AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2025WR040528

Leveraging Spatiotemporal Redundancy for Sensor Data Imputation in Water Distribution Networks

Ang Xu[1], Avi Ostfeld[2], Yu Shao[1], Tuqiao Zhang[1], Shipeng Chu[1], Yu Tian[3], and Dewu Jian[4]

[1]College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, China, [2]Civil and Environmental Engineering Technion, Israel Institute of Technology, Haifa, Israel, [3]China Institute of Water Resources and Hydropower Research, Beijing, China, [4]Central & Southern China Municipal Engineering Design & Research General Institute, Co., Ltd., Wuhan, China

Key Points:

- Spatiotemporal water distribution network data exhibit intrinsic redundancy with low-rank algebraic structure
- A theoretical framework including complex missing patterns and low-rank prior-based approaches for sensor data imputation is developed
- Experimental results on real-world data sets demonstrate the effectiveness of the presented imputation method in diverse missing scenarios

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Chu,
chushipeng@zju.edu.cn

Citation:

Xu, A., Ostfeld, A., Shao, Y., Zhang, T., Chu, S., Tian, Y., & Jian, D. (2025). Leveraging spatiotemporal redundancy for sensor data imputation in Water Distribution Networks. *Water Resources Research*, *61*, e2025WR040528. https://doi.org/10.1029/2025WR040528

Received 18 MAR 2025
Accepted 8 SEP 2025

Author Contributions:

Conceptualization: Ang Xu, Shipeng Chu
Funding acquisition: Avi Ostfeld, Yu Shao, Shipeng Chu
Investigation: Shipeng Chu
Methodology: Ang Xu
Project administration: Tuqiao Zhang
Resources: Yu Shao, Tuqiao Zhang, Yu Tian, Dewu Jian
Supervision: Shipeng Chu
Writing – review & editing: Ang Xu, Avi Ostfeld, Shipeng Chu

Abstract The rapid digital transformation of Water Distribution Networks (WDNs) has led to the collection of multi-sensor time series with high temporal and spatial resolution. However, missing data poses a significant challenge, undermining the usability and effectiveness of data-driven applications. Performing missing data imputation is essential to enhance data quality and support intelligent management. This study first reveals that WDN sensor data in tensor form inherently exhibit spatiotemporal redundancy across three dimensions: inter-sensor similarity, intra-day regularity, and daily recurrence. The redundancy can be algebraically characterized by the low-rank structure of WDN tensor data, providing a robust foundation for imputation. Based on these findings, a novel Low-rank Autoregressive Tensor Completion (LATC) approach is proposed to efficiently impute spatiotemporal WDN data. The LATC combines autoregressive regularization with standard low-rank tensor completion, effectively capturing both global redundancy and local correlation of multi-sensor WDN data. Finally, the LATC is validated on four real-world and simulated WDN data sets under eight different missing scenarios. Extensive experiments show that the LATC significantly outperforms state-of-the-art baseline methods, achieving accurate imputation even under severe corruption and complex missing patterns.

Plain Language Summary With the rapid digital transformation of Water Distribution Networks (WDNs), there is increasingly more large-scale data being collected with fine temporal resolution and high sensor coverage. However, the inevitable missing data issue makes numerous data-driven intelligent applications suffer from incorrect responses. Therefore, performing reliable imputations on spatiotemporal WDN data has become an essential step before further applications. The present study aims to fill this knowledge gap by proposing a theoretical framework and a methodological approach. First, it is revealed that WDN tensor data exhibit inherent redundancy in all modes, including inter-sensor similarity, intra-day regularity, and daily recurrence. This spatiotemporal redundancy can be algebraically characterized by low-rank, enabling the use of low-rank prior-based approaches in imputation tasks. Building on this framework, a general low-rank autoregressive tensor completion approach is presented for efficient WDN data imputation. Finally, extensive experiments on real-world WDN data sets prove the effectiveness and superiority of the proposed approach, particularly in scenarios involving severe structural corruption and complex missing patterns. This study is the first to comprehensively address issues concerning spatiotemporal data imputation in real-world WDNs and is expected to serve as a starting point for further exploration.

1. Introduction

Water Distribution Networks (WDNs) serve as an essential urban infrastructure and are accountable for the reliable and efficient supply of clean water to residential and industrial users (Zhou et al., 2022). The continuous expansion of urbanization (Sanchez et al., 2020) and rapid population growth (Sivagurunathan et al., 2022) have significantly increased water resource consumption and placed higher loads on WDNs, which urge the implementation of digital transformation (Boyle et al., 2022) to enhance water supply management. WDN digitization is accelerating due to remarkable advances in cost-effective data acquisition and wireless data transfer technologies (Eggimann et al., 2017; Oberascher et al., 2022). Therefore, a growing number of multi-sensor observations (e.g., pressure, flow, water quality) are collected with fine temporal resolution and high spatial sensor coverage.

These spatiotemporal time series, which reflect the underlying states and dynamics of WDNs, establish the foundation for a variety of downstream tasks and decision-making processes in Smart Water Networks (SWNs)

(Eggimann et al., 2017; Fu et al., 2022), such as hydraulic model calibration (Chu et al., 2020; Zhou et al., 2023), anomaly detection (Z. Hu et al., 2022; Zhou et al., 2019), demand forecasting (Salloom et al., 2021; Zanfei, Brentan et al., 2022), optimal scheduling (Hajgató et al., 2020; Salomons & Housh, 2020), to name but a few. However, in real-world WDNs, the problem of missing values is virtually ubiquitous and inevitable (Osman et al., 2018) for various reasons, including sensor malfunction, transmission failure, insufficient sampling, and human damage. This problem undermines the usability and effectiveness of data-driven applications that rely on complete data entries or time series (Zanfei, Menapace et al., 2022). For example, the presence of missing values or outliers greatly diminishes the accuracy and robustness of real-time hydraulic models, resulting in calibrated parameters that fail to reflect the actual state of the WDNs (Chu et al., 2021). Furthermore, in short-term water demand forecasting, the reliability of predicted results is highly sensitive to data quality, particularly when forecasting for numerous sensors simultaneously (Wu et al., 2023). Incomplete data inputs can hinder the model's ability to extract input features, causing substantial output errors. Thus, accurately estimating and filling in missing values—i.e., imputation—on spatiotemporal WDN data sets is vital for the success of SWNs. However, this task remains challenging due to complex missing mechanisms and varying degrees of missingness.

There are several approaches to dealing with missing values in the WDN domain, ranging from simple deletion techniques to complex imputation algorithms. Osman et al. (2018) developed a "top-down bottom-up" two-pronged method to select the appropriate imputation model based on the category and percentage of missing data. Indeed, such stepwise decisions and manual model selection introduce considerable inefficiency and often result in unstable performance when applied to large-scale WDN data sets. To obtain a robust calibration of the nodal water demand in WDN, Chu et al. (2021) offset the absence of data and outliers by the predicted value from historical measurements in real-time. However, as noted by the authors, the relatively simple nature of the prediction algorithm limits its validity in handling continuous missing time events. Moreover, Zanfei, Menapace et al. (2022) analyzed diverse conventional imputation methods, including statistical and machine learning algorithms, to tackle missing water demand data. It was found that catching intra-day seasonality is crucial for estimating incomplete time series with large gaps. Nevertheless, these methods did not achieve acceptable results because of the high variability in real-world WDN time series.

Overall, with the development of WDN digitization, missing value imputation has become an essential step for fully leveraging the data, attracting great interest in recent years. However, the study on the imputation approach is still in its infancy. These remaining open issues can be summarized in three key points. First, existing works have not yet investigated the complicated missing patterns in real-world WDN data. Beyond the well-known Random Missing (RM), the spatiotemporal data suffer from considerable structural corruption, especially the condition of a sensor losing observations for one day and all sensors losing observations over several consecutive time points, which is fatal to the regular operation of SWNs. Second, existing works almost impute missing values for individual sensors or in a specific application, requiring repeated model matching and parameter tuning due to limited generalizability across diverse sensors and missing patterns. When applied to large-scale WDN data containing hundreds of sensors, this process becomes both computationally expensive and operationally burdensome. The desired imputation model for SWNs will impute missing values quickly and precisely, thereby supporting all applications following real-time data acquisition without requiring custom imputation in each application. Third, existing works have overlooked the abundant correlations and dependencies across the temporal and spatial dimensions that can be exploited to better estimate missing values. For instance, sensor time series typically show strong global patterns associated with intrinsic daily recurrence, and observations collected from neighboring sensors over short periods exhibit similar trends.

This study aims to comprehensively address these issues concerning spatiotemporal WDN data imputation. Specifically, three principal missing patterns are summarized to characterize the features of observation loss caused by different mechanisms. An in-depth analysis of WDN data characteristics reveals the inherent global redundancy (or algebraical low-rankness) and local correlation of multi-sensor time series organized as a third-order tensor (sensors $\times$ intervals $\times$ days). This spatiotemporal redundancy manifests itself in all tensor modes, including inter-sensor similarity, intra-day regularity, and daily recurrence. It provides a theoretical ground for missing value imputation in real-world WDNs through Low-Rank Tensor Completion (LRTC). Based on these properties, a general approach named Low-rank Autoregressive Tensor Completion (LATC) is developed for accurate and effective imputation under diverse missing scenarios. By combining the autoregressive processes with the truncated tensor Nuclear Norm (NN)-based model, LATC can better capture the global low-rank structure and local temporal continuity underlying spatiotemporal WDN data. Extensive experiments on real-

world data sets demonstrate the effectiveness and superiority of LATC for large-scale WDN data imputation, outperforming several state-of-the-art baseline models, especially in scenarios with severe structural corruption and complicated missing patterns. Additionally, the results of parameter analysis and ablation study confirm the importance of truncation operation in characterizing spatiotemporal redundancy and of autoregressive regularization in ensuring temporal continuity.

The remainder of this study is structured as follows. Section 2 describes the theoretical and methodological framework for spatiotemporal WDN data imputation. Sections 3 and 4 present the experimental setup and corresponding results, respectively. Section 5 explains the mechanisms of LATC's two key components in modeling WDN data and explores its other potential applications in SWNs. Finally, Section 6 concludes the study and suggests directions for future research.

