Environmental Science and Ecotechnology 23 (2025) 100512

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Augmented machine learning for sewage quality assessment with limited data

Jia-Qiang Lv [a, b, 1], Wan-Xin Yin [c, 1], Jia-Min Xu [b], Hao-Yi Cheng [b], Zhi-Ling Li [a], Ji-Xian Yang [a], Ai-Jie Wang [a, b], Hong-Cheng Wang [a, b, *]

[a] State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin, 150090, China
[b] School of Civil and Environmental Engineering, Harbin Institute of Technology Shenzhen, Shenzhen, 518055, China
[c] College of the Environment, Liaoning University, Shenyang, 110036, China

ARTICLE INFO

Article history:
Received 3 June 2024
Received in revised form
12 November 2024
Accepted 12 November 2024

Keywords:
Machine learning
Hybrid model
Mechanistic augmentation
Sewer system
Sulfide and methane

ABSTRACT

Physical, chemical, and biological processes within sewers significantly alter sewage composition during conveyance. This leads to the formation of sulfide and methane—compounds that contribute to sewer corrosion and greenhouse gas emissions. Reliable modeling of these compounds is essential for effective sewer management, but the development of machine learning (ML) models is hindered by differences in data accessibility and sampling frequencies of water quality variables. Here we present a mechanistically enhanced hybrid (ME-Hybrid) model that combines mechanistic modeling with data-driven approaches. This model harmonizes datasets with varying sampling frequencies and generates synthetic samples for ML training, thereby enhancing the monitoring of methane and sulfide in sewers. The optimal ME-Hybrid model integrates the backpropagation neural network with mechanistic frequency harmonization. We demonstrate that the ME-Hybrid model outperforms pure ML and linear interpolation in capturing fluctuating trends and extremes of sulfide concentrations, achieving a coefficient of determination (R^2) of 0.94. Synthetic samples generated through mechanistic augmentation closely approximate real samples in modeling performance, statistical distribution, and data structure. This enables the model to maintain high predictive accuracy ($R^2 > 0.76$) for sulfide even when trained on only 50 % of the dataset. Additionally, the ME-Hybrid model successfully assesses sewer methane concentrations with an R^2 of 0.94, validating its applicability and generalization ability. Our results provide a reliable methodological framework for modeling and prediction under data scarcity. By facilitating better monitoring and management of sewer systems, the ME-Hybrid model aids in the development of strategies that minimize environmental impacts, enhance urban resilience, and ultimately lead to sustainable urban water systems.

© 2024 Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Urban sewer systems, described as the veins of cities, are responsible for collecting and transporting sewage to wastewater treatment plants for purification [1,2]. These underground sewers are essential for safeguarding public health and the environment by protecting groundwater quality, preventing contamination, and upholding fundamental safety standards [3,4]. Despite their significance, sewer systems face challenges such as inadequate regulation and maintenance due to lacking accurate water quality assessments [5]. Corrosive damage, primarily due to sulfidic activity, threatens the sewer infrastructure, reducing its lifespan and increasing the risk of leaks [6,7]. The organic components in sewers are degraded during conveyance, resulting in a low carbon-to-nitrogen ratio, which impedes nutrient removal in conventional secondary treatment processes in wastewater treatment plants [8,9]. Furthermore, methane emissions, which significantly contribute to the greenhouse effect and pose an explosion risk, are often underestimated [1,10]. Therefore, developing robust water quality monitoring methods in sewers is essential for sustainable

[*] Corresponding author. State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin, 150090, China.
E-mail address: wanghongcheng@hit.edu.cn (H.-C. Wang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ese.2024.100512

2666-4984/© 2024 Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

water resource management [11].

Recent advancements in machine learning (ML) technologies have introduced considerable potential for applications in the environmental field. Many studies have employed ML techniques to analyze water, soil, and atmospheric data [12–14]. In the context of sewer system failures, including leaks [15,16], blockages [17,18], and overflows [19], substantial ML-based studies have already been conducted. Furthermore, ML techniques have been utilized to detect the accumulation of fats and oils in sewer networks [20], predict microbial-induced corrosion [21], and simulate the transformation of dissolved organic matter in sewage [22]. However, as decision-making tools, ML models heavily rely on high-quality and abundant data for calibration to minimize uncertainties. Data collection challenges have hindered the widespread use of ML models calibrated with real-world data [23]. Sulfide and methane production in sewers is influenced by a combination of hydraulic conditions, water quality, and sewer structure [9], making the absence of models calibrated with real-world data even more pronounced. Due to limitations in sensor acquisition, installation, maintenance, and failure, datasets with small sample sizes and irregular sampling frequencies impede the successful development of ML models for predicting sulfide and methane concentrations [24,25]. Thus, innovative approaches are needed to address the challenges of modeling sewer water quality with constrained datasets.

Data augmentation techniques help mitigate challenges related to data scarcity by enriching datasets by generating additional data [26,27]. Based on distribution theory, perturbation analysis, or neural network models, these techniques have proven effective in various experimental scenarios [3,27]. Wang et al. employed the distribution-based StyleGAN2-ADA algorithm to generate artificial algae images, addressing issues of data imbalance and insufficient training images, which significantly enhanced the efficiency of ML in algae classification [28]. Similarly, Dong et al. proposed a novel framework for wastewater prediction in constructed wetlands, where virtual samples generated using the PSOVSG method improved the prediction accuracy for ammonium nitrogen and chemical oxygen demand [3]. Although these methods have been shown to improve ML performance, the generated synthetic data cannot exceed the information in the original samples, and it remains challenging to interpret the relationships between the generated variables and other factors. Mechanistic models of sewers can simulate variations in wastewater quality and elucidate transformation processes between components based on established equations. Their key advantage lies in their ability to calibrate reaction kinetics with only a small amount of data [29,30]. Mechanistic models have been successfully applied to simulate sulfide and methane generation under varying sewage conditions [31,32] and biofilm growth [33,34] and to predict the impact of chemicals on sewer microbial communities, thereby optimizing dosing strategies [35]. Sun et al. used a biofilm model to simulate the effects of different sulfate and soluble chemical oxygen demand concentrations on sulfide and methane yield parameters [31]. Liang et al. developed a sewer process model that effectively simulated the impact of nitrate on sulfide control [33]. Nevertheless, existing studies typically focus on simulations under stable hydraulic conditions, with minimal variability in environmental factors such as temperature (T) and retention time (RT). Real sewers are subject to highly complex and variable conditions. Unlike ML models, mechanistic models cannot incorporate arbitrary variables, making it difficult to account for these missing variables. This limitation renders the model calibration time-consuming and less efficient [25,36].

Notably, hybrid models that combine mechanistic models with ML techniques represent an advanced approach capable of leveraging the strengths of both methods [37,38]. Due to data acquisition limitations within sewers, such methods' application in sewer water quality management remains relatively limited. Liang et al. recently integrated ML algorithms with a sewer process model, using the process model to generate large amounts of synthetic data for ML to predict and control hydrogen sulfide [39]. Nonetheless, considering the spatial variability across different sewers, all mechanistic models must undergo calibration and validation with field data before practical application. When ML models effectively capture such spatial variations, modeling time can be significantly reduced. Furthermore, many studies overlook details such as data collection frequency. Due to varying difficulty levels in obtaining data, collection frequencies often differ. For instance, easily measured parameters such as flow velocity (u), pH, and oxidation-reduction potential (ORP) contrast with resource-intensive variables like sulfate and methane concentrations [40]. Li et al. employed smoothing techniques to transform dense datasets into uniformly distributed time intervals and supplemented sparser datasets via linear interpolation [41]. However, the downsized dataset may affect model training accuracy and impair model performance [26]. The interpolation technique may fail to capture the intricate interactions between variables, potentially distorting the data structure [42]. In such cases, mechanistic models can act as a form of data augmentation by generating additional samples to fill gaps in the dataset, thereby increasing the volume of underrepresented data [43]. ML methods can identify complex, nonlinear relationships that mechanistic models cannot fully capture [44].

This study developed a mechanistically enhanced hybrid (ME-Hybrid) model that predicted sulfide concentrations in sewers by combining the mechanistic model with ML techniques. The ME-Hybrid model was designed to handle datasets with different sampling frequencies and to elevate prediction accuracy. This research is threefold. Initially, the advantages of the ME-Hybrid model over pure ML models and linear interpolation were assessed, and the effects of multiple ML algorithms were investigated within this framework. Subsequently, the study explored the contribution of incorporating the mechanistic model into the hybrid model and provided a detailed interpretive analysis. Furthermore, the study elucidated the impact of data constraints and framework adaptability on the hybrid model. Leveraging the convergence of ML and expert knowledge offers novel insights and methodological advances that have potential to enhance water quality management in sewers.

