Environmental Science and Pollution Research (2020) 27:24594–24603
https://doi.org/10.1007/s11356-019-05490-7

RECENT ADVANCES AND NOVEL CONCEPTS IN ENVIRONMENTAL TECHNOLOGIES

Models for predicting carbonaceous disinfection by-products formation in drinking water treatment plants: a case study of South Korea

Nirmal Kumar Shahi[1] · Minsoo Maeng[1] · Seok Dockko[1]

Received: 29 March 2019 / Accepted: 15 May 2019 / Published online: 26 June 2019

Abstract

Chlorination in a drinking water treatment plant is the critical process for controlling harmful pathogens. However, the reaction of chlorine with organic matter forms undesirable, harmful, and halogenated disinfection by-products. Carbonaceous disinfection by-products, such as trihalomethanes (THMs) and haloacetic acids (HAAs), are genotoxic or carcinogenic and are reported at high concentration in drinking water. This study is aimed at developing a mathematical model for predicting concentration levels of THMs and HAAs in drinking water treatment plants in South Korea because no previous attempts to do so have been reported for the country. The THMs concentration levels ranged from 29 to 39 μg/L, and those for the HAAs from 6 to 7 μg/L. Multiple regression models, i.e., both linear and nonlinear, for THMs and HAAs were developed to predict their concentration levels in water treatment plants using datasets (January 2015 to December 2016) from three treatment plants located in Seoul, South Korea. The constructed models incorporated principal factors and interactive and higher-order variables. The principal factor variables used were dissolved organic carbon, ultraviolet absorbance, residual chlorine, bromide, contact time, chlorine dose and temperature for treated water, and pH for both raw and treated water at the plant. The linear models for both THMs and HAAs were found to give acceptable fits with measured values from the water treatment plants and predictability values were found to be 0.915 and 0.772, respectively. The models developed were validated with a later dataset (January 2017 to July 2017) from the same water treatment plants. In addition, the models were applied to two different water treatment plants. Application and validation results of the constructed model showed no significant differences between predicted and observed values.

Keywords Multiple regression analysis · Carbonaceous disinfection by-products (C-DBPs) · Trihalomethanes (THMs) · Haloacetic acids (HAAs) · Interactive variable · Validation

Introduction

The disinfection process is one of the crucial steps in drinking water treatment plants (DWTPs) to reduce waterborne diseases by inactivating harmful pathogens and microorganisms. Chlorine is widely used as a disinfectant in South Korea and elsewhere because it is highly efficient in preventing pathogens and microorganisms and cost-effectiveness(Abdullah et al. 2003; Uyak et al. 2005). However, chlorine reacts with natural organic matter (NOM) present in source water and forms undesirable carbonaceous and nitrogenous disinfection by-products (C- and N-DBPs) (Sérodes et al. 2003; Chowdhury et al. 2010; Maeng et al. 2018). Epidemiological studies conducted repeatedly in laboratory animals have shown that ingestion of chlorinated by-products containing water causes bladder, colon, and rectal cancer (Morris et al. 1992; WHO 2005; Brown et al. 2011). In addition, toxicological studies have shown that ingestion of certain DBPs causes cancer in the liver and kidneys, as well as adverse reproductive and developmental disorders in laboratory animals (Babaei et al. 2015; Krasner et al. 2017). Among all the DBPs formed, C-DBPs, i.e., trihalomethanes (THMs) and haloacetic acids (HAAs) are found in higher concentration levels in DWTPs (Richardson et al. 2007; Hua et al. 2015) and are considered potentially carcinogenic (Golfinopoulos et al. 1998;

Responsible editor: Marcus Schulz

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11356-019-05490-7) contains supplementary material, which is available to authorized users.

✉ Seok Dockko
dockko@dankook.ac.kr

[1] Department of Civil and Environmental Engineering, Dankook University, Yongin-si, Gyeonggi-do 448-701, Republic of Korea

Golfinopoulos and Arhonditsis 2002; Uyak et al. 2005; Platikanov et al. 2012). Moreover, bromine-containing species are reported to be more geno- and cytotoxic than their chlorinated form and are of concern (Krasner et al. 2017). These DBPs can enter the human body through ingestion of drinking water, inhalation, and dermal contact during regular indoor activities such as showering, bathing, swimming, and cooking (Chowdhury et al. 2010; Chowdhury et al. 2011; Domínguez-Tello et al. 2017). Thus, several DBPs are regulated by international regulatory agencies worldwide. The US Environmental Protection Agency developed a Disinfectants/DBP (D/DBP) rule in 1998 and set minimum contaminant levels of 80 μg/L for THMs (Uyak et al. 2005; Singh and Gupta 2012) and 60 μg/L for HAAs (Ged et al. 2015). The formation of THMs and HAAs depends on the quality of the source water and the treatment process, i.e., chlorine dose, contact time between chlorine and organic matters, pH, water temperature, and other factors (Sadiq and Rodriguez 2004; Fooladvand et al. 2011). Continuous monitoring throughout the operation of DWTP is needed to ensure compliance with the guidelines. More than 100 predictive models have been developed because of active research on DBPs (Chowdhury et al. 2009; Domínguez-Tello et al. 2017). These models are based on either laboratory or field scale data and have shown varying levels of predictive capabilities. Most of the mathematical models developed are empirical in nature and are site specific, which means their predictive capabilities for different environmental conditions and treatment processes remain inappropriate (Elshorbagy et al. 2000; Uyak et al. 2007; Ata et al. 2015). On the other hand, seasonal, locational, and temporal factors, and the complexity of the reaction between organic matter and chlorine and the formation of DBPs makes it very difficult to develop mechanistic models (Semerjian et al. 2009; Kulkarni and Chellam 2010). Most of the models lack interacting parameters (Sohn et al. 2004). Mathematical models that are developed and based on real DWTPs and distribution systems, and which consider all the water quality parameters and operating variables that can predict THMs and HAAs, are very useful tools as alternatives to field measurements. Laboratory tests for the measurement of DBPs are very time consuming and expensive. Predictive models can provide quick and reasonable estimates and can help in making decisions to optimize the treatment process (Westerhoff et al. 2000; Mukundan and Van Derson 2014; Lin et al. 2018).