2. Methodology

2.1. Notations

To facilitate the following study, some mathematical notations outlined by Kolda and Bader (2009) are first introduced. Matrices are indicated with bold uppercase letters (e.g., $\boldsymbol{X} \in \mathbb{R}^{M\times N}$), vectors with bold lowercase letters (e.g., $\boldsymbol{x} \in \mathbb{R}^{M}$), and scalars with lowercase letters (e.g., x). Given a matrix $\boldsymbol{X} \in \mathbb{R}^{M\times N}$, its (m,n)th element is denoted by $x_{m,n}$ and its Frobenius norm is defined as $\|\boldsymbol{X}\|_F = \sqrt{\sum_{m,n} x_{m,n}^2}$. The singular value decomposition is defined as $\boldsymbol{X} = \boldsymbol{U}\sum \boldsymbol{V}^\top$, where $\sum(\boldsymbol{X})$ is a vector containing the singular values of $\boldsymbol{X}$ in descending order and $\sigma_i(\boldsymbol{X})$ represents ith largest singular value.

When expanding to a third-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{M\times I\times J}$, its (m,i,j)th element is denoted by $x_{m,i,j}$, and the inner product with another tensor of identical size is calculated by $\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{m,i,j} x_{m,i,j}\, y_{m,i,j}$. The Frobenius norm on a tensor is defined as $\|\boldsymbol{\mathcal{X}}\|_F = \sqrt{\sum_{m,i,j} x_{m,i,j}^2}$. The kth-mode $(k = 1,2,3)$ unfolding of $\boldsymbol{\mathcal{X}}$ is denoted by $\boldsymbol{\mathcal{X}}_{(k)}$ (i.e., $\boldsymbol{\mathcal{X}}_{(1)} \in \mathbb{R}^{M\times(IJ)}$, $\boldsymbol{\mathcal{X}}_{(2)} \in \mathbb{R}^{I\times(MJ)}$, $\boldsymbol{\mathcal{X}}_{(3)} \in \mathbb{R}^{J\times(MI)}$). Here, $IJ = I \times J$, $MJ = M \times J$, and $MI = M \times I$ are introduced for notational simplicity. Correspondingly, the folding operator $\text{fold}_k(\cdot)$ converts the kth-mode unfolding matrix to the origin tensor; in this way, the folding operator satisfies $\text{fold}_k\left(\boldsymbol{\mathcal{X}}_{(k)}\right) = \boldsymbol{\mathcal{X}}$. For ease of understanding, the 1-mode, 2-mode, and 3-mode unfolding are referred to as the "sensors" mode, "intervals" mode, and "days" mode unfolding, respectively, in the following.

2.2. Problem Statement

In a real-world WDN, the spatiotemporal time series data set collected from M sensors over IJ consecutive time points is organized as a matrix $\boldsymbol{D}$:

$$\boldsymbol{D} = \begin{bmatrix} d_{1,1} & \cdots & d_{1,IJ} \\ \vdots & \ddots & \vdots \\ d_{M,1} & \cdots & d_{M,IJ} \end{bmatrix} \in \mathbb{R}^{M\times(IJ)}, \tag{1}$$

where the rows correspond to individual sensors and the columns represent time points. In particular, I denotes the time points per day, and J denotes the total number of days. Since $\boldsymbol{D}$ frequently has substantial structural corruption and missing values, the partially observed matrix is denoted as $\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D})$. Generally, outliers can also be treated as missing data and subsequently imputed. The operator $\mathcal{P}_{\boldsymbol{\Omega}} : \mathbb{R}^{M\times(IJ)} \mapsto \mathbb{R}^{M\times(IJ)}$ represents a projection that retains the elements in the observation index set $\boldsymbol{\Omega}$ and sets all others to zero:

$$[\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D})]_{m,n} = \begin{cases} d_{m,n}, \text{if } (m,n) \in \boldsymbol{\Omega}, \\ 0, \text{otherwise}, \end{cases} \tag{2}$$

where $[\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D})]_{m,n}$ indicates its (m,n)th elements with $m = 1,\ldots,M$ and $n = 1,\ldots,IJ$.

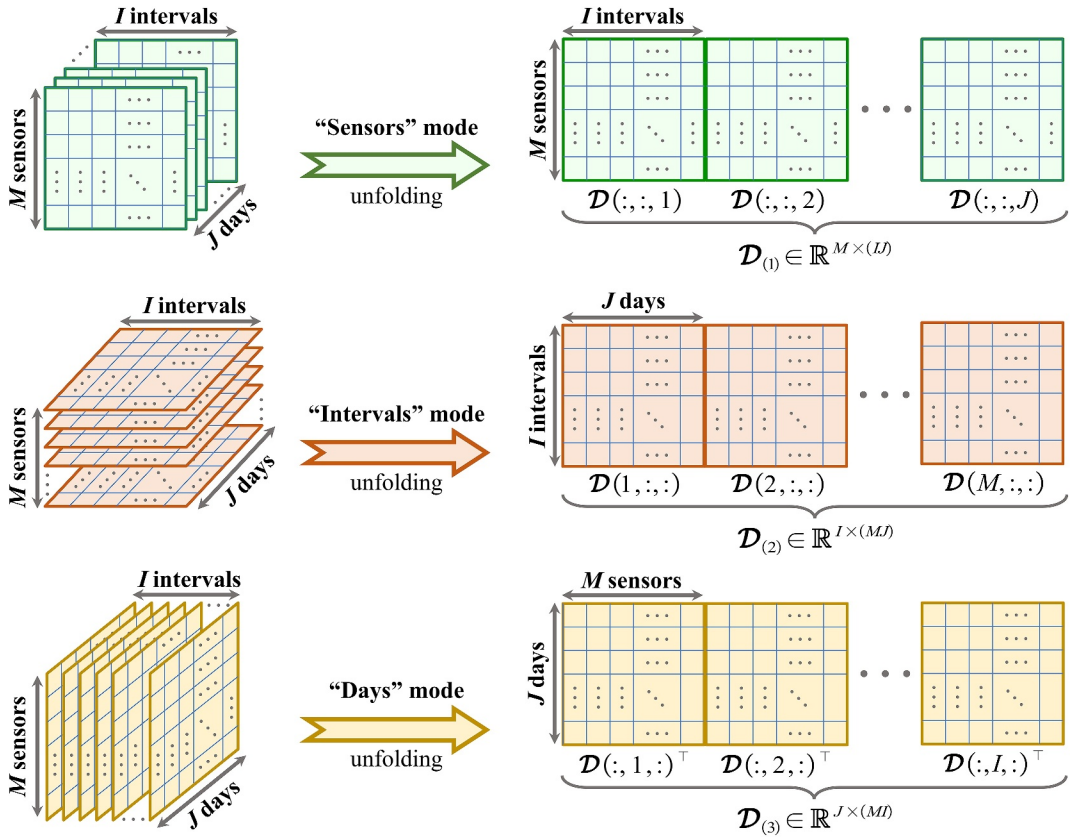

Figure 1. Illustration of unfolding Water Distribution Network tensor data.

The main idea of spatiotemporal WDN data imputation is to extract higher-order correlations and dependencies from the partial observations, which can be further leveraged to estimate the missing entries. Given the noticeable day-to-day patterns and similarity in WDN time series, the observation matrix $\boldsymbol{D}$ (sensors × time points) is transformed into a third-order tensor $\boldsymbol{\mathcal{D}}$ (sensors × intervals × days) by introducing an additional "day" dimension. As depicted in Figure 1, the tensorization step enables the comprehensive capture of global information across all three modes, including the observation matrix (i.e., the "sensors" mode unfolding).

2.3. Complicated Missing Patterns

An in-depth investigation of the missing patterns and mechanisms in spatiotemporal WDN data is necessary for developing a practical imputation approach. Rubin (1975) introduced a classification system, still widely used today, that categorizes missing values according to the relationships between measured variables and the likelihood of missingness: missing completely at random, missing at random, and missing not at random. This system has been applied on several WDN data imputation studies (Osman et al., 2018; Zanfei, Menapace et al., 2022), requiring strict assumptions, and focusing more on the nature of the missing data rather than on imputation efficiency.

To this end, the missing patterns of spatiotemporal WDN data are classified into three categories according to the features of observation loss caused by different mechanisms in real-world data sets, as illustrated in Figure 2. The first category is RM. Power supply fluctuations and packet loss may give rise to cases of random data loss, where individual data points are missing unpredictably during the acquisition and transmission process. The second category is Long-range Missing (LM). In the context of sensor operation, malfunctions and routine maintenance can introduce non-random data gaps that compromise the observation integrity over extended periods. Block Missing (BM) is the third category, which presents a steep challenge due to the unavailability of all sensor data

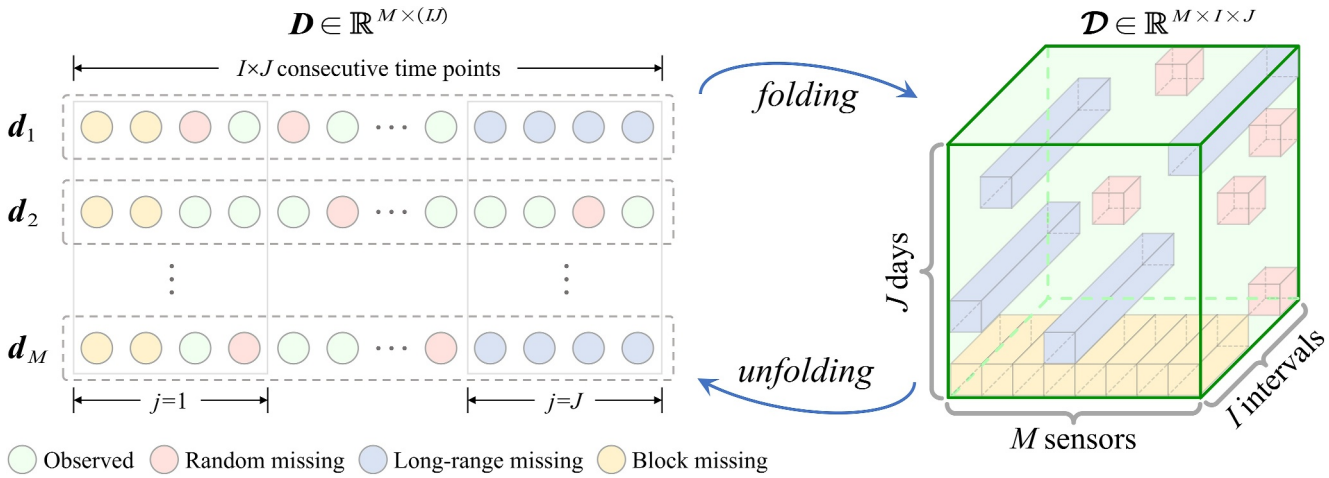

Figure 2. Visualization of three missing patterns in spatiotemporal Water Distribution Network data, presented in matrix and tensor form.

within a specific time window. Such scenarios emerge when software or hardware failures affect the shared SCADA system.

