Environmental Science & Technology

Article

pubs.acs.org/est

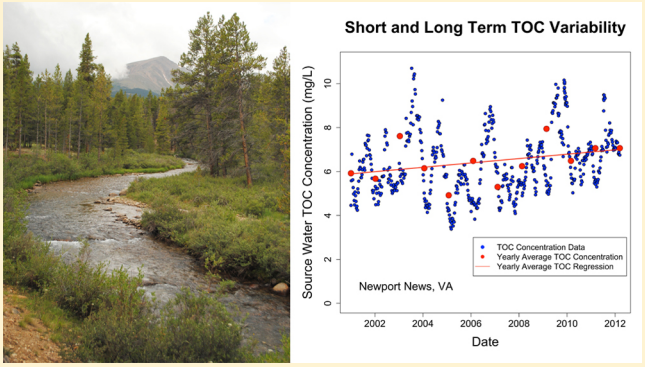
Modeling Source Water TOC Using Hydroclimate Variables and Local Polynomial Regression

Carleigh C. Samson,* Balaji Rajagopalan, and R. Scott Summers

University of Colorado Boulder, Department of Civil, Environmental, and Architectural Engineering, College of Engineering & Applied Science, 428 UCB, 1111 Engineering Drive, Boulder, Colorado 80309-0428, United States

Supporting Information

ABSTRACT: To control disinfection byproduct (DBP) formation in drinking water, an understanding of the source water total organic carbon (TOC) concentration variability can be critical. Previously, TOC concentrations in water treatment plant source waters have been modeled using streamflow data. However, the lack of streamflow data or unimpaired flow scenarios makes it difficult to model TOC. In addition, TOC variability under climate change further exacerbates the problem. Here we proposed a modeling approach based on local polynomial regression that uses climate, e.g. temperature, and land surface, e.g., soil moisture, variables as predictors of TOC concentration, obviating the need for streamflow. The local polynomial approach has the ability to capture non-Gaussian and nonlinear features that might be present in the relationships. The utility of the methodology is demonstrated using source water quality and climate data in three case study locations with surface source waters including river and reservoir sources. The models show good predictive skill in general at these locations, with lower skills at locations with the most anthropogenic influences in their streams. Source water TOC predictive models can provide water treatment utilities important information for making treatment decisions for DBP regulation compliance under future climate scenarios.



1. INTRODUCTION

Variations in source water quality can affect the ability of drinking water utilities to meet regulations and provide safe potable water. Variations can be short-term, such as seasonal patterns, or long-term, related to changes in land use or climate change. Source water dissolved organic matter (DOM) is an important water quality component, as certain organic matter (OM) fractions react with chlorine, the most commonly used disinfectant in water treatment, to form disinfection byproducts (DBPs).[1] Some DBPs are of health concern and two groups of DBPs, total trihalomethanes (TTHM), and five haloacetic acids (HAA5), are regulated under the EPA Stage 2 D/DBP Rule.[2] Under this rule, the maximum contaminant level (MCL) for TTHM is set at 80 μg/L and that for HAA5 at 60 μg/L. DBP formation begins in water treatment plant (WTPs) and continues into the distribution system, as U.S. EPA regulations for systems that utilize surface water (SW) sources also require a detectable disinfectant residual throughout the distribution system, i.e., secondary disinfection. The Stage 2 D/DBP Rule became effective in 2012, requiring DBP compliance monitoring at locations with the highest formation.[2] For TTHMs this is typically at the end of the distribution system. Organic carbon (OC), both as total OC (TOC) and dissolved OC (DOC), concentration, is the most commonly used OM measure and has been shown to be well related to TTHM and HAA5 formation, especially when other water quality and treatment factors are controlled,[3] and has been used in regression models to predict the formation of these DBPs.[4,5] In addition to the TOC concentration, other factors, in particular, higher water temperature, lead to increased DBP formation.[4,6,7]

To control DBP formation, WTPs typically remove OM prior to disinfection, which can be more effective than directly removing preformed DBPs.[8] WTPs that utilize SW must also remove OM to meet the required percent TOC removal based on source water TOC and alkalinity concentrations.[9] Conventional surface WTPs remove OM in coagulation−flocculation−sedimentation−filtration processes. In this process, coagulants, added to the water during the rapid mix process, react with particles and OM to form flocs, which settle out in a sedimentation basin, removing a part of the OM.[10,11]

WTPs that cannot meet either the distribution system-based DBP MCLs or the TOC removal requirement through conventional SW treatment must use advanced treatment options. Two examples of advanced treatment are granular activated carbon (GAC), to remove additional levels of OM

Received: February 5, 2016
Accepted: March 21, 2016
Published: March 21, 2016

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

and therefore reduce DBP formation, and chloramines as a secondary disinfectant. Compared to free chlorine, chloramines are less reactive with OM forming fewer regulated DBPs and are used in the distribution system to meet residual disinfectant requirements.[12,13]

Variations in SW TOC concentrations impact a WTP's ability to meet regulations and may influence decisions on implementing advanced treatment options, which have significant costs and operational complexities. Identifying OM sources is important in understanding variability in source water TOC concentrations. OM can enter SWs due to mobilization of leaf litter, crop residue and OM in soil,[14,15] and its transport is affected by soil and topographic conditions.[16,17] Soil leaching is the principle source of DOM in nonwastewater impacted SWs.[18] We refer the reader to Köhler et al.,[14] Worrall et al.,[19] and Christ and David,[20] for information about DOM production in soil. The amount and type of catchment area vegetation also influences TOC.[16,21]

Previous studies have investigated temporal and spatial changes in SW OC[14,15,18,21−35] and changes in terrestrial OC export to SW.[16,36] The SW TOC concentrations vary seasonally, especially in regions with snowmelt driven source,[36] including semiarid regions[37] and boreal regions,[14,22] where they peak during spring and early summer. Increased SW DOM has been observed during periods of elevated temperatures.[28,29,38] Temperature influences DOM dissolution and desorption and microbial activity, consequently, affecting OM production, decay, and mobilization.[39] During rainfall events increased SW OC concentrations have been observed.[31,35] Heavy rainfall increases surface runoff, therefore increasing TOC transport to SW.[38] However, droughts have also been found to increase SW DOC.[40] A short period of heavy rainfall followed by a long drought tends to increase SW DOC[32]—wherein rainfall leads to vegetation growth and the subsequent dry period leads to its demise and decay, that is then mobilized in the following rainy period. While a large quantity of annual DOM transport occurs during rainfall events, rainfall intensity and frequency can decrease this relationship.[39] Interplay between temperature and precipitation is important, as Köhler et al.[22] observed increased SW TOC during warm summers in wet years, but not in dry years.

