Membrane and Water Treatment, *Vol. 16, No. 4 (2025) 195-202*
https://doi.org/10.12989/mwt.2025.16.4.195

Development of a real-time THMs prediction technology based on artificial intelligence and sensors

Yunseok Choi and Doo-il Kim*

Civil and Environmental Engineering, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, 16890, Republic of Korea

(Received May 11, 2025, Revised June 8, 2025, Accepted June 24, 2025)

Abstract. This study presents the development of a real-time prediction model for trihalomethanes (THMs) using deep learning and sensor-based data from a drinking water treatment plant. As chlorine remains the most widely used disinfectant due to its cost-effectiveness, the formation of carcinogenic THMs has become a critical concern, especially with multi-point chlorine injection strategies. Traditional THMs prediction models have faced limitations due to the complex nature of organic matter and the lack of real-time data availability. In this study, a deep learning model utilizing artificial neural networks (ANNs) was trained on 171 data points comprising real-time flow and water quality variables. Model performance was evaluated using mean absolute error (MAE), mean squared error (MSE), and R^2 metrics under different activation functions including rectified linear unit (ReLU), hyperbolic tangent (tanh), and exponential linear unit (ELU). The ReLU-based model achieved the best performance with an R^2 of 0.76, indicating reliable prediction without clear over- or underestimation tendencies. Variants of the model with reduced input variables and adjusted output ranges were also tested for model improvement. These modifications showed reduced error variability and enhanced model stability, albeit with slightly reduced R^2 values (approximately 0.72). This approach demonstrates the potential of real-time artificial intelligence (AI)-driven THMs forecasting to support optimized chlorine dosing and regulatory compliance in drinking water treatment facilities. Unlike previous studies focused solely on accuracy, this research emphasizes practical applicability using sensor-acquirable parameters, providing a scalable and real-time decision-support tool for THMs management.
Keywords: activation function; artificial intelligence; real time chlorine injection control; THMs prediction; water quality sensor

1. Introduction

Chlorine, ozone, chlorine dioxide, and ultraviolet (UV) are used as disinfectants in drinking water treatment plants. Among these methods, chlorine has been the optimal choice because of its high efficiency and cost-effectiveness and provides residual disinfection unlike ozone or UV: remaining active in the distribution system over time. Since the first chlorination disinfection was conducted in 1908 at Bubbly Creek in Chicago, USA, chlorine disinfection has been widely adopted and has become the most commonly used disinfection method worldwide (McGuire 2013). However, chlorine reacts with organic substances in water to form carcinogenic disinfection by-products (DBPs) (Bae *et al.* 2020a, Singer 1999). The main chlorination by-products include trihalomethanes (THMs) and haloacetic acids (HAAs) (Bae *et al*. 2020b, Singer 1999, Singer and Reckhow 2011). The formation of THMs is influenced by factors such as pH (higher pH favors THMs formation), temperature (higher temperature increases THMs generation), chlorine dose, contact time, the type and concentration of organic matter, and the presence of bromide (which leads to the formation of brominated THMs) (Singer and Reckhow 1999, USEPA 2006, Richardson 2003). The formation of THMs depends on the multiple characteristics of the natural organic matter, including molecular weight distribution, aromaticity, degree of saturation, and hydrophilicity. Therefore, it is difficult to accurately predict the generation of disinfection by-products (Singer and Reckhow 1999, Safeer *et al*. 2022, Singh and Gupta 2012).

In drinking water treatment plants, chlorine is injected into the clear well and the chlorine dosage has often been adjusted based on the operator's acquired experience, aiming to meet the residual chlorine regulation (AWWA 2006). Recently, due to the deterioration of raw water quality caused by low source water quality caused by eutrophication and urbanization (Lee *et al*. 2024, Oh *et al*. 2023b, Youn *et al*. 2024), chlorine is sometimes injected at the receiving well or between the sedimentation basin and the filtration unit. As chlorine usage increases, the potential for DBP formation also rises, and there have been cases where the concentration of THMs temporarily exceeded regulatory limits in South Korea (Lee *et al*. 2007, Oh *et al*. 2023a). The recent trend of the multiple point chlorine injection in the treatment process has made THMs management even more difficult. During chlorine dosing, both residual chlorine level and THMs concentration should be considered in real time, which is not feasible in practice. The real-time estimation of the THMs formation is required to control the residual chlorine level and THMs concentration by optimizing the chlorine dosing. There are a few studies

*Corresponding author, Professor,
E-mail: dikim21@dankook.ac.kr

ISSN: 2005-8624 (Print), 2092-7037 (Online)

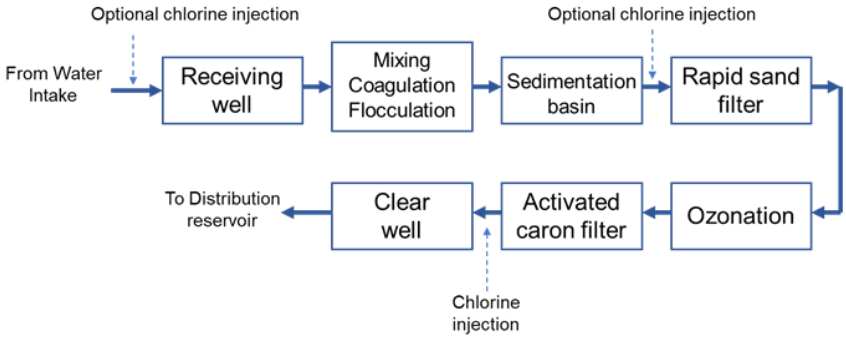

Fig. 1 Process diagram of the drinking water treatment

that have attempted to predict THMs concentrations using mathematical models based on raw water qualities and operational parameters (Brown 2009, Hogue *et al*. 2023, Osorio *et al*. 2011, Tsitsifli and Kanakoudis 2020). However, these models have limitations in real-time prediction due to the limited availability of influencing parameters.

