Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2020.25583

191 (2020) 372–380
July

Prediction models evaluating the production of trichloroacetaldehyde, a by-product of chlorination disinfection of raw water in northeast China

Xu Jiang[a,b], Munan Zhao[a], Hengyue Bao[a], Yazhuo Wang[a], Zhenfeng Shi[a], Chongwei Cui[a,*]

[a]*State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin 150090, China, Tel. +8618845181188; email: 18845181188@126.com (C. Cui), Tel. +13936512676; email: jiangxu921@126.com (X. Jiang), Tel. +15512850829; email: 15512850829@163.com (M. Zhao), Tel. +18845721532; email: 994849742@qq.com (H. Bao), Tel. +17839223070; email: 254997235@qq.com (Y. Wang), Tel. +13796040088; email: szfhit@126.com (Z. Shi)*
[b]*Harbin Water Supply Group Co., Ltd, Harbin 150001, China*

Received 14 September 2019; Accepted 3 February 2020

ABSTRACT

Raw water after chlorination disinfection will produce a variety of chlorinating organic compounds, known as disinfection by-products (DBPs). After investigation, the main DBPs exceeding the standard in northeast China is trichloroacetaldehyde (CH) stipulated in the standards for drinking water quality (GB5749-2006). However, the scale of water treatment plants in such areas is mostly smaller than 5×10^4 t/d and due to the problems of testing capacity and cost, the detection of CH cannot be realized. Once the raw water quality changes, the safety of the factory water quality cannot be guaranteed. In this paper, chlorination experiments were carried out on the raw water of the Mopanshan Reservoir in cold areas of northeast China, which contained a high level of natural organic matter. Multiple linear regression analysis was used to conduct statistical analysis on the results, and the formation model of CH based on water temperature (T), pH, turbidity, chlorine dose (Cl) and permanganate index was established. It can provide a good way to predict the formation of CH in water treatment plants with similar raw water characteristics and disinfection methods in cold areas of northeast China.

Keywords: Natural organic matter; Chlorination disinfection; Trichloroacetaldehyde; Disinfection by-products; Prediction models

1. Introduction

Trihaloacetaldehyde is an important disinfection by-product (DBP) formed during the chlorination process, and it's the next (to trihalomethanes and haloacetic acids) most prevalent DBP in drinking water [1]. CH exists in the form of chloral hydrate in water [2]. Due to its potential carcinogenic risk and other toxicity [3], the maximum contaminant level of CH in the Chinese Environmental Protection Ministry (GB5749-2006) is 0.01 mg/L [4].

Taking Harbin in northeast China as an example, its raw water is from Mopanshan Reservoir, located in the natural primitive forest region. Due to the vast upstream area of water source and the lush forest vegetation, the drinking water source is rich in natural organic matter, which is proved to be mainly humic acid (HA) and fulvic acid (FA), thus, it contains a high disinfection by-products formation potential (DBPFP) [5]. A new approach has been recently developed to remove intermediate halogenated aromatic DBPs by granular activated carbon (GAC) adsorption [6–8], instead of removing the organic matter (DBPFP) by GAC adsorption. However, the conventional treatment process adopted by the water plant cannot effectively remove the organic matter in raw water [9].

In this paper, drinking water from the Mopanshan Reservoir was disinfected after the process of coagulation–sedimentation–filtration. According to the United States Environmental Protection Agency (U.S. EPA) Method 551.1 [10], sampling was carried out for 36 consecutive months, and the main DBP that was found that is harmful to humans

* Corresponding author.

was CH. In northern China, many water plants have similar treatment practices and raw water characteristics to Harbin, all of which involve high levels of CH (>0.01 mg/L). Numerous studies have emerged during the last decades that use linear regression techniques or non-linear regression analyses to establish relationship models between the generation of DBPs and the indicators associated with chlorination processes such as total organic carbon, temperature, pH, and chlorine amount [11]. To effectively predict the production of CH and to better help water supply enterprises ensure the quality of finished water leaving the plant, it is necessary to establish a model relating to the raw water characteristics and treatment processes [11]. Due to the above reasons, in this study, a multiple regression analysis was used to establish a prediction model for CH by combining the chlorine dose and the main indicators that are routinely measured. This will be helpful in effectively controlling the generation of DBPs in the water treatment process by predicting the formation potential of CH, testing for it, and confirming its applicability.

2. Materials and methods

2.1. Sample

All samples concerned were from Harbin Pingfang Treatment Plant. The raw water comes from the Mopanshan Reservoir in the virgin forest, 180 km away from the water plant, which supplies water to nearly 3.4 million people and with a design capacity of 9×10^5 t/d. The conventional water treatment process including coagulation–sedimentation–filtration–disinfection, and the disinfection process was carried out before entering the water distribution systems, the liquid chlorine was used as a disinfectant. Fig. 1 shows each unit process and the sampling point locates in the water distribution pump room.