The aim of this study was to develop a mathematical model for predicting THMs and HAAs, based on multiple regression analysis and using water quality parameters of both raw and treated water and operational conditions from three DWTPs located in Seoul, South Korea. Models were validated using more recent data from the three treatment plants and were applied to two different DWTPs for evaluating their predictability capability. Most of the mathematical models that were reported previously lack principal factors such as dissolved organic carbon (DOC), bromide ion (Br^-), and chlorine dose (Chowdhury et al. 2009; Bond et al. 2014). Besides these, most of the models do not consider interactive variables (effect of two or more varying together). The THM models suggested by Amy et al. (1987), Golfinopoulos et al. (1998) and Uyak et al. (2005) have good predictability ($R^2 = 0.90$, 0.98, and 0.98, respectively) and are based on raw water characteristics. Raw water characteristics do not represent treated water characteristics. In addition, the models do not consider reaction time and chlorine dose. The objective of this study was to develop models that would address the shortcomings that existed in previously published research works. In addition, this work is the first of its kind to develop mathematical models using multiple regression analysis for both THMs and HAAs in South Korea. The model developed in this research can be considered robust because both raw water and treated water characteristics, along with most of the water quality and operational parameters with interactive parameters, are incorporated.

Materials and methods

Description of DWTPs

Seoul has a population of 10.178 million and is served by six DWTPs (SMG 2017). The Han River, which is the second longest river, serves as the main source of raw water to all these DWTPs. More than 3 million m^3 of water is needed daily for citizens residing in Seoul from all six DWTPs (Fig. 1). Each day, a total of 4.44 million m^3 of water is processed and supplied by all DWTPs, as shown in Table S1. The Seoul Metropolitan Government (SMG) monitors water quality and operation parameters every day to ensure compliance with the National Drinking Water Standard Guideline for the safety of the citizens. The treatment process combines conventional processes, i.e., prechlorination, coagulation, flocculation, sedimentation, filtration, and postchlorination, with advanced treatment processes, i.e., ozonation and powdered activated carbon treatment (Fig. S1).

Mathematical model development

For the purpose of the model development, water quality and operational data for both raw water and treated water were collected from 2015 to 2016 for three DWTPs (Gangbuk, Gwangam, and Yeongdeungpo). Data obtained were based on monthly analyses of water samples. These datasets included 120 and 66 measured values for THMs and HAAs, respectively, along with other water quality and operational parameters. Water quality parameters include DOC, ultraviolet absorbance (UV_{254}), Br^- ion concentration, temperature and residual chlorine, THMs, and HAAs for treated water. Likewise, operational parameters such as pH and prechlorine dose for

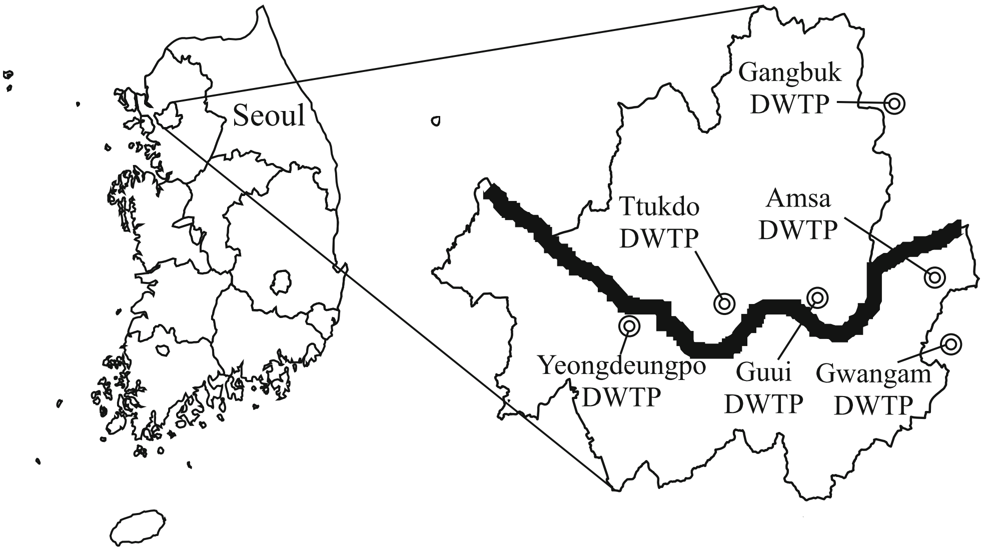

Fig. 1 Study area with all five drinking water treatment plants considered

raw water, postchlorine dose, pH, temperature, and contact time were included. The models, which included at least five principal factors, i.e., predictor variables of the seven (DOC, UV_{254}, Br^-, chlorine dose, temperature, contact time, and pH) showed high predictability for THMs and HAAs (Ged et al. 2015). In addition, the effect of two or more variables (interactive) and higher-order variables needed to be incorporated. Multiple regression analysis was carried out using Minitab 18 statistical software (Minitab, LLC, USA) and Excel (Microsoft Office 2016's), to develop both linear and nonlinear (power) models. For the THMs, a forward selection process was used, and for the HAAs, a backward elimination process was carried out. Before multiple regression analysis, the statistical significance of each direct, quadratic, and interactive predictor variable was verified using a Pearson correlation matrix at a 95% significance level. The models investigated here include the principal factors, and interactive and higher-order factors for both linear and nonlinear forms. The principal factor models are direct and in their very simplest form (Chowdhury et al. 2011). The generalized form of the mathematical models for predicting the THMs and HAAs values are presented in Eqs. 1 and 2, where y represents the THMs and HAAs, β represents the model coefficient, x represents the predictor variables, and ε represents the residuals or errors.