Recently, artificial intelligence (AI) technologies have advanced rapidly, expanding their applications across various fields. By utilizing deep learning, it becomes possible to incorporate a wide range of raw water qualities and operational parameters from drinking water treatment plants, thereby overcoming the limitations of conventional methods. Such complex problems of controlling residual chlorine and THMs formation at the same time may be resolved through the use of AI (Hogue *et al*. 2023). Furthermore, by collecting large volumes of data from numerous sensors and analyzing them with AI, it is possible to significantly improve THMs control technologies across different chlorination stages such as pre-, mid-, and post-chlorination.

Research to develop technologies for optimizing water treatment processes using AI has become more active (Sikder *et al*. 2023, Hogue *et al*. 2023, Singh and Gupta 2012, Lin 2020, Mahato 2022). Especially, extensive research has been conducted to reduce chemical costs by minimizing disinfectant usage and to prevent human error using automated water quality analysis in the disinfection process (AWWA 2006, Rovira *et al*. 2020, Koo *et al*. 2024). Various AI models have been developed to predict DBP levels. Singh and Gupta (2012) predicted THMs formation in chlorinated water using three models: artificial neural networks (ANN), support vector machines (SVM), and gene expression programming (GEP). They compared the performance of each model using a dataset, which included pH, temperature, contact time, and bromide concentration in chlorinated water as input variables. Each model successfully represented the nonlinear relationships between disinfection conditions and THMs formation. Hong *et al.* (2020) used a GRA-RBF(Radial Basis Function)- ANN model, which combines grey relational analysis (GRA) with RBF-ANN, to predict THMs formation in tap water. Input variables for the study included DOC, NO_2^--N, NH_4^+-N, NO_3^--N, residual chlorine, pH, and temperature. The GRA-RBF-ANN model outperformed the RBF-ANN model by identifying an optimal model with fewer variables, thereby saving time and cost in implementation. Mahato and Gupta (2022) collected data from five drinking water treatment plants across India over two seasons and compared THMs prediction performance using ANN, SVM, and MLR. Their results showed that ANN had superior performance compared to other AI techniques. The study confirmed that ANN-based THMs prediction is highly useful for managing and controlling DBP levels in drinking water systems.

Therefore, this study aims to develop a real-time THMs prediction model based on deep learning that utilizes sensor-acquirable data reflecting raw water qualities and operational parameters of the treatment processes. Especially, this study includes to perform a real-time prediction of THMs concentration using sensor-acquired data from actual water treatment plants, thereby enabling instant control of chlorine dosage, unlike previous studies that have focused on improving the accuracy of THMs prediction using developed AI models only.

2. Numerical simulation

2.1 Target drinking water treatment plant

A drinking water treatment plant in South Korea that supplies drinking water to the public after treating an average of 200,000 m^3/day is selected as a model plant. The treatment process consists of a receiving well, coagulation and flocculation, sand filtration, ozone and granular activated carbon treatment, and a clear well as shown at Fig. 1. Chlorine could be injected at the receiving well, between the sedimentation basin and filtration unit, and at the clear well. In the plant, water quality and quantity are managed in real time using an automatic water quality monitoring system (Nanochem, Seoul, Korea). The automatically measured data and average values for each treatment process are summarized in Table 1. Water quality data were collected from January 1, 2020 to May 31, 2023. THMs were analyzed weekly using a gas chromatography system (Agilent, CA, USA).

2.2 Partitioning the dataset for training and validation

From the dataset, 137 data points (80%) were randomly selected and used as the training dataset. The remaining 20% (34 data points) was used as the validation dataset. For

Table 1 Mean and standard deviation of train dataset

Process	Data	Unit	Mean	Std. Div.
Raw water	Water intake	1,000m^3	9331.73	689.48
	Turbidity	NTU	6.67	7.82
	pH	-	7.89	0.32
	Alkalinity	ppm as $CaCO_3$	51.70	8.24
	Ammonia	ppm	0.05	0.04
	TOC	ppm	2.17	0.36
Receiving well	Turbidity	NTU	6.59	7.25
	pH	-	7.61	0.18
	Electrical conductivity	μS/cm	466.88	302.07
	Temperature	℃	13.54	7.68
Mixing basin	pH	-	7.42	0.11
	Alkalinity	ppm as $CaCO_3$	46.21	8.97
	Residual chlorine	ppm	0.21	0.09
Sedimentation basin	Turbidity	NTU	0.62	0.80
	pH	-	7.30	0.13
	Alkalinity	ppm as $CaCO_3$	42.41	7.63
Filter basin	Turbidity	NTU	0.05	0.01
Clear well	Turbidity	NTU	0.05	0.00
	pH	-	7.16	0.12
	Temperature	℃	13.74	7.84
Activated carbon sedimentation basin	Influent Turbidity	NTU	0.06	0.01
	TOC	ppm	0.81	0.25
	Effluent Turbidity	NTU	0.06	0.01
Clear well	chlorine conc. after injection	ppm	1.23	0.46

the training dataset, normalization was applied as part of the AI model development process. A summary of the statistical characteristics of the training data is presented in Table 1.