Notably, this is the first time that LM and BM have been clearly defined in WDN data analysis, reflecting structural corruption along the temporal and spatial dimensions. Adequate attention should be directed toward these patterns, as their uncertainty and risk potentially disrupt many data-driven applications. Therefore, a general imputation approach for SWNs must be capable of handling missing data with complicated patterns.

2.4. Spatiotemporal Data Redundancy

2.4.1. Data Redundancy Analysis

The tensorization step is a powerful tool that allows each unfolding mode to be analyzed separately, potentially revealing similarities or redundancies that may not be immediately perceived in matrix form. To demonstrate the inherent multi-mode data redundancies in spatiotemporal WDN data, visualizations using the real-world "Z-city flow" data set (see Section 3.1 for details) are presented in Figure 3.

Inter-sensor similarity: Figure 3a displays the time series curves of five flow sensors over a one-day period. Despite being positioned at different locations in WDN, all sensors exhibit similar patterns of variation in their readings due to the interconnected topology of the pipe network. This sensor similarity highlights the spatial redundancy in the multi-sensor observation data.

Intra-day regularity: Focusing on sensor #5 from Figures 3a and 3c illustrates its observations at six consecutive time points over a 1-month period. The curves show a consistent decline from 22:15 to 23:30 each day, aligning with typical residential water consumption patterns. This intra-day regularity indicates data redundancy between adjacent sampling intervals.

Daily recurrence: Focusing again on sensor #5, Figure 3e presents data collected over a 1-week period, with daily observations plotted as individual curves. These curves exhibit recurrent patterns throughout the week, characterized by an "M" shape with dual peaks during the morning and evening rush hours. This daily recurrence demonstrates the redundancy in the sensor's readings on different days.

2.4.2. Data Redundancy Representation

In this study, these multi-mode data redundancies are collectively referred to as global redundancy, which characterizes the intrinsic structure of spatiotemporal WDN data. Global redundancy indicates that information is redundant and can be compressed across all tensor modes, a property that is algebraically represented as "low rankness" (Zhang et al., 2024). As depicted in Figures 3b, 3d, and 3f, the singular value decomposition is performed on the three tensor mode unfoldings (i.e., $\boldsymbol{D}_{(1)}$, $\boldsymbol{D}_{(2)}$ and $\boldsymbol{D}_{(3)}$ in Figure 1). The singular values of each

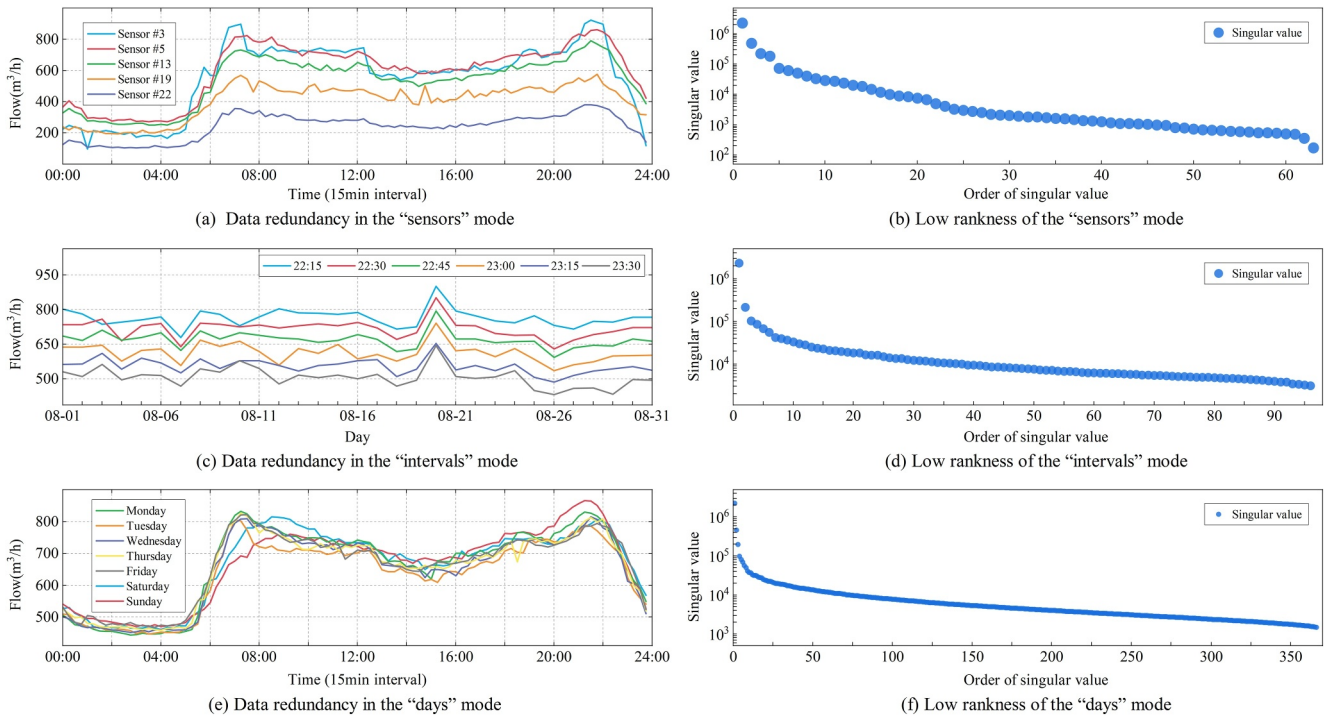

Figure 3. Illustration of multi-mode data redundancies and low-rank properties of spatiotemporal Water Distribution Network data.

unfolding matrix are dominated by only a handful of large ones, indicating that fewer fundamental patterns can describe most information. By imposing low-rank prior constraints, reliable estimates for missing values are obtained by exploiting global information from observations and latent correlations of other modes. For example, BM data is no longer regarded as structural corruption in the "days" mode and can now be effectively imputed using daily recurrence.

2.5. Spatiotemporal Data Imputation

In the following subsections, a low-rank prior based on Truncated Nuclear Norm (TNN) minimization is first applied to capture the global redundancy in WDN tensor data. Given the intrinsic temporal continuity in sensor time series, a novel autoregressive regularization is subsequently introduced into the tensor completion model. This combination enables the imputation of missing values with high efficiency and accuracy, fully harnessing the potential of spatiotemporal WDN data.

2.5.1. Truncation Operation for Global Redundancy

As described previously, spatiotemporal WDN data is characterized by global redundancy or algebraically exhibits low-rank properties. For this reason, LRTC models (Yuan & Zhang, 2016) have great potential for WDN data imputation. There is growing evidence that tensor representation outperforms matrix representation in various research areas, such as the recovery of image, video, and traffic data (Asif et al., 2016; Liu et al., 2013; Lu et al., 2020). The tensorization step converts the missing value estimation task in SWNs into a standard tensor completion problem as follows:

$$\min_{\mathcal{X}} \operatorname{rank}(\mathcal{X}) \quad \text{s.t.} \begin{cases} \mathcal{X} = \mathcal{Q}(\boldsymbol{Y}), \\ \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D}), \end{cases} \tag{3}$$

where $\boldsymbol{Y} \in \mathbb{R}^{M\times(IJ)}$ is introduced to retain the observation information as an auxiliary variable, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{M\times I\times J}$ is the low-rank tensor to be reconstructed based on partial observations. The forward tensorization operator $\mathcal{Q}(\cdot)$ is utilized to split the temporal dimension into (intervals, days)-indexed combinations, that is, $\boldsymbol{\mathcal{X}} = \mathcal{Q}(\boldsymbol{Y}) \in \mathbb{R}^{M\times I\times J}$. Conversely, an inverse operator $\mathcal{Q}^{-1}(\cdot)$ is defined to unfold the resulting tensor back into the original matrix by $\boldsymbol{Y} = \mathcal{Q}^{-1}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{M\times(IJ)}$.

In the optimization problem (3), minimizing the tensor rank can describe global redundancy across different dimensions. However, a central issue in LRTC is the appropriate definition of the tensor rank (Wang et al., 2021) since its direct calculation is NP-hard (Håstad, 1990). Over the past decade, numerous studies have sought alternative approximations of the rank function to overcome this difficulty. The NN $\|\boldsymbol{\mathcal{X}}\|_*$, which is the tightest convex surrogate for the tensor rank, has been widely used in LRTC (Lu et al., 2019) due to its effectiveness in preserving the inherent structure of tensors. In recent years, more studies have suggested that the nonconvex approximation of the rank function, namely the TNN $\|\boldsymbol{\mathcal{X}}\|_{r,*}$, produces superior estimation accuracy than the NN $\|\boldsymbol{\mathcal{X}}\|_*$ (Huang et al., 2014; Nie et al., 2022; Xue et al., 2018). Formal definitions of the NN and TNN for matrices and tensors are outlined below.

Definition 1 (Nuclear Norm & Truncated Nuclear Norm for Matrices). The NN of a given matrix $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ is defined as the summation of its singular values:

$$\|\boldsymbol{X}\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{X}). \tag{4}$$

Considering a positive integer $r < \min\{m,n\}$, its TNN is defined as the summation of $\min\{m,n\} - r$ minimum singular values:

$$\|\boldsymbol{X}\|_{r,*} = \sum_{i=r+1}^{\min\{m,n\}} \sigma_i(\boldsymbol{X}). \tag{5}$$

Definition 2 (Nuclear Norm & Truncated Nuclear Norm for Tensors). In alignment with the work of Liu et al. (2013), the NN/TNN for any tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{M\times I\times J}$ is defined as the weighted sum of NN/TNN for all the unfolding matrices along each mode, expressed as Equations 6 and 7 respectively:

$$\|\boldsymbol{\mathcal{X}}\|_* = \sum_{k=1}^{3} \alpha_k \left\|\boldsymbol{\mathcal{X}}_{(k)}\right\|_*, \tag{6}$$

$$\|\boldsymbol{\mathcal{X}}\|_{r,*} = \sum_{k=1}^{3} \alpha_k \left\|\boldsymbol{\mathcal{X}}_{(k)}\right\|_{r,*}, \tag{7}$$

where α_k denotes the non-negative weight parameter on the corresponding unfolding matrix $\boldsymbol{\mathcal{X}}_{(k)}$ with $\sum_{i=1}^{3} \alpha_k = 1$, $r \in \mathbb{N}_+$ serves as the truncation threshold and satisfies $r < \min\{M, I, J\}$.

