Environmental Science and Pollution Research (2022) 29:32845–32854
https://doi.org/10.1007/s11356-021-18299-0

RESEARCH ARTICLE

Modeling the formation of trihalomethanes in rural and semi-urban drinking water distribution networks of Costa Rica

Daniel Enrique Kelly-Coto[1] · Alejandra Gamboa-Jiménez[2] · Diana Mora-Campos[2] · Pablo Salas-Jiménez[2] · Basilio Silva-Narváez[3] · Joaquín Jiménez-Antillón[4] · Macario Pino-Gómez[4] · Luis Guillermo Romero-Esquivel[4]

Received: 13 October 2021 / Accepted: 20 December 2021 / Published online: 12 January 2022

Abstract

Chlorination is one of the most important stages in the treatment of drinking water due to its effectiveness in the inactivation of pathogenic organisms. However, the reaction between chlorine and natural organic matter (NOM) generates harmful disinfection by-products (DBPs), such as trihalomethanes (THMs). In this research, drinking water quality data was collected from the distribution networks of 19 rural and semi-urban systems that use water sources as springs, surfaces, and a mixture of both, in three provinces of Costa Rica from April 2018 to September 2019. Twelve models were developed from four data sets: all water sources, spring, surface, and a mixture of spring and surface waters. Linear, logarithmic, and exponential multivariate regression models were developed for each data set to predict the concentration of total trihalomethanes (TTHMs) in the distribution networks. Concentrations of TTHMs were found between <0.20 and 91.31 µg/L, with chloroform being the dominant species accounting for 62% of TTHMs on average. Turbidity, free residual chlorine, total organic carbon (TOC), dissolved organic carbon (DOC), and ultraviolet absorbance at 254 nm (UV_{254}) showed a significant correlation with TTHMs. In all the data sets the linear models presented the best goodness-of-fit and were moderately robust. Four models, the best of each data set, were validated with data from the same systems, and, according to the criteria of R^2, standard error (SE), mean square error (MSE), and mean absolute error (MAE), spring water and mixed spring/surface water models showed a satisfactory level of explanation of the variability of the data. Moreover, the models seem to better predict TTHM concentrations below 30 µg/L. These models were satisfactory and could be useful for decision-making in drinking water supply systems.

Keywords Drinking water · Disinfection · Disinfection by-products · Trihalomethanes · Predictive models · Multiple regression analysis

Introduction

Disinfection is one of the most important stages in water treatment to reduce the content of pathogenic material. In most of the world, chlorine disinfection is the most widely used method for its high effectiveness in preventing pathogenic microorganisms and its low cost (Mazhar et al. 2020). However, chlorine can react with natural organic matter (NOM) present in water from supply sources and generate disinfection by-products (DBPs) such as trihalomethanes (THMs) (Richardson and Plewa 2020). The formation of THMs is influenced by several factors: operational variables (e.g., pH, type and disinfectant dose, residence time), environmental conditions (e.g., water temperature and seasonal variation), and water characteristics (e.g., type and concentration of NOM, bromide ion concentration) (Al-Tmemy et al. 2018).

Responsible Editor: Ester Heath

✉ Daniel Enrique Kelly-Coto
dkellycoto@gmail.com

[1] Environmental Engineering Licentiate Program, School of Chemistry, Instituto Tecnológico de Costa Rica, P.O. Box 159, Cartago, Costa Rica

[2] Environmental Analysis Laboratory, School of Environmental Sciences, National University, P.O. Box 86-3000, Heredia, Costa Rica

[3] Basic Sciences, National Technical University, P.O. Box 1902-4050, Alajuela, Costa Rica

[4] Environmental Protection Research Center (CIPA), School of Chemistry Instituto Tecnológico de Costa Rica , P.O. Box 159, Cartago, Costa Rica

Various researchers have reported adverse human health effects from exposure to THMs, for example, bladder cancer (Costet et al. 2011), colorectal cancer (Rahman et al. 2010), miscarriage, and congenital anomalies (Wright et al. 2017). In addition, some THMs are classified as possibly carcinogenic (IARC 2021) . Therefore, maximum contaminant level (MCL) has been established for drinking water. The U.S. Environmental Protection Agency (US EPA) establishes an 80 μg/L MCL for total THMs (TTHMs) that include chloroform, bromoform, dibromochloromethane, and bromodichloromethane (US EPA 1998). In Costa Rica, MCL of 200 μg/L, 100 μg/L, 100 μg/L, and 60 μg/L, respectively, are established (MINSA 2018).

Monitoring of THMs is important to avoid the aforementioned adverse effects and for compliance with legislation. However, the most common method for THM determination by gas chromatography is expensive and time consuming (Mukundan and Van Dreason 2014). As a tool for decision making, multiple prediction models have been developed. These models can be generated from laboratory or field data by collecting samples at the treatment plant and/or distribution network (Sadiq et al. 2019). For the first case, they have the advantage that many variables can be controlled; however, it does not contemplate certain aspects that occur on a real scale (Chowdhury et al. 2009). The models obtained with field data have the advantage of contemplating variables such as the influence of the infrastructure of the distribution networks; however, they are specific to each site (Shahi et al. 2020) and therefore cannot be generalized to any context (Semerjian et al. 2009). The prediction models can be classified into mechanistic ones based on the kinetics of chlorine reactions, and empirical ones (Kumari and Gupta 2015). The DBP empirical models are based on the water quality, operational and environmental conditions that influence its formation. The models are developed using statistical regression or artificial neural networks (Sadiq et al. 2019). Accordingly with the same study, the generation of empirical models benefit in understanding the factors that contribute to the formation of THMs and are a tool for decision-making.

