

Time series imputation in wastewater treatment plant monitoring: a multi-plant evaluation under heterogeneous dataset structures and missingness patterns

Hyeongju Park^{a,b}, Seongbong Heo^a, Sangchul Lee^{c,**}, Young Mo Kim^{a,b,*}

^a Department of Civil and Environmental Engineering, Hanyang University, Seoul, Seongdong-gu, 04763, Republic of Korea

^b Developing a Team Response Using Digital Construction to Mitigate Disasters Related to Climate Change (BK21 FOUR), Hanyang University, Seoul, Seongdong-gu, 04763, Republic of Korea

^c Department of Environmental Science & Ecological Engineering, College of Life Sciences & Biotechnology, Korea University, Seoul, Seongbuk-gu, 02841, Republic of Korea

ARTICLE INFO

Keywords:

Multivariate time-series imputation
Full-scale wastewater treatment plants
Consecutive block missingness
Cross-plant evaluation
Model robustness

ABSTRACT

Data-driven operation of wastewater treatment plants (WWTPs) requires reliable monitoring data, yet missing values result from sensor failures and operational interruptions. Despite numerous proposed imputation models, WWTP evaluations rely on a single facility, randomly distributed point missingness, aggregate performance without variable-level analysis, and limited statistical analysis. This study addresses these limitations through the first multi-plant evaluation of six imputation approaches (linear interpolation, KNN, MICE, CSDI, Transformer, and SAITS) across 14 WWTP datasets. Models were evaluated at 5–40% missingness under non-consecutive and 6-day block patterns using R^2 and nRMSE. Performance varied markedly across facilities and variables. Even the best-performing model yielded R^2 of 0.09–0.74, while linear interpolation remained competitive for weakly correlated effluent variables and multivariate models benefited from stronger influent correlations. Kendall's W indicated reproducible rate-dependent responses across plants, while ANOVA and OLS regression showed significant effects of both missingness pattern and rate. Performance was lower under consecutive than non-consecutive missingness at the same rate and declined with increasing rate, with model-specific sensitivity. Transformer and SAITS showed low sensitivity to missingness pattern and rate with a favorable balance between accuracy and computational efficiency. In short records, consecutive block missingness produced a counterintuitive inversion, with R^2 being lower at a 5% missingness rate than at 40%, whereas nRMSE increased consistently with the missingness rate. This discrepancy indicates that rate-based evaluation in short records with consecutive missingness depends on the selected metric. These findings distinguish dataset-driven from model-specific effects and guide the selection and evaluation of robust imputation strategies.

* Corresponding author. Department of Civil and Environmental Engineering, Hanyang University, Seongdong-gu, Seoul, 04763, Republic of Korea

** Corresponding author. Department of Environmental Science & Ecological Engineering, College of Life Sciences & Biotechnology, Korea University, Seoul, Seongbuk-gu, 02841, Republic of Korea.

E-mail addresses: slee2024@korea.ac.kr (S. Lee), youngmo@hanyang.ac.kr (Y.M. Kim).

<https://doi.org/10.1016/j.wri.2026.100394>

Received 1 July 2026; Received in revised form 10 August 2026; Accepted 23 August 2026

Available online 24 August 2026

2212-3717/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

1. Introduction

Developments in machine learning (ML) have facilitated the adoption of data-driven monitoring and control frameworks within industrial sectors, particularly wastewater treatment plants (WWTPs) [1,2]. These systems are widely used for managing complex and nonlinear processes under variable operating conditions to ensure effluent compliance [3,4]. However, the effectiveness of data-driven systems depends critically on the availability of comprehensive datasets [5]. Given that sensor limitations and operational hurdles frequently result in missing data [6,7], data incompleteness can disrupt real-time monitoring and operational control in WWTPs [8]. Robust imputation techniques are vital in preserving the integrity and predictive reliability of monitoring infrastructures.

Various multivariate time-series imputation techniques have been proposed, spanning foundational statistical methods and conventional machine learning algorithms such as k-nearest neighbors (KNN) and multiple imputation by chained equations (MICE) [9]. Recent advancements have highlighted the advantages of deep learning (DL) architectures in capturing complex temporal dependencies and inter-variable correlations [10–12]. Multivariate time-series imputation has been widely studied across domains such as healthcare, transportation, and environmental monitoring [13–15]. Many existing studies rely on standardized datasets collected under relatively stable conditions [5,16].

WWTP monitoring datasets differ from these domains in two key aspects. First, data characteristics vary substantially across plants, including differences in distributions, variability, and inter-variable correlations, reflecting variations in treatment processes, operation systems, and catchment conditions [17,18]. Second, missing data in WWTP monitoring systems often occur as consecutive blocks rather than isolated random points, arising from practical monitoring constraints such as sensor malfunction, interference, and manual sampling [19,20]. These factors complicate the imputation process, as across-plant variability may degrade model generalizability while realistic missingness patterns, such as consecutive blocks, may further increase the difficulty of reconstruction.

Research into WWTP monitoring has increasingly focused on imputation models [21–23]. However, experimental coverage remains limited. Most studies are based on a single WWTP or a small number of datasets, failing to address how statistical heterogeneity across different facilities affects model efficacy [21,24,25]. Additionally, missingness is typically simulated as randomly distributed points or a single continuous block [6,25], disregarding multiple consecutive missing blocks observed in real-world monitoring data.

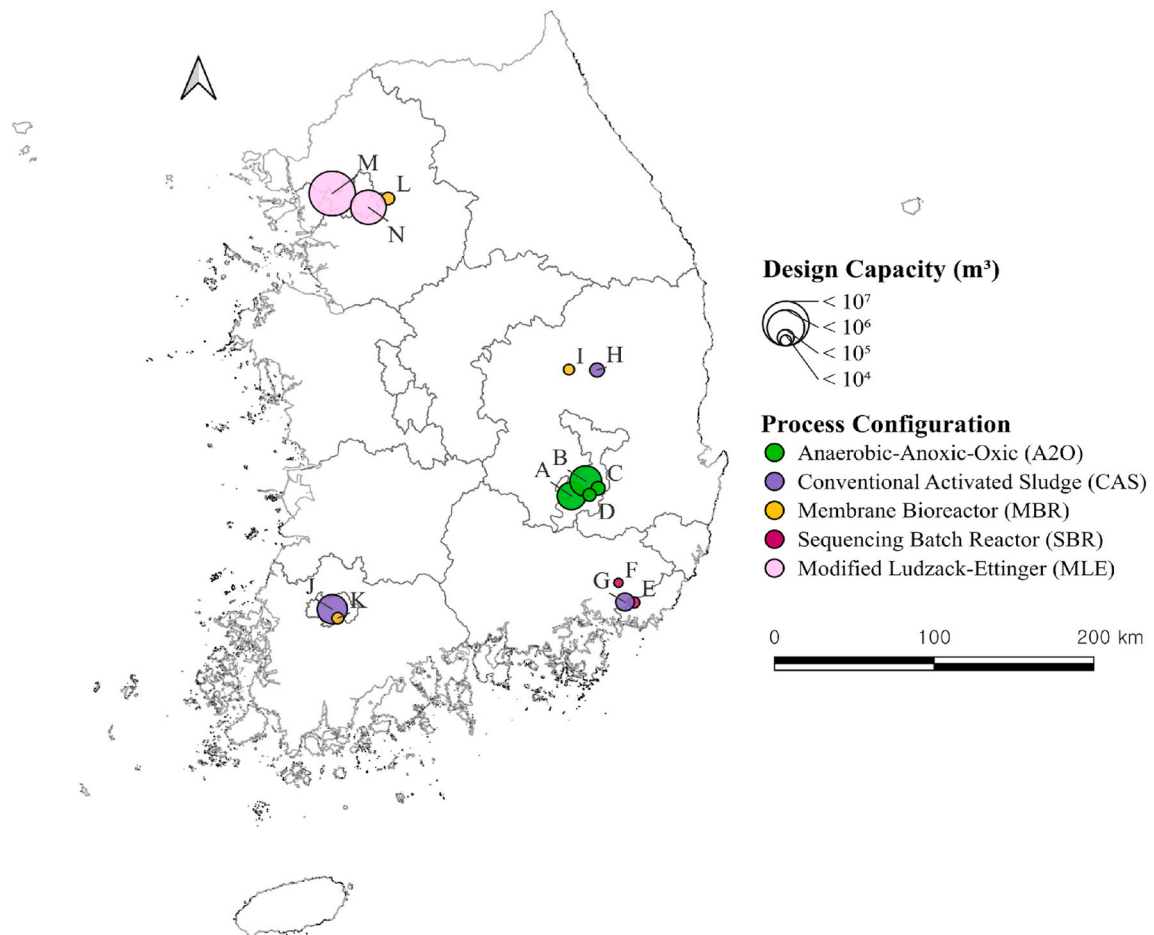


Fig. 1. Geographic distribution of the collected WWTP datasets, including design capacities and process configurations.

While practitioners have long recognized that short monitoring records, weak inter-variable correlations, and prolonged missing periods adversely affect imputation accuracy, these effects have rarely been systematically quantified across multiple full-scale WWTPs operating under heterogeneous conditions.

Furthermore, analytical gaps persist in multivariate time-series imputation studies. Most studies have compared model performance within specific datasets, with limited analysis of performance variability across datasets [13,26]. Consequently, the factors governing performance differences across models and datasets remain insufficiently understood. Moreover, imputation performance is often evaluated at the dataset level or using representative variables, potentially obscuring variable-level differences that are critical in multivariate monitoring systems [6,27]. Because imputation performance can depend on variable-specific statistical properties, variable-level evaluation is needed to better characterize imputation reliability among variables [28,29]. In addition, many imputation studies have also used R^2 as a primary evaluation metric [15,16,26], although its interpretation may depend on the variance and temporal distribution of the evaluated observations. Complementary metrics are therefore needed to distinguish reconstruction accuracy.

To the best of our knowledge, previous studies on multivariate environmental time-series imputation have relied mostly on single-facility or synthetically idealized datasets, leaving the influence of plant-specific heterogeneity on imputation performance largely unquantified. Rather than proposing a new imputation algorithm, this study focuses on identifying and quantifying practical factors governing imputation performance under realistic WWTP monitoring conditions. Specifically, this study aims to: (i) compare the imputation performance and computational efficiency of six models across 14 full-scale WWTP datasets; (ii) analyze how cross-plant and within-plant dataset heterogeneity drives variation in imputation performance; and (iii) assess the impact of missingness conditions, including patterns (non-consecutive vs. consecutive block missingness) and rates, on imputation performance. Through this multi-plant evaluation, this study seeks to delineate the operational boundaries of current imputation models by statistically validating factors previously characterized only empirically. The comparison is also intended to inform how imputation is assessed under WWTP operational missingness conditions.

2. Materials and methods

2.1. Data collection

Monitoring data were obtained from 14 full-scale WWTPs across South Korea (Fig. 1). The plants represented diverse operational conditions, including different geographic locations, process configurations, and design capacities. The process configurations

Table 1
Summary of collected datasets and variables.