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}^{'} + \varepsilon \tag{1}$$

$$y = \beta_0 \cdot \sum_{j=1}^{k} x_{ij}^{'\,\beta_j} + \varepsilon \tag{2}$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$. The $n > k$ and x'_{ij} denotes the ith observation of independent variable x_j. The independent variables x'_{ij} includes principal factors (e.g., DOC, pH, and T), interactive variables (effect of two or more varying together e.g., $T \times t$, UV×DOC × $\log Cl_T$) and higher-order variables (e.g., quadratic i.e., T^2, Cl^2). The models for both THMs and HAAs were developed based on the particular values of the independent variables x'_{ij} ($x_{i1}, x_{i2}, \ldots, x_{ik}$). Equations 1 and 2 are the generalized form of the linear and nonlinear models, respectively, for prediction of both THMs and HAAs. Their goodness-of-fit and performance were compared by performing F tests, Student's T test, the coefficient of determination (R^2), the standard error (SE, Eq. 3), the mean square error (MSE, Eq. 4), and the Durbin–Watson statistic (d, Eq. 5).

$$\mathrm{SE} = \sqrt{\frac{\sum_1^{\mathrm{n}}(\mathrm{Observed}-\mathrm{Predicted})^2}{\mathrm{n}}} \tag{3}$$

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_1^{\mathrm{n}}(\mathrm{Observed}-\mathrm{Predicted})^2 \tag{4}$$

$$d = \frac{\sum_1^{n}(e_n - e_{n-1})^2}{\sum_1^{n} e_n^2} \tag{5}$$

In Eq. 5, e is the residual value and is calculated by subtracting the predicted value from the observed value.

To determine the significance of the difference between the measured and predicted values, an F test was performed. For the F value > 0.5, Student's T test with equal variance was performed. In contrast, if the F value < 0.5, Student's T test with unequal variance was performed. If the p value from the T test is < 0.05, the two datasets, i.e., measured and predicted, do not have statistical similarities or are not equivalent. On the other hand, if the p value is > 0.05, the two datasets are equivalent or do not have significant statistical differences.

Model validation and applicability

The validation process determines or confirms how sound and effective the models are. In this study, it shows the stability and reasonableness of the THM and HAA models. For validation and applicability, the developed models were subjected to two different types of tests: (i) comparisons of measured and predicted values by performing internal evaluations, i.e., on more recent or additional datasets from the same DWTPs on which the models were based (calibration), and (ii) comparisons of measured and predicted values by performing external evaluations, i.e., on datasets from different DWTPs. The developed models were used to predict both THMs and HAAs for the additional datasets (January 2017 to July 2017) obtained from three DWTPs, as well as external datasets (January 2016 to December 2016). Analyses were done to calculate the R^2, SE, and MSE values. A T test was performed on the predicted models to determine the biases by calculating the t value and t critical or p value. The values were compared and if t-calculated < t-critical or the p value > 0.05, bias was considered not significant and vice versa.

Results and discussion

The occurrence of THMs and HAAs

The range and average levels of THMs and HAAs in treated water from the three DWTPs are summarized in Fig. 2 and data were collected from 2015 to 2016. The formation of THMs is ranked for the DWTPs as Gangbuk DWTP > Gwangam DWTP > Yeongdeungpo DWTP. For the HAAs, there were no significant differences between the three DWTPs. The observed maximum values of THMs were 33 μg/L, 39 μg/L, and 29 μg/L for Gwangam, Gangbuk, and Yeongdeungpo, respectively. Very low values for HAAs were observed in all three DWTPs. The maximum values were found to be 6 μg/L, 6 μg/L, and 7 μg/L in Gwangam, Gangbuk, and Yeongdeungpo, respectively. The measured value for THMs were higher and dispersed compared with HAAs throughout the year because of high hydrophobic fraction of NOM compared with hydrophilic fraction (Bond et al. 2012). Figure 3 shows seasonal variations of THMs and HAAs in treated water. For THMs, high values were observed during summer (June to August) and at the beginning of the autumn season (September) for all three DWTPs, especially in Gangbuk DWTP. This may be because of temperature changes and organic matter present in the source water. Although the temperature in autumn is slightly lower than in summer, the water is rich in organic matter. The main reason could be the rapid decay of vegetation (Kumari and Gupta 2015). Similarly, lower values were observed during the winter season (December to February). In contrast, HAA values were observed to be higher during spring (March to May). The major HAA species to contribute to the higher concentrations is dichloroacetic acid (DCAA)(Rodriguez et al. 2004) and shows high concentration levels during spring. In addition, DCAA is not affected by the pH levels of the source and treated water. The decrease in the concentration of HAAs during the summer and autumn seasons may be attributed to microbial activities. It has been reported that microorganisms do degrade HAAs over time (Zhou and Xie 2002; Rodriguez et al. 2004).

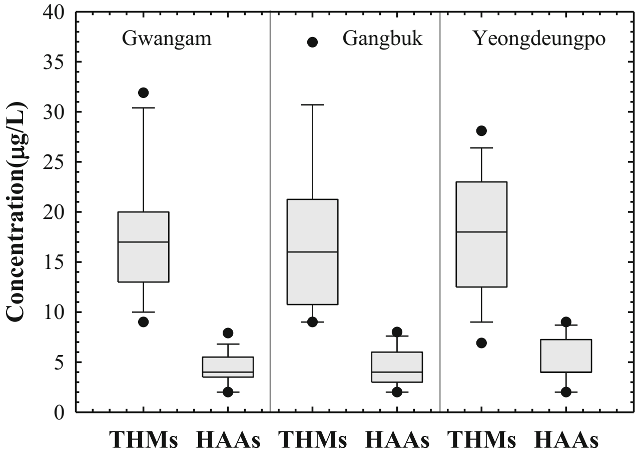

Fig. 2 THMs and HAAs concentration in treated water from three DWTPs. HAAs is HAAs+2

Correlation of independent variables with THMs and HAAs formation

In this research, the models were built by considering principal factors and interactive and higher-order variables. The correlation matrices for both THMs' and HAAs' formation with selected variables were obtained using Pearson's correlation test and are shown in Table 1. A positive and very strong correlation was observed for temperature (quadratic form) and an interactive variable (UV_{254}, temperature, reaction time, and total chlorine dose, i.e., $UV_{254} \times T^2 \times t \times Cl_T$) with THMs formation ($r = 0.888$ and 0.878, respectively). This indicated that higher-order and interactive variables have the largest influence on the formation of THMs. Besides this, the temperature is the variable which has the highest influence compared with other variables. This observation was also reported in other studies (Babaei et al. 2015; Kumari and Gupta 2015). The increase in temperature increases the reaction rate between organic matter and residual disinfection. Temperature acts as an energy source and activates the reaction (Kumari and Gupta 2015). Negative and very good correlations were observed between pH_{avg} (average value of raw water and treated water pH) and THMs formation ($r = -0.709$). The residual chlorine (Cl_R) and postchlorine dose (Cl_{post}) showed moderate correlation ($r = 0.581$ and 0.509, respectively). Compared with other variables, the Br^- concentration does not show good correlation and was found to be negative ($r = -0.124$). This result may be attributed to the very low concentration of