All samples were continuously monitored from January 2015 to February 2018 and samples were taken eight times a month. Samples of DBPs were dispensed into 100 mL brown glass bottles, and 0.1 g ascorbic acid and sodium thiosulfate were added. Typical data, including permanganate index (COD_{Mn}), turbidity, pH, and temperature were detected by using in-line sampling. The chlorine amount in the water plant was determined according to the actual amount of chlorine consumed by the plant. All sample vials were rinsed with tap water, washed with ultrapure water, and placed in an oven at 150°C for 2 h. After sampling, the bottles were stored in a dark environment at 4°C and brought back to the laboratory for analysis.

2.2. Analytical procedure

According to the requirements of the experiment, the main reagents and drugs to be used in the detection and analysis include CH, trichloroacetic acid and trichloromethane (TCM) standards (Brilliant Technology Co. Ltd., Beijing), methyl tert-butyl ether and methanol (chromatographic purity) (Anfu Scientific Instruments Co. Ltd., Shanghai). All reagents were used directly without further purification process. The experimental water was prepared by Millipore Milli-Q pure water system (resistivity ≥ 18.2 Ω cm). The glassware needed in the experiment was washed by ultrasound for 15 min, washed with tap water, and then washed with ultra-pure water 3 times, dried at 130°C for 24 h.

The detection methods of the main parameters used in the experiments include [12]: (1) CH: using Agilent GC7890B gas chromatography (Fairborn Precision Instruments Co. Ltd., Shanghai), chromatography column models for DB - 624, chromatographic column size was 30 m × 0.25 mm × 1.4 m. Split-flow injection (10:1) was used in the test process and the injection quantity was 1 μL, the temperature of the injection port was 200°C and the detector temperature was 250°C, the airflow control in 60.0 mL/min, hydrogen flow at 2.0 mL/min. High purity nitrogen was used as a carrier gas, the flow rate was 30.0 mL/min, the standard recovery rate was 102.4%, and the precision was 3.21%. (2) Turbidity: using Hashan 2100 turbidity meter, scattering light spectrophotometry method was carried. (3) pH: using Hashan 2100 turbidity meter, the glass electrode method was carried. (4) Permanganate index: by using the potassium permanganate oxidation method.

2.3. Data set

In this study, the raw water and finished water in 2015–2019 were monitored, according to environmental quality

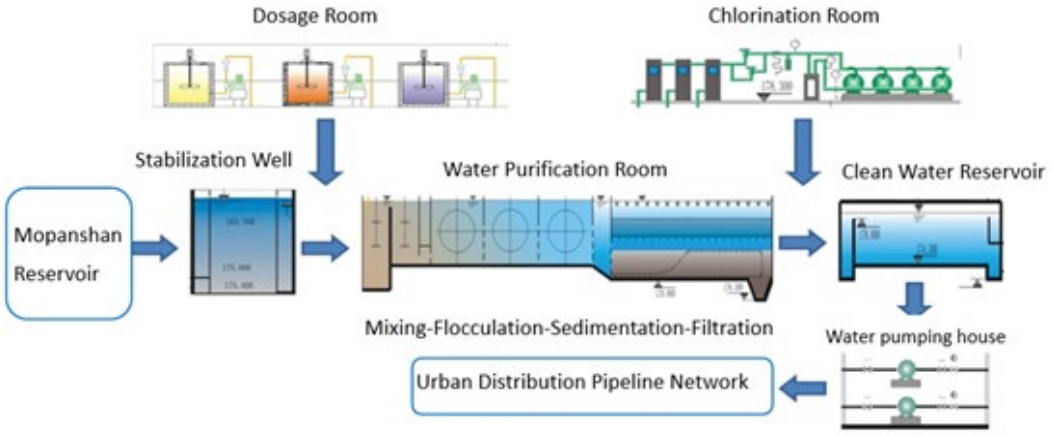

Fig. 1. Harbin Pingfang treatment plant, flow chart and sample collection point in water pumping house.

standards for surface water, 109 indicators were continuously monitored, among them, the indicators with large fluctuation range were temperature (T), pH, turbidity (TD), COD_{Mn}, the above indexes are also the indexes that are generally concerned by water treatment plants. In combination with the chlorine dose (Cl), the predictive relationship between each index and the amount of CH produced in finished water can be established. No CH was detected in raw water, so the background value can be counted as zero in the calculation process, and the relationship between the concentration of CH and the variable factors is more direct. Also, studies have shown that bromine has a certain effect on the formation of DBPs, but no bromine was detected in raw water. The chlorine dose was based on the measurement method commonly used by water treatment plants in China, COD_{Mn} is used to reflect the organic matter in water. It is an important parameter for water supply enterprises in China to master the water quality of raw water at present. Since there was no change in the operating time of the water plant disinfection (30 min), the disinfection residence time was not considered.