WWTP	Design capacity (m ³)	Process configuration	Number of observations	Monitoring period	Variables
A	520,000	A2O	361	Jan 2021– Dec 2021	Rainfall, Temp,
B	680,000	A2O	362	Jan 2021– Dec 2021	Q _i , BOD _i , TOC _i , SS _i , TN _i , TP _i , BOD _o , TOC _o , SS _o , TN _o , TP _o
C	47,000	A2O	362	Jan 2021– Dec 2021	
D	33,750	A2O	362	Jan 2021– Dec 2021	
E	15,000	SBR	365	Jan 2021– Dec 2021	
F	1600	SBR	365	Jan 2023– Dec 2023	
G	145,000	CAS	365	Jan 2023– Dec 2023	
H	54,000	CAS	365	Jan 2023– Dec 2023	
I	9000	MBR	365	Jan 2023– Dec 2023	
J	600,000	CAS	365	Jan 2021– Dec 2021	
K	16,000	MBR	365	Jan 2021– Dec 2021	
L	32,000	MBR	2828	Mar 2015–Dec 2022	
M	1,630,000	MLE	1079	Jan 2021–May 2024	
N	900,000	MLE	1247	Jan 2021–May 2024	

a i indicates influent variables, and o indicates effluent variables.

b A2O = anaerobic–anoxic–oxic; CAS = conventional activated sludge; SBR = sequencing batch reactor; MBR = membrane bioreactor; MLE = modified Ludzack–Ettinger.

included anaerobic–anoxic–oxic (A2O), conventional activated sludge (CAS), membrane bioreactor (MBR), sequencing batch reactor (SBR), and modified Ludzack–Ettinger (MLE). Monitoring periods varied among WWTPs according to data availability, ranging from one year to multiple years of daily observations. The WWTP characteristics, monitoring periods, record lengths (n), and variables included in each dataset are summarized in [Table 1](#).

A consistent set of meteorological, hydraulic, influent, and effluent variables was used across the WWTP datasets. Water quality variables comprised biochemical oxygen demand (BOD), total organic carbon (TOC), suspended solids (SS), total nitrogen (TN), and total phosphorus (TP) in the influent and effluent streams. Influent flow, influent water temperature, and precipitation were also included.

Influent and effluent variables were aligned by date, and observations recorded on the same day were treated as concurrent. The sensitivity of the results to this temporal alignment assumption was evaluated in an additional analysis ([Supplementary Methods S1](#)).

2.2. Dataset characterization

Dataset characterization was conducted to quantify statistical and structural heterogeneity among the 14 WWTP datasets. Statistical, inter-variable, and temporal descriptors were calculated to characterize both overall dataset structure and variation among variables.

2.2.1. Dataset statistical characteristics

The statistical properties of each dataset were quantified using descriptors of variability, distributional shape, and multivariate structure, consistent with standard multivariate time-series characterization [30]. Univariate characteristics were represented by the maximum coefficient of variation (CV), skewness, and kurtosis across variables. Dataset-level variability was further represented by the mean univariate CV (MCV_{Avg}) and trace-based multivariate CV (MCV_{Trace}) [31]. Mardia's multivariate skewness and kurtosis were calculated to characterize multivariate distributional shape [32].

Inter-variable dependence and temporal similarity were characterized using Pearson correlation, dynamic time warping (DTW) [33], and principal component analysis (PCA), respectively. Pairwise Pearson correlation values were summarized using mean absolute values, and pairwise DTW distances were normalized and averaged to represent dataset-level characteristics. The cumulative variance explained by the first two principal components ($PC1 + PC2$) was used to indicate the extent to which the multivariate structure was concentrated in a low-dimensional representation.

Definitions and interpretations of these descriptors are provided in [Table S1](#).

2.3. Imputation experiments

Imputation experiments were conducted separately for each WWTP dataset. For each dataset, synthetic missingness was introduced according to predefined missingness rates and patterns, and the resulting incomplete data were imputed using six approaches. All models were evaluated using the same masking, scaling, imputation, and scoring procedure. The overall experimental and evaluation framework is summarized in [Fig. 2](#).

2.3.1. Missingness scenarios and synthetic missingness generation

Missingness scenarios were defined according to missingness rate and pattern [23]. Imputation performance was evaluated under increasing missingness rates of 5%, 10%, 15%, 20%, 30%, and 40%.

Two missingness patterns were considered: non-consecutive missing points and consecutive missing blocks. Consecutive block missingness was included to reflect operational data gaps observed in full-scale WWTP monitoring datasets [21,24]. Based on the observed distribution of missing intervals ([Fig. S1](#)), a six-day block length was selected to represent the most prevalent non-measurement duration in practice [34]. As a supplementary sensitivity analysis, the model comparison was repeated using block lengths of 3 and 14 days ([Supplementary Methods S2](#)).

To evaluate multivariate imputation performance, all datasets were complete before synthetic masking and were used as ground truth. Synthetic missingness was generated by masking observations in all variables except precipitation, which was treated as an exogenous meteorological covariate rather than an imputation target. Masking locations were assigned randomly across the monitoring period separately for each variable and without regard to the observed value, allowing occasional overlap in missingness across variables. The synthetic missingness was therefore treated as missing completely at random (MCAR) [35], with temporal clustering imposed under the consecutive pattern. For each masked variable, the target number of missing observations was determined from the dataset length and the predefined missingness rate. Under the non-consecutive pattern, individual observations were randomly selected across the full monitoring period. Under the consecutive pattern, missing observations were generated as six-day blocks. The resulting number of blocks varied with dataset length and missingness rate. When the target number was not divisible by six, one additional shorter block was generated for the remainder.

Each missingness scenario was repeated over 50 independent iterations using different masking locations to reduce timing-related sampling variability. Masked observations were excluded from model input and retained only as reference values for performance evaluation.

2.3.2. Imputation models and implementation

Six imputation models were selected to represent distinct paradigms for multivariate time-series imputation ([Table 2](#)). Linear

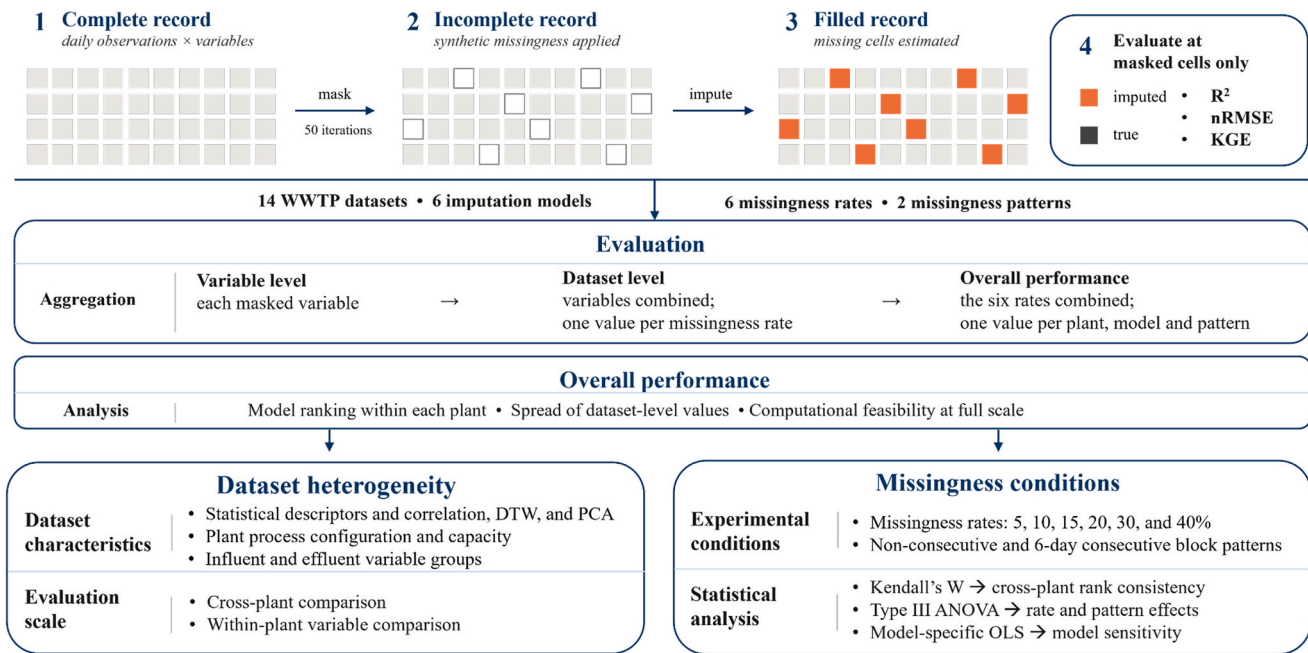


Fig. 2. Experimental and evaluation framework for assessing imputation performance across heterogeneous WWTP datasets and missingness conditions.

interpolation served as a univariate statistical baseline model reflecting standard operational monitoring systems. The ML-based approaches, KNN and MICE [36], leverage multivariate relationships. The DL models comprised the conditional score-based diffusion model for imputation (CSDI) [10], Transformer [37], and self-attention-based imputation for time series (SAITS) [38]. CSDI reconstructs missing values through conditional iterative denoising, whereas Transformer and SAITS use attention mechanisms to capture temporal and inter-variable dependencies. The methodological characteristics of the six models are summarized in Table 2.

For each masking iteration, every model received only the incomplete dataset generated under the corresponding missingness scenario (Fig. 2). Before imputation, each variable was standardized using the mean and standard deviation of the observations that remained after masking, and the standardization was inverted before scoring. Linear interpolation, KNN, and MICE estimated missing values from the available remaining observations, whereas CSDI, the Transformer, and SAITS obtained their internal parameters from the observed entries of the same incomplete record. The original observations at the masked positions were withheld from every model and used only for performance evaluation.

Model hyperparameters were selected separately for each WWTP dataset. The search ranges and selected values are provided in Table S2, and the search procedure is described in Supplementary Methods S3.

All models were implemented in Python 3.9.25. Linear interpolation and KNN were implemented in scikit-learn 1.6.1, MICE in fancyimpute 0.7.0, and DL models in PyTorch 2.5.1 (CUDA 12.4) within the PyPOTS 1.0 framework [39].

2.4. Multi-plant performance evaluation

Imputation performance was evaluated using synthetic missing data generated under predefined scenarios across all WWTP datasets.

2.4.1. Imputation performance and computational efficiency

Imputation performance was assessed using three metrics: the coefficient of determination (R^2), a range-normalized root mean squared error (nRMSE), and Kling–Gupta efficiency (KGE) [40]. R^2 was used as the primary metric to assess how well the imputed values reproduced the observed variability. nRMSE and KGE were used as complementary metrics, providing information on error magnitude relative to the variable range and overall agreement, respectively. The equations are provided in Supplementary Methods S4.

Computational efficiency was compared using the total run time of one iteration. The run time covers the whole procedure, from parameter estimation to the filling of the masked positions, and was recorded for every condition.

