International Journal of Environmental Science and Technology (2022) 19:5275–5288
https://doi.org/10.1007/s13762-021-03392-1

ORIGINAL PAPER

Exploring applicability of artificial intelligence and multivariate linear regression model for prediction of trihalomethanes in drinking water

J. K. Mahato[1] · S. K. Gupta[1]

Received: 15 January 2021 / Revised: 16 April 2021 / Accepted: 11 May 2021 / Published online: 21 May 2021

Abstract

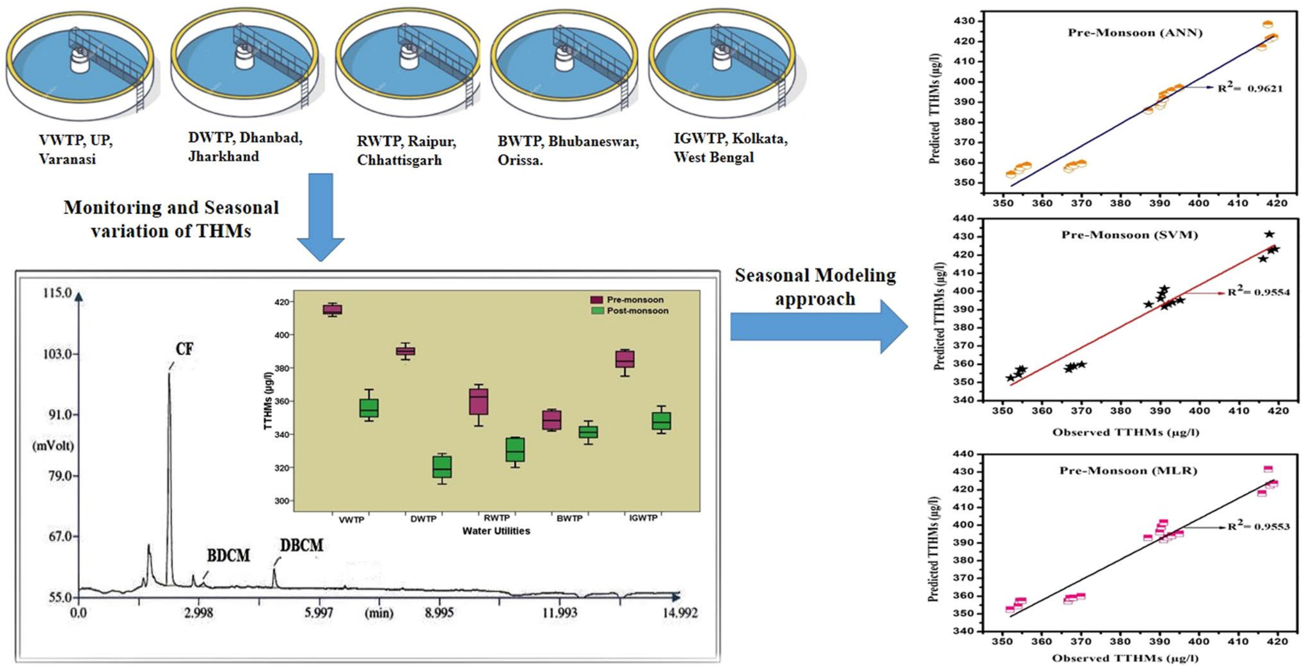
The present study describes the application of artificial intelligence-based modeling approach to predict trihalomethanes in drinking water supplies. The samples were collected from five major water utilities located in five different states across India for two seasons to establish the baseline. Trihalomethane formation was correlated with various operation parameters and exceeded the prescribed guideline value of the World Health Organization and the Bureau of Indian Standard. Chloroform was found to be the most predominant compound > 90% contribution to total trihalomethanes. The seasonal variation assessment revealed that the trihalomethanes level was relatively 1.12 ± 0.074 times higher in pre-monsoon than post-monsoon. The correlation analysis confirmed, total organic carbon followed by dissolved organic carbon is a major organic precursor responsible for trihalomethane formation. Monitoring these compounds is essential to ensure public safety but cannot be regularly determined due to the involvement of sophisticated instruments and the procedure. The artificial intelligence-based modeling approach could prove to be a good tool for instant prediction of trihalomethanes with better accuracies. An artificial neural network and support vector machine was employed using Python® and MATLAB® respectively, whereas, for multivariate linear regression, SPSS® was used. The value of coefficient and comparison of performance data indicated that artificial neural networks gave the most promising results, followed by support vector machine and multivariate linear regression. The study could prove to be very useful for regulatory agencies to manage and control trihalomethanes' levels in drinking water supplies.

Editorial responsibility: Lifeng Yin.

✉ S. K. Gupta
sunil@iitism.ac.in

[1] Department of Environmental Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India 826004

Graphic abstract



Keywords Artificial neural network · Disinfection by-products · Natural organic matter · Support vector machine

Introduction

As one of the fastest developing countries, India grips only 4% of global potable water resources, supporting around 17% of the earth's population (Chakraborty 2017). The major sources of drinking water supply in India are the surface reservoir, making it even more challenging to provide potable water due to microbiological contamination (Mahato and Gupta 2020; Marais et al. 2019). Chlorine is a predominant disinfectant used to date, which results in the formation of trihalomethanes (THMs) via reacting with natural organic matter (NOM) (Fig. 1) (Al-Tmemy et al. 2018; Mahato et al. 2019; Hur et al. 2014; Hong et al. 2008). These compounds include chloroform ($CHCl_3$) (CF), bromoform ($CHBr_3$) (BF), dibromochloromethane ($CHBr_2Cl$) (DBCM), and bromodichloromethane ($CHBrCl_2$) (BDCM), are of great concern as they were classified as potential human carcinogens (Padhi et al. 2019; Li and Mitch 2018). Previous findings showed that the concentration range of THMs (231–511 μg/l) in the Indian drinking water distribution system is greatly influenced by the seasonal and spatial variations (Kumari and Gupta 2015, 2018; Mishra and Dixit 2013; Thacker et al. 2002). Similarly, in other countries like Pakistan (575–595 μg/l) (Abbas et al. 2015), Japan (378 μg/l) (Imo et al. 2007), Canada (137.8–141 μg/l) (Milot et al. 2000; Rodriguez et al. 2003a), Turkey (96–102 μg/l) (Uyak et al. 2005), and China (92.77 μg/l) (Ye et al. 2011), wide

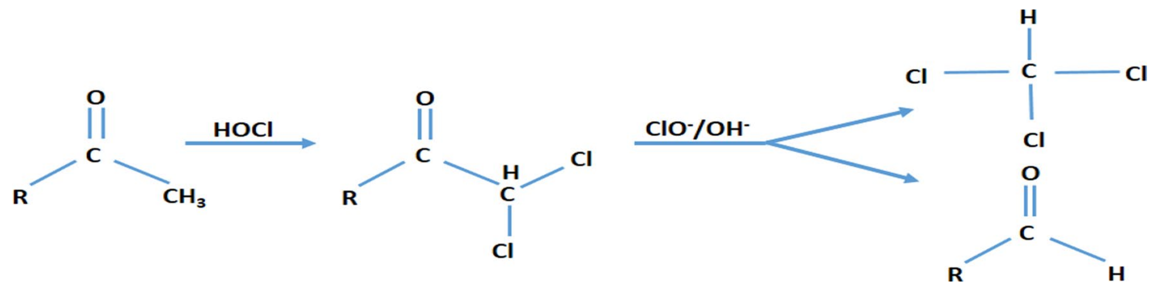


Fig. 1 Reaction mechanism of $CHCl_3$ formation in chlorinated drinking water

fluctuations in THMs levels were recorded in their water supplies.

These changes were also arisen due to the variation in their precursor concentration (NOM) and other operational parameters (pH, residual chlorine [RC], temperature) (Padhi et al. 2019; Kumari and Gupta 2015; Rathbun 1996). The Bureau of Indian Standard (BIS) (BIS 2012) set the permissible limit for individual THMs, i.e., CF (200 µg/l), BDCM (60 µg/l), DBCM (100 µg/l), and BF (100 µg/l), which is similar to that of proposed by the World Health Organization (WHO) except CF (200 µg/l) (Cotruvo 2017). The United States Environmental Protection Agency (USEPA) also established the guideline value (80 µg/l) only for total THMs (TTHMs) (USEPA 2018).