It can be seen from Table 1 that there are 374 sets of data, among which data are collected 4 times per month, and the average value is taken as the monthly collection value of the corresponding year, and a total of 95 sets of data are sorted out. Also, we adopt multiple regression analysis methods to establish CH formation model [13], considering that Heilongjiang Province has significant seasonal characteristics, especially the unique climate characteristics of northeast China in winter, so for the four distinct levels (seasons) that individually affect the dependent variable (CH), one dummy variable [autumn and winter (AW), the period was from September to February] was defined [14]. For example, the values of the dummy variable AW are:

$$\mathrm{AW} = \begin{cases} 1 \text{ if the collected data were the autumn and winter season} \\ 0 \text{ otherwise} \end{cases}.$$

3. Results and discussion

The focus of this study is that the data used in regression analysis have the regional characteristics of water quality in northeast China, all monitoring data were obtained from the Harbin monitoring sub-center of national urban water quality monitoring center. According to the range, mean value and standard deviation of the variables shown in Table 1. The raw water temperature (T) in the plant ranged from 2.3°C to 23.8°C, the pH value varied from 6.6 to 7.1, while the TD varied from 0.55 to 6.37 NTU, the chlorine dose (Cl) was lying in the range of 1.4–2.0 mg/L and COD_{Mn} ranged from 2.72 to 6.48 mg/L. It is important to note that T (2.3°C–23.8°C) with mean 10.72°C, TD (0.55–6.37 NTU) with mean 1.993 NTU and COD_{Mn} (2.72–6.48 mg/L) with mean 3.94 mg/L have a wide range of changes in above variables, mainly occurred in spring when the ice in the reservoir area starts to melt and in summer when the annual rainfall is large, while the ranges of Cl and pH were rather narrow.

The results show that CH was detected in the water samples, and almost all of them exceed the standard limit (0.01 mg/L) of the drinking water standards of China. At present, most of the literature studies mainly focus on TCM [15], so we refer to the research methods of TCM to study whether there is a correlation between CH and the above five variables. The variables were tested for normality, the Kolmogorov–Smirnov (K–S) test [16] was used to test the goodness of fit to the normal distribution based on the null hypothesis H_0: the random sample has the normal distribution, with unspecified mean and variance and the alternative hypothesis H_a: the distribution function of the sample is not random. The results of the normalization of the variables including CH content, COD_{Mn}, temperature (T), pH, chlorine dose (Cl) and turbidity (TD) (in the purpose of eliminating dimensional influence, the logarithm of each variable was used for calculation) are shown in Table 2, and the variables are not subject to normality by K–S normality test.

The correlation matrix between all the variables (the variable capacity is 95) is given in Table 3. CH is significantly and positively correlated with chlorine dose (r = 0.311) and pH (r = 0.375), respectively, which shows that the pH and the dosage of chlorine could promote the formation of CH. However, CH has a weak correlation with TD (r = 0.168), T (r = –0.012) and COD_{Mn} (r = –0.023), respectively. In addition, the chlorine dose has a positive correlation with TD (r = 0.348), T (r = 0.377), COD_{Mn} (r = 0.390), and the TD has a positive correlation with T (r = 0.459), COD_{Mn} (r = 0.242) and pH (r = 0.231), statistically means that an interaction may exist between some of the two independent variables. Hence, the relationship between the mean value of the dependent variable (CH) and one of the independent variables is dependent upon the value of the other independent variable.

Covariance analysis was used to judge whether the formation of CH is influenced by season [17]. Covariance analysis requires the dependent variable (CH) subject to normality

Table 1
Range of variables, the means and the standard deviations

Variables	Number of observations	Min.	Max.	Mean	Standard deviation
CH (mg/L)	374	0.0020	0.0430	0.0139	0.0067
Chlorine dose (Cl) (mg/L)	374	1.14	2.03	1.53	0.1844
pH	374	6.00	7.90	6.88	0.2159
Turbidity (TD) (NTU)	374	0.288	114.000	1.993	7.5888
Temperature (T) (°C)	374	0.41	23.80	10.72	5.7772
COD_{Mn} (mg/L)	374	1.60	7.68	3.94	0.8617
Autumn and winter (AW)	374	0	1	0.24	0.425

Table 2
Estimation of goodness of fit to the normal distribution

Variables	ln(CH)	ln(Cl_2)	ln(pH)	ln(TD)	ln(T)	ln(COD_{Mn})
Variable capacity	95	95	95	95	95	95
Test statistic	0.066	0.075	0.048	0.089	0.092	0.123
Asymp. Sig. (2-tailed)	0.200[a,b]	0.200[a,b]	0.200[a,b]	0.059[a]	0.044[a]	0.001[a]

[a]Lilliefors significance correction
[b]Lower bound of the true significance
Asymp. Sig. (2-tailed): asymptotic significance (2-tailed), a two-sided approximation of *P*

Table 3
Simple correlation between ln(CH), ln(Cl), ln(pH), ln(TD), ln(T), and ln(COD_{Mn})

	ln(CH)	ln(Cl_2)	ln(pH)	ln(TD)	ln(T)	ln(COD_{Mn})
ln(CH)	1.000					
ln(Cl_2)	0.311[b]	1.000				
ln(pH)	0.375[b]	–0.025	1.000			
ln(TD)	0.168	0.348[b]	0.231[a]	1.000		
ln(T)	–0.012	0.377[b]	–0.147	0.459[b]	1.000	
ln(COD_{Mn})	–0.023	0.390[b]	–0.190	0.241[a]	0.347[b]	1.000

