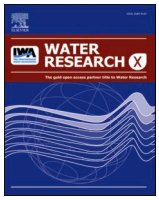
Water Research X 23 (2024) 100230

Contents lists available at ScienceDirect

Water Research X

journal homepage: www.sciencedirect.com/journal/water-research-x



Integrating machine learning algorithm with sewer process model to realize swift prediction and real-time control of H_2S pollution in sewer systems

Zhensheng Liang [a,b], Wenlang Xie [a,b], Hao Li [a,b], Yu Li [c], Feng Jiang [a,b,*]

[a] School of Environmental Science & Engineering, Guangdong Provincial Key Lab of Environmental Pollution Control and Remediation Technology, Sun Yat-sen University, Guangzhou, 510275, China
[b] Guangdong Provincial International Joint Research Center on Urban Water Management and Treatment, Sun Yat-sen University, Guangzhou, 510275, China
[c] School of Environment, Guangdong Provincial Key Laboratory of Chemical Pollution and Environmental Safety & MOE Key Laboratory of Theoretical Chemistry of Environment, South China Normal University, Guangzhou, 510006, China

ARTICLE INFO

Keywords:
Sulfide pollution in sewers
Validated dynamic sewer process model
Machine learning algorithm
Swift prediction model
Real-time control

ABSTRACT

The frequent occurrence of safety incidents in sewer systems due to the emergency toxicity of hydrogen sulfide (H_2S) necessitate timely and efficient prediction, early warning and real-time control. However, various factors influencing H_2S generation and emission leads to a substantial computational burden for the existing dynamic sewer process models and fails to timely control the H_2S exposure risk. The present study proposed a swift prediction model (SPM) that combined the validated dynamic sewer process model (the biofilm-initiated sewer process model, BISM) with a high-speed machine learning algorithm (MLA), achieving accurately and swiftly predict the dissolved sulfide (DS) concentration and H_2S concentration in a specific sewer network. Based on Gradient Boosting Decision Tree-based SPM, the simulated concentrations of DS and H_2S are 1.95 mg S/L and 214 ppm, respectively, which are closely to the field-measured values of 1.82 mg S/L and 219 ppm. Notably, SPM achieved a computation time of less than 0.3 s, and a significant improvement over BISM (> 5000 s) for the same task. Moreover, the real-time and dynamic dosing scheme facilitated by SPM outperformed the conventional constant dosing scheme provided by dynamic sewer process model, which significantly improved the H_2S control completion rate from 69 % to 100 %, and achieved a significant reduction in chemical dosage. In conclusion, the integration of dynamic sewer process models with MLA addresses the inadequacy of monitoring data for MLA training, and thus pursues swift prediction of H_2S generation and emission, and achieving real-time, effective, and economic control of H_2S in complex sewer networks.

1. Introduction

Efficiently controlling hydrogen sulfide (H_2S) pollution in urban sewer systems is a paramount challenge for sewage management authorities. Previous research has primarily focused on sewer corrosion caused by H_2S (Jiang et al. 2015; Liang et al. 2016; Zhang et al. 2008; Zhang et al. 2023b), while less attention given to its emergency toxicity. It is crucial to highlight that an H_2S concentration of 100 ppm has the potential to inflict damage on the human nervous system, while at 700 ppm, it poses a severe risk of causing sudden death (Guidotti 2010). In urban sewer systems, H_2S concentrations demonstrate notable variability, with the potential to rapidly shift from 0 ppm to 1000 ppm within a period as short as 1 min (Juan et al. 2017). These elevated H_2S levels pose a severe hazard to the safety of sewer maintenance personnel. Therefore, it is crucial to implement timely and effective control measures to mitigate this risk.

Currently, the prevailing method for H_2S control involves flushing and chemical dosing with nitrates, alkalis, iron salts, and dissolved oxygen (Liang et al. 2023; Zhang et al. 2008; Zhang et al. 2023b). Chemical dosing is commonly employed due to its rapid effectiveness and strong control capabilities. However, conventional constant dosing or flow-paced dosing methods prove inadequate in addressing the highly dynamic fluctuations of H_2S in sewer systems (Zhang et al. 2023b). This poses a risk of underdosing or overdosing, resulting in poor effective control or chemical wastage (Liang et al. 2019b; Sharma et al. 2008; Zhang et al. 2023b). This highlights the necessity for a control measure capable of timely and dynamically adjusting chemical dosages to effectively mitigate H_2S pollution in sewer systems.

* Corresponding author.
E-mail address: jiangf58@mail.sysu.edu.cn (F. Jiang).

https://doi.org/10.1016/j.wroa.2024.100230
Received 16 April 2024; Received in revised form 4 June 2024; Accepted 16 June 2024
Available online 17 June 2024

To address this issue, researchers have developed dynamic sewer process models aimed at dynamically predicting the generation and emission of H_2S in sewer systems, providing a foundation for optimizing chemical dosing schemes. Until now, three main dynamic sewer process models, namely Wastewater Aerobic/anaerobic Transformation in Sewers, SeweX, and Biofilm-Initiated Sewer Process Model (BISM), have been developed and successfully employed to predict H_2S generation and emission in sewer systems (Huisman and Gujer 2002; Liang et al. 2019b; Sharma et al. 2008). The models serve as valuable tools for sewage management authorities, offering accurate predictions of sulfide production and enabling the optimization of dosing schemes, thereby improving overall sulfide control efficiency in sewer systems. However, in order to enhance the accuracy of the dynamic sewer process model, it is necessary to augment grid density and reduce simulation step length, which further increases the time lags in mathematical models in biology (Frank et al. 2020; MacDonald 2013). This presents challenges in swiftly predicting H_2S emissions or determining real-time appropriate dosages during emergency H_2S toxicity events, thus limiting their utility in sewer management.

Machine learning (ML) has emerged as a powerful tool for addressing nonlinear relationships in complex systems, offering swift data processing and prediction capabilities (Ray 2019). In the field of water environment, machine learning algorithms (MLAs) have already been successfully applied to analogous issues, such as water quality monitoring and prediction (Francesco et al. 2017; Guo et al. 2015), water quality early warning and accident diagnosis (Dovzan et al. 2015; Mika et al. 2013), and intelligent control (Bernardelli et al. 2020; Canete et al. 2021; Félix et al. 2019). In the field of sewer network systems, Zhang et al. (2023a) and Jiang et al. (2021) have conducted comprehensive monitoring efforts, employing ML techniques to forecast the concentrations of fat, oil, and grease in the sewer network, as well as the levels of chemical oxygen demand, total nitrogen, and total phosphorus in the influent of wastewater treatment plants. Their efforts have resulted in outstanding simulation outcomes. As we know, ML is a type of data driven algorithm. In the context of sewer systems, numerous factors influence the generation and emission of H_2S, including flow rate, temperature, organic matter concentration, sulfate concentration, dissolved oxygen levels, etc. (Hvitved-Jacobsen et al. 2013; Liang et al. 2019b). The expensive and challenging maintenance of monitoring equipment (Eerikinen et al. 2020) makes it difficult to collect data on a large scale, over an extended period, and with multiple indicators. Consequently, there is an insufficient quantity of relevant data for ML to predict the generation and emission patterns of H_2S in sewer systems.