In the literature, most models predicting the formation of THMs have been developed in temperate and urban zones, for example, in Quebec, Canada (Rodriguez et al. 2000); New York, USA (Mukundan and Van Dreason 2014); and Seoul, South Korea (Shahi et al. 2020). Moreover, models have been reported for systems located in semi-arid areas like the city of Ahvaz, Iran (Babaei et al. 2015) and Wassit Province Southeast Iraq (Al-Tmemy et al. 2018), the Mediterranean region in Lebanon (Semerjian et al. 2009), and in few cases in tropical regions, for example, in Thailand (Feungpean et al. 2015). In general, considering that the NOM present in the different water sources is influenced by autochthonous and allochthonous production, it is expected to find differences in the nature of NOM depending on the region (Edzwald and Tobiason 2011). Therefore, it is expected to develop models for the different sites. The present research is the first attempt to develop a THM prediction in Costa Rica and to the best of the authors' knowledge in the Central American and Caribbean region. Furthermore, this study was focused on rural and semi-urban areas, where no studies were found in the literature.

In Costa Rica, 93% of the population received drinking water in 2019 (PEN and CONARE 2020). Moreover, in the same year, 19.4% of homes in rural and semi-urban areas were supplied with water by local Associations Administrators of Aqueduct and Sewerage Systems (ASADAs in Spanish) (Sánchez-Hernández 2019). In addition, in 2016, 14.3% of the population was supplied by 24 municipalities and the rest by duly organized public companies (AyA 2016). The main water sources used are groundwater, springs, surface water, and the mixture of the two latter ones. In all cases, chlorine disinfection is the method used (Arellano-Hartig et al. 2020). In general, due to economic and analytical capacity limitations, monitoring of THMs is scarce, mainly at the ASADAs and municipal levels. Thus, the objective of this study was to develop a series of prediction models of TTHMs in the distribution systems of rural and semi-urban areas supplied by springs, surface water, and the mixture of both sources. This is the first study of its kind carried out in the country and is expected to serve as a tool for decision-making in the aqueducts regarding their operation and parameters to be monitored.

Materials and methods

Study site and drinking water systems

The study was performed in three different zones of the country (Fig. 1). The sites present a dry season from December to March, a rainy season from May to October, and two months of transition, April and November (Manso et al. 2005). Nineteen small distribution systems of rural or semi-urban areas were selected. The population of most of the systems ranges from 328 to 8000 inhabitants. The length of the distribution networks ranges from 1.2 to 13 km. The raw water sources of the systems were surface (6), springs (6), and a mixture of both (7). The surface water and the mixture of water sources were treated with conventional treatment systems (2), slow sand filtration (1), screening or sedimentation (5), multi-stage filtration (1), and coarse-layered filtrations (2). The water was chlorinated in 16 cases with solid $Ca(ClO)_2$, in one case with liquid NaClO, and in two systems generated in situ by electrolysis. In this study, mainly in spring water, chlorination was the only treatment;

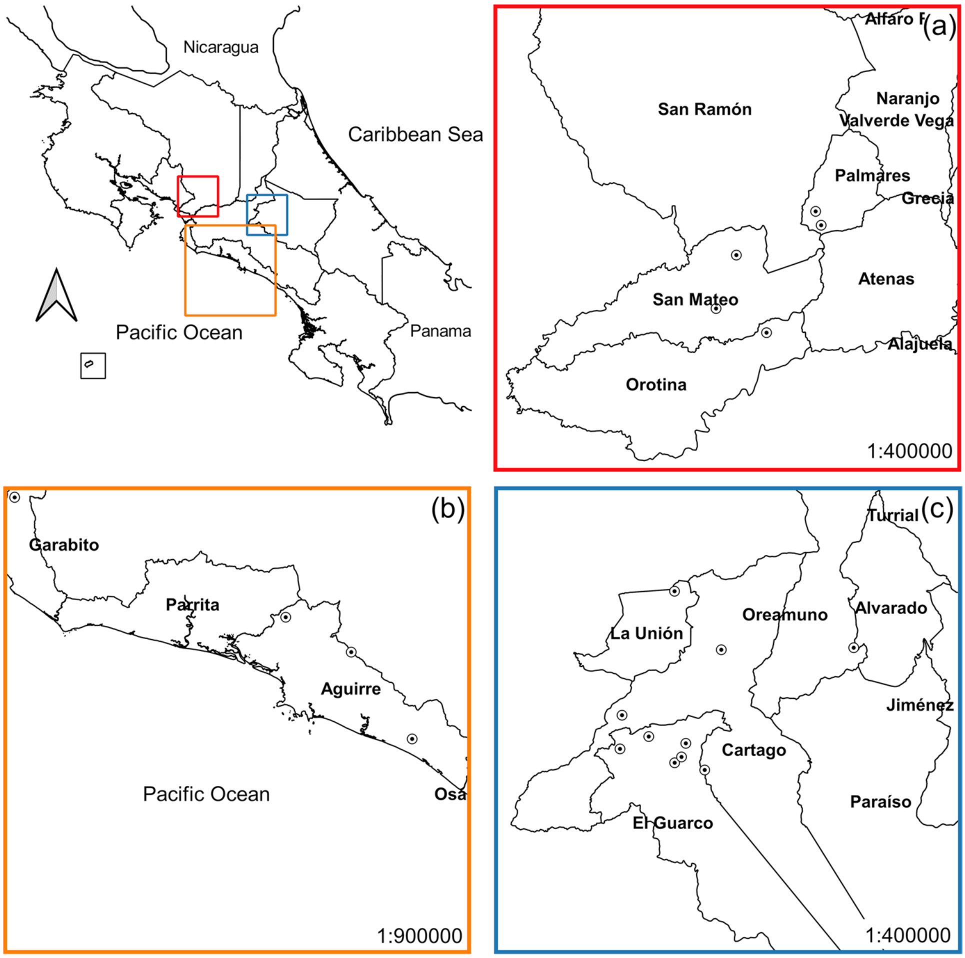

Fig. 1 Study site including all the drinking water systems in three provinces of the country: (**a**) Alajuela, (**b**) Puntarenas, and (**c**) Cartago

therefore, water subjected solely to chlorination was considered as treated water.