[a]Correlation is significant at the 0.05 level (2-tailed)
[b]Correlation is significant at the 0.01 level (2-tailed)

and without abnormal value under each group of fixed factors (AW) under classification, it also requires to satisfy that fixed factor have no interaction with covariate variable (T) and the above two conditions can be tested by the test of homogeneity of variances. With the aid of SPSS software, the grouping variable (AW) was used for factor variable, the temperature (T) was used as a concomitant variable, the test results of normality, linearity, homogeneity, and interactivity are shown in Tables 4 and 5. It can be seen from Table 4 that the values of ln(CH) in different groups have passed the K–S normality test, and both of the linear goodness of fit of ln(CH) and ln(T) under different grouping conditions have exceeded 0.9 (R^2 = 0.936, 0.922 respectively), so it can be seen that CH has passed the normality and linearity test. Also, according to the *P*-value of the homogeneity test [18] in the first row of Table 5, the null hypothesis (H_0: the two sets of data come from the same sample) is accepted at the significance level of 0.05 (P = 0.150 > 0.05), that is, the two sets of data are with homogeneity of variance. According to the interaction test in the second row, the *P*-value (p = 0.668) is greater than the significance level of 0.05, indicating that the seasonal dummy variable (AW) and ln(T) have no interaction. Finally, the boxplot was drawn. It can be seen that ln(CH) has no outliers in each group. Thus can undertake the covariance test (ANCOVA) [19], the results are shown in Table 6, in the third row of Table 6, *P*-value (P = 0.000 < 0.05) shows that under the condition of the significance level of 0.05, AW passed the significance test, so it indicates that the dummy variable (AW) has an impact on CH content after excluding the influence of water temperature.

By observing the scatter plots between the dependent variable and independent variables, it can be found that it's unable to specify a better relationship between the dependent

Table 4
Test of normality and linearity

AW		K–S			R^2	
		Statistic	*Df*	Sig.	$R^{2\,a}$	Adjusted R^2
ln(CH)	0	0.091	49	0.200	0.937	0.936
	1	0.128	46	0.058	0.924	0.922

Df: Degrees of freedom
Sig.: Level of significance
[a]Linear regression through the origin

Table 5
Homogeneity and interaction test

Variable	*F*	*P*
ln(CH)	2.105	0.150
AW × ln(T)[a]	0.185	0.668

Significance level is 0.05
[a]Interaction between AW and T

variable and independent variables, therefore, to preliminarily screen the independent variables' form, the curve fitting model between a dependent variable and independent variables is considered. As shown in Table 7, the ln(CH) and five independent variables were fitted with linear, inverse, quadratic and cubic polynomial models respectively, the values in the table are the goodness of fit of the corresponding model [20], *F* statistic value and its corresponding *P*-values, significance test *P*-value of the corresponding parameter of each model, in this paper, the initial variable form is finally determined by integrating the three indexes of the goodness

Table 6
Analysis of covariance (ANCOVA) and least square difference test for detecting differences among the seasons of the year for CH concentrations

Source	Sum of squares	Df	Mean square	*F*	*P*
Model	1,799.652[a]	3	599.884	3,229.596	0.000
ln(T)	0.035	1	0.035	0.190	0.664
AW	109.441	2	54.720	294.599	0.000
Error	17.089	92	0.186		
Total	1,816.741	95			

[a]R^2 = 0.991 (Adjusted R^2 = 0.990)
Significance level is 0.05
Df: Degrees of freedom

Table 7
Selection of independent variable forms

Variables	Equation	Model summary			Parameter significant estimates		
		R^2	*F*	Sig.	b1(sig.)	b2(sig.)	b3(sig.)
ln(Cl)	Linear	0.899	839.681	0.000	0.000		
	Inverse	0.897	821.126	0.000	0.000		
	Quadratic	0.980	2,262.011	0.000	0.000	0.000	
	Cubic	0.989	2,817.971	0.000	0.000	0.000	0.000
ln(pH)	Linear	0.989	8,690.633	0.000	0.000		
	Inverse	0.991	10,579.23	0.000	0.000		
	Quadratic	0.992	5,644.055	0.000	0.000	0.000	0.000
	Cubic	0.992	5,650.407	0.000	0.000	0.366	0.000
ln(TD)	Linear	0.052	5.198	0.025	0.025		
	Inverse	0.001	0.070	0.791	0.791		
	Quadratic	0.098	5.044	0.008	0.819	0.033	
	Cubic	0.436	23.745	0.000	0.579	0.000	0.000
ln(T)	Linear	0.930	1,255.674	0.000	0.000		
	Inverse	0.914	997.641	0.000	0.000		
	Quadratic	0.985	3,027.887	0.000	0.000	0.000	
	Cubic	0.989	2,812.915	0.000	0.000	0.000	0.000
ln(COD_{Mn})	Linear	0.975	3,722.965	0.000	0.000		
	Inverse	0.975	3,706.952	0.000	0.000		
	Quadratic	0.990	4,509.037	0.000	0.000	0.000	
	Cubic	0.990	3,176.683	0.000	0.000	0.001	0.014