2. Materials and methods

2.1. Description of the dataset collection

We built five laboratory-level sewer systems to simulate sewage quality transformation processes (Supplementary Material Fig. S1). The devices were cylinders made of Plexiglas and were equipped with water inlets, water outlets, and sampling ports. The mixing disturbance of sewage was carried out by motor-driven agitators. The motor speeds of five systems were set to 40, 80, 120, 160, and 200 rpm, resulting in u-values of 0.06, 0.11, 0.17, 0.22, and 0.28 m s^{-1}, respectively, which were consistent with u-values observed in actual sewer trunks [45]. The sewage was collected weekly from a local septic tank (Shenzhen, China) and mixed with configured sewage as simulated sewage for the experiments. See Text S1(Supplementary Material) for details of sewage quality.

The sewer system initially utilized a continuous flow intake pattern, with the hydraulic residence time set at 12 h. After 100 days of operation to stabilization, a batch intake mode with a 12-h cycle was implemented. Total chemical oxygen demand (TCOD), sulfate, sulfide, and methane were measured inside the reactors at

RTs of 2, 4, 6, 8, 10, and 12 h. They were considered as low-frequency variables. Before analysis, sewage was collected from the reactor outlet and replenished with equal sewage. The concentration of TCOD was determined via the rapid digestion-spectrophotometric method. A quantitative amount of sewage was taken to a brown headspace bottle, and mercuric chloride was added to inhibit microbial activity. After 24 h of standing, gas phase methane was measured via gas chromatography (GC-8890 Agilent, United States) equipped with a flame ionization detector. The gas-phase methane concentration was subsequently converted to methane concentration in the sewage based on the gas-liquid equilibrium [40]. Quantitative sewage was mixed with sodium hydroxide and antioxidant solution in sampling vials and corked without leaving space above the liquid to prevent sulfide volatilization or oxidation. The sulfide concentration was then determined using the methylene blue spectrophotometric method [46]. An appropriate volume of sewage was treated with zinc chloride to form the zinc sulfide precipitate. After filtration through a 0.45 μm cellulose acetate membrane, the sulfate concentration was quantified using ion chromatography (ICS-2000 Dionex, United States). T, ORP, dissolved oxygen (DO), and pH of the sewage were measured using a multi-parameter analyzer (HQ2100 Hach, United States) at hourly intervals. These indicators, along with *u* and sewage RTs, were used as high-frequency variables. Overall, 240 high-frequency data and 120 low-frequency data were collected. The statistical analysis is presented in Table S1 (Supplementary Material).

2.2. Mechanistic model

In this study, the SeweX [47] model was optimized by introducing the release of hydrogen sulfide and methane from the liquid to the gaseous state. We used the improved model as the mechanistic model. The mechanistic model encompasses the critical steps of hydrolysis, fermentation, acidification, sulfation, and methanogenesis and can be used to simulate the contamination transformations in sewers (Supplementary Materials Tables S2–S5). These kinetic parameters were set based on the results of previous studies and mainly followed the default values used by Sun et al. [31]. However, several critical parameters, such as TCOD hydrolysis, sulfide production, and methane production rates, were calibrated using the simulated annealing algorithm, and the detailed procedure is referenced (Supplementary Material Text S2). Table S6 (Supplementary Material) presents the program code used to identify the optimal values of the mechanistic model parameters based on the simulated annealing algorithm. The code was implemented with Python 3.8, utilizing key libraries such as NumPy, SciPy, and Matplotlib.

2.3. Machine learning algorithm

Machine learning algorithms exhibit marked performance differences due to factors such as algorithmic assumptions, data distributions, and dataset sizes [48,49]. We selected eight typical ML algorithms to evaluate the applicability of the proposed framework for multiple algorithms and identify the optimal algorithm (Supplementary Material Fig. S2). These algorithms included linear regression (LR1), logistic regression (LR2), support vector regression (SVR), decision tree (DT), random forest (RF), back propagation neural network (BPNN), recurrent neural network (RNN), and long short-term memory (LSTM).

Specifically, LR1 assumes a linear relationship between sulfide and factors such as DO and pH. It employs the least squares method to minimize the squared differences between the actual and predicted sulfide values. LR2 is the generalization of LR1, which typically uses a logistic likelihood loss function. SVR is a geometrical approach that achieves maximum fitting by determining the optimal hyperplane in a high-dimensional space composed of multiple features, such as TCOD, sulfate, and ORP. This method aims to bring sulfide sample points as close as possible to this hyperplane [50]. DT makes decisions through a tree-like structure, which includes several internal and leaf nodes [51]. The internal nodes represent attribute tests that affect sulfide concentrations. Each node contains a set of samples divided into sub-nodes based on the results of attribute tests. Ultimately, the leaf nodes correspond to the decision results of sulfide concentrations. RF is an ensemble model consisting of numerous DTs, where the final sulfide concentration is made by averaging or voting based on the results of each DT [52]. The BPNN, RNN, and LSTM are three neural network models. The BPNN consists of an input, hidden, and output layer. It is trained using a backpropagation algorithm. This algorithm iteratively updates the parameters by transferring the sulfide prediction error backward to the neurons [53]. The RNN is a neural network specialized in processing sequential information, where recursive connections in the network nodes allow previous water quality information to be remembered. The LSTM is a variant of the RNN and is designed with three gating mechanisms to address the problem of gradient vanishing and gradient explosion faced by traditional RNNs when dealing with long sequence data [41].

2.4. Development of the mechanistically enhanced hybrid model

Three modeling strategies were developed for comparison (Fig. 1). The first strategy (M1) model was to develop the ML model only using high-frequency data. The input variables were *u*, RT, ORP, DO, pH, and T. These indicators were measured or calculated by sensors in the actual sewers. The output was sulfide concentration, which was later changed to methane concentration to validate the framework's applicability (Section 3.5). The purpose was to test the viability of using easily accessible indicators for assessing sewage quality. The second strategy (M2) model was to develop the ML model using the full range of indicators, adding TCOD and sulfate as inputs, compared to the M1 model. The M2 model employed downsampling techniques to align high-frequency data with low-frequency data. The final strategy was the ME-Hybrid model. It initially used the same dataset as the M2 model. The intermediate process then generated virtual data for ML training through the mechanistic model. Data on RT, TCOD, sulfate, sulfide, and methane were employed to calibrate the mechanistic model. The simulated annealing algorithm optimized the kinetic parameters of the mechanistic model to achieve the best model simulation effects (Supplementary Material Text S2). The mechanistic model was used to generate virtual data for TCOD, sulfate, sulfide, and methane at the desired RTs, which were used to boost the sampling frequency for low-frequency data. The ME-Hybrid model combined the mechanistic model and ML, where the input and output indicators for ML were the same as for the M2 model.

The dataset was randomly divided into a training set (70 %) and a test set (30 %). For the ME-Hybrid model, synthetic samples were utilized for training data, while the mechanistic model was calibrated using only the training set to prevent test set leakage. Each variable was normalized to range between 0 and 1 to ensure uniformity in the scale of data points, thereby enhancing the efficiency and stability in processing and analyzing the data. For the development of eight ML algorithms, the grid search algorithm was employed to traverse key hyperparameter combinations and select the optimal model configuration (Supplementary Material Table S8). In this process, ten random splits of the dataset were tested considering the impact of dataset splitting on the model results. We calculated these models' root mean square error (*RMSE*)