Monitoring of THMs all through the treatment process is vital for the management of quality control and to ensure the compliance of regulatory standards. The development of predictive models was proven to be a more effective and instant approach (Rodriguez et al. 2000). Usually, the predictive models establish the empirical and mechanistic relationship between the level of THMs and the operational parameters (Ye et al. 2011; Di Cristo et al. 2013), whereas some of the models are based on statistical regression equations and described the formation of THMs kinetics (Rathbun 1996; Rodriguez et al. 2000; Elshorbagy et al. 2000; Sadiq and Rodriguez 2004; Milot et al. 2000; Amy et al. 1987). Previous findings demonstrated that the artificial intelligence (AI)-based modeling approach can provide greater set prediction accuracies even in low quantities of data than the conventional multiple linear regression (MLR) model (Peleato et al. 2018; Kulkarni and Chellam 2010; Uyak et al. 2005). However, the application of artificial neural network (ANN) and its comparative assessment with support vector machine (SVM) and MLR model to predict THMs in drinking were not explored earlier. It was also noticed that most of the studies are done on laboratory-generated simulated water, which differs from the actual drinking water utilities (Ye et al. 2011; Milot et al. 2000). Hence, our study's emphasis is to generate models based on the real water collected from different water treatment plants (WTPs) located in India's various regions. The objective of the study includes (1) develop the AI-based THMs models from filed scale real data (2) the comparative assessment of these machine learning approach with conventional MLR model and (3) investigate the correlation of various operational parameters on THMs formation.

The above study was carried out during pre-monsoon (PrM) (April to June) and post-monsoon (PoM) (October to December) seasons in the year 2016–2018 in five different states of India, i.e., Jharkhand, Utter Pradesh Chhattisgarh, Orissa, and West Bengal.

Materials and methods

Sampling protocol

Five major drinking water utilities from the city of four different contiguous states of Jharkhand were considered for this study, i.e., (1) Water Treatment Plant, Belatand, Dhanbad, Jharkhand (DWTP), (2) Water Treatment Plant, Bhelupur, Varanasi, Utter Pradesh (VWTP), (3) Water treatment plant Ravanbhata, Raipur, Chhattisgarh (RWTP), (4) Water Treatment Plant Palasuni, Bhubaneshwar Orissa (BWTP), and (5) Indira Gandhi Water Treatment Plant, Barrackpore, West Bengal (IGWTP). Triplicate samples of raw (intake) and treated water (supply water) from these WTPs were collected during PrM and PoM seasons in the year 2016–2018. A total of 150 samples were first analyzed to establish THMs levels. The description and location details of the study area are illustrated in Table 1 and Fig. 2, respectively. These utilities follow conventional water treatment processes

Table 1 General characteristics of utilities under the study

Name of water utility	Latitude and longitude	Source river	Plant capacity (MLD)	Coagulant used	Disinfectant used
VWTP	25°17′54.81″N 82°59′43.51″E	Ganga River	125	Alum and PACl	Chlorine gas
DWTP	23°50′36.62″N 86°26′27.96″E	Barakar River	77	Alum and Lime	Chlorine gas
RWTP	21°12′59.43″N 81°37′50.07″E	Kharoon River	277	Alum, and PACl	Chlorine gas and bleaching powder
BWTP	20°18′15.72″N 85°51′48.98″E	Kuakhai River	41	Lime, Alum, and PACl	Chlorine gas
IGWTP	22°47′2.18″N 88°20′35.79″E	Hooghly River	916	Alum	Chlorine gas

MLD Million litter per day, *PACl* Poly aluminum chloride

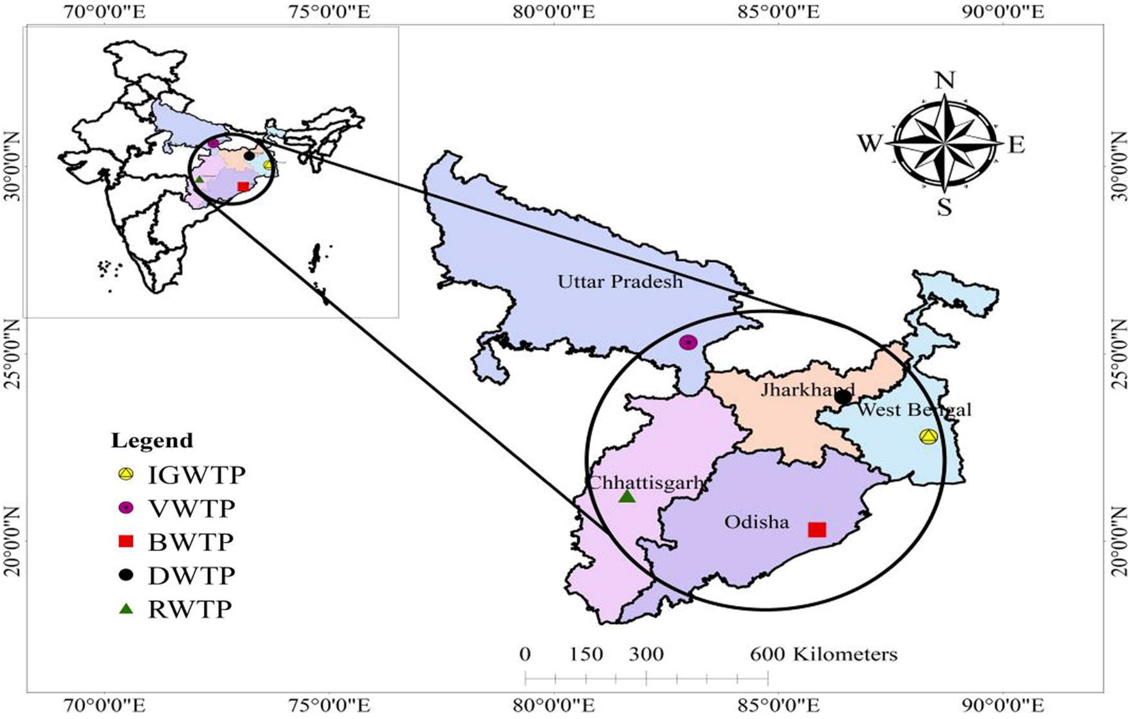


Fig. 2 Location details of drinking water utilities selected for the study

comprised of coagulation–flocculation, sedimentation, sand filtration, and chlorination.

Analytical method

The monitoring of physicochemical parameters was done as per the standard protocols of APHA, 2012. Total organic carbon (TOC) and dissolved organic carbon (DOC) (sample filtered through a 0.45 µm filter) were analyzed by a TOC analyzer (TOC-L CSH; Make: Shimadzu, Japan). Specific ultraviolet absorption (SUVA), which is an indicator of the aromatic character of NOM, was determined by the ratio of UV_{254} and DOC concentration, expressed as $L\ mg^{-1}\ m^{-1}$. The concentration of THMs was determined by USEPA method 552.1 using a combination of liquid–liquid extraction and gas chromatograph electron capture detector (GC-ECD) (Thermo Fisher, CERES 800 plus) (Hodegeson 1990). The GC-ECD conditions used for analysis are given in Table 2.

Table 2 Operating conditions for analysis of THM through the GC

Item	Condition
GC	CHEMITO CERES-800 PLUS with Ni^{63} ECD
Column	Packed column
Oven temperature	40 °C
Injector temperature	200 °C
Detector temperature	250 °C
Carrier gas	Nitrogen
Carrier gas flow	60 (ml/min)

Quality assurance and quality control (QA/QC)

As to ensure the consistency of the analytical results, the blank sample was prepared and analyzed to determine the presence of background contamination. Sample injection was performed in triplicate for the precision of measurement, and the average value was considered the final value. In case the relative percent difference between the two samples tends to surpassed ± 10%, the instrument was considered out of calibration and recalibrated.