Sig.: Level of significance

of fit, significance test and coefficient significance test of the model, as can be seen in Table 7, which are ln(Cl), $(\ln(Cl))^2$, $(\ln(Cl))^3$, ln(pH), $(\ln(pH))^2$, 1/ln(pH), ln(T), $(\ln(T))^2$, $(\ln(T))^3$, ln(COD_{Mn}), $(\ln(COD_{Mn}))^2$, $(\ln(COD_{Mn}))^3$. Among which, TD and CH only pass the coefficient test of the linear fitting, and their goodness of fit is very low, indicating that there is no obvious and commonly known relationship between ln(CH) and ln(TD), so the TD was not introduced in this model.

In this paper, considering that the five independent variables, as well as the seasonal variable, have interactions, which will influence CH content, so use the single-variable linear model analysis of SPSS to judge the interactions of variables [21], the results showed that ln(TD) and ln(T), ln(TD) and ln(COD_{Mn}), ln(T) and ln(COD_{Mn}) do have interactions, also, considering the effect of initial CH content on current CH content, therefore, the first-order lag term of CH $(\ln(CH))_{t-1}$ is introduced.

In this paper, the stepwise regression method [22] was used to establish multiple regression model, the method of stepwise regression equations according to the sum of squares of partial regression and select the variables from the equation, so that the explanatory variables retained in the model are both important and not severely multicollinearity. We introduced a total of 17 independent variables obtained above and obtained the optimal set of explanatory variables through four stepwise regression. The model test, variance analysis table, and equation significance are shown in Tables 8 and 9. From these relationships, the following

Table 8
Model test and analysis of variance (ANOVA)

Model summary[a,b]					
R	R^{2b}		Adjusted R^2		Standard errors
0.998[a]	0.997		0.997		0.2552
ANOVA[a,b]					
	Sum of squares	Df	Mean square	*F*	Sig.
Regression	1,796.137	4	449.034	6,896.436	0.000
Residual	5.860	90	0.065		
Total	1,801.997	94			

[a]represent for ln(CH)
[b]means linear regression through the origin
Df: Degrees of freedom
Sig.: Level of significance

Table 9
Regression coefficients, standard errors, *t*-values and level of significance for the multiple regression model of CH concentrations

Variables		Unstandardized coefficients		Standardized coefficients			Collinearity statistics	
		B	Std. error	Beta	*t*	Sig.	Tolerance	VIF
1	$(\ln(CH))_{t-1}$	0.998	0.007	0.998	146.782	0.000	1.000	1.000
2	$(\ln(CH))_{t-1}$	0.753	0.068	0.753	11.027	0.000	0.009	113.879
	1/ln(pH)	–2.076	0.576	–0.246	–3.602	0.001	0.009	113.879
3	$(\ln(CH))_{t-1}$	0.696	0.068	0.696	10.273	0.000	0.008	122.797
	1/ln(pH)	–2.868	0.606	–0.340	–4.736	0.000	0.007	137.804
	$(\ln(Cl))^2$	0.830	0.264	0.042	3.138	0.002	0.210	4.758
4	$(\ln(CH))_{t-1}$	0.676	0.067	0.676	10.029	0.000	0.008	125.570
	1/ln(pH)	–2.853	0.596	–0.338	–4.789	0.000	0.007	137.826
	$(\ln(Cl))^2$	1.177	0.312	0.059	3.774	0.000	0.146	6.843
	$(\ln(COD_{Mn}))^3$	–0.062	0.031	–0.040	–2.018	0.047	0.092	10.905

Df: Degrees of freedom
Sig.: Level of significance
B: Regression coefficients
Std. Error: Standard errors
Beta: Regression coefficients

equation can be deduced, in which the variables of water temperature and seasonal variable are excluded, indicating that although these two variables may have an impact on the content of CH, their effects are not obvious or do not have an impact on other variables, so they are not introduced into the equation in the stepwise regression method.

$$\ln(CH)_t = 0.676 \times \ln(CH)_{t-1} - 2.853 \times \frac{1}{\ln(pH)_t} + 1.177 \times (\ln(Cl_2)_t)^2 - 0.062 \times (\ln(COD_{Mn})_t)^3$$

From Table 8 we can see that the fitting degree of the model (R^2 = 0.997, adjust R^2 = 0.997), which shows that the model has 99.7% interpretability, furthermore, it can be found that the *P*-value of the analysis of variance (*F*-test) of the model is 0.000, which is significantly smaller than the significance level of 0.05, therefore the hypothesis that the population regression coefficient is 0 can be significantly rejected, that is to say, there is a linear relationship between dependent variables and independent variables in the established model, which means that the model is valid. In the fourth row of Table 9, all the *P*-values of the coefficient significance test of all independent variables are less than 0.05, so the coefficient significance of the four variables is not 0, and all the parameters of the model are significant.