2.3 Evaluation of model performance

The performance of the model can be evaluated and compared using correlation coefficients and error metrics. In this study, mean absolute error (MAE, Eq. (1)) and mean squared error (MSE, Eq. (2)) were used to compare the AI model's performance on the training and validation datasets. For the final performance evaluation, the correlation coefficient R^2 was used. The R^2 value is calculated using the sum of squares total (SST), sum of squares error (SSE), and sum of squares regression (SSR) as shown at Eq. (3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \quad (2)$$

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - y^m)^2} \quad (3)$$

where, y_i is actual value. y_i' is predicted value. y^m is mean of actual value. n is the number of data points.

3. Results and discussion

3.1 THMs prediction model development

Neurons in the hidden layers sum the weighted input values including bias, and this weighted sum passes through an activation function (Eq. (4)), which enables the network to shift within the space of data non-linearity.

$$y = \sum_{i=1}^{n} f(w_i x_i + b) \quad (4)$$

where, y is output value. w_i is weighted vector. x_i is scaled input vector. b is bias. To prevent underfitting and overfitting, the model's architecture was carefully tuned, including the number of layers, neurons, training epochs, the choice of the activation function for model training (Deng 2023, Alvarez and Salzmann 2016, Asthana *et al.* 2017). The model was configured with 2 hidden layers and 64 neurons per layer. To evaluate error changes over training epochs, the number of epochs was set to 1,000. The model training was conducted using rectified linear unit (ReLU), hyperbolic tangent (tanh) and exponential linear unit (ELU) activation function.

Initially, the AI model was implemented and evaluated using the ReLU activation function. The MAE and MSE of the training and validation datasets for the model, according to the number of training epochs, are shown in Fig. 2. Both training and validation errors decreased sharply between 0 and approximately 200 epochs; however, no further performance improvement was observed beyond 200 epochs. For the training error, it was observed that the model converged toward a THMs value of zero as the number of epochs increased. In contrast, the validation error converged around 150 epochs, with the MAE stabilizing at approximately 7 of THMs and the MSE at approximately 75.

After around 100 epochs, the validation error plateaued while the training error continued to decrease, indicating potential overfitting. A large discrepancy between training and validation errors indicates that the model may be overfitting to the training data, which can significantly degrade prediction performance on validation. Additionally, the number of training epochs affects the total model training time. Excessive epochs may lead to inefficiency in model development. Therefore, a feature to automatically limit the number of epochs was implemented in the model to address issues arising from insufficient data. When more data becomes available in the future, this restriction can be lifted to retrain the model at an optimal number of epochs.