Modeling approach

For this study, two machine learning techniques, viz., ANN and SVM were employed for the prediction of THMs formation and compared with the conventional MLR model. A set of five water quality parameters, namely pH, temperature, RC, TOC, and $UV_{254,}$ were used as independent variables and total THMs (TTHMs) as dependent variables. For maintaining the measurement accuracy, triplicated samples were collected each time during sample collection. The average of the three sample readings was reported as a single observed value for each parameter. A total of 150 observations from various WTPs included triplicate samples that were averaged and confined to 50 observations, which were considered input data for the models' development and validation, wherein, for the modeling 60% of the data set and validation 40% of the data set were used, i.e., subsets of input and output data had the dimensions of 30 samples × 5 independent variables × 1 dependent variable, and 20 samples × 5 independent variables × 1 dependent variables, respectively. To meet the required algorithm and facilitate network learning, data normalization is essential before starting the training process. There are many methods for normalizing the input data, like external normalization, along channel, across channel, and mixed channel. In the present work, the input data were normalized using the min–max normalization method as stated in Eq. (1). This method has the advantage of preserving exactly all relationships in the data. It actually normalizes the raw values in the range of 0–1 for better prediction.

$$\text{Normalized data} = (L\text{–Min})/(\text{Max} - \text{Min}) \quad (1)$$

where L is the raw value, Max and Min are the maximum and minimum of raw values, respectively.

The ANN is based on complex biological neural systems of the human brain, having certain theoretical advantages over the conventional modeling approach (MLR). It arises from the field of artificial intelligence and consists of several layers of processing elements with their nodes (neuron) (Rodriguez et al. 2003b; Singh and Gupta 2012). These neurons are arranged in an input layer that receives a signal input, one or many hidden layers that process the information actively, as well as an output layer that responds to the network (Fig. 3). Elements of different layers are highly interconnected by weighted links through which information may pass. The number of these elements in the input and output layer mainly depends on the number of input and output variables used in the specific problem to be solved. In the present work, a three-layer ANN was implemented with backpropagation algorithm in Python (3.7.1) by using Jupyter Notebook integrated development environment (IDE) with Sklearn library. The backpropagation algorithm has demonstrated several advantages to having the potential for determining networks with arbitrary mapping projections (Cook and Wolfe 1991; Rodriguez et al. 2003b). Hence, this algorithm was used to supervise the learning algorithm by changing the hyperparameters, viz., learning rate (LR) and momentum term (MT) to yield the best convergence. Moreover, the logistic relu activation function was used to activate the hidden and output layer. The number of nodes in the hidden layer for the optimal neural network was determined by optimization of hyperparameters using trial and error methods (Azadi and Karimi-Jashni 2016). The input layer consists of five nodes, i.e., pH, Temp., RC, TOC, and UV_{254}, whereas the output layer has one node – TTHMs. The practical applications of ANNs require the correct selection of LR and MT to separate the noisy data and avoid over-fitting problems. The degree of

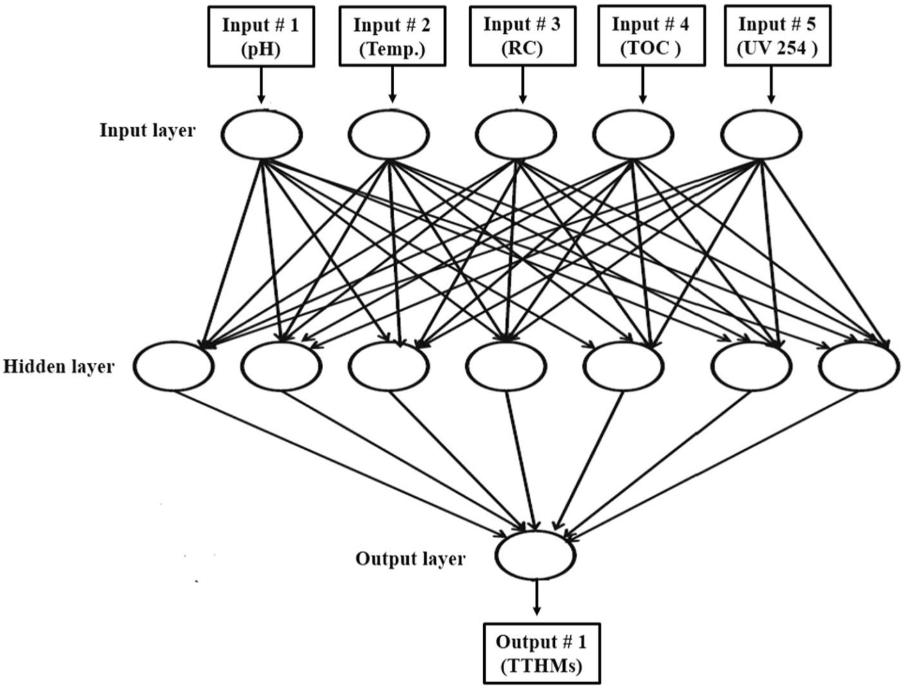


Fig. 3 Basic structure of the ANN model

Table. 3 Degree of correlation and MSE at various MT and LR

	MT-0.9	LR-0.01	MT-0.7	LR-0.02	MT-0.5	LR-0.03
R^2	0.962		0.954		0.952	
MSE	28.36		34.83		35.35	

correlation (R^2) and MSE at various LR and MT is given in Table 3. A maximum of 10,000 iterations was performed to achieve the optimum network. In-depth theory and mathematical details of learning and estimation of the parameter are broadly explained in the previous literature (Peleato et al. 2018; Singh and Gupta 2012).

SVM is a well-known supervised machine learning technique based on structural risk minimization (SRM), the theory of statistical learning (Singh and Gupta 2012; Vapnik 2013). It acts as a binary classifier to find the maximal margin (hyperplane) between two classes. In this approach, the original data points from the input space are mapped into a high or even infinite-dimensional feature space using a suitable kernel function (class of algorithms for pattern analysis), where the hyperplane is constructed. It can deal with a large number of features to find the optimal hyperplane from which the distance to all the data points is minimum and reduce the model dimensions and estimated errors: the theory and mathematical concept of the SVM model described in detail by Haykin (1999). The implementation of SVM was performed in MATLAB (9.5).

MLR is a relatively advanced concept of simple linear regression, used in various research fields to establish the strength of a linear relationship between a set of independent variables and dependent variables (Rodriguez et al. 2003b). This relationship can be described by following equation form (Rodriguez et al. 2003b).

$$Y = \sum \beta_o + \sum_{i=1}^{m} \beta_i X_i \tag{2}$$

where Y and X_i represent the dependent and independent variables, respectively, with m denoting the number of independent variables considered, β_o and β_i are the intercept and partial slope coefficients, respectively, providing prediction for the value of Y. In this approach, predictor variables were classified first according to their statistical significance and then including one variable at a time at different steps. The MLR model works on the ordinary least squares (OLS) method, which minimized the vertical distances of the sum of squared from the observed data points to the line (Neter et al. 1990). The SPSS (IBM, 21.0) was used to perform the implementation of the MLR model.

Sensitivity analysis

Sensitivity analyses were carried out using various statistical metrics, that is, R^2 (predicted vs. observed), root means square error (RMSE), means absolute percentage error (MAPE), and Index of Agreement (IA) to evaluate the performance of developed models. The equation for the determination of MAPE, RMSE, and IA is indicated in Eqs. (3) (4) and (5), respectively,

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i \right)^2} \tag{4}$$

where Y_i and $\hat{Y}_i$ are the observed and predicted value of TTHMs, and n is the number of samples.

$$\text{IA} = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} \left[\left| O_i - O_m \right| + \left| P_i - P_m \right| \right]^2} \tag{5}$$

where O_i and P_i are observed and predicted TTHMs concentration and N is the number sample tested. O_m and P_m represent the means of the observed and predicted total trihalomethane concentration.