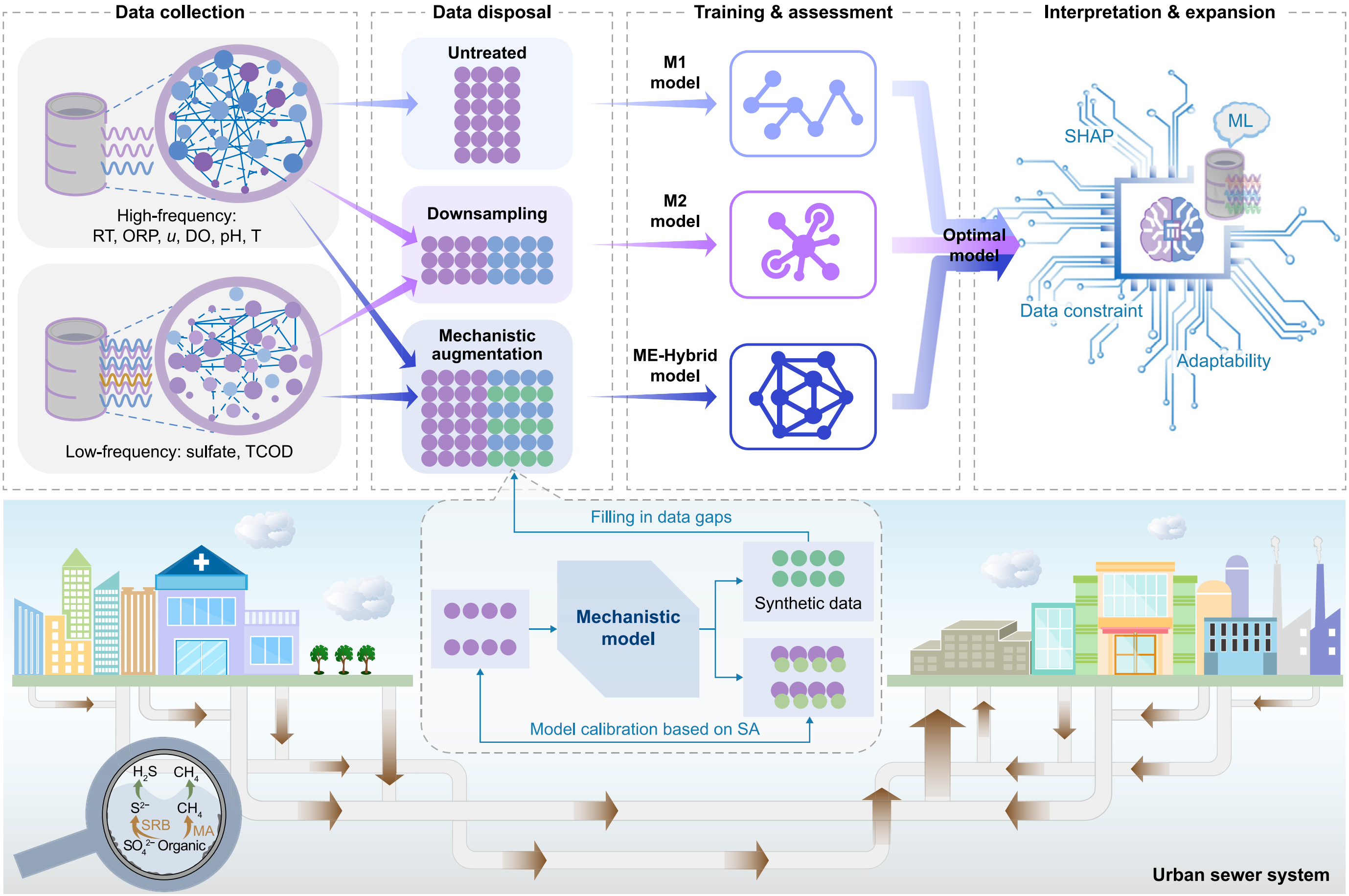

Fig. 1. Schematic diagram of the mechanistically enhanced hybrid (ME-Hybrid) model compared to the first strategy (M1) and the second strategy (M2) models for processing datasets with different sampling frequencies. RT: retention time, ORP: oxidation reduction potential, u: flow velocity, DO: dissolved oxygen, T: temperature, TCOD: total chemical oxygen demand, ML: machine learning, SHAP: Shapley additive explanations, SA: simulated annealing, MA: methanogenic archaea, SRB: sulfate-reducing bacteria.

values. Subsequently, the hyperparameter with the smallest average *RMSE* was selected as the model hyperparameter.

Three regression-based metrics were used to assess model performance: coefficient of determination (R^2), *RMSE*, and mean absolute percentage error (*MAPE*). This study quantitatively revealed the importance of various indicators in predicting sulfide concentrations through Shapley additive explanations (SHAP) analysis [54,55]. This is achieved by calculating the average of the absolute SHAP values for all samples. The swarm summary plot is colored by the level of feature values to illustrate the indicators and their impact on the prediction results. Supervised clustering is the hierarchical clustering of data points based on their SHAP interpretations. As the SHAP values quantify the contribution of each feature to each sample, supervised clustering can help to identify sets of samples that exhibit similar model behaviors, thereby facilitating the analysis and explanation of the predictions in specific scenarios.

3. Results and discussion

3.1. Performance and evaluation of the M1 and M2 models

The efficacy of ML models in predicting sulfide concentrations was competent, and a further improvement in accuracy was still necessary (Fig. 2). The M2 models performed more robustly than the M1 models, achieving a maximum R^2 value of 0.84 compared to 0.76. The *RMSE* and *MAPE* decreased from 0.87 mg S L^{-1} and 0.07 to 0.71 mg S L^{-1} and 0.057, respectively. This comparison suggests that the inclusion of low-frequency variables brought predictive capability. Sulfate and TCOD concentrations were identified as pivotal determinants in sulfide prediction, owing to their role in reducing sulfate to sulfide by sulfate-reducing bacteria. Significant differences between eight models were observed despite utilizing the same training set (Fig. 2). Within the scope of developed M1 and M2 models, SVR, DT, and RF demonstrated higher accuracy than other algorithms. The R^2 of M2-SVR, M2-DT, and M2-RF reached 0.84, 0.75, and 0.76, respectively. M2-SVR may offer advantages due to its simple structure and low complexity of parameter tuning, and it is particularly adept at sulfide prediction tasks through its kernel techniques and optimal hyperplane [48,56]. Conversely, the M2-BPNN achieved an inferior R^2 value of 0.71. The deep learning models (M2-RNN and M2-LSTM) were further degraded, with their R^2 of 0.66 and 0.68, *RMSE* of 1.05 mg S L^{-1} and 1.01 mg S L^{-1}, and *MAPE* of 0.081 and 0.086, respectively (Fig. 2). These models struggle to capture sulfide fluctuations and exhibited larger errors for extreme values (Fig. 3), demonstrating that deep learning models do not invariably outperform SVR, particularly with small datasets [26]. Despite improvements with the addition of low-frequency data, LR1 and LR2 did not achieve R^2 values above 0.50, and their MAPE remained over 0.10. These models cannot track the continuous fluctuations of sulfide concentrations (Fig. 3).

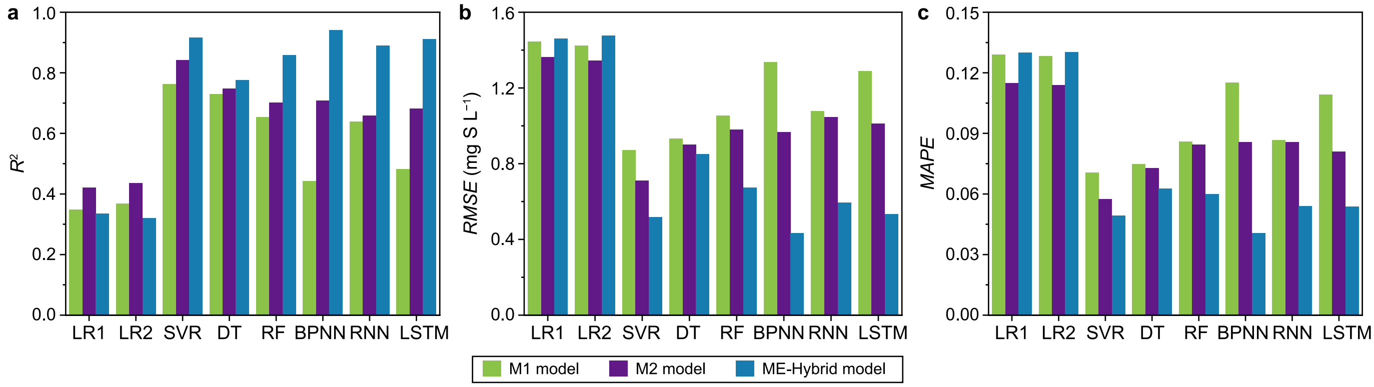

Fig. 2. Performance evaluation of eight machine learning algorithms for sulfide prediction within the framework of the first strategy (M1), the second strategy (M2), and the mechanistically enhanced hybrid (ME-Hybrid) models, respectively. **a**, Coefficient of determination (R^2); **b**, Root mean square error (*RMSE*); **c**, Mean absolute percentage error (*MAPE*). LR1: linear regression, LR2: logistic regression, SVR: support vector regression, DT: decision tree, RF: random forest, BPNN: back propagation neural network, RNN: recurrent neural network, LSTM: and long short-term memory.

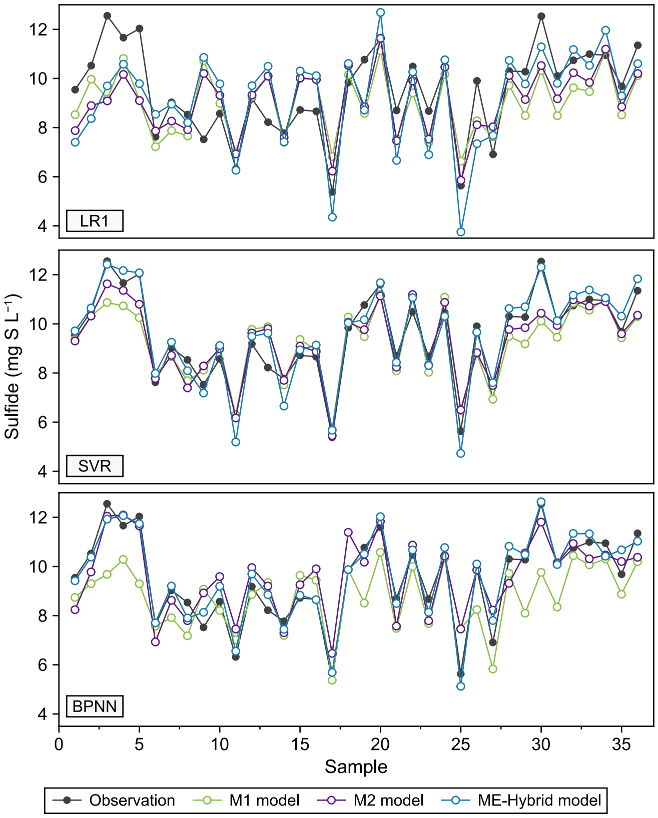

Fig. 3. Comparison of sulfide observations with predictions made by linear regression (LR1), support vector regression (SVR), and back propagation neural network (BPNN) algorithms in the framework of the first strategy (M1), the second strategy (M2), and the mechanistically enhanced hybrid (ME-Hybrid) model. The predictions of the other five algorithms are shown in Fig. S3 (Supplementary Material).

