Environmental Science
Water Research & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Water Res. Technol.*, 2023, **9**, 2745

Predictive capability of THM models for drinking water treatment and distribution†

Derek Hogue, *[a] Pitu B. Mirchandani[b] and Treavor H. Boyer*[a]

Received 28th April 2023,
Accepted 21st August 2023

DOI: 10.1039/d3ew00308f

rsc.li/es-water

Research and practice suggest markers of drinking water quality such as trihalomethanes (THMs), can change during treatment and distribution, potentially elevating health risk of end users. Models have been developed to predict THM formation at drinking water treatment plants (DWTP), in drinking water distribution systems (DWDS), and to a lesser extent, building premise plumbing (PP). The goal of this research was to evaluate the performance of published THM models and their development methodology, with the purpose of improving future THM model development. Water quality variable data were collected from literature and used as inputs for collected models. Mean and variance of model prediction values were used to measure THM model performance compared to THM data trends from literature. The research found differences in model formulation, water quality variable selection, and model development practices, despite evaluated models being statistical in nature. These differences lead to substantial inconsistencies in model output behavior. Diversity of data used for model development was found to be the most important factor for generalizable model prediction capabilities. Following these findings, a new framework was proposed to encourage novel strategies, data sharing, and collaboration among researchers and practitioners to improve THM model development, application, and performance. Potential use of machine learning techniques for future model development was also discussed based on findings.

Water impact

The potential health risks of disinfection byproducts (DBPs) are a primary concern within the scope of drinking water treatment and distribution. Regulated DBPs including trihalomethanes (THMs) are of particular importance due to regulatory and carcinogenicity concerns. It has been demonstrated that THM concentrations can increase during drinking water distribution, and ultimately cause increased health risk to end users. This problem may be enhanced in green buildings as lower water use leads to greater stagnation, an increased THM formation. THM models have been useful for predicting changes in THM concentrations during water treatment and distribution, however there has been limited development for premise plumbing application due to greater challenges imposed by differences in physiochemical phenomena influencing THM formation. Further, THM model development for the past 30 years has focused primarily on statistical models fitted for system specific data. This research evaluates the generalizability of recent regression based THM models to identify useful strategies for future THM model development. Further, it offers a framework that promotes a more cohesive system of data and model development reporting that aims to facilitate greater progress and support novel data-science based approaches to the challenges introduced by premise plumbing systems in particular. Accurate prediction of THM formation in premise plumbing will allow us to promote sustainable water management practices while also considering the associated health implications.

1. Introduction

Ensuring drinking water quality is crucial for maintaining public health. Disinfection of drinking water is an important step of drinking water treatment that ensures inactivation of pathogens. However, disinfectants such as chlorine (Cl_2), can form carcinogenic disinfection byproducts (DBP), such as trihalomethanes (THMs), increasing risk to consumers. The regulation of THMs by the USEPA underlines the significance of maintaining safe THM concentrations in drinking water distribution systems (DWDS) for public health. There is an expanding body of research that demonstrates changes in residual Cl_2 and increases in THMs can occur during the treatment and distribution of drinking water.[1,2] Recent innovations in building design that promote water efficiency may also contribute to degraded drinking water quality.[3] When not appropriately accounted for, lower water use in "green" buildings can create increased water stagnation times, leading to lower chlorine concentration and higher disinfection byproduct formation.[4–7] Understanding how THMs change at each point during water treatment and

[a] *School of Sustainable Engineering and the Built Environment (SSEBE), Arizona State University, PO Box 873005, Tempe, AZ 85287-3005, USA.*
E-mail: dahogue@asu.edu, thboyer@asu.edu; Tel: +1 480 965 7447
[b] *School of Computing and Augmented Intelligence (SCAI), Arizona State University, 699 S Mill Ave, Tempe, AZ 85281, USA*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3ew00308f

distribution is vital for maintaining public health. Although *in situ* water quality sensing capabilities are available, it is often too costly or not possible to measure certain water quality variables, such as THMs, in real-time. Model-based estimation of water quality variables may be a solution to the feasibility challenges faced by drinking water treatment plants (DWTP), DWDS, and building plumbing especially when combined with real-time sensor data.

Numerous models have been developed with the objective of predicting THM based on water quality variables attributed to their formation. Model development is referred to as the methodology used to generate models based on research goals, and generated models are referred to as the final published model or models of a study. Most models have been statistical in formulation where data is used to fit independent explanatory variables (water quality variables), to dependent variables (*i.e.*, THM), based on correlations in the data. Data used for fitting the models is referred to as training data in this research, and data used for model validation is referred to as test data. Replicative validation is the practice of evaluating model accuracy with training data, while predictive validation is the practice of evaluating model accuracy with data not used for model development. Model evaluation is referred to as the process of measuring the outputs of a model compared to an expected output. Generalizability is the ability of a model to produce a reasonable output given new or unseen data from the same type of application (*e.g.*, DWDS model given new or unseen DWDS data). Generalizable models in this context would predict similar THM concentrations for different systems within the same application, given the appropriate water quality variable data. A considerable body of research has been conducted in the past 30 years dedicated to DBP water quality modeling at drinking water treatment plants (DWTP), and within DWDS.[8,9] However, there has been limited model development for prediction of THM in building premise plumbing (PP). Studies suggest that THM and other water quality variables (*e.g.*, Cl_2, UV_{254}, and cATP), can change significantly from DWDS to PP.[10–13] Therefore, PP models were considered in this study despite the lack of models within literature.

The performance of THM models developed for one application and applied to another application (*e.g.*, DWTP data applied to DWDS model), are not known since there are considerable differences in system conditions.[4] Further, THM models generated since 2010 have been primarily statistical in formulation,[8] where formulation refers to the mathematical basis of the model. Although statistical models are useful tools, they may not provide widespread applicability to different systems due to the application specific nature of their development. A preliminary review of recent THM models found that research was inconsistent with explanatory variable usage, data preparation, statistical analysis, and validation of the models. More broadly, there was a limited accessibility to the data and methods used for model development. Easily accessible data not only provides transparency, but also provides the opportunity to develop and employ more generally applicable models through the use of more diverse data sets. For accurate prediction within real drinking water systems, water age, flow conditions, and seasonal variations in water quality variables are important considerations; however, many models do not consider these factors. There is a need for investigation and understanding of water quality modeling practices which may benefit future THM model development.

The goal of the research was to evaluate existing THM models to understand how model development (*i.e.*, variable selection, statistical assessment, data collection) impacted prediction capability and generalizable application of the models. The desired outcome was to inform and advance future THM model development. The specific objectives were to: (1) compile models developed for THM prediction in DWTP, DWDS, and PP systems, (2) compile data from literature of common water quality variables used in THM models for various conditions, (3) apply the water quality variables to the collected models and quantitatively evaluate performance, (4) identify characteristics that impact variance of model outputs, and (5) make recommendations for future model development.

2. Research approach

2.1 THM models

A literature review of THM models was conducted using Scopus and Google Scholar.[14–25] Search results were restricted to include articles with titles or abstracts containing the keywords "THM"/"trihalomethanes", "model", and "water". Results were restricted to include at least one secondary term, such as; "DBP", "chlorine", "treatment plant", "distribution system", or "premise plumbing" to be included within the scope of the search. Using both inclusion criteria narrowed the search to more relevant articles. All searches were limited to papers published after 2009 to focus on most recent advances in model development not covered by prior studies.[8,9] Studies with models developed for natural waters with high NOM, alkalinity concentrations (Alk), or other characteristics not commonly found in drinking water systems were not included. Mechanistic models that included unique or system specific parameters were excluded from this study. Multiple models were selected from the same publication if there were significant differences in modeling approach, formulation, and/or application. Fourteen models were found that fit the search criteria. Models were divided into three categories based on their intended application: DWTP, DWDS, or PP.

2.2 Water quality variable data

Water quality variable data were compiled through a literature review in Scopus.[18,19,23,26–39] Search results were restricted to articles with titles or abstracts containing combinations of the words "trihalomethanes", "drinking water", and/or "model". Searches were further classified

based on the inclusion of "treatment plant", "distribution system", "distribution network", "building plumbing", or "premise plumbing". Data inclusion was limited to articles with reported number of samples (n), mean, and standard deviation (SD) values. Exceptions were made in the case of limited available data for a particular water quality variable, reported ranges were used by assuming SD = (max − min)/6. There was no restriction on the year of publication for water quality variable data. Data were categorized based on application: DWTP or DWDS. There was insufficient data in literature to create a complete set of water quality variable conditions for PP application.