Results and discussion

Concentration range of THMs species at five water utilities

The descriptive statistics of all the THMs species under the present study are given in Table 4. During this investigation, the highest concentration of TTHMs was found in VWTP for both the season. This may be attributed to the difference in THMs precursor content, RC, temperature, and other operational parameters (Padhi et al. 2019; Rathbun 1996). It may also be greatly affected by the geographical distribution and climatic conditions of the WTPs. The investigated range of

Table 4 Descriptive statistics of THM species at five utilities

THM species (µg/L)								
Water utilities	Pre-monsoon				Post-monsoon			
	CF	BDCM	DBCM	TTHMs	CF	BDCM	DBCM	TTHMs
VWTP								
Min	376	16	14	411	321.8	13.8	8	348
Max	385	20	18	419	330.5	19.5	17.5	367
Mean	380.85	18.25	15.52	414.62	325.97	16.92	13.14	356.03
SD	3.18	1.39	1.56	2.82	3.37	1.72	3.02	6.63
DWTP								
Min	354	15	8.3	381	283.50	15.00	8.00	310.00
Max	367	20	16	395	291.40	20.90	16.70	328.38
Mean	360.17	16.9	12.33	389.4	287.80	18.39	13.78	319.97
SD	4.53	1.61	2.47	4.06	2.81	2.21	2.86	6.87
RWTP								
Min	314	20	10	345	290.70	16.00	6.00	320.00
Max	329.7	23.7	18.2	370	299.70	23.40	18.60	338.25
Mean	324.26	21.71	14.22	360.19	295.94	20.13	13.92	329.99
SD	5.33	1.32	3.09	8.75	2.79	2.21	4.21	7.32
BWTP								
Min	312	18	5	342	301.40	19.00	10.20	334.00
Max	324	22	11	355	313.70	23.00	12.60	348.00
Mean	319.72	20.33	8.45	348.6	308.08	21.46	11.71	341.25
SD	4.02	1.33	1.77	5.14	3.93	1.15	0.75	4.48
IGWTP								
Min	344	17	10	375	305.00	18.40	11.70	340.50
Max	360	20.7	13.6	391	319.90	22.20	15.00	357.00
Mean	353.08	18.83	12.13	384.08	313.69	20.53	13.80	348.02
SD	5.17	1.26	1.26	5.54	4.49	1.41	1.22	5.78

TTHMs is consistent with the results obtained by Kumari and Gupta (2015) during their study of various water utilities situated in the Eastern part of India. Similarly, the higher concentration range of TTHMs was also monitored in other countries like Pakistan (575–595 µg/l) (Abbas et al. 2015) and Japan (378 µg/l) (Imo et al. 2007). Throughout the study, it was also noticed that the CF was the predominant compound among all four THMs species, which surpassed the WHO (300µg/l) and BIS (200µg/l) drinking water guideline value. The other two (BDCM and DBCM) were found well within the BIS and WHO standards, i.e., 60 and 100 µg/l, respectively. The BF was not detected in any of the water utilities because bromide ions were found below the detectable limit (BDL) (<0.1 mg/l). Source water with BDL bromide ions forms more chlorinated THMs than brominated THMs (Barrett et al. 2000; Nikolaou et al. 1999; Chowdhury et al. 2011; Imo et al. 2007; Lebel and Williams 1995), which can also be seen in the present study.

The percentage distribution of THMs species in various utilities and GG chromatograms (Fig. 4a, b) illustrates that CF shared more than 90% of TTHMs, followed by BDCM and DBCM. More than 90% of the THMs in the chlorinated drinking water supplies typically consisted of CF, while BDCM and DBCM contribute up to 2.1–14%. The observation of the present study was good in line with Zhang et al. (2011), where they reported that CF's contribution was up to 94% to that of other THMs compound in 13 WTPs of China.

Periodic fluctuations of THMs and their precursors in processed water

Assessment of periodic fluctuations in TTHMs and their precursors observed in all the water utilities are shown in Fig. 5a–d. There was substantial variation observed in the mean value of these species. The concentration range of

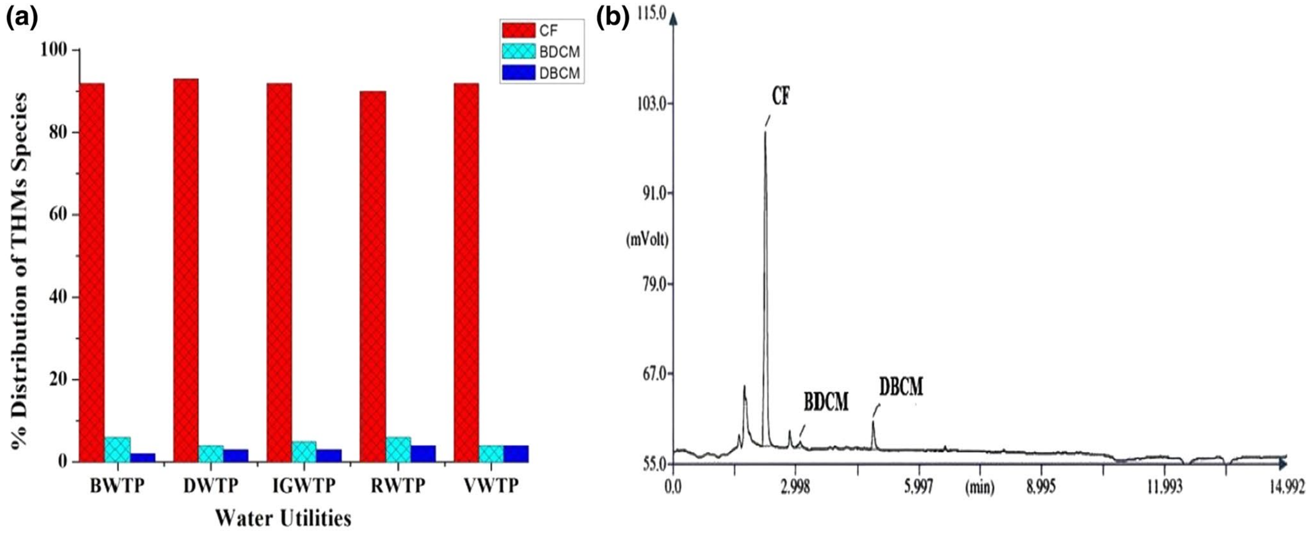


Fig. 4 **a** Percentage distribution of THMs species and **b** representative chromatogram of THMs of VWTPs