3.2. Performance and insights of the mechanistically enhanced hybrid model

The number of mixed datasets was doubled by generating synthetic samples using the mechanistic augmentation technique. The hybrid dataset was then fed into the ML model. Incorporating the mechanistic model improved the accuracy of most ME-Hybrid models (Fig. 2). Deep learning algorithms leveraged their powerful network structure and learning capabilities to gain excellent model performance, with R^2 values surpassing 0.85 and MAPE confined to within 5.5 %. The hybrid model based on the BPNN performed the best, with an increase in R^2 by 11.8 % to 0.94 compared to the M2-SVR model. Additionally, *RMSE* and *MAPE* were reduced to 0.43 mg S L^{-1} and 0.041, respectively (Fig. 2), indicating that ME-Hybrid-BPNN was adept at predicting local peaks and troughs in sulfide concentrations (Fig. 3). The RNN and LSTM models were also significantly improved due to the inclusion of synthetic data. Their R^2 increased from below 0.7 to around 0.9, and their MAPE reduced by more than 30 %. However, the hybrid framework was not applicable to all ML algorithms. The performance of the LR1 and LR2 models declined after integrating the mechanistic model. This decline was due to two main factors: The inclusion of synthetic samples increased the complexity of the dataset, and the assumption of a linear relationship between sulfide and the input variables was not realistic.

Fig. 4a demonstrates the absolute value of the relative error of the BPNN algorithm, and it can be found that the M2-BPNN model had greater errors for low sulfide samples, even more than 30 %, in both the training and test sets. The ME-Hybrid model, on the other hand, effectively overcame this shortcoming with all relative errors of less than 15 %. After adding the generated samples to the original dataset, the percentage of samples with low sulfide samples in the hybrid dataset increased (Supplementary Material Fig. S4), which prompted the BPNN model to optimize the prediction pattern of low sulfide samples. Furthermore, this study verified the feasibility of linear interpolation (Fig. 4b). The performance of the interpolation-based models was comparable to that of the M2 model. The most effective algorithm was SVR, which achieved sulfide prediction with an R^2 of 0.85, *RMSE* of 0.69 mg S L^{-1}, and *MAPE* of 0.063. The BPNN and LSTM models with interpolation processing did not show significant performance improvements compared to the ME-Hybrid models. For the BPNN algorithm using interpolation, the R^2 was 0.71, *RMSE* was 0.96 mg S L^{-1}, and *MAPE* was 0.094. In contrast, the ME-Hybrid-BPNN model increased the R^2 by 32.4 %, and *RMSE* and *MAPE* decreased by 54.9 % and 52.2 %, respectively. The R^2 of the RNN model using interpolation achieved 0.81, but this was still lower than the 0.89 attained by the ME-Hybrid-RNN model. The box plot analysis illustrated in Fig. 4c reveals that both real and synthetic sulfide datasets exhibit similar distribution characteristics, but the synthetic data have a wider range of distributions. For example, the synthetic samples contained low sulfide concentrations that were not observed. The

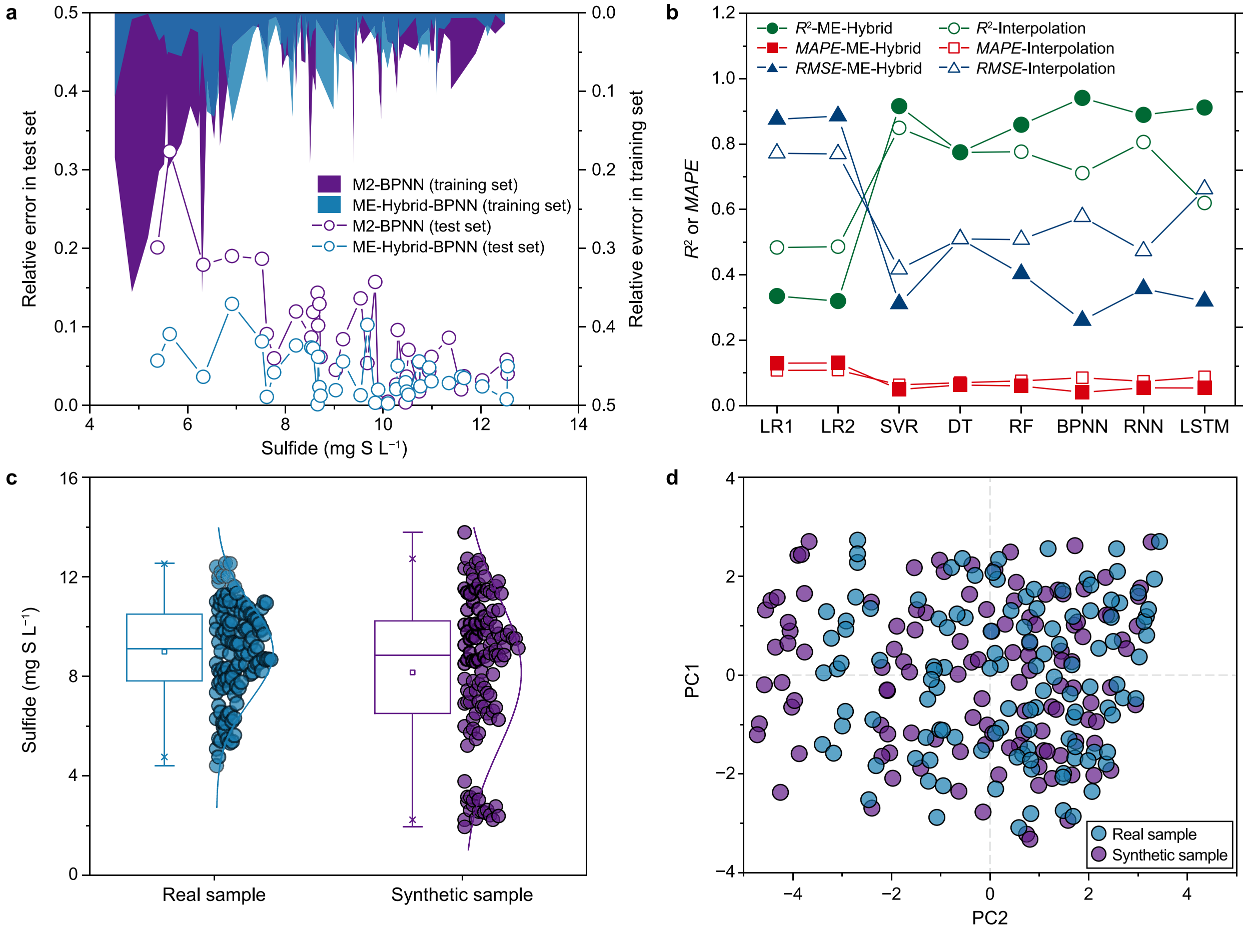

Fig. 4. Explanation of the mechanistically enhanced hybrid (ME-Hybrid) model performance enhancement. **a**, Absolute values of the relative error in predicting sulfide concentrations by back propagation neural network (BPNN) algorithm under the second strategy (M2) and ME-Hybrid models. **b**, Performance comparison for interpolation-based models and ME-Hybrid models to predict sulfide concentrations. **c**, Box plots and distributions of the real samples and synthetic samples. **d**, Distribution of the real and synthetic samples on PC1 and PC2. R^2: coefficient of determination, *RMSE*: root mean square error, *MAPE*: mean absolute percentage error, PC: principal component.

prediction for samples with low sulfide concentrations was improved due to the hybrid dataset (Supplementary Material Fig. S5). Moreover, the principal component (PC) analysis indicated that the PC distributions of the two datasets fail to distinguish significantly (Fig. 4d). The first two PCs (PC1 and PC2) are linear combinations of the original variables, which highlights that the synthetic samples preserve the relationships between the real sample variables and mimic the real-world observations.

