https://doi.org/10.71573/g8qmc295

Modeling residual chlorine and disinfection by-products (DBPs) dynamics in urban sewers during COVID-19 disinfection practices: A comparative analysis of process-based and data-driven approaches

Xuhao Wang [1], *Chunyan Wang* [1], *Yi Liu* [1,]

[1]*Tsinghua University, School of Environment, Beijing, China*
**Corresponding author email: yi.liu@tsinghua.edu.cn*

Abstract

The COVID-19 pandemic has intensified chlorine-based disinfection, elevating risks from residual chlorine and disinfection by-products (DBPs) in sewers. Using a pilot-scale sewer system with MS2 bacteriophage (SARS-CoV-2 surrogate), we stimulated wastewater collection and transportation process, and compared process-based (reaction kinetics) and data-driven models (random forest, decision tree, deep learning, general additive model, stacked model) under static (collection) and dynamic (transport) scenarios. The experimental results showed that the changes in residual chlorine and DBPs concentration in the dynamic scenario were more complex than in the static scenario, and higher residual chlorine dose didn't accelerate the inactivation of the virus. According to analyses, data-driven models showed superior accuracy for residual chlorine prediction (R^2 +0.03) but poorer robustness for DBPs (MAE +0.35 vs. process-based), while process-based models exhibited smaller RMSE increases (2.91 vs. 5.31) when predicting DBPs versus chlorine, reflecting their adaptability to complex chlorine-organic matter interactions driving DBP formation. Uncertainty analysis revealed data-driven models' sensitivity to high initial residual chlorine and DBPs doses. As the global situation is still rapidly evolving with a more frequent outbreak of epidemic events, our study provides a tool for estimating hazardous substances production caused by sterilization behavior for pandemic prevention.

Highlights

- A pilot-scale physical sewer system is used to stimulate residual chlorine and DBPs dynamics.
- Data-driven models in principle outperform process-based models in DBPs prediction.
- Higher initial dose levels may lead to more prediction uncertainties for data-driven models.

Introduction

The COVID-19 pandemic has significantly increased the use of chlorine-based disinfectants, leading to concerns about the accumulation of disinfection by-products (DBPs) in urban wastewater and their potential ecological impacts (Zhang et al., 2022). Environmental risks via wastewater treatment are evident, as chlorine residuals have been reported to exceed regulatory limits in some areas during the pandemic (Chu et al., 2021). However, current models for predicting these risk substances in wastewater have some limitations. Unlike water supply systems, wastewater scenarios for quantitative analysis are hard to generalize due to a broader range of influencing factors. The complex composition of pollutants in wastewater makes traditional process-based models inadequate for accurately predicting multiple reactions among substances. Hence, researchers in recent years have focused on data-driven models to improve their usefulness through trade-offs between accuracy and interpretability. This study aims to compare process-based and data-driven models (Onyutha and Kwio-Tamale, 2022) in predicting the concentration changes of residual chlorine and DBPs in wastewater treatment processes. Using a pilot-scale sewer system, we evaluated these models' performance in simulating chlorine decay and DBP formation under static and dynamic scenarios. The study highlights the strengths and limitations of each modeling approach, providing insights into managing COVID-19-related hazardous substances in urban wastewater.

Methodology

Experimental design

A pilot sewer system composed of a set of water tanks, a 1200-meter sewage pipe network, and a reflux device to perform downscale simulations. In this study, a static scenario was designed to stimulate the process of sewage from different sources first entering collection facilities such as hospital pre-disinfection tanks and septic tanks. In this process, risk substances will stay in a relatively fixed space. Accordingly, a dynamic scenario was set up to stimulate the process of sewage transmission in the pipe network, featuring their spatial movement at a relatively stable flow rate. Both scenarios used real wastewater and sodium hypochlorite as the materials. Due to safety reasons, the MS2 bacteriophage was used as the surrogate for the COVID-19 virus.

Sample collection

In both scenarios, the residual chlorine and pH were monitored by four real-time residual chlorine monitors, and the flow velocity was monitored by two ultrasonic flowmeters. The water samples were collected from 4 sampling ports at 360, 480, 600, and 720 m (A, B, C, D) along the pipe to measure DBPs (trichloromethane, dichloromethane), MS2 bacteriophage titer, and TOC.
The collection schemes in the two scenarios are different due to disparate sewage processes. In the static scenario, the sampling interval of the static scenario is fixed to every 2 hours, while the sampling interval of the dynamic scenario was adjusted according to the change of the flow rate based on the Lagrange particle tracking method.