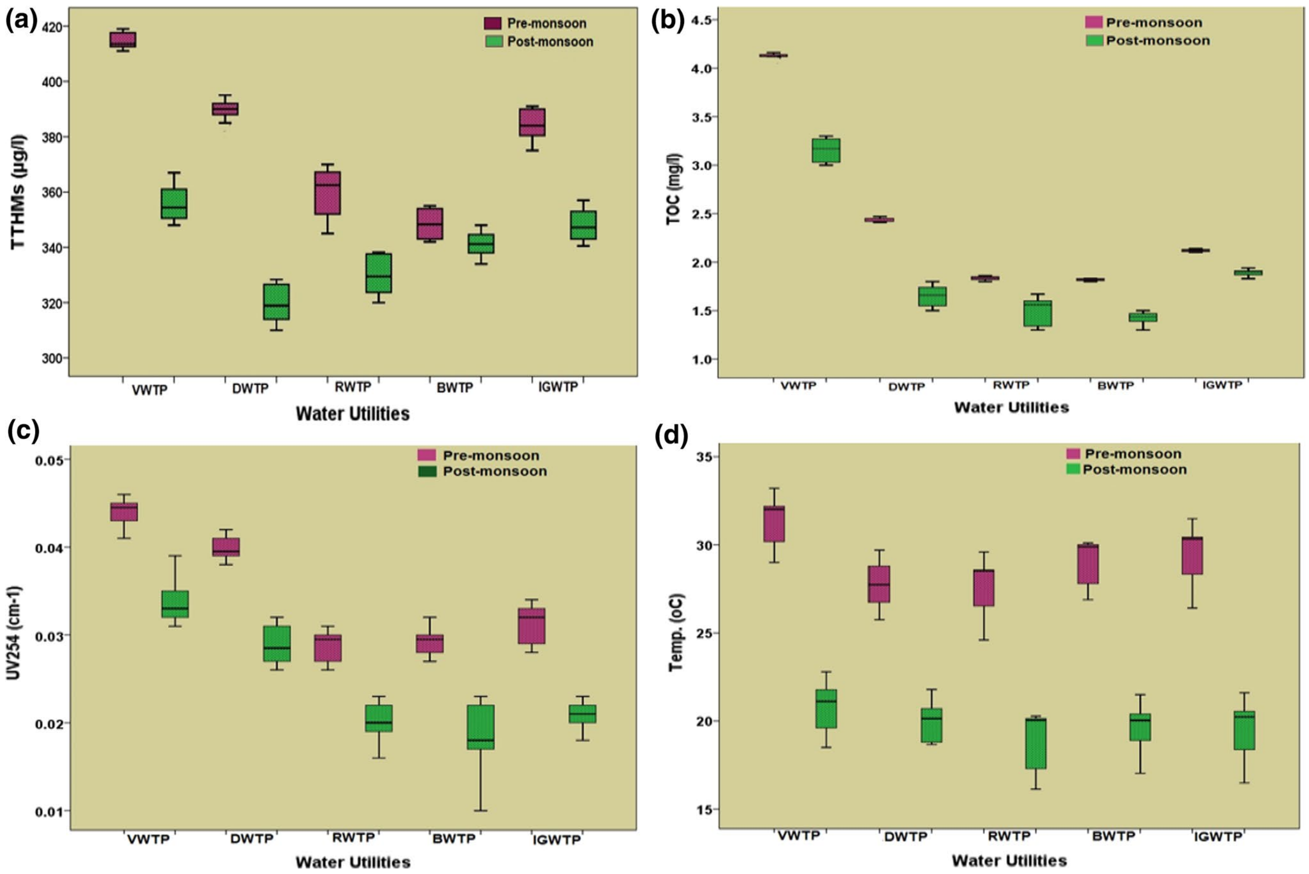


Fig. 5 Box and whisker plot of periodic fluctuations of **a** TTHMs, **b** TOC, **c** UV_{254} and **d** temperature at various utilities

TTHMs was spotted 1.12 ± 0.074 times higher in PrM than PoM. The TTHMs were in the order of VWTP followed by DWTP, IGWTP, RWTP, and BWTP for PrM, while in PoM, it was again VWTP followed by IGWTP, BWTP, RWTP, and DWTP. Organic content (TOC, UV_{254}) and the temperature were also appeared to be higher during the PrM, which may favor the higher THMs formation (Nikolaou et al. 1999). According to Rodriguez and Serodes (2001), rates of chlorine decay are high at elevated temperatures. Hence, it required higher doses of chlorine for treatment this season, which ultimately reacts with available NOM, thus providing more THMs (Uyak et al. 2008) in processed water. Besides, high organic content in water will also require a higher chlorine dose (Rodriguez and Serodes 2001). Temperature and NOM in water during PoM observed slightly lower, resulting in lesser chlorine demand (Rodriguez et al. 2003b); thus, comparatively lower THMs formed in this season. The observation in the present study is good in line with the finding of Wei et al. (2010), Rodriguez et al. (2004), and Toroz and Uyak (2005) for the drinking water distribution system.

Correlation analysis

In order to investigate the effects of NOM and other operational parameters on THMs formation, the Pearson correlation matrix was established (Kumari and Gupta 2015) (Table 5a, b).

Effect of NOM (TOC, DOC, and UV_{254})

TOC, DOC, and UV_{254} are essential surrogate measures of NOM, act as a key precursor for THMs formation (Padhi et al. 2019; Li and Mitch 2018; Sung et al. 2000). The Pearson correlation test confirmed that all these surrogates are strong and significantly correlated with TTHMs and each other. The THM formation rate is equal to that consumption of TOC, thus increasing in organic content of water, upswing the formation of THMs (Chang et al. 2001); Hasani et al. 2010; Arora et al. 1997). It was being reported previously that a water sample with high TOC can produce more THMs if enough RC is available (Babcock and Singer 1979). DOC constitutes approximately 83–98% of TOC in water and generally more representative of the soluble organic carbon than TOC (Owen et al. 1993). The strong and significant correlation between TOC and DOC under the study also supports this observation. Thus, concerning THMs formation, DOC follows the same trend parallel to TOC (Westerhoff et al. 2000; Müller 1998). UV_{254} is another essential key surrogate of NOM after TOC and DOC, provides an insight into the nature of organic content, and liable to form the THMs (Edzwald et al. 1985). The correlation coefficients of TOC with TTHMs were slightly higher than the DOC and UV_{254}, indicating TOC as more influential parameters. Moreover, it was also noticed that a slow reaction between chlorine and NOM results in the formation of THMs under second-order reaction to TOC, especially for the long-term (Draper and

Table 5 Pearson correlation matrix of variables with TTHMs

	pH	Temp	Alkalinity	RC	TOC	DOC	UV254		TTHMS
(a) Pre-monsoon									
pH	1	.525**	.415**	− .395**	.180	.130	.150		.406**
Temp		1	.597**	− .199	.594**	.508**	.510**		.584**
Alkalinity			1	− .028	.711**	.704**	.664**		.733**
RC				1	.546**	.585**	.555**		.471**
TOC					1	.977**	.973**		.949**
DOC						1	.963**		.929**
UV254							1		.927**
TTHMs									1
(b) Post-monsoon									
pH	1	.076	.766**	.342*	.608**	.581**	.725**	.513**	
Temp		1	− .237	− .328*	.425**	.436**	.199	.572**	
Alkalinity			1	.306*	.321*	.303*	.441**	.636**	
RC				1	.597**	.550**	.739**	.373*	
TOC					1	.935**	.931**	.815**	
DOC						1	.872**	.781**	
UV254							1	.719**	
TTHMs								1	

*Correlation is significant at the 0.05 level (2-tailed)

**Correlation is significant at the 0.01 level (2-tailed)

Smith 1998). Thus, it is a multistage process that operates through an initial reaction of TOC with residual chlorine followed by many possible pathways to produce THMs. The second step is found to be rate determining through which the reactive chlorinated intermediates are formed in the initial step (Trussell and Umphres 1978). With respect to NOM, DOC and UV_{254} were found second and third most influential parameters after TOC responsible for THMs formation, respectively. A similar investigation was also reported by Hua et al. (2015).

Effect of pH and alkalinity

In the present investigation, pH and alkalinity have shown a moderate and statistically significant correlation with TTHMs, respectively (Table 5a-b). pH showed a positive correlation with THMs; in other words, increasing in pH formation of THMs also increases (Roccaro et al. 2014; Hong et al. 2013; Kim et al. 2003). The oxidation process of chlorine is more prevalent in alkaline pH required more chlorine may support the greater THMs formation. In contrast, acidic pH lowered the reactivity of the chlorine pathway and strongly disfavored the THMs formation (Navalon et al. 2008). Besides, during the chlorination process, when chlorine comes in contact with water leads to the formation of hypochlorous acid (HOC1) and a hypochlorite ion ($OC1^-$). The formation of these two species is pH-dependent, as in acidic conditions, HOC1 is found to be dominated, whereas in alkaline pH $OC1^-$ (Uyak et al. 2005). Many researchers also widely accepted that base-catalyzed reactions play a major role in THM formation (Reckhow et al. 1990). In this regard, pH, and alkalinity seems to be an important operational parameter in controlling the THMs formation. The observation of the present study was well supported by the finding of Kim et al. (2003) and Oliver and Lawrence (1979).

Effect of temperature

THMs formation is proportional to the temperature; the higher the temperature greater the formation (Hua and Reckhow 2008). It was observed that every 10°C increase in the temperature doubles the rate, enhancing the activation energy of the reaction between organic matter and residual disinfectant (Engerholm and Amy 1983; Chowdhury and Champagne 2008). During the period of study, moderate relation was obtained between temperature and TTHMs. This observation is also good in line with the result of seasonal variation where PrM gives rise to the greater formation of THMs than PoM due to variation in temperature. Krasner (1999) also reported that the formation of THMs was higher during summer when there was high temp.