3.3. Interpretive analysis of the mechanistically enhanced hybrid model

The ME-Hybrid model, complemented by interpretable ML techniques, can explain predictions. This increases the potential of the hybrid model as a decision-making tool [57]. The impact and contribution of various variables to the ME-Hybrid model was elucidated by integrating correlation analysis, feature importance assessment, summary bee-swarm plot, and supervised clustering heatmap. The correlation heatmap revealed that RT, sulfate, pH, and TCOD were the four most impactful variables on sulfide concentrations (Fig. 5a). A positive correlation was registered between RT and sulfide (0.74), suggesting a temporal dimension to sulfide production. Conversely, sulfate, pH, and TCOD were inversely related to sulfide concentrations (−0.70, −0.56, and −0.50, respectively), indicating the conversion dynamics of these components in sewers.

The SHAP analysis revealed the intricate dynamics of variables within the ME-Hybrid model. The order of factors affecting sulfide was RT > ORP > Sulfate > TCOD > *u* > DO > pH > T (Fig. 5b). The mean SHAP value of 0.137 for RT was significantly higher than the other factors. The increase in sulfide and methane is mainly due to increased hydraulic residence time, which has been confirmed [58,59]. Along the sewer line, there was a gradual transition from fermentation reactions to methanogenesis and sulfate reduction reactions, resulting in higher sulfide concentrations at the end of sewers [2]. ORP was the second most significant factor affecting water quality. Its ease of measurement makes it a critical indicator for developing soft measurement models for water quality. Deng et al. found that lower ORP values were more favorable for the production of sulfides [60], but this was not entirely consistent with the results shown in the swarm plot (Fig. 5c). They observed ORP values ranging from −100 to −270 mV. Considering the applicable ORP ranges for sulfate reduction (−50 to −250 mV) and methane production (−175 to −400 mV) [61], the ORP range observed in this study (−259 to −348 mV) suggested that as ORP increases, there might be a shift from methane production to sulfate reduction process. The importance of sulfate and TCOD was reasserted, confirming the superiority of the M2 model compared to the M1 model in predicting sulfide concentrations. Typical sewage usually contains limited sulfate and sufficient organic substrate, highlighting

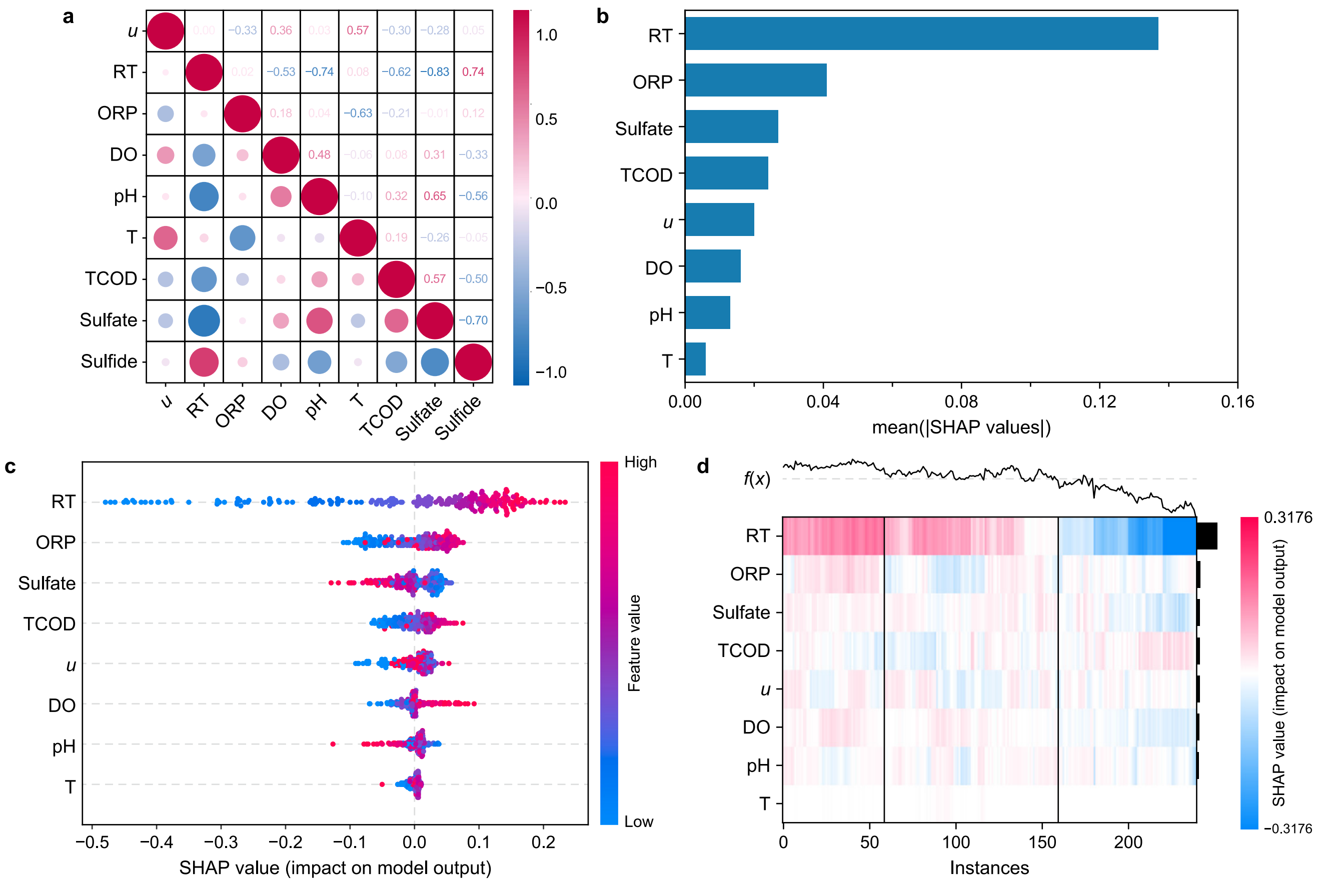

Fig. 5. Interpretive analysis of the mechanistic-enhanced hybrid model. **a**, Heatmap of Pearson correlation coefficients between model variables. **b**, Ranking of the importance of input variables for sulfide prediction. **c**, Global Shapley additive explanations (SHAP) interpretation by the summary bee-swarm plot. **d**, Supervised clustering heatmap for samples. *u*: flow velocity, RT: retention time, ORP: oxidation reduction potential, DO: dissolved oxygen, T: temperature, TCOD: total chemical oxygen demand.

sulfate's greater role than TCOD in sulfide production [62]. Increased *u* positively influenced the production of sulfide. On the one hand, sulfate-reducing bacteria predominantly reside in the surface layer of the biofilm [63]. Increasing the *u* reduces the thickness of the interfacial diffusion boundary layer, thereby enhancing the transport of sulfate, which is a limiting factor for sulfide production rates [45,64]. On the other hand, this study observed that the *u* has different patterns of influence on sulfate reduction and methanogenic activity (Supplementary Material Fig. S10). Methanogenic bacteria are mainly present in the inner layer of the biofilm [63]. The rough biofilm structure and the low oxygen transfer efficiency may be more favorable for the survival and competition of methanogenic bacteria at a lower *u*.

Fig. S7 (Supplementary Material) illustrates the feature importance rankings and the corresponding SHAP summary plots for the M1-SVR and M2-SVR models. Both models successfully identified the importance of RT and ORP. Nonetheless, in the M1 model, sulfate and TCOD were not included as inputs. The absence of important metrics caused the M1 model to lower the predictive performance for sulfide. In contrast, the M2 model failed to adequately capture the importance of TCOD and sulfate. The importance shares of sulfate and TCOD in the ME-Hybrid model were 9.7 % and 8.6 %, respectively, while their importance percentages were 6.9 % and 7.1 % in the M2-SVR model, respectively. The M2 model overemphasized the significance of RT and neglected key water quality indicators. The inclusion of synthetic data promoted the importance of low-frequency metrics. As shown in Fig. S7d (Supplementary Material), the distributions of all points on both sides of the centerline for sulfate and TCOD were insignificant or even disordered. In other words, after data augmentation using the mechanistic model, the mapping relationships of sulfate and TCOD for sulfide were identified with clearer distributions on both sides of the centerline. Supervised clustering produced an array with the same dimensions as the original data by transforming it into SHAP values. The direction and strength of the feature's influence on sulfide concentration were indicated by different colors (Fig. 5d). The sulfide output results ($f(x)$) showed that the left segment clustering was a grouping of high sulfide concentration samples, with RT, ORP, and sulfate mainly having positive effects. As the SHAP values of ORP and sulfate turned negative, the modeled results in the middle section were pulled down. Lower RT levels in the latter section were the most important reason for limiting sulfide concentrations.

3.4. *Assessment of the mechanistically enhanced hybrid model under data constraints*

The scarcity of datasets is the primary challenge for ML model development, which directly determines the effectiveness of the developed model [12]. Faced with this limitation, the ME-Hybrid model leveraged the mechanistic augmentation to expand the training dataset. This study kept the synthetic sample size consistent with the original dataset. It revealed the potential of the ME-Hybrid model under data constraints by adjusting the ratio of the training and test sets.