Water sampling and analytical procedures

Water samples from the 19 systems were collected from three different sampling campaigns, in the dry, transition, and rainy seasons, respectively. The study period was between April 2018 and September 2019. Each sampling day, four samples, at different points of the distribution network, were taken as recommended by the local legislation (MINSA 2018). Specifically, the sampling points were at the exit of the chlorination storage tank (minimum estimated contact time design of 30 min) and the beginning, the middle, and the end of the distribution network.

Total and dissolved organic carbon, TOC and DOC, respectively, were determined using a Teledyne Tekmar TOC Fusion model device following the SM5310 C method of the Standard Methods (APHA et al. 2017). The limit of detection and quantification were 0.03 and 0.05 mg C/L, respectively. For the determination of DOC, the samples were filtered using a cellulose nitrate membrane of 0.45 µm. The ultraviolet absorbance at 254 nm (UV_{254}) was determined using a spectrophotometer Shimadzu model UV 1800 ENG120V with a 1-cm optical length and following the 5910B method of the Standard Methods (APHA et al. 2017). From the ratio of UV_{254} values to DOC concentrations, the specific ultraviolet light absorbance (SUVA) was calculated.

Total THMs (TTHMs) were calculated as the sum of chloroform, bromoform, dibromochloromethane, and bromodichloromethane. These substances were determined following method 6040 D (APHA et al. 2017) using Agilent 7890A equipment with an electron capture detector (ECD) and solid phase microextraction with a polydimethylsiloxane (PDMS) fiber. The THMs were analyzed using a calibration curve of 6 standards in a range between (0–10) µg/L ($r^2 > 0.995$). Helium was used as a carrier gas (4 mL/min) and a ZB-624 capillary column (length: 105 m, ID: 0.53 mm, layer thickness: 3.00 µm). The initial oven temperature was 35 °C and the final temperature was 250 °C with an increment of 5 °C/min. The detection and quantification limits of chloroform, bromoform, dibromochloromethane, and bromodichloromethane were 0.2 µg/L, 0.06 µg/L, 0.07 µg/L, and 0.06 µg/L and 0.6 µg/L, 0.2 µg/L, 0.2 µg/L, and 0.2 µg/L, respectively.

In the field, pH was determined at all sampling points using Hanna HI 8–124 equipment and free chlorine was determined using a colorimeter (Pocket Colorimeter II, Hach) following the DPD method (N, N-diethyl-p-phenylenediamine). Turbidity and apparent color were determined in the laboratory in less than 24 h after sampling using 2100Q and DR900 equipment (both Hach). In all cases, the methods of the Standard Methods (APHA et al. 2017) or those recommended by the equipment manufacturers were followed.

Mathematical model development

The models were developed using the data from the water samples taken at the exit of the chlorinated water storage tank and in the distribution network of each system. The models have developed from four data sets accordingly to the source water of the systems: (1) all sources, (2) spring, (3) surface, and (4) mixture of surface and spring waters refer as mixed. Before the analysis, an aleatory code was assigned to each sample, and with the help of Minitab 17 statistical software, each database was randomly divided into two groups: calibration data (70% of the total) and validation data (30% of the total). A similar procedure was reported by Golfinopoulos and Arhonditsis (2002) for the development of multivariate regression models for the prediction of THMs in a water treatment plant in Greece.

Initially, the normality of TTHMs and variables like temperature, pH, turbidity, color, free residual chlorine, TOC, DOC, and UV_{254} reported by Sadiq et al. (2019) as potentially influential in the formation of THMs were evaluated using the Anderson–Darling test (Ryan 2007). As it will be discussed later, the variables presented a non-normal distribution as shown in Table S1 (Online Resource 1); therefore, as recommended by Kargaki et al. (2020) for non-parametric data, the Spearman correlation test with a significance level (α) of 0.05 was used. Using this test, the Spearman correlation coefficient (r_s) and their respective p-value were determined. Similar to Chowdhury et al. (2008) applied criteria for Pearson's correlation coefficient in THM model development, in the present research an r_s below 0.3 means weak correlation, between 0.3 and 0.7 moderate and greater than 0.7 strong correlation. Furthermore, the correlation was considered statistically significant if the $p\text{-value} < 0.05$ and vice versa.

Multiple regression analysis was performed in the Minitab 17 statistical software program for the development of linear and non-linear models. TTHM concentrations were considered as the dependent variables, while the other water quality parameters were considered as the independent variables. Once the potential variables to include in the models were identified, as recommended by Feungpean et al. (2015), the stepwise method was used to identify the significant variables in the explanation of variability provided by the model. In the stepwise method, each of the variables is included or excluded when evaluating the p-value of the F test, against the alpha values to enter or leave the model considering a significance level of 0.05.

To find the model that represents the best performance and goodness-of-fit of the data, for each data set, linear and non-linear models were generated. Transformations were applied in the dependent and/or independent variables (e.g., square root, exponential, logarithmic) (Pardoe 2012). In all cases, data exclusion criteria were used, such as studentized residual deleted greater than 3, high leverage points, Cook's distance, and DFTIS (Acuña-Fernández 2004).

Subsequently, for the models obtained, the statistical assumptions were evaluated: normality, constant variance or homoscedasticity, and independence (Acuña-Fernández 2004) (Figs. S1–S4, Online Resource 1). In addition, for the comparison of performance between the models, the statistical results were analyzed: R^2, R^2 adjusted, the significance of the model (F test), Durbin–Watson statistic, average standard error (SE; Eq. (1)), average square error (MSE; Eq. (2)), and mean absolute error (MAE; Eq. (3)).

$$\mathrm{SE} = \sqrt{\frac{\sum_1^n \left(\mathrm{TTHM_M} - \mathrm{TTHM_P}\right)^2}{\mathrm{n}}} \tag{1}$$

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_1^{\mathrm{n}} \left(\mathrm{TTHM_M} - \mathrm{TTHM_P}\right)^2 \tag{2}$$

$$\mathrm{MAE} = \frac{\sum_1^n \left|\mathrm{TTHM_M} - \mathrm{TTHM_P}\right|}{\mathrm{n}} \tag{3}$$

where TTHM_M indicates the measured TTHMs, TTHM_P indicates the predicted TTHMs by the models, and n refers to the number of observations evaluated. The SE, MSE, and MAE units are µg/L corresponding to the TTHM units.