Effect of RC

The elevated range of RC present in treated water consequently increased the formation of chlorinated THMs (Chowdhury and Champagne 2008). However, the availability of organics beyond the chlorination breakpoint is so less than the THMs were not found to increase significantly after that point (Sung et al. 2000; Chowdhury and Champagne 2008). Pearson correlation test in this study indicated that RC has positively correlated with TTHMs. Hence, the THMs yield attains higher value due to the greater availability of RC (El-Dib and Ali 1995). This result appeared to be inconsistent with the finding of Al-Tmemy et al. (2018), Uyak et al. (2005), and Wei et al. (2010). Pearson correlation matrix of variables with TTHMs during PrM was found to exhibit similar trends as PoM.

A seasonal modeling approach for THMs formation

Modeling plays a very crucial role in predicting THMs formation in water supply systems. The study emphasizes the use of both conventional and models based on artificial intelligence to explore their accuracy and feasibility. The traditional modeling approach is based on multilinear regression, while machine language employs ANN and SVM to model the THMs formation in drinking water. At first, data of PrM season were utilized for model development and PoM data for validation studies. But, surprisingly, all the three models failed by giving significantly lower values of $R^2=0.5619$ (ANN), $R^2=0.5678$ (SVM), and $R^2=0.5670$ (MLR) (Fig. 6a–c). This indicates that the model developed from the PrM season cannot predict THMs in PoM owing to the seasonal variation in water quality parameters, especially the change in temperature, which largely influences the rate of THM formation (Rodriguez et al. 2003b; Hua and Reckhow 2008). To overcome the lacunae, separate models were developed to predict THMs during both PrM and PoM seasons (Fig. 7a–f). The performance data indicated that out of all the three models, ANN gave the most promising results with $R^2=0.9621$, followed by SVM ($R^2=0.9554$) and MLR ($R^2=0.9553$). The applicability of ANN can further be justified by significantly lower values of RMSE and MAPE than other models (Table 6). Moreover, the observed value of IA, closer to unity (0.99), also confirmed the better compliance of ANN than SVM and MLR models. This may be attributed to the higher generalization capacity of

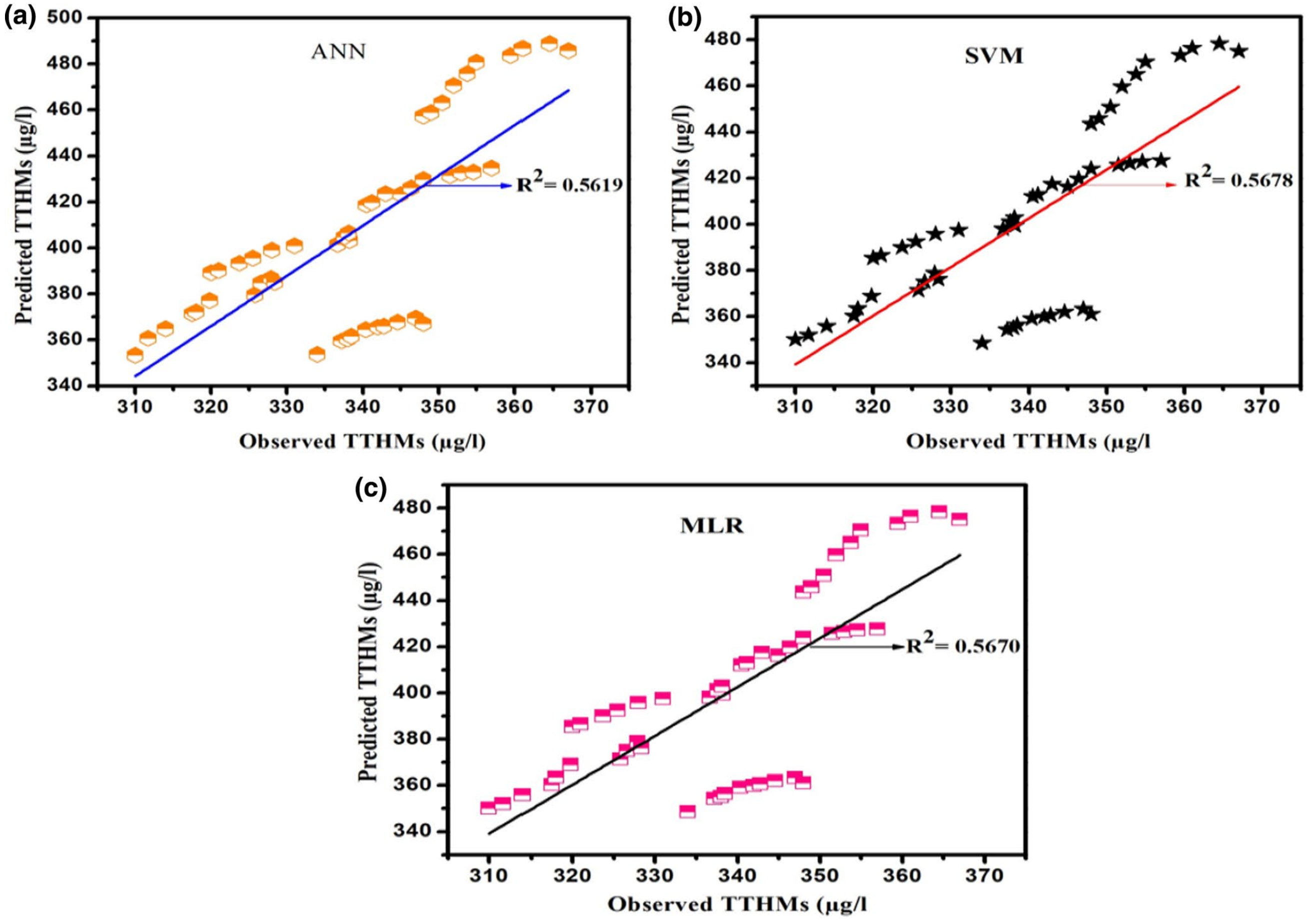


Fig. 6 Model plots of TTHMs using various approach **a** ANN, **b** SVM, and **c** MLR

ANN and its increased tolerance to noisy data (Rodriguez et al. 2003b; Milot et al. 2002; Hashem and Karkory 2007). Ye et al. 2011) also modeled DBPs in the drinking water of China using artificial neural networks and reported that the performance of the ANN model was excellent ($r > 0.84$). Significantly higher correlation for ANN in our study may be attributed to the precise calculation in neural networking, which eliminates the chances of any biased prediction on account of uneven distribution of modeling and testing data sets.

SVM and MLR models were also used in the study to model THMs in drinking water. The results dictated poor performance wrt ANN; however, close linearity between observed and predicted values was obtained for both SVM ($R^2 = 0.9554$) and MLR ($R^2 = 0.9553$). The values corresponding to MAPE and RMSE (Table 6) were also comparatively higher for SVM and MLR, indicating lesser suitability of these models than ANN. The variation in the models' performance may be due to the application of different prediction algorithms in machine language-based models. Hong et al. (2016) have developed an MLRs model for predicting THMs in the water distribution network of China, where they observed this regression model exhibited good accuracy and precision, as well as 86–97 % of the calculated fell within ±25% of the measured values. However, it is essential to note that the developed models were site-specific, and the predictive capabilities may vary according to the changes in environmental conditions.