In contrast to the LRTC with NN (LRTC-NN) approach, which minimizes all singular values simultaneously, the LRTC with TNN (LRTC-TNN) approach focuses on the minimization of the smaller singular values. This strategy more precisely reflects the actual rank structure and preserves the principal components. By minimizing TNN as the objective function, the optimization problem (3) is reformulated into a LRTC-TNN model:

$$\min_{\boldsymbol{\mathcal{X}}} \ \|\boldsymbol{\mathcal{X}}\|_{r,*} \quad \text{s.t.} \begin{cases} \boldsymbol{\mathcal{X}} = \mathcal{Q}(\boldsymbol{Y}), \\ \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D}). \end{cases} \tag{8}$$

2.5.2. Autoregressive Regularization for Local Correlation

A major limitation of LRTC-based models is their inability to capture local temporal patterns in non-stationary WDN time series despite characterizing global redundancy. In practice, the observed flow and pressure values frequently exhibit temporary fluctuations due to special events or random noise, which deviate more or less from the global trend. Figure 3c clearly shows the local correlation and smoothness between temporally adjacent observations, especially when water demand has an apparent increase (e.g., August 20) or decrease (e.g., August 7). Therefore, it is essential to explicitly impose temporal continuity on the model to ensure that the imputation results remain accurate and relevant over time.

Recent advances have proposed feasible solutions by introducing regularization schemes that incorporate such temporal continuity to improve imputation performance while avoiding overfitting (Chen & Sun, 2021; Takeuchi et al., 2017). To better encode strong local correlation, autoregressive processes are incorporated as a novel regularization term that accounts for temporal variation of the variable matrix $\boldsymbol{Y}$:

$$\|\boldsymbol{Y}\|_{\boldsymbol{C},\mathcal{H}} = \sum_{m,t}\left(y_{m,t} - \sum_{i} c_{m,i}\, y_{m,t-h_i}\right)^2, \tag{9}$$

where $\boldsymbol{C} \in \mathbb{R}^{M\times d}$ is a learnable coefficient parameters matrix and $\mathcal{H} = \{h_1, \cdots, h_d\}$ is a time lag set. $\|\boldsymbol{Y}\|_{\boldsymbol{C},\mathcal{H}}$ quantifies the accumulated sum of autoregressive errors incurred in fitting each time series $\boldsymbol{y}_m$ using coefficient vector $\boldsymbol{c}_m$. Upon estimating $\boldsymbol{C}$, minimizing temporal variation fosters enhanced temporal continuity of $\boldsymbol{Y}$.

2.5.3. Data Imputation Approach

Based on the above theoretical foundations, the LATC approach (X. Chen et al., 2022) is employed for spatio-temporal WDN data imputation. To model both global redundancy and local correlation, LATC integrates TNN minimization of the completed tensor with temporal variation minimization of the unfolding time series matrix:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{X}},\boldsymbol{Y},\boldsymbol{C}} \ \|\boldsymbol{\mathcal{X}}\|_{r,*} + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\boldsymbol{C},\mathcal{H}} \\ &\text{s.t.}\begin{cases} \boldsymbol{\mathcal{X}} = \mathcal{Q}(\boldsymbol{Y}), \\ \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D}), \end{cases} \end{aligned} \tag{10}$$

where λ is the positive weight parameter that regulates the trade-off between TNN and temporal variation.

The straightforward and default technique for solving LRTC-based models is the Alternating Direction Method of Multipliers (ADMM) algorithm, which decomposes large optimization problems into smaller tractable sub-problems and facilitates ease of parallelization. However, the new optimization problem Equation 10 remains unsolvable due to the introduced autoregressive coefficient matrix $\boldsymbol{C}$, rendering the ADMM algorithm incapable of ensuring convergence (C. Chen et al., 2016). An alternating minimization technique is implemented to decompose Equation 10 into two subproblems. Let $(\boldsymbol{\mathcal{X}}^{\circ}, \boldsymbol{Y}^{\circ}, \boldsymbol{C}^{\circ})$ denote the given initial values and ℓ be the iteration count. The variables $\left\{\left(\boldsymbol{\mathcal{X}}^{\ell}, \boldsymbol{Y}^{\ell}, \boldsymbol{C}^{\ell}\right)\right\}_{\ell\in\mathbb{N}}$ are updated by iteratively solving Equation 11 and Equation 19. Specifically, $\boldsymbol{C}^{\ell}$ is fixed, and the following optimization problem is solved to update $\boldsymbol{\mathcal{X}}^{\ell+1}$ and $\boldsymbol{Y}^{\ell+1}$:

$$\begin{aligned} &\boldsymbol{\mathcal{X}}^{\ell+1}, \boldsymbol{Y}^{\ell+1} := \arg\min_{\boldsymbol{\mathcal{X}},\boldsymbol{Z}} \ \|\boldsymbol{\mathcal{X}}\|_{r,*} + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\boldsymbol{C}^{\ell},\mathcal{H}} \\ &\text{s.t.}\begin{cases} \boldsymbol{\mathcal{X}} = \mathcal{Q}(\boldsymbol{Y}), \\ \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{Y}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D}). \end{cases} \end{aligned} \tag{11}$$

With $\boldsymbol{C}^{\ell}$ fixed, the subproblem Equation 11 naturally transforms into a standard LRTC-TNN problem, which the ADMM can solve similarly to Hu et al. (2013) and Liu et al. (2013). Following the ADMM, the augmented Lagrangian function of Equation 11 is formulated as

$$\mathcal{L}(\boldsymbol{\mathcal{X}}, \boldsymbol{Y}, \boldsymbol{C}^{\ell}, \boldsymbol{\mathcal{T}}) = \|\boldsymbol{\mathcal{X}}\|_{r,*} + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\boldsymbol{C}^{\ell},\mathcal{H}} + \frac{\rho}{2}\|\boldsymbol{\mathcal{X}} - \mathcal{Q}(\boldsymbol{Y})\|_F^2 + \langle \boldsymbol{\mathcal{X}} - \mathcal{Q}(\boldsymbol{Y}), \boldsymbol{\mathcal{T}} \rangle, \tag{12}$$

where $\rho > 0$ is a penalty parameter and $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{M \times I \times J}$ is the dual variable. Subsequently, one of the variables is iteratively updated while keeping the other two fixed:

$$\boldsymbol{\mathcal{X}}^{\ell+1,v+1} := \arg\min_{\boldsymbol{\mathcal{X}}} \mathcal{L}\left(\boldsymbol{\mathcal{X}}, \boldsymbol{Y}^{\ell+1,v}, \boldsymbol{C}^{\ell}, \boldsymbol{\mathcal{T}}^{\ell+1,v}\right), \tag{13}$$

$$\boldsymbol{Y}^{\ell+1,v+1} := \arg\min_{\boldsymbol{Y}} \mathcal{L}\left(\boldsymbol{\mathcal{X}}^{\ell+1,v+1}, \boldsymbol{Y}, \boldsymbol{C}^{\ell}, \boldsymbol{\mathcal{T}}^{\ell+1,v}\right), \tag{14}$$

$$\boldsymbol{\mathcal{T}}^{\ell+1,v+1} := \boldsymbol{\mathcal{T}}^{\ell+1,v} + \rho\left(\boldsymbol{\mathcal{X}}^{\ell+1,v+1} - \mathcal{Q}\left(\boldsymbol{Y}^{\ell+1,v+1}\right)\right), \tag{15}$$

where v is the number of iterations in the ADMM. The detailed solutions of Equations 13 and 14 are discussed and demonstrated below. It is noteworthy that observation consistency is preserved by enforcing the fixed constraint $\mathcal{P}_{\boldsymbol{\Omega}}\left(\boldsymbol{Y}^{\ell+1}\right) := \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D})$.

1. Update Variable $\boldsymbol{\mathcal{X}}$

   The optimization problem concerning $\boldsymbol{\mathcal{X}}$ involves TNN minimization. According to Equation (7), each $\boldsymbol{\mathcal{X}}_k$ must be solved independently, and its closed-form solution is determined by

$$\begin{aligned} \boldsymbol{\mathcal{X}}_k &:= \arg\min_{\boldsymbol{\mathcal{X}}} \alpha_k \left\|\boldsymbol{\mathcal{X}}_{(k)}\right\|_{r,*} + \frac{\rho}{2}\left\|\mathcal{Q}^{-1}(\boldsymbol{\mathcal{X}}) - \boldsymbol{Y}^{\ell+1,v}\right\|_F^2 + \left\langle \mathcal{Q}^{-1}(\boldsymbol{\mathcal{X}}) - \boldsymbol{Y}^{\ell+1,v}, \mathcal{Q}^{-1}\left(\boldsymbol{\mathcal{T}}^{\ell+1,v}\right)\right\rangle \\ &= \arg\min_{\boldsymbol{\mathcal{X}}} \alpha_k \left\|\boldsymbol{\mathcal{X}}_{(k)}\right\|_{r,*} + \frac{\rho}{2}\left\|\boldsymbol{\mathcal{X}}_{(k)} - \left(\mathcal{Q}\left(\boldsymbol{Y}^{\ell+1,v}\right)_{(k)} - \boldsymbol{\mathcal{T}}_{(k)}^{\ell+1,v}/\rho\right)\right\|_F^2 \\ &= \mathrm{fold}_k\left(\mathcal{G}_{r,\alpha_k/\rho}\left(\mathcal{Q}\left(\boldsymbol{Y}^{\ell+1,v}\right)_{(k)} - \boldsymbol{\mathcal{T}}_{(k)}^{\ell+1,v}/\rho\right)\right), \end{aligned} \tag{16}$$

   where $\mathcal{G}_{\cdot}(\cdot)$ denotes the generalized Singular Value Thresholding (SVT) related to TNN minimization. The detailed solution process of Equation 16 is given in Text S1 in Supporting Information S1. Subsequently, the updated value of $\boldsymbol{\mathcal{X}}$ can be obtained by

$$\boldsymbol{\mathcal{X}}^{\ell+1,v+1} := \sum_{k=1}^{3} \alpha_k \boldsymbol{\mathcal{X}}_k \tag{17}$$

2. Update Variable $\boldsymbol{Y}$

   In terms of $\boldsymbol{Y}^{\ell+1,v+1}$, Equation 14 is formulated as

$$\begin{aligned} \boldsymbol{Y}^{\ell+1,v+1} &:= \arg\min_{\boldsymbol{Y}} \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\boldsymbol{C}^{\ell},\mathcal{H}} + \frac{\rho}{2}\left\|\boldsymbol{\mathcal{X}}^{\ell+1,v+1} - \mathcal{Q}(\boldsymbol{Y})\right\|_F^2 - \left\langle \mathcal{Q}(\boldsymbol{Y}), \boldsymbol{\mathcal{T}}^{\ell+1,v}\right\rangle \\ &= \arg\min_{\boldsymbol{Y}} \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\boldsymbol{C}^{\ell},\mathcal{H}} + \frac{\rho}{2}\left\|\boldsymbol{Y} - \mathcal{Q}^{-1}\left(\boldsymbol{\mathcal{X}}^{\ell+1,v+1} + \boldsymbol{\mathcal{T}}^{\ell+1,v}/\rho\right)\right\|_F^2. \end{aligned} \tag{18}$$

3. Update Variable $\boldsymbol{C}$

   According to the result of $\boldsymbol{Y}^{\ell+1,\Upsilon}$, the update of $\boldsymbol{C}^{\ell+1}$ is obtained by solving a least squares problem as follows:

$$\boldsymbol{C}^{\ell+1} := \arg\min_{\boldsymbol{C}} \left\|\boldsymbol{Y}^{\ell+1,\Upsilon}\right\|_{\boldsymbol{C},\mathcal{H}}, \tag{19}$$

   where Υ is the maximum iteration in the ADMM.