Models' validation and applicability

The best model obtained for each data set was validated using the excluded data used to obtain the models (30% of the total data). For validation, predicted TTHMs and those measured were compared using the criteria: R^2, SE, and MSE (Shahi et al. 2020). In addition, as the study mentioned, a T-test was performed to determine a significant difference between the mean of the TTHMs measured and the predicted by the models. A test of equal variances was performed to determine whether equal variance could be assumed in the T-test. Next, the T-test was performed by calculating the t-value and its respective p-value. The values were compared, and if the $p\text{-value} > 0.05$, the difference between the measured and predicted values was considered as non-significant and vice versa.

Results and discussion

Water quality parameters

Table 1 presents the main characteristics of the treated/chlorinated water of the 19 systems. In general, the water quality was maintained from the outlet of the chlorinated water storage tank to the end of the network. The temperature range is typical for tropical countries and the pH values were close to 7. The turbidity and color of all samples were relatively low, indicating that the efficiency of the treatments and/or that the water sources were good. Similarly, in most cases, TOC and DOC were quite low. Moreover, UV$_{254}$ indicates a low presence of humic substances, and SUVA, in most cases less than 2 L/mg·m, suggests non-humic NOM and low molecular weight aliphatic compounds. Furthermore, only slight seasonal variation was found in the water NOM-related parameters (Fig. S5, Online Resource 1).

The low values in the above parameters related to NOM justify the low concentrations of TTHMs, where only two samples slightly exceeded the 80 μg/L regulated by the US EPA (US EPA 1998), despite the relatively high free chlorine (within the local regulation, i.e., 0.3 to 0.6 mg/L). Moreover, chloroform, even though at low concentration (10.60 ± 13.86 μg CHCl$_3$/L), in most of the samples accounted for around 62% of the different THM species. In addition, the species CHBrCl$_2$, CHBr$_2$Cl, and CHBr$_3$ were frequently found, but at much lower concentrations (i.e., < 2 μg/L). Such speciation of THMs has been reported in other studies (Sérodes et al. 2003). In general, in all the parameters (except in pH and free residual chlorine), surface water values at least double spring water ones, and the mixed and the whole data set values were in between. That is expected as surface water is highly influenced by allochthonous and autochthonous production, and the effect is also observed in the whole and the mixed water data sets. Furthermore, the higher concentration of precursor (e.g., TOC, UV$_{254}$) is reflected in higher THM concentration.

Correlation of independent variables with THMs in treated water

The Anderson–Darling statistical test (Ryan 2007) showed that the dependent (TTHM concentrations) and most of the

Table 1 Water characteristics of the data sets

Parameters	Whole data set ($N=216$)	Spring water data set ($N=70$)	Surface water data set ($N=67$)	Mixed water data set ($N=79$)
Temperature (°C)	22.0[a] ± 5.7[b] 17.0[c]–32.1[d]	22.0 ± 5.9[b] 17.8–31.3[d]	23.6 ± 6.0 17.0–31.1	21.1 ± 4.1 17.9–32.1
pH	7.30 ± 1.01 5.94–8.17	7.30 ± 0.92 6.00–7.94	7.27 ± 0.80 6.19–7.90	7.30 ± 1.08 5.94–8.17
Turbidity (NTU)	0.35 ± 0.79 < 0.01[e]–7.88	0.18 ± 0.27 < 0.01[e]–1.53	0.87 ± 1.35 < 0.01[e]–6.66	0.46 ± 0.68 < 0.01[e]–7.88
Apparent color (U Pt–Co)	3.15 ± 8.64 < 0.01[e]–31.30	0.01 ± 4.44 < 0.01[e]–11.00	4.67 ± 17.23 < 0.01[e]–31.30	4.11 ± 7.19 < 0.01[e]–30.97
Free residual chlorine (mg/L)	0.45 ± 0.37 < 0.02[e]–1.64	0.41 ± 0.34 < 0.02[e]–1.64	0.53 ± 0.41 0.04–1.64	0.43 ± 0.38 < 0.02[e]–1.13
TOC (mg/L)	0.50 ± 0.38 0.16–4.81	0.35 ± 0.18 0.16–2.42	0.79 ± 0.58 0.32–4.81	0.51 ± 0.23 0.18–3.52
DOC (mg/L)	0.48 ± 0.37 0.10–4.74	0.30 ± 0.22 0.10–2.42	0.66 ± 0.56 0.25–4.74	0.45 ± 0.28 0.10–3.47
UV$_{254}$ (cm^{-1})	0.0082 ± 0.0093 0.0004–0.0861	0.0046 ± 0.0034 0.0004–0.0478	0.0153 ± 0.0245 0.0043–0.0861	0.0091 ± 0.0078 0.0017–0.0829
SUVA (L/mg·m)	1.99 ± 1.31 0.15–14.06	1.58 ± 1.14 0.15–14.06	2.31 ± 1.44 0.82–4.74	2.25 ± 1.65 0.26–8.88
TTHM (μg/L)	10.64 ± 15.24 < 0.20[e]–91.31	7.22 ± 6.79 < 0.20[e]–24.62	19.91 ± 27.90 < 0.20[e]–91.31	10.65 ± 17.12 < 0.20[e]–65.45

[a] Median

[b] Interquartile range ($IQR = Q_3 - Q_1$)

[c] Minimum

[d] Maximum

[e] Detection limit

independent variables presented a non-normal distribution across all data sets (p-value < 0.05) (Table S1, Online Resource 1). This is expected because the data comes from systems with different operational characteristics. The data presented a positively skewed distribution, which is characterized by having a large amount of data in the low ranges of the parameter compared to the higher ranges. Therefore, to evaluate the correlation between the variables, Spearman's non-parametric test was used (Kurajica et al. 2020).