Despite the limitation of the dataset size, the ME-Hybrid model demonstrated strong adaptability and accuracy. As the proportion of the training set decreased from 70 % to 30 %, the ME-Hybrid model's R^2 decreased from 0.94 to 0.49, *RMSE* increased from 0.43 mg S L^{-1} to 1.32 mg S L^{-1}, and *MAPE* increased from 0.041 to more than 0.10 (Fig. 6). It is worth noting that to facilitate adequate learning and prevent underfitting, the allocation of the training set typically spans between 70 % and 80 %. Nevertheless, this study found that even with a training set proportion of 50 %, the ME-Hybrid model still achieved commendable performance, with the R^2, *RMSE*, and *MAPE* values of 0.76, 0.89 mg S L^{-1}, and 0.083, respectively (Fig. 6). This was attributed to the mechanistic model generating synthetic samples for training the ML model. The real sample volume constituted only one-third of the mixed dataset used for ML training. The mechanistic model exhibited a similar trend, with R^2 values plunging from 0.87 to 0.30 and *RMSE* and *MAPE* increasing from 0.64 mg S L^{-1} and 0.058 to 1.54 mg S L^{-1} and 0.148, respectively. When the training set proportion fell below 50 %, the R^2 of the mechanistic model dropped below 0.50, and the MAPE exceeded 0.10. With the reduction in training sample size, the mechanistic model may not be calibrated. The low-quality synthetic data negatively affected the ME-Hybrid model. In summary, the results underscore that the ME-Hybrid model is well-suited for data constraints and can maintain acceptable accuracy on small-scale datasets. Nonetheless, the ME-Hybrid model may be compromised when the amount of data is extremely limited and insufficient to support the mechanistic model.

3.5. Applicability of the mechanistically enhanced hybrid model

The ME-Hybrid model exhibited robust performance for sulfide concentrations. Nonetheless, the potential for its application to alternative targets remains to be explored. Methane is a significant contributor to climate change [1], but its emissions from sewers have been underestimated in prior assessments [10,65]. There is a risk of explosion when the methane concentration in the air exceeds 5 %. Therefore, the effective assessment of methane concentrations is a non-negligible part of sewer management [65].

Eight ML algorithms were employed in the M1, M2, and ME-Hybrid models, respectively (Supplementary Material Fig. S9). The ME-Hybrid model predicted methane more accurately compared to the M1 and M2 models. The optimal ML algorithms under the three frameworks were M1-SVR, M2-SVR, and ME-Hybrid-BPNN. The predictive performance metrics R^2, *RMSE*, and *MAPE* of M1-SVR for methane were 0.82, 0.75 mg L^{-1}, and 0.354, respectively. These values of M2-SVR were 0.84, 0.70 mg L^{-1}, and 0.331, respectively. Compared to the M2-SVR model, the performance of the ME-Hybrid-BPNN model was improved significantly, with a 15.8 % increase in R^2 (0.94) and reductions of 44.9 % in *RMSE* (0.42 mg L^{-1}) and 43.4 % in *MAPE* (0.193) (Fig. 7b). Notably, in the test set, the highest methane concentration is 16 times higher than the lowest concentration (Fig. 7a), suggesting that the ME-Hybrid model was effective in capturing methane concentration fluctuations and demonstrated robust resilience against disturbances. Further analysis revealed no significant difference between the performance of M1 and M2 models in methane prediction (Supplementary Material Fig. S9). SHAP analysis identified RT, *u*, DO, and pH as the four highly correlated variables with the methane concentrations. Their importance ranking further emphasized the status of these high-frequency variables as the main factors (Supplementary Material Fig. S10). This finding suggests that methane production is more sensitive to indicators such

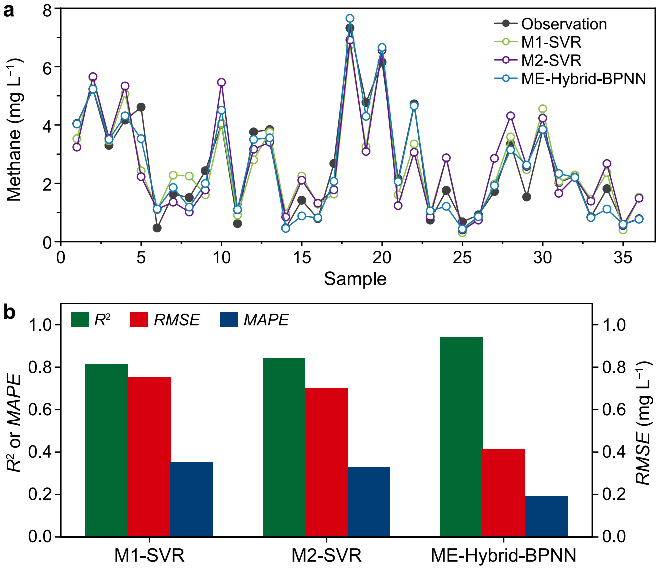

Fig. 7. Generalization capability of the mechanistically enhanced hybrid (ME-Hybrid) model. **a**, Comparison of the methane observations with predictions made by the M1-SVR model, M2-SVR model, and ME-Hybrid-BPNN model. **b**, Performance evaluation for methane prediction. SVR: support vector regression, M1-SVR: the first strategy model based on SVR, M2-SVR: the second strategy model based on SVR, ME-Hybrid-BPNN: the ME-Hybrid based on back propagation neural network, R^2: coefficient of determination, *RMSE*: root mean square error, *MAPE*: mean absolute percentage error.

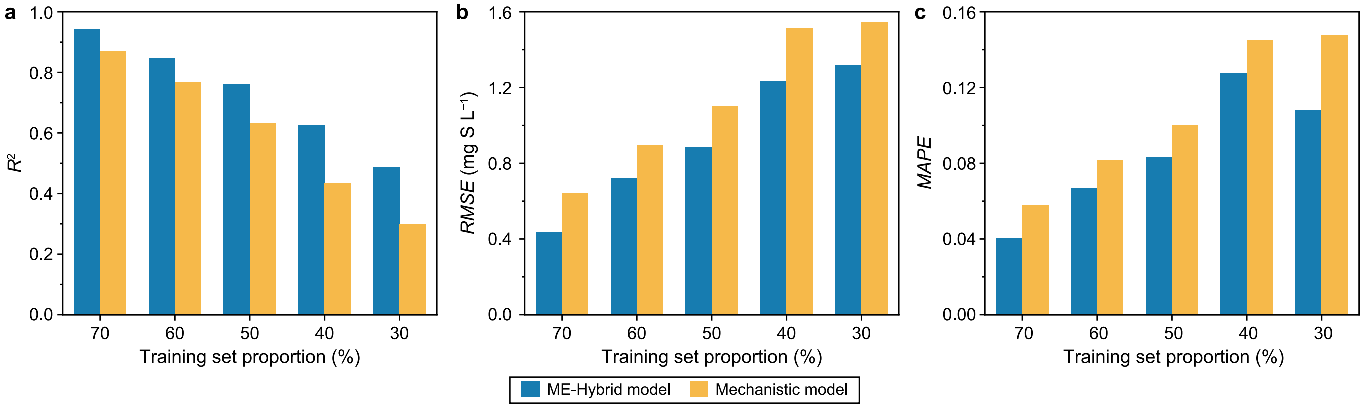

Fig. 6. Coefficient of determination (R^2, **a**), root mean square error (*RMSE*, **b**), and mean absolute percentage error (*MAPE*, **c**) for sulfide prediction by the mechanistically enhanced hybrid (ME-Hybrid) model and mechanistic model under different proportions of training set.

as *u*, pH, and DO, reflecting the stringent requirements of methanogenic bacteria for environmental conditions. In summary, the ME-Hybrid model accurately assessed methane concentrations and provided interpretable insights into the determinants, demonstrating this proposed framework's flexibility and adaptability.

3.6. Environmental implications and outlooks

The imbalance in variable availability hampers the effective integration of datasets for ML modeling, thereby impeding efficient sewer management [13]. Conventional interpolation methods are inadequate for time series data with missing values exceeding a 25 % threshold [42]. To address this issue, this study developed a mechanistic model to augment low-frequency sampling data, utilizing ML to capture the nonlinear relationships. The mechanistic model successfully simulated matter dynamics across time scales without synchronizing metrics at a single point in time [66]. Mechanistic augmentation techniques generated complete and independent samples and addressed gaps in the datasets with different sampling frequencies [43,67]. This study demonstrated the potential of mechanistic augmentation techniques, which exhibited greater credibility and interpretability than statistical methods or black-box models [68,69], as the mechanistic model maintained the nonlinear relationships guided by knowledge [70]. Notably, the ME-Hybrid model demonstrated high data utilization efficiency and achieved satisfactory performance when the share of the training set was only 50 % (Fig. 6). This highlighted the feasibility of improving model performance through data resource allocation in data-constrained situations [71]. The advantage of hybrid model is attributed to combining two modeling paradigms. The mechanistic augmentation increased the diversity of information in the hybrid dataset by generating samples that were not observed, such as sewage-characterizing changes in a early period [28]. ML models can learn patterns from synthetic samples while accounting for the effects of variables not present in the mechanistic model [56].