Therefore, this study developed a Swift Prediction Model (SPM) that combines dynamic sewer process models BISM and MLA for swift prediction of H_2S pollution in sewer systems, aiming to provide a scientific basis for real-time and efficient control of H_2S by sewage management authorities. Initially, a large amount of simulated data was generated using BISM validated with field tested data. Subsequently, the aforementioned simulated data and MLA were employed to construct and validate the SPM, and field-measured data from large-scale sewer systems were used to ensure its accuracy and swiftness. Finally, the validated SPM was applied to simulate real-time dynamic dosing control of H_2S for various locations and chemicals.

2. Results and discussion

2.1. Performance of simulation on accuracy and efficiency

2.1.1. Simulation of dissolved sulfide (DS) and H_2S

Given that H_2S gas in sewer systems originates from DS in sewage, reducing the concentration of DS can serve as an indirect method to control H_2S pollution. Accordingly, the model was also utilized to simulate the DS concentration. Four extensive MLAs were employed to construct SPM, including Gradient Boosting Decision Tree (GBDT), Multiple Linear Regression (MLR), Random Forest Regression (RFR) and Artificial Neural Network (ANN) (Zhu et al. 2022). The simulation results obtained from the GBDT-based SPM demonstrated the most favorable fitting performance with the results obtained from validated BISM, suggesting that SPM can be used for DS and H_2S simulation of real sewer system (Figs. S1-S2 and Figs. 1 and 2). Specifically, in DS simulation, both the GBDT-based and ANN-based SPMs exhibited the most accurate fitting results in comparison to the validated BISM simulation within the training set. The GBDT-based SPM achieved an R-squared (R^2) value of 0.919 and an Root Mean Square Error (RMSE) value of 0.012 mg S/L, while the ANN-based SPM achieved an R^2 value of 0.900 and an RMSE value of 0.015 mg S/L. Conversely, the RFR-based and MLR-based SPMs exhibited noticeable deviations from the validated BISM simulation results, with R^2 values of 0.429 and 0.548, and RMSE values of 0.086 mg S/L and 0.068 mg S/L, respectively. In the validation set (Fig. 1), the GBDT-based SPM maintained high accuracy in DS simulation, achieving an R^2 value of 0.921 and an RMSE value of 0.014 mg S/L. Similarly, the ANN-based SPM also showed good fitting results, with an R^2 value of 0.914 and an RMSE value of 0.016 mg S/L, while the simulation results based on RFR and MLR remained unsatisfactory, with R^2 values ranging from 0.310 to 0.444. Additionally, the GBDT-SPM exhibited the most favorable fitting performance for H_2S (Fig. 2), with R^2 values surpassing 0.994 and RMSE values below 39 ppm for both the training and validation sets. The subsequent best fitting algorithms were RFR, MLR, and ANN. The results emphasize that compared with the simulation results of validated BISM, the GBDT-based SPM exhibited good fitting performance for both performance for DS and H_2S concentrations in the sewer system.

2.1.2. Computation time

For the sulfide control technology in sewer systems, the computation time of the model is crucial for the real-time development of sulfide control schemes and the efficient implementation of chemical dosages. Dynamic sewer process model such as BISM typically require over an h to complete calculations due to the intricate biological, chemical, and physical processes involved in the production and emission of H_2S in sewer systems. However, the trade-off for enhancing the computational efficiency of dynamic sewer process models is often a sacrifice in accuracy (Frank et al. 2020). In contrast, MLA-based SPMs are driven by data and do not prioritize specific water quality migration and transformation processes. Instead, they emphasize the accuracy of output results, leading to exceptional computational efficiency. For instance, Yan et al. (2018) utilized the support vector machine learning method to optimize the computation time required by the MIKE FLOOD model, reducing it from 25 h to 2.1 milliseconds.

As shown in Table 1, MLA-based SPMs possess a distinct advantage in handling vast amounts of data and swiftly assessing DS and H_2S concentrations within sewer systems. MLA-based SPMs can accomplish predictions within a second or even shorter, resulting in a remarkable improvement in computational efficiency. This capability provides a practical approach for real-time adjustment of dosing levels for H_2S control in sewer systems.

2.2. Model application for sulfide simulation

To validate the predictive capability of the SPM, we compared the simulation results of the GBDT-based SPM with the validated dynamic sewer process model BISM using field-measured data. As illustrated in Fig. 3, both the SPM and validated BISM demonstrated commendable fitting performance to the measured data of DS and H_2S concentrations. The average concentrations of DS and H_2S obtained both from the SPM and validated BISM were 1.95 mg S/L and 214 ppm, respectively. These values closely corresponded to the measured average concentrations of 1.82 mg S/L and 219 ppm. Moreover, both the SPM and validated BISM successfully replicated the production and emission patterns of DS and H_2S, accurately predicting the timing of their peaks and valleys. It is noteworthy that in this simulation, SPM swiftly simulated DS and H_2S