2.4.2. Performance aggregation and comparative evaluation

Performance metrics were calculated at the synthetically masked positions of each variable. For each combination of WWTP, model, missingness pattern, missingness rate, and variable, performance was summarized as the mean and standard deviation across 50 masking iterations. For dataset-level evaluation, the variable-level mean values were averaged across all variables subjected to synthetic masking for each missingness condition. These values represented performance for each WWTP, model, missingness pattern, and missingness rate.

For cross-plant comparison, overall performance for each WWTP, model, and missingness pattern was obtained by averaging the dataset-level values across the six missingness rates. These rate-integrated values were used as representative WWTP-level performance measures for model comparison and ranking. Models were ranked within each WWTP, and mean ranks across the 14 WWTP datasets were used to assess the reproducibility of relative model performance across facilities.

For within-plant analysis, variable-level performance was evaluated separately for each model, missingness pattern, and missingness rate, enabling comparisons among variables under the same missingness condition.

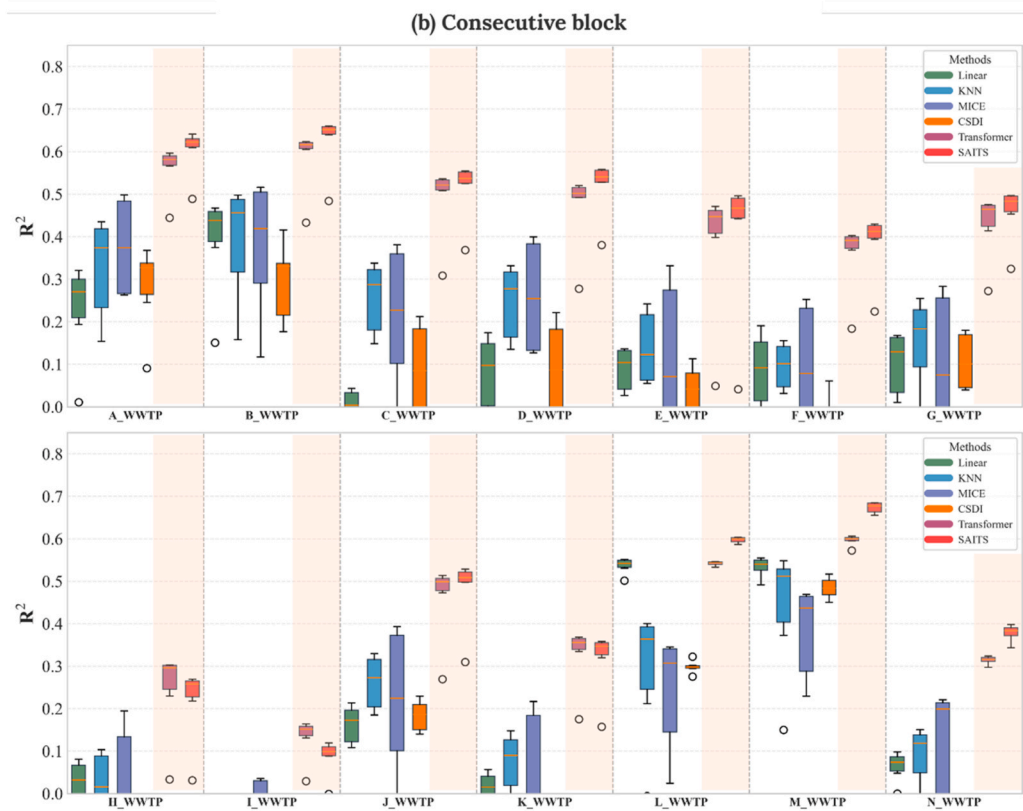
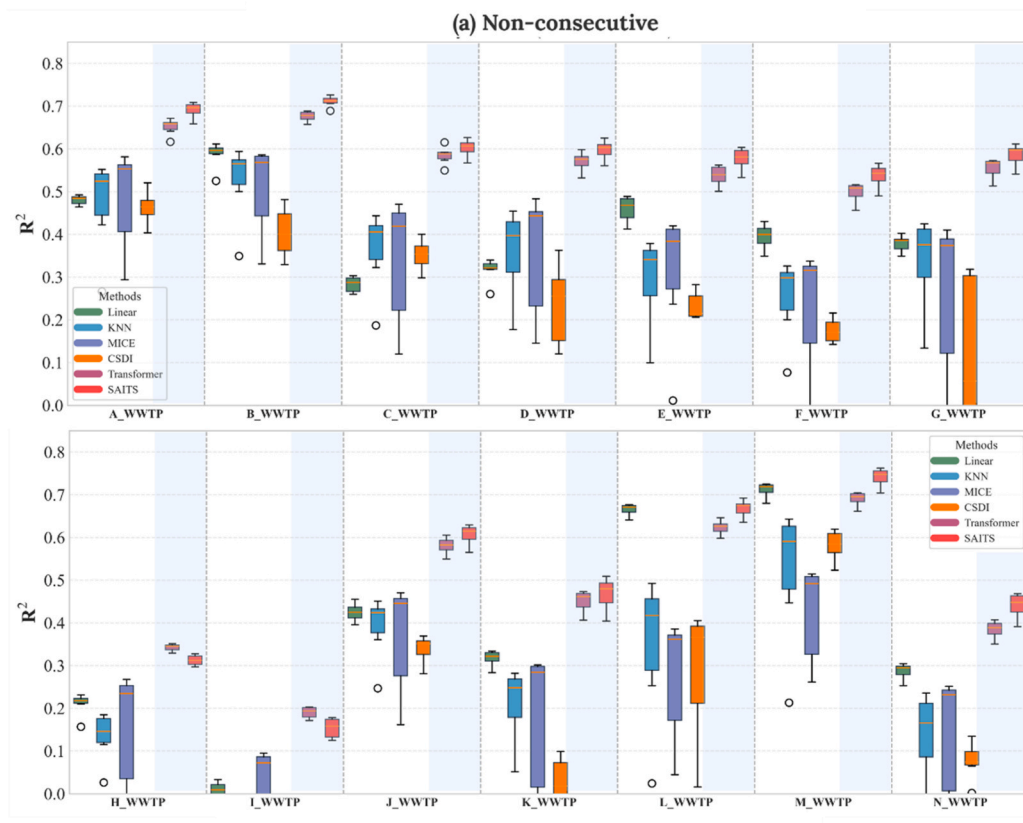
2.5. Statistical analysis

Statistical analyses were conducted using dataset-level performance values to quantify cross-plant reproducibility and characterize performance variation associated with dataset and missingness conditions. The analytical procedures and the performance values used

Table 2

Overview of imputation models evaluated in this study and their modeling characteristics.

Model	Category	Core architecture	Multivariate modeling	Temporal modeling	Key concept
Linear interpolation	Statistical	Interpolation	No	Local interpolation	Adjacent-value interpolation
KNN	ML	Distance-based	Yes	No	Similarity-based averaging
MICE	ML	Regression-based	Yes	Limited	Iterative chained regression
CSDI	DL	Diffusion-based generative model	Yes	Yes	Probabilistic imputation via score-based diffusion
Transformer	DL	Self-attention architecture	Yes	Yes	Attention-based sequence modeling
SAITS	DL	Self-attention architecture	Yes	Yes	Attention-based imputation for time-series data



(caption on next page)

Fig. 3. Comparison of R^2 across models and WWTP datasets. Boxplots illustrate R^2 distributions across missing rates under (a) non-consecutive and (b) consecutive block missingness.

for each analysis are described in [Supplementary Methods S5–S6](#). All analyses were implemented using SciPy 1.13.1 and statsmodels 0.14.6.

2.5.1. Reproducibility of the missingness-rate effect across plants

To assess whether the effect of missingness rate was consistent across the 14 WWTP datasets, Kendall's coefficient of concordance (W) [41] was calculated separately for each model and missingness pattern. The six missingness rates were treated as ranked objects and the 14 WWTP datasets as raters, so that W quantifies the consistency of the rate ordering among plants. Reference thresholds of 0.5 and 0.7 were used for moderate and strong concordance, and statistical significance was evaluated using the chi-squared approximation.

2.5.2. Effects of missingness pattern and rate

Type III analysis of variance was applied separately to R^2 and nRMSE to evaluate the effects of missingness pattern, missingness rate, and their interaction on imputation performance. The ordinary least squares (OLS) model included missingness pattern, mean-centered missingness rate, their interaction, model type, and WWTP identity as fixed effects, with WWTP serving as a blocking factor.

2.5.3. Model-specific sensitivity to missingness conditions

To quantify how individual models responded to missingness pattern and rate, the OLS structure separately to each imputation model. The intercept difference represented sensitivity to missingness pattern, the rate slope represented sensitivity to increasing missingness rate, and the interaction term represented the difference in rate sensitivity between patterns.

3. Results and discussion

3.1. Overall imputation performance across models

Among the six imputation models evaluated using 14 full-scale WWTP datasets, SAITS and Transformer consistently outperformed the other models across datasets ([Fig. 3](#) and [Table S3](#)). SAITS ranked first in 12 of the 14 datasets under non-consecutive missingness and in 11 of the 14 under consecutive block missingness, while Transformer ranked first in the remaining datasets. The joint distribution of R^2 and nRMSE further distinguished the models across datasets and missingness rates ([Fig. 4](#)). Under both missingness patterns, SAITS and Transformer were concentrated in the region corresponding to higher R^2 and lower nRMSE, with comparatively limited dispersion across the 14 datasets and six missingness rates. In contrast, linear interpolation, KNN, MICE, and particularly CSDI showed broader distributions that extended toward lower or negative R^2 and higher nRMSE values. Consecutive block missingness generally shifted the distributions toward poorer performance and increased their spread, although this deterioration was less pronounced for SAITS and Transformer.

For operational deployment in WWTP systems, computational cost is also a critical factor ([Table S4](#)). For the longest record (WWTP L, $n = 2828$), the total runtime was 1439 s for CSDI, 65.6 s for SAITS, 37.2 s for Transformer, and less than 1 s for linear interpolation, KNN, and MICE. All models were therefore completed within the daily compilation interval of the present datasets. However, their computational demands diverged as record length increased. Relative to the one-year WWTP A record, parameter estimation time on WWTP L, which contained approximately eight times as many observations, increased by 2.9 times for Transformer, 2.7 times for SAITS, and 7.3 times for CSDI ([Table S4](#)). Multi-year records or finer sampling intervals would therefore constrain CSDI before the other models. Overall, attention-based models achieved the highest imputation accuracy at favorable computational cost, consistent with recent benchmarks on multivariate time-series imputation [26].

Despite the consistent top ranking of SAITS and Transformer, the absolute performance of individual models varied substantially across WWTP datasets and missingness conditions ([Fig. S2](#)). Averaged across the six missingness rates, SAITS achieved dataset-level R^2 values ranging from 0.15 in WWTP I to 0.74 in WWTP M under non-consecutive missingness and from 0.09 to 0.67 for the same plants under consecutive block missingness. Moreover, consecutive block missingness ([Fig. 3b](#)) yielded lower R^2 than non-consecutive patterns ([Fig. 3a](#)) for all models. These variations across datasets, missingness rates, and missingness patterns indicate that imputation outcomes in WWTP cannot be fully characterized by model-level comparison alone. The following sections therefore analyze dataset characteristics (Section 3.2) and missingness conditions (Section 3.3) as two complementary perspectives on these performance differences.