Increasing SW OM concentrations in recent decades have been observed in North America and Europe,[18,19,26,28,29,37,41] indicating increasing DBP precursors in SW sources, potentially forcing utilities into expensive treatment options. Thus, a robust and simple modeling tool that can simulate and predict current and future TOC variability for a given source water is important for such decision-making. Understanding relationships between climate and source water TOC could allow for the use of climate change projections to predict future TOC concentrations, which can help project DBP formation[4,5,42] in the finished water and the distribution system.

Several SW OC models[15,18,24,25,27,33,34,36] have been developed. Recently, regression models have been developed for TOC using runoff and temperature as predictors,[14] and using Normalized Difference Vegetation Index (NDVI), runoff and fraction of area covered by bogs.[21] All of these models incorporated SW catchment hydrology, exploiting the strong relationships between OM and streamflow.[30,31] However, streamflow data are not widely available and many free-flowing streams are impacted by human activities such as diversions, dams, and reservoirs. Therefore, streamflow-based TOC models are not easily developed or applicable to model TOC under climate change where streamflow data are hard to generate.

Motivated by this need, the objective of this work was to develop a unique statistical methodology for predicting SW TOC concentrations that directly uses climate and land surface predictors, bypassing the need for streamflow. We demonstrated the utility of the methodology by applying it at three case study locations covering diverse climate regions, SW sources, and treatment processes. The models developed in this study select the most influential climate and land surface factors on SW TOC in each SW catchment to allow future predictions of TOC variability. This demonstrated methodology can be applied to other SW catchments to develop models with appropriate predictors. In this paper, the data sets and the development of climate and land surface predictors of TOC are first described followed by the methodology development. The Results and Discussion section describes the selected climate and land surface predictors and the model performance.

2. STUDY REGION DATA SETS AND DEVELOPMENT OF PREDICTOR VARIABLES

2.1. Case Studies. Three water utilities were used as case studies in this study: Greater Cincinnati Water Works in Cincinnati, Ohio, Newport News Waterworks in Newport News, Virginia, and City of Boulder Water Utilities Division in Boulder, Colorado. Monthly source water TOC concentrations for these facilities are shown in the Supporting Information (SI) Figure S1; additional source water quality, climate and land surface data for each case study location are summarized in Table S1.

The Harwood's Mill WTP in Newport News, Virginia treats SW from the Chickahominy River and five reservoirs, using coagulation, sedimentation, ozonation, and biofiltration. This plant supplies approximately 91% of the Newport News Waterworks drinking water. Approximately weekly source water TOC concentration data from January 2001 to March 2012 were used in this study.

The Greater Cincinnati Water Works' Richard Miller WTP treats Ohio River water and supplies approximately 88% of the customers. The Ohio River is an impacted source water with substantial anthropogenic influence. Streamflow in the Ohio River is controlled by various navigation dams. The Miller Plant uses conventional SW treatment followed by GAC as an advanced treatment. The source water TOC concentration data used are approximately monthly from January 1988 to December 2001 and approximately weekly from January 2002 to April 2007.

The City of Boulder's Betasso WTP is the primary drinking water facility serving Boulder residents and uses conventional SW treatment. It receives its source water from two reservoirs, Lakewood Reservoir and Barker Reservoir, located west of Boulder at an elevation of approximately 2500 m above sea level. Approximately weekly source water TOC concentration data from the Lakewood Reservoir (the primary source water) from January 1995 to April 2013 were used in this study. High OM peaks in source water occur in spring months, when snowmelt occurs, compared with relatively low concentrations during the rest of the year. Beggs et al.[37] show similar results from a nearby utility. For this analysis, April−July source water TOC concentrations were modeled, as these months have greater TOC concentration and variability than other months, causing concern for increased DBP formation during treatment.

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

2.2. Development of Predictor Variables. Predictor variables were developed using climate and land surface data sets. For each case study, predictor variable data sets include temperature, precipitation, NDVI, and Palmer Drought Severity Index (PDSI) data—NDVI captures vegetation,[21,43] a main source of OM, and PDSI captures soil moisture,[44] a property affecting mobilization of OM from the soil, respectively. For each study location, daily temperature and precipitation data were collected from the National Oceanic and Atmospheric Administration (NOAA) Daily Global Historical Climatology Network (GNCN-Daily) and supplemented by Weather Underground (http://www.wunderground.com/), monthly and 15-day averaged NDVI data for the grid nearest to the location of the source water were obtained from the International Research Institute (IRI) for Climate and Society Data Library (http://iridl.ldeo.columbia.edu/), and monthly average PDSI data for the nearest climate division were obtained from NOAA. For each location, a suite of predictor variables were developed using averages, or totals in the case of precipitation, to best represent the climate and land surface scenarios which would impact source water TOC concentration.

Predictor variables representing seasonal and short-term climatic influences were developed, including temperature averages ranging from 7-day averages to 30-day averages and precipitation totals ranging from 7-day totals to 30-day totals. To account for delays in the resulting effect on TOC transport to SWs after periods of high or low temperature, one-month temperature averages that occurred one to three months prior were also included in the predictor variable set. Relationships between raw lake water DOC and monthly temperature averages, ranging from 1-month to 36-month averages, and monthly precipitation totals, ranging from 1-month to 24-month totals, have been investigated previously.[41] Both rainfall intensity and frequency impacts DOM transport;[39] therefore, the frequency of precipitation events and the intervening dry spell were also included in the suite of predictor variables. PDSI is a widely used measure of the severity of drought in a region,[44] which is believed to influence OM production[14] and play a role in OM decomposition,[41] therefore affecting subsequent OM transport from soil to water. We refer readers to Evans et al.,[41] Kalbitz et al.,[45] and Dai et al.[46] for further information about the relationship between PDSI, soil moisture, and OM production and transport. NDVI is a measure of vegetation determined by the detection of reflected visible and near-infrared sunlight by the vegetation. It has shown consistent correlation with vegetation biomass and dynamics[47] and has been used previously in TOC predictive regression models.[21] Month-averages of PDSI and NDVI values were included in the predictor set.