The data were compiled in RStudio by combining normal distributions of each data entry using the rnorm() function. The data were cleaned to remove improbable extrema (*e.g.*, pH values were limited to values between 6.5 and 8.5, and negative concentration values were removed). The fitdistrplus package was used to determine the best fitting distribution for each set of water quality variable data, and corresponding distribution parameters. The descdist() function was used to produce Cullen and Fey plots for the water quality variable data. The plots aided in choosing distributions that provided the best fit of the water quality variable data. Distribution parameters for the associated distributions were determined using the fitdist() function using the maximum likelihood estimate. The function rtrunc() was used to produce Monte Carlo (MC) simulated data sets of size n = 100 000 based on the appropriate distribution type and associated parameters for each water quality variable data set. Simulated water quality variable data were analyzed using descriptive statistics including mean, SD, 90% confidence interval (CI), and coefficient of variation (CV), which was calculated as SD/mean.

2.3 Model evaluation

Fig. 1 illustrates the research approach outlined in this section. Model evaluation was conducted using the simulated water quality variable data sets as inputs for the 14 THM models. Data sets for both applications were applied to all models for unbiased comparison of model performance. From the MC simulation, each water quality variable consisted of n random values from the respective distribution described in section 2.2. Since the simulated water quality variable data sets were used as inputs for the models, and the resulting model outputs were data sets of equal size (n = 100 000). Descriptive statistics of model output data were presented, including mean predicted THM, SD, ratios of mean predicted THM concentration/mean THM concentration from collected data (THM_p/THM_m), and CV. Mean, variance, and skewness of model outputs were compared to mean, variance, and skewness of THM data to quantitatively rank model performance. Graphical representations for the model output data were produced to visually compare to the distribution of THM data from the literature. In addition to quantitative evaluation of model outputs, qualitative comparisons were made to better understand impactful aspects of model development.

3. Results and discussion

3.1 THM models

Of the 14 THM models compiled from the literature review, 8 were developed for DWTP application, 5 were developed for DWDS application, and 1 was developed for PP application. Table 1 shows model development information. The models were developed using some type of correlation test followed by a multiple linear regression method to

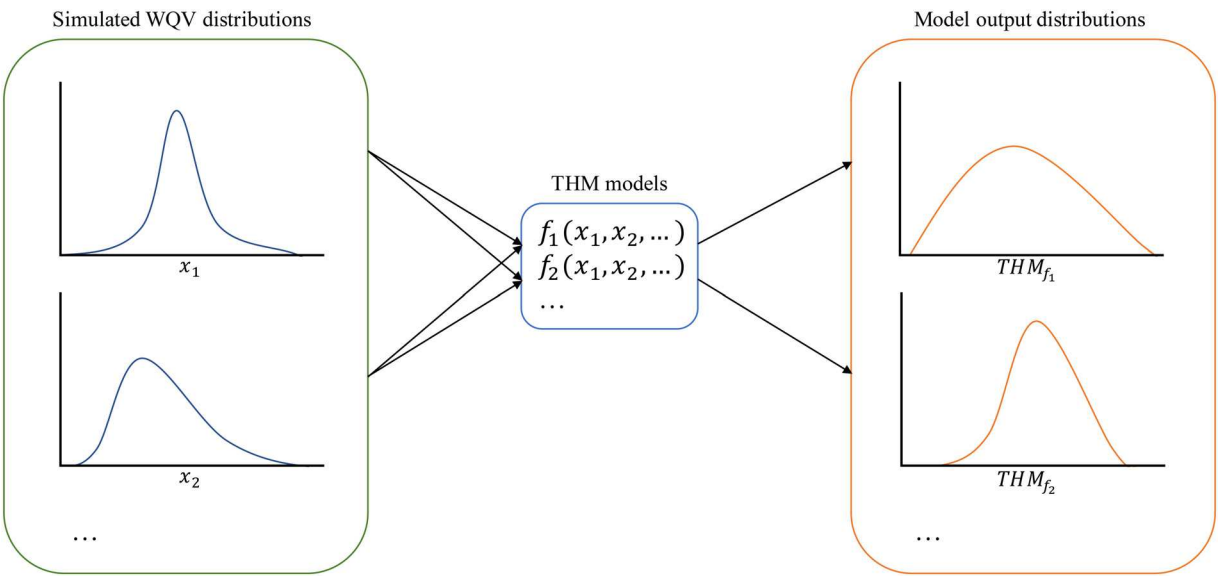

Fig. 1 Simplified schematic of research methodology for evaluating THM model performance. Curves represent probability density functions (pdf), for water quality variables (x_i), and predicted THM model output pdf (THM_i), for corresponding models (f_i).

Table 1 Development methodology of the evaluated THM models for DWTP, DWDS, and PP application. NA is used where replicative validation only was performed

Application	Model #	Source/year	Parameter estimation	Training/test%	Validation method	Sensitivity analysis
DWTP	1	Hong *et al.*, 2016 (ref. 14)	Stepwise multi linear regression	NA	Replicative with independent sample *t*-test	Not conducted
	2	Kumari and Gupta, 2015 (ref. 15)	Multi linear regression, Pearson correlation	Not discussed	Predictive with test data	Not conducted
	3	Kumari and Gupta, 2015 (ref. 15)	Multi linear regression, Pearson correlation	Not discussed	Predictive with test data	Not conducted
	4	Roth and Cornwell, 2018 (ref. 16)	Multi linear regression, Pearson correlation	NA	Replicative, and residual analysis	Not conducted
	5	Shahi *et al.*, 2020 (ref. 17)	Multi linear regression, Pearson correlation	Not discussed	Predictive with test and independent data	Not conducted
	6	Godo-Pla *et al.*, 2021 (ref. 18)	Multi linear regression with outlier detection	80/20	Predictive with test data	Differential based sensitivity analysis
	7	Domínguez-Tello *et al.*, 2017 (ref. 19)	Multi linear regression, Pearson correlation	Not discussed	Predictive, test and independent data	Not conducted
	8	Chen and Westerhoff, 2010 (ref. 20)	Multi non-linear regression, MSE correlation	NA	Replicative, RMSE	Independent variable sensitivity analysis
DWDS	9	Wert *et al.*, 2012 (ref. 21)	Linear correlation	NA	Predictive with independent data	Independent variable sensitivity analysis
	10	Cong *et al.*, 2012 (ref. 22)	Empirical parameter estimates based on experimental data	NA	Replicative, RMSE	Not conducted
	11	Osorio *et al.*, 2011 (ref. 23)	Multi linear regression, bivariate correlation, and one-way ANOVA test	NA	Replicative, RMSE	Not conducted
	12	Tsitsifli and Kanakoudis, 2020 (ref. 25)	*K–S* test, Pearson correlation estimate for multi linear regression	NA	Replicative, one-way ANOVA test	Not conducted
	13	Domínguez-Tello *et al.*, 2017 (ref. 19)	Multi linear regression, Pearson correlation	Not discussed	Predictive, test and independent data	Not conducted
PP	14	Chowdhury *et al.*, 2011 (ref. 24)	Significant factors analysis *via* numerical and graphical techniques	65/35	Predictive with test data	Not conducted

determine statistical coefficients, with the exception of models 9, 10, and 14. Only two of the eight models that conducted replicative validation discussed the training/test data split. Only three of the models conducted sensitivity analyses, and only model 6 development considered and removed outlier data.

Table 2 shows the model formulation, number of data, and data source. The number of water quality variables utilized ranged from 2 to 7 and the number of data used to fit the models ranged from 35 to 893. Predictive validation was conducted on 8 of the 14 models, while replicative validation was conducted on the remaining models (*i.e.*, goodness of fit was measured with training data only). Data from full-scale drinking water systems were used for development for 11 of the models, while the other 3 models were developed with bench-scale experimental data.

The distribution of models developed for the different applications investigated in this research highlighted the disparity between predictive THM models for PP application compared to DWTP and DWDS application. DWTP models are important for evaluating process selection and performance, and ensuring regulatory requirements are met, however they may not capture the true concentration of THM at the tap. DWDS models bridge the gap between DWTP and PP systems, but may not capture the significant changes in THM concentration within buildings.[6] Other THM models exist outside of the one used in this research; however they tend to be mechanistic in nature.[40–42] Mechanistic models require system-specific parameters for accurate prediction and may not capture the impact of omitted water quality variables on THM formation. For example, reaction rate coefficients have been used in some models to relate changes in water quality variable concentrations to THM formation.[22,43] Differences in physicochemical and biological characteristics between systems would affect the relative impact of water quality variables on THM formation, as well as the values of their reaction coefficients. For these reasons, mechanistic models may not be as useful for predicting water quality behavior in different systems. The difference in number of models developed for PP compared to other applications shows that disproportionate effort has gone toward developing THM models for DWTP and DWDS. To aid in future development of THM models, the collected models were analyzed for attributes that contributed to their predictive capabilities.