3.2. Dataset-related variability in imputation performance

3.2.1. Cross-plant heterogeneity in dataset characteristics

The fourteen WWTP datasets exhibited pronounced statistical divergence in distributional shape, variability, and inter-variable relationships (Max kurtosis: 3.90–62.94; MCVavg: 0.20–0.40; mean correlation: 0.10–0.39) ([Tables S5 and S6](#), and [Fig. S3](#)). This variability in statistical characteristics can affect imputation performance variation across datasets [42,43].

In other environmental monitoring domains, such as air [44] and freshwater systems [45], statistical characteristics tend to be

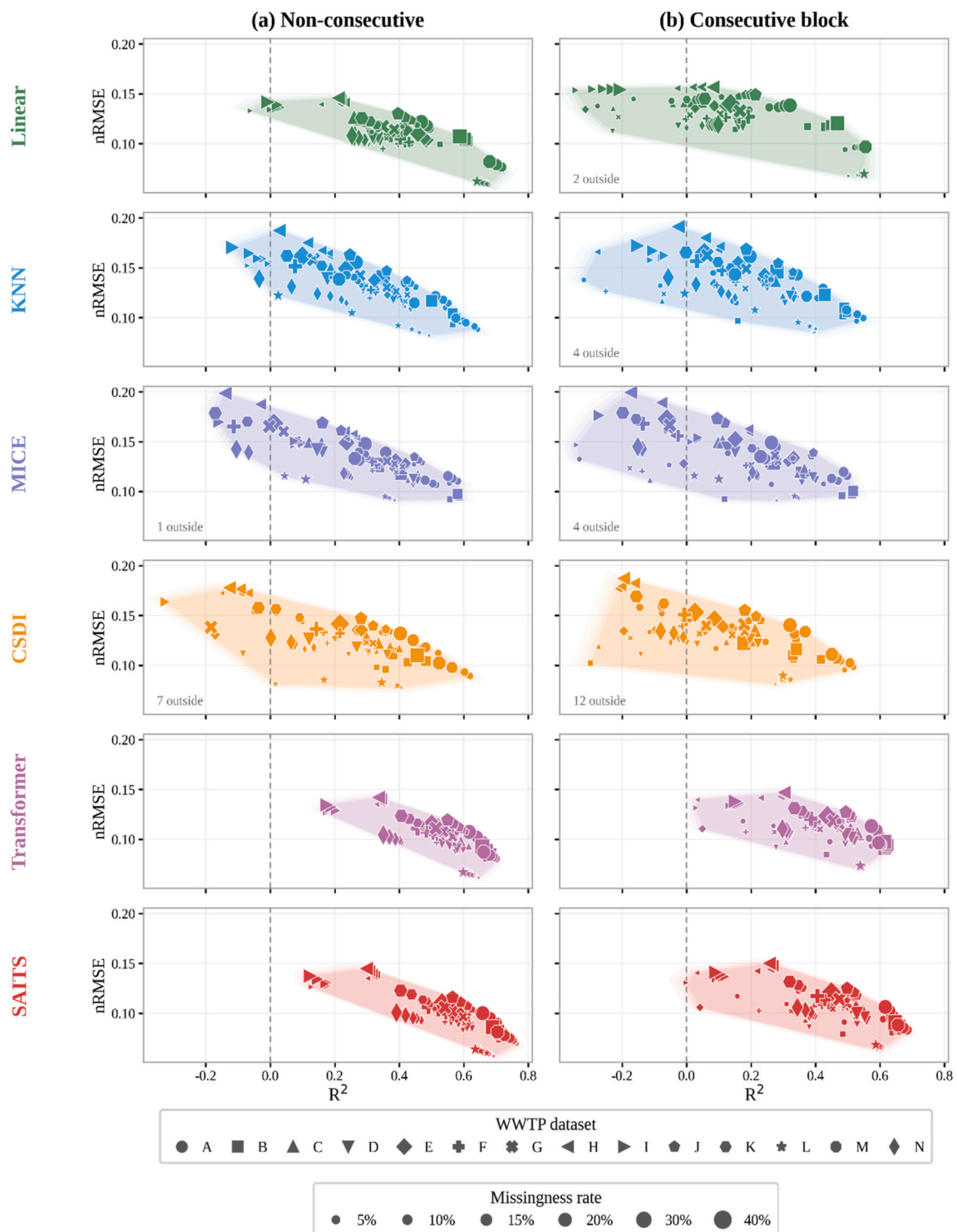


Fig. 4. Joint distribution of R^2 and nRMSE across WWTP datasets, missingness rates, and imputation models under non-consecutive and consecutive block missingness.

relatively similar across sites, as monitoring targets are not subject to engineered treatment processes. This similarity supports stable imputation performance with low cross-site variability. Accordingly, Wang et al. [46] reported high imputation performance ($R^2 > 0.95$) across multiple monitoring locations, with consistently low variability. In contrast, WWTP datasets displayed pronounced site-specific heterogeneity in their statistical profiles (Tables S5 and S6, and Fig. S3). Driven by diverse treatment processes and

operational conditions. For example, Fig. S4 compares four WWTPs within the same region, revealing substantial differences in value ranges and temporal variability despite similar regional conditions.

The marked heterogeneity characteristic of WWTP datasets arises from diverse factors [43]. These include influent characteristics, operational control strategies, process configurations and plant capacity. Consequently, statistical properties varied across plants and lacked a consistent pattern. Statistical divergence also remained within the same process configuration, indicating that no single factor sufficiently explained the observed differences. Sequencing batch reactors (SBRs), such as WWTPs E and F, exhibited higher variability (MCV_{avg} : 0.37–0.40) than the remaining 12 continuous-flow WWTPs (MCV_{avg} : 0.20–0.38). This pattern is consistent with the phase-dependent fluctuations of SBR systems [47]. However, even within the same SBR process, distributional characteristics differed between WWTP E and F, with WWTP F exhibiting higher maximum skewness and kurtosis (3.386 and 19.755) than WWTP E (2.123 and 7.432). This difference may partly related plant scale, as the design capacity of WWTP E exceeds that of WWTP F by over tenfold, potentially moderating water-quality fluctuations and distributional extremes [48]. These findings highlight the multifactorial and plant-specific nature of statistical heterogeneity in WWTP datasets, which was more pronounced than that reported for other environmental monitoring systems. Such heterogeneity provides a plausible basis for the differences in imputation performance observed among WWTPs and supports the need for multi-plant evaluation when assessing model robustness under full-scale operating conditions.

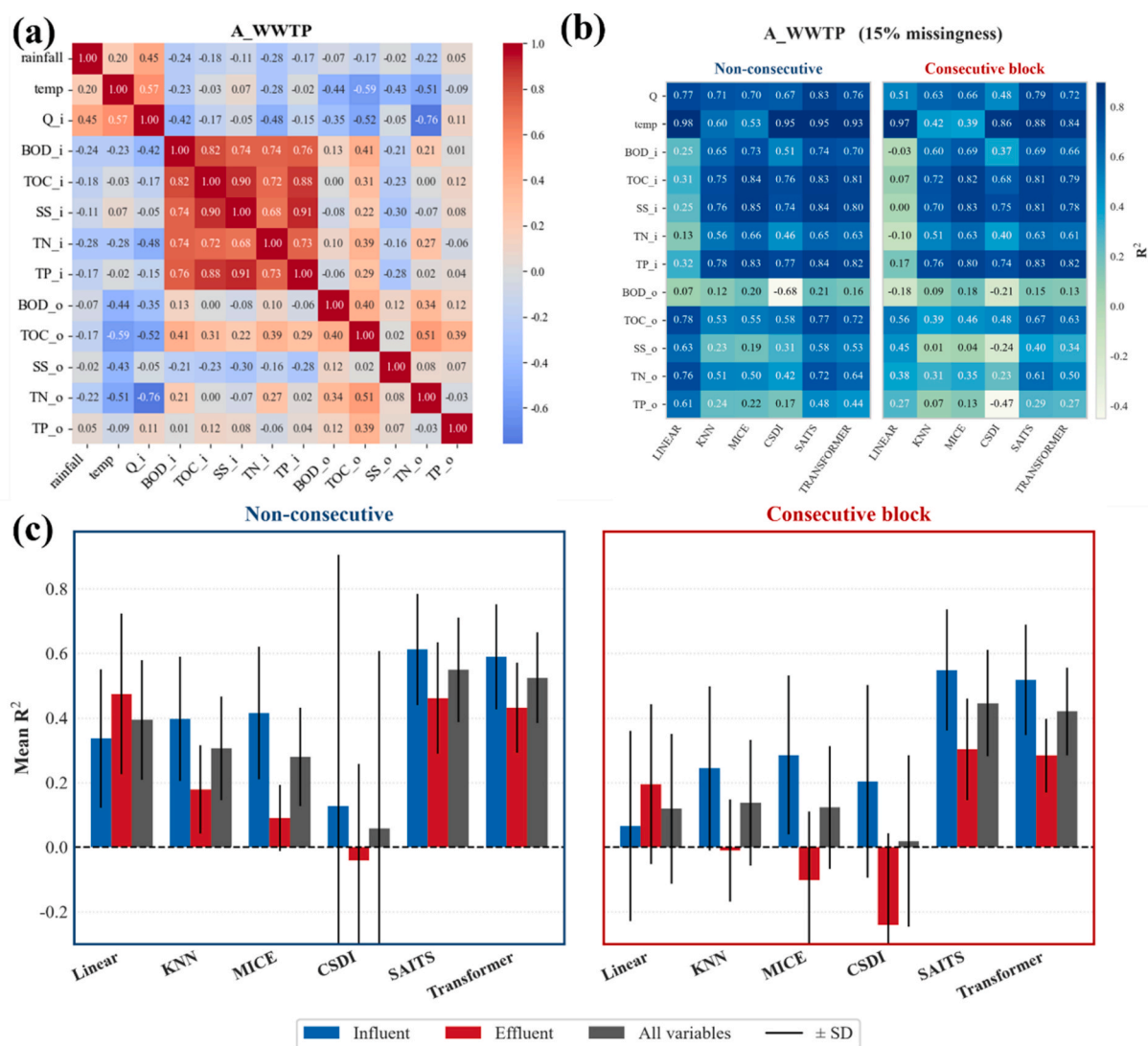


Fig. 5. Influence of dataset structure on variable-level imputation performance: (a) Pearson correlation matrix of monitoring variables for WWTP A; (b) Variable-level imputation performance (R^2) for WWTP An under non-consecutive and consecutive block missingness; (c) Average imputation performance across fourteen WWTPs, grouped by influent and effluent variables.