The derivation process for solving Equations 18 and 19 is beyond the scope of this paper. Closed-form solutions are provided in Text S2 and Text S3 in Supporting Information S1, respectively.

The implementation of the LATC Imputer for estimating missing values in spatiotemporal WDN data is summarized in Algorithm 1. Three parameters play crucial roles in the algorithm: weight parameter λ (balancing TNN and temporal variation), learning rate parameter ρ (controlling the ADMM and the SVT) and integer-wise truncation parameter r for TNN. For parameters $\alpha_1, \alpha_2, \alpha_3$, the same weights are assigned to the three tensor unfolding matrices, rather than tuning these parameters for each data set and missing scenario, as this process would be computationally expensive. The time lag set $\mathcal{H}$ and convergence threshold ϵ are case-specific. For convenience, a new parameter c is introduced, defined by $c = \lambda/\rho$. Consequently, $c = 1$ indicates the equal significance of both norms in the objective function.

Algorithm 1. LATC Imputer

Input: incomplete matrix $\boldsymbol{D}$, index set $\boldsymbol{\Omega}$, and parameters $c, r, \rho, \mathcal{H}, \epsilon$

Output: recovered matrix $\widehat{\boldsymbol{D}}$

Initialize $\boldsymbol{\mathcal{C}}^0$ as random small values and $\boldsymbol{\mathcal{T}}^{0,0}$ as zeros. Configure $\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Y}}^{0,0}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D})$, $\lambda = c \cdot \rho$, $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$, $\ell = 0$, and $\Upsilon = 3$.

while *not converged* **do**

  for $\nu = 1$ **to** Υ **do**

    $\rho = \min\{1.05 \times \rho, \rho_{\max}\}$;

    for $k = 1$ **to** 3 **do**

      Compute $\boldsymbol{\mathcal{X}}_k^{\ell+1,\nu+1}$ by Equation (16);

    Update $\boldsymbol{\mathcal{X}}^{\ell+1,\nu+1}$ by Equation (17);

    Update $\boldsymbol{\mathcal{Y}}^{\ell+1,\nu+1}$ by Equation (18);

    Update $\boldsymbol{\mathcal{T}}^{\ell+1,\nu+1}$ by Equation (15);

    Ensure observation consistency by enforcing $\mathcal{P}_{\boldsymbol{\Omega}}\left(\boldsymbol{\mathcal{Y}}^{\ell+1,\nu+1}\right) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D})$;

    Update $\boldsymbol{\mathcal{C}}^{\ell+1}$ by Equation (19);

  Compute $\widehat{\boldsymbol{X}}^{\ell+1} = \mathcal{Q}^{-1}\left(\boldsymbol{\mathcal{X}}^{\ell+1,\Upsilon}\right)$;

  Compute $e^{\ell+1} = \left\|\widehat{\boldsymbol{X}}^{\ell+1} - \widehat{\boldsymbol{X}}^{\ell}\right\|_F / \|\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{D})\|_F$;

  if $e^{\ell+1} < \epsilon$ **then**

    Set $\hat{d}_{m,n} = \hat{x}_{m,n}, \forall (m,n) \notin \boldsymbol{\Omega}$;

    Converge.

  $\ell \coloneqq \ell + 1$;

In SWNs, data-driven downstream applications have fundamental requirements for imputation. First, model training, calibration, or updating relies on extensive, high-quality historical data sets. Second, complete and real-time input data guarantees the model's proper operation. Considering the above, two methods for parameter optimization and real-time data imputation are proposed in Text S4 in Supporting Information S1, which are necessary for the practical application of LATC. Therefore, the LATC framework is anticipated to serve as a general approach to consistently achieve robust and reliable estimation results for multiple imputation-related tasks.

3. Experiment

3.1. Data Set Descriptions

To investigate the imputation applicability of LATC for spatiotemporal WDN data, two real-world, large-scale data sets provided by the water utility of Z city—a typical city in northern China—are selected. Figure 4

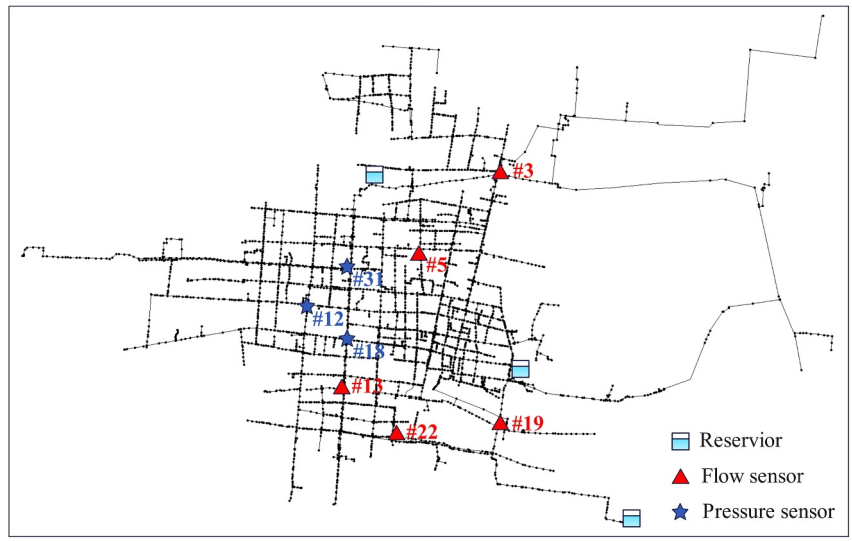

Figure 4. Schematic of Z-city pipe network topology. This graph labels the locations of the flow and pressure sensors shown in Figures 3 and 10, respectively.

presents the schematic of Z-city pipe network topology. As key hydraulic parameters of SWNs, flow and pressure data must be collected and uploaded in real-time during WDN operation. Two simulated data sets are also generated using a public network (see Figure S1 in Supporting Information S1 for details) to contrast with the real-world data. A summary of the four data sets is presented in Table 1. For clarity, "*Z-city Flow*" and "*Z-city Pres*" denote the real-world flow and pressure data sets collected from Z-city's WDN, whereas "*Sim-net Flow*" and "*Sim-net Pres*" represent the simulated flow and pressure data sets generated based on the public network. These abbreviations are used consistently throughout the paper. In particular, these data set sizes differ in the temporal and spatial dimensions, permitting a comprehensive analysis.

3.2. Missing Data Scenarios

To test the imputation effectiveness of LATC for spatiotemporal WDN data, several missing scenarios are created based on three missing patterns: RM, LM, and BM, as explained in Section 2.3. The first is multiple scenarios consisting of individual missing pattern data at different rates, including 30% RM, 60% RM, 30% LM, 30% BM,

Table 1
Description of Four Spatiotemporal Water Distribution Network Data Sets

Data set	Description	Composition of sensors	Data size
Z-city Flow (m^3/h)	This data set provides instantaneous flow rate/water demand values collected from 63 sensors over 1 year (from 8 September 2023 to 7 September 2024) at 15-min intervals. It contains 3.41% missing values.	3 sensors monitoring water plants outflows; 5 sensors monitoring scheduling valves flows; 24 sensors monitoring pipeline flows; 31 sensors monitoring end-user consumption flows.	Tensor: 63 × 96 × 366; Matrix: 63 × 35,136
Z-city Pres (m)	This data set records pressure values collected from 52 sensors over 6 months (from 1 March 2024 to 31 August 2024) at 15-min intervals. It contains 2.87% missing values.	5 sensors monitoring water plants outlet pressure; 8 sensors monitoring Pressure before or after the valve; 25 sensors monitoring pipeline pressures; 14 sensors monitoring end-user pressures.	Tensor: 52 × 96 × 184; Matrix: 52 × 17,664
Sim-net Flow (m^3/h)	This data set contains instantaneous flow rate/water demand values from 120 sensor junctions over 2 months at 15-min intervals. Gaussian noise with a standard deviation of 3% is added to the simulated values of each sensor.	3 sensors monitoring water plants outflows; 5 sensors monitoring scheduling valves flows; 30 sensors monitoring pipeline flows; 82 sensors monitoring end-user consumption flows.	Tensor: 120 × 96 × 61; Matrix: 120 × 5,856
Sim-net Pres (m)	This data set contains pressure values from 52 sensor junctions over 2 months at 15-min intervals. Gaussian noise with a standard deviation of 0.5 m is added to the simulated values of each sensor.	3 sensors monitoring water plants outlet pressure; 8 sensors monitoring Pressure before or after the valve; 28 sensors monitoring pipeline pressures; 13 sensors monitoring end-user pressures.	Tensor: 52 × 96 × 61; Matrix: 52 × 5,856

and 60% BM. Specifically, the LM gap (indicating the length of consecutive missing in a time series) set is defined as $\{96, 48, 24\}$ and the BM window (indicating the length of entire missing for all sensors) set as $\{2, 4, 6\}$. Recognizing that real-world WDN data sets exhibit a combination of missing patterns, the following Mixed Missing (MM) scenarios are also considered:

- 30% MM consisting of 10% RM, 10% LM, and 10% BM;
- 50% MM consisting of 20% RM, 15% LM, and 15% BM;
- 70% MM consisting of 30% RM, 20% LM, and 20% BM.