Conclusion

The study established the concentration range of THMs and their precursors in drinking water utilities of five different Indian states. The study highlighted the need to adopt effective control measures for bringing down the high concentration of THMs to their permissible limit. THMs concentration showed a strong correlation with temperature followed by pH and NOM. Conclusive

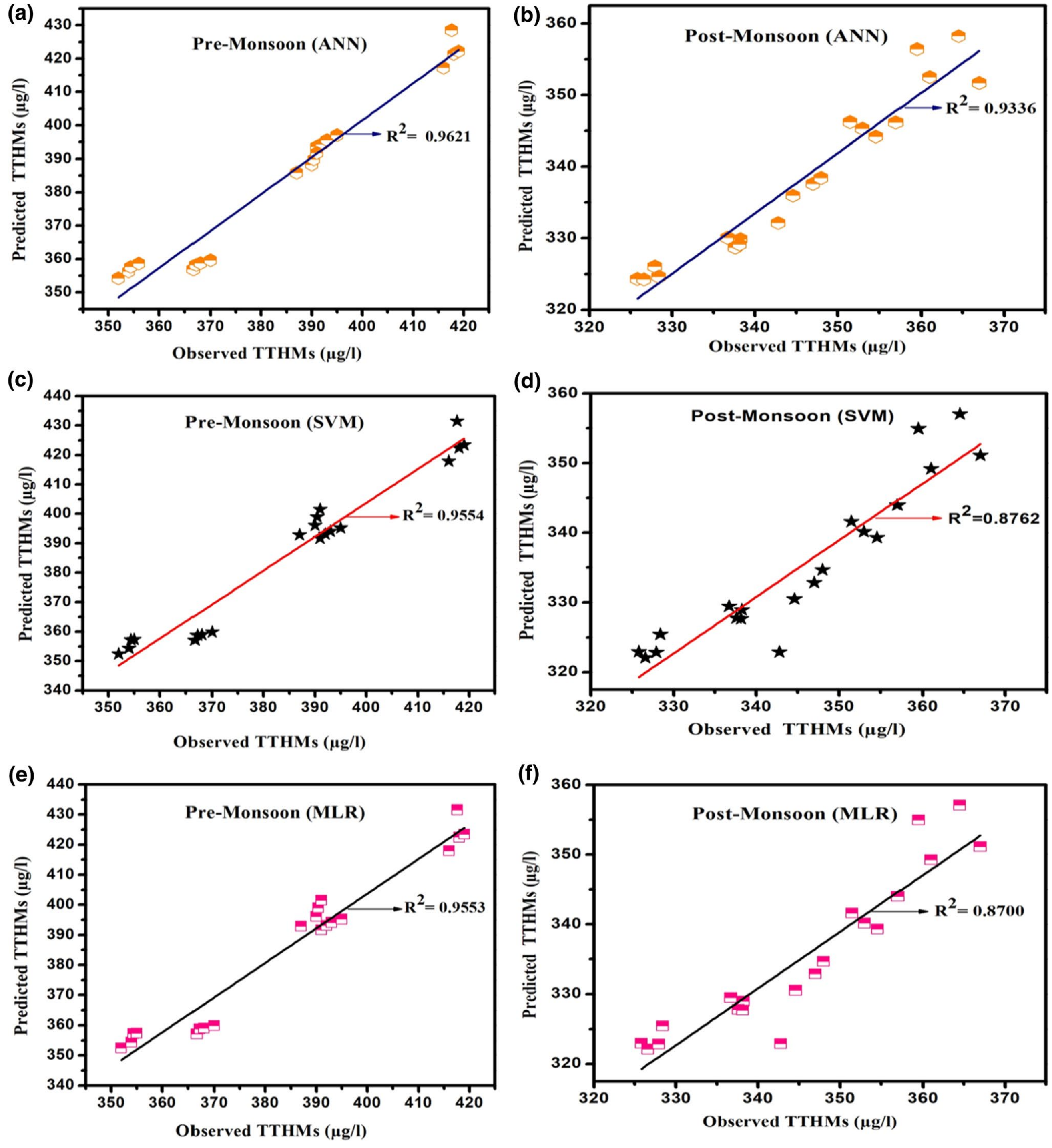


Fig. 7 **a**–**f** Season-wise validation model plots of TTHMs using various approach ANN, SVM, and MLR

Table 6 Descriptive performance of ANN, SVM, and MLR for both the season

	ANN		SVM		MLR	
	PrM	PoM	PrM	PoM	PrM	PoM
R^2	0.9621	0.933	0.955	0.876	0.955	0.870
IA	0.99	0.889	0.981	0.808	0.981	0.804
RMSE	5.33	8.16	6.53	11.01	6.54	11.21
MAPE	1.07	2.12	1.32	2.88	1.32	2.93

evidence from the analysis of performance data of various models dictated that the prediction of THMs through AAN was found relatively more precise than SVM and MLR models, hence, can be invariably adopted for quality control in drinking water supplies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13762-021-03392-1.

Acknowledgements The authors acknowledge the Department of Environmental Science & Engineering, Indian Institute of Technology (ISM), Dhanbad, India, to support the research work.

Authors' contribution JKM contributed to conceptualization, methodology, validation, investigation, data curation, and writing—original draft. SKG contributed to conceptualization, methodology, resources, writing—review and editing, and supervision.

Funding Not applicable.

Availability of data and materials Data collected and analyzed in this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethical approval No ethical (human or animal) approval was required to conduct the study.

References

Abbas S, Hashmi I, Rehman MSU, Qazi IA, Awan MA, Nasir H (2015) Monitoring of chlorination disinfection by-products and their associated health risks in drinking water of Pakistan. J Water Health 13(1):270–284

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(12):3393–3399

Amy GL, Chadik PA, Chowdhury ZK (1987) Developing models for predicting trihalomethane formation potential and kinetics. J Am Water Works Assoc 79(7):89–97

Arora H, LeChevallier MW, Dixon KL (1997) DBP occurrence survey. J Am Water Works Assoc 89(6):60–68

Azadi S, Karimi-Jashni A (2016) Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: a case study of Fars province. Iran Waste Manag 48:14–23

Babcock DB, Singer PC (1979) Chlorination and coagulation of humic and fulvic acids. J Am Water Works Assoc 71(3):149–152

Barrett SE, Krasner SW, Amy GL (2000) Natural organic matter and disinfection by-products: characterization and control in drinking water: an overview

BIS ISDWS (2012) Bureau of Indian Standards. New Delhi, pp 2–3

Chakraborty R (2017) Issues of water in India and the Health Capability Paradigm. Ethics Sci Environ Polit 17:41–50

Chang EE, Chiang PC, Ko YW, Lan WH (2001) Characteristics of organic precursors and their relationship with disinfection by-products. Chemosphere 44(5):1231–1236

Chowdhury S, Champagne P (2008) An investigation on parameters for modeling THMs formation. Global Nest J 10(1):80–91

Chowdhury S, Rodriguez MJ, Sadiq R (2011) Disinfection byproducts in Canadian provinces: associated cancer risks and medical expenses. J Hazard Mater 187(1–3):574–584

Cook DF, Wolfe ML (1991) A back-propagation neural network to predict average air temperatures. AI applications in natural resource management (USA)

Cotruvo JA (2017) 2017 WHO guidelines for drinking water quality: first addendum to the fourth edition. J Am Water Works Assoc 109(7):44–51

Di Cristo C, Esposito G, Leopardi A (2013) Modelling trihalomethanes formation in water supply systems. Environ Technol 34(1):61–70

Draper NR, Smith H (1998) Applied regression analysis, vol 326. Wiley

Edzwald JK, Becker WC, Wattier KL (1985) Surrogate parameters for monitoring organic matter and THM precursors. J Am Water Works Assoc 77(4):122–132

El-Dib MA, Ali RK (1995) THMs formation during chlorination of raw Nile river water. Water Res 29(1):375–378

Elshorbagy WE, Abu-Qdais H, Elsheamy MK (2000) Simulation of THM species in water distribution systems. Water Res 34(13):3431–3439

Engerholm BA, Amy GL (1983) A predictive model for chloroform formation from humic acid. J Am Water Works Assoc 75(8):418–423

Hasani A, Jafari MA, Torabifar B (2010) Trihalomethanes concentration in different components of water treatment plant and water distribution system in the north of Iran, pp 887–892

Hashem M, Karkory H (2007) Artificial neural networks as alternative approach for predicting trihalomethane formation in chlorinated waters. In: Eleventh international water technology conference

Haykin S (1999) Neural networks: a comprehensive foundation. McMaster University, Hamilton

Hodegeson JW (1990) Determination of chlorination disinfection by products and chlorinated solvents in drinking water by liquid-liquid extraction and gas chromatography with electron-capture detection. Environmental monitoring systems laboratory office of research and development method, p 551

Hong HC, Wong MH, Mazumder A, Liang Y (2008) Trophic state, natural organic matter content, and disinfection by-product formation potential of six drinking water reservoirs in the Pearl River Delta, China. J Hydrol 359(1–2):164–173