Data processing

In total, 1440 residual chlorine and pH data as well as 36 DBPs, MS2 bacteriophage, and TOC data were collected. The sliding window method and rolling window method were jointly used for data processing to fulfill the requirements of data volume during the formulation of data-driven models. We compared the data measured at a certain time node with all the data measured after that time node to obtain the time series of changing course. The amount of processed data generated from the original samples' data can be expressed by the following formula:

$$PD_{i,n} = \frac{n(n-1)}{2}$$

Where *PD* is the amount of processed data, *i* is the variables we observed, and *n* is the amount of original observation. In this study, a minimum volume of 48 observations (DBPs, TOC, titer of MS2 bacteriophage data in static scenario) were processed into 1,128 data for model training by using our data processing methods to meet the requirements of data volume for the formulation of most data-driven models (Helm et al., 2024). In dynamic scenario, 96 observations of DBPs, TOC, and titer of MS2 bacteriophage data were processed into 4,560 data. The data from sampling port A will be used as the validation set, while the data collected from the other three ports will be used as the training set.

Model formulation and evaluation

Two process-based models and five data-driven models were selected to formulate prediction models of different hazardous substances in different scenarios. For the process-based model, first-order reaction kinetics were used for the residual chlorine decay prediction, and second-order reaction kinetics were used for DBPs formation prediction model (Clark et al., 2001). For data-driven models, the decision tree, random forest, deep learning, and general additive model were utilized. Additionally, a stacked model using these four models as base learners and a general linear model as the meta-learner were selected.

The model evaluation was divided into 3 steps. The first step was comparing R^2, RMSE, and ranking MAE across different models, second step was selecting the two most important variables of each model. The last step was analyzing prediction uncertainties under the impact of the two most important variables in each model. Due to the variety of model structures, multiple methods were utilized to select important variables in the second step. For two process-based models, the two important variables were selected by picking the variables with higher p-values. For decision tree and random forest models, the SHAP values of each variable were calculated to pick out the two most important variables. For the other three models, the variable relative importance, scaled importance, and importance percentage were used to select the two most important variables.

Results and discussion

Different changing characteristics of risk substances were manifested between the two scenarios. The concentration of residual chlorine changed the most rapidly within 0 - 4 hours in the static scenario and 4-6 hours in the dynamic scenario. Compared with the existing scientific understanding of chlorine decay dynamics in the water supply process, the concentration change of residual chlorine in the static scenario is more consistent with the first-order reaction kinetics, while the results in the dynamic scenario are more fitting with other mechanisms such as multi-phase reaction models (Jedas-Hecart et al., 1992; USEPA, 1992). The concentration of DBPs ultimately dropped to 1.05 µg/L in the static scenario, while it was 4.38 µg/L at the end of the dynamic scenario. By observing the inflection points of residual chlorine curves and DBPs curves, the number of inflection points of the DBPs curve under a dynamic scenario is the largest, and the inflection points of the DBPs curve under the same scenario are more than the residual chlorine curve. The titer of MS2 bacteriophage decreased to 0 after 6 hours in the static scenario and 2 hours in the dynamic scenario. The initial titer in the two scenarios was of the same order of magnitude, although the initial residual chlorine concentration was higher in the dynamic scenario, the MS2 bacteriophage was deactivated faster under experimental conditions.

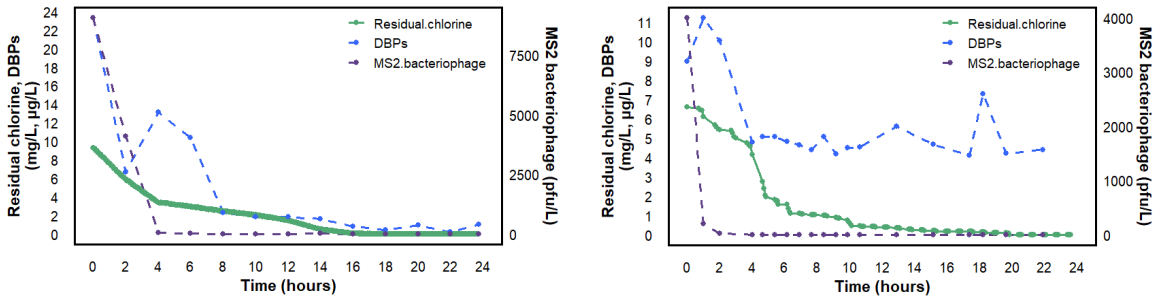

Figure 1. Changes in residual chlorine and DBP concentration and titer of MS2 bacteriophage over time. (Left: static scenario; right: dynamic scenario)

