that will need to occur on seasonal time scale can now be estimated, helping to smooth the operational transition for the enhanced coagulation regulation. In addition, because DBP formation is directly related to TOC during

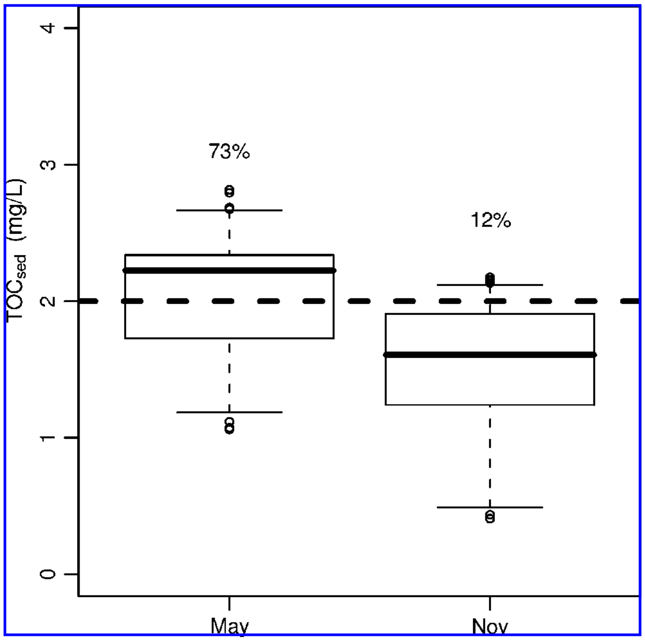

FIG. 5. TOC_{sed} concentrations simulated by the model (box plots) and a limit threshold (dotted line). Values above box plots correspond to the probability of exceeding the 2 mg/L threshold in a given month.

chlorination, a better idea of potential DBP concentrations can be reached.

Water quality changes from postsedimentation to finished water

The scatterplot of postsedimentation pH with its corresponding values in the finished water is shown in Fig. 6. The local smoother (gray line) shows that the average finished water pH was 7.5 for waters with pH_{sed} values below 8. This reflects the utility practice of increasing the pH after sedimentation to minimize corrosion in the distribution system. Thus, although the pH_{sed} model is useful, if a process is going to be added after sedimentation, a new model should be fit, including the relevant chemical additions, to capture the relationship from sedimentation to finished water.

In the comparison of TOC_{sed} and alk_{sed} to their respective finished water values, both variables showed tight scatters around the one-to-one line (not shown), indicating that there was not much of a change from postsedimentation to finished water. The best fit line between TOC_{sed} vs. the finished water TOC produced a slope equal to 0.88 (with intercept equal to zero), revealing that an additional 12% of the TOC was removed between the sedimentation basin and the finished water. In addition, the mean error (ME, or average of the residuals) was slightly negative (−0.24 mg/L). These results reflect removal of particulate organic matter by the filter or biodegradation by biomass that can accumulate on filter media when chlorine is not present. The slope of the alkalinity plot was 1.0 (with intercept set to zero), indicating that there was no change in alkalinity. The ME value was slightly positive for alkalinity (4.8 mg/L), reflecting some base addition for corrosion control.

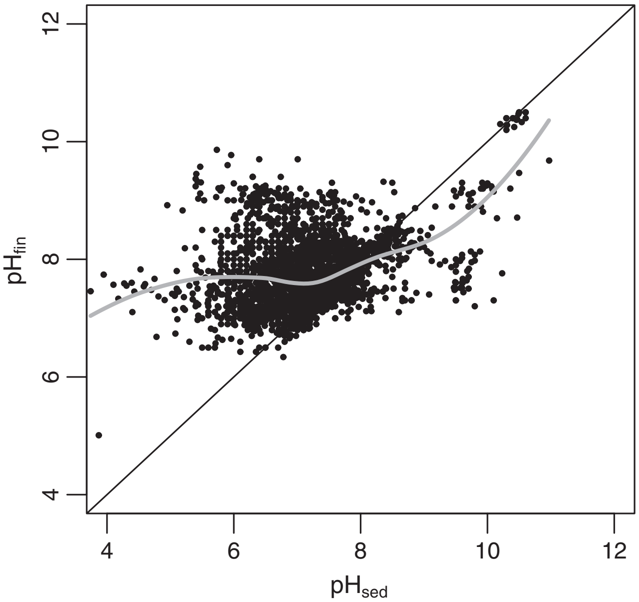

FIG. 6. Scatterplot of finished pH (pH_{fin}) vs. postsedimentation pH (pH_{sed}). Gray line is local smoother.