Thus, a large suite of 18 climate and land surface predictors was computed (Table S2). The temperature variables are named "*TXD*", corresponding to the *X*-day average temperature prior to the TOC observation, were *X* is equal to 7, 15 and 30, and "*T30DYM*", corresponding to the 30-day average temperature prior to the TOC observation with a *Y*-month lag, where *Y* is 1, 2, and 3. Similarly, the precipitation totals are labeled as "*PXD*", corresponding to the *X*-day total precipitation prior to the TOC observation where *X* is 7, 15, and 30, and the precipitation frequencies include "*ddweek*" and "*ddmonth*", which are the number of dry days in the week and month prior to the TOC observation. The PDSI variables are named "*PDSIXM*" for the average PDSI *X* months prior to the TOC observation, and similarly, the NDVI variables are "*NDVIXM*" for the average NDVI *X* months prior. In both cases, *X* is equal to 1, 2, and 3. Finally, "*NDVI*" is the variable representing the average NDVI at the time of the TOC observation.

3. PROPOSED MODEL

3.1. Local Polynomial Regression. Preliminary analysis of the source water TOC concentrations at each case study location and the 18 predictor variables showed nonlinear relationships between TOC and predictors (Figure S3), suggesting that any modeling approach should have the ability to capture these nonlinear relationships. Therefore, we proposed to replace the linear aspect of the generalized linear model (GLM)[48] with a nonlinear functional estimation based on local polynomials.[49] The local polynomial regression model is in the form:

$$Y_i = \mu(x_i) + \varepsilon_i \quad (1)$$

In this, the function μ is estimated "locally" for any desired point x. A small set of neighbors ($K = \alpha N$; N is the total number of data points, and α is a value in the range of 0 to 1) of x are identified and a polynomial of order p is fitted via weighted least-squares method—wherein, the nearest neighbors are assigned highest weight and the farthest the least, using a bisquare or tricubic weight function.[49] The fitted polynomial is used to estimate the response variable Y at the desired point x. This process is repeated for all desired points of estimate. Note that if α and p are set to 1 and the neighbors assigned equal weights, then this reduces to the standard linear regression. The local estimation method provides an additional degree of flexibility to the GLM framework, making it GNLM (or Generalized NonLinear Model). The choice of α and p are obtained using a Generalized Cross Validation criteria (GCV), which penalizes higher order models and strives for parsimony. The GCV can be used to obtain the local polynomial parameters (α and p) and also the best set of predictors.[50] Local polynomial based GLMs have been widely used—for seasonal streamflow forecasting,[51−53] flood frequency estimation,[54] turbidity threshold exceedance modeling,[55,56] and for modeling attributes of stream temperature.[57]

The residuals from the above models are assumed to be uncorrelated; however, often the predictors are not efficient at capturing the autocorrelation present in the data. In the linear regression framework the residuals are modeled as another companion model at the second level of hierarchy. Both the models are fitted together, also known as regression with correlated errors.[58−60] Another approach is to include lagged values of the dependent variable along with the suite of predictors for auto regressive models with external variables[61,62]—this is possible if the data of the dependent variable is continuous in time, such as the case with TOC data at Harwood's Mill WTP. If not continuous, such as the case with Miller WTP, then we propose fitting a best model to the residuals separately using the same suite of predictors, except the predictors selected for the TOC model. Then the TOC model and the residual model are added together to create an additive model. In this research, we apply both these approaches for modeling residual correlation in TOC.

3.2. Model Validation. The best model is fitted for the TOC data at each of the three plants separately—which involves obtaining the best alpha (α), p, the best subset of predictors, and the appropriate link. The goodness of fit of the modeled source water TOC is visually inspected along with the

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

Table 1. Summary of the Best Local Polynomial Regression Models for each Case Study

case studies		predictor variables	link function	alpha (α)	p (degree)	gcv score	NSE statistic	hypothesis test p-value
Harwood's Mill WTP, Newport News, VA		−T30D3M −P7D −previous TOC concentration (lag 1)	log	0.97	2	0.004	0.92	9.79×10^{-06}
Miller WTP, Cincinnati, OH		**base model:** −T30D2M −PDSI1M	inverse	0.11	1	0.030	0.51 (for additive model)	2.77×10^{-08}
		residual model: −NDVI1M	identity (Gaussian family)	0.06	1	0.179		1.94×10^{-04}
Betasso WTP, Boulder, CO	April and May	−T15D −PDSI1M −PDSI3M	inverse	0.35	1	0.069	0.82	0.0367
	June and July	−T30D1M −P30D −PDSI1M	log	0.60	1	0.057	0.75	0.0576

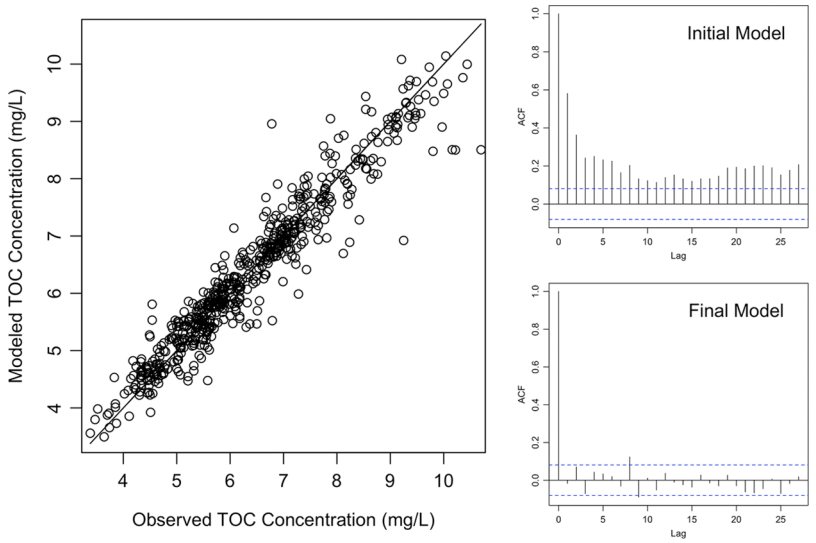


Figure 1. Final model (with lag 1 predictor): Scatterplot of modeled and observed source water TOC concentration for Harwood's Mill WTP in Newport News, Virginia, with a 1:1 line as reference (left); autocorrelation function for the initial model (without lag 1 predictor) residuals with 95% confidence intervals as dashed lines (upper right) and the autocorrelation function for the final model (with lag 1 predictor) residuals (lower right). The final model shown here has three predictor variables: T30D3M, P7D and the lag 1 (previous TOC concentration).