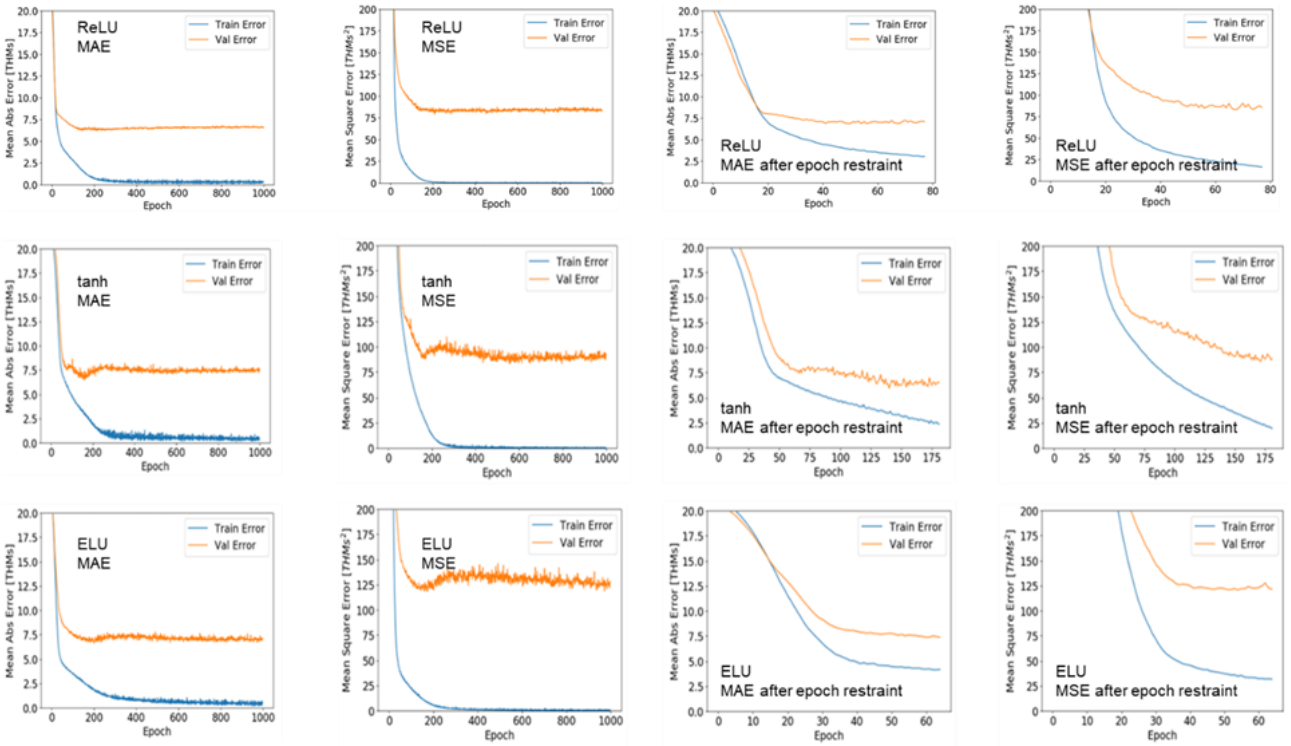

Fig. 2 Training(Train) and validation(Val) errors of ReLU, tanh, and ELU

When the activation function was changed to the tanh and ELU functions, the MAE and MSE of the resulting models did not show clear convergence to a specific value, even with continued training epochs, unlike the model using the ReLU function. This lack of convergence was more pronounced in the MSE graphs, where differences of up to 25 were observed even after 200 epochs. For the training error, both models showed values close to zero after 200 epochs. In terms of validation error, both models exhibited approximate MAE values around 7.5, which did not differ significantly from the ReLU-based model. However, in terms of MSE for the validation error, the model using the tanh function showed an approximate value of 100, while the ELU-based model showed about 125. These results indicate significantly higher error values of tanh and ELU based model compared to the ReLU function model.

Fig. 3 shows comparison of true values with predictions using R^2 for ReLU, tanh, and ELU in the 5 independent simulations. The THMs prediction performance of the ReLU model ranged from R^2=0.73 to R^2=0.79, which is somewhat lower compared to previous studies, where artificial neural network models achieved R^2 values close to 0.90 (Singh and Gupta 2012, Hong *et al*. 2020, Mahato and Gupta, 2022). This difference stems from the inability to classify organic matter in detail, time lag between sensor and THM lab analysis, and the lack of bromide (Br^-) concentration. Since some of these parameters cannot be obtained by sensors at drinking water treatment plants, a slight decrease in prediction accuracy is inevitable.

For the tanh-based model, the prediction performance ranged from R^2 =0.52 to R^2=0.60, and an average value of R^2=0.55, indicating a significant decrease in performance compared to the ReLU model. This low prediction performance is likely due to the vanishing gradient problem associated with the tanh activation function (Wang *et al*. 2019). The ELU based model showed a prediction performance ranging from R^2=0.67 to R^2=0.73, with an average of R^2=0.70, which was better than the tanh-based model but still lower than the ReLU based model. Although the ELU function was designed to overcome the limitations of the ReLU function, the slightly lower performance is presumed to result from the increased model complexity caused by considering negative-valued outputs, even though the dataset used in this study had a limited number of samples and did not include negative values (Eberlein *et al*. 2025). Overall, the models using tanh and ELU based model showed larger training and validation errors in comparison to the ReLU based model.

3.2 THMs forecasting model optimized through variable restraining

In the development of predictive models using AI, the number of input variables can significantly affect the model's performance. A reduction in the number of input variables is likely to facilitate the prediction of THMs. A variable restraining THMs prediction model was developed using the ReLU activation function and its performance was compared to the 24 variables model in section 3.1. As shown in Table 2, the 14 variables used in the model are carefully selected, which include flow and water qualities closely related to THMs formation at each treatment stage (Singer and Reckhow 1999, USEPA 2006).

As shown at Fig. 4, the MAE and MSE of a variable restraining THMs prediction model were not significantly different from those of the ReLU model in section 3.1.