Table 2 Collected THM models developed for DWTP, DWDS, and PP application. Test data refers to data separated from training data prior to parameter fitting, and independent data refers to data from outside system(s)

Application	Model #	Source/year	Model	# WQV	# data	Data source
DWTP	1	Hong *et al.*, 2016 (ref. 14)	$\mathrm{THMs} = 10^{-2.534}(\mathrm{DOC})^{0.369}(\mathrm{Br})^{0.212}\left(\frac{\mathrm{Cl_{2,d}}}{\mathrm{DOC}}\right)^{0.400}(T)^{0.662}(\mathrm{pH})^{2.364}(\mathrm{RT})^{0.305}$	6	243	Bench scale experimental data
	2	Kumari and Gupta, 2015 (ref. 15)	$\mathrm{THM} = -150.833 + 40.948(\mathrm{pH}) + 6.153(T) - 13.876(\mathrm{Cl_{2,r}}) + 8.100(\mathrm{RT}) + 6.221(\mathrm{TOC}) + 292.308(\mathrm{UV_{254}})$	6	46	*In situ* DWTP and DWDS data
	3	Kumari and Gupta, 2015 (ref. 15)	$\mathrm{THM} = 33.436(\mathrm{pH})^{0.062}(T)^{0.069}(\mathrm{Cl_{2,r}})^{-0.048}(\mathrm{RT})^{0.018}(\mathrm{TOC})^{0.079}(\mathrm{UV_{254}})^{0.045}$	6	46	*In situ* DWTP and DWDS data
	4	Roth and Cornwell, 2018 (ref. 16)	$\mathrm{THM} = 10^{1.2146}(\mathrm{Cl_{2,d}})^{0.3897}(\mathrm{RT})^{0.3142}(\mathrm{UV_{254}})^{0.1381}$	3	66	Bench scale experimental data
	5	Shahi *et al.*, 2020 (ref. 17)	$\mathrm{THM} = 85.928 - 5.2\times10^{-4}(\mathrm{UV_{254}}\times\mathrm{DOC}\times\log(\mathrm{Cl_{2,d}}))^2 - 6.2\times10^{-2}(\mathrm{Br}+2) + 1.66\times10^{-5}(\mathrm{Cl_{2,r}})^2 + 3.87\times10^{-6}(\mathrm{Cl_{2,post}})^2 - 10.25(\mathrm{pH}) + 7\times10^{-3}(T)^2 + 8.42\times10^{-5}(\mathrm{UV_{254}}\times(T)^2\times\mathrm{RT}\times\mathrm{Cl_{2,d}})$	7	120	*In situ* DWTP data
	6	Godo-Pla *et al.*, 2021 (ref. 18)	$\mathrm{THM} = 6.18(\mathrm{UV_{254}}+1)^{3.64}(\mathrm{TOC})^{0.462}(\mathrm{Cl_{2,d}})^{0.420}(\mathrm{Br}+1)^{0.471}(T)^{0.169}(\mathrm{pH})^{0.048}(\mathrm{RT})^{0.298}$	7	573	*In situ* DWTP and DWDS data
	7	Domínguez-Tello *et al.*, 2017 (ref. 19)	$\mathrm{THM} = 165 - 21.3(\mathrm{pH}) + 0.232(\mathrm{Br}) + 5.84(\mathrm{Cl_{2,d}}\times\mathrm{RT}\times T\times\mathrm{UV_{254}})$	6	198	*In situ* DWTP and DWDS data
	8	Chen and Westerhoff, 2010 (ref. 20)	$\mathrm{THMFP} = 1147(\mathrm{UV_{254}})^{0.83}(\mathrm{Br}+1)^{0.27}$	2	210	*In situ* WTP data
DWDS	9	Wert *et al.*, 2012 (ref. 21)	$\mathrm{THM} = 0.035(\mathrm{TOC})^{1.098}(\mathrm{Cl_2})^{0.152}(T)^{0.609}(\mathrm{pH})^{1.601}(\mathrm{RT})^{0.263}$	5	172	*In situ* DWTP data
	10	Cong *et al.*, 2012 (ref. 22)	$\mathrm{THM} = (11.1(\mathrm{TOC}) + 20.06) - ((11.1(\mathrm{TOC}) + 20.06) - \mathrm{THM_0}) \times \exp\left(\frac{k_0 C_0}{7.5\times10^7(0.7(\mathrm{TOC}) - 2.2(C_0))\times\exp\left(\frac{-6500}{T}\right)}\times\left(\exp\left(-7.5\times10^7(0.7(\mathrm{TOC}) - 2.2(C_0))\times\exp\left(-\frac{6500}{T}\right)(\mathrm{RT})\right) - 1\right)\right)$	5	49	Bench scale experimental data
	11	Osorio *et al.*, 2011 (ref. 23)	$\sqrt{\mathrm{THM}} = -28.826 + 1.583(\mathrm{TOC}) + 2.713(\log(\mathrm{cond})) - 1.307(\log(\mathrm{bicarb})) + 3.744(\mathrm{Cl_2}) + 2.427(\mathrm{pH}) + 0.102(T)$	6	893	*In situ* DWDS data
	12	Tsitsifli and Kanakoudis, 2020 (ref. 25)	$\log(\mathrm{THM}) = -3.84 + 0.633(\mathrm{pH}) - 0.1056(\mathrm{TOC})^{-2}$	2	35	*In situ* DWDS data
	13	Domínguez-Tello *et al.*, 2017 (ref. 19)	$\mathrm{TTHM} = 14.9 + 1.01(\mathrm{TTHM_{Ef}}) + 0.20(\mathrm{pH_{DS}}) - 0.104(\mathrm{Cl_{2,d}}\times\mathrm{RT}\times T\times\mathrm{UV_{254}})$	5	280	*In situ* DWTP and DWDS data
PP	14	Chowdhury *et al.*, 2011 (ref. 24)	$\mathrm{THM_{PP}} = 21.4 + 36.9(\mathrm{Cl_2}) + 0.986(\mathrm{THM_{DS}}) + 0.59(\mathrm{TOC}) - 1.83(T) - 1.21((\mathrm{TOC} - 4.1)(T - 18.7))$	4	350	*In situ* PP data

Model development techniques were inconsistent between studies, and generally lacked the execution of important considerations such as outlier data calculation and sensitivity analyses. Outlier data may skew model correlation parameters leading to less accurate prediction capabilities. Sensitivity analyses are important for understanding bias of model outputs. The parameter estimation methods were mainly based on Pearson correlation tests, with the exception of models 1, 6, 11, 12, and 14, which used more comprehensive parameter estimation techniques. There were no clear trends in model performance based on model development methodologies. Further, the differences between methodologies of the models highlights the inconsistencies between model development and reporting.

Water quality variable selection was different for each of the models due to differences in model development. A summary of water quality variable usage for each model is presented in Table 3. Each study used a correlation test to determine which water quality variables were significant predictors for THM formation. The differences between models highlight the relative differences in phenomena affecting THM formation in different applications. The amount of data used for model training and validation also varied between models. For the studies that did not explicitly declare number of data, best estimates were used based on number of data points on graphs or sampling protocol descriptions. Both the amount of data and the diversity of data are important considerations for regression fitting since they can impact model accuracy and prediction capabilities under different conditions. Predictive validation is important for understanding the prediction accuracy of the model under conditions outside of the training data range. Water quality variable selection, amount of training data, data diversity, and model validation are further discussed in later sections. Overall, the differences in model development emphasize the lack of consistency with model approach. Recommendations are provided in section 4.