Temperature and pH showed non-significant and weak correlations (p-value > 0.05, $r_s < 0.3$) in all data sets (Table 2), expected as both parameters were relatively stable (Table 1). This differs from those reported by Al-Tmemy et al. (2018) for treated water from five treatment plants in Iraq where they found a significant and moderate correlation for both parameters. Accordingly, an increase in temperature tends to increase the reaction rate between organic matter and chlorine, and the THM concentrations increase with pH because many hydrolysis reactions, which occur in basic medium, promote their formation.

Turbidity presented a weak correlation in all data sets ($r_s < 0.3$) and was significant (p-value < 0.05) only in the whole data set and surface water data set (Table 2). Tsitsifli and Kanakoudis (2020) reported a greater correlation between turbidity and TTHMs ($r = 0.553$) for two treatment plants using surface sources. About apparent color, a low and significant positive correlation in the surface water data set was observed; in the others, the correlation was not significant (Table 2). Abdel Azeem et al. (2014) reported that Pearson correlation coefficient between THMs and color was between 0.87 and 0.93 for treated water at four treatment plants in Egypt.

Free residual chlorine showed a significant correlation in the whole data set and the spring and mixed water data sets (Table 2). In addition, the correlation was moderate and positive in all data sets. Contrary, some authors reported negative correlations between this parameter and TTHMs (Feungpean et al. 2015; Kumari and Gupta 2015). This inverse correlation can be attributable to radial diffusion and wall consumption of residual chlorine while THMs form (Kumari and Gupta, 2015). However, similar to the present study, positive and significant correlations have been attributed to the covariance of operational parameters or interactions between parameters (Salam et al. 2020).

The NOM, TOC, and DOC presented a moderate positive correlation ($0.3 < r_s < 0.7$) and significant (p-value < 0.05) in all data sets (Table 2), which agrees with the correlation values reported by several authors between 0.47 and 0.57 (Kumari and Gupta 2015; Shahi et al. 2020). Considering that chlorine reacts with NOM to produce THMs, the trend is that as TOC and DOC increase, the concentration of THM increases, as long as sufficient free residual chlorine is available (Kumari and Gupta 2015). Also, it was found that UV_{254} presented a significant and moderate positive correlation in the whole data set and surface water data set; however, in the other data sets, the correlation was weak and not significant. Similar, significant, and moderate observations were reported by other researchers for UV_{254} and THMs

Table 2 Spearman correlation between TTHMs and the independent variables

Parameters	Statistic	Whole data set (N = 216)	Spring water data set (N = 70)	Surface water data set (N = 67)	Mixed water data set (N = 79)
Temperature	r_s	0.042	−0.072	0.150	−0.051
	p-value	0.548	0.568	0.242	0.654
pH	r_s	−0.010	−0.034	−0.065	0.108
	p-value	0.884	0.785	0.613	0.347
Turbidity	r_s	0.146	0.050	0.321	−0.182
	p-value	0.036	0.687	0.010	0.111
Apparent color	r_s	0.135	0.189	0.164	−0.275
	p-value	0.058	0.128	0.199	0.023
Free residual chlorine	r_s	0.392	0.432	0.220	0.489
	p-value	< 0.001	< 0.001	0.083	< 0.001
TOC	r_s	0.454	0.330	0.325	0.380
	p-value	< 0.001	0.007	0.009	< 0.001
DOC	r_s	0.492	0.366	0.370	0.430
	p-value	< 0.001	0.003	0.003	< 0.001
UV_{254}	r_s	0.337	0.224	0.357	0.113
	p-value	< 0.001	0.071	0.004	0.325
SUVA	r_s	0.014	−0.109	0.104	−0.256
	p-value	0.842	0.386	0.417	0.024

(Semerjian et al. 2009; Kumari and Gupta 2015). Finally, the SUVA only presented a significant, but low negative correlation in the mixed water data set (Table 2). Other studies have reported low and negative correlations for SUVA, but not significant (Babaei et al. 2015).

Modeling THM formation within the distribution system

As shown in Table 3, linear, logarithmic, and exponential models were developed for each type of water. All models were significant (p-value < 0.05 of F-test), and in most cases, the Durbin-Watson value was found between 1.5 and 2.5 as recommended in the literature to avoid autocorrelation problems (Tsitsifli and Kanakoudis 2020). The models presented a wide range of adjusted R^2, from 0.132 to 0.687 indicating a varied performance and adjustment of the data.

The most appropriated models (in bold in Table 3) were selected not only because of the values of the coefficient of determination but also for statistical parameters related to the error (i.e., SE, MSE, MAE). For the whole data set, spring and mixed water data sets, the models 1, 4, and 10, respectively, presented the lowest values of SE, MSE, and MAE and they were selected although they presented a slightly lower R^2. However, in these models, the R^2 of 0.448, 0.657, and 0.531, respectively (Table 3), remain satisfactory and comparable to those reported by several authors (Babaei et al. 2015; Feungpean et al. 2015; Tsitsifli and Kanakoudis 2020). In the surface water data set, model 7 presented the lowest value of SE, MSE, and MAE, and the highest value of R^2 (Table 3). Therefore, models 1, 4, 7, and 10, all linear, were selected as the ones with the best performance and goodness-of-fit. Among those models, a greater goodness-of-fit is observed in those of spring waters (of higher quality) followed by the model of the mixed water data set, then the model of the whole data set and lower performance in the case of the surface water data set. In general, those models can be considered moderately robust and could be improved by including some parameters and operational variables that affect the formation of THMs in distribution networks (e.g., bromide ion, contact time, chlorine dose) (Nikolaou et al. 2004).