corresponding Nash-Sutcliffe Efficiency (NSE)[63] value. The NSE is an efficiency criteria used to assess model performance, similar to the well-recognized coefficient of determination R^2 statistic, and is defined by the following equation:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \quad (2)$$

where, in this study, Y_i are the observed TOC concentrations, $\hat{Y}_i$ are the predicted TOC concentrations, and $\bar{Y}$ is the mean of the observed TOC concentrations. A NSE value of 1 indicates a perfect fit for the model. Model diagnostics are performed; here the model residuals are checked for homoscedasticity, normal distribution and autocorrelation. Normality is tested by creating Q−Q plots and histograms of model residuals with a fitted normal distribution. Correlation is tested by plotting and analyzing the autocorrelation function for the residuals. A goodness of fit test is also conducted to compare that the local polynomial regression model is significantly better than a global linear regression model.[49]

Finally, the skill of the model is tested using a drop-10% cross-validation method. This is done by conducting 500 simulations in which 10% of the historic source water TOC concentrations are dropped, the model is fit to the remaining 90% of the data, and then the dropped points are predicted using the new model. The median root-mean-square error (RMSE) and the NSE value for comparing the predicted dropped points to the true dropped points for each simulation are reported. This is a robust approach for evaluating model performance and has been used in several water quality modeling studies.[52,64−66]

4. RESULTS AND DISCUSSION

4.1. Summary of Results. A summary of the regression models for the three case studies is presented in Table 1. This includes the best set of predictors, the Gamma family link

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

function, neighborhood size alpha (α), degree of polynomial (p), the NSE statistic of modeled TOC and observed TOC concentrations, and the p-value of the hypothesis test comparing the local polynomial regression model to the corresponding global linear model. Details of the residual model are also presented for the Miller WTP case study, where an additive model was developed to treat autocorrelation in the base model's residuals. The results from each case study are described in sequence below.

4.2. Case Study 1: Harwood's Mill Water Treatment Plant, Newport News, Virginia. The regression model selected for the Harwood's Mill source water TOC concentration used three predictor variables: the 30-day Average Temperature with a 3-month Lag (*T30D3M*), the 7-day Total Precipitation (*P7D*) and the Previous TOC Concentration (*lag 1*). A second order polynomial ($p = 2$) was selected with a neighborhood size ($\alpha = 0.97$) close to using all the observations, and a log link function was found to be optimal. This indicates that there are local nonlinearities (the reason for a second order polynomial) and the log link function captures high values well.

To demonstrate the utility of the lag 1 predictor, the initial model without this variable was evaluated. The observed TOC concentration and modeled values (without the lag 1 predictor variable) are plotted (Figure S4) and while the model estimates the observed values very well (NSE = 0.87), the residuals from the model are skewed (Figure S5) and exhibit significant autocorrelation (Figure 1). The presence of autocorrelation among the residuals is problematic as it allows one model prediction to affect the following model prediction, but can be corrected by including information about previous concentrations in the model.[67] Thus, the lag 1 predictor was incorporated in the final model.

The modeled versus observed TOC concentration ($n = 587$) for the final model with the lag 1 predictor is shown in Figure 1. This scatter is tighter along the 1:1 line compared to the previous model and has a higher NSE of 0.92. The significant reduction in autocorrelation in the residuals is illustrated in Figure 1, which compares the autocorrelation function for the initial model to that of the final model, where the autocorrelation in the residuals is virtually absent. The final model residuals are also normally distributed (Figure S5). The inclusion of the lag 1 predictor variable captures the variability of source water TOC very well and satisfies the assumptions of the residuals. To test the model performance in a blind forecasting mode, the RMSE and NSE from 500 simulations of drop-10% cross validation were calculated (Figure S6); the median RMSE and NSE are 0.19 and 0.91, respectively.

The results of this regression model indicate that temperature, precipitation, and the prior TOC concentrations provide significant information to model TOC variability at current time. While the temperature variable selected describes temperature 3 months prior to the TOC observation, the 7-day precipitation variable selected suggests that recent rainfall may play an important role in transporting OM to the Harwood's Mill WTP. It is interesting to note that PDSI variables were selected in the preliminary model, which did not incorporate a lag 1 predictor variable, suggesting that while soil moisture may still be an important physical factor affecting TOC transport from soil to SW, much of this relationship may be captured in previous TOC concentrations, therefore eliminating PDSI as a predictor.

4.3. Case Study 2: Miller Water Treatment Plant, Cincinnati, Ohio. The regression model selected for the Miller Plant source water TOC concentration is an additive model, in which two predictor variables, the 30-day Average Temperature with a 2-month Lag (*T30D2M*) and the PDSI 1 Month Prior (*PDSI1M*), are used to model the source water TOC concentration, creating the base model. The residuals from this base model are then modeled using another local polynomial regression model with the NDVI 1 Month Prior (*NDVI1M*) as the sole predictor variable. Both models are added together to create the additive regression model. The best polynomial order selected was one for both models, and the neighborhood size selected was 0.11 and 0.06 for the base and residual models, respectively. The smaller neighborhood size indicates substantial local nonlinearities, which can be captured using a local linear model. The data for this source water were not continuous, unlike the previous case study; thus, the lag 1 predictor cannot be incorporated meaningfully. Hence, it was necessary to model the residuals separately and create an additive model. Additive models have been introduced in various forms, such as the class of generalized additive models[68] and nonparametric regression additive models.[69,70]

Plotting the observed and modeled TOC concentrations ($n = 407$) from the base model shows the model performance is poor (NSE = 0.39) with overestimation of lower values and significant underestimation of high values (Figure S7). In addition, model diagnostics illustrate that the residuals show significant autocorrelation and skew (Figure S7). To address the structure in the residuals, a second local polynomial regression model was created in which best predictor variables and the parameters (α and p) were selected. Since the residuals are unbounded a normal distribution assumption is appropriate. The best predictor selected was the NDVI 1 Month Prior (*NDVI1M*) with a polynomial of order one (i.e., local linear) and a neighborhood size of 0.06—indicating local nonlinearities. The modeled TOC concentration from the additive "full" model (estimates from base model + estimates from the residual model) plotted against the observed concentrations demonstrates a better fit (with NSE = 0.51) than that from just the base model (Figure S8). Furthermore, the residuals show decreased autocorrelation with a slight autocorrelation at lag 1, and the residuals are close to normal distribution (Figure S8). The median RMSE and median NSE from the drop-10% cross validation are 0.21 and 0.27, respectively (Figure S9).