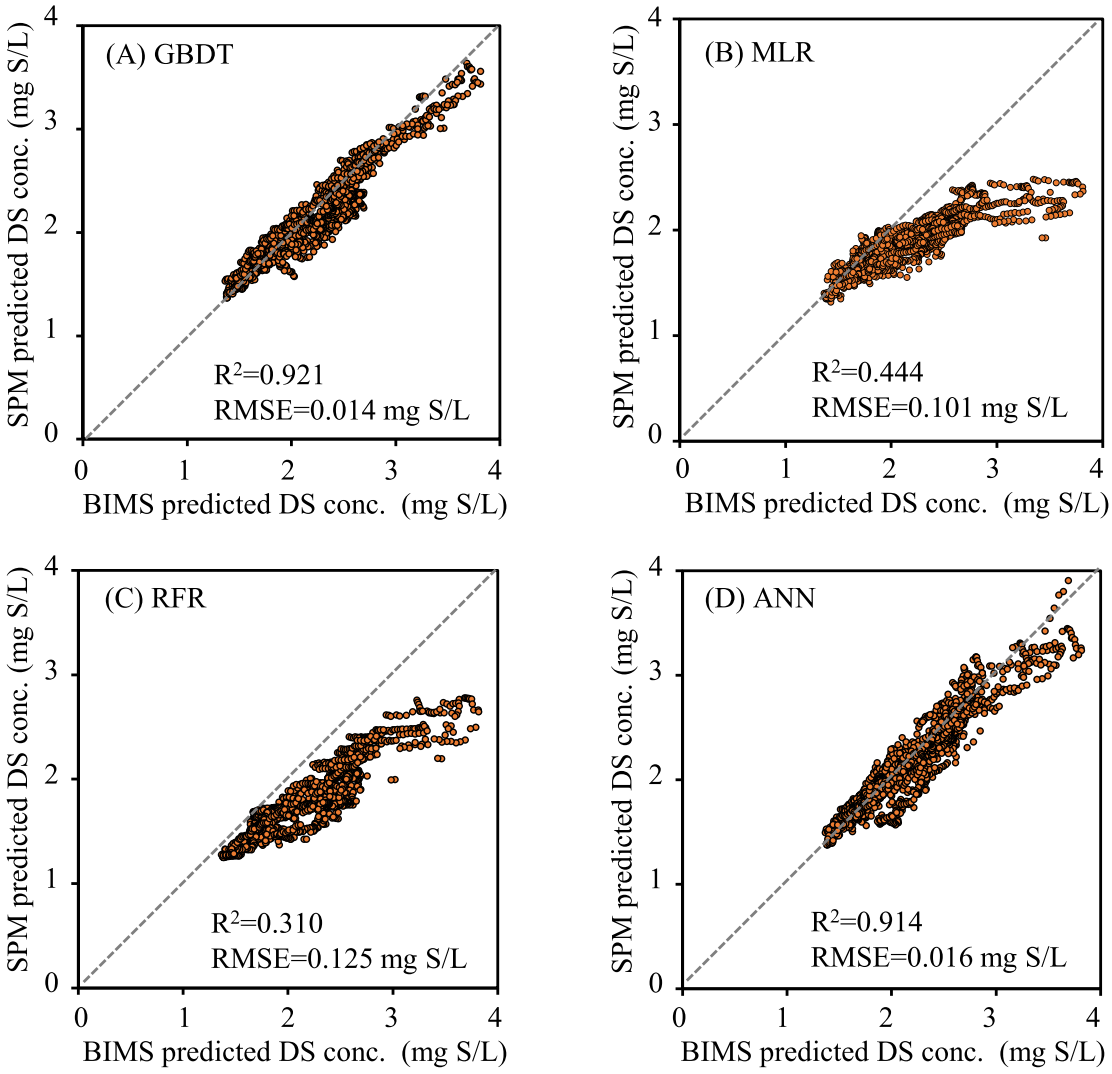


Fig. 1. Simulation results of the SPM based on machine learning algorithms for DS in the testing set: GBDT (A), MLR (B), RFR (C), ANN (D). Note: The data is sourced from simulated data generated by the validated BISM model.

within a mere 0.2 s and 0.3 s, respectively. This standed in stark contrast to the computational times of validated BISM, which were more than 4200 s and 5000 s for the same simulations, showcasing a significant increase in computational efficiency by thousands of times.

These findings underscored the robustness of the SPM, which integrates the validated BISM with MLAs, in effectively simulating and predicting variations in DS and H_2S concentrations within a complex sewer system. Despite the multitude of factors influencing the production and emission of H_2S in sewer systems, as well as the significant concentration fluctuations, the SPM provides a scientific foundation for the swift and efficient management of H_2S pollution.

2.3. Model application for sulfide control

Based on the accuracy and efficiency of GBDT-based SPM, we optimized the sulfide control scheme of Harbour Area Treatment Scheme Stage 1 Sewage Conveyance System (HATS 1) in Hong Kong through simulation using GBDT-based SPM. The specified control targets for HATS 1 outlet were established as maintaining DS concentration $\leq$ 1 mg S/L and $H_2S \leq$ 20 ppm.

2.3.1. Determination of the optimal dosing location

Selecting appropriate dosing points for controlling the effluent DS concentration is crucial, as it directly influences sulfide control effectiveness and chemical costs. Accordingly, SPM was employed to evaluate the optimal nitrate addition location by simulating the individual effects of a fixed dosing rate of solid calcium nitrate dosing at Preliminary Treatment Works (PTWs) located at TKW, KT, TKO, SKW, CW, and TY on the concentrations of DS and H_2S at the outlet of HATS 1 (i.e., Stonecutters Island Sewage Treatment Works (SCISTWs)). The simulated results indicate that, under equivalent nitrate dosing, the optimal effectiveness in reducing both DS and H_2S concentration was observed when nitrate was added at the TKW PTW (Fig. S3). For instance, as illustrated in Fig. 4, dosing nitrate at a rate of 3150 kg/h at TKW PTW led to a substantial decrease in the average DS concentration from 1.94 mg S/L to 0.95 mg S/L, representing a notable decrease of 51 %. In contrast, other locations only achieved a decrease of 2 %–33 % in the effluent DS concentration. These findings align with the study conducted by Gutierrez et al. (2010), which found that dosing nitrate near the end of the sewer was more effective in reducing the effluent DS concentration, leading to a 42 % reduction in nitrate consumption. Similarly, the most effective H_2S control occurred when nitrate was dosed at TKW PTW, resulting in a significant 53 % reduction in the average H_2S

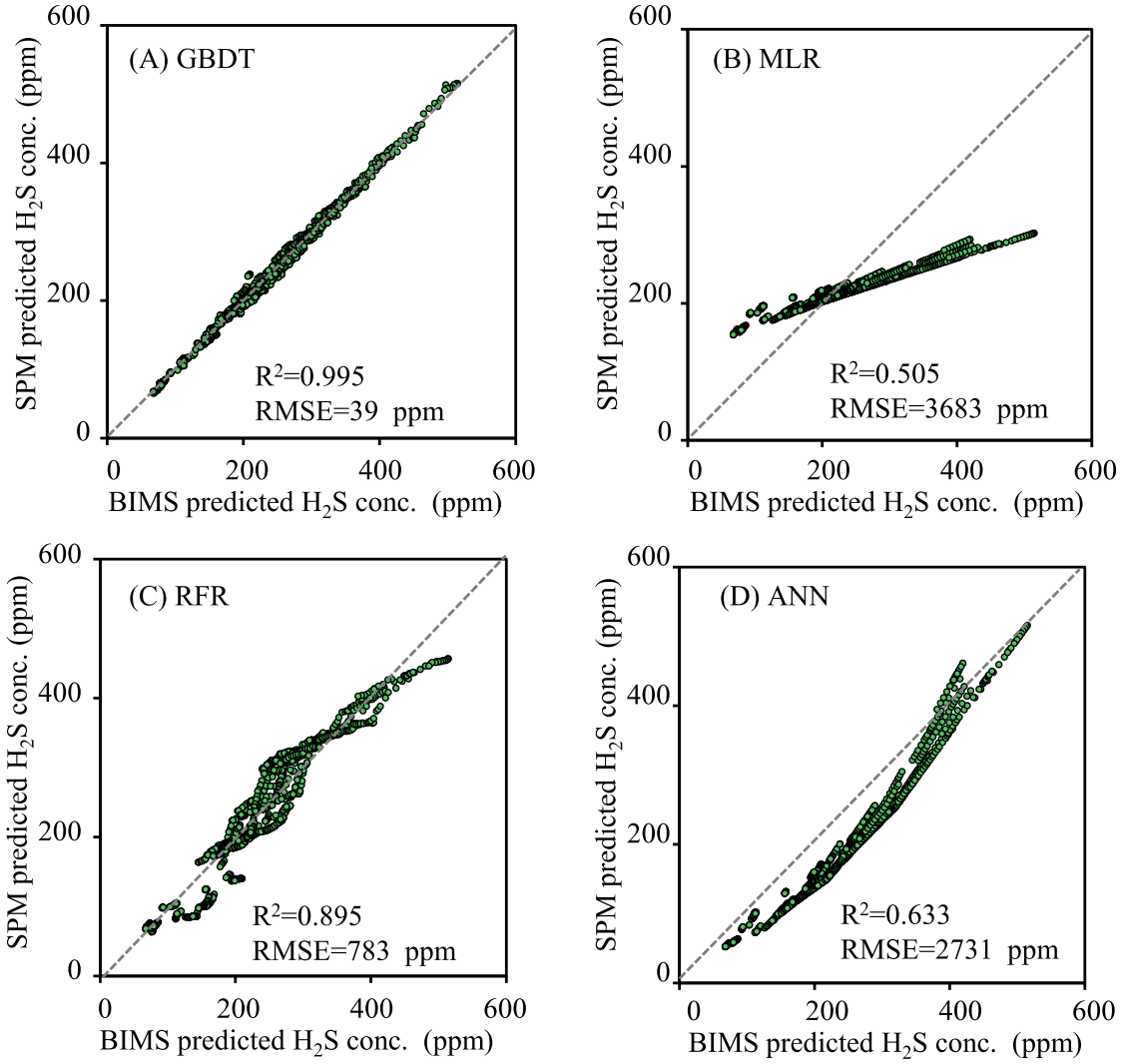


Fig. 2. Simulation results of the SPM based on machine learning algorithms for H_2S in the testing set: GBDT (A), MLR (B), RFR (C), ANN (D). Note: The data is sourced from simulated data generated by the validated BISM model.