Hong H, Xiong Y, Ruan M, Liao F, Lin H, Liang Y (2013) Factors affecting THMs, HAAs and HNMs formation of Jin Lan Reservoir water exposed to chlorine and monochloramine. Sci Total Environ 444:196–204

Hong H, Song Q, Mazumder A, Luo Q, Chen J, Lin H, Yu H, Shen L, Liang Y (2016) 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 38(6):1303–1312

Hua G, Reckhow DA (2008) DBP formation during chlorination and chloramination: effect of reaction time, pH, dosage, and temperature. J Am Water Works Assoc 100(8):82–95

Hua G, Reckhow DA, Abusallout I (2015) Correlation between SUVA and DBP formation during chlorination and chloramination of NOM fractions from different sources. Chemosphere 130:82–89

Hur J, Lee BM, Lee S, Shin JK (2014) Characterization of chromophoric dissolved organic matter and trihalomethane formation potential in a recently constructed reservoir and the surrounding areas–Impoundment effects. J Hydrol 515:71–80

Imo TS, Oomori T, Toshihiko M, Tamaki F (2007) The comparative study of trihalomethanes in drinking water. Int J Environ Sci Technol 4(4):421–426

Kim J, Chung Y, Shin D, Kim M, Lee Y, Lim Y, Lee D (2003) Chlorination by-products in surface water treatment process. Desalination 151(1):1–9

Krasner SW (1999) Chemistry of disinfection by-product formation. Formation and control of disinfection by-products in drinking water, pp 27–52

Kulkarni P, Chellam S (2010) Disinfection by-product formation following chlorination of drinking water: artificial neural network models and changes in speciation with treatment. Sci Total Environ 408(19):4202–4210

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

Kumari M, Gupta SK (2018) Age dependent adjustment factor (ADAF) for the estimation of cancer risk through trihalomethanes (THMs) for different age groups-A innovative approach. Ecotoxicol Environ Saf 148:960–968

Lebel GL, Williams DT (1995) Differences in chloroform levels from drinking water samples analysed using various sampling and analytical techniques. Int J Environ Anal Chem 60(2–4):213–220

Li XF, Mitch WA (2018) Drinking water disinfection byproducts (DBPs) and human health effects: multidisciplinary challenges and opportunities, pp 1618–1689

Mahato JK, Gupta SK (2020) Modification of Bael fruit shell and its application towards Natural organic matter removal with special reference to predictive modeling and control of THMs in drinking water supplies. Environ Technol Innov 18:100666

Mahato JK, Kumar A, Gupta SK (2019) Efficiency evaluation of alternative disinfectant for the removal of THMs precursors in drinking water supplies of India. In: AIP conference proceedings, vol 2091, no 1. AIP Publishing LLC, p 020005

Marais SS, Ncube EJ, Msagati TAM, Mamba BB, Nkambule TTI (2019) Assessment of trihalomethane (THM) precursors using specific ultraviolet absorbance (SUVA) and molecular size distribution (MSD). J Water Process Eng 27:143–151

Milot J, Rodriguez MJ, Sérodes JB (2000) Modeling the susceptibility of drinking water utilities to form high concentrations of trihalomethanes. J Environ Manag 60(2):155–171

Milot J, Rodriguez MJ, Sérodes JB (2002) Contribution of neural networks for modeling trihalomethanes occurrence in drinking water. J Water Resour Plan Manag 128(5):370–376

Mishra ND, Dixit SC (2013) Trihalomethanes formation potential in surface water of Kanpur, India. Chem Sci Trans 2:821–828

Müller U (1998) THM in distribution systems. Water Supply 16(3):121–131

Navalon S, Alvaro M, Garcia H (2008) Carbohydrates as trihalomethanes precursors. Influence of pH and the presence of Cl- and Br- on trihalomethane formation potential. Water Res 42(14):3990–4000

Neter J, Wasserman W, Kutner MH (1990) Applied statistical models. Richard D. Irwin Inc., Burr Ridge

Nikolaou AD, Kostopoulou MN, Lekkas TD (1999) Organic by-products of drinking water chlorination. Global Nest Int J 1(3):143–156

Oliver BG, Lawrence J (1979) Haloforms in drinking water: a study of precursors and precursor removal. J Am Water Works Assoc 71:161–163

Owen DM, Amy GL Chowdhury ZK (1993) Characterization of natural organic matter and its relationship to treatability. Foundation and American Water Works Association

Padhi RK, Subramanian S, Mohanty AK, Satpathy KK (2019) Comparative assessment of chlorine reactivity and trihalomethanes formation potential of three different water sources. J Water Process Eng 29:100769

Peleato NM, Legge RL, Andrews RC (2018) Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. Water Res 136:84–94

Rathbun RE (1996) Regression equations for disinfection by-products for the Mississippi, Ohio and Missouri rivers. Sci Total Environ 191(3):235–244

Reckhow DA, Singer PC, Malcolm RL (1990) Chlorination of humic materials: byproduct formation and chemical interpretations. Environ Sci Technol 24(11):1655–1664

Roccaro P, Korshin GV, Cook D, Chow CW, Drikas M (2014) Effects of pH on the speciation coefficients in models of bromide influence on the formation of trihalomethanes and haloacetic acids. Water Res 62:117–126

Rodriguez MJ, Serodes JB (2001) Spatial and temporal evolution of trihalomethanes in three water distribution systems. Water Res 35(6):1572–1586

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(2):57–73

Rodriguez MJ, Milot J, Sérodes JB (2003a) Predicting trihalomethane formation in chlorinated waters using multivariate regression and neural networks. J Water Supply Res Technol AQUA 52(3):199–215

Rodriguez MJ, Vinette Y, Sérodes JB, Bouchard C (2003b) Trihalomethanes in drinking water of greater Québec region (Canada): occurrence, variations and modelling. Environ Monit Assess 89(1):69–93

Rodriguez MJ, Sérodes JB, Levallois P (2004) Behavior of trihalomethanes and haloacetic acids in a drinking water distribution system. Water Res 38(20):4367–4382

Sadiq R, Rodriguez MJ (2004) Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. Sci Total Environ 321(1–3):21–46

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

Sung W, Reilley-Matthews B, O'Day DK, Horrigan K (2000) Modeling DBP formation. J Am Water Works Assoc 92(5):53–63

Thacker NP, Kaur P, Rudra A (2002) Trihalomethane formation potential and concentration changes during water treatment at Mumbai (India). Environ Monit Assess 73(3):253–262

Toroz I, Uyak V (2005) Seasonal variations of trihalomethanes (THMs) in water distribution networks of Istanbul City. Desalination 176(1–3):127–141

Trussell RR, Umphres MD (1978) The formation of trihalomethanes. J Am Water Works Assoc 70(11):604–612

US EPA (2018) 2018 edition of the drinking water standards and health advisories. EPA 822-S-12-001

Uyak V, Toroz I, Meric S (2005) Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul. Desalination 176(1–3):91–101

Uyak V, Ozdemir K, Toroz I (2008) Seasonal variations of disinfection by-product precursors profile and their removal through surface water treatment plants. Sci Total Environ 390(2–3):417–424

Vapnik V (2013) The nature of statistical learning theory. Springer

Wei J, Ye B, Wang W, Yang L, Tao J, Hang Z (2010) Spatial and temporal evaluations of disinfection by-products in drinking water distribution systems in Beijing, China. Sci Total Environ 408(20):4600–4606

Westerhoff P, Debroux J, Amy GL, Gatel D, Mary V, Cavard J (2000) Applying DBP models to full-scale plants. J Am Water Works Assoc 92(3):89–102

Ye B, Wang W, Yang L, Wei J, Xueli E (2011) Formation and modeling of disinfection by-products in drinking water of six cities in China. J Environ Monit 13(5):1271–1275

Zhang J, Yu J, An W, Liu J, Wang Y, Chen Y, Tai J, Yang M (2011) Characterization of disinfection byproduct formation potential in 13 source waters in China. J Environ Sci 23(2):183–188