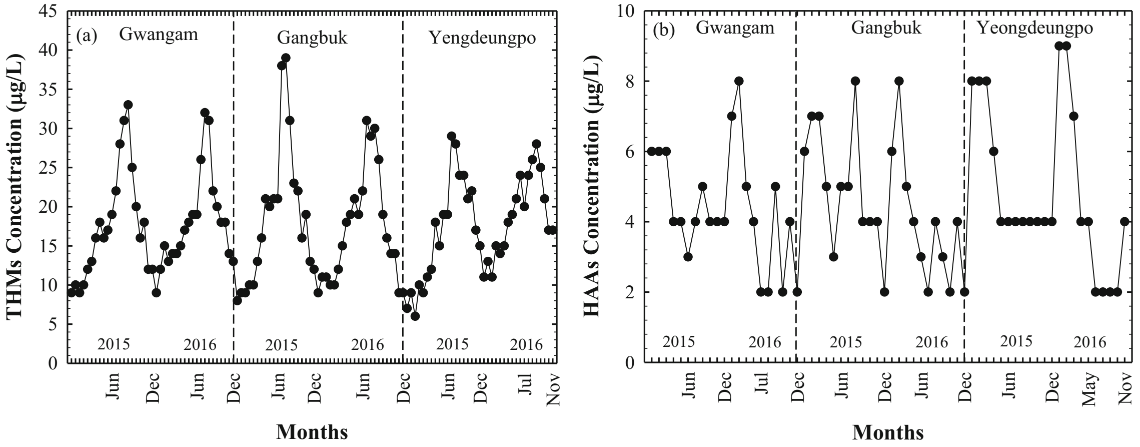

Fig. 3 Monthly variation of THMs and HAAs in three DWTPs

Br^- in treated water. However, the models that included Br^- as an independent variable exhibit a better degree of accuracy than the models that excluded Br^-(Ged et al. 2015). The models excluding the Br^- concentration resulted in the overprediction of THMs for low concentrations of Br^- and underprediction for high concentrations of Br^-. The interactive variable with higher order (quadratic form) shows a negative and very low correlation with THMs formation. Similarly, an attempt was made to find the correlation of primary factors and interactive and higher-order variables with HAAs' formation in treated water. For HAAs, interactive variables such as log ($Cl_T \times DOC/(T \times t)$) and log ($Cl_T \times DOC/(T \times t \times pH_{avg})$) show positive and very good correlation ($r = 0.707$ and 0.706, respectively). For principal factors such as DOC and pH_{avg}, the correlation was found to be positive and moderate ($r = 0.47$ and 0.448, respectively). The ratio of DOC and residual chlorine (DOC/Cl_R) and bromide ion and residual chlorine ($(Br + 1)/Cl_R$) shows positive correlation ($r = 0.576$ and 0.325, respectively). In contrast with THMs, temperature shows negative and moderate correlation ($r = -0.482$) with HAAs' formation. The $\log Cl_R$ also shows negative and does not show good correlation ($r = -0.213$).

Mathematical models for DBPs within DWTPs

The variables (principal factors and interactive and higher-order) that were considered for the mathematical models are shown in Table 2. Before selecting variables, different variables and their combinations were tried to develop both linear and nonlinear models with the best statistical outputs and the accuracy of predictions vs observed values for THMs and HAAs (Table S2). Based on the results, both linear and nonlinear models for THMs and HAAs were developed and a comparative analysis (statistical test) was carried out to determine the best model.

Table 1 Correlation of THMs and HAAs with independent variables (principal factors and interactive and higher-order variables)

THMs predictor variables	Correlation coefficient (r)	HAAs predictor variables	Correlation coefficient (r)
$UV \times T^2 \times t \times Cl_T$	0.878	$\log Cl_R$	− 0.213
T^2	0.888	DOC	0.47
pH_{avg}	− 0.709	$(DOC - 100)_{(\text{if DOC} > 1000)}$[a]	0.508
Cl_{post}	0.509	(DOC/Cl_R)	0.576
Cl_R	0.581	($(Br + 1)/Cl_R$)	0.325
Br + 2	− 0.124	pH_{avg}	0.448
$(UV \times DOC \times \log Cl_T)^2$	− 0.07	$(T - 5)_{(\text{if } T > 5)}$[b]	− 0.482
		log ($Cl_T \times DOC/(T \times t)$)	0.707
		log ($Cl_T \times DOC/(T \times t \times pH_{avg})$)	0.706

Significant value is set at 0.05

[a] DOC values greater than 1000 were optimized by subtracting 100 from the original value, whereas values less than or equal to 1000 were used as is

[b] Temperature (T) values greater than 5 °C were optimized by subtracting 5 from the original value, whereas values less than or equal to 5 °C were used as is

Table 2 Models coefficients and statistical analysis results for the linear and nonlinear models for prediction of THMs and HAAs

	THMs		HAAs	
	Linear	Nonlinear	Linear	Nonlinear
Statistical analysis				
Number	120	120	66	66
F Test	1.093	1.263	1.295	1.21
T Test (*p* value)	1	0.803	1	0.652
R^2	0.915	0.852	0.772	0.554
SE	2.085	2.35	0.997	1.113
MSE	4.06	5.52	0.843	1.538
Durbin–Watson	1.554	1.207	2.064	1.662
Model coefficients				
β_0	85.928	1.156×10^3	2.284×10^3	3.594×10^{-14}
β_1	-5.2×10^{-4}	− 0.032	− 48.812	58.141
β_2	-6.2×10^{-2}	− 0.026	0.014	− 11.476
β_3	1.66×10^{-5}	0.199	0.007	1.946
β_4	3.87×10^{-6}	0.155	− 7.205	10.456
β_5	− 10.25	− 4.81	28.069	0.078
β_6	7×10^{-3}	0.039	297.66	− 4.243
β_7	8.42×10^{-5}	0.109	0.149	0.339
β_8	–	–	-5.032×10^3	79.875
β_9	–	–	5.035×10^3	− 61.297