Table 4 Descriptive statistics for DWTP WQV data, and DWDS WQV data gathered from literature review. $Cl_{2,r}$ and $Cl_{2,d}$ are chlorine residual and chlorine dose, respectively. Br is bromide ion concentration, *T* is temperature, DOC is dissolved organic carbon, TOC is total organic carbon, RT is residence time, UV_{254} is ultraviolet absorbance at 254 nm wavelength, Alk is alkalinity, cond is conductivity, and THM it total trihalomethanes

Variable	Mean	SD	5%	95%	CV
DWTP					
$Cl_{2,d}$ (mg L^{-1})	2.51	0.39	1.87	3.16	0.16
$Cl_{2,r}$ (mg L^{-1})	1.16	0.16	0.90	1.41	0.14
Br (mg L^{-1})	0.39	0.17	0.17	0.74	0.44
pH	7.56	0.22	7.19	7.92	0.03
T (°C)	15.45	4.71	7.66	23.25	0.30
DOC (mg L^{-1})	1.80	0.23	1.42	2.19	0.13
TOC (mg L^{-1})	1.71	0.78	0.69	3.18	0.46
RT (hours)	21.18	15.46	3.03	52.79	0.73
UV_{254} (1 cm^{-1})	0.03	0.03	0.01	0.09	1.00
Alk (mg L^{-1} $CaCO_3$)	127.40	21.65	91.84	163.10	0.17
THM (μg L^{-1})	35.63	23.03	8.46	81.04	0.65
DWDS					
cond (μS cm^{-1})	959.39	398.79	310.89	1643.28	0.42
$Cl_{2,r}$ (mg L^{-1})	0.76	0.38	0.17	1.44	0.50
Br (mg L^{-1})	0.31	0.15	0.07	0.56	0.48
pH	7.61	0.4	6.93	8.26	0.05
T (°C)	19.10	6.33	8.5	29.58	0.33
DOC (mg L^{-1})	0.62	0.31	0.13	1.16	0.50
TOC (mg L^{-1})	2.02	0.99	0.44	3.74	0.49
RT (hours)	37.45	30.76	4.29	99.46	0.82
UV_{254} (1 cm^{-1})	0.02	0.02	0.01	0.06	1.00
Bicarbonate (mg L^{-1} $CaCO_3$)	204.53	78.19	76.39	334.99	0.38
THM (μg L^{-1})	36.97	23.1	9.19	82.19	0.62

3.2 Water quality variable data

Descriptive statistics for water quality variable data from the MC simulation can be seen in Table 4a and b. The CV values for water quality variables were within 10% of each other for DWTP and DWDS data sets with the exception of chlorine residual ($Cl_{2,r}$), pH, and dissolved organic carbon (DOC). Most of the water quality variable data was taken from multiple sources; however, DWTP DOC, residence time (RT), and alkalinity (Alk), and DWDS DOC, and bicarbonate

Table 3 WQV use among THM models

Application	Model #	$Cl_{2,d}$	$Cl_{2,r}$	Br	pH	Temp	DOC	TOC	RT	UV_{254}	cond	Alk	bicarb	THM
DWTP	1	x		x	x	x	x		x					
	2		x		x	x		x	x	x				
	3		x		x	x		x	x	x				
	4		x						x	x				
	5	x	x	x	x	x		x	x	x				
	6	x		x	x	x		x	x	x				
	7	x		x	x	x			x	x				
	8			x						x				
DWDS	9		x		x	x		x	x					
	10	x				x		x	x					x
	11		x		x	x		x			x		x	
	12				x			x						
	13		x		x	x				x				x
PP	14		x			x		x						x
	Total	5	8	5	10	11	1	9	9	8	1	0	1	3

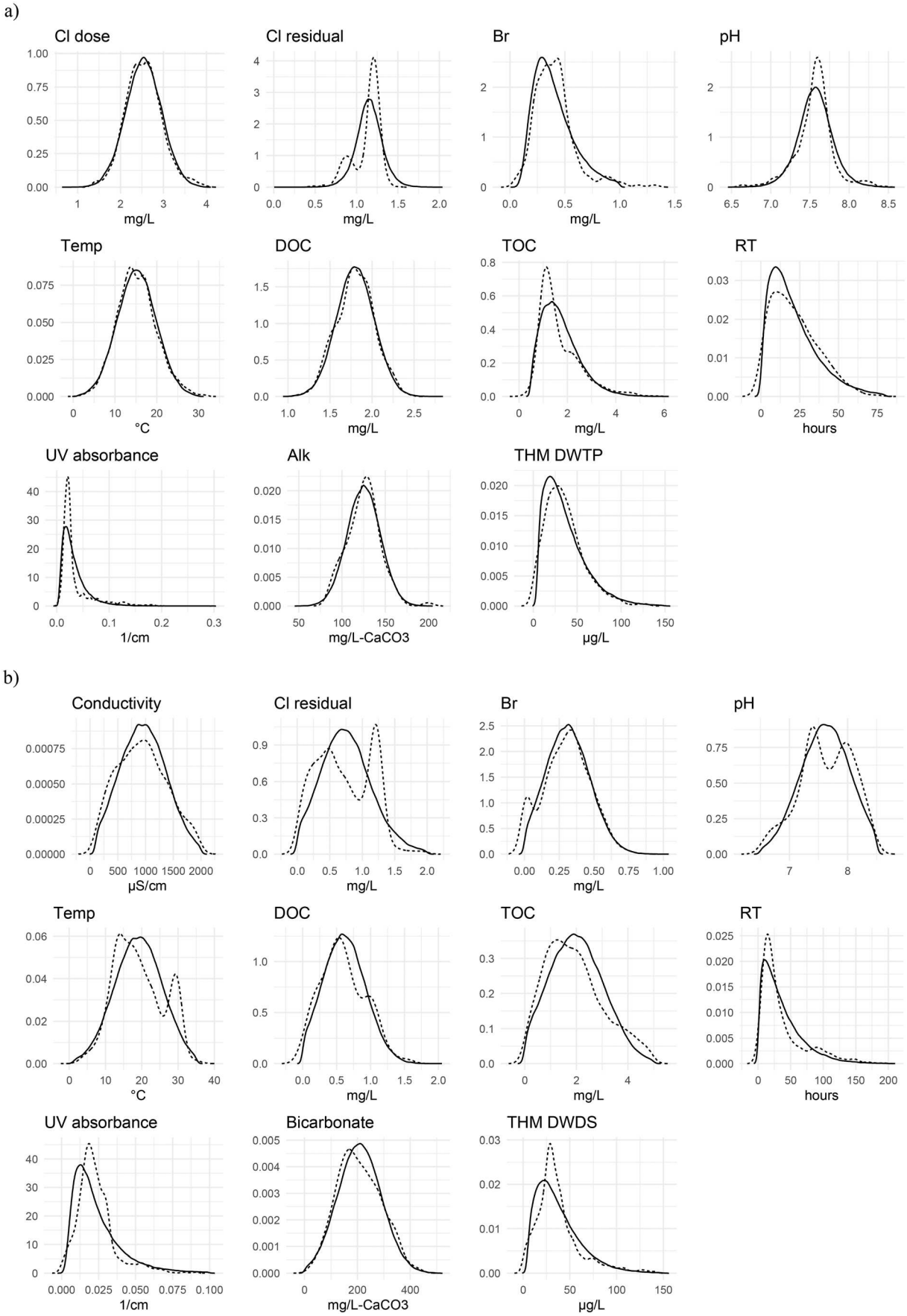

Fig. 2 Probability density functions (pdf), for raw WQV data (dashed line), and fitted distribution pdf (solid line), from MC simulation of a) DWTP data, and b) DWDS data.

Table 5 Mean, sd, 90% CI, ratio of mean predicted THM value/mean THM value from literature, and coefficient of variation (CV), for each model using a) DWTP data, and b) DWDS data generated from MC simulation

Application	Model #	Mean	sd	5%	95%	THM_p/THM_m	CV
a)							
DWTP	1	17.88	6.54	8.20	29.50	0.50	0.37
	2	422.44	131.21	263.70	690.17	11.86	0.31
	3	41.81	2.36	37.93	45.71	1.17	0.06
	4	25.22	7.45	13.29	37.86	0.71	0.30
	5	50.84	56.22	2.66	145.21	1.43	1.11
	6	59.23	22.16	28.15	99.59	1.66	0.37
	7	155.82	205.26	12.60	514.82	4.37	1.32
	8	71.62	43.89	23.97	155.53	2.01	0.61
DWDS	9	17.97	11.13	5.32	39.43	0.50	0.62
	10	39.03	8.83	27.66	55.75	1.10	0.23
	11	96.54	42.96	34.94	173.91	2.71	0.44
	12	8.02	3.05	3.97	13.45	0.23	0.38
	13	51.95	22.63	25.05	96.11	1.46	0.44
PP	14	—	—	—	—	—	—
b)							
DWTP	1	24.66	10.18	10.22	43.30	0.67	0.41
	2	583.50	253.90	301.43	1091.07	15.78	0.44
	3	42.60	3.00	37.30	46.92	1.15	0.07
	4	29.02	8.70	15.34	44.06	0.78	0.30
	5	95.19	122.58	13.13	304.67	2.57	1.29
	6	73.33	29.45	29.53	125.91	1.98	0.40
	7	247.32	333.02	15.91	842.84	6.69	1.35
	8	53.19	29.82	18.20	112.87	1.44	0.56
DWDS	9	26.84	17.72	4.06	60.57	0.73	0.66
	10	42.63	11.08	24.91	61.85	1.15	0.26
	11	91.11	53.81	18.91	191.35	2.46	0.59
	12	9.61	6.25	1.91	22.28	0.26	0.65
	13	51.99	22.63	25.09	96.07	1.41	0.44
PP	14	53.34	28.79	12.73	106.42	1.44	0.54

(bicarb), were based on one data source due to limited data available in the literature. Due to very limited PP data (*i.e.*, only one study), DWDS data were used for PP model analysis. Overall, there was a lack of consistency on data reporting between studies. For example, data were reported as a range of minimum and maximum values, confidence interval, or mean and SD either with or without number of data points. The exercise of collecting data from different studies was valuable in determining variability and average conditions for DWTP and DWDS water quality variables; however, the results could have been improved from increased reporting, and standardized reporting practices.