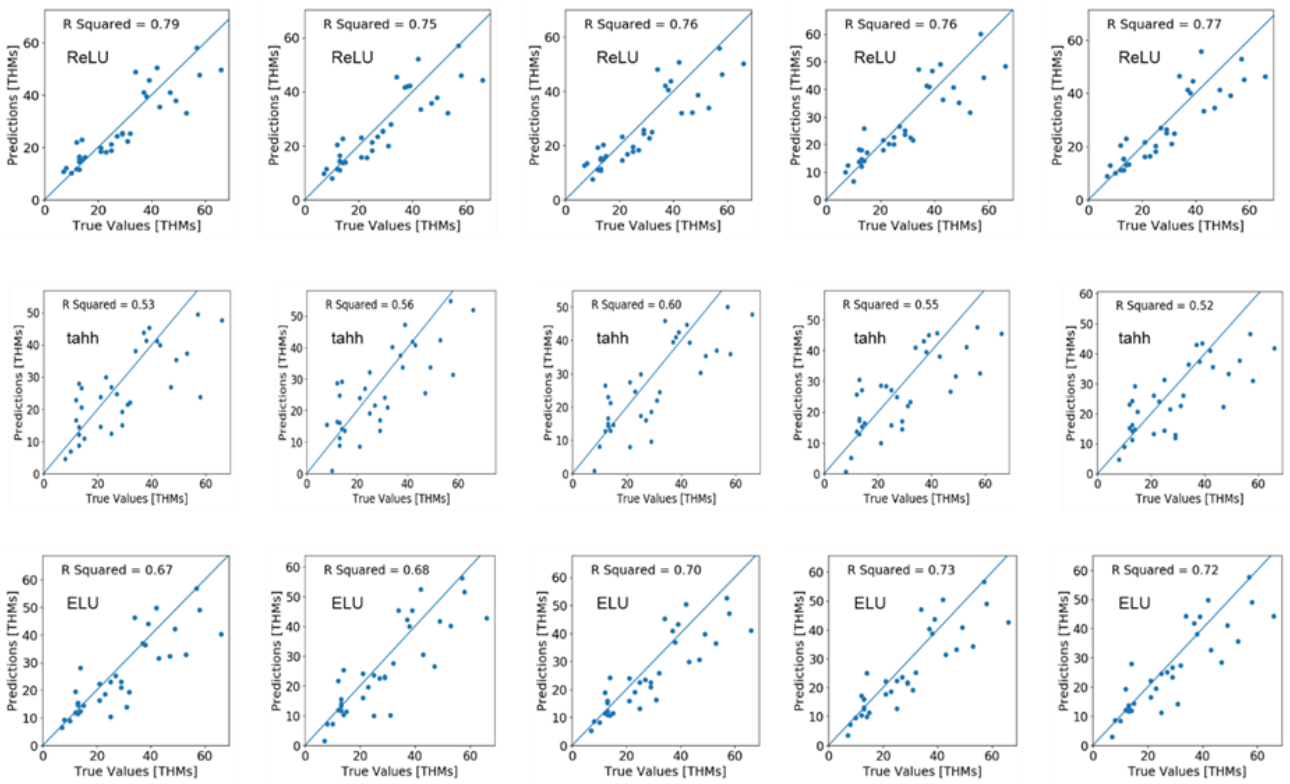

Fig. 3 Comparing true values with predictions for ReLU, tanh, and ELU functions

Table 2 Train dataset statistics of a variable-adjusted THMs prediction model (n=137)

Process	Data	Unit	Mean	Std. Div.
Raw water	Water intake	1,000m^3	9331.73	689.47
	Turbidity	NTU	6.67	7.82
	pH	-	7.89	0.32
	Alkalinity	ppm as $CaCO_3$	51.70	8.24
	Ammonia	ppm	0.05	0.04
	TOC	ppm	2.17	0.36
Receiving well	Turbidity	NTU	6.59	7.25
	pH	-	7.61	0.18
	Electrical conductivity	μS/cm	466.88	302.07
	Temperature	℃	13.54	7.68
Clear well	Turbidity	NTU	0.05	0.00
	pH	-	7.16	0.12
	Temperature	℃	13.74	7.84
	Residual chlorine	ppm	1.23	0.46

After 200 training epochs, the MAE validation error initially reached approximately 5 but then slightly increased and failed to stabilize. Similarly, the MSE reached around 40 and subsequently increased, indicating that the error did not converge in a stable manner.

The method of limiting the model's automatic number of epochs involves constraining the number of training cycles to the point where the error reaches its minimum, which was determined to be 70 epochs. After applying this constraint, the MAE and MSE graphs showed the smallest difference between training and validation errors compared to all previously developed ReLU models. The training and validation MAEs converged to approximately 2.5 and 5, respectively, resulting in a difference of about 2.5. This is half the MAE gap observed in the ReLU based model, where the training and validation errors differed by around 5. Similarly, the MSE values at the end of training were approximately 25 and 50 for the training and validation data sets, respectively, again showing a twofold reduction in the error gap. This improvement is presumed to be the result from model simplification achieved through parameter adjustment. To evaluate the impact of this simplification on predictive performance, the same model was simulated five times independantly as shown at Fig. 5.

The predictive performance of the variable restraining THMs prediction model showed R^2 values ranging from R^2=0.71 to R^2=0.73, with an average of R^2=0.72, indicating minimal variation across repeated simulations. Although model simplification resulted in a slight decrease in overall prediction accuracy, it appears to have improved the stability of the model's performance in response to changes in the training data. For a predictive model to be applied in actual drinking water treatment plants, maintaining stable performance despite data fluctuations is essential. Therefore, variable restraining is expected to be an important consideration when implementing THMs prediction models in real world treatment facilities.

3.3 THMs prediction model with adjusted output range

In predictive model development using artificial

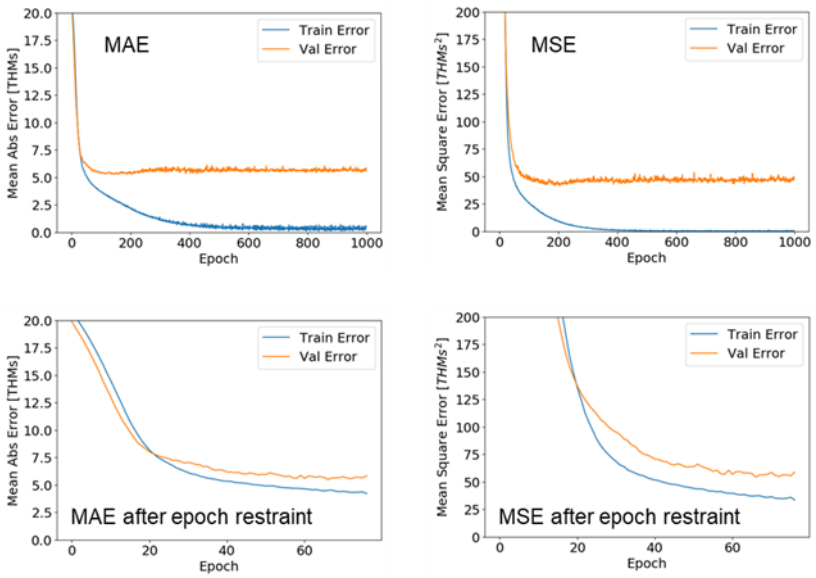

Fig. 4 Training(Train) and validation(Val) error of the variable restraining model