Summary and Conclusions

Statistical methods of regression are useful in modeling efforts, especially when the underlying relationships are complex. Nonparametric regression methods have shown promising results in a variety of water management applications (e.g., Prairie *et al.*, 2005; Grantz *et al.*, 2007), and this article extends their influence to modeling conventional drinking water treatment (coagulation/flocculation/sedimentation) of surface waters. In this investigation, the nonparametric local polynomial models outperformed their parametric linear least-squares counterparts in terms of fit and predictive capability. This was most pronounced with the pH model, and was attributed to the nonlinear relationship found between pH and one of the predictors.

We recognize that the concept of "locally" fitting a model, as in nonparametric regression, requires a paradigm shift from the traditional way of thinking about statistical models. Traditional linear regression results in a single model equation, whereas in nonparametric regression, the equation is "locally" evaluated at each desired point. The dynamic nature of nonparametric models provides valuable flexibility in capturing any arbitrary underlying feature (i.e., linear or nonlinear). The additional computing time required is barely an issue with recent advances in computing power. If the underlying relationships being modeled are known to be linear, then it is practical to use linear regression. However, we point out that this is not known a priori, and often times the data exhibits local nonlinear features; thus, the nonparametric approach provides an attractive alternative.

As water quality standards heighten, being able to estimate intermediate and finished water quality is important. In many cases, decision-making tools are developed to help utilities weigh various options, such as additional processes, as they plan for their future. Predictive models, such as those developed in this article, could be useful in estimating variables of consequence that could be inputs in a decision-making tool.

A framework, such as the one presented here, provides additional means for advancing water treatment planning and adaptation.

Acknowledgments

The authors acknowledge the AwwaRF project 3115, "Decision Tool to Help Utilities Develop Simultaneous Compliance Strategies," for partial financial support on this research effort. We also thank the five utilities that provided additional data. We are grateful to six anonymous reviewers whose comments greatly improved the manuscript.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

Archer, A.D., and Singer, P.D. (2006). An evaluation of the relationship between SUVA and NOM coagulation using the ICR database. *J. AWWA* 98, 110.

Bowman, A.W., and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Carroll, R.J., and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman & Hall.

Craven, P., and Whaba, G. (1979). Optimal smoothing of noisy data with spline functions. *Numer. Math*. 31, 377.

Dawson, C.W., and Wilby, R.L. (2001). Hydrological modeling using artificial neural networks. *Prog. Phys. Geog.* 25, 80.

Diggle, P.J., Kung-Yee, L., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Edwards, M. (1997). Predicting DOC removal during enhanced coagulation. *J. AWWA* 89, 78.

Gangopadhyay, S., Harding, B.L., Rajagopalan, B., Lukas, J.L., and Fulp, T.J. (in press). A non-parametric approach for paleohydrologic reconstruction of annual streamflow ensembles. *Water Resour. Res.*

Grantz, K., Rajagopalan, B., Clark, M., and Zagona, E. (2005). A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resour. Res*. 41, W10410.

Grantz, K., Rajagopalan, B., Clark, M., and Zagona, E. (2007). Water management applications of climate-based hydrologic forecasts: Case study of the Truckee-Carson River Basin. *J. Water Res. Pl. ASCE* 133, 339.

Gupta, A.K., and Shrivastava, R.K. (2006). Uncertainty analysis of conventional water treatment plant design for suspended solids removal. *J. Environ. Eng. ASCE* 132, 1413.

Harrington, G.W., Chowdhury, Z.K., and Owen, D.M. (1992). Developing a computer model to simulate DBP formation during water treatment. *J. AWWA* 84, 78.

Hastie, T.J., and Tibshirani, R.J. *Generalized Additive Models*. London: Chapman & Hall.

Helsel, D.R., and Hirsch, R.M. (1995). *Statistical Methods in Water Resources*. New York: Elsevier.

Hidalgo, H.G., Piechota, T.C., and Dracup, J.A. (2000). Alternative principal components regression procedures for dendrohydrologic reconstructions. *Water Resour. Res*. 36, 3241.