For the static scenario, the deep learning model performed the highest R^2 and lowest MAE in the prediction of residual chlorine, and the random forest model achieved the highest R^2, the lowest RMSE, and the lowest MAE. In the dynamic scenario, whether for the prediction of residual chlorine or DBPs, the stacked model achieves the best performance in both R^2 and RMSE (Table 1). Nevertheless, when confronted with different scenarios, the predictive performance of certain data-driven models varies considerably. For instance, the RMSE for GAM and deep learning models in predicting DBPs increased by 34.12 and 13.77 from dynamic to static scenario, and MAE also increased by 13.96 and 7.48 respectively, higher than those in dynamic scenarios, and 13.77 and 7.48 for deep learning models. For the process-based model, the change of RMSE and MAE were 2.71 and 1.82, respectively. The key variables selection results show that the initial dose of chlorine/DBPs and reaction time were evaluated as the two most important variables for each model selected in this study.

Table 1. The performance and important variables of each selected model for predicting different hazards under different scenarios.

Model	Scenario	Hazardous Substance	Model Performance			Important variables	
			R^2	RMSE	MAE	Most important	Second important
First-order reaction kinetic	Static	Residual chlorine	0.8	0.29	0.18	Initial dose	Reaction time
		DBPs(THMs)	NA	NA	NA	NA	NA
	Dynamic	Residual chlorine	0.84	0.3	0.14	Initial dose	Reaction time
		DBPs(THMs)	NA	NA	NA	NA	NA
Second-order reaction kinetic	Static	Residual chlorine	NA	NA	NA	NA	NA
		DBPs(THMs)	-16.41	4.56	3.29	Initial dose	Reaction time
	Dynamic	Residual chlorine	NA	NA	NA	NA	NA
		DBPs(THMs)	-8.87	1.85	1.47	Initial dose	Reaction time
Decision tree	Static	Residual chlorine	0.77	0.51	0.23	Initial dose	Reaction time
		DBPs(THMs)	-1.48	1.98	1.1	Initial dose	Reaction time
	Dynamic	Residual chlorine	0.98	0.11	0.03	Initial dose	Reaction time
		DBPs(THMs)	0.49	0.42	0.26	Initial dose	Reaction time
Random forest	Static	Residual chlorine	0.88	0.37	0.18	Initial dose	Reaction time
		DBPs(THMs)	0.12	1.6	1	Initial dose	Reaction time
	Dynamic	Residual chlorine	0.99	0.08	0.03	Initial dose	Reaction time
		DBPs(THMs)	0.71	0.33	0.25	Initial dose	Reaction time
Deep learning	Static	Residual chlorine	0.88	0.72	0.1	Initial dose	Reaction time
		DBPs(THMs)	-297.8	14.2	7.8	Initial dose	Reaction time
	Dynamic	Residual chlorine	0.98	0.11	0.05	Initial dose	Reaction time
		DBPs(THMs)	0.47	0.43	0.32	Initial dose	Reaction time

General additive model (GAM)	Static	Residual chlorine	0.55	0.72	0.62	Initial dose	Reaction time
		DBPs(THMs)	-1919.85	34.68	14.38	Initial dose	Reaction time
	Dynamic	Residual chlorine	0.58	0.5	0.27	Initial dose	Reaction time
		DBPs(THMs)	0.18	0.56	0.42	Initial dose	Reaction time
Stacked model	Static	Residual chlorine	0.86	0.38	0.29	Initial dose	Reaction time
		DBPs(THMs)	-3.23	2.27	1.6	Initial dose	Reaction time
	Dynamic	Residual chlorine	0.99	0.07	0.26	Initial dose	Reaction time
		DBPs(THMs)	0.89	0.23	0.17	Initial dose	Reaction time

Our uncertainty analysis showed that when the inputs of data-driven models are relatively high initial concentrations and relatively short reaction times, the prediction always had a high relative error compared with actual observations. When the initial concentration of the risk substance decreases, the relative error of the prediction also decreases (Figure 2).