3.2.2. Within-plant heterogeneity in variable correlations

In addition to the cross-plant differences in overall dataset characteristics, substantial heterogeneity was also evident within individual WWTP datasets. In particular, inter-variable correlation structures showed a consistent contrast between influent and effluent variables across all plants (Fig. S5). Inter-variable correlations among influent variables (BOD_i , TOC_i , SS_i , TN_i , and TP_i) were consistently stronger than those among effluent variables (BOD_o , TOC_o , SS_o , TN_o , and TP_o). For example, in WWTP A, the correlation coefficient was 0.74 between BOD_i and SS_i , compared with 0.12 between BOD_o and SS_o (Fig. 5a). This contrast reflects process-related mechanisms in which influent variables share common pollution sources, resulting in strong correlations [49], whereas treatment processes progressively decouple these relationships through physical, biological, and chemical transformations [50].

Hydraulic retention time may cause temporal misalignment because same-day influent and effluent measurements are not hydraulically corresponding observations [51]. Moreover, a fixed temporal shift cannot fully represent plant hydraulics because residence-time distributions are influenced by mixing characteristics and flow variability [51,52]. A sensitivity analysis of temporal alignment showed that a one-day influent shift produced only minor changes in R^2 across models and missingness patterns (Fig. S6). In practical applications, explicit correction for hydraulic retention is difficult because the influent–effluent time lag is plant-specific and varies with operating conditions. Together, these findings suggest that treatment-induced decoupling more plausibly explains the weaker effluent correlations than temporal misalignment resulting from the same-day alignment of influent and effluent observations.

These within-plant differences in correlation structure were reflected in variable-level imputation performance (Fig. 5b and c). The WWTP A dataset further exemplifies this evaluation bias, as imputation performance differed across individual variables, with the most significant divergence occurring between influent and effluent variables (Fig. 5b). Comparable variable-level heterogeneity was observed in the other datasets (Fig. S7). Importantly, the relative performance of influent and effluent variables differed among models (Fig. 5c). For example, under non-consecutive missingness, KNN and Transformer architectures achieved higher R^2 values for influent variables compared to effluent variables (KNN: 0.32–0.43 vs. 0.08–0.30; Transformer: 0.46–0.57 vs. 0.28–0.54). Conversely, the linear model exhibited an inverse trend, performing better on effluent variables (0.22–0.65) than on influent ones (0.06–0.20). These opposing model-specific patterns can be obscured when performance is summarized only at the plant level by averaging across variables. The mechanisms underlying these model-specific differences are discussed in Section 3.2.3.

3.2.3. Model generalizability under cross-plant and within-plant heterogeneity

The model-specific discrepancies documented above are primarily attributable to differences in each model's capacity to exploit temporal and multivariate dependencies. Variable-level performance showed a consistent distinction between univariate (linear interpolation) and multivariate models. Linear interpolation generally achieved higher performance for effluent than for influent variables, whereas most multivariate models performed better for influent variables. This discrepancy is attributable to the contrasting properties of the two data streams; specifically, influent variables were characterized by expansive ranges, significant temporal oscillations (Fig. S4b), and robust inter-variable correlations, features that bolster the utility of multivariate models [53]. By comparison, effluent variables had narrower distributions, lower temporal variability (Fig. S4c), and weaker inter-variable correlations. Under these relatively stable conditions, linear interpolation remained competitive by relying on local temporal continuity [54,55]. Conversely, KNN and MICE, which rely more heavily on inter-variable relationships [56], showed reduced relative performance for effluent variables, where correlations were weaker. These results suggest that the relative contribution of temporal and multivariate information differs between influent and effluent variables [53]. Accordingly, robust imputation in WWTP systems requires models capable of jointly leveraging temporal and multivariate information across variables.

Among the multivariate models, Transformer and SAITS showed the most robust performance across variable groups (Fig. 5b and Fig. S7). Averaged across the 14 datasets, their relative performance against linear interpolation differed substantially between influent and effluent variables (Fig. 5c). For effluent variables, SAITS and Transformer achieved R^2 of 0.46 and 0.43, respectively, compared with 0.47 for linear interpolation under non-consecutive missingness; under consecutive block missingness, the R^2 reached 0.30 and 0.28, exceeding linear interpolation at 0.20. For influent variables, SAITS and Transformer substantially outperformed linear interpolation under both non-consecutive missingness (0.54 and 0.52 vs. 0.16) and consecutive block missingness (0.50 and 0.48 vs. –0.16). Thus, SAITS and Transformer were the only models that maintained high R^2 across both variable groups. The performance of attention-based models (SAITS and Transformer) derives from their joint modeling of temporal and multivariate relationships, ensuring stable performance despite within-plant heterogeneity [23,57].

Despite being engineered to integrate both temporal and multivariate dependencies, CSDI exhibited the least robustness among the six models evaluated, with dataset-level R^2 ranging from –1.58 to 0.58, the widest variability observed across plants. This diminished resilience likely stems from its diffusion-based architecture, which reconstructs signals via iterative denoising based on learned joint distributions [10]. Such mechanisms are sensitive to the distributional shifts and variability present across diverse datasets [10,58], constraining CSDI's efficacy within highly heterogeneous WWTP environments.

These findings indicate that cross-plant and within-plant heterogeneity in statistical properties constitute primary determinants of imputation performance in WWTP systems. Even SAITS, the best-performing model, exhibited dataset-level R^2 values ranging from 0.15 to 0.74 under non-consecutive missingness and from 0.09 to 0.67 under consecutive block missingness (Fig. S2), demonstrating that plant-specific dataset characteristics impose fundamental constraints on imputation performance that model architecture alone cannot fully overcome.

3.3. Effects of missingness conditions on imputation performance

3.3.1. Effects of missingness rate and patterns

Missingness-rate rankings were broadly reproducible across the 14 WWTP datasets. Kendall's W was at least 0.50 for R^2 across both missingness patterns for all models except CSDI and at least 0.81 for nRMSE across all models and patterns (Fig. 6; Table S7). WWTP A therefore provides a representative illustration of the common rate-dependent response observed across the 14 datasets, although its absolute performance level remained facility-specific (Fig. 7; Table S3).

For missingness rate, imputation performance decreased with increasing missingness rate across all models under non-consecutive missingness, although the magnitude of this degradation was model-dependent. In WWTP A, R^2 fell from 0.54 to 0.27 for KNN and from 0.56 to 0.29 for MICE as missingness increased from 5% to 40%, whereas the corresponding decreases in imputation performance were smaller for SAITS and Transformer, from 0.69 to 0.66 and from 0.66 to 0.62, respectively (Fig. 7a). The downward trend remained consistent across all datasets, with KNN showing near-perfect concordance in its nRMSE-based rate ranking across plants ($W = 1.00$; Fig. 6). This trajectory aligns with established literature documenting systematic performance erosion under increasing missingness rates [15,16,59].

Regarding the impact of missingness patterns, consecutive missingness consistently resulted in lower overall performance than non-consecutive patterns at every rate and for every model. In WWTP A, SAITS achieved the highest R^2 at every missingness rate under both patterns, with these ranging from approximately 0.65–0.71 under non-consecutive missingness (Fig. 7a) and from 0.49 to 0.64 under consecutive block missingness (Fig. 7b). Wu et al. [60] reported that consecutive missingness poses significant challenges for multivariate time-series imputation. The disruption of temporal continuity intrinsic in consecutive patterns severely compromises overall imputation performance [61].

Analysis of variance confirmed that both the missingness rate and the missingness pattern significantly affected imputation performance (Table S8). Under R^2 , the pattern effect exceeded the rate effect ($F = 61.31$ vs. 10.66), and under nRMSE the rate effect was larger ($F = 377.85$ vs. 260.61). Given that consecutive data gaps are a frequent occurrence in WWTP operations [19,20], robustness under consecutive block missingness remains particularly important for practical application.

3.3.2. Limitations of rate-based evaluation and metric choice under consecutive block missingness

A counterintuitive trend emerged under consecutive block missingness. Unlike the expected decline in imputation efficacy with increasing missingness rate, as previously reported for single continuous gaps [61], the pattern was inverted under certain conditions (Fig. 7b). In WWTP A, R^2 reached its lowest value at 5% missingness (R^2 : 0.01–0.49), yielding lower accuracy than at 40% missingness (R^2 : 0.20–0.62). A similar reversal was observed across the 11 datasets with approximately one year of monitoring records, whereas it was not consistently reproduced in the three multi-year datasets (Fig. S8).

The inversion reveals a critical limitation of rate-based imputation evaluation. In imputation benchmarks the missingness rate is usually the single axis along which difficulty is varied, and performance is reported as a variance-referenced statistic such as R^2 [15,16,26]. Because the same pattern occurred for linear interpolation, which has no learned parameters, the inversion cannot be attributed to model training and instead reflects how the missingness scenario is constructed and evaluated. With fixed six-day blocks, a one-year record contains approximately three blocks at 5% missingness but approximately 24 blocks at 40%. At low rates, the evaluated positions are therefore concentrated within a few temporally correlated segments and may capture only part of the variability associated with seasonal, rainfall-driven, and operational changes in WWTP data [7,62]. Because R^2 is referenced to the variance of the masked evaluation subset rather than that of the complete record, adding blocks can raise R^2 partly by making that subset more representative rather than by improving reconstruction. This issue is less pronounced under non-consecutive masking, where evaluated positions are distributed throughout the record at every rate.

The record-length and block-length analyses supported this interpretation. Short records exhibited the inversion more consistently than multi-year records (Fig. 8a, Fig. S8), while longer blocks intensified the trend by reducing the number of independent blocks at a fixed missingness rate (Fig. S9). With three-day blocks, in contrast, no model produced a negative R^2 at any missingness rate. Although this behavior arose within the evaluation design, it is directly relevant to short operational records in which only a few consecutive gaps are available for assessment.

Under nRMSE the same comparison produced the expected trend. Error increased monotonically with the missingness rate for every model under both patterns and was lowest at 5% (Fig. 8, Fig. S10), the rate at which R^2 indicated the poorest performance. Reconstruction was therefore not more difficult at low rates; what differed was the quantity by which each metric is normalized. nRMSE is normalized by the range of the complete record, which is identical in every condition, whereas R^2 is normalized by the variance of the observations selected for evaluation, which the missingness condition itself determines. The structure of the gaps had a further effect: at a fixed rate of 5%, lengthening the gaps from 3 to 14 days raised nRMSE from 0.096 to 0.105 for SAITS and from 0.100 to 0.110 for Transformer (Fig. S9). Under consecutive gaps, a missingness rate paired with a single variance-referenced statistic is therefore insufficient, and at least one metric normalized to the complete record should be reported alongside it.

3.3.3. Model sensitivity to missingness rate and pattern

The nRMSE-based OLS analysis revealed model-specific responses to missingness pattern and rate (Table 3). Consecutive blocks remove temporally adjacent observations that support local reconstruction [63], whereas increasing the missingness rate reduces the overall amount of observed information [64]. These two dimensions of data loss produced distinct model-specific sensitivity profiles. Particularly, the interaction term was not significant for any model, indicating that the rate slope was essentially unchanged between patterns and that consecutive missingness acted as a shift in level rather than a change in the rate of degradation.