Through a more detailed analysis of each of the chosen models, it can be determined which are the most influential variables in the formation of THMs by type of water source. Thus, model 1, similar to models reported by Kumari and Gupta (2015), includes the variables pH, free residual chlorine, DOC, and UV_{254}. In the case of the spring water data set, model 4, free residual chlorine, DOC, and turbidity were included; the latter variable has also been used in THM prediction models (Al-Tmemy et al. 2018). Finally, in the surface and mixed water data sets, models 7 and 10, free

Table 3 TTHM predictive models for various data sets

Data set	Model		R^2	Adjusted R^2	F-test (p-value)	SE	MSE	MAE	Durbin-Watson	n
Whole data	**1**	Linear:$\mathrm{TTHM} = (5.10 - 0.618 \bullet \mathrm{pH} + 2.791 \bullet \mathrm{Cl} + 1.703 \bullet \mathrm{DOC} + 23.5 \bullet \mathrm{UV}_{254})^2$	**0.448**	**0.429**	**25.43 (<0.001)**	**8.67**	**74.93**	**6.58**	**1.54**	**131**
	2	Logarithmic:$\mathrm{TTHM} = 10^{4.149} \bullet \mathrm{pH}^{-3.028} \bullet \mathrm{Cl}^{0.5984} \bullet \mathrm{DOC}^{0.7598}$	0.568	0.557	52.14 (<0.001)	11.94	142.60	7.22	1.59	123
	3	Exponential:$\mathrm{TTHM} = e^{(3.625-0.318 \bullet \mathrm{pH}+1.496 \bullet \mathrm{Cl}+0.5818 \bullet \mathrm{DOC})}$	0.331	0.314	19.42 (<0.001)	14.08	198.12	8.83	1.77	122
Spring water	**4**	Linear:$\mathrm{TTHM} = 1.826 + 4.76 \bullet \mathrm{T} + 6.84 \bullet \mathrm{Cl} + 2.686 \bullet \mathrm{DOC}$	**0.657**	**0.628**	**22.97 (<0.001)**	**2.32**	**5.38**	**1.93**	**1.56**	**40**
	5	Logarithmic:$\mathrm{TTHM} = 10^{5.812} \bullet \mathrm{pH}^{-5.598} \bullet \mathrm{Cl}^{0.6366} \bullet \mathrm{DOC}^{0.3077} \bullet \mathrm{UV}_{254}{}^{-0.1356}$	0.718	0.687	22.91 (<0.001)	2.78	7.73	2.08	1.58	41
	6	Exponential:$\mathrm{TTHM} = e^{\left(4.124-0.315 \bullet \mathrm{pH}+0.913 \bullet \mathrm{Cl}-15.14 \bullet \mathrm{UV}_{254}\right)}$	0.281	0.224	4.94 (0.005)	4.82	23.23	3.35	1.42	42
Surface water	**7**	Linear:$\mathrm{TTHM} = 0.52 + 14.03 \bullet \mathrm{Cl} + 14.28 \bullet \mathrm{DOC}$	**0.342**	**0.306**	**9.60 (<0.001)**	**12.22**	**149.33**	**10.09**	**1.84**	**40**
	8	Logarithmic:$\mathrm{TTHM} = 10^{1.4938} \bullet \mathrm{Cl}^{0.416} \bullet \mathrm{DOC}^{0.656}$	0.328	0.291	8.78 (0.001)	13.92	193.82	10.77	0.84	39
	9	Exponential:$\mathrm{TTHM} = e^{(2.065+0.776 \bullet \mathrm{Cl}+0.359 \bullet \mathrm{DOC})}$	0.177	0.132	3.97 (0.027)	14.67	221.06	11.19	1.61	40
Mixed water	**10**	Linear:$\mathrm{TTHM} = -19.17 + 23.97 \bullet \mathrm{Cl} + 35.66 \bullet \mathrm{TOC}$	**0.531**	**0.508**	**23.23 (<0.001)**	**6.71**	**45.04**	**5.50**	**2.12**	**44**
	11	Logarithmic:$\mathrm{TTHM} = 10^{-5.47} \bullet \mathrm{pH}^{8.12} \bullet \mathrm{AP}^{-02282} \bullet \mathrm{Cl}^{0.788} \bullet \mathrm{DOC}^{1.482}$	0.630	0.593	17.04 (<0.001)	21.96	482.40	10.80	1.74	45
	12	Exponential:$\mathrm{TTHM} = e^{(-8.71+1.084 \bullet \mathrm{pH}+3.366 \bullet \mathrm{Cl}+0.964 \bullet \mathrm{DOC})}$	0.501	0.466	14.39 (<0.001)	18.94	358.57	11.88	2.14	47

TTHM total trihalomethanes (μg/L), *Cl* free residual chlorine (mg/L), UV_{254} ultraviolet absorption at 254 nm (cm^{-1}), *TOC* total organic carbon (mg/L), *DOC* dissolved organic carbon (mg/L), *AP* apparent color (U Pt–Co), *T* turbidity (NTU) residual chlorine and organic matter content such as DOC and TOC, respectively, are observed as influential.

Validation of THM models

Table 4 presents the validation results, R^2, SE, MSE, and MAE, as well as the results of the *T*-test for each model. The values of R^2 were between 0.359 and 0.772, which demonstrated a satisfactory level of explanation of the observed variability and are comparable with those reported by Golfinopoulos and Arhonditsis (2002) (i.e., 0.37 to 0.54). Similar to the calibration phase, SE, MSE, and MAE results showed that models 4 and 10 (spring and mixed water, respectively) performed better. Also, the bias of the four models determined by a *T*-test (Shahi et al. 2020) indicated no statistically significant difference between the predicted and measured average values (p-value > 0.05; Table 4). Furthermore, Fig. 2 shows that most of the data are within the prediction interval for all the models. In the case of the whole data set and surface water (Fig. 2a and c) the data tend to move away from the line of best fit above 30 µg/L. In the case of the models for spring water and the mixed water (Fig. 2b and d), with lower TTHM concentrations, the data tend to distribute more evenly. Therefore, these models seem to perform better at TTHM concentrations lower than 30 µg/L.