Note that the same missing ratio is assigned to each element in the LM gap set for all missing scenarios containing LM, and the same applies to the BM window set.

In this study, a certain percentage of observations is manually masked at specific locations to simulate the missing cases. The masked data set shares a function similar to the test data set in machine learning. By comparing the imputed values with the masked ones, the optimal parameters $\overset{*}{c}$ and $\overset{*}{r}$ that perform best under a chosen evaluation metric can be identified.

Symmetric Mean Absolute Percentage Error (SMAPE) and Root Mean Square Error (RMSE) are adopted as evaluation metrics:

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|d_i - \hat{d}_i|}{(|d_i| + |\hat{d}_i|)/2} \tag{20}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{d}_i)^2}, \tag{21}$$

where d_i and $\hat{d}_i$ are masked observations values and imputed values, respectively. Lower values of SMAPE and RMSE imply that the imputation model produces results closer to the actual values. Conversely, higher values suggest poorer imputation performance. Given that many flow values are near-zero or relatively small, the relative error can be substantial when measured using MAPE. In contrast, SMAPE offers a more balanced error measure for dealing with flow data sets.

3.3. Baseline Models and Parameter Settings

The imputation performance of LATC on spatiotemporal WDN data sets is primarily influenced by two key parameters: the trade-off coefficient c (defined by λ/ρ) and the rank truncation r. The parameter search spaces are uniformly set to a wide range as $c \in \{1/10, 1/5, 1, 5, 10\}$ and $r \in \{5, 10, 15, 20, 25, 30\}$ for all experiments. Furthermore, a preliminary test is conducted to determine the appropriate values of other parameters. Specifically, ρ is selected from $\{1 \times 10^{-4}, 1 \times 10^{-5}\}$ and $\mathcal{H}$ is preset as $\{1, 2, 3, 4, 5, 6\}$ for all data sets.

To our knowledge, no research has explored the usability of low-rank prior-based models on large WDN data sets. Therefore, three well-known and proven approaches in other domains are selected to evaluate the improvement of LATC, namely HaLRTC, TRMF, and LAMC-TNN. Two popular machine learning approaches are also used as a contrast: KNN and Missforest. All experimental code and detailed configurations are publicly available at GitHub (Xu, 2025a). A summary of the five baseline models is as follows.

- *HaLRTC*: High-accuracy LRTC (Liu et al., 2013), which minimizes the NN of tensor via ADMM, has been widely used as a benchmark LRTC in numerous imputation studies.
- *TRMF*: Temporal Regularized Matrix Factorization (H.-F. Yu et al., 2016) introduces an autoregressive regularization temporal regularizer into the default matrix factorization model. The matrix rank is searched from $\{5, 10, 20, 30, 40\}$ for all data sets.
- *LAMC*: Low-rank Autoregressive Matrix Completion with TNN, which also incorporates an autoregressive regularization term into the default matrix completion model (Y. Hu et al., 2013), can be viewed as a matrix variant of LATC. It shares the same parameter search space as LATC.

Table 2
Imputation Performance Comparison (in Symmetric Mean Absolute Percentage Error/Root Mean Square Error) on Z-City Flow and Sim-Net Flow Data Set

Data set	Scenario	LATC	HaLRTC	TRMF	LAMC	KNN	Missforest
Z-city Flow	30% RM	**10.92/60.69**	13.89/100.5	13.71/86.20	15.54/108.2	18.93/187.0	12.77/101.4
	60% RM	**11.93/73.55**	15.65/124.0	14.59/97.26	19.53/183.9	24.93/340.2	15.96/164.9
	30% LM	14.68/**117.7**	16.46/142.3	18.99/143.1	16.65/131.1	21.75/240.6	**13.91**/126.5
	30% BM	**13.24/93.81**	16.01/132.2	39.61/447.1	18.49/176.5	41.93/470.3	41.54/483.5
	60% BM	**14.67/120.2**	18.62/163.9	75.02/765.3	26.11/297.5	42.57/472.0	42.66/494.0
	30% MM	**13.10/87.87**	15.42/118.8	21.66/213.7	16.60/107.7	28.67/325.7	25.70/340.1
	50% MM	**13.34/95.48**	16.00/132.1	25.03/286.0	17.59/120.7	30.75/364.7	26.29/350.8
	70% MM	**14.70/115.9**	18.14/159.3	37.90/451.7	19.34/158.5	34.39/410.5	30.51/377.1
Sim-net Flow	30% RM	2.855/6.03	3.202/5.94	3.832/6.09	3.261/6.18	**2.394/6.01**	2.437/6.20
	60% RM	2.902/**5.91**	3.552/6.12	4.253/6.21	3.574/6.29	2.826/6.55	**2.760**/6.84
	30% LM	2.896/**6.24**	3.625/6.54	3.945/6.40	3.306/6.46	**2.399**/6.32	2.446/6.49
	30% BM	**2.925/5.82**	4.163/6.56	11.63/26.20	12.01/25.52	44.91/110.6	43.94/115.1
	60% BM	**3.166/5.83**	5.957/8.86	51.28/136.6	26.41/74.19	43.04/108.5	42.40/112.6
	30% MM	**2.960/5.90**	3.810/6.33	5.348/9.22	6.686/15.02	21.00/72.49	20.05/78.63
	50% MM	**2.893/5.79**	3.742/6.23	5.619/10.52	10.77/32.53	18.83/67.74	17.70/72.14
	70% MM	**2.964/5.99**	4.209/6.90	14.46/42.16	8.308/21.44	24.52/78.44	24.64/84.77

Note. Values highlighted in boldface represent the best performance within each respective row.

- *KNN*: K-Nearest Neighbor (Beretta & Santaniello, 2016) is a classical imputation algorithm that estimates missing values by identifying the k nearest neighbors to the missing data point and calculating their average. The parameter k is searched from $\{5, 10, 15, 20, 30\}$ for all data sets.
- *Missforest*: Missforest (Stekhoven & Bühlmann, 2012) is a non-parametric approach that utilizes the random forests algorithm to predict and replace missing values iteratively.

4. Results

4.1. Imputation Performance

Tables 2 and 3 summarize the imputation performance of LATC and baseline models on four spatiotemporal WDN data sets. The results demonstrate that the LATC model consistently outperforms the baseline models under both BM and MM scenarios, with a particularly significant advantage. For RM and LM imputations, the performance of each model varied across data sets; however, LATC generally achieves the best or near-best results on all data sets. On the two real-world data sets (*Z-city Flow* and *Z-city Pres*), the rate and pattern of missing data are observed to considerably influence all models. In particular, as the missing rate increases, the SMAPE/RMSE values rise accordingly, and structural missing patterns (e.g., LM and BM) pose more complex challenges than the simpler RM pattern.

To illustrate the imputation performance of various models clearly, visualization examples are provided under a representative and complex missing scenario (i.e., 50% MM) on *Z-city Flow* data set and *Z-city Pres* data set, as shown in Figures 5 and 6. One can easily find that even in severe, structural, and MM scenarios, the true long-term trends and detailed information are successfully reconstructed by the LATC model based on the captured global low-rank structure and local temporal continuity of spatiotemporal WDN data.

Compared to HaLRTC, LATC shows greater robustness to the increasing missing rate, and as displayed in Figures 5a and 5b, its estimation for missing values is in better agreement with the actual values. The superiority of LATC over the two matrix-based models, TRMF and LAMC, indicates that tensor structure is more capable of characterizing abundant and underlying multi-mode correlations. Although TRMF gives acceptable results in RM

Table 3
Imputation Performance Comparison (in Symmetric Mean Absolute Percentage Error/Root Mean Square Error) on Z-City Pres and Sim-Net Pres Data Sets

Data set	Scenario	LATC	HaLRTC	TRMF	LAMC	KNN	Missforest
Z-city Pres	30% RM	1.969/0.852	2.262/0.954	**1.684/0.765**	2.047/0.920	2.094/0.926	1.781/0.844
	60% RM	2.197/0.944	2.696/1.125	**1.892/0.843**	2.465/1.129	3.125/1.394	2.228/1.063
	30% LM	2.247/0.961	2.604/1.089	2.073/0.927	2.150/0.973	2.195/0.968	**1.864/0.882**
	30% BM	**2.885/1.251**	3.180/1.339	5.897/2.485	3.271/1.445	5.375/2.270	10.42/4.017
	60% BM	**3.148/1.345**	3.660/1.509	11.94/4.591	4.234/1.859	5.450/2.300	5.919/2.474
	30% MM	**2.358/1.044**	2.654/1.132	3.276/1.505	2.518/1.156	3.605/1.698	3.575/1.794
	50% MM	**2.477/1.083**	2.809/1.184	3.518/1.705	2.643/1.200	3.552/1.658	3.495/1.723
	70% MM	**2.739/1.185**	3.250/1.348	5.830/2.741	3.081/1.372	4.168/1.875	3.993/1.900
Sim-net Pres	30% RM	1.322/0.472	1.393/0.503	1.375/0.492	1.393/0.498	**1.321/0.470**	1.374/0.491
	60% RM	**1.322/0.471**	1.468/0.535	1.453/0.523	1.582/0.591	1.600/0.613	1.560/0.563
	30% LM	1.310/0.470	1.445/0.530	1.368/0.492	1.390/0.499	**1.306/0.468**	1.364/0.490
	30% BM	**1.300/0.467**	1.645/0.628	2.849/1.157	3.875/1.890	14.29/5.884	13.86/6.109
	60% BM	**1.325/0.467**	2.114/0.841	14.64/6.176	7.393/3.800	14.81/5.840	13.72/5.942
	30% MM	**1.326/0.471**	1.493/0.546	1.575/0.589	2.156/0.955	7.473/4.052	7.060/4.162
	50% MM	**1.322/0.469**	1.538/0.565	2.128/0.884	2.251/1.023	7.433/3.995	6.968/4.062
	70% MM	**1.315/0.469**	1.712/0.665	4.544/2.373	2.653/1.265	8.650/4.478	8.273/4.643

Note. Values highlighted in boldface represent the best performance within each respective row.

and LM scenarios for specific data sets, as seen from the gray rectangles in Figures 5c and 6c, its dependence on the autoregressive process alone leads to unreasonable estimation for BM data. In addition to low-rank matrix-based models, two machine learning approaches, KNN and Missforest, fail for BM imputation tasks due to a lack of available neighbor entries or sufficient reference information for accurate prediction. Furthermore, their sensitivity to outliers and reliance on black-box algorithms (Rudin, 2019) further limit the reliability of missing WDN data imputation.