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

Letterman, R.D., Amirtharajah, A., and O'Melia, C.R. (1999). Coagulation and flocculation. In R.D. Letterman, Ed., *Water Quality and Treatment: A Handbook of Community Water Supplies*, 5th ed. New York: McGraw-Hill, Inc.

Loader, C. (1999). *Local Regression and Likelihood*. New York: Springer.

Loader, C. (2004). Locfit: local regression, likelihood and density estimation. R package version 1.1-9. Available at: http://cm.bell-labs.com/stat/project/locfit/.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. New York: Chapman & Hall.

McGuire, M.J., McLain, J.L., and Obolensky, A. (2002). *Information Collection Rule Data Analysis*. Denver, CO: AwwaRF.

Obolensky, A., and Singer, P.C. (2008). Development and interpretation of disinfection byproduct formation models using the Information Collection Rule database. *Environ. Sci. Technol*. 42, 5654.

Owosina, A. (1992). Methods for assessing the space and time variability of groundwater data. M.S. Thesis, Utah State University, Logan, Utah.

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

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

Rajagopalan, B., Grantz, K., Regonda, S., Clark, M., and Zagona, E. (2005). Ensemble streamflow forecasting: methods and applications. In U. Aswathanarayana, Ed., *Advances in Water Science Methodologies.* The Netherlands: Taylor and Francis.

Rajagopalan, B., and Lall, U. (1999). A nearest neighbor bootstrap resampling scheme for resampling daily precipitation and other weather variables. *Water Resour. Res*. 35, 3089.

Rao, C.R., and Toutenburg, H. (1999). *Linear Models: Least Squares and Alternatives*. New York: Springer.

Regonda, S.K., Rajagopalan, B., and Clark, M. (2006a). A new method to produce categorical streamflow forecasts. *Water Resour. Res*. 42, W09501.

Regonda, S.K., Rajagopalan, B., Clark, M., and Zagona, E. (2006b). A multi-model ensemble forecast framework: application to spring seasonal flows in the Gunnison River Basin. *Water Resour. Res*. 42, W09404.

Regonda, S.K., Rajagopalan, B., Lall, U., Clark, M., and Moon, Y. (2005). Local polynomial method for ensemble forecast of time series. *Nonlinear Proc. Geoph*. 12, 397.

Schulze, F.H., Wolf, H., Jansen, H.W., and Van Der Veer, P. (2005). Applications of artificial neural networks in integrated water management: fiction or future? *Water Sci. Technol.* 52, 21.

Shariff, R., Cudrak, A., and Stanley, S.J. (2004). Lime softening clarifier modeling with artificial neural networks. *J. Environ. Eng. Sci*. 3, S69.

Solarik, G., Summers, R.S., Soh, J., Swanson, W.J., Chowdhury, Z.K., and Amy, G.L. (2000). Extensions and verification of the water treatment plant model for disinfection by-product formation. In S.E. Barrett, S.W. Krasner, and G.L. Amy, Eds., *ACS Symposium Series 761: Natural Organic Matter and Disinfection By-Products, Characterization and Control in Drinking Water.* Washington, DC: American Chemical Society.

Towler, E., Rajagopalan, B., Seidel, C., and Summers, R.S. (2009). Simulating ensembles of source water quality using a k-nearest neighbor resampling approach. *Environ. Sci. Technol.* doi: 10.1021/es8021182.

U.S. Environmental Protection Agency. (2000). *ICR Auxiliary 1 Database Version 5.0. Query Tool Version 2.0 (CD-ROM),* EPA 815-C-00-002, Washington, DC: U.S. EPA.

Weiss, R.E. (2005). *Modeling Longitudinal Data.* New York: Springer.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall.

Yu, R.F., Kang, S.F., Liaw, S.L., and Chen M. C. (2000). Application of artificial neural network to control the coagulant dosing in water treatment plant. *Water Sci. Technol.* 42, 403.

Zachman, B., Rajagopalan, B., and Summers, R.S. (2007). Modeling NOM breakthrough in GAC adsorbers using nonparametric regression techniques. *Eviron. Eng. Sci*. 24, 1280.

Zhang, Q., and Stanley, S.J. (1999). Use of neural networks for process modeling and control of coagulation. *J. Environ. Eng. ASCE* 125, 153.