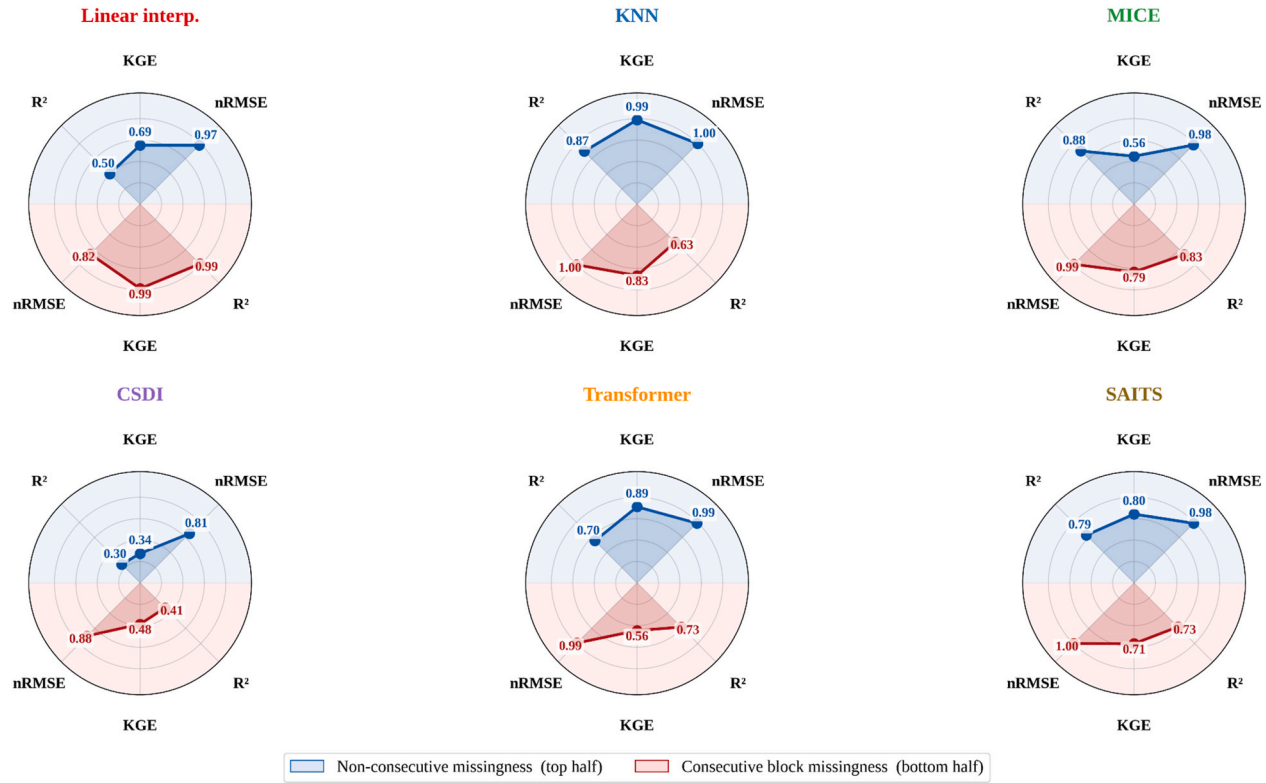


Fig. 6. Model performance under non-consecutive and consecutive block missingness. Radar plots display R², KGE, and nRMSE for each model, with blue (upper half) indicating non-consecutive patterns and red (lower half) indicating consecutive block patterns.

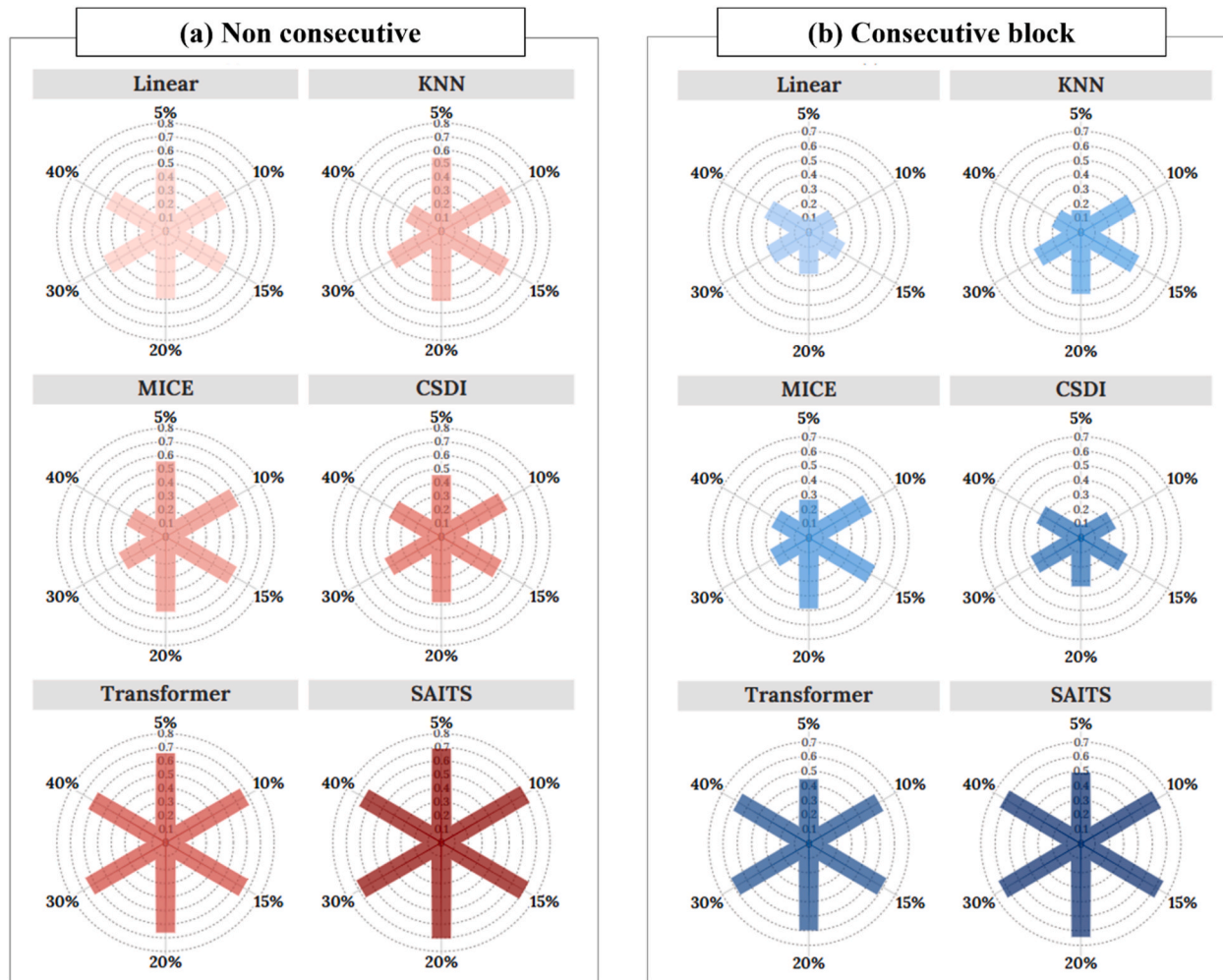


Fig. 7. Imputation performance (R^2) across varying missingness rates: (a)non-consecutive missingness; (b) consecutive block missingness.

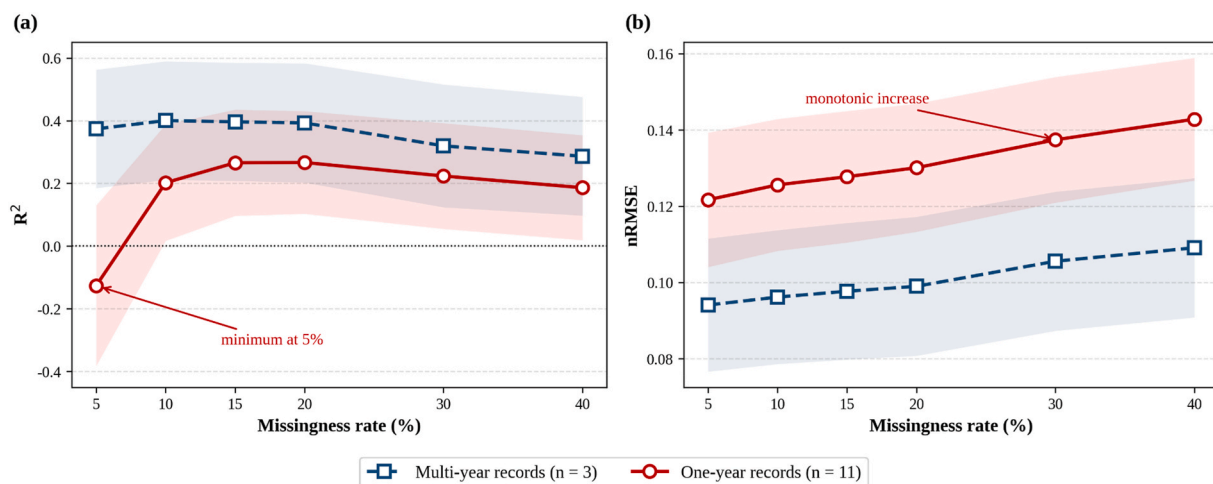


Fig. 8. Comparison of R^2 and nRMSE responses to missingness rate between one-year and multi-year records under consecutive block missingness.

Table 3

Model-specific effects of missingness pattern and rate on nRMSE estimated using separate OLS regressions.

Imputation model	$\Delta\text{Intercept}^a$	$\beta_{\text{non-consecutive}}^b$	$\Delta\beta$ (interaction) ^c	Rate effect over 5–40% ^d
Linear	−0.0189***	0.00027	0.00015	0.0095
KNN	−0.0076**	0.00031	0.00003	0.0109
MICE	−0.0076*	0.00036	0.00003	0.0126
CSDI	−0.0066	0.0004	−0.00006	0.0140
Transformer	−0.0049	0.00099	0.00003	0.0347
SAITS	−0.0020	0.00121	−0.00004	0.0424

*Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

^a $\Delta\text{Intercept}$ represents the estimated intercept difference between non-consecutive and consecutive missingness patterns, calculated as non-consecutive minus consecutive. For nRMSE, a negative value indicates lower error under non-consecutive missingness.

^b $\beta_{\text{non-consecutive}}$ denotes the change in nRMSE per one-percentage-point increase in missingness rate under the non-consecutive pattern.

^c $\Delta\beta$ represents the difference in the missingness-rate slope between consecutive and non-consecutive patterns, corresponding to the pattern-by-rate interaction.

^d Rate effect gives the change in nRMSE produced across the missingness-rate range examined ($\beta_{\text{non-consecutive}} \times 35$), expressed in the same units as $\Delta\text{Intercept}$ so that the two effects can be compared directly.

Linear interpolation, predicated on temporal continuity, proved exceptionally vulnerable to consecutive missingness ($\Delta\text{Intercept} = -0.0189$, $p < 0.001$). In contrast, its rate effect was the smallest of the six, 0.0095 over the 5–40% range, half the pattern effect (Table 3). This profile is consistent with its reliance on local temporal continuity: consecutive gaps remove the adjacent observations required for interpolation [54,65], whereas increasing the number of isolated missing points had a comparatively smaller effect.