Applying the proposed ME-Hybrid model in real-world sewers is anticipated to reduce the need for spatial sample collection and enhance the accuracy of water quality predictions. The RT simulated in this study effectively reflects variations in sewer length. In long sewage segments, intensive sampling typically necessitates substantial human and material resources. Traditional interpolation methods are unsuitable for acquiring mid-segment water quality data. The mechanistic model in this study can increase sewer sampling points and the water quality data required for these points, further enabling ML to effectively train and predict water quality changes along the sewer. Therefore, this model can potentially be an important tool for managing or mitigating sewer corrosion and greenhouse gas emissions. Although the ME-Hybrid model has successfully assessed sulfide and methane concentrations in sewers, its extension to other areas of water environment management, such as wastewater treatment scenarios, is urgently desired and expected. In these scenarios, fundamental principles such as carbon conversion and nitrogen conversion could be available in existing mechanistic models [67,72], and the sampling frequency of the variables is usually not uniform [24,41]. This hybrid framework is expected to address frequency discrepancies in datasets through its mechanistic component and enhance the prediction of wastewater treatment targets using ML algorithms, which may improve the efficiency of raw data utilization and help manage water treatment processes.

Admittedly, while the findings presented in this study are significant, it is important to acknowledge its limitations. Although the ME-Hybrid model is theoretically expected to improve the performance of ML models, the experimental results indicated only modest gains. Specifically, for the prediction of sulfide and methane, the R^2 values increased by 11.8 % and 15.8 %. This outcome was primarily attributed to the dataset's limitations. The sampling frequency of high-frequency data was only twice that of low-frequency data, and the increase in data collection frequency from 2 h to 1 h through the mechanistic model was insufficient to significantly improve prediction performance. Therefore, when applying the ME-Hybrid model to real-world sewers, methods such as Monte Carlo simulation or stochastic perturbations could be employed to fully leverage the data generation capabilities of mechanistic models. Alternatively, increasing the frequency of raw data collection may provide the ML model with richer dynamic information. These improvements will yield greater performance gains in more complex dynamic systems.

4. Conclusion

This study presented the ME-Hybrid model by integrating the mechanistic and ML models and evaluated its effectiveness using sulfide and methane concentrations in sewers as representatives. The mechanistic component could generate samples to fill gaps in the original data, thereby harmonizing datasets with irregular sampling frequencies. This potential surpassed interpolation and has been underappreciated in previous research. It was demonstrated that mechanistic samples can effectively substitute real samples. The accuracy of the ML model was significantly improved by incorporating mechanistic samples into the ML model training set. This reduced the dependence of model training on real sample size. This work offered insights into enhancing model development in data-constrained situations, highlighting the promising integration of mechanistic knowledge with ML approaches to analyze environmental systems. Future research should evaluate the applicability of the ME-hybrid model across diverse scenarios.

CRediT authorship contribution statement

Jia-Qiang Lv: Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Wan-Xin Yin:** Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jia-Min Xu:** Project administration, Investigation, Data curation. **Hao-Yi Cheng:** Validation, Methodology, Investigation. **Zhi-Ling Li:** Methodology, Investigation, Formal analysis, Data curation. **Ji-Xian Yang:** Validation, Investigation, Formal analysis. **Ai-Jie Wang:** Resources, Methodology, Investigation, Data curation. **Hong-Cheng Wang:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition, Data curation, Conceptualization.

Declaration of interests

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.

Acknowledgement

We gratefully acknowledge the financial support by the National Natural Science Foundation of China (No.52321005, No. 52293445) and Shenzhen Science and Technology Program (KCXFZ20211020163404007, Grant No. KQTD20190929172630447).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2024.100512.

References

[1] X. Cen, H. Duan, Z. Hu, et al., Multifaceted benefits of magnesium hydroxide dosing in sewer systems: impacts on downstream wastewater treatment processes, Water Res. (2023) 247.

[2] P. Jin, X. Shi, G. Sun, et al., Co-variation between distribution of microbial communities and biological metabolization of organics in urban sewer systems, Environ. Sci. Technol. 52 (3) (2018) 1270–1279.

[3] Q. Dong, S. Bai, Z. Wang, et al., Virtual sample generation empowers machine learning-based effluent prediction in constructed wetlands, J. Environ. Manag. 346 (2023).

[4] I. Pikaar, K.R. Sharma, S. Hu, et al., Reducing sewer corrosion through integrated urban water management, Science 345 (6198) (2014) 812–814.

[5] D. Huang, X. Liu, S. Jiang, et al., Current state and future perspectives of sewer networks in urban China, Front. Environ. Sci. Eng. 12 (3) (2018).

[6] Y. Liu, Z. Zuo, H. Li, et al., *In-situ* advanced oxidation of sediment iron for sulfide control in sewers, Water Res. 240 (2023).

[7] L. Zhang, Y.-Y. Qiu, K.R. Sharma, et al., Hydrogen sulfide control in sewer systems: a critical review of recent progress, Water Res. (2023) 240.

[8] J.-M. Xu, Y.-L. Sun, X.-D. Yao, et al., Highly efficient coremoval of nitrate and phosphate driven by a sulfur-siderite composite reactive filler toward secondary effluent polishing, Environ. Sci. Technol. 57 (43) (2023) 16522–16531.

[9] J. Chen, H. Wang, W. Yin, et al., Deciphering carbon emissions in urban sewer networks: bridging urban sewer networks with city-wide environmental dynamics, Water Res. 256 (2024).

[10] P. Jin, Y. Gu, X. Shi, et al., Non-negligible greenhouse gases from urban sewer system, Biotechnol. Biofuels 12 (2019).

[11] G. Fu, Y. Jin, S. Sun, et al., The role of deep learning in urban water management: a critical review, Water Res. 223 (2022).

[12] S. Zhong, K. Zhang, M. Bagheri, et al., Machine learning: new ideas and tools in environmental science and engineering, Environ. Sci. Technol. 55 (19) (2021) 12741–12754.

[13] X. Liu, D. Lu, A. Zhang, et al., Data-driven machine learning in environmental pollution: gains and problems, Environ. Sci. Technol. 56 (4) (2022) 2124–2133.

[14] B. Xu, H. Yu, Z. Shi, et al., Knowledge-guided machine learning reveals pivotal drivers for gas-to-particle conversion of atmospheric nitrate, Environ. Sci. Ecotechnol. 19 (2024).

[15] M. Goodarzi, S. Vazirian, A machine learning approach for predicting and localizing the failure and damage point in sewer networks due to pipe properties, J. Water Health. 22 (3) (2024) 487–509.

[16] S. Ma, N. Elshaboury, E. Ali, et al., Proactive exfiltration severity management in sewer networks: a hyperparameter optimization for two-tiered machine learning prediction, Tunn. Undergr. Space Technol. (2024) 144.

[17] U. Iqbal, M.Z. Bin Riaz, J. Barthelemy, et al., Artificial Intelligence of Things (AIoT)-oriented framework for blockage assessment at cross-drainage hydraulic structures, Aust. J. Water Resour. (2023) 1–11.

[18] E. Gul, M.J.S. Safari, O.F. Dursun, et al., Ensemble and optimized hybrid algorithms through Runge Kutta optimizer for sewer sediment transport modeling using a data pre-processing approach, Int. J. Sediment Res. 38 (6) (2023) 847–858.

[19] Z. Yin, Y. Saadati, M.H. Amini, et al., Forecasting and optimization for minimizing combined sewer overflows using Machine learning frameworks and its inversion techniques, J. Hydrol. (2024) 628.

[20] N.S. Simmons, J.J. Ducoste, Fat, oil, and grease sewer waste management system: a modeling platform for simulating the formation of FOG deposits in sewer networks, J. Environ. Eng. 150 (4) (2024).

[21] M. Zounemat-Kermani, A. Aldallal, Predicting microbiologically influenced concrete corrosion in self-cleansing sewers using meta-learning techniques, Corrosion 80 (4) (2024) 338–348.

[22] F. Hou, S. Liu, W.-X. Yin, et al., Machine learning for high-precision simulation of dissolved organic matter in sewer: overcoming data restrictions with generative adversarial networks, Sci. Total Environ. (2024) 947.

[23] F.C.A. Mendes, F. Pierre, C. Valentin, et al., Modelling an urban wastewater system via a space-time multivariate calibration to understand and improve water bodies quality, Water Sci. Technol. 90 (5) (2024) 1433–1450.

[24] M. Alvi, T. French, R. Cardell-Oliver, et al., Enhanced deep predictive modelling of wastewater plants with limited data, IEEE Trans. Ind. Inf. 20 (2) (2023) 1920–1930.

[25] S. Huang, J. Xia, Y. Wang, et al., Water quality prediction based on sparse dataset using enhanced machine learning, Environ. Sci. Ecotechnol. 20 (2024).