Trihalomethane models

The linear and nonlinear models for THMs are shown in Eqs. 6 and 7, respectively.

$$\mathrm{THMs} = \beta_0 + \beta_1(\mathrm{UV} \times \mathrm{DOC} \times \log\mathrm{Cl_T})^2 + \beta_2(\mathrm{Br} + 2) + \beta_3(\mathrm{Cl_R})^2 + \beta_4(\mathrm{Cl_{Post}})^2 + \beta_5\left(\mathrm{pH_{avg}}\right) + \beta_6(T)^2 + \beta_7\left(\mathrm{UV} \times T^2 \times t \times \mathrm{Cl_T}\right), \tag{6}$$

$$\mathrm{THMs} = \beta_0 \times \left((\mathrm{UV} \times \mathrm{DOC} \times \log\mathrm{Cl_T})^2\right)^{\beta_1} \times (\mathrm{Br} + 2)^{\beta_2} \times \left((\mathrm{Cl_R})^2\right)^{\beta_3} \times \left((\mathrm{Cl_{post}})^2\right)^{\beta_4} \times \left(\mathrm{pH_{avg}}\right)^{\beta_5} \times \left((T)^2\right)^{\beta_6} \times \left(\mathrm{UV} \times T^2 \times t \times \mathrm{Cl_T}\right)^{\beta_7}, \tag{7}$$

where THMs, DOC, Br + 2, $\mathrm{Cl_R}$, $\mathrm{Cl_{post}}$, and $\mathrm{Cl_T}$ are in μg/L, *T* is in degrees Celsius, *t* is in hours, and β_0, β_1, β_2, β_3, β_4, β_5, β_6, and β_7 are model statistical coefficients. The effects of bromide ion were expressed as Br + 2 to avoid zero prediction values for THMs when the value of bromide ion was under the detection level. The size of the sample (*n*), *F* test, *T* test, coefficient of determination (R^2), standard error (SE), mean square error (MSE), Durbin–Watson statistic (*d*), and the model statistical coefficients values are summarized in Table 2. Student's *T* test results for both linear and nonlinear (*p* value > 0.05, i.e., 1 and 0.803, respectively) show no significant statistical difference between observed and predicted values. In addition, the analysis of variance (ANOVA) result showed that both linear and nonlinear models were statistically significant (*p* value = 0.000). The coefficient of determination for the linear model ($R^2 = 0.915$) was found to be greater than for the nonlinear model ($R^2 = 0.852$). In contrast, the observed values of SE and MSE for linear models (2.085 and 4.06, respectively) were found to be lower than the values for the nonlinear model (2.350 and 5.52, respectively). This suggests that the linear model performs better than the nonlinear model for THMs. This result is supported by the *d* value. The value of *d* is preferred to be in the range between 1.5 and 2.5 for a statistically best-fit model (Uyak et al. 2007; Kumari and Gupta 2015; Domínguez-Tello et al. 2017). The value of *d* was found to be 1.554 for the linear model and 1.207 for the nonlinear model. This indicated that the linear model is statistically the best-fit model. Figure 4 a shows the plot for the observed and predicted THMs values in the three DWTPs. The model predicted most of the peak observed values consistently.

Haloacetic acid model

Table 2 summarizes both the linear and nonlinear HAA models' statistical coefficient and regression analysis results. Although the dataset used for these models was relatively small because of the unavailability of all the independent predictive variables, and the treated water concentrations were low, this study still attempted to develop a model for HAAs. Very limited research has been conducted to develop the HAA model. The model suggested by Sérodes et al. 2003 has good predictive capability ($R^2 = 0.92$), but the model did not consider the pH levels and was not validated. Similarly, the model developed by Nikolaou et al. in 2004 has very low predictability ($R^2 = 0.28$) and did not incorporate TOC, DOC, and temperature. The ANOVA analysis of the models formulated in this research shows the models to be statistically significant (*p* value = 0.000). The models formulated are as follows:

$$\begin{aligned}(\mathrm{HAAs}+2) = {} & \beta_0 + \beta_1 \log \mathrm{Cl_R} + \beta_2 \mathrm{DOC} + \beta_3 (\mathrm{DOC}{-}100)_{\mathrm{For\ DOC}>1000} + \beta_4 \left(\frac{\mathrm{DOC}}{\mathrm{Cl_R}}\right) + \beta_5 \left(\frac{(\mathrm{Br}+1)}{\mathrm{Cl_R}}\right) \\ & + \beta_6 \mathrm{pH_{avg}} + \beta_7 (T{-}5)_{\mathrm{For}\ T>5} + \beta_8 \log(\mathrm{Cl_T} \times \mathrm{DOC}/(T \times t)) + \beta_9 \log\left(\mathrm{Cl_T} \times \mathrm{DOC}/\left(T \times t \times \mathrm{pH_{avg}}\right)\right),\end{aligned} \tag{8}$$

$$\begin{aligned}(\mathrm{HAAs}+2) = {} & \beta_0 \times \log \mathrm{Cl_R}^{\beta_1} \times \mathrm{DOC}^{\beta_2} \times (\mathrm{DOC}{-}100)^{\beta_3}_{\mathrm{For\ DOC}>1000} \times \left(\frac{\mathrm{DOC}}{\mathrm{Cl_R}}\right)^{\beta_4} \times \left(\frac{(\mathrm{Br}+1)}{\mathrm{Cl_R}}\right)^{\beta_5} \times \mathrm{pH_{avg}}^{\beta_6} \\ & \times (T{-}5)^{\beta_7}_{\mathrm{For}\ T>5} \times \log(\mathrm{Cl_T} \times \mathrm{DOC}/(T \times t))^{\beta_8} \times \log\left(\mathrm{Cl_T} \times \mathrm{DOC}/\left(T \times t \times \mathrm{pH_{avg}}\right)\right)^{\beta_9},\end{aligned} \tag{9}$$