Comparing the results of the Miller WTP regression model to those of the Harwood's Mill WTP regression model, it is apparent that the climate and land surface variables do not capture as much of the source water TOC variability, suggesting that other significant factors need to be considered. The Ohio River, the source water for the Miller WTP, drains catchment areas for which almost half of the area is agricultural or urban,[71] and there are 20 navigational dams along the river affecting its flow. In addition, wastewater discharges enter the waterway at many points with the closest just 17.5 km upstream of the Miller WTP's water intake source. Municipal wastewater discharge and agricultural and urban runoff can be important contributions to source water OM. The predictor variables selected do suggest that the temperature a few months prior is an important variable, as was in the Harwood's Mill WTP regression model. Recent soil moisture, represented by PDSI, also proved to be an important predictor variable, as this indicates terrestrial primary production, microbial decomposi-

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

tion of OM, as well as OM transport to the Ohio River. Lastly, recent vegetation coverage, represented by NDVI, was selected as a predictor variable for the model, which suggests that recent terrestrial primary production is an important source of OM for this river.

4.4. Case Study 3: Betasso Water Treatment Plant, Boulder, Colorado. Dramatic seasonal variation in source water TOC concentration, specifically in spring and early summer months, presents a challenge for the Betasso WTP in meeting DBP regulations. The physical processes influencing the transport of OM to SW in the months of April and May is prominently driven by snowmelt, while precipitation plays a greater role in June and July. Preliminary analysis also suggested important climate predictors for TOC differed in April and May versus June and July, with different relationships observed between TOC and temperature variables, in particular (Figure S2). Hence, two separate models are created for TOC concentrations in Apr−May and Jun−Jul. TOC concentrations were typically very low during the remainder of the year, and generally do not present any threat of DBP concentrations exceeding their regulatory MCLs.

For both the Apr−May and Jun−Jul model, all of the NDVI predictor variables were removed because the NDVI data set only included data until 2006, and its incorporation would drastically reduce the size of the TOC data set. From the preliminary analysis, the incorporation of NDVI would improve the regression models' abilities to estimate source water TOC, and further analysis may produce improved regression models.

4.4.1. April and May. The best model selected consisted of three predictor variables: the 15-day Average Temperature (*T15D*), the PDSI 1 Month Prior (*PDSI1M*) and the PDSI 3 Months Prior (*PDSI3M*). The inverse link function with a first order polynomial and neighborhood size of 0.35 indicates local nonlinearities present in the relationship. The scatterplot of modeled and observed TOC concentrations (n = 78) is shown in Figure 2 and the corresponding NSE is 0.83. The residuals exhibit no significant autocorrelation and are normally distributed (Figure S10), indicating that the model captured almost all of the variability in TOC with white noise residuals, in compliance with the theoretical framework. The performance of the model in a prediction mode from the drop-10% cross validation has median RMSE of 0.72 and median NSE of 0.62 (Figure S11).

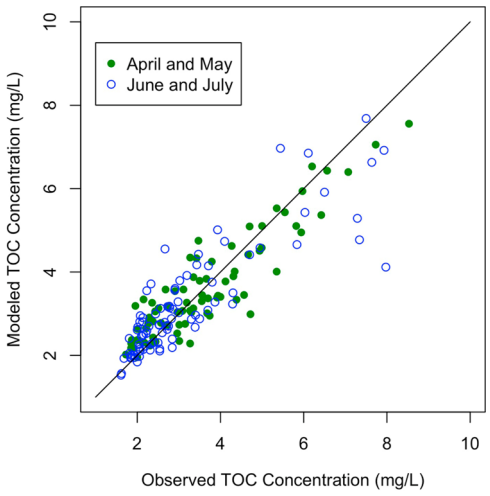


Figure 2. Scatterplot of modeled and observed source water TOC concentration for the Betasso WTP in Boulder, Colorado in Apr−May and Jun−Jul, with a 1:1 line as reference (left).

Since snowmelt is the main driver of TOC during this period, the 15-day temperature average is an appropriate predictor, as it controls the quantity of snowmelt thus influencing the surface runoff transporting OM to the source water reservoirs. If enough melt occurs, then water can travel to the base of the snow cover and to surface of the underlying soil, where it is either available for runoff and/or infiltration.[72] Soil moisture influences the soil's capacity for infiltration and therefore, also influences the quantity of runoff−with higher soil moisture before snowmelt period reduces infiltration of the snowmelt and increases runoff. Therefore, the model's selection of PDSI, both one month prior and three months prior, suggests the importance of soil moisture in these physical processes.

4.4.2. June and July. The best model for June and July TOC concentration consisted of three predictor variables: the 30-day Average Temperature with a 1-Month Lag (*T30D1M*), the 30-day Total Precipitation (*P30D*), and the PDSI 1 Month Prior (*PDSI1M*)—with log link function, local linear polynomial and neighborhood of 0.6, indicating local nonlinearities. The modeled and observed TOC concentrations (n = 89) are shown in Figure 2 and the corresponding NSE of 0.75 is very good as the modeled values are close to the observed with an underestimation of higher values. The residuals show no significant autocorrelation and exhibit normal distribution (Figure S12) indicating that the model is quite effective at capturing almost all of the variability in the TOC concentrations. The drop-10% cross validation has a median RMSE of 0.60 and median NSE of 0.61 (Figure S13), both indicating very good model performance in a true forecasting mode.

In the early summer months of June and July, some snowmelt may still occur depending on the temperatures during the preceding spring months. The temperature predictor variable selected by the model captures the temperatures during spring months, which directly influences the quantity of snowmelt. Rainfall becomes the main physical process leading to surface runoff during these months, which is captured in the selected precipitation variable. As in the April and May regression model, PDSI is an important predictor variable, because soil moisture influences the soil's capacity for infiltration, which in turn affects the surface runoff transporting OM to the reservoirs supplying water to the Betasso WTP.