Conclusions

Several TTHM models were developed for the tropical Costa Rican rural and semi-urban chlorinated water. The TTHM concentrations ranged between < 0.20 and 91.31 µg/L with chloroform ($CHCl_3$) accounting on average for 62% of the total. Depending on the data set, several parameters, including turbidity, total organic carbon (TOC), dissolved organic carbon (DOC), free residual chlorine, and ultraviolet absorbance at 254 nm (UV_{254}), presented significant correlation (p-value < 0.05). Four linear models presented the best goodness-of-fit and were moderately robust. From the validation stage, it was found that according to the criteria of R^2, standard error (SE), mean square error (MSE), and mean absolute error (MAE), spring water and mixed spring/

Table 4 Validation of proposed models for the prediction of TTHMs in the distribution systems

Data set	Model	*N*	R^2	SE	MSE	MAE	*t*-value	*p*-value	Significance
Whole data	1	50	0.393	10.05	101.05	6.95	0.56	0.576	No
Spring water	4	14	0.598	2.83	8.03	2.31	– 1.05	0.303	No
Surface water	7	17	0.359	15.84	250.85	11.20	0.96	0.346	No
Mixed water	10	16	0.772	4.40	19.33	3.50	0.44	0.665	No

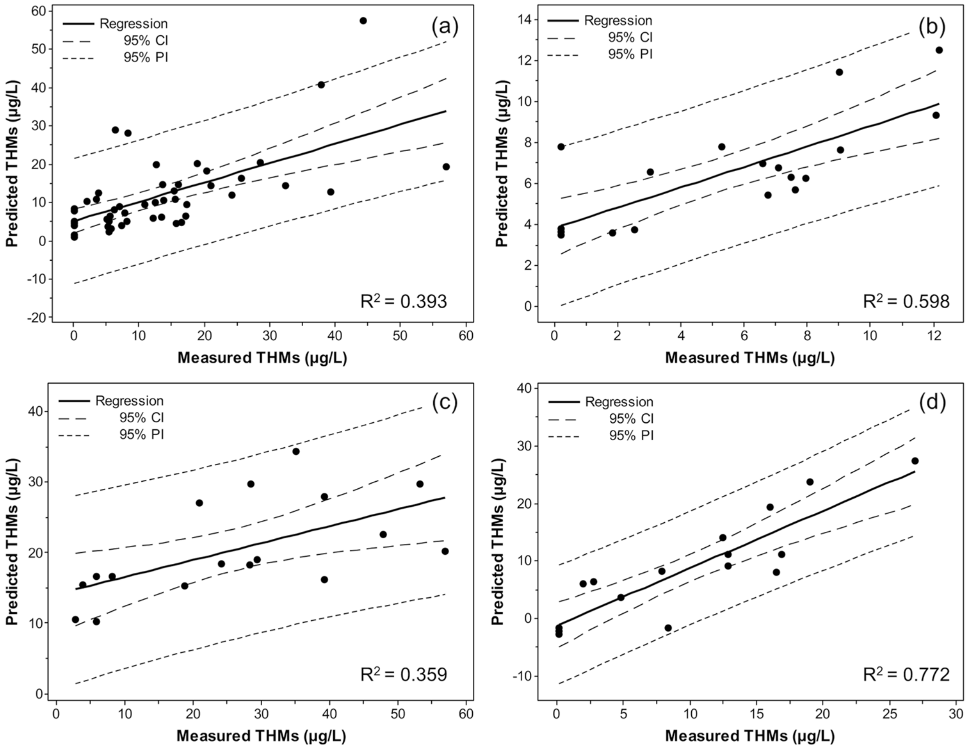

Fig. 2 Validation of models and comparison of measured vs. predicted TTHM concentrations for (**a**) model 1, (**b**) model 4, (**c**) model 7, and (**d**) model 10. *CI* confidence interval, *PI* prediction interval

surface water models showed a satisfactory level of explanation of the variability of the data. Moreover, all the models seem to better predict TTHM concentrations below 30 μg/L. Therefore, considering the specific chlorinated water characteristics (low NOM and TTHMs produced) the models developed could be useful for decision-making in drinking water supply systems.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-021-18299-0.

Acknowledgements The authors are thankful to Manuel Rodríguez (Laval University), Guillermo Calvo (ITCR), and Nirmal Kumar Shahi (Dankook University) for their valuable guidance about the model's development.

Author contribution KC collaborated on field sampling and analysis, analyzed and interpreted the data, initial ideas, and development of the models, wrote the initial draft, and wrote, reviewed, and edited the final manuscript. MC, SJ, SN, and PG collaborated on field sampling and analysis. GJ collaborated on THM analysis and data quality. JA contributed to TOC, DOC, and UV254 analysis and data quality. RE collaborated on field sampling and analysis, initial ideas of the research, the methodology, and model design, and wrote, reviewed, and edited the final manuscript. All authors read and approved the final manuscript.

Funding This study was funded by Consejo Nacional de Rectores (CONARE), National University of Costa Rica, National Technical University, and Instituto Tecnológico de Costa Rica.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

Abdel Azeem SM, Burham N, Borik MG, El Shahat MF (2014) Trihalomethanes formation in water treatment plants and distribution lines: a monitoring and modeling scheme. Toxicol Environ Chem 96:12–26. https://doi.org/10.1080/02772248.2014.922565

Acuña-Fernández E (2004) Regression analysis. Department of Mathematics, Universidad de Puerto Rico, Mayaguez (in Spanish)

Al-Tmemy WB, Alfatlawy YF, Khudair SH (2018) Seasonal variation and modeling of disinfection by-products (DBPs) in drinking water distribution systems of Wassit Province Southeast Iraq. J Pharm Sci Res 10:3393–3399

APHA, AWWA, WEF (2017) Standard methods for the examination of water and wastewater, 23rd edn. America Public Health Association, Washington, DC

Arellano-Hartig F, Garita-Incer A, González-Jiménez A, García-Fernández R, Quesada-Rodríguez J, Villalobos-Jiménez A (2020) National diagnosis of operating entities (survey 2017–2020). AyA, San José (in Spanish)

AyA (2016) National policy for the drinking water subsector of Costa Rica 2017–2030. Comisión Interinstitucional, San José (in Spanish)

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. https://doi.org/10.2166/wh.2015.211

Chowdhury S, Champagne P, James McLellan P (2008) Factors influencing formation of trihalomethanes in drinking water: results from multivariate statistical investigation of the Ontario drinking water surveillance program database. Water Qual Res J 43:189–199. https://doi.org/10.2166/wqrj.2008.022