where β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7, β_8, and β_9 are statistical model coefficients, HAAs, DOC, Br + 1, $\mathrm{Cl_R}$, and $\mathrm{Cl_T}$ are in μg/L, *T* is in degrees Celsius, and *t* is in hours. The HAAs concentration values were expressed as HAAs +2 to avoid zero values of prediction and to enhance correlation with independent variables. Student's *T* test results for both the linear and nonlinear models (*p* value > 0.05, i.e., 1 and 0.652, respectively) show no significant statistical difference between the observed and predicted values. It is noted that the coefficient of determination for the linear model ($R^2 = 0.772$) is higher than for the nonlinear model ($R^2 = 0.652$). The SE and MSE were found to be 0.997 and 0.843, respectively, for the linear model and are lower than for the nonlinear model. Although the *d* values for both models are in the range of 1.5

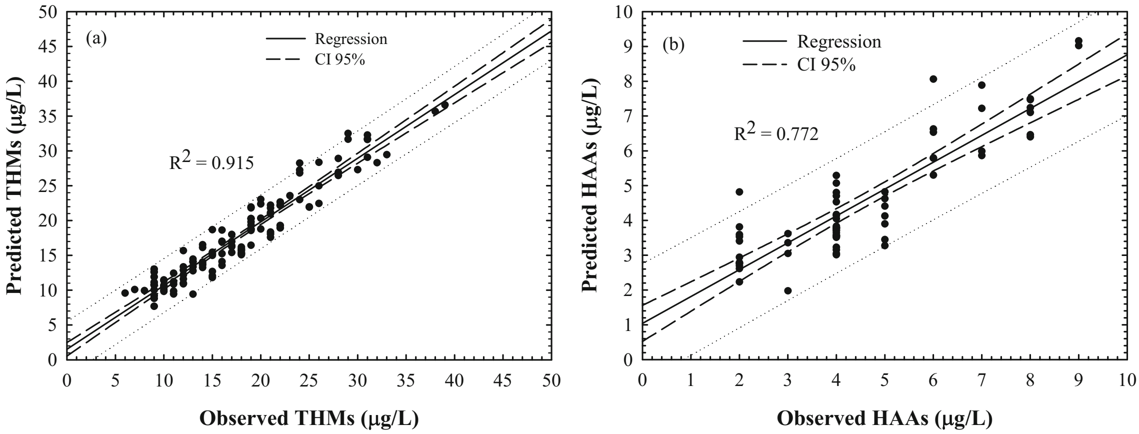

Fig. 4 Calibration of predicted vs. observed concentration **a** THMs linear model and **b** HAAs linear model

Table 3 Validation and application of linear models for THMs and HAAs

	Number	R^2	SE	MSE	*t*-value	*t*-critical	*p* value	Significant
THM linear Model								
Validation	33	0.914	2.308	5.325	− 0.164	1.998	0.87	No
Application								
Amsa DWTP	24	0.84	4.293	18.427	1.372	2.013	0.177	No
Guui DWTP	17	0.872	4.033	16.264	1.358	2.037	0.184	No
HAAs linear Model								
Validation	18	0.794	0.841	0.778	− 0.601	2.032	0.552	No
Application								
Amsa DWTP	12	0.684	2.971	8.83	− 1.786	2.074	0.088	No
Guui DWTP	16	0.704	1.037	1.076	− 0.861	2.042	0.396	No