KNN and MICE showed the opposite profile to linear interpolation and were the most sensitive to increasing missingness rate. Although KNN and MICE have shown competitive performance in previous studies [42], their strong rate sensitivity indicates reduced robustness as the amount of missing data increases. Raising the missingness rate across the 5–40% range increased nRMSE by 0.0347 for KNN and 0.0424 for MICE, the two largest increases among the models and about three times those of Transformer and SAITS, whereas the pattern shifted nRMSE by only 0.0049 and 0.0020 (Table 3). This sensitivity stems from increased missingness, which reduces the density of neighboring observations for KNN and constrains the data available for iterative estimation in MICE, exacerbating error propagation [36,56].

Transformer and SAITS alone showed low sensitivity to both dimensions, without being weighted toward either. The pattern shifted nRMSE by 0.0076 for both models, and raising the missingness rate across the 5–40% range added 0.0109 and 0.0126, respectively (Table 3). Self-attention architecture enables the capture of long-range temporal and multivariate dependencies [37,57]. This ability is consistent with their comparatively stable performance under increasing missingness rates and consecutive gaps. In contrast, CSDI showed no statistically distinguishable response to either missingness condition (Table 3), an inconsistency likely attributable to the significant performance variability observed across different facilities (Fig. 3).

The block-length sensitivity analysis supported the same leading-model ordering. SAITS achieved the lowest nRMSE across all nine combinations of block length (3, 6, and 14 days) and missingness rate (5%, 20%, and 40%), with Transformer ranking second throughout (Fig. S9). These rankings were unchanged across block lengths and were consistent with the leading positions reported in the main comparison (Table S9). Notably, SAITS remained the most robust model even under 14-day consecutive gaps, demonstrating

that its advantage persisted as the duration of missingness increased.

A limitation of this study is that the synthetic masking followed an MCAR mechanism (Section 2.3.1). In real-world, operational WWTP monitoring, missingness may instead be associated with sensor condition, maintenance, operating states, or extreme observations, corresponding more closely to either missing at random (MAR) or missing not at random (MNAR) mechanisms. Nevertheless, missingness was introduced across all monitored variables rather than a single target variable, reducing the multivariate information available for reconstruction and providing a relatively stringent test under the assumed MCAR setting. Future work should evaluate mechanism-dependent scenarios based on observed sensor failures and operational interruptions.

4. Conclusions

Across 14 full-scale WWTP datasets, six imputation models were compared under multiple missingness scenarios. Cross-plant dataset heterogeneity, within-plant differences in variable correlations, and consecutive block missingness limited model generalizability beyond what a single-facility evaluation can detect. Three principal conclusions emerged. First, cross-plant heterogeneity and within-plant variable asymmetries showed that imputation performance was governed not only by model choice, but also by dataset-specific conditions that bounded attainable accuracy. Second, statistical analyses confirmed that the missingness pattern determined the level of performance attained, while the missingness rate governed the change within a pattern. Rate-based evaluation alone is therefore insufficient to assess imputation reliability. Under consecutive block missingness the two metrics disagreed on which rate was more difficult to impute. Performance under consecutive gaps should therefore be reported under more than one metric, at least one of which is referenced to the complete record. Third, the comparison quantified model-specific sensitivity to these conditions. Linear interpolation was the most sensitive to the missingness pattern and the least to the rate while KNN and MICE showed the reverse, whereas Transformer and SAITS showed lower sensitivity to both and maintained relatively stable performance across heterogeneous datasets. The main contribution of this study is not the identification of a universally superior algorithm, but the quantification of dataset- and missingness-driven constraints across operational WWTPs. These findings caution against generalizing single-facility accuracy to other operational contexts and indicate that evaluations intended to inform model choice should resolve performance at the variable level. Future work should develop adaptive imputation strategies that account for plant-specific data structures and operational missingness mechanisms.

CRedit authorship contribution statement

Hyeonju Park: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft. **Seongbong Heo:** Methodology, Visualization, Writing – review & editing. **Sangchul Lee:** Methodology, Writing – review & editing. **Young Mo Kim:** Conceptualization, Project administration, Supervision, Writing – review & editing.

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.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25485864). This research was also supported by the Climate Resilient R&D Project for Water-Related Disaster Management through the Korea Environment Industry & Technology Institute (KEITI), funded by the Ministry of Climate, Energy and Environment (MCEE) (No. RS-2024-00332378), and by a Korea University Grant.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.wri.2026.100394>.

Data availability

Data will be made available on request.

References

- [1] M. Alvi, D. Batstone, C.K. Mbamba, P. Keymer, T. French, A. Ward, J. Dwyer, R. Cardell-Oliver, Deep learning in wastewater treatment: a critical review, *Water Res.* 245 (2023) 125449, <https://doi.org/10.1016/j.watres.2023.120518>.

- [2] M.W. Lee, S.H. Hong, H. Choi, J.H. Kim, D.S. Lee, J.M. Park, Real-time remote monitoring of small-scaled biological wastewater treatment plants by a multivariate statistical process control and neural network-based software sensors, *Process Biochemistry* 43 (2008) 1107–1113, <https://doi.org/10.1016/J.PROCBIO.2008.06.002>.
- [3] M. Bahramian, R.K. Dereli, W. Zhao, M. Giberti, E. Casey, Data to intelligence: the role of data-driven models in wastewater treatment, *Expert Syst. Appl.* 217 (2023), <https://doi.org/10.1016/j.eswa.2022.119453>. Elsevier Ltd.
- [4] D.J. Dürrenmatt, W. Gujer, Data-driven modeling approaches to support wastewater treatment plant operation, *Environ. Modell. Software* 30 (2012) 47–56, <https://doi.org/10.1016/j.envsoft.2011.11.007>.
- [5] D. Jiang, H. Yang, H. Cao, D. Xu, A missing data imputation method for industrial soft sensor modeling, *J. Process Control* 152 (2025) 103485, <https://doi.org/10.1016/J.JPROCONT.2025.103485>.
- [6] H. Han, M. Sun, H. Han, X. Wu, J. Qiao, Univariate imputation method for recovering missing data in wastewater treatment process, *Chin. J. Chem. Eng.* 53 (2023) 201–210, <https://doi.org/10.1016/j.cjche.2022.01.033>.
- [7] K.B. Newhart, R.W. Holloway, A.S. Hering, T.Y. Cath, Data-driven performance analyses of wastewater treatment plants: a review, *Water Res.* 157 (2019) 498–513, <https://doi.org/10.1016/j.watres.2019.03.030>. Elsevier Ltd.
- [8] Y. Peng, E. Li, Basis expansion imputation network for incomplete multivariate sensor data in wastewater treatment system, *J. Water Proc. Eng.* 63 (2024) 105536, <https://doi.org/10.1016/J.JWPE.2024.105536>.
- [9] A. Rácz, A. Gere, Comparison of missing value imputation tools for machine learning models based on product development cases studies, *LWT (Lebensm.-Wiss. & Technol.)* 221 (2025) 117585, <https://doi.org/10.1016/j.lwt.2025.117585>.
- [10] Y. Tashiro, J. Song, Y. Song, S. Ermon, CSDI: conditional score-based diffusion models for probabilistic time series imputation, *Adv. Neural Inf. Process. Syst.* 30 (2021) 24804–24816.
- [11] J. Yoon, J. Jordon, M. Van Der Schaar, GAIN: Missing Data Imputation Using Generative Adversarial Nets, 2018.
- [12] Y.F. Zhang, P.J. Thorburn, W. Xiang, P. Fitch, SSIM - a deep learning approach for recovering missing time series sensor data, *IEEE Internet Things J.* 6 (2019) 6618–6628, <https://doi.org/10.1109/JIOT.2019.2909038>.
- [13] I.P. Banjara, S. Poudel, K. Pariyar, D. Upreti, A. Zafeirakou, S.R. Paudel, AI for enhanced water quality data imputation: a deep learning perspective, *Environ. Sci. Adv.* 4 (6) (2025) 817–823, <https://doi.org/10.1039/d4va00367e>. Royal Society of Chemistry.
- [14] Y. Sun, J. Li, Y. Xu, T. Zhang, X. Wang, Deep learning versus conventional methods for missing data imputation: a review and comparative study, *Expert Syst. Appl.* 227 (2023) 120201, <https://doi.org/10.1016/j.eswa.2023.120201>.
- [15] J. Wang, W. Du, Y. Yang, L. Qian, W. Cao, K. Zhang, W. Wang, Y. Liang, Q. Wen, Deep learning for multivariate time series imputation: a survey. <http://arxiv.org/abs/2402.04059>, 2024.
- [16] M. Alwateer, E.S. Atlam, M.M.A. El-Raouf, O.A. Ghoneim, I. Gad, Missing data imputation: a comprehensive review, *J. Comput. Commun.* 12 (2024) 53–75, <https://doi.org/10.4236/jcc.2024.1211004>.
- [17] L. Benedetti, D. Bixio, P.A. Vanrolleghem, Benchmarking of WWTP design by assessing costs, effluent quality and process variability, *Water Sci. Technol.* 54 (2006) 95–102, <https://doi.org/10.2166/wst.2006.809>.
- [18] S.J. Judd, The status of industrial and municipal effluent treatment with membrane bioreactor technology, *Chem. Eng. J.* 305 (2016) 37–45, <https://doi.org/10.1016/j.ccej.2015.08.141>.
- [19] C. Kazadi Mbamba, P. Keymer, M. Alvi, S.O.N. Topalian, F. Ud Din, D.J. Batstone, Enhancing data quality in wastewater processes: missing data imputation with deep variational autoencoders and genetic algorithms, *Comput. Chem. Eng.* 199 (2025) 109123, <https://doi.org/10.1016/j.compchemeng.2025.109123>.
- [20] Y. Liu, G. Mancuso, L. Petrotto, S. Lavrnić, Z. Dong, Y. Tian, J. Zhang, A. Toscano, AI-driven solutions in wastewater treatment and agricultural reuse systems: a comprehensive review, *J. Environ. Manag.* 393 (2025) 127008, <https://doi.org/10.1016/j.jenvman.2025.127008>.
- [21] A.H. Ba-Alawi, J. Loy-Benitez, S.Y. Kim, C.K. Yoo, Missing data imputation and sensor self-validation towards a sustainable operation of wastewater treatment plants via deep variational residual autoencoders, *Chemosphere* 288 (2022), <https://doi.org/10.1016/j.chemosphere.2021.132647>.
- [22] H. Han, M. Li, J. Qiao, Q. Yang, Y. Peng, Filter transfer learning algorithm for missing data imputation in wastewater treatment process, *IEEE Trans. Knowl. Data Eng.* 35 (2023) 12649–12662, <https://doi.org/10.1109/TKDE.2023.3270118>.
- [23] M. Sun, F. Li, H. Han, Attention-guided position-sensitive multiple imputation for wastewater treatment process, *IEEE Trans. Ind. Inf.* 20 (2024) 14459–14468, <https://doi.org/10.1109/TII.2024.3452282>.
- [24] H.G. Han, M.T. Sun, X.L. Wu, F.Y. Li, Double-cycle weighted imputation method for wastewater treatment process data with multiple missing patterns, *Sci. China Technol. Sci.* 65 (2022) 2967–2978, <https://doi.org/10.1007/s11431-022-2163-1>.
- [25] J. He, Y. Li, Y. Zhao, K. Wang, A deep learning method for missing data imputation in wastewater treatment processes under dynamic weather conditions, *J. Water Proc. Eng.* 76 (2025), <https://doi.org/10.1016/j.jwpe.2025.108203>.
- [26] W. Du, J. Wang, L. Qian, Y. Yang, Z. Ibrahim, F. Liu, Z. Wang, H. Liu, Z. Zhao, Y. Zhou, W. Wang, K. Ding, Y. Liang, B.A. Prakash, Q. Wen, TSI-Bench: Benchmarking Time Series Imputation, 2024. <http://arxiv.org/abs/2406.12747>.
- [27] C. Zhang, X. Fu, D. Qiu, H. Badihi, H. Gu, Robust imputation of missing photovoltaic power data using a Weather- and context-aware hybrid transformer framework, *Renew. Energy* (2025) 124576, <https://doi.org/10.1016/J.RENENE.2025.124576>.
- [28] M. Kazijevs, M.D. Samad, Deep imputation of missing values in time series health data: a review with benchmarking, *J. Biomed. Inf.* 144 (2023) 104440, <https://doi.org/10.1016/J.JBI.2023.104440>.
- [29] Y. Luo, Evaluating the state of the art in missing data imputation for clinical data, *Briefings Bioinf.* 23 (2022), <https://doi.org/10.1093/bib/bbab489>.
- [30] C.C. Aggarwal, *Data mining: the textbook 1*, Springer., New York, 2015.
- [31] S. Aerts, G. Haesbroeck, C. Ruwet, Multivariate coefficients of variation: Comparison and influence functions, *J. Multivariate Anal.* 142 (2015) 183–198, <https://doi.org/10.1016/j.jmva.2015.08.006>.
- [32] K.V. Mardia, Measures of multivariate skewness and kurtosis with applications, *Biometrika* 57 (1970) 519–530, <https://doi.org/10.1093/biomet/57.3.519>.
- [33] Q. Zhang, C. Zhang, L. Cui, X. Han, Y. Jin, G. Xiang, Y. Shi, A method for measuring similarity of time series based on series decomposition and dynamic time warping, *Appl. Intell.* 53 (6 53) (2022) 6448–6463, <https://doi.org/10.1007/s10489-022-03716-9>.
- [34] M. Von Sperling, M.E. Verbyla, S.M. Oliveira, *Assessment of Treatment Plant Performance and Water Quality Data: a Guide for Students, Researchers and Practitioners*, IWA Publishing, 2020.
- [35] D.B. Rubin, Inference and missing data, *Biometrika* 63 (1976) 581–592, <https://doi.org/10.1093/biomet/63.3.581>.
- [36] S. van Buuren, K. Groothuis-Oudshoorn, Mice: multivariate imputation by chained equations in R, *J. Stat. Software* 45 (3) (2011) 1–67, <https://doi.org/10.18637/jss.v045.i03>.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you Need, 2017. <http://arxiv.org/abs/1706.03762>.
- [38] W. Du, D. Côté, Y. Liu, SAITS: self-attention-based imputation for time series, *Expert Syst. Appl.* 219 (2023), <https://doi.org/10.1016/j.eswa.2023.119619>.
- [39] W. Du, P. Research, Y. Yang, L. Qian, J. Wang, Q. Wen, PyPOTS: a python toolkit for machine learning on partially-observed time series, *J. Mach. Learn. Res.* 23 (2023) 1–7. <https://arxiv.org/pdf/2305.18811>.
- [40] H.V. Gupta, H. Kling, K.K. Yilmaz, G.F. Martinez, Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling, *J. Hydrol.* 377 (2009) 80–91, <https://doi.org/10.1016/J.JHYDROL.2009.08.003>.
- [41] P. Legendre, *Coefficient of Concordance, Encyclopedia of Research Design*, vol. 1, 2010, pp. 164–169.
- [42] L.E. Alejo-Sanchez, A. Márquez-Grajales, F. Salas-Martínez, A. Franco-Arcega, V. López-Morales, O.A. Acevedo-Sandoval, C.A. González-Ramírez, R. Villegas-Vega, Missing data imputation of climate time series: a review, *MethodsX* 15 (2025) 103455, <https://doi.org/10.1016/J.MEX.2025.103455>.
- [43] M. Li, Y. Wu, Y. Jiang, L. Peng, Y. Jiang, G. Huang, J. Zhang, Enhancing wind speed data imputation accuracy in wind resource assessment: comparative study of emerging multivariate time series neural network models, *Int. J. Green Energy* (2026), <https://doi.org/10.1080/15435075.2026.2619437>.