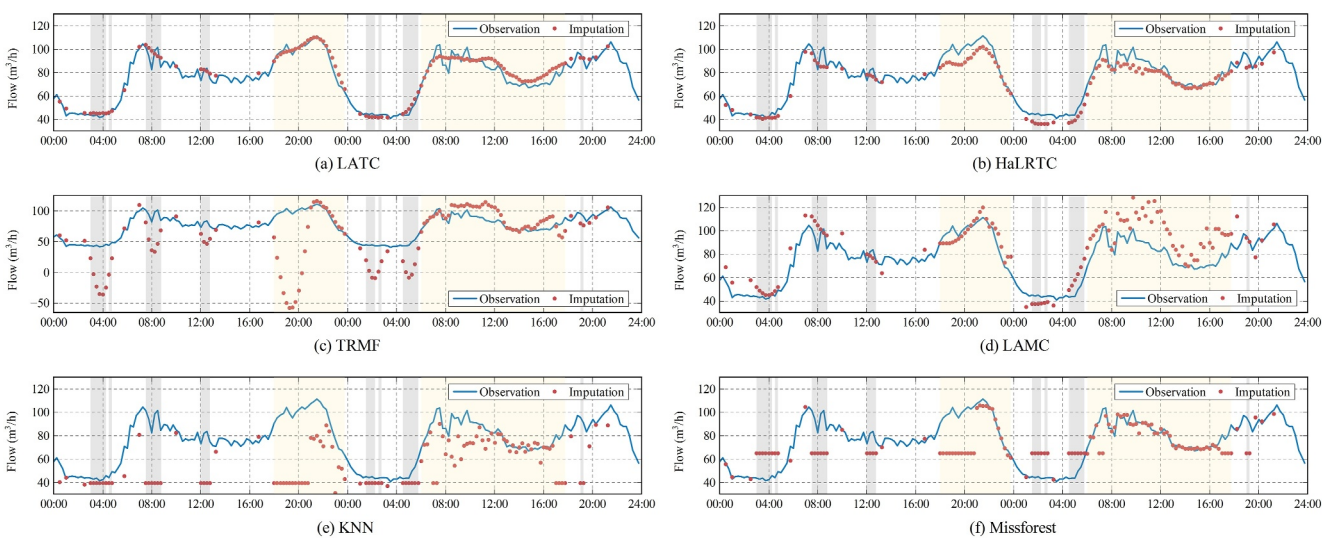

Figure 5. Visualizations of imputation performance comparison between Low-rank Autoregressive Tensor Completion and baseline models under 50% Mixed Missing missing scenario on *Z-city Flow* data set. In these visualizations, red dots represent imputed values at masked positions, blue curves represent observed data, light yellow rectangles indicate Long-range Missing within a specific gap, and gray rectangles indicate Block Missing within a specific window. Examples correspond to sensor #4 from September 4 to 5, 2024.

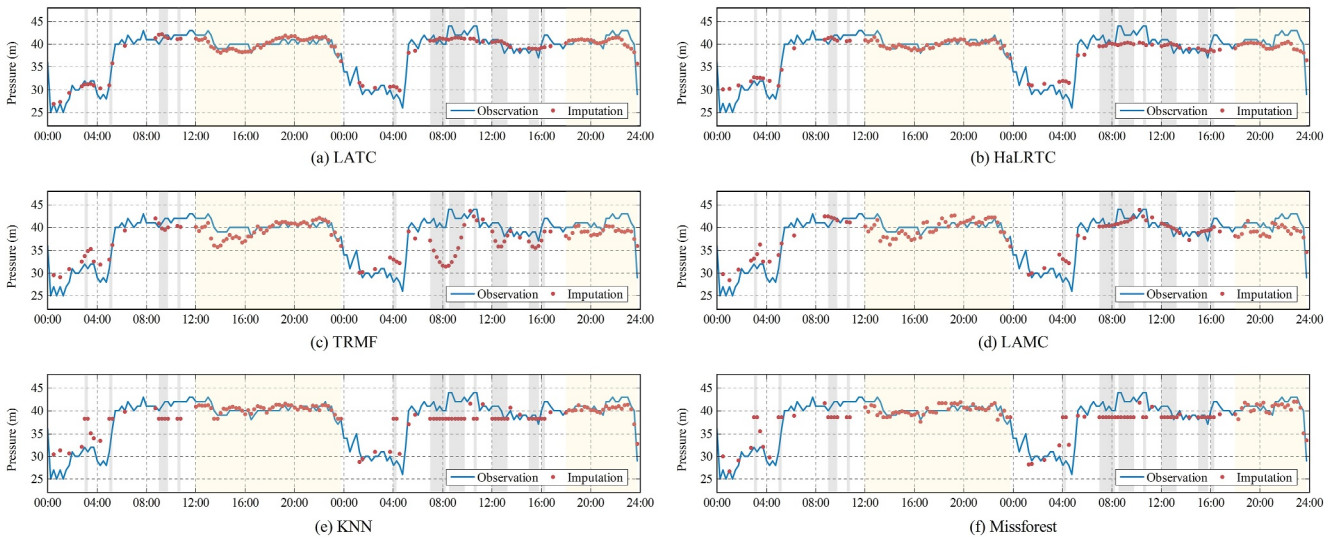

Figure 6. Visualizations of imputation performance comparison between Low-rank Autoregressive Tensor Completion and baseline models under 50% Mixed Missing missing scenario on *Z-city Pres* data set. Examples correspond to sensor #6 from June 7 to 8, 2024.

Furthermore, beyond the promising imputation performance, the computational efficiency of LATC is also validated. As detailed in Appendix A, a comprehensive comparison of LATC and baseline models is conducted across eight missing scenarios and four data sets, along with an in-depth efficiency analysis. The results demonstrate that LATC achieves a favorable balance between runtime and accuracy. Specifically, LATC consistently outperforms TRMF and Missforest by a large margin in terms of running time, while maintaining comparable or slightly better efficiency than LAMC and KNN. Although HaLRTC exhibits the fastest runtime, this comes at the cost of poor imputation accuracy under complex missing patterns. Benefiting from its lightweight yet effective design, LATC demonstrates stable and efficient performance across different data sets and missing scenarios, which confirms its practical applicability for real-time imputation tasks in SWNs.

4.2. Effect of Truncation and Autoregressive Regularization

4.2.1. Parameter Analysis

In this study, two fundamental properties—global low-rank structure and local temporal continuity—are leveraged to model spatiotemporal WDN data with missing values. To better capture the importance of the truncation operation and autoregressive regularization in missing WDN data imputation, heat maps of LATC imputation performance are presented across diverse missing scenarios. Through a comparison of results from both real-world and simulated flow data sets, as shown in Figures 7 and 8, it is observed that the LATC model achieves optimal performance with large coefficient c and truncation r on *Z-city Flow* data set, whereas it performs best with small coefficient c and truncation r on *Sim-net Flow* data set. This result validates the importance of minimizing temporal variation for flow data imputation tasks in real-world WDNs. In other words, autoregressive regularization can effectively characterize the underlying time-varying system behaviors and model strong local patterns in actual sensor time series, which are often accompanied by noise, randomness, and fluctuations. In addition, real-world network data exhibit more complex temporal demand patterns and spatial dependencies than simulated data, resulting in a relatively large truncation for accurately representing global redundancy of *Z-city Flow* data set.

Alternatively, the same experiment is conducted on two additional data sets to provide a more thorough evaluation of LATC. Detailed plots of *Z-City Pres* data set and *Sim-Net Pres* data set are presented in Figures S2 and S3 in Supporting Information S1. The results also suggest that autoregressive regularization plays a significant role in real-world pressure data imputation tasks across all BM and MM scenarios. In contrast, the truncation mainly determines the imputation quality for the LM scenario.