Probability density functions (PDF) for raw water quality variable data and fitted distributions are presented graphically in Fig. 2a and b. The aim was to produce distributions which represented standard conditions in DWTP and DWDS while also accounting for variability seen in real systems. Since data were collected as n, mean, and SD, data sets with greater n had greater bias in the shape of the raw data PDF, and subsequently the descriptive statistics. This can be seen in Fig. 2a and b where some raw water quality variable PDF have multimodal distributions (*e.g.*, $Cl_{2,r}$, T_{DWDS}). The modes correspond to relatively large data sets with significant differences in mean values. Another important consideration for water quality variables in the investigated systems is interdependency. Interdependency is when a change in one variable is correlated to a change in another variable (*e.g.*, Cl_2 and THM are affected by RT). Due to limited data availability, interdependency was not a consideration for this work. Despite the differences in data reporting and the resulting bias, the fitted distributions accurately reproduced general trends of the raw data and were consistent with trends found in literature.

3.3 Model performance

Descriptive statistics for model output data are presented in Table 5a and b. Graphic presentation of DWTP and DWDS model output PDF with their respective data sets are shown in Fig. 3a and b. Mean, SD, and CV of the model outputs were generally greater using DWDS data compared to DWTP data. Since the WQV data was not from a single source, comparison of model outputs was made between collected THM data. For analysis, model outputs were considered reasonable if the THM_p/THM_m was between 0.5 to 2.0. Five of the models did not meet this criterion when DWTP data were applied (models 2, 7, 8, 11, and 12), and five of the models did not meet this criterion when DWDS data were applied (models 2, 5, 7, 11, and 12). The weighted sum of absolute difference between mean, variance, and skewness (*i.e.*, the first three moments), between model output data and THM data for both DWTP and DWDS was calculated for each model:

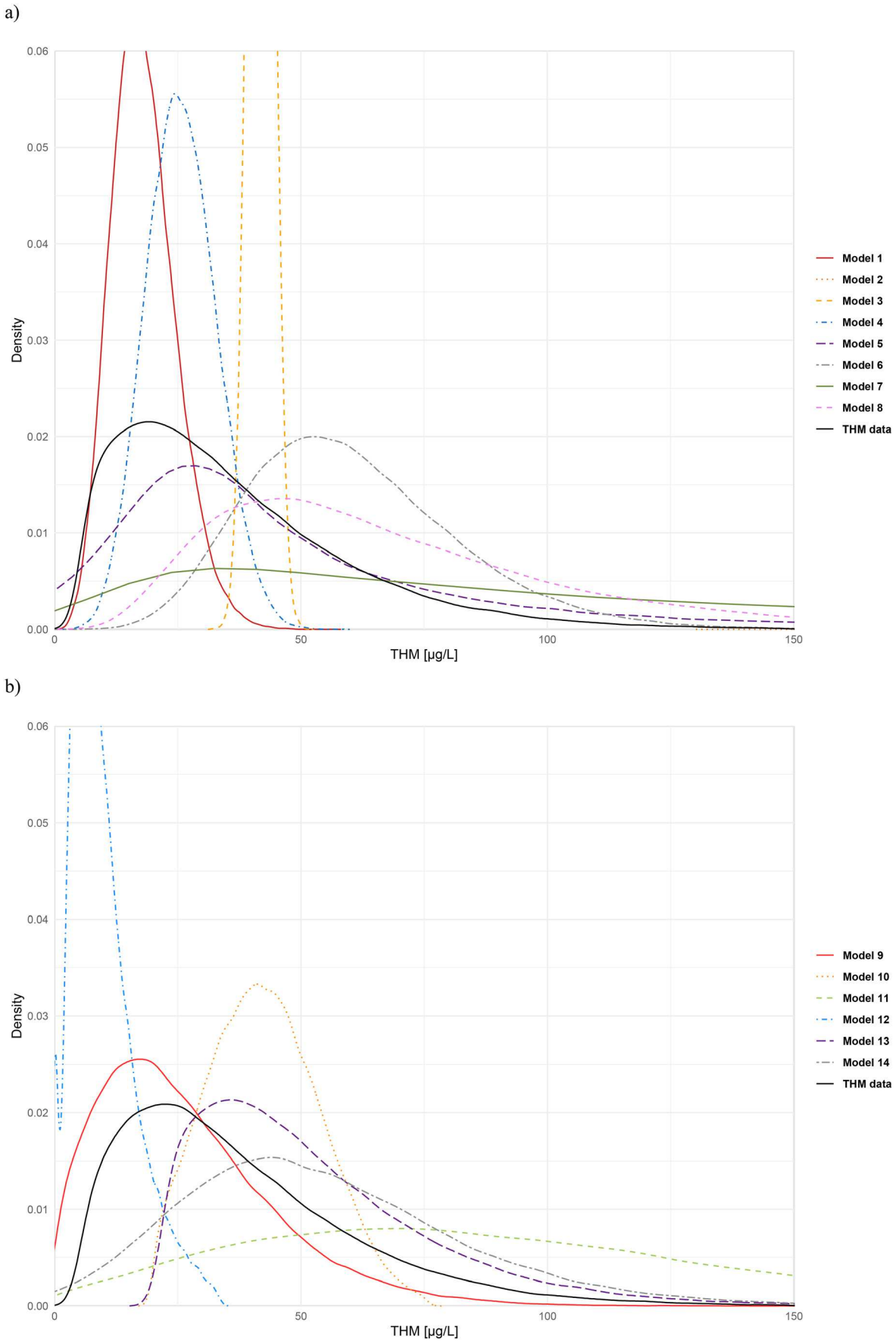

Fig. 3 Probability density functions (pdf) for a) DWTP model outputs with DWTP WQV data, and b) DWDS and PP model outputs with DWDS WQV data. The black lines represent distribution of THM data from literature.

$$\text{performance} = \sum_{j=1}^{2}\sum_{i=1}^{3}\frac{w_i \times \left|\text{mom}_{i,\text{THM},j} - \text{mom}_{i,\text{model}}\right|}{2}$$, where w designates the assigned weight, i designates the moment, and j designates the THM data (either DWDS or DWTP). The weights were chosen as 1.0, 0.1, and 1.0 for the first, second, and third moments, respectively. This performance value was used to quantitatively compare performance of each model to a reasonable estimate of what could be seen in practice. This method was chosen because it allowed for comparison of data shape and data distribution in addition to mean predicted THM. Based on this evaluation, the best performing models were models 6, 9, 13, and 14. These models demonstrated reasonable mean predicted THM values as well as reasonable variance and distribution shape compared to THM data, as seen in Fig. 3a and b. Although the variance in THM was not directly correlated to input variables, the goal of performance evaluation in this manner was to provide a balanced comparison. To understand why there were substantial differences between model outputs, water quality variable selection, training data variance, and model validation were compared.

The impact of variable selection on prediction accuracy for THM models was previously explored by Ged *et al.*[8] The research found that the most accurate THM models included at least 5 of the 7 following water quality variables as explanatory variables: DOC, UV_{254}, bromide ion concentration (Br,) pH, $Cl_{2,d}$, RT, and *T*. These were the most commonly used water quality variables among the models in this study as shown in Table 3. However, the number and type of water quality variables used by the models were found to have no significant correlation with model output variance for this work. It was also hypothesized that greater variance in water quality variable data would lead to greater variance in model output however, no trends between water quality variable variance and model output variance were found. The lack of correlation between model output variance and water quality variable selection/number of water quality variables demonstrates the differences in phenomena affecting THM formation between systems. The correlation between water quality variables and THM formation are system dependent due to differences in water quality profiles, differences in physical aspects of the systems including temperature, usage patterns, and pipe geometry, and differences in biofilm and subsequent effects.[40,44,45] The differences in phenomena affecting THM formation subsequently impact which water quality variables are included in the models (*i.e.*, which variables are statistically significant during model development), as well as the correlation coefficients for the variables. As a result of the development methods and statistical formulation of the models, the behavior between input data and model output variance is unique to each model. This demonstrates the lack of generalizable applicability of statistical models.