In addition to the passing of the significance test and the significance test of the coefficient of the above model, the multiple regression equation also requires the model to satisfy the residual normality, sequence irrelevance and homoskedasticity, therefore, three properties of the model residuals are tested in this paper, these results are shown in Figs. 2–4, Fig. 2 is the P-P diagram of residual normality test, Fig. 3 shows the heteroscedasticity test of residuals [23], the *x*-coordinate

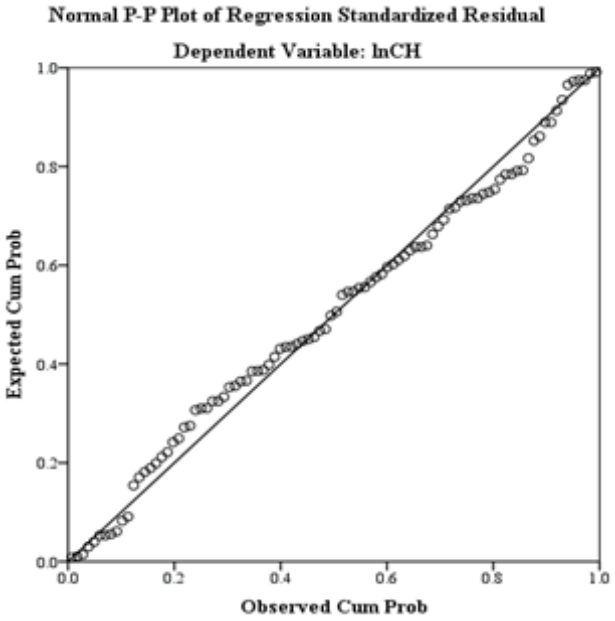

Fig. 2. Normal probability plot for the multiple regression model of CH (P-P diagram of residual normality test).

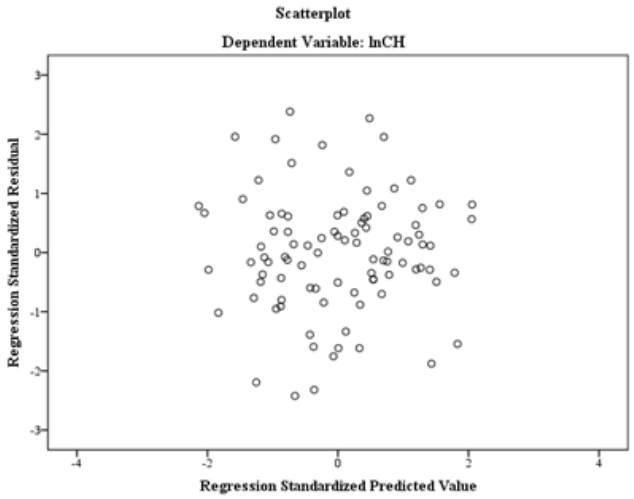

Fig. 3. Normal probability plot for the multiple regression model of CH. Expected normal values vs. residuals.

is the predicted value of the model, the *y*-coordinate is the residual value of the model, Fig. 4 shows the sequence correlation test of residuals [24], the abscissa is the first-order lag term of residuals, and the ordinate is the residuals, it can be seen from the Fig. 4 that there is no obvious linear trend, so the residuals do not have sequence correlation. To sum up, all the test indexes of the model constructed are qualified.

By using the above equation, 95 sets of data were fitted, Figs. 5a and b show that the predicted values have a high fitting degree with the real values, the absolute value of its error mostly within 0.5, and the real values' overall tendency is consistent with that of the predicted values. Model validation

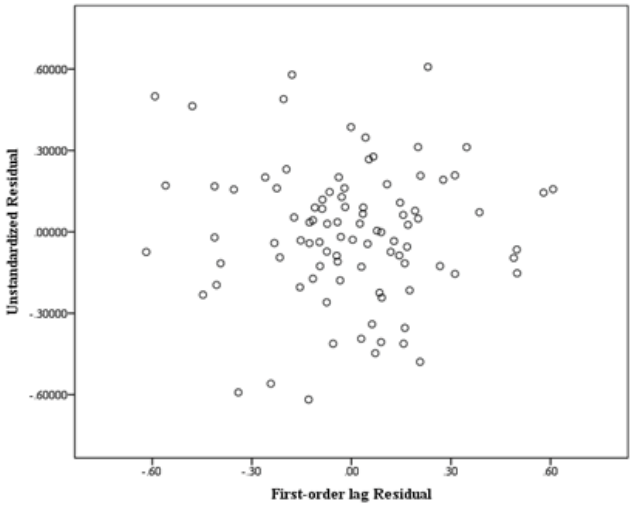

Fig. 4. Serial correlation test of residuals.