Table 1
The computation time of the machine learning-based SPM and the dynamic sewer process model BISM on the validation set.

		SPM				Validated BISM
		GBDT	MLR	RFR	ANN	
Computation time (*sec*)	DS	0.133	0.005	0.327	0.991	>4200
	H_2S	0.252	0.004	0.254	1.146	>5000

concentration, while the other locations only achieved a decrease of 2 %−34 % (Fig. S3B).

As illustrated in Fig. S4A, a fixed nitrate dosing method generally allows for achieving an average DS concentration of less than 1.0 mg S/L, but there was still significant fluctuation in DS concentration, ranging from 0.4 to 1.9 mg S/L. Moreover, the H_2S concentration failed to reach the targeted 20 ppm by such a dosing method. With the increasing dosage of nitrate, the reduction of H_2S reached a plateau limit at approximately 50 ppm (Fig. S3B). One important reason is that fixed dosing cannot adapt to the dynamic changes in water quantity, water quality and biofilm formation within the sewer system in real-time, leading to either insufficient or excessive dosing. To minimize DS fluctuation (Fig. S4A), it was necessary to dynamically adjust the dosing rate in real-time.

2.3.2. Sulfide control by real-time and dynamic dosing of nitrate at TKW PTW

With the high computational efficiency, SPM can swiftly determine the optimal dosing rate, enabling dynamic control of sulfide. In contrast, validated dynamic sewer process models such as BISM require an h or even longer to calculate the relationship between dosing rate and sulfide concentration under specific water quality and quantity conditions. Compared to a conventional constant dosage method, the dynamic dosing through SPM has reduced the fluctuation range of DS concentration and increased the compliance rate from 54 % to 86 % (Fig. 5A). Furthermore, the dynamic dosing through SPM has also improved the control effect of H_2S (Fig. 5B).

Importantly, SPM not only optimizes sulfide control effectiveness but also reduces the required chemical dosage, mitigating issues of excessive or insufficient chemical dosages. As illustrated in Fig.5C, the average dosing rate of $Ca(NO_3)_2$ (solid) with SPM was 3070 kg/h, which was 2.5 % lower than the sulfide control scheme provided by validated BISM.

It should be noted that, whether using the fixing dosage method simulated by validated BISM or the dynamic dosing method simulated by SPM, the application of nitrate only to TKW as a sulfide control

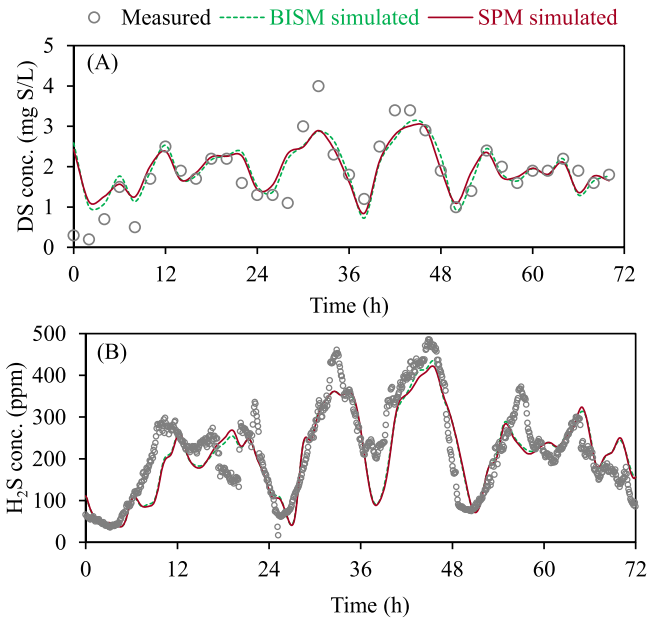


Fig. 3. Simulation results of validated BISM and GBDT-based SPM for measured DS (A) and H_2S (B) concentrations at the sewer system outlet. Note: The data is sourced from field-measured data.

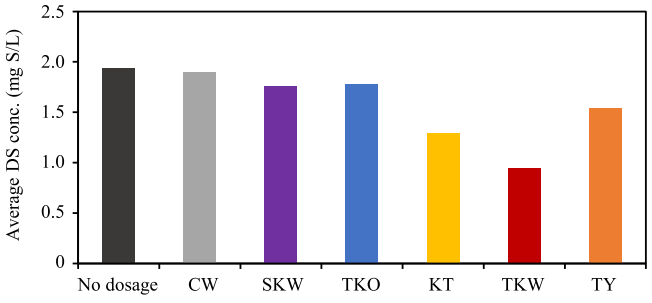


Fig. 4. Prediction of the control effectiveness of $Ca(NO_3)_2$ reagent added at a rate of 3150 kg/h at different locations on the effluent concentrations of DS at the sewer system outlet.

strategy does not result in a DS concentration at the outlet of HATS 1 below 1 mg S/L at all times. Moreover, the concentration of H_2S still failed to meet the target concentration ($\leq$ 20 ppm). Therefore, additional sulfide control measures need to be adopted to effectively control DS and H_2S.

2.3.3. *Sulfide control by real-time and dynamic dosing of nitrate at both TKW PTW and TY PTW*

Nitrate has demonstrated efficacy in managing sulfide levels within sewer systems (Liang et al. 2023). However, despite the $Ca(NO_3)_2$ (solid) dosage reaching 9000 kg/h, the average H_2S concentration consistently exceeds 20 ppm, as indicated in Fig. S3B. Moreover, sporadic occurrences show that the outlet DS concentration surpasses 1 mg S/L. This occurrence can be attributed to the combined influence of tunnels E and F on sulfide levels in the effluent, as illustrated in Fig. 6. Therefore, relying solely on nitrate dosing at the TKW PTW for sulfide control within tunnel E cannot guarantee compliance with DS concentration standards ($\leq$ 1 mg S/L) at all times (Fig. 5B). Consequently, this part introduced SPM to assess the concurrent application of nitrate at both the TKW PTW and TY PTW for sulfide management (a detailed comprehensive methodology description was described in Text S1).

As illustrated in Fig. 5A, the simultaneous dynamic addition of

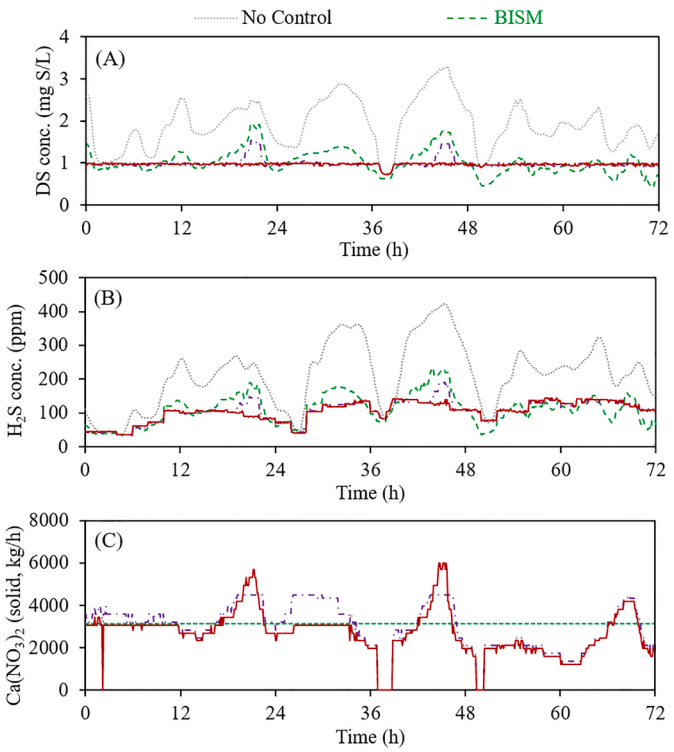


Fig. 5. The sulfide control effects of the nitrate dosing schemes simulated using validated BISM and SPM: DS (A), H_2S (B), and dosing rate (C).