Table 3 Train dataset statistics of result value restraining model

Process	Water intake	1,000m^3	9373.91	747150
Raw water	Turbidity	NTU	6.27	5.81
	pH	-	7.90	0.31
	Alkalinity	ppm as $CaCO_3$	51.31	8.43
	Ammonia	ppm	0.06	0.04
	TOC	ppm	2.17	0.36
	Turbidity	NTU	6.01	4.04
Receiving well	pH	-	7.61	0.17
	Electrical conductivity	μS/CM	442.61	296.82
	Temperature	℃	13.00	7.40
	pH	-	7.43	0.10
Mixing basin	Alkalinity	ppm as $CaCO_3$	45.75	9.08
	Residual chlorine	ppm	0.22	0.09
	Turbidity	NTU	0.53	0.65
Sedimentation basin	pH	-	7.31	0.13
	Alkalinity	ppm as $CaCO_3$	41.82	7.82
	Turbidity	NTU	0.05	0.01
Filter basin	Turbidity	NTU	0.05	0.00
Clear well	pH	-	7.17	0.11
	Temperature	℃	13.23	7.51
	Ozone Eff. Turbidity	NTU	0.06	0.01
Activated carbon sedimentation basin	TOC	ppm	0.81	0.24
	Turbidity	NTU	0.06	0.01
	Chlorine conc. after injection	ppm	1.18	0.45
Clear well	Water intake	1,000m^3	9373.91	747150

intelligence, the range of target output values can influence the improved accuracy of model. Although this approach is not suitable for identifying instances where THMs exceed regulatory limits, this study was conducted to investigate whether such adjustment could have an impact on prediction accuracy. Note that in both the ReLU based and variable adjusted THMs prediction models (Fig. 3 and Fig. 5), significant discrepancies between actual and predicted values were observed when THMs concentrations exceeded 40 µg/L. Therefore, THMs prediction model with a adjusted output range was compared with the models in section 3.1. To investigate how adjusting output range affects predictive performance, data points with THMs concentrations over 40 µg/L were excluded. As a result, the total number of data samples was reduced by 21, resulting in an adjusted dataset of 150 samples. The training dataset consisted of a randomly selected 80% (n=120) of the adjusted data, as shown in Table 3.

As shown at Fig. 6, the MAE and MSE of the adjusted output range THMs prediction model did not show significant differences in training error compared to previous models. However, in the validation phase, the error decreased up to approximately 100 epochs, after which it began to increase until around 200 epochs. The validation MAE reached a minimum of approximately 4, then increased and stabilized at around 5. Similarly, the MSE reached a minimum of about 45 epochs before increasing and settling around 50 epochs. This error increase with prolonged training indicates that excessive epochs can not only extend model development time but also negatively affect predictive performance. Therefore, the number of automatic training epochs was limited to 70, just before the point where the validation error began to increase.

To assess prediction outcomes at specific target values, the adjusted output range THMs prediction model was simulated five times using the same architecture. The prediction results of 5 independant simulations are shown in Fig. 7. The predictive performance of the adjusted output range THMs prediction models showed R^2 values ranging from R^2=0.71 to R^2=0.74, with an average R^2 of 0.72, which

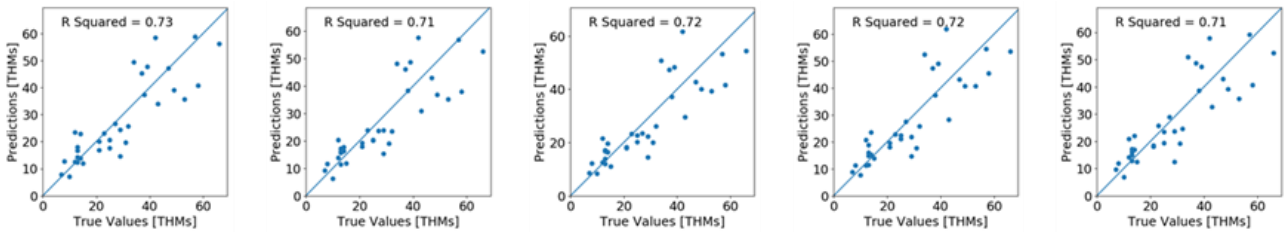

Fig. 5 Comparing true values with predictions for the variable restraining model

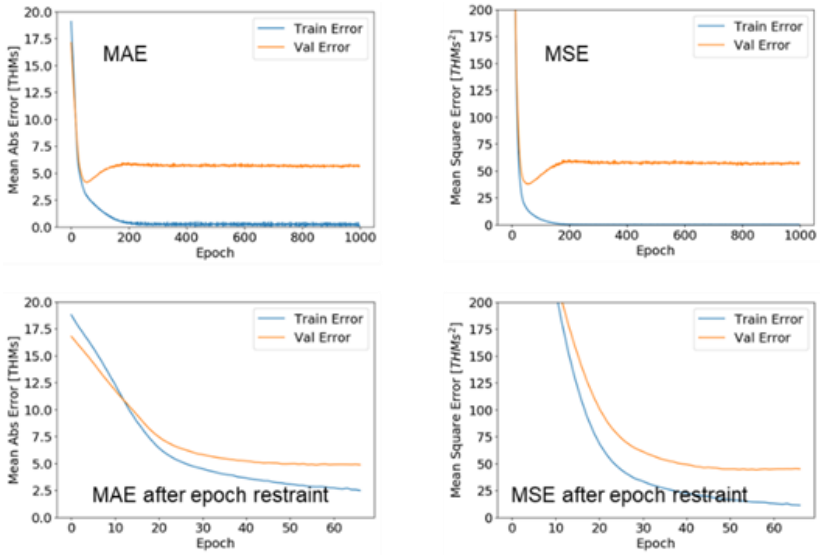

Fig. 6 Training(Train) and validation(Val) error of the adjusted output range model