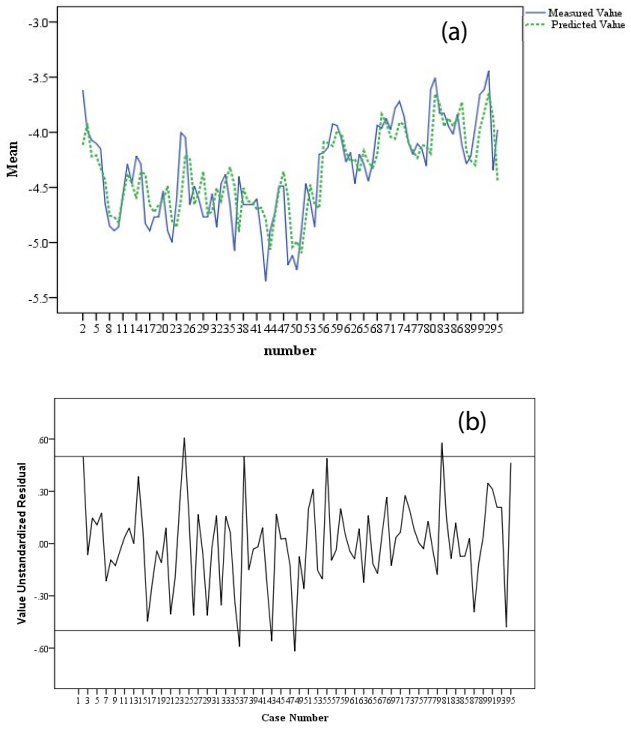

Fig. 5. The goodness of fit of the model of the training set: predicted and measured values of CH (a) and validation of the model: predicted and measured values of CH (b).

is settled by using data from January 2019 to June 2019, as shown in Figs. 6a and b. Fig. 6b shows that the test error of the absolute value is less than 0.006, the goodness of fit is 0.31 so that the model has a certain prediction ability. The establishment of a multiple regression model is to obtain a better CH formation tendency, but because the multiple regression model itself has some defects such as multicollinearity, the model does not necessarily have a high accuracy.

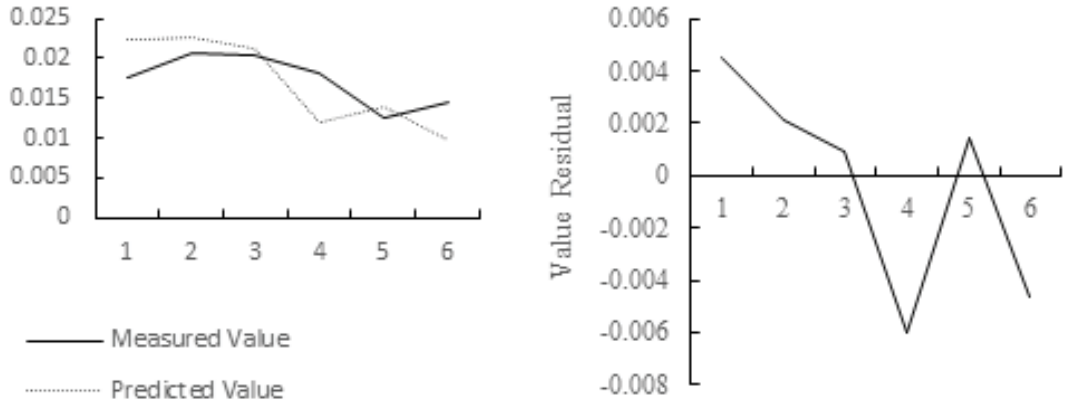

Fig. 6. The goodness of fit of the model of the test set: predicted and measured values of CH (a) and validation of the model: predicted and measured values of CH (b).

4. Conclusions

The CH concentration of reservoir water disinfected with chlorine in cold areas of northeast China was predicted by multiple regression analysis. The parameters chosen here are generally conventional index used to evaluate water quality in China's water plants, which has a certain universality. Therefore, we infer that this equation form can be used to provide water plants with similar raw water characteristics and disinfection methods for CH concentration calculations.

In the established model, although the logarithmic form of various variables is adopted, so that the variation tendency of the logarithmic form is consistent with the original function. Therefore, according to the equation, the formation of CH is affected by the initial content of CH, the pH value, the chlorine dose, and COD_{Mn}. Specific analysis is as follows: the COD_{Mn} has a negative effect on the formation of CH, but the effect is not significant; The logarithm of the current pH value, the current chlorine dose and the initial concentration of CH in raw water have a positive effect on the logarithm of the current CH content, among which, the pH value and the chlorine dose have the most positively affection on the CH content.

Also, it must be noted that the complexity of the formation of the CH makes it difficult to establish a truly universally applicable model. The current model is limited to field data within a specific dataset. And for a wider range of applications, the recalibration process must be considered and the optimization technique was introduced to modify the coefficient values. Besides, we need to recognize that this paper believes that the HA and FA are the precursors of CH [1]. However, studies have shown that the yield and rate of different precursors for CH are different. Therefore, in the view of the need to develop approaches of universal applicability, future modeling structures should focus on the use of different precursor types and the relationship between them and raw water characteristics. In this direction, DOC and UV_{254} seem to be reliable alternative indicators for characterizing organic precursors, since it has proved to be more widely used in experimental and simulated data, further research will be conducted in the future.