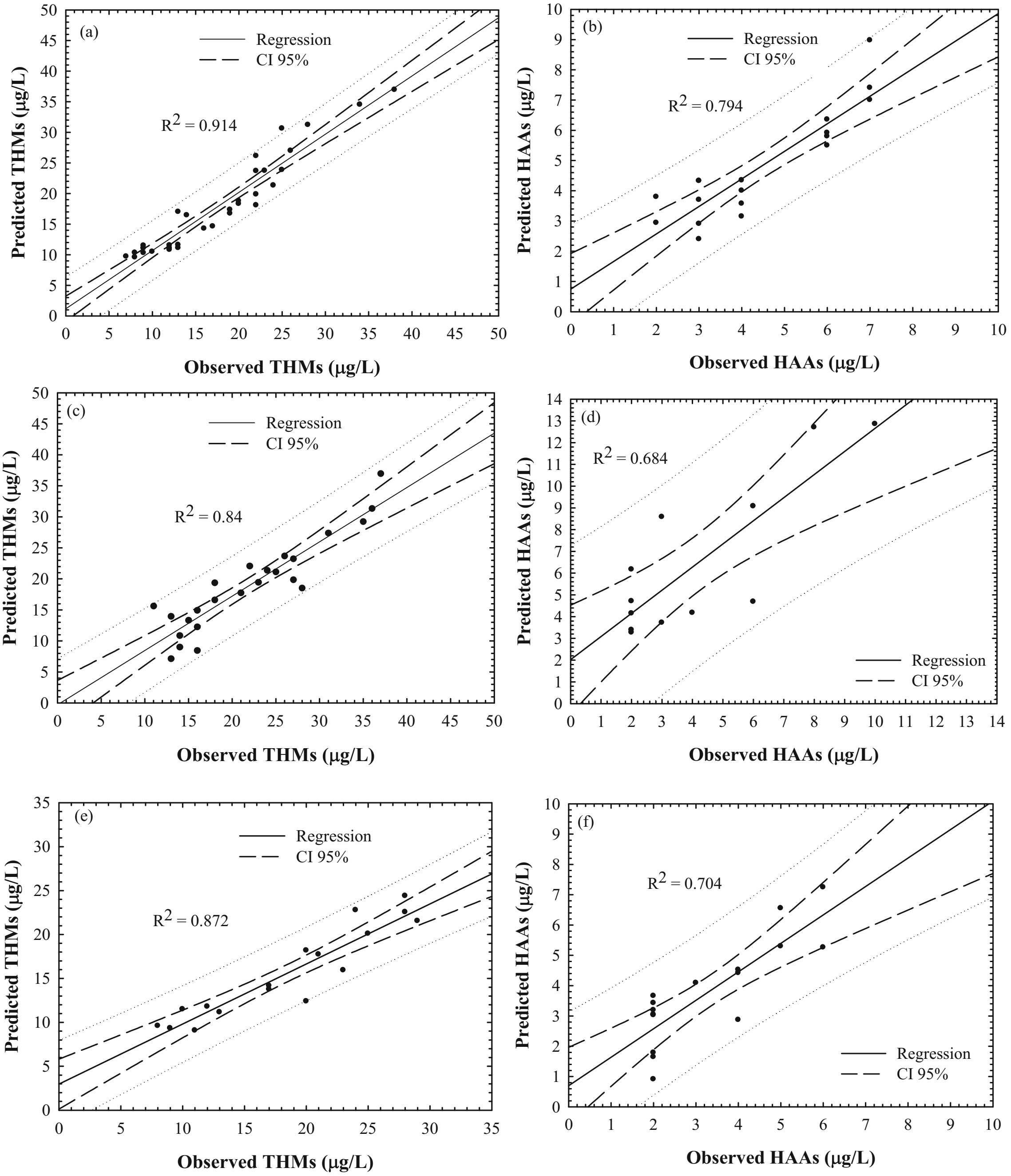

Fig. 5 Predicted vs. observed concentration **a** THMs validation, **b** HAAs validation, **c** THMs application Amsa DWTP, **d** HAAs application Amsa DWTP, **e** THMs application Guui DWTP, and **f** HAAs application Guui DWTP

to 2.5, the model selection was made based on the R^2, SE, and MSE values. Based on the statistical analysis, the linear model was adopted in this research study. Figure 4 b shows the predicted vs observed HAAs.

Validation of THM and HAA model

To demonstrate or confirm the effectiveness of the models (Eqs. 6 and 8) for which they are intended, the models must be validated. For validation, data from the same DWTPs for January 2017 to July 2017 were used. The concentration of THMs and HAAs were predicted using new independent datasets and were compared with measured values. Analyses were done to determine R^2, SE, and MSE for the validation dataset and the results are summarized in Table 3. The R^2 for THMs and HAAs were found to be 0.914 and 0.794, respectively, for the validation dataset. Both models showed a good correlation with the measured and predicted values. The bias of both THM and HAA models was demonstrated using Student's *T* test. The *t*-critical values for a two-tailed test at the significance level of 0.05 were found to be 1.998 and 2.032 with degrees of freedom of 64 and 34, respectively. The *t*-calculated values for THMs (− 0.164) and HAAs (− 0.601) from the *T* test results were found to be less than the *t*-critical. Moreover, the *p* values for both THMs (0.87) and HAAs (0.552) were greater than 0.05. This suggests that the measured and predicted values did not display significant differences. The plots between measured and predicted values are shown in Fig. 5a, b for THMs and HAAs, respectively.

Application of THMs and HAAs to different DWTPs

The models developed were applied in two different DWTPs (Amsa and Guui) to evaluate their suitability, soundness, and effectiveness. The data from these two DWTPs were not used during model development and validation. The statistical results are shown in Table 3. A similar approach to the model validation was applied for the application of the models to these two DWTPs. Both THM and HAA models showed slight decreases in performance, as indicated by the coefficient of determination values. The values obtained were 0.840 and 0.872 for the THM model, and 0.684 and 0.704 for the HAA model at the Amsa and Guui DWTPs, respectively. Compared with the HAA model, the THM model showed better performance. The *T* test was performed to determine bias. For both DWTPs, the *t*-calculated values were found to be less than *t*-critical, and the *p* values were also greater than 0.05. The *t*-calculated values obtained for the THM model were 1.372 for the Amsa DWTP and 1.358 for the Guui DWTP. Likewise, for the HAA model, the values obtained were − 1.786 for the Amsa DWTP and − 0.861 for the Guui DWTP. These statistical results suggest that the measured and predicted values do not display significant differences. The measured and predicted value plots for the THMs and HAAs for the Amsa and Guui DWTPs are shown in Fig. 5c–f. However, the data generation for this model application study was limited. Overall, the models for both THMs and HAAs showed moderate to very good predictability.

Conclusion

This research study developed predictive models for both THMs and HAAs. Stepwise multiple regression analysis was used to develop both models. The approach used for the model development provided critical information regarding predictor variables. The quadratic form (temperature) and interactive variable (UV_{254}, temperature, reaction time, and total chlorine dose, i.e., $UV_{254} \times T^2 \times t \times Cl_T$) for THMs formation, and interactive variables such as log ($Cl_T \times DOC/(T \times t)$) and log ($Cl_T \times DOC/(T \times t \times pH_{avg})$) for HAAs, show more significance effects than their respective principal variables. The incorporation of higher-order and interactive variables enhances the predictability of the models. This study also indicated that a better understanding of the effects of interacting and higher-order variables is needed. For both THMs and HAAs, linear models were found to show better performance than nonlinear models. The validation and application of models showed no significant differences between measured and predicted values. These models may be useful in the identification of strategies and decision-making to improve the treatment and disinfection process of drinking water in South Korea and to other places with similar climatic conditions and treatment processes.

Funding information This paper is supported by Korea Ministry of Environment as "Global Top Project (2016002110007)."