nitrate at both the TKW PTW and TY PTW demonstrated superior sulfide control in the effluent compared to dosing solely at the TKW PTW, maintaining a DS concentration below 1 mg S/L throughout all periods. The average dosage necessary for this improvement is 2707 kg $Ca(NO_3)_2$/h, which was 14.1 % lower than the conventional constant dosing scheme provided by validated BISM. Despite the improved efficacy in controlling H_2S with this approach (Fig. 5B), the average H_2S concentration reaches up to 103 ppm, exceeding the required standards ($\leq$ 20 ppm). Therefore, more robust control measures must be implemented to mitigate the H_2S pollution effectively.

2.3.4. *Optimization of dosing scheme*

According to Zhang et al. (2023b), raising the pH of wastewater has proven to be an effective method for H_2S control. Therefore, in this study, the addition of alkalis has been further adopted to enhance the efficacy of H_2S control at the outlet of HATS 1. We developed a model to determine the necessary dosage of NaOH (solid) to achieve the target pH of the sewage under specific alkalinity and pH conditions (detailed in Text S2 and Fig. S5). This model allows for accurate calculation of the required NaOH (solid) dosage to reach the desired pH. Subsequently, employing the SPM model, we simulated a dynamic dosing scheme involving $Ca(NO_3)_2$ and NaOH at the TKW PTW and TY PTW location to effectively control DS and H_2S at the outlet of HATS 1.

Based on simulation results, the combined dynamic dosing scheme of nitrate and alkalis effectively achieved sulfide control. Although with a fixed dosing scheme at a rate of 3150 kg/h for $Ca(NO_3)_2$ and 1650 kg/h for NaOH, the average concentration of H_2S reduced to 20 ppm. However, this approach still carried the risk of overdosing or underdosing of chemicals, resulting in H_2S concentration fluctuations ranging from 0 to 105 ppm, with a compliance rate of only 69 %. The dynamic nitrate and alkali dosing scheme implemented through the SPM model proved to be more effective in controlling sulfide. The average H_2S concentration reduced to 17 ppm, exhibiting minimal fluctuations and mostly remaining below the target range of 10–20 ppm. This dynamic scheme achieved an impressive compliance rate of 100 %, which cannot be

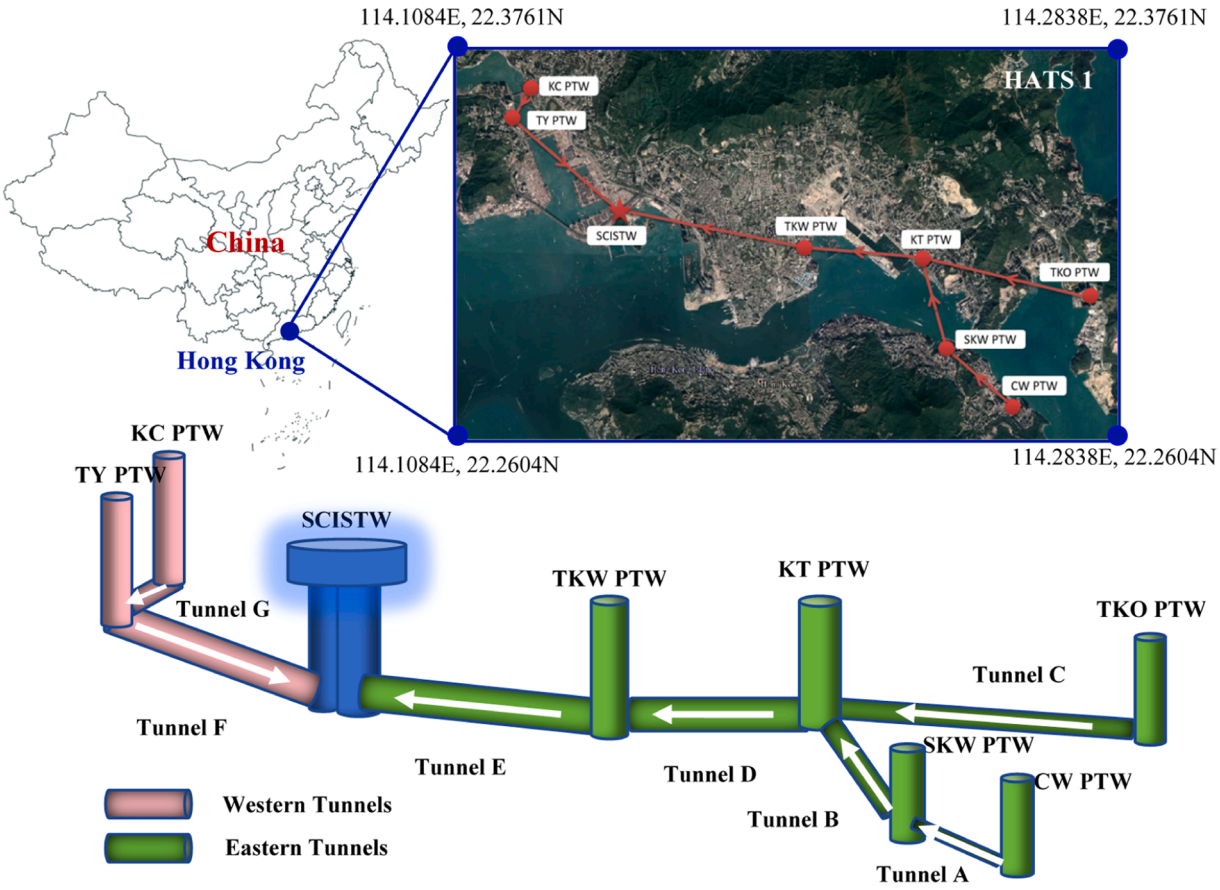


Fig. 6. Schematic diagram of the study area.

achieved by the dynamic sewer process model. Moreover, in comparison to the conventional constant dosing scheme provided by validated BISM, the implementation of SPM led to cost savings in chemical usage. The dosage rates for nitrate and NaOH were optimized at 2707 kg/h and 1339 kg/h, respectively, resulting in savings of 14 % and 19 % chemicals from the constant dosing scheme, respectively.

The aforementioned findings highlight the efficacy of the SPM model in achieving real-time, effective, and cost-efficient control of H_2S in complex sewer networks, thereby showing substantial application potential. It is imperative to underscore that the calibration and validation of the BISM are indispensable in practical scenarios, as they are essential for utilizing the BISM to generate simulation data for training MLA. This process facilitates the accurate real-time forecasting and online management of hydrogen sulfide contamination in complex sewer networks.