Comparing results for all case study locations allows for identification of common predictors. A temperature variable was selected as a predictor in every model, although the time period represented by the variables ranges from a 15-day temperature average (T15D) to a 30-day average temperature 3 months prior to the TOC observation (T30D3M), suggesting the time scale of temperature impacting SW TOC concentrations may be based on the physical processes leading to TOC transport to SW. PDSI one month prior to the TOC observation (PDSI1M) was selected as a predictor for both Boulder models and the Cincinnati model, and was initially selected for the Newport News model before the lag 1 predictor was incorporated, illustrating the influence of soil moisture on SW TOC concentrations in three different geographic locations and suggesting that recent soil moisture, measured as PDSI, may be an important predictor of TOC concentrations in other watersheds. The three case studies demonstrate that the climate and land surface variables with the strongest impact on the mobilization and transport of TOC to SW can vary in different watersheds. This methodology can be

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

applied in other SW catchments to determine appropriate model parameters, which may not be the specific parameters selected in the three case studies presented.

The three case study locations illustrate the varying degrees at which a local polynomial regression model with climate and land surface predictors can model SW TOC variability avoiding reliance on streamflow data used in previous models.[14,15,18,21,24,25,27,33,34,36] The three case study SW sources do not have long residence times; this modeling approach may also have limited results if applied to reservoir sources with long residence times. The Cincinnati case study demonstrates the limitations to this modeling approach when substantial anthropogenic impacts on SW OM are present. The predictor suite utilized does not account for any OM sources from wastewater discharge or urban and/or agricultural runoff, and therefore this modeling technique will have limited predictive skill if applied to watersheds with these OM sources. The results for the Newport News model and the Boulder models suggest that this modeling approach offers good predictive skill for watersheds dominated by OM with little anthropogenic sources; applying it to similar watersheds should allow for predictions on future TOC variability using climate prediction models. The ability to predict future TOC variability may become increasingly important as potential climate change scenarios threaten to increase SW TOC, increasing the potential for DBP formation and the challenge for water utilities to meet regulations and protect the public that they serve.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.6b00639.

Detailed information about source water quality and climate data for each case study, the predictor variable suite, relationships between TOC and predictors, traditional linear regression models and generalized linear models, results for the case study models, model diagnostics, and drop-10% validation method results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*carleigh.samson@colorado.edu (C.C.S.).

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

For providing the source water quality data, we thank the staff at: the City of Boulder's Water Quality and Environmental Services team at the Department of Public Works, the Newport News Waterworks, and the Greater Cincinnati Water Works. We also thank the three anonymous reviewers for their insightful comments that significantly helped improve the manuscript. This publication was developed under partial support for the second and third authors from the Assistance Agreement No. RD 83586501-0 awarded by the U.S. Environmental Protection Agency. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and EPA does not endorse any products or commercial services mentioned in this publication.

■ REFERENCES

(1) Liang, L.; Singer, P. C. Factors Influencing the Formation and Relative Distribution of Haloacetic Acids and Trihalomethanes in Drinking Water. *Environ. Sci. Technol.* **2003**, *37* (13), 2920−2928.

(2) US Environmental Protection Agency. *National Primary Drinking Water Regulations: Stage 2 Disinfectants and Disinfection Byproducts Rule* **2006**, *71*, 388−493.

(3) Pifer, A. D.; Fairey, J. L. Suitability of Organic Matter Surrogates to Predict Trihalomethane Formation in Drinking Water Sources. *Environ. Eng. Sci.* **2014**, *31* (3), 117−126.

(4) Obolensky, A.; Singer, P. C. Development and Interpretation of Disinfection Byproduct Formation Models Using the Information Collection Rule Database. *Environ. Sci. Technol.* **2008**, *42* (15), 5654−5660.

(5) Solarik, G.; Summers, R. S.; Sohn, J.; Swanson, W. J.; Chowdhury, Z. K.; Amy, G. L. Extensions and Verifications of the Water Treatment Plant Model for Distribution By-Product Formation. In *Natural Organic Matter and Disinfection By-Products Characterization and Control in Drinking Water*; Barrett, S. E., Krasner, S. W., Amy, G. L., Eds.; Washington, DC, 2000; pp 47−66.

(6) Summers, R. S.; Hooper, S. M.; Shukairy, H. M.; Solarik, G.; Owen, D. Assessing DBP yield: uniform formation conditions. *J. Am. Water Works Assn.* **1996**, *88* (6), 80−93.

(7) Amy, G. L.; Chadik, P. A.; Chowdhury, Z. K. Developing Models for Predicting Trihalomethane Formation Potential and Kinetics. *J. Am. Water Works Assn.* **1987**, *79* (7), 89−97.

(8) Clark, R. M.; Adams, J. Q.; Lykins, B. W., Jr. DBP Control in Drinking Water: Cost and Performance. *J. Environ. Eng.* **1994**, *120* (4), 759−782.

(9) US Environmental Protection Agency. *National Primary Drinking Water Regulations: Disinfectants and Disinfection Byproducts* **1998**, *63*, 69390−69476.

(10) Letterman, R. D.; Yiacoumi, S. Coagulation and Flocculation. In *Water Quality and Treatment: A Handbook on Drinking Water*; Edzwald, J. K., Ed.; New York, **2011**; pp 8-1−8-6.

(11) Gregory, R.; Edzwald, J. K. Sedimentation and Flotation. In *Water Quality and Treatment: A Handbook on Drinking Water*; McGraw-Hill Companies, Inc.: New York, 2011; pp 9.1−9.14.

(12) Bougeard, C. M. M.; Goslan, E. H.; Jefferson, B.; Parsons, S. A. Comparison of the disinfection by-product formation potential of treated waters exposed to chlorine and monochloramine. *Water Res.* **2010**, *44* (3), 729−740.

(13) Hua, G.; Reckhow, D. A. Comparison of disinfection byproduct formation from chlorine and alternative disinfectants. *Water Res.* **2007**, *41* (8), 1667−1678.

(14) Köhler, S. J.; Buffam, I.; Seibert, J.; Bishop, K. H.; Laudon, H. Dynamics of stream water TOC concentrations in a boreal headwater catchment: Controlling factors and implications for climate scenarios. *J. Hydrol.* **2009**, *373* (1−2), 44−56.