References

Abdullah MP, Yew CH, Ramli MS (2003) Formation, modeling and validation of trihalomethanes (THM) in Malaysian drinking water: a case study in the districts of Tampin, Negeri Sembilan and Sabak Bernam, Selangor, Malaysia. Water Res 37:4637–4644

Amy GL, Chadik PA, Chowdhury ZK (1987) Developing models for predicting trihalomethane formation potential kinetics. J Am Water Works Assoc 79:89–97

Ata S, Wattoo FH, Din MI, Wattoo MHS, Qadir MA, Tirmizi SA, Abdullah P (2015) Critical study of multiple regression modelling for monitoring of haloacetic acids in water reservoirs. Arab J Sci Eng 40:101–108

Babaei AA, Atari L, Ahmadi M, Ahmadiangali K, Zamanzadeh M, Alavi N (2015) Trihalomethanes formation in Iranian water supply systems: predicting and modeling. J Water Health 13:859–869

Bond T, Goslan EH, Parsons SA, Jefferson B (2012) A critical review of trihalomethane and haloacetic acid formation from natural organic matter surrogates. Environ Technol Rev 1:93–113

Bond T, Huang J, Graham NJD, Templeton MR (2014) Examining the interrelationship between DOC, bromide and chlorine dose on DBP formation in drinking water – a case study. Sci Total Environ 470-471:469–479

Brown D, Bridgeman J, West JR (2011) Predicting chlorine decay and THM formation in water supply systems. Rev Environ Sci Biotechnol 10:79–99

Chowdhury S, Champagne P, McLellan PJ (2009) Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. Sci Total Environ 407:4189–4206

Chowdhury S, Rodriguez MJ, Serodes J (2010) Model development for predicting changes in DBP exposure concentrations during indoor handling of tap water. Sci Total Environ 408:4733–4743

Chowdhury S, Rodriguez MJ, Sadiq R, Serodes J (2011) Modeling DBPs formation in drinking water in residential plumbing pipes and how water tanks. Water Res 45:337–347

Domínguez-Tello A, Arias-Borrego A, García-Barrera T, Gómez-Ariza JL (2017) A two-stage predictive model to simultaneous control of trihalomethanes in water treatment plants and distribution systems: adaptability to treatment processes. Environ Sci Pollut Res 24: 22631–22648

Elshorbagy WE, Abu-Qadais H, Elsheamy MK (2000) Simulation of THM species in water distribution system. Water Res 34:3431–3439

Fooladvand M, Ramavandi B, Zandi K, Ardestani M (2011) Investigation of trihalomethanes formation potential in Karoon River water, Iran. Environ Monit Assess 178:63–71

Ged EC, Chadik PA, Boyer TH (2015) Predictive capability of chlorination disinfection byproducts models. J Environ Manag 149:253–262

Golfinopoulos SK, Arhonditsis GB (2002) Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. Chemosphere 47:1007–1018

Golfinopoulos SK, Xilourgidis NK, Kostopoulou MN, Lekkas TD (1998) Use of a multiple regression model for predicting trihalomethane formation. Water Res 32:2821–2829

Hua G, Reckhow DA, Abusallout I (2015) Correlation between SUVA and DBP formation during chlorination and chloramination of NOM fractions from different sources. Chemosphere 13:82–89

Krasner SW, Cantor KP, Weyer PJ, Hildesheim M, Amy G (2017) Case study approach to modeling historical disinfection by-product exposure in Iowa drinking waters. J Environ Sci 58:183–190

Kulkarni P, Chellam S (2010) Disinfection by-product formation following chlorination of drinking water: artificial neural network models and changes in speciation with water. Sci Total Environ 408:4202–4210

Kumari M, Gupta SK (2015) Modeling of trihalomethanes (THMs) in drinking water supplies: a case study of eastern part of India. Environ Sci Pollut Res 22:12615–12623

Lin J, Chen X, Zhu A, Hong H, Liang Y, Sun H, Lin H, Chen J (2018) Regression models evaluating THMs, HAAs and HANs formation upon chlorination of source water collected from Yangtze River Delta region, China. Ecotoxicol Environ Saf 160:249–256

Maeng M, Shahi NK, Shin G, Son H, Kwak D, Dockko S (2018) Formation characteristics of carbonaceous and nitrogenous disinfection by-products depending on residual organic compounds by CGS and DAF. Environ Sci Pollut Res

Morris RD, Audet D-M, Angelilo IF (1992) Chlorination, chlorination by-products and cancer: a meta analysis. Am J Public Health 82: 955–963

Mukundan R, Van Derson R (2014) Predicting trihalomethanes in the New York city water supply. J Environ Qual 43:611–616

Nikolaou AD, Lekkas TD, Golfinopoulos SK (2004) Kinetics of the formation and decomposition of chlorination by-products in surface waters. Chem Eng J 100:139–148

Platikanov S, Martín J, Tauler R (2012) Linear and non-linear chemometric modeling of THM formation in Barcelona's water treatment plant. Sci Total Environ 432:365–374

Richardson SD, Plewa MJ, Wagner ED, Schoeny R, DeMarini DM (2007) Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: a review and roadmap for research. Mutat Res 636:178–242

Rodriguez MJ, Sérodes JB, Levallois P (2004) Behavior of trihalomethanes and haloacetic acids in a drinking water distribution system. Water Res 38:4367–4382

Sadiq R, Rodriguez MJ (2004) Disinfection by-products(DBPs) in drinking water system and predictive models for their occurrence. Sci Total Environ 321:21–46

Semerjian L, Dennis J, Ayoub G (2009) Modeling the formation of trihalomethanes in drinking waters of Lebanon. Environ Monit Assess 149:429–436

Sérodes JB, Rodriguez MJ, Li H, Bouchard C (2003) Occurrence of THMs and HAAs in experimental chlorinated waters of the Quebec City area (Canada). Chemosphere 51:253–263

Singh KP, Gupta S (2012) Artificial intelligence based modeling for predicting the disinfection by-products in water. Chemom Intell Lab 114:122–131

SMG (2017) Seoul Tap Water Arisu. Seoul Metropolitan Government

Sohn J, Amy G, Cho J, Lee Y, Yoon Y (2004) Disinfection decay and disinfection by-products formation model development: chlorination and ozonation by-products. Water Res 38:2461–2478

Uyak V, Toroz I, Meriç S (2005) Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul. Desalination 176:91–101

Uyak V, Ozdemir K, Toroz I (2007) Multiple linear regression modeling of disinfection by-products formation in Istanbul drinking water reservoirs. Sci Total Environ 378:269–280

Westerhoff P, Debroux J, Amy GL, Gatel D, Mary V, Cavard J (2000) Applying DBP models to full-scale plants. J Am Water Works Assoc 92:89–102

WHO (2005) Trihalomethane in drinking water: background document for development of WHO guidelines for drinking water quality. World Health Organization

Zhou HJ, Xie YFF (2002) Using BAC for HAA removal-part 1: batch study. J Am Water Works Assoc 94:194–200

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.