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. https://doi.org/10.1016/j.scitotenv.2009.04.006

Costet N, Villanueva CM, Jaakkola JJK, Kogevinas M, Cantor KP, King WD, Lynch CF, Nieuwenhuijsen MJ, Cordier S (2011) Water disinfection by-products and bladder cancer: is there a European specificity? A pooled and meta-analysis of European caseecontrol studies. Occup Environ Med 68:379–385. https://doi.org/10.1136/oem.2010.062703

Edzwald JK, Tobiason JE (2011) Chemical principles, source water composition, and watershed protection. In: Edzwald JK (ed) Water Quality & Treatment: A Handbook on Drinking Water, 6th edn. McGraw-Hill Education, New York, p 3.1-3.76

Feungpean M, Panyapinyopol B, Elefsiniotis P, Fongsatitkul P (2015) Development of statistical models for trihalomethane (THM) occurrence in a water distribution network in Central Thailand. Urban Water J 12:275–282. https://doi.org/10.1080/1573062X.2013.871042

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. https://doi.org/10.1016/S0045-6535(02)00058-9

IARC (2021) Agents classified by the IARC monographs. https://monographs.iarc.who.int/list-of-classifications. Accessed 7 Oct 2021

Kargaki S, Iakovides M, Stephanou EG (2020) Study of the occurrence and multi-pathway health risk assessment of regulated and unregulated disinfection by-products in drinking and swimming pool waters of Mediterranean cities. Sci Total Environ 739:139890. https://doi.org/10.1016/j.scitotenv.2020.139890

Kumari M, Gupta S (2015) Modeling of trihalomethanes (THMs) in drinking water supplies: a case of study of eastern part of India. Environ Sci Pollut Res 22:12615–12623. https://doi.org/10.1007/s11356-015-4553-0

Kurajica L, Ujević Bošnjak M, Novak Stankov M, Kinsela AS, Štiglić J, Waite DT, Capak K (2020) Disinfection by-products in Croatian drinking water supplies with special emphasis on the water supply network in the city of Zagreb. J Environ Manage 276:111360. https://doi.org/10.1016/j.jenvman.2020.111360

Manso P, Ortiz W, Fallas J (2005) The precipitation regime in Costa Rica. Ambientico 144:7–8 (in Spanish)

Mazhar MA, Khan NA, Ahmed S, Khan AH, Hussain A, Rahisuddin CF, Yousefi M, Ahmadi S, Vambol V (2020) Chlorination disinfection by-products in municipal drinking water — a review. J Clean Prod 273:123159. https://doi.org/10.1016/j.jclepro.2020.123159

MINSA (2018) Regulation for the quality of drinking water. Diario Oficial La Gaceta 69:1–49

Mukundan R, Van Dreason R (2014) Predicting trihalomethanes in the New York City water supply. J Environ Qual 43:611–616. https://doi.org/10.2134/jeq2013.07.0305

Nikolaou AD, Golfinopoulos SK, Arhonditsis GB, Kolovoyiannis V, Lekkas TD (2004) Modeling the formation of chlorination by-products in river waters with different quality. Chemosphere 55:409–420. https://doi.org/10.1016/j.chemosphere.2003.11.008

Pardoe I (2012) Applied regression modeling. John Wiley & Sons, New Jersey

PEN, CONARE (2020) Chapter 10: balance 2020 harmony with nature. In: State of the Nation Report. San José, pp 339–380 (in Spanish)

Rahman MB, Driscoll T, Cowie C, Armstrong BK (2010) Disinfection by-products in drinking water and colorectal cancer: a meta-analysis. Int J Epidemiol 39:733–745. https://doi.org/10.1093/ije/dyp371

Richardson SD, Plewa MJ (2020) To regulate or not to regulate? What to do with more toxic disinfection by-products? J Environ Chem Eng 8:103939. https://doi.org/10.1016/j.jece.2020.103939

Rodriguez MJ, Sérodes J, Morin M (2000) Estimation of water utility compliance with trihalomethane regulations using a modelling approach. J Water Supply Res Technol - AQUA 49:57–73. https://doi.org/10.2166/aqua.2000.0006

Ryan TP (2007) Modern engineering statistics. John Wiley & Sons, New Jersey

Sadiq R, Rodriguez MJ, Mian HR (2019) Empirical models to predict disinfection by-products (DBPs) in drinking water: an updated review. In: Nriagu J (ed) Encyclopedia of Environmental Health. Elsevier, Oxford, pp 324–338

Salam E, Bassam AF, Salwan A, Nadhir AA (2020) Modeling of trihalomethane compounds formation in Baghdad water supply network. Sci Rev Eng Environ Sci 29:136–144 . https://doi.org/10.22630/PNIKS.2020.29.2.12

Sánchez-Hernández L (2019) Current situation of community organizations that provide drinking water and sanitation services. State of the Nation Program, San José (in Spanish)

Semerjian L, Dennis J, Ayoub G (2009) Modeling the formation of trihalomethanes in drinking waters of Lebanon. Environ Monit Assess 149:429–436. https://doi.org/10.1007/s10661-008-0219-4

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. https://doi.org/10.1016/S0045-6535(02)00840-8

Shahi NK, Maeng M, Dockko S (2020) Models for predicting carbonaceous disinfection by-products formation in drinking water treatment plants: a case study of South Korea. Environ Sci Pollut Res 27:24594–24603. https://doi.org/10.1007/s11356-019-05490-7

Tsitsifli S, Kanakoudis V (2020) Developing THMs' predictive models in two water supply systems in Greece. Water (switzerland) 12:1422. https://doi.org/10.3390/w12051422

Epa US (1998) National primary drinking water regulations: disinfectants and disinfection byproducts. Fed Reg 63:69390–69476

Wright JM, Evans A, Kaufman JA, Rivera-Núñez Z, Narotsky MG (2017) Disinfection by-product exposures and the risk of specific cardiac birth defects. Environ Health Perspect 125:269–277. https://doi.org/10.1289/EHP103