(15) Futter, M. N.; Butterfield, D.; Cosby, B. J.; Dillon, P. J.; Wade, A. J.; Whitehead, P. G. Modeling the mechanisms that control in-stream dissolved organic carbon dynamics in upland and forested catchments. *Water Resour. Res.* **2007**, *43* (2), W02424.

(16) Ågren, A.; Buffam, I.; Jansson, M.; Laudon, H. Importance of seasonality and small streams for the landscape regulation of dissolved organic carbon export. *J. Geophys. Res.* **2007**, *112* (G3), G03003.

(17) Dosskey, M. G.; Bertsch, P. M. Forest sources and pathways of organic matter transport to a blackwater stream: a hydrologic approach. *Biogeochemistry* **1994**, *24* (1), 1−19.

(18) Hejzlar, J.; Dubrovský, M.; Buchtele, J.; Růzicka, M. The apparent and potential effects of climate change on the inferred concentration of dissolved organic matter in a temperate stream (the Malše River, South Bohemia). *Sci. Total Environ.* **2003**, *310* (1−3), 143−152.

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

(19) Worrall, F.; Burt, T.; Adamson, J. Can climate change explain increases in DOC flux from upland peat catchments? *Sci. Total Environ.* **2004**, *326* (1−3), 95−112.
(20) Christ, M. J.; David, M. B. Temperature and moisture effects on the production of dissolved organic carbon in a Spodosol. *Soil Biol. Biochem.* **1996**, *28* (9), 1191−1199.
(21) Larsen, S.; Andersen, T.; Hessen, D. O. Climate change predicted to cause severe increase of organic carbon in lakes. *Global Change Biol.* **2011**, *17* (2), 1186−1192.
(22) Köhler, S. J.; Buffam, I.; Laudon, H.; Bishop, K. H. Climate's control of intra-annual and interannual variability of total organic carbon concentration and flux in two contrasting boreal landscape elements. *J. Geophys. Res.* **2008**, *113* (G3), G03012−G03012.
(23) Worrall, F.; Burt, T. P. The effect of severe drought on the dissolved organic carbon (DOC) concentration and flux from British rivers. *J. Hydrol.* **2008**, *361* (3−4), 262−274.
(24) Erlandsson, M.; Buffam, I.; Fölster, J.; Laudon, H.; Temnerud, J.; Weyhenmeyer, G. A.; Bishop, K. Thirty-five years of synchrony in the organic matter concentrations of Swedish rivers explained by variation in flow and sulphate. *Global Change Biol.* **2008**, *14* (5), 1191−1198.
(25) Yurova, A.; Sirin, A.; Buffam, I.; Bishop, K.; Laudon, H. Modeling the dissolved organic carbon output from a boreal mire using the convection-dispersion equation: Importance of representing sorption. *Water Resour. Res.* **2008**, *44* (7), W07411.
(26) Monteith, D. T.; Stoddard, J. L.; Evans, C. D.; de Wit, H. A.; Forsius, M.; Høgåsen, T.; Wilander, A.; Skjelkvåle, B. L.; Jeffries, D. S.; Vuorenmaa, J.; et al. Dissolved organic carbon trends resulting from changes in atmospheric deposition chemistry. *Nature* **2007**, *450*, 537−540.
(27) Canham, C. D.; Pace, M. L.; Papaik, M. J.; Primack, A. G. B.; Roy, K. M.; Maranger, R. J.; Curran, R. P.; Spada, D. M. A Spatially Explicit Watershed-Scale Analysis of Dissolved Organic Carbon in Adirondack Lakes. *Ecol. Appl.* **2004**, *14* (3), 839−854.
(28) Worrall, F.; Burt, T.; Shedden, R. Long term records of riverine dissolved organic matter. *Biogeochemistry* **2003**, *64* (2), 165−178.
(29) Freeman, C.; Evans, C. D.; Monteith, D. T. Export of organic carbon from peat soils. *Nature* **2001**, *412*, 6849−6785.
(30) Kendall, K. A.; Shanley, J. B.; McDonnell, J. J. A hydrometric and geochemical approach to test the transmissivity feedback hypothesis during snowmelt. *J. Hydrol.* **1999**, *219* (3−4), 188−205.
(31) Hinton, M. J.; Schiff, S. L.; English, M. C. The significance of storms for the concentration and export of dissolved organic carbon from two Precambrian Shield catchments. *Biogeochemistry* **1997**, *36*, 67−88.
(32) Schindler, D. W.; Curtis, P. J.; Bayley, S. E.; Parker, B. R.; Beaty, K. G.; Stainton, M. P. Climate-induced changes in the dissolved organic carbon budgets of boreal lakes. *Biogeochemistry* **1997**, *36* (1), 9−28.
(33) Boyer, E. W.; Hornberger, G. M.; Bencala, K. E.; McKnight, D. Overview of a simple model describing variation of dissolved organic carbon in an upland catchment. *Ecol. Modell.* **1996**, *86* (2−3), 183−188.
(34) Tipping, E. CHUM: a hydrochemical model for upland catchments. *J. Hydrol.* **1996**, *174*, 305−330.
(35) Meyer, J. L.; Tate, C. M. The Effects of Watershed Disturbance on Dissolved Organic Carbon Dynamics of a Stream. *Ecology* **1983**, *64* (1), 33−44.
(36) Boyer, E. W.; Hornberger, G. M.; Bencala, K. E.; McKnight, D. M. Effects of asynchronous snowmelt on flushing of dissolved organic carbon: a mixing model approach. *Hydrol. Processes* **2000**, *14*, 3291−3308.
(37) Beggs, K. M. H.; Billica, J. A.; Korak, J. A.; Rosario-Ortiz, F. L.; McKnight, D. M.; Summers, R. S. Spectral evaluation of watershed DOM and DBP precursors. *J. Am. Water Works Assn.* **2013**, *105*, E173−E188.
(38) Delpla, I.; Jung, A.-V.; Baures, E.; Clement, M.; Thomas, O. Impacts of climate change on surface water quality in relation to drinking water production. *Environ. Int.* **2009**, *35* (8), 1225−1233.
(39) Xu, N.; Saiers, J. E. Temperature and Hydrologic Controls on Dissolved Organic Matter Mobilization and Transport within a Forest Topsoil. *Environ. Sci. Technol.* **2010**, *44* (14), 5423−5429.
(40) Hope, D.; Billett, M. F.; Cresser, M. S. A review of the export of carbon in river water: Fluxes and processes. *Environ. Pollut.* **1994**, *84* (3), 301−324.
(41) Evans, C. D.; Monteith, D. T.; Cooper, D. M. Long-term increases in surface water dissolved organic carbon: Observations, possible causes and environmental impacts. *Environ. Pollut.* **2005**, *137* (1), 55−71.
(42) Sadiq, R.; Rodriguez, M. J. Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. *Sci. Total Environ.* **2004**, *321* (1−3), 21−46.
(43) Tucker, C. J.; Sellers, P. J. Satellite remote sensing of primary production. *Int. J. Remote Sens.* **1986**, *7* (11), 1395−1416.
(44) Alley, W. M. The Palmer Drought Severity Index: Limitations and Assumptions. *J. Clim. Appl. Meteorol.* **1984**, *23* (7), 1100−1109.
(45) Kalbitz, K.; Solinger, S.; Park, J. H.; Michalzik, B.; Matzner, E. Controls on the dynamics of organic matter in soils: a review. *Soil Sci.* **2000**, *165*, 277−304.
(46) Dai, A.; Trenberth, K. E.; Qian, T. A Global Dataset of Palmer Drought Severity Index for 1870−2002: Relationship with Soil Moisture and Effects of Surface Warming. *J. Hydrometeorol.* **2004**, *5* (6), 1117−1130.
(47) Pettorelli, N.; Vik, J. O.; Mysterud, A.; Gaillard, J.-M.; Tucker, C. J.; Stenseth, N. C. Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends Ecol. Evol.* **2005**, *20* (9), 503−510.
(48) McCullagh, P.; Nelder, J. A. *Generalized Linear Models*; Chapman & Hall: New York, 1989.
(49) Loader, C. *Local Regression and Likelihood*; Springer: New York, 1999.
(50) Regonda, S.; Rajagopalan, B.; Lall, U.; Clark, M.; Moon, Y. I. Local polynomial method for ensemble forecast of time series. *Nonlin. Processes Geophys.* **2005**, *12* (3), 397−406.
(51) 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* (9), W09404.
(52) Rajagopalan, B.; Bracken, C.; Prairie, J. A multisite seasonal ensemble streamflow forecasting technique. *Water Resour. Res.* **2010**, *46* (3), W03532.
(53) 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* (10), W10410.
(54) Apipattanavis, S.; Rajagopalan, B.; Lall, U. Local Polynomial−Based Flood Frequency Estimator for Mixed Population. *J. Hydrol. Eng.* **2010**, *15* (9), 680−691.
(55) Towler, E.; Rajagopalan, B.; Summers, R. S.; Yates, D. An approach for probabilistic forecasting of seasonal turbidity threshold exceedance. *Water Resour. Res.* **2010**, *46* (6), W06511.
(56) Towler, E.; Rajagopalan, B.; Gilleland, E.; Summers, R. S.; Yates, D.; Katz, R. W. Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. *Water Resour. Res.* **2010**, *46* (11), W11504.
(57) Caldwell, J.; Rajagopalan, B.; Danner, E. Statistical Modeling of Daily Water Temperature Attributes on the Sacramento River. *J. Hydrol. Eng.* **2015**, *20* (5), 04014065.
(58) Opsomer, J.; Wang, Y.; Yang, Y. Nonparametric Regression with Correlated Errors. *Stat. Sci.* **2001**, *16* (2), 134−153.
(59) Smith, M.; Wong, C.-M.; Kohn, R. Additive nonparametric regression with autocorrelated errors. *J. R. Stat. Soc. B* **1998**, *60* (2), 311−331.
(60) Stone, C. J. Additive Regression and Other Nonparametric Models. *Ann. Stat.* **1985**, *13* (2), 689−705.
(61) Agiakloglou, C. Resolving spurious regressions and serially correlated errors. *Empir. Econ.* **2013**, *45* (3), 1361−1366.