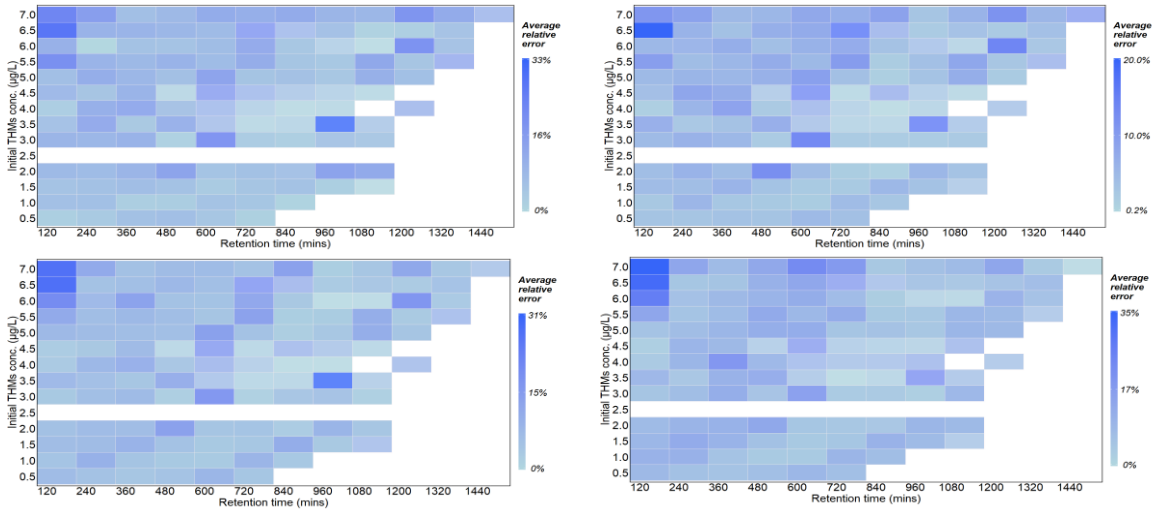

Figure 2. Effect of initial dose and reaction time on error in concentration prediction of THMs by four data-driven models in dynamic scenarios. (Top left: Decision tree; top right: Random forest; bottom left: Deep learning; bottom right: GAM) effect of initial dose and reaction time on the error when four data-driven models predict the concentration of different risk substances in different scenarios.

Conclusions and future work

In conclusion, data-driven models can achieve higher R^2, yet generate greater RMSE and MAE than process-based models when data volume decreases. This indicates that data-driven models demonstrate a greater upper limit of accuracy in predicting the concentration change of residual chlorine and DBPs in diverse sewage environments, but may be lacking in robustness compared with the process-based models. Our study also found that the prediction error does not show significant variation characteristics from the perspective of vital variables, which may indicate that the uncertainty of models is affected by other fundamental factors such as the model's own structure or parameter settings. Additionally, through a public health and environmental management lens, the behavior of residual chlorine and DBPs in sewage systems exhibits greater intricacy during collection and transportation compared to water supply networks. This phenomenon is driven by both the variable wastewater matrix and COVID-19 pandemic-induced hyperchlorination. Crucially, our findings challenge conventional disinfection paradigms: increased chlorine dosing (tested via MS2 phage surrogates) showed diminishing returns in viral inactivation efficiency, urging a re-evaluation of disinfection protocols in crisis scenarios.

Our studies have some limitations. Future research should focus on refining predictive models by leveraging advanced technologies to investigate the behavior of hazardous substances in complex wastewater systems. These efforts will provide a reliable scientific foundation for evidence-based policy-making, enabling more effective and sustainable management of disinfection practices and their environmental impacts.

References

1. Zhang, W., Dong, T., Ai, J., Fu, Q., Zhang, N., He, H., ... & Wang, D. (2022). Mechanistic insights into the generation and control of Cl-DBPs during wastewater sludge chlorination disinfection process. Environment International, 167, 107389.
2. Chu W, Fang C, Deng Y, Xu Z (2021). Intensified disinfection amid COVID-19 pandemic poses potential risks to water quality and safety. Environmental Science & Technology, 55(7): 4084–4086 doi:10.1021/acs.est.0c04394
3. Onyutha, C., & Kwio-Tamale, J. C. (2022). Modelling chlorine residuals in drinking water: a review. International journal of Environmental Science and Technology, 19(11), 11613-11630.

4. Helm, W., Zhong, S., Reid, E., Igou, T., & Chen, Y. (2024). Development of gradient boosting-assisted machine learning data-driven model for free chlorine residual prediction. Frontiers of Environmental Science & Engineering, 18(2), 17.
5. Clark, R. M. , Rossman, L. A. , Sivaganesan, M. , & Schenck, K. M. . (2001). Modeling chlorine decay and the formation of disinfection by-products (dbps) in drinking water.
6. Jadas-Hecart, A., El Morer, A., Stitou, M., Bouillot, P., and Legube, B. (1992). —Modelisation de la Demande en Chlore D'une Eau Traitee." Water Reserves, 26(8), 1073.
7. United States Environmental Protection Agency (USEPA). (1992). Water treatment plant simulation program user's manual, Version 1.21. Drinking Water Technology Branch, Drinking Water Standards Division, Office of Ground Water and Drinking Water, Malcolm Pirnie, Inc.