Another important consideration is the vague nature of many commonly used water quality variables. For instance, TOC is an aggregate measurement for carbonaceous constituents and is considered a crucial water quality variable for THM prediction since organic carbon is one of the reactants in the formation of THMs. Theoretically two systems could have identical TOC concentrations, however the carbonaceous species could be significantly different. It has been demonstrated that differences in TOC characteristics (*e.g.*, humic acid *vs.* fulvic acid composition) impact rate, and potential of THM formation.[46,47] Similarly, other water quality variables such as conductivity provide somewhat ambiguous characterization of water chemistry. Further, the previously discussed system specific differences present in DWTP and DWDS are more pronounced in PP systems due to stochastic flow conditions, higher surface area to volume ratio, and differences in building design.[5,48,49] Therefore, it can be concluded that statistically formulated models will not have generalizable application for different applications, or even different systems.

To understand the impact of training data on model prediction capabilities, Table 6 was constructed which provides reported data ranges used for model development compared to 90% CI for the data used in this study for model performance evaluation. Some of the studies had limited, or no reporting for the training data. Reporting descriptive statistics for model training data gives the reader a better understanding for how the model was developed, and the ranges which the model is expected to be most accurate. In general, models with more diverse training data (*i.e.*, larger ranges between min and max), tended to achieve better results for THM_p/THM_m and CV. However, unreported data values and differences in water quality variable sensitivity due to differences in correlation coefficients make it difficult to compare some data.

The results of this research demonstrate that there are significant differences in THM model development, evaluation, and reporting among studies. This research showed that comprehensive data was more important than number of data for model performance when applied to independent data. Similarly, models developed for highly specific application may struggle to perform well outside their training data ranges. THM models have been developed in a relatively similar manner for the past 30 years.[8,50] Statistical models provide value for utilities and consumers; however, it has been demonstrated that they have many drawbacks. With the advent of novel modeling techniques in the area of machine learning, there is much to explore outside of the realm of statistical models.

There have been a growing amount of research exploring the use of ML based techniques for predicting THMs.[51–57] Most of the research uses some type of artificial neural network (ANN) based model to develop non-linear relationships between water quality variables and THM concentration. The ML based approaches show promise by demonstrating lower error compared to their multiple linear regression based model counterparts.[51,55,56] Additionally, Zhang *et al.*, 2023 demonstrated that conducting a stepwise multiple linear regression for selection of significant input variables prior to

Table 6 A comparison of model training data ranges compared to data ranges used for model performance evaluation in this research. CI stands for confidence interval (*e.g.*, 90% of the data lies within the given range), NR stands for none reported. Italicized values are reported data which was not used as explanatory variables in the corresponding model. * = a direct value was not given, ** = mean and sd reported, provided ranges were based on assumption that ±2 (sd) = 90% CI

Model #	Data type	$Cl_{2,d}$	$Cl_{2,r}$	Br	pH	*T*	DOC	TOC	RT	UV_{254}	cond	bicarb	THM
DWDS	90% CI	1.85–3.18	0.90–1.38	0.16–0.75	7.18–7.93	7.85–22.55	1.41–2.17	0.69–3.86	2.59–52.58	0.01–0.22	—	—	8.71–80.19
DWTP	90% CI	—	0.16–1.43	0.07–0.57	6.95–8.27	8.35–29.64	0.12–1.07	0.44–3.73	4.46–97.78	0.01–0.06	298.44–1653.19	80.06–337.06	9.08–84.96
1	min–max	*	—	0.09–0.648	6.0–8.0	10–30	1.3–10.34	—	6–168	—	—	—	5.01–76.65
2	min–max	—	NR	—	NR	NR	—	NR	NR	NR	—	—	231–484
3	min–max	—	NR	—	NR	NR	—	NR	NR	NR	—	—	231–484
4	min–max	—	0.2–1.5	—	7.2	22	—	—	0–168	NR	—	—	35–135
5	min–max	NR	NR	NR	NR	NR	—	—	NR	NR	—	—	29–39
6	80% CI	1.13–1.29	—	0.22–0.51	7.22–7.72	11.28–25.00	—	0.81–2.78	0.00–46.40	0.014–0.0278	—	—	0.00–46.92
7	min–max	0.70–5.80	—	0.02–0.176	6.50–7.80	10.6–26.6	—	—	0.10–3.25	0.017–0.076	—	—	22.6–125.5
8	min–max	—	—	0.0–1.0	—	—	0.6–23.0	—	—	0.01–0.48	—	—	—
9	80% CI	2.93–4.24	—	—	7.54–7.76	12.1–16.4	—	2.32–3.54	2.4–37.0	—	—	—	8–55
10	min–max	0.0–0.4				15–30		3.5–5.5	0–34				NR
11	Mean, sd**	—	0.18–0.96	—	NR	NR	—	0.53–3.19	—	—	397.4–1490.1	123.05–287.91	17.22–132.31
12	min–max	—	*0.16–0.80*	—	7.3–8.9	—	—	0.31–39.5	—	—	*419.0–1141.0*	—	0.48–68.35
13	min–max	2.97–6.31	—	0.020–0.176	6.73–7.75	10.6–26.6	—	—	19.7–30.0	0.017–0.076	—	—	27.3–130.1
14	min–max	—	0.39–2.34	—	7.00–8.02	11.0–28	—	1.2–12.6	—	0.019–0.14	65–496	—	—

ML model training allowed for more efficient training and implementation of ML model.[51] Efficient implementation of ML models is particularly useful for real-time prediction of THM. Conducting correlation tests can also provide more detailed insight into the significant factors impacting THM formation for specific systems. Other sensitivity analysis techniques such as exclusion of variables, input variable differential analysis, and model weight analysis can provide insights into input variable importance even when correlation test are not conducted prior to training.[58,59] As software and hardware capabilities continue to improve, ML techniques will undoubtedly provide more accurate models for the prediction of THMs. ML techniques also have the potential to generate more generalizable models compared to regression models due to their ability to develop higher order relationships between water quality variables. This may be especially useful in PP model applications since there are more factors influencing THM formation such as water usage, pipe material, and temperature, potentially leading to highly non-linear relationships between water quality variables and THM formation.[6,12] It has been demonstrated that regression-based models can reasonably predict THM for specific systems, but further exploration of ML techniques for THM modeling seems like the most promising avenue of exploration.

There are certain applications that may benefit from the use of a simple statistical model, however novel approaches could provide improved insight into THM production mechanisms, and more generally applicable models. Models with greater applicability have the potential for far greater impact on the improvement of human health than models developed for a specific system. Further, data sharing and collaboration could increase the pace of THM model development. Many studies have attempted similar approaches with varying levels of success. The exercise of producing statistical models for THM in drinking waters has been demonstrated, now it is time to explore new approaches.

3.4 Limitations and future research

This research relied on assumptions that may not translate in practice, such as independent behavior of water quality variables, raw water quality variable data were representative of most systems, and differences in model CV were directly comparable. In practice, water quality variables are dependent on each other in a complex manner that follows general and system-specific trends. For example, RT impacts formation of THM, and consumption of Cl_2. The exact relationship between the variables is system specific due to water quality profiles varying by location. Further, the physicochemical phenomena impacting the relationships is different for different systems. In future work, a large enough data set with proper characterization may allow for consideration of the interdependence between the water quality variables. The other major assumption was that each model had directly comparable CV. This is a difficult

comparison to make since each model used different combinations of water quality variables. For example, model 12 only used two water quality variables, while model 6 used seven water quality variables. Consideration of these differences may be possible in the future if the recommendations provided in section 4 are utilized. Even with these limitations, this research was able to derive meaningful lessons for future THM model development. Moving forward, the work presented in this research would benefit from collection and application of water quality data from multiple applications and sources. This would allow for more accurate characterization of the data used, as well as better understanding of the interdependence of water quality variables, how they differ among location and application, how different models respond, and more accurate comparison of model output behaviors.

4. Proposed framework for THM model development

The following framework is proposed for future development of THM models to (1) promote clarity and consistency with respect to data reporting and model development methodology, (2) allow water quality from different sources to be accessed and utilized, and (3) improve THM prediction capabilities. These guidelines would improve understanding of THM formation in all applications discussed and are especially important for ML based THM models.

For data reporting, it is proposed that the following be included in the research:

1. High level description of data including geographic region(s) of collection, sampling timeline, sampling frequency, and any anomalies in the data;
2. Descriptive statistics of data used for model development including amount of data collected, mean, median, SD, and 90 or 95% CI, or equivalent;
3. Clear presentation and description of units for each explanatory variable;
4. Inclusion of raw data in accessible form (*e.g.*, .csv file or github link);
5. Inclusion of any code used for data cleaning, sorting, transformation, *etc.*;
6. Description of uncertainty with associated measuring techniques.