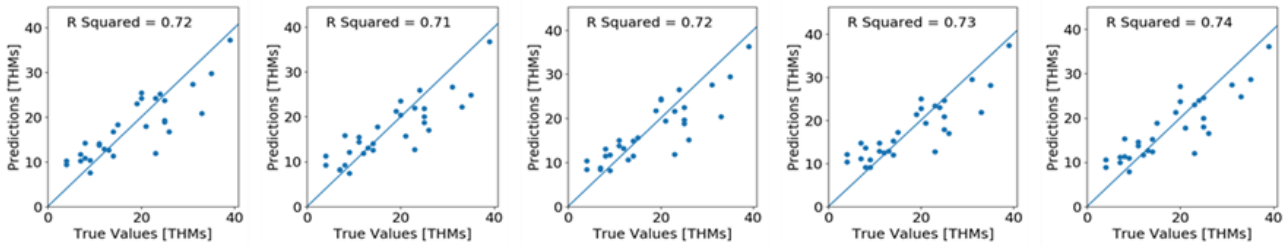

Fig. 7 Comparing true values with predictions for the adjusted output range model

is lower than that of the model in section 3.1. The adjusted output range model did not yield any enhancement in model accuracy. Although data points located far from the median may impair the predictive capability of a model, this effect was not observed in the current study. However, such a result cannot be generalized, as the impact is case-specific. Thus, careful consideration is essential during model development.

4. Conclusions

To prevent the unintentional exceedance of regulatory limits for THMs due to excessive chlorine dosing, a predictive model was developed using real-time sensor data in conjunction with artificial intelligence. Among the tested activation functions, the model using ReLU showed the best performance, achieving a predictive accuracy of $R^2 = 0.76$ without clear tendencies toward over- or under-estimation. In contrast, models based on tanh and ELU activation functions yielded lower performance, with R^2 values of 0.55 and 0.70, respectively.

To explore potential improvements, a model with a reduced number of input variables (14 instead of 24) was developed. Although this led to a slight decrease in accuracy (from $R^2 = 0.76$ to $R^2 = 0.72$), reducing the number of variables can enhance system operational efficiency and is one of critical considerations when deploying THM prediction models in real-world drinking water treatment facilities.

Additionally, attempts to improve performance by adjusting output range did not yield any benefit, with an average R^2 of 0.72.

Despite the slightly lower accuracy compared to previous studies, the proposed approach offers a practical advantage by relying solely on variables that can be measured in real time using sensors at drinking water treatment plants.

References

Alvarez, J.M. and Salzmann M. (2016), "Learning the number of neurons in deep networks", *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016*), Barcelona, Spain, December.

Asthana, S., Pandit, A. and Bhardwaj, A. (2017), "Analysis of multiple hidden layer vs. accuracy in performance using back propagation neural network", *Indian J. Sci. Technol*., **10**(4), 1-4. https://doi.org/10.17485/ijst/2017/v10i4/110899.

AWWA (2006), *Water chlorination/chloramination practices and principles (Manual of Water Supply Practices–M20, 2nd ed.)*, AWWA, Denver, CO, U.S.A.

Bae, Y., Pasteris, J.D., Giammar, D.E. (2020a), "Impact of orthophosphate on lead release from pipe scale in high pH, low alkalinity water", *Water Res*., **177**(15), 115764. https://doi.org/10.1016/j.watres.2020.115764.

Bae, Y., Pasteris, J.D., Giammar, D.E. (2020b), "The ability of phosphate to prevent lead release from pipe scale when switching from free chlorine to monochloramine", *Environ. Sci. Technol*., **54**(2), 879-888. https://doi.org/10.1021/acs.est.9b06019.

Brown, D. (2009), "The management of trihalomethanes in water supply systems", Ph.D. Dissertation, University of Birmingham, Birmingham, U.K.

Deng, Y., Zhou, X., Shen, J., Xiao, G., Hong, H., Lin, H., Wu, F. and Liao, B.Q. (2021), "New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water", *Sci. Total Environ*., **772**, 145534. https://doi.org/10.1016/j.scitotenv.2021.145534.

Deng, T. (2023), "Effect of the number of hidden layer neurons on the accuracy of the back propagation neural network", *Highl. Sci. Eng. Technol*., **74**, 462-468. https://doi.org/10.54097/nbra6h45.

Eberlein, J., Rodriguez, D. and Harrison. R. (2025), "The effect of data complexity on classifier performance", *Empire Softw. Eng*., **30**(16), 1-23. https://doi.org/10.1007/s10664-024-10554-5.

Hong, H., Zhang, Z., Guo, A., Shen, L., Sun, H., Liang, Y., Wu, F. and Lin, H. (2020), "Radial basis function artificial neural network(RBF-ANN) as well as the hybrid method of RBF-ANN and grey relational analysis able to well predict trihalomethanes levels in tap water", *J. Hydrol*., **591**, 125574. https://doi.org/10.1016/j.jhydrol.2020.125574.