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421

(62) Granger, C. W. J.; Newbold, P. Spurious regressions in econometrics. *J. Econometrics* **1974**, *2* (2), 111−120.
(63) Nash, J. E.; Sutcliffe, J. V. River flow forecasting through conceptual models part I — A discussion of principles. *J. Hydrol.* **1970**, *10* (3), 282−290.
(64) Towler, E.; Rajagopalan, B.; Summers, R. S. Using Parametric and Nonparametric Methods to Model Total Organic Carbon, Alkalinity, and pH after Conventional Surface Water Treatment. *Environ. Eng. Sci.* **2009**, *26* (8), 1299−1308.
(65) Zachman, B. A.; Rajagopalan, B.; Summers, R. S. Modeling NOM Breakthrough in GAC Adsorbers Using Nonparametric Regression Techniques. *Environ. Eng. Sci.* **2007**, *24* (9), 1280−1296.
(66) Alberto, W. D.; del Pilar, D. M.; María del Pilar, D.; María Valeria, A.; Valeria, A. M.; Fabiana, P. S.; Cecilia, H. A.; María de los Ángeles, B.; de los Ángeles, B. M. Pattern Recognition Techniques for the Evaluation of Spatial and Temporal Variations in Water Quality. A Case Study. *Water Res.* **2001**, *35* (12), 2881−2894.
(67) Gardner, M. W.; Dorling, S. R. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmos. Environ.* **1999**, *33* (5), 709−719.
(68) Hastie, T.; Tibshirani, R. Generalized Additive Models. *Stat. Sci.* **1986**, *1* (3), 297−310.
(69) Linton, O. B.; Härdle, W. Estimation of additive regression models with known links. *Biometrika* **1996**, *83* (3), 529−540.
(70) Buja, A.; Hastie, T.; Tibshirani, R. Linear Smoothers and Additive Models. *Ann. Stat.* **1989**, *17* (2), 453−510.
(71) Beaulieu, J. J.; Shuster, W. D.; Rebholz, J. A. Nitrous Oxide Emissions from a Large, Impounded River: The Ohio River. *Environ. Sci. Technol.* **2010**, *44* (19), 7527−7533.
(72) Colbeck, S. C. A simulation of the enrichment of atmospheric pollutants in snow cover runoff. *Water Resour. Res.* **1981**, *17* (5), 1383−1388.

DOI: 10.1021/acs.est.6b00639
Environ. Sci. Technol. 2016, 50, 4413−4421