3. Conclusions

In order to achieve real-time, dynamic, and effective control of H_2S in sewer systems, this study introduces a swift prediction model for H_2S prediction that integrates a validated dynamic sewer process model and MLA. The method of training MLA using simulation data generated by BISM validated by field-measured data can swiftly predict the DS concentration and H_2S concentration in a specific sewer network. The computation time of GBDT-based SPM for a case simulation was less than 0.3 s, representing a significant improvement in applicability over dynamic sewer process models, which are thousands of times slower. The method of training MLA using simulation data generated by BISM validated by field-measured data can swiftly predict the DS concentration and H_2S concentration in a specific sewer network. The computation time of GBDT-based SPM for a case simulation was less than 0.3 s, representing a significant improvement in applicability over dynamic sewer process models, which are thousands of times slower. Overall, the newly developed SPM proves adept at swiftly and accurately predicting the production and discharge patterns of H_2S, enabling real-time, effective, and economic control. With the continuous expansion of sewer systems, H_2S pollution has intensified. The SPM can not only provide various strategies for sulfide control and mitigates the occurrence of H_2S-related safety accidents but also holds significant potential for future developments in sewage management and safety operation and maintenance. To enhance the model's generalizability and performance, a follow-up study may be exploring the utilization of simulation data for data filtering, augmentation, and reinforcement learning.

4. Materials and methods

4.1. Case for simulation

The primary focus of the study is to enhance the efficiency of H_2S pollution control in sewer system, such as the HATS 1 in Hong Kong, China. This region is characterized by a marine subtropical monsoon climate. Under this climate, the sewer network in the region is vulnerable to sulfate reduction reactions, leading to significant H_2S pollution. As depicted in Fig. 6, the sewer network consists of seven tunnels, incorporating seven PTWs and a centralized sewage treatment plant (i.e., SCISTWs). The entire system spans a total length of 23.3 km, with depths ranging from 70 to 160 m and pipe diameters ranging from 1.2 to 3.5 m.

During the operational period of HATS 1, the occurrence of H_2S contamination was discovered, prompting a thorough investigation by the relevant regulatory authorities. A comprehensive 72-h campaign was conducted by relevant authorities at seven inlet points (KC, TY, TKW, KT, TKO, SKW, and CK PTWs) and one outlet point (wet well at

SCISTW) within HATS 1. This effort resulted in the acquisition of a substantial amount of field-measured data, including soluble chemical oxygen demand (SCOD), sulfate (SO_4^{2-}), dissolved oxygen (DO), pH, temperature (Temp), volatile fatty acid (VFA), alkalinity (ALK), total chemical oxygen demand (TCOD), total suspended solid (TSS), volatile suspended solid (VSS), soluble ammonia (NH_3—N), total Kjeldahl nitrogen (TKN), nitrate(NO_3^-), nitrite (NO_2^-), sewage flowrate, total sulfide (TS), DS and H_2S concentration. The main objective of this campaign was to assess the extent of H_2S pollution in HATS 1, identify sources and key factors contributing to H_2S generation, and propose effective solutions to mitigate sulfide pollution and associated risks in the sewer systems. Accordingly, this study utilized the aforementioned field-measured data as the data sources for subsequent research. Further details can be found in the literature (Liang et al. 2019a).

4.2. Model establishment, validation and application

The SPM established in this study represents a novel approach that integrates a validated dynamic sewer process model with MLA. The dynamic sewer process model employed in this study was based on our previously developed BISM model, which has proven effective in predicting sulfide production in sewer systems. Its success is evident from its application in multiple sewer systems in Hong Kong (Liang et al. 2019a; Liang et al. 2019b). Despite its accurate simulation of sulfide pollution in sewer systems, calibrated and validated using field-measured data, a drawback of the BISM is its lengthy computation time. This limitation hinders its practical use for real-time control simulation of sulfide. Therefore, the primary goal of the SPM is to swiftly and accurately predict concentrations of DS and H_2S in sewer systems. Given the temporal and spatial disparities inherent in diverse sewer systems, it is imperative that all models undergo on-site data calibration and validation procedures prior to their implementation. Dynamic sewer process models, as such BISM, are designed to simulate the migration and transformation processes of various substructures. To ensure accurate simulation, it is crucial to calibrate key parameters using on-site measured data from the target sewer. Conversely, MLA function as a black box model that also requires training and validation with substantial amounts of target sewer data before being utilized. However, the limited on-site measured data from HATS 1, spanning only three consecutive days, is insufficient for training MLA. Therefore, this study leverages the validated BISM model to generate simulation data for HATS 1, facilitating the training and validation of the MLA model. Subsequently, the validated MLA model is confirmed through comparison with on-site measured data. Additionally, the SPM allowed for the swift determination of optimization schemes for real-time H_2S control. The flowchart of the SPM is illustrated in Fig. 7, outlining the sequential steps involved in its implementation:

Step 1: Simulation data generation and preprocessing. A large amount of simulation data was generated from dynamic sewer process model BISM validated by field-measured data undergoes specific preprocessing methods (detailed in Section 4.2.1). Afterwards, standardize the simulation data and divide the dataset into training and testing sets in a 4:1 ratio.