[26] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inf. Assoc. 25 (10) (2018) 1419–1428.

[27] S. Wei, Z. Chen, S.K. Arumugasamy, et al., Data augmentation and machine learning techniques for control strategy development in bio-polymerization process, Environ. Sci. Ecotechnol. 11 (2022).

[28] W.H. Chan, B.S.B. Fung, D.H.K. Tsang, et al., A freshwater algae classification system based on machine learning with StyleGAN2-ADA augmentation for limited and imbalanced datasets, Water Res. 243 (2023).

[29] H.-W. Lin, Y. Lu, R. Ganigue, et al., Simultaneous use of caustic and oxygen for efficient sulfide control in sewers, Sci. Total Environ. 601 (2017) 776–783.

[30] J. Vollertsen, T. Hvitved-Jacobsen, Z. Ujang, et al., Integrated design of sewers and wastewater treatment plants, Water Sci. Technol. 46 (9) (2002) 11–20.

[31] J. Sun, B.-J. Ni, K.R. Sharma, et al., Modelling the long-term effect of wastewater compositions on maximum sulfide and methane production rates of sewer biofilm, Water Res. 129 (2018) 58–65.

[32] F. Zan, Z. Liang, F. Jiang, et al., Effects of food waste addition on biofilm formation and sulfide production in a gravity sewer, Water Res. 157 (2019) 74–82.

[33] Z. Liang, D. Wu, G. Li, et al., Experimental and modeling investigations on the unexpected hydrogen sulfide rebound in a sewer receiving nitrate addition: mechanism and solution, J. Environ. Sci. 125 (2023) 630–640.

[34] Z.-S. Liang, L. Zhang, D. Wu, et al., 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 (2019) 94–103.

[35] S.A. Naudin, A.A. Ferran, P.H. Imazaki, et al., Development of an *in vitro* biofilm model for the study of the impact of fluoroquinolones on sewer biofilm microbiota, Front. Microbiol. 15 (2024).

[36] M.C. Nicoletti, L.C. Jain, R.C. Giordano, Computational intelligence techniques as tools for bioprocess modelling, optimization, supervision and control, in: M.D.C. Nicoletti, L.C. Jain (Eds.), Computational Intelligence Techniques for Bioprocess Modelling, Supervision and Control, 2009, pp. 1–23.

[37] L. Faure, B. Mollet, W. Liebermeister, et al., A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models, Nat. Commun. 14 (1) (2023).

[38] H. Duan, J. Li, Z. Yuan, Making waves: knowledge and data fusion in urban water modelling, Water Res. X 24 (2024).

[39] Z. Liang, W. Xie, H. Li, et al., Integrating machine learning algorithm with sewer process model to realize swift prediction and real-time control of H_2S pollution in sewer systems, Water Res. X 23 (2024).

[40] A. Guisasola, D. de Haas, J. Keller, et al., Methane formation in sewer systems, Water Res. 42 (6–7) (2008) 1421–1430.

[41] K. Li, H. Duan, L. Liu, et al., An integrated first principal and deep learning approach for modeling nitrous oxide emissions from wastewater treatment plants, Environ. Sci. Technol. 56 (4) (2022) 2816–2826.

[42] C. Betancourt, C.W.Y. Li, F. Kleinert, et al., Graph machine learning for improved imputation of missing tropospheric ozone data, Environ. Sci. Technol. 57 (46) (2023) 18246–18258.

[43] W. Quaghebeur, E. Torfs, B. De Baets, et al., Hybrid differential equations: integrating mechanistic and data-driven techniques for modelling of water systems, Water Res. 213 (2022).

[44] S. Yao, C. Zhang, H. Yuan, Emerging investigator series: modeling of wastewater treatment bioprocesses: current development and future opportunities, Environ. Sci. Water Res. Technol. 8 (2) (2022) 208–225.

[45] Z. Zuo, D. Ren, L. Qiao, et al., Rapid dynamic quantification of sulfide generation flux in spatially heterogeneous sediments of gravity sewers, Water Res. 203 (2021).

[46] Ministry of Environmental Protection, Monitoring and Analytical Methods of Water and Wastewater, 4th ed. China Environmental Science Press, Beijing, 2016.

[47] K. Sharma, R. Ganigue, Z. Yuan, pH dynamics in sewers and its modeling, Water Res. 47 (16) (2013) 6086–6096.

[48] J. Lv, L. Du, H. Lin, et al., Enhancing effluent quality prediction in wastewater treatment plants through the integration of factor analysis and machine learning, Bioresour. Technol. (2024) 393.

[49] Y.-Q. Wang, H.-C. Wang, Y.-P. Song, et al., Machine learning framework for intelligent aeration control in wastewater treatment plants: automatic feature engineering based on variation sliding layer, Water Res. (2023) 246.

[50] V. Nourani, P. Asghari, E. Sharghi, Artificial intelligence based ensemble modeling of wastewater treatment plant using jittered data, J. Clean. Prod. (2021) 291.

[51] R. Huang, C. Ma, J. Ma, et al., Machine learning in natural and engineered water systems, Water Res. 205 (2021).

[52] Y. Tian, X. Yang, N. Chen, et al., Data-driven interpretable analysis for polysaccharide yield prediction, Environ. Sci. Ecotechnol. 19 (2024).

[53] B. Yang, Z. Xiao, Q. Meng, et al., Deep learning-based prediction of effluent quality of a constructed wetland, Environ. Sci. Ecotechnol. 13 (2023).

[54] Z. Li, Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost, Comput. Environ. Urban Syst. 96 (2022).

[55] J. Park, W.H. Lee, K.T. Kim, et al., Interpretation of ensemble learning to predict water quality using explainable artificial intelligence, Sci. Total Environ. (2022) 832.

[56] Z. Wang, J. Feng, M. Liang, et al., Prediction model and application of machine learning for supersaturated total dissolved gas generation in high dam discharge, Water Res. 220 (2022).

[57] Y. Cha, J. Shin, B. Go, et al., An interpretable machine learning method for supporting ecosystem management: application to species distribution models of freshwater macroinvertebrates, J. Environ. Manag. (2021) 291.

[58] A. Guisasola, K.R. Sharma, J. Keller, et al., Development of a model for assessing methane formation in rising main sewers, Water Res. 43 (11) (2009) 2874–2884.

[59] J. Sun, S. Hu, K.R. Sharma, et al., Impact of reduced water consumption on sulfide and methane production in rising main sewers, J. Environ. Manag. 154 (2015) 307–315.

[60] Q. Deng, S. Li, M. Yao, et al., Study on the factors of hydrogen sulfide production from lignite bacterial sulfate reduction based on response surface method, Sci. Rep. 13 (1) (2023).

[61] Y. Yuan, G. Zhang, H. Fang, et al., Microbial spatial distribution and corrosion evaluation in urban sewer systems with different service lives, Eng. Fail. Anal. 139 (2022).

[62] Y. Liu, B.-J. Ni, R. Ganigue, et al., Sulfide and methane production in sewer sediments, Water Res. 70 (2015) 350–359.

[63] J. Sun, S. Hu, K.R. Sharma, et al., Stratified microbial structure and activity in sulfide- and methane-producing anaerobic sewer biofilms, Appl. Environ. Microbiol. 80 (22) (2014) 7042–7052.

[64] Z. Zuo, Z. Sun, Y. Zhang, et al., In situ exploration of the sulfidogenic process at the water-sediment interface in sewers: mechanism and implications, Acs Es&T Engineering 1 (3) (2021) 415–423.

[65] X. Yan, J. Sun, A. Kenjiahan, et al., Rapid and strong biocidal effect of ferrate on sulfidogenic and methanogenic sewer biofilms, Water Res. (2020) 169.

[66] S. Freguia, K. Sharma, O. Benichou, et al., Sustainable engineering of sewers and sewage treatment plants for scenarios with urine diversion, J. Hazard Mater. (2021) 415.

[67] M. Alvi, D. Batstone, C.K. Mbamba, et al., Deep learning in wastewater treatment: a critical review, Water Res. (2023) 245.

[68] Q. Wei, X. Li, M. Song, Reconstruction of irregular missing seismic data using conditional generative adversarial networks, Geophysics 86 (6) (2021) V471–V488.

[69] R.-Z. Xu, J.-S. Cao, Y. Wu, et al., An integrated approach based on virtual data augmentation and deep neural networks modeling for VFA production prediction in anaerobic fermentation process, Water Res. 184 (2020).

[70] J.S. Anderson, T.J. McAvoy, O.J. Hao, Use of hybrid models in wastewater systems, Ind. Eng. Chem. Res. 39 (6) (2000) 1694–1704.

[71] L. Peng, H. Wu, M. Gao, et al., TLT: recurrent fine-tuning transfer learning for water quality long-term prediction, Water Res. 225 (2022).

[72] D.S. Lee, P.A. Vanrolleghem, J.M. Park, Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant, J. Biotechnol. 115 (3) (2005) 317–328.