Hogue, D., Mirchandani, P.B. and Boyer, T.H. (2023), "Predictive capability of THM models for drinking water treatment and distribution", *Environ. Sci. Water Res*., **9**, 2745-2759. https://doi.org/10.1039/D3EW00308F.

Koo, J., Kim, Y. and Lee, D. (2024), "Optimization of chlorine injection schedule in water distribution networks using genetic algorithm", *Water*, **16**(3), 486. https://doi.org/10.3390/w16030486.

Lee, B., Kang C. and Gil, K. (2024), "Effects of upstream pollution patterns on the water quality of Paldang lake", *Membr. Water Treat*., **15**(4), 185-192. https://doi.org/10.12989/mwt.2024.15.4.185.

Lee, D., Kim, Y., Kim, S., Lee, K. and Park, H. (2007), "Effect of advanced treatment process for residual chlorine decay and THM formation in water distribution system", *J. Korean Soc. Environ. Eng*., **29**(4), 419-424

Lin, H, Dai, Q., Zheng, L., Hong, H., Deng, W. and Wu, F. (2020), "Radial basis function artificial neural network able to accurately predict disinfection by-product levels in tap water: Taking haloacetic acids as a case study", *Chemosphere*, **248**, 125999. https://doi.org/10.1016/j.chemosphere.2020.125999.

Mahato, J.K. and Gupta, S.K. (2022), "Exploring applicability of artificial intelligence and multivariate linear regression model for prediction of trihalomethanes in drinking water", *Int. J. Environ. Sci. Technol*., **19**, 5275-5288. https://doi.org/10.1007/s13762-021-03392-1.

McGuire, M.J. (2013), *The Chlorine Revolution: Water Disinfection and The Fight To Save Lives*, AWWA, Denver, CO, U.S.A.

Oh, H., Jang, P., Hyung, J., Koo, J. and Maeng, S. (2023a), "Experimental and model study on the mixing effect of injection method in UV/H_2O_2 process", *Membr. Water Treat*., **14**(3), 129-140. https://doi.org/10.12989/mwt.2023.14.3.129

Oh, S., Wang, J., Kim J. and Cho, Y. (2023b), The effects of algal-derived organic matters (AOMs) and chlorinated AOMs on the survival and behavior of zebrafish, *Membr. Water Treat*., **14**(3), 141-146. https://doi.org/10.12989/mwt.2023.14.3.141

Osorio, F., Ribes, D., González-Martínez, A., Poyatos, J.M. and García, P. (2011), "A model for predicting THM presence in networks of water supply systems", *WIT Trans. Built Environ*., **117**, 233-240. https://doi.org/10.2495/SAFE110211.

Rovira, M., Menéndez, J. and Montserrat, S. (2020), "Automated chlorine dosage in a simulated drinking water treatment plant: A real case study", *Appl. Sci*., **10**(11), 4035. https://doi.org/10.3390/app10114035.

Richardson, S.D. (2003), "Disinfection by-products and other emerging contaminants in drinking water", *TRAC*, **22**(10), 666-684. https://doi.org/10.1016/S0165-9936(03)01003-3.

Singh, P. and Gupta, A.B. (2012), "Application of artificial neural network, support vector machine, and gene expression programming for the prediction of THMs in drinking water", *J. Water Health*, **10**(1), 43-56. https://doi.org/10.3390/app12041771.

USEPA (2006), *Stage 2 disinfectants and disinfection byproducts rule (DBPR)*, Washington DC, USA

Sikder, R., Zhang, T. and Ye, T. (2023), "Predicting THM formation and revealing its contributors in drinking water treatment using machine learning", *ACS ES&T Water*, **3**(5), 1234-1245. https://doi.org/10.1021/acsestwater.3c00020.

Singer, P. and Reckhow, D. (2011), *Water Quality and Treatment 6th ed*, AWWA, Denver, CO, U.S.A.

Singer, P.C. (1999), *Formation and Control of Disinfection by-Products in Drinking Water*, AWWA, Denver, CO, U.S.A.

Singh, K.P. and Gupta, S. (2012), "Artificial intelligence based modeling for predicting the disinfection by-products in water", *Chemom. Intell. Lab. Syst*., **114**, 122-131. https://doi.org/10.1016/j.chemolab.2012.03.014.

Safeer, S., Pandey, R.P., Rehman, B., Safdar, T., Ahmad, I., Hasan, S.W. and Ullah, A. (2022), "A review of artificial intelligence in water purification and wastewater treatment: Recent advancements", *J. Water Process Eng*., **49**, 102974. https://doi.org/10.1016/j.jwpe.2022.102974.

Tsitsifli, S. and Kanakoudis, V. (2020), "Developing THMs' predictive models in two water supply systems in Greece", *Water*, **12**(5), 1422. https://doi.org/10.3390/w12051422.

Wang, X., Qin, Y., Wang, Y., Xiang, S. and Chen, H. (2019), "ReLTanh: An activation function(s) with vanishing gradient resistance for SAE-based DNNs and its application to rotating machinery fault diagnosis", *Neurocomputing*, **363**, 88-98. https://doi.org/10.3390/w12051422.

Youn, H., Kang, C. and Gil, K. (2024), "Water quality management strategy based on organic matter characteristics of streams and lakes in the Namhan river watershed", *Membr. Water Treat*., **15**(3), 99-106. https://doi.org/10.12989/mwt.2024.15.3.099

JL