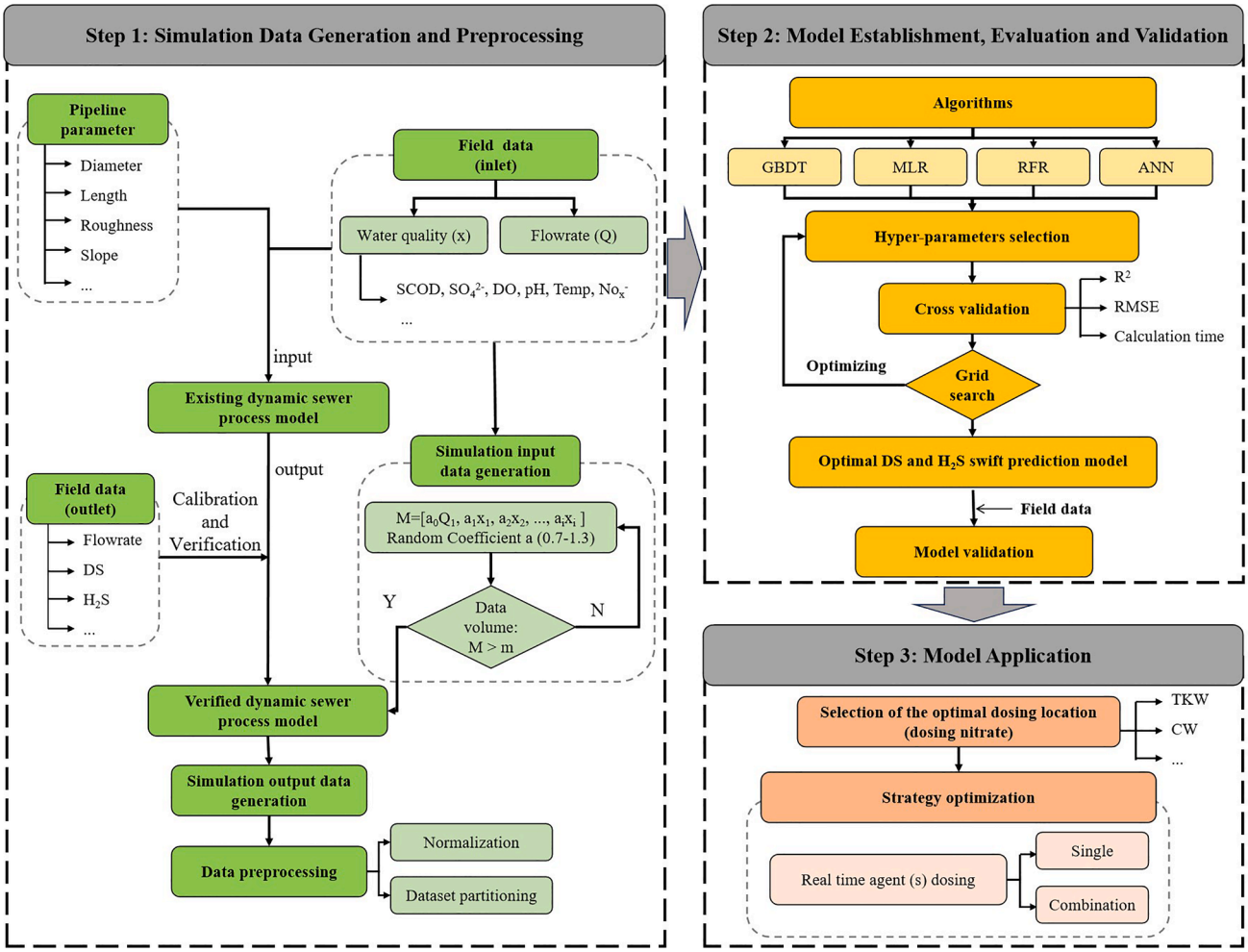


Fig. 7. Flowchart of SPM.

Step 2: Model establishment, evaluation and validation. Utilize simulated data to train and test the MLAs, assessing their accuracy and efficiency in predicting DS and H_2S. Following this, validate the performance of the selected MLA using field-measured data to confirm its reliability and suitability in real sewer systems.
Step 3: Model application. The selected MLA-based SPM was applied to simulate real-time dynamic dosing of optimization schemes for sulfide control in the sewer systems.

4.2.1. Simulation data generation and preprocessing

As a general principle, the reliability of simulation results obtained through ML tends to increase with larger volumes of data. In our previous research, we successfully utilized BISM, a dynamic sewer process model validated by field-measured data, to accurately predict the production and fluctuation patterns of the sulfide in HATS 1 (Liang et al. 2019b). Additionally, key influencing factors were identified through this application (Liang et al. 2019b).

Therefore, to address the lack of monitoring data in HATS 1, the approach adopted involved integrating limited field-measured data with validated BISM to generate a large amount of simulation data. The specific methodology is as follows: First, to account for variations in water quantity and quality in real sewer systems, the data measured onsite for three consecutive days were cyclically processed. The input parameters, including flow rate, organic matter, sulfate, etc., underwent multiplication by a random fluctuation coefficient "a" (ranging from 0.7 to 1.3) to generate 10,800 sets of simulated input data. Secondly, the simulated input data were used as the model input for validated BISM to simulate the corresponding concentrations of DS and H_2S at the outlet. Finally, the total of 10,800 sets of simulated inlet and outlet data were combined to form a comprehensive dataset for training and validating subsequent ML models.

Furthermore, to mitigate the impact of substantial numerical variations in the data, we employed normalization for all the data using the formula (1). This normalization formula uniformly scales all the simulated data to the interval [0,1].

$$X_i^{'} = \frac{x_i - x_{i,min}}{x_{i,max} - x_{i,min}} \tag{1}$$

Where $X_i^{'}$represents the normalized data, $x_i$represents the data before normalization, $x_{i,max}$ represents the maximum value in the original dataset, and $x_{i,min}$ represents the minimum value in the original dataset.

4.2.2. Model establishment, evaluation and validation

The MLA investigations detailed in this paper were conducted utilizing the Jupyter notebook framework and Python programming language (Computer configuration: 11th Gen Intel (R) Core (TM) i5–1135G7 @ 2.40 GHz, 2.42 GHz, 16GB RAM). Four extensive MLAs were employed to construct SPM, including GBDT, MLR, RFR and ANN (Zhu et al. 2022). The GBDT, MLR and RFR algorithms were implemented in Python using the Scikit-learn package, while the ANN algorithm was implemented using the advanced deep learning library, Keras. Detailed hyperparameters for each algorithm are available in Text S3 and Table S1.

To evaluate the performance of the SPM, we utilized the following indicators for assessment. Firstly, the computation time of the models was evaluated using the time calculation function within the Python software package. Additionally, R^2 and RMSE were utilized to evaluate the predictive capability of the models. R^2 is a commonly used statistical parameter that quantifies the extent to which the modeling process captures the variability in the data. It is calculated using the formula (2). A higher R^2 value indicates a stronger predictive ability of the model, with a value closer to 1 being desirable. RMSE is a measure that quantifies the disparity between predicted and actual values. It is calculated using the formula (3). The magnitude of RMSE is influenced by the measurement scale, but generally, a smaller RMSE suggests a more accurate match between predicted and actual values.

$$R^2 = 1 - \frac{\sum_1^m (y_i - \hat{y}_i)^2}{\sum_1^m (y_i - \bar{y})^2} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{m} * \sum_1^m (y_i - \hat{y}_i)^2} \tag{3}$$

Where m represents the number of calculated samples in the dataset, $\bar{y}$ represents the average of the actual measurement results in the dataset, y_i and $\hat{y}_i$ represent the actual measurement results and the model prediction results for i data point, respectively.

In this study, the dataset was initially partitioned into input variables and output variables. Specifically, the input data for the model included water quantity and quality parameter from all wastewater intakes, while the model's output consisted DS and H_2S data at the outlet location. Subsequently, 10,800 sets of simulated data were split into a training dataset and a validation dataset, maintaining a ratio of 4:1. The first 8640 sets of data were assigned to the training dataset, employed for training the SPM. To identify the optimal parameters for the SPM using specific algorithms, cross-validation and grid search techniques were employed. Additionally, the remaining 2160 sets of data were reserved for the validation of the SPM, ensuring an unbiased evaluation of its performance.

4.2.3. Model application

To evaluate the swiftness of the SPM in predicting DS and H_2S concentrations, as well as its viability for real-time dynamic dosing, a series of application simulations was undertaken. Initially, The SPM was utilized to simulate the field-measured data in HATS 1, and a comparative analysis was conducted between the simulation results of the SPM and validated BISM models. Subsequently, certain simulations encompassed the identification of optimal dosing locations, the implementation of real-time nitrate dosing for sulfide (both DS and H_2S) control at the outlet of HATS 1, and the optimization of sulfide control schemes through the combined dosing of nitrate and alkalis.

Considering the associated hazards of H_2S and the expertise of relevant management personnel, it was determined that the sulfide control targets for the outlet of HATS 1 should be DS $\leq$ 1 mg S/L and H_2S $\leq$ 20 ppm. To ensure the optimality of the real-time dynamic dosing sulfide control scheme, an optimal dosing planning model for DS and H_2S concentration at the outlet of HATS 1 was established based on the SPM.

The relationship between water quality and quantity conditions in the sewer system, chemical dosages, and the predicted concentrations of DS and H_2S at the outlet of HATS 1 are presented by formulas (4) and (5). Additionally, considering the constraint conditions (formula (6)) requiring the DS and H_2S concentrations to be reduced below the target values, a minimum dosing model can be formulated to determine the optimal dosing quantity.

$$DS_i = f(x, y, z_i) \tag{4}$$

$$H_2S_i = g(x, y, z_i) \tag{5}$$

$$\min(\mathrm{sum}(x, y)) \ \text{s.t.} \begin{cases} DS_i \leq m \\ H_2S_i \leq n \end{cases} \tag{6}$$

Where, x and y represent the dosage of nitrate chemicals and NaOH chemicals, respectively. $z_i (i = 1, 2, 3, \ldots, i)$ represents a set of input water quality and quantity data. f and g represent the concentration prediction methods for DS and H_2S provided by SPM, respectively. DS_i and H_2S_i represent the predicted concentrations of DS and H_2S at the outlet of HATS 1, respectively. m and n represent the target concentration of DS in the effluent and H_2S emissions, respectively.