References

[1] L. Barrott, Chloral hydrate: formation and removal by drinking water treatment, J. Water Supply Res. Technol. AQUA, 6 (2014) 381–390.
[2] A. Dąbrowska, J. Nawrocki, Controversies about the occurrence of chloral hydrate in drinking water, Water Res., 43 (2009) 2201–2208.
[3] I. Zimoch, E. Lobos, Evaluation of health risk caused by chloroform in drinking water, Desal. Water Treat, 57 (2016) 1027–1033.
[4] China National Standardization Management Committee, Standards for Drinking Water Quality GB 5749-2006, Chinese Ministry of Health, Beijing, 2006, p. 4.
[5] M. Rajca, A. Wlodyka-Bergier, M. Bodzek, T. Bergier, MIEX (R) DOC process to remove disinfection by-product precursors, Desal. Water Treat., 64 (2017) 372–377.
[6] J. Jiang, X. Zhang, X. Zhu, Y. Li, Removal of intermediate aromatic halogenated DBPs by activated carbon adsorption: a new approach to controlling halogenated DBPs in chlorinated drinking water, Environ. Sci. Technol., 51 (2017) 3435–3444.
[7] J. Jiang, W. Li, X. Zhang, J. Liu, X. Zhu, A new approach to controlling halogenated DBPs by GAC adsorption of aromatic intermediates from chlorine disinfection: effects of bromide and contact time, Sep. Purif. Technol., 203 (2018) 260–267.
[8] J. Jiang, X. Zhang, A smart strategy for controlling disinfection by-products by reversing the sequence of activated carbon adsorption and chlorine disinfection, Sci. Bull., 63 (2018) 1167–1169.
[9] P. Roy, D. Kumar, M. Ghosh, A. Majumder, Disinfection of water by various techniques - comparison based on experimental investigations, Desal. Water Treat., 57 (2016) 28141–28150.
[10] B.K. Koudjonou, G.L. LeBel, Halogenated acetaldehydes: analysis, stability and fate in drinking water, Chemosphere, 64 (2006) 795–802.
[11] R. Sadiq, M. Rodriguez, Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review, Sci. Total Environ., 321 (2004) 21–46.
[12] J. Xu, Z. Munan, J. Feng, C. Chongwei, Study on Model Prediction of Trichloromethane Generation as a by-product of Chlorination of Raw Water in Northeast China, Journal of Harbin Institute of Technology, Harbin, China, 2020, (in Chinese).
[13] K. Preacher, P. Curran, D. Bauer, Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis, J. Edu. Behav. Stat., 31 (2006) 437–448.
[14] J. Angrist, Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice, J. Bus. Econ. Stat., 19 (2001) 2–16.

[15] J. Lin, X. Chen, Z. Ansheng, H. Hong, Y. Liang, H. Sun, H. Lin, J. Chen, Regression models evaluating THMs, HAAs and HANs formation upon chloramination of source water collected from Yangtze River Delta Region, China, Ecotoxicol. Environ. Saf., 160 (2018) 249–256.
[16] A. Justel, D. Pena, R. Zamar, A multivariate Kolmogorov-Smirnov test of goodness of fit, Stat. Probab. Lett., 35 (1997) 251–259.
[17] J. Willett, A. Sayer, Using covariance structure-analysis to detect correlates and predictors of individual change over time, Psychol. Bull., 116 (1994) 363–381.
[18] D. Gebregiorgis, D. Rayner, H. Linderholm, Does the IOD independently influence seasonal monsoon patterns in Northern Ethiopia?, Atmosphere-Basel, 10 (2019), doi: 10.3390/atmos10080432.
[19] B. Byrne, R. Shavelson, B. Muthen, Testing for the equivalence of factor covariance and mean structures - the issue of partial measurement invariance, Psychol. Bull., 105 (1989) 456–466.
[20] G. Cheung, R. Rensvold, Evaluating goodness-of-fit indexes for testing measurement invariance, Struct. Equation Model., 9 (2002) 233–255.
[21] Y. Kirby, R. McNew, J. Kirby, R. Wideman, Evaluation of logistic versus linear regression models for predicting pulmonary hypertension syndrome (Ascites) using cold exposure or pulmonary artery clamp models in broilers, Poultr. Sci., 76 (1997) 392–399.
[22] T. Burkholder, R. Lieber, Stepwise regression is an alternative to splines for fitting noisy data, J. Biomech., 29 (1996) 235–238.
[23] J. Lin, Y. Zhao, H. Wang, Heteroscedasticity diagnostics in varying-coefficient partially linear regression models and applications in analyzing Boston housing data, J. Appl. Stat., 42 (2015) 2432–2448.
[24] L. Godfrey, Alternative approaches to implementing Lagrange multiplier tests for serial correlation in dynamic regression models, Comput. Stat. Data Anal., 51 (2007) 3282–3295.