- [44] X. Zhang, P. Zhou, A transferred spatio-temporal deep model based on multi-LSTM auto-encoder for air pollution time series missing value imputation, *Future Gener. Comput. Syst.* 156 (2024) 325–338, <https://doi.org/10.1016/j.future.2024.03.015>.
- [45] Z. Chen, H. Xu, P. Jiang, S. Yu, G. Lin, I. Bychkov, A. Hmel'nov, G. Ruzhnikov, N. Zhu, Z. Liu, A transfer learning-based LSTM strategy for imputing large-scale consecutive missing data and its application in a water quality prediction system, *J. Hydrol.* 602 (2021), <https://doi.org/10.1016/j.jhydrol.2021.126573>.
- [46] M. Wang, K. Yang, J. He, J. Qi, Y. Chen, H. Gu, R. Liu, S. Wu, Z. Du, Enhanced buoy-based water quality time series imputation in coastal waters, *J. Hydrol.* 663 (2025) 134071, <https://doi.org/10.1016/J.JHYDROL.2025.134071>.
- [47] A.E.C.O.M. Metcalf & Eddy, *Wastewater Engineering Treatment and Resource Recovery*, McGraw-Hill Education, 2014.
- [48] J.T. Bunce, D.W. Graham, A simple approach to predicting the reliability of small wastewater treatment plants, *Water* 11 (2019) 11, <https://doi.org/10.3390/W11112397>, 2019.
- [49] J. Parkas, F. Polesel, M. Kjos, P.A. Carvalho, T. Ciesielski, X. Flores-Alsina, S.F. Hansen, A.M. Booth, Monitoring and modelling of influent patterns, phase distribution and removal of 20 elements in two primary wastewater treatment plants in Norway, *Sci. Total Environ.* 725 (2020) 138420, <https://doi.org/10.1016/J.SCIOTOTENV.2020.138420>.
- [50] S. Platikanov, S. Rodriguez-Mozaz, B. Huerta, D. Barceló, J. Cros, M. Batle, G. Poch, R. Tauler, Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV absorption measurements, *J. Environ. Manag.* 140 (2014) 33–44, <https://doi.org/10.1016/j.jenvman.2014.03.006>.
- [51] M. Majewsky, T. Gallé, M. Bayerle, R. Goel, K. Fischer, P.A. Vanrolleghem, Xenobiotic removal efficiencies in wastewater treatment plants: residence time distributions as a guiding principle for sampling strategies, *Water Res.* 45 (2011) 6152–6162, <https://doi.org/10.1016/j.watres.2011.09.005>.
- [52] M. Ahnert, V. Kuehn, P. Krebs, Temperature as an alternative tracer for the determination of the mixing characteristics in wastewater treatment plants, *Water Res.* 44 (2010) 1765–1776, <https://doi.org/10.1016/j.watres.2009.11.049>.
- [53] M. Lu, Z.H. Li, Y. Jiang, Effluent temporal collective behaviors of a wastewater treatment plant community, *Sci. Total Environ.* 787 (2021) 147694, <https://doi.org/10.1016/J.SCIOTOTENV.2021.147694>.
- [54] M. Lepot, J.B. Aubin, F.H.L.R. Clemens, Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment, *Water* 9 (2017) 9, <https://doi.org/10.3390/w9100796>.
- [55] S. Moritz, T. Bartz-Beielstein, imputeTS: time series missing value imputation in R, *RELC J.* 9 (1) (2017) 207–218, <https://doi.org/10.32614/rj-2017-009>.
- [56] A. Jadhav, D. Pramod, K. Ramanathan, Comparison of performance of data imputation methods for numeric dataset, *Appl. Artif. Intell.* 33 (2019) 913–933, <https://doi.org/10.1080/08839514.2019.1637138>.
- [57] J. Zhao, C. Rong, C. Lin, X. Dang, Multivariate time series data imputation using attention-based mechanism, *Neurocomputing* 542 (2023) 126238, <https://doi.org/10.1016/j.neucom.2023.126238>.
- [58] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Adv. Neural Inf. Process. Syst.* 33 (2020) 6840–6851.
- [59] T.P. Mesquita, D.M.P.F. Silva, A.M.N.C. Ribeiro, I.R.R. Silva, C.J.A. Bastos-Filho, R.P. Monteiro, A comparative analysis of deep learning-based methods for multivariate time series imputation with varying missing rates, *ETCM 2024 - 8th Ecuador Technical Chapters Meeting 1* (2024), <https://doi.org/10.1109/ETCM63562.2024.10746216>.
- [60] R. Wu, S.D. Hamshaw, L. Yang, D.W. Kincaid, R. Etheridge, A. Ghasemkhani, Data imputation for multivariate time series sensor data with large gaps of missing data, *IEEE Sens. J.* 22 (2022) 10671–10683, <https://doi.org/10.1109/JSEN.2022.3166643>.
- [61] T.T.H. Phan, A. Bigand, P. Caillault, A new fuzzy logic-based similarity measure applied to large gap imputation for uncorrelated multivariate time series, *Applied Computational Intelligence and Soft Computing* 2018 (2018), <https://doi.org/10.1155/2018/9095683>.
- [62] J. Kinnunen, P.M. Rossi, I. Herrmann, A.K. Ronkanen, E. Heiderscheidt, Factors affecting effluent quality in on-site wastewater treatment systems in the cold climates of Finland and Sweden, *J. Clean. Prod.* 404 (2023) 136756, <https://doi.org/10.1016/J.JCLEPRO.2023.136756>.
- [63] Y. Wang, X. Xu, L. Hu, J. Fan, M. Han, A time series continuous missing values imputation method based on generative adversarial networks, *Knowl. Base Syst.* 283 (2024) 111215, <https://doi.org/10.1016/J.KNOSYS.2023.111215>.
- [64] J. Choi, K.J. Lim, B. Ji, Robust imputation method with context-aware voting ensemble model for management of water-quality data, *Water Res.* 243 (2023), <https://doi.org/10.1016/j.watres.2023.120369>.
- [65] X. Che, H.K. Zhang, Z.B. Li, Y. Wang, Q. Sun, D. Luo, H. Wang, Linearly interpolating missing values in time series helps little for land cover classification using recurrent or attention networks, *ISPRS J. Photogrammetry Remote Sens.* 212 (2024) 73–95, <https://doi.org/10.1016/J.ISPRSJPRS.2024.04.021>.