In this study, real-time dynamic dosing sulfide control simulations

were conducted using the field-measured data. To assess and compare the effectiveness and efficiency of the SPM and the validated BISM, a comprehensive examination of the disparities between the two models was undertaken. Metrics such as the compliance rate, dosing quantity, and model computation time were employed to evaluate their performance.

CRediT authorship contribution statement

Zhensheng Liang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition. **Wenlang Xie:** Writing – original draft, Investigation. **Hao Li:** Investigation. **Yu Li:** Writing – review & editing, Writing – original draft. **Feng Jiang:** Writing – review & editing, Supervision, Methodology, Conceptualization, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

The authors acknowledge the supports from the National Natural Science Foundation of China (52300124), the China Postdoctoral Science Foundation (2023M734002), the Key Research and Development Program of Guangzhou (2023B03J0007), and the Science and Technology Planning Project of Guangdong Province (2021A0505020010).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.wroa.2024.100230.

References

Bernardelli, A., Marsili-Libelli, S., Manzini, A., Stancari, S., Venier, S., 2020. Real-time model predictive control of a wastewater treatment plant based on machine learning. Water Sci. Technol. 81 (11), 2391–2400.
Canete, J.F.D., Saz-Orozco, P.D., Gómez-De-Gabriel, J., Baratti, R., Rivas-Blanco, I., 2021. Control and soft sensing strategies for a wastewater treatment plant using a neuro-genetic approach. Comput. Chem. Eng. 144, 107146.
Dovzan, D., Logar, V., Skrjanc, I., 2015. Implementation of an evolving fuzzy model (eFuMo) in a monitoring system for a waste-water treatment process. IEEE Trans. Fuzzy Syst. 23 (5), 1761–1776.
Eerikinen, S., Haimi, H., Mikola, A., Vahala, R., 2020. Data analytics in control and operation of municipal wastewater treatment plants: qualitative analysis of needs and barriers. Water Sci. Technol. 82 (12), 2681–2690.
Félix, H.-d.-O., Gaudioso, E., Duro, N., Dormido, R., 2019. Machine learning weather soft-sensor for advanced control of wastewater treatment plants. Sensors 19 (14), 3139.
Francesco, G., Stefano, P., Giovanni, E., Rudy, G., Giovanni, D.M., 2017. Machine learning algorithms for the forecasting of wastewater quality indicators. Water (Basel) 9 (2), 105.
Frank, M., Drikakis, D., Charissis, V., 2020. Machine-learning methods for computational science and engineering. Computation 8 (1), 15.
Guidotti, T., 2010. Hydrogen sulfide: advances in understanding human toxicity. Int. J. Toxicol. 29 (6), 569.
Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y.M., Park, J.P., Kim, J.H., Cho, K.H., 2015. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. J. Environ. Sci. 32, 90–101.
Gutierrez, O., Sutherland-Stacey, L., Yuan, Z.G., 2010. Simultaneous online measurement of sulfide and nitrate in sewers for nitrate dosage optimisation. Water Sci. Technol. 61 (3), 651–658.
Huisman, J., Gujer, W., 2002. Modelling wastewater transformation in sewers based on ASM3. Water Sci. Technol. 45 (6), 51–60.
Hvitved-Jacobsen, T., Vollertsen, J., Nielsen, A.H., 2013. Sewer Processes: Microbial and Chemical Process Engineering of Sewer Networks. CRC press, pp. p221–p225.
Jiang, G.M., Sun, J., Sharma, K.R., Yuan, Z.G., 2015. Corrosion and odor management in sewer systems. Curr. Opin. Biotechnol. 33, 192–197.
Juan, G., Antonio, V.R., Luis, C., José, C., 2017. Evaluation of sulfide control by air-injection in sewer force mains: field and laboratory study. Sustainability 9 (3), 402.
Jiang, Y., Li, C., Zhang, Y., Zhao, R., Yan, K., Wang, W., 2021. Data-driven method based on deep learning algorithm for detecting fat, oil, and grease (FOG) of sewer networks in urban commercial areas. Water Res. 207, 117797.
Liang, S., Zhang, L., Jiang, F., 2016. Indirect sulfur reduction via polysulfide contributes to serious odor problem in a sewer receiving nitrate dosage. Water Res. 100, 421–428.
Liang, Z.S., Sun, J.L., Chau, H.K.M., Leong, E.I.M., Wu, D., Chen, G.H., Jiang, F., 2019a. Experimental and modelling evaluations of sulfide formation in a mega-sized deep tunnel sewer system and implications for sewer management. Environ. Int. 131, 105011.
Liang, Z.S., Wu, D.P., Li, G.B., Sun, J.L., Jiang, F., Li, Y., 2023. Experimental and modeling investigations on the unexpected hydrogen sulfide rebound in a sewer receiving nitrate addition: mathematical and solution. J. Environ. Sci. 125, 630–640.
Liang, Z.S., Zhang, L., Wu, D., Chen, G.H., Jiang, F., 2019b. Systematic evaluation of a dynamic sewer process model for prediction of odor formation and mitigation in large-scale pressurized sewers in Hong Kong. Water Res. 154, 94–103.
MacDonald, N., 2013. Time Lags in Biological Models. Springer Science & Business Media, p. 27.
Mika, L., Ilkka, L., Yrj, H., 2013. Advanced monitoring platform for industrial wastewater treatment: multivariable approach using the self-organizing map. Environ. Modell. Soft. 48 (10), 193–201.
Ray, S., 2019. A Quick Review of Machine Learning algorithms. 2019 International conference On Machine learning, Big data, Cloud and Parallel Computing (COMITCon). IEEE, pp. 35–39.
Sharma, K.R., Yuan, Z.G., de Haas, D., Hamilton, G., Corrie, S., Keller, J., 2008. Dynamics and dynamic modelling of H_2S production in sewer systems. Water Res. 42 (10–11), 2527–2538.
Yan, J., Jin, J.M., Chen, F.R., Yu, G., Yin, H.L., Wang, W.J., 2018. Urban flash flood forecast using support vector machine and numerical simulation. J. Hydroinf. 20 (1), 221–231.
Zhang, L., De Schryver, P., De Gusseme, B., De Muynck, W., Boon, N., Verstraete, W., 2008. Chemical and biological technologies for hydrogen sulfide emission control in sewer systems: a review. Water Res. 42 (1), 1–12.
Zhang, Y., Li, C., Duan, H., Yan, K., Wang, J., Wang, W., 2023a. Deep learning based data-driven model for detecting time-delay water quality indicators of wastewater treatment plant influent. Chem. Eng. J. 467, 143483.
Zhang, L., Qiu, Y.Y., Sharma, K.R., Shi, T., Song, Y.R., Sun, J.L., Liang, Z.S., Yuan, Z.G., Jiang, F., 2023b. Hydrogen sulfide control in sewer systems: a critical review of recent progress. Water Res. 240, 120046.
Zhu, M.Y., Wang, J.W., Yang, X., Zhang, Y., Zhang, L.Y., Ren, H.Q., Wu, B., Ye, L., 2022. A review of the application of machine learning in water quality evaluation. Eco-Environ. Health 1, 107–116.