For model development methodology and reporting, it is proposed that the following be included in the research:

1. Detailed description of rationale behind model development approach;
2. Description of novelty provided by model development;
3. Description of data usage during model development (*e.g.*, 60% used for fitting/training, 20% used for testing, and 20% used for validation);
4. Inclusion of any code used for model development;
5. Inclusion of model validation predictive performance through with test data and/or independent data.

With this framework, it is envisioned that future research on THM modeling will serve not only a local purpose (*e.g.*, municipality), but also a global purpose to advance the field of water quality modeling. In particular, data sharing will allow models to be trained and validated using more diverse data sets, leading to more generalizable models. Modeling THM formation within PP is more challenging than DWTP or DWDS systems due to differences in physicochemical conditions, biological conditions, and stochastic water usage patterns. These differences may lead to different water quality variables needing to be considered. For example, it has been shown that copper pipes can catalyze THM formation, while PEX pipes may leach organic carbon.[60,61] With greater amount of data and better system characterization, higher level evaluation of system-specific characteristics could be evaluated. With these practices, it is intended that a more cohesive, multi-disciplinary approach will be encouraged, leading to greater progress in the field of THM modeling. Additionally, larger data sets would facilitate the exploration of machine learning based models to address the problem of generalizable models. Machine learning techniques have the capability of addressing the complex mechanisms leading to THM in all systems discussed.

5. Conclusion

Key findings of this research were:

- There has been disproportionately limited THM model development for PP application compared to DWTP and DWDS application.
- Although most THM models are statistical in formulation, there are inconsistencies with reporting of data and model development methodologies between THM studies.
- There were considerable differences between THM model performance due to differences in model development including intended application, water quality variable selection, amount of, and diversity of data used for training.
- THM modeling approach has primarily been focused on regression-based models for the past 30 years, however ML based models demonstrate promise to increase the accuracy and generalizability of THM models. To foster more unified THM modeling efforts, a new framework for model development was proposed to encourage novel strategies, data sharing, and collaboration.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

This research was supported by NSF award # 2027444 Proactive Water Quality Management of Water Networks in Buildings.

References

1 I. Fisher, G. Kastl, A. Sathasivan and R. Catling, Modelling chlorine residual and trihalomethane profiles in water distribution systems after treatment including pre-chlorination, *J. Environ. Chem. Eng.*, 2021, **9**(4), 105686.

2 P. Charisiadis, S. S. Andra, K. C. Makris, C. A. Christophi, D. Skarlatos and V. Vamvakousis, *et al.*, Spatial and seasonal variability of tap water disinfection by-products within distribution pipe networks, *Sci. Total Environ.*, 2015, **506–507**, 26–35.

3 W. J. Rhoads, A. Pruden and M. A. Edwards, Survey of green building water systems reveals elevated water age and water quality concerns, *Environ. Sci.: Water Res. Technol.*, 2016, **2**(1), 164–173.

4 R. Julien, E. Dreelin, A. J. Whelton, J. Lee, T. G. Aw and K. Dean, *et al.*, Knowledge gaps and risks associated with premise plumbing drinking water quality, *AWWA Water Sci.*, 2020, **2**(3), e1177.

5 E. Leslie, J. Hinds and F. I. Hai, Causes, Factors, and Control Measures of Opportunistic Premise Plumbing Pathogens—A Critical Review, *Appl. Sci.*, 2021, **11**(10), 4474.

6 M. Salehi, T. Odimayomi, K. Ra, C. Ley, R. Julien and A. P. Nejadhashemi, *et al.*, An investigation of spatial and temporal drinking water quality variation in green residential plumbing, *Build. Environ.*, 2020, **169**, 106566.

7 K. Yamamoto, N. Kakutani, A. Yamamoto and Y. Mori, A case study on the effect of storage of advanced treated water in a building's plumbing system on trihalomethane levels, *Bull. Environ. Contam. Toxicol.*, 2007, **79**(6), 665–669.

8 E. C. Ged, P. A. Chadik and T. H. Boyer, Predictive capability of chlorination disinfection byproducts models, *J. Environ. Manage.*, 2015, **149**, 253–262.

9 S. Chowdhury, P. Champagne and P. J. McLellan, Models for predicting disinfection byproduct (DBP) formation in drinking waters: A chronological review, *Sci. Total Environ.*, 2009, **407**(14), 4189–4206.

10 L. Wang, Y. Chen, S. Chen, L. Long, Y. Bu and H. Xu, *et al.*, A one-year long survey of temporal disinfection byproducts variations in a consumer's tap and their removals by a point-of-use facility, *Water Res.*, 2019, **159**, 203–213.

11 R. Richard, K. A. Hamilton, P. Westerhoff and T. H. Boyer, Physical, Chemical, and Microbiological Water Quality Variation between City and Building and within Multistory Building, *ACS ES&T Water*, 2021, **1**(6), 1369–1379.

12 M. Zheng, C. He and Q. He, Fate of free chlorine in drinking water during distribution in premise plumbing, *Ecotoxicology*, 2015, **24**(10), 2151–2155.

13 S. Masters, J. Parks, A. Atassi and M. A. Edwards, Distribution system water age can create premise plumbing corrosion hotspots, *Environ. Monit. Assess.*, 2015, **187**(9), 559.

14 H. Hong, Q. Song, A. Mazumder, Q. Luo, J. Chen and H. Lin, *et al.*, Using regression models to evaluate the formation of trihalomethanes and haloacetonitriles via chlorination of source water with low SUVA values in the Yangtze River Delta region, China, *Environ. Geochem. Health*, 2016, **38**(6), 1303–1312.

15 M. Kumari and S. K. Gupta, Modeling of trihalomethanes (THMs) in drinking water supplies: a case study of eastern part of India, *Environ. Sci. Pollut. Res.*, 2015, **22**(16), 12615–12623.

16 D. K. Roth and D. A. Cornwell, DBP Impacts From Increased Chlorine Residual Requirements, *J. AWWA*, 2018, **110**(2), 13–28.

17 N. K. Shahi, M. Maeng and S. Dockko, Models for predicting carbonaceous disinfection by-products formation in drinking water treatment plants: a case study of South Korea, *Environ. Sci. Pollut. Res.*, 2020, **27**(20), 24594–24603.

18 L. Godo-Pla, P. Emiliano, M. Poch, F. Valero and H. Monclús, Benchmarking empirical models for THMs formation in drinking water systems: An application for decision support in Barcelona, Spain, *Sci. Total Environ.*, 2021, **763**, 144197.

19 A. Domínguez-Tello, A. Arias-Borrego, T. García-Barrera and J. L. Gómez-Ariza, 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.*, 2017, **24**(28), 22631–22648.

20 B. Chen and P. Westerhoff, Predicting disinfection by-product formation potential in water, *Water Res.*, 2010, **44**(13), 3755–3762.

21 E. C. Wert, J. Bolding, D. J. Rexing and R. E. Zegers, Real-time modeling of trihalomethane formation in a full-scale distribution system, *J. Water Supply: Res. Technol.–AQUA*, 2012, **61**(6), 352–363.

22 L. Cong, Y. J. Yang, Y. Jieze, Z. Tu-qiao, M. Xinwei and S. Weiyun, Second-Order Chlorine Decay and Trihalomethanes Formation in a Pilot-Scale Water Distribution Systems, *Water Environ. Res.*, 2012, **84**(8), 656–661.

23 F. Osorio, D. Ribes, A. Gonzalez-Martinez, J. M. Poyatos and P. Garci, A model for predicting THM presence in networks of water supply systems, *WIT Trans. Built Environ.*, 2011, **117**, 233–340.

24 S. Chowdhury, M. J. Rodriguez, R. Sadiq and J. Serodes, Modeling DBPs formation in drinking water in residential plumbing pipes and hot water tanks, *Water Res.*, 2011, **45**(1), 337–347.

25 S. Tsitsifli and V. Kanakoudis, Total and Specific THMs' Prediction Models in Drinking Water Pipe Networks, *Environmental Sciences Proceedings*, 2020, **2**(1), 55.

26 S. K. Golfinopoulos, N. K. Xilourgidis, M. N. Kostopoulou and T. D. Lekkas, Use of a multiple regression model for predicting trihalomethane formation, *Water Res.*, 1998, **32**(9), 2821–2829.

27 J. Sohn, D. Gatel and G. Amy, Monitoring and Modeling of Disinfection By-Products (DBPs), *Environ. Monit. Assess.*, 2001, **70**(1), 211–222.

28 V. Uyak, I. Toroz and S. Meriç, Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul, *Desalination*, 2005, **176**(1), 91–101.

29 R. S. Chaves, D. Salvador, P. Nogueira, M. M. Santos, P. Aprisco and C. Neto, *et al.*, Assessment of Water Quality Parameters and their Seasonal Behaviour in a Portuguese Water Supply System: a 6-year Monitoring Study, *Environ. Manage.*, 2022, **69**(1), 111–127.
30 S. K. Golfinopoulos and G. B. Arhonditsis, Quantitative assessment of trihalomethane formation using simulations of reaction kinetics, *Water Res.*, 2002, **36**(11), 2856–2868.
31 M. J. Rodriguez, Y. Vinette, J.-B. Sérodes and C. Bouchard, Trihalomethanes in Drinking Water of Greater Québec Region (Canada): Occurrence, Variations and Modelling, *Environ. Monit. Assess.*, 2003, **89**(1), 69–93.
32 D. Mouly, E. Joulin, C. Rosin, P. Beaudeau, A. Zeghnoun and A. Olszewski-Ortar, *et al.*, Variations in trihalomethane levels in three French water distribution systems and the development of a predictive model, *Water Res.*, 2010, **44**(18), 5168–5179.
33 S. Tsitsifli and V. Kanakoudis, Developing THMs' Predictive Models in Two Water Supply Systems in Greece, *Water*, 2020, **12**(5), 1422.
34 D. E. Kelly-Coto, A. Gamboa-Jiménez, D. Mora-Campos, P. Salas-Jiménez, B. Silva-Narváez and J. Jiménez-Antillón, *et al.*, Modeling the formation of trihalomethanes in rural and semi-urban drinking water distribution networks of Costa Rica, *Environ. Sci. Pollut. Res.*, 2022, **29**(22), 32845–32854.
35 M. P. Abdullah, C. H. Yew and M. S. Ramli, 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.*, 2003, **37**(19), 4637–4644.
36 M. Feungpean, B. Panyapinyopol, P. Elefsiniotis and P. Fongsatitkul, Development of statistical models for trihalomethane (THM) occurrence in a water distribution network in Central Thailand, *Urban Water J.*, 2015, **12**(4), 275–282.
37 Y. Zhang, D. Martinez, C. Collins, N. Graham, M. R. Templeton and J. Huang, *et al.*, Modelling of haloacetic acid concentrations in a United Kingdom drinking water system, *J. Water Supply: Res. Technol.–AQUA*, 2011, **60**(5), 275–285.
38 A. Domínguez-Tello, A. Arias-Borrego, T. García-Barrera and J. L. Gómez-Ariza, Seasonal and spatial evolution of trihalomethanes in a drinking water distribution system according to the treatment process, *Environ. Monit. Assess.*, 2015, **187**(11), 662.
39 H. H. Chang, H. H. Tung, C. C. Chao and G. S. Wang, Occurrence of haloacetic acids (HAAs) and trihalomethanes (THMs) in drinking water of Taiwan, *Environ. Monit. Assess.*, 2010, **162**(1), 237–250.
40 J. Xu, C. Huang, X. Shi, S. Dong, B. Yuan and T. H. Nguyen, Role of drinking water biofilms on residual chlorine decay and trihalomethane formation: An experimental and modeling study, *Sci. Total Environ.*, 2018, **642**, 516–525.
41 M. A. Palmegiani, A. J. Whelton, J. Mitchell, P. Nejadhashemi and J. Lee, New developments in premise plumbing: Integrative hydraulic and water quality modeling, *AWWA Water Sci.*, 2022, **4**(2), e1280.
42 G. R. Abhijith, L. Kadinski and A. Ostfeld, Modeling Bacterial Regrowth and Trihalomethane Formation in Water Distribution Systems, *Water*, 2021, **13**(4), 463.
43 A. Sathasivan, G. Kastl, S. Korotta-Gamage and V. Gunasekera, Trihalomethane species model for drinking water supply systems, *Water Res.*, 2020, **184**, 116189.
44 C. Zhang, C. Li, X. Zheng, J. Zhao, G. He and T. Zhang, Effect of pipe materials on chlorine decay, trihalomethanes formation, and bacterial communities in pilot-scale water distribution systems, *Int. J. Environ. Sci. Technol.*, 2017, **14**(1), 85–94.
45 R. Richard and T. H. Boyer, Pre- and post-flushing of three schools in Arizona due to COVID-19 shutdown, *AWWA Water Sci.*, 2021, **3**(5), e1239.
46 H. V.-M. Nguyen, H.-S. Lee, S.-Y. Lee, J. Hur and H.-S. Shin, Changes in structural characteristics of humic and fulvic acids under chlorination and their association with trihalomethanes and haloacetic acids formation, *Sci. Total Environ.*, 2021, **790**, 148142.
47 M. Y. Z. Abouleish and M. J. M. Wells, Trihalomethane formation potential of aquatic and terrestrial fulvic and humic acids: examining correlation between specific trihalomethane formation potential and specific ultraviolet absorbance, *Environ. Chem.*, 2012, **9**(5), 450–461.
48 G. R. Calle, I. T. Vargas, M. A. Alsina, P. A. Pastén and G. E. Pizarro, Enhanced Copper Release from Pipes by Alternating Stagnation and Flow Events, *Environ. Sci. Technol.*, 2007, **41**(21), 7430–7436.
49 K. Lautenschlager, N. Boon, Y. Wang, T. Egli and F. Hammes, Overnight stagnation of drinking water in household taps induces microbial growth and changes in community composition, *Water Res.*, 2010, **44**(17), 4868–4877.
50 G. Amy, Survey on Bromide in Drinking Water and Impacts on DBP Formation, Denver, American Water Works Research Foundation Report, 1994.
51 J. Zhang, D. Ye, Q. Fu, M. Chen, H. Lin and X. Zhou, *et al.*, The combination of multiple linear regression and adaptive neuro-fuzzy inference system can accurately predict trihalomethane levels in tap water with fewer water quality parameters, *Sci. Total Environ.*, 2023, **896**, 165269.
52 A. A. Babaei, Y. Tahmasebi Birgani, Z. Baboli, H. Maleki and K. Ahmadi, Using water quality parameters to prediction of the ion-based trihalomethane by an artificial neural network model, *Environ. Monit. Assess.*, 2023, **195**(8), 917.
53 K. Liu, T. Lin, T. Zhong, X. Ge, F. Jiang and X. Zhang, New methods based on a genetic algorithm back propagation (GABP) neural network and general regression neural network (GRNN) for predicting the occurrence of trihalomethanes in tap water, *Sci. Total Environ.*, 2023, **870**, 161976.
54 C. N. Okoji, A. I. Okoji, M. S. Ibrahim and O. Obinna, Comparative analysis of adaptive neuro-fuzzy inference system (ANFIS) and RSRM models to predict DBP (trihalomethanes) levels in the water treatment plant, *Arabian J. Chem.*, 2022, **15**(6), 103794.

55 A. Alver, E. Başt̨ürk and A. Kılıç, Development of adaptive neuro-fuzzy inference system model for predict trihalomethane formation potential in distribution network simulation test, *Environ. Sci. Pollut. Res.*, 2021, **28**(13), 15870–15882.

56 J. K. Mahato and S. K. Gupta, Exploring applicability of artificial intelligence and multivariate linear regression model for prediction of trihalomethanes in drinking water, *Int. J. Environ. Sci. Technol.*, 2022, **19**(6), 5275–5288.

57 I. Kropp, A. Nejadhashemi, R. Julien, J. Mitchell and A. Whelton, A machine learning framework for predicting downstream water end-use events with upstream sensors, *Water Sci. Technol.: Water Supply*, 2022, **22**, 6427–6442.

58 B. Mrzygłód, M. Hawryluk, M. Janik and I. Olejarczyk-Wożeńska, Sensitivity analysis of the artificial neural networks in a system for durability prediction of forging tools to forgings made of C45 steel, *Int. J. Adv. Manuf. Technol.*, 2020, **109**(5), 1385–1395.

59 Z. Zhang, M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda and H. Goyal, *et al.*, Opening the black box of neural networks: methods for interpreting neural network models in clinical applications, *Ann. Transl. Med.*, 2018, **6**(11), 216.

60 E. R. Blatchley 3rd, D. Margetas and R. Duggirala, Copper catalysis in chloroform formation during water chlorination, *Water Res.*, 2003, **37**(18), 4385–4394.

61 D. L. Tolofari, S. V. Masters, T. Bartrand, K. A. Hamilton, C. N. Haas and M. Olson, *et al.*, Full factorial study of pipe characteristics, stagnation times, and water quality, *AWWA Water Sci.*, 2020, **2**(5), e1204.