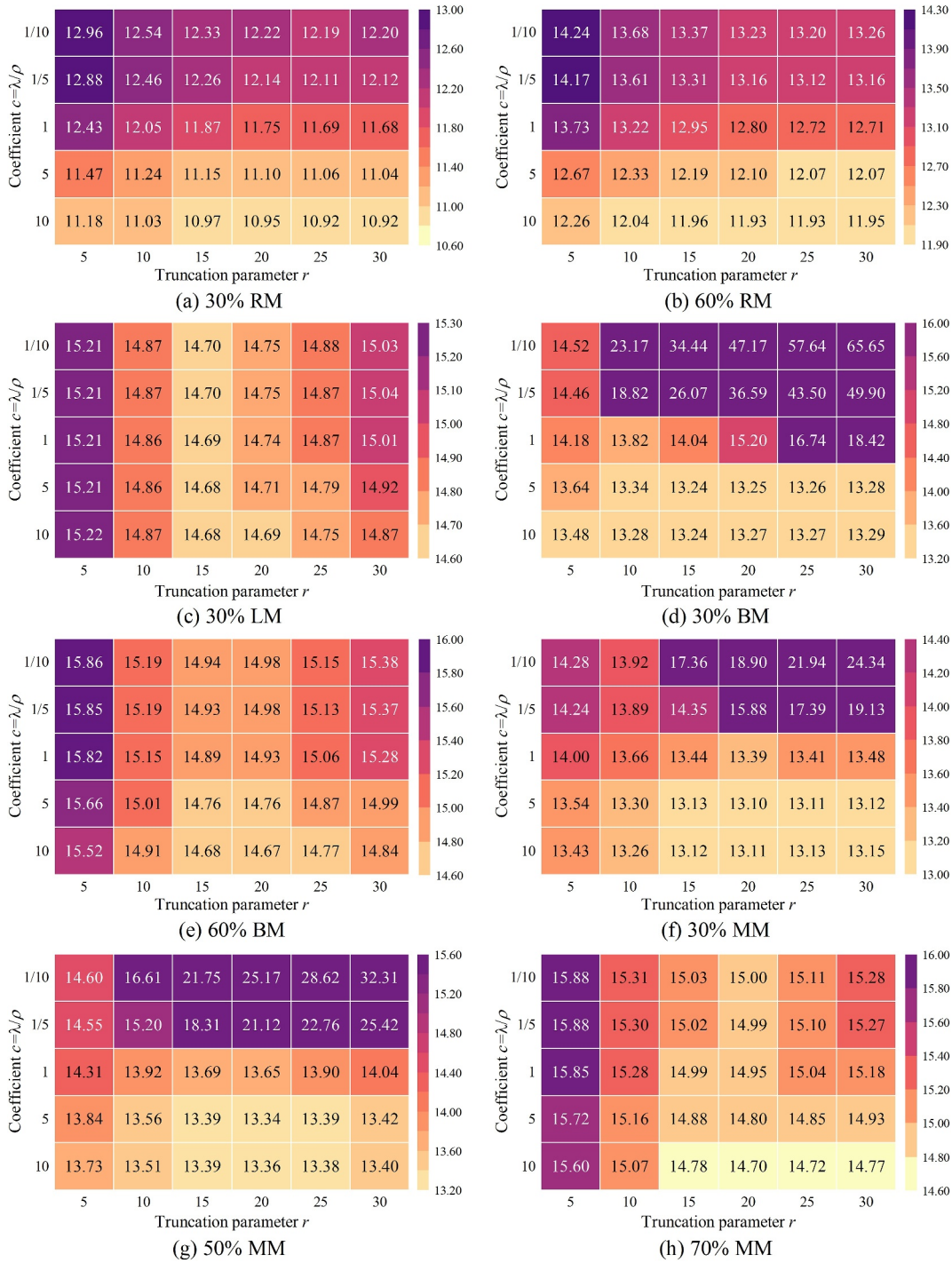

Figure 7. SMAPEs of Low-rank Autoregressive Tensor Completion imputation on *Z-city Flow* data set, with $\rho = 1 \times 10^{-5}$ for (c), (e) and (h) and $\rho = 1 \times 10^{-4}$ for other missing scenarios. The corresponding parameters of minimum Symmetric Mean Absolute Percentage Error are: (a) $c = 10, r = 25$; (b) $c = 10, r = 20$; (c) $c = 5, r = 15$; (d) $c = 10, r = 15$; (e) $c = 10, r = 20$; (f) $c = 5, r = 20$. (g) $c = 5, r = 20$; (h) $c = 10, r = 20$.

4.2.2. Ablation Study

To better investigate the impacts of the truncation operation and autoregressive regularization on imputation, two variants of LATC are designed: LATC with NN (LATC-NN) and LRTC with TNN (LRTC-TNN). Specifically, LATC-NN substitutes TNN minimization with NN minimization, while LRTC-TNN omits the temporal

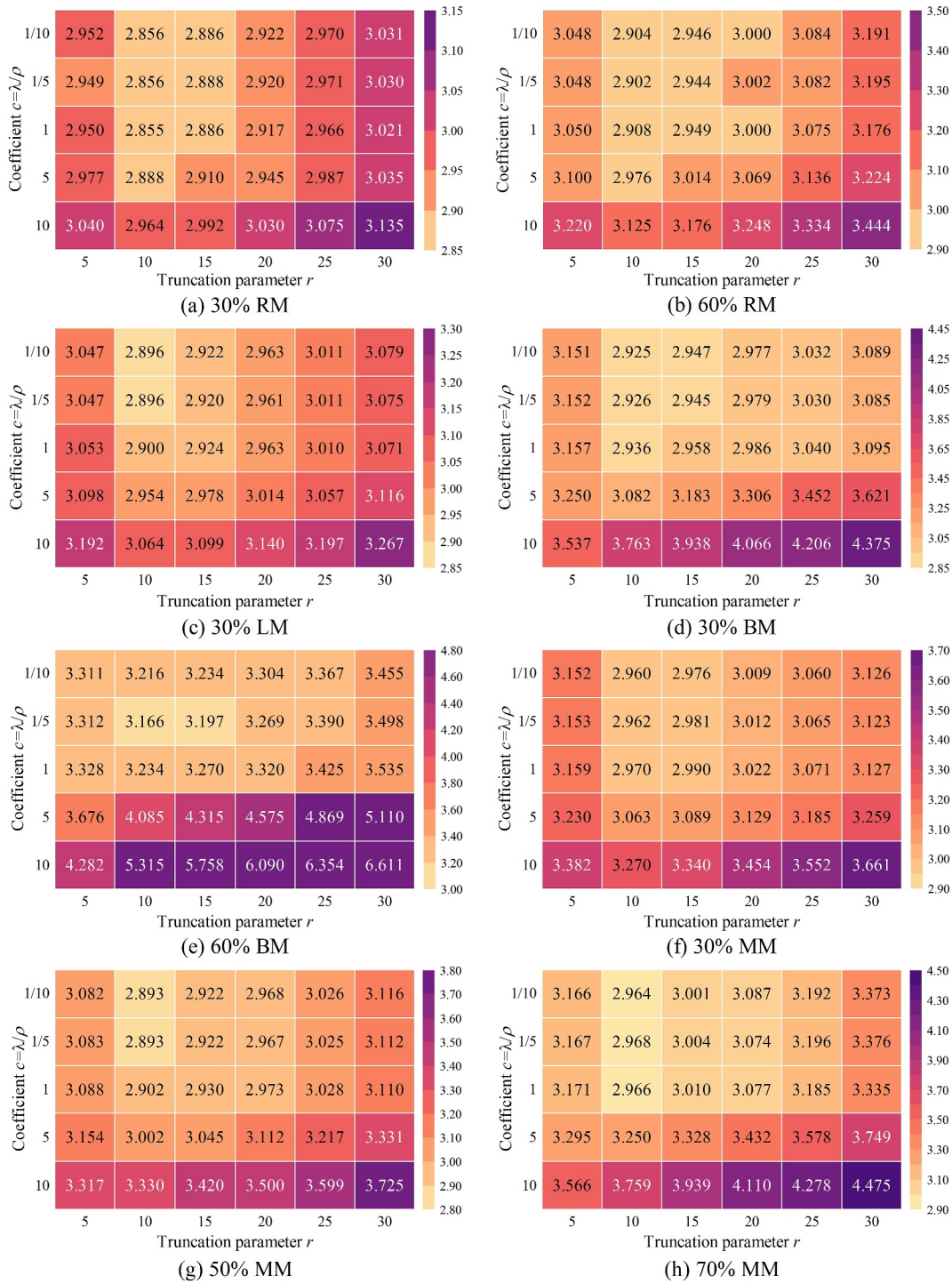

Figure 8. SMAPEs of Low-rank Autoregressive Tensor Completion imputation on *Sim-net Flow* data set, with $\rho = 1 \times 10^{-5}$ for all missing scenarios. The corresponding parameters of minimum Symmetric Mean Absolute Percentage Error are: (a) $c = 1, r = 10$; (b) $c = 1/5, r = 10$; (c) $c = 1/10, r = 10$; (d) $c = 1/10, r = 10$; (e) $c = 1/5, r = 10$; (f) $c = 1/10, r = 10$; (g) $c = 1/5, r = 10$; (h) $c = 1/10, r = 10$.

regularization term. The parameters c for LATC-NN and r for LRTC-TNN are optimized using the same search space as that of LATC. The results in Figure 9 show that LATC consistently outperforms two other models across all missing scenarios on real-world data sets.

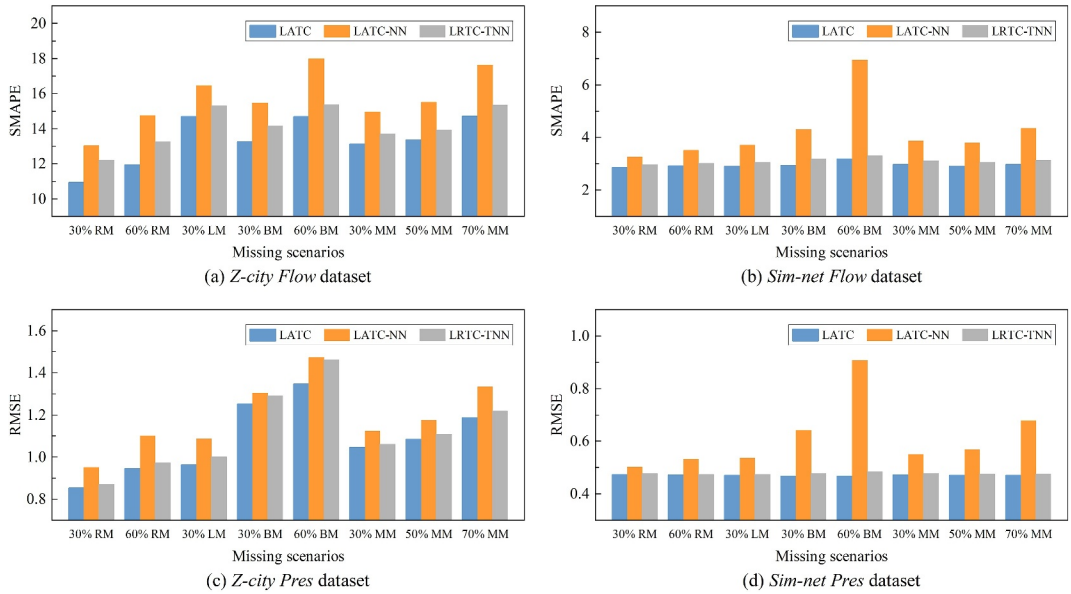

Figure 9. Bar charts of imputation performance under eight missing scenarios on four data sets. (a) and (b) report SMAPE results on *Z-city Flow* data set and *Sim-net Flow* data set, respectively; (c) and (d) report RMSE results on *Z-city Pres* data set and *Sim-net Pres* data set, respectively.

Compared to LRTC-NN, LATC achieves a 10.8%–19.0% improvement in SMAPE on *Z-city Flow* data set and a 4.0%–14.2% improvement in RMSE on *Z-city Pres* data set. This result is further validated and explained through a case analysis of the LM scenario, as shown in Figure 10. When a sensor loses observations for an entire day, the LATC model more accurately reconstructs LM data. Imputation examples in Figures 10a and 10b highlight the remarkable advantages of TNN minimization over NN minimization in capturing low-rank structure and modeling global redundancy. In particular, the pressure curve exhibits a noticeable drop at 5:45 a.m. because of the surge in water demand during the morning peak hour, and only LATC identifies this dynamic pattern. As illustrated in Figures 10c and 10d, the LM data is estimated by effectively exploiting the principal correlations from neighboring sensors and similar days.

Additionally, the strong local correlation in real-world WDN time series is encoded by minimizing the temporal variation, and the experimental results corroborate the effectiveness of this strategy. Compared to LRTC-TNN, LATC demonstrates up to a 10.6% improvement in SMAPE on *Z-city Flow* data set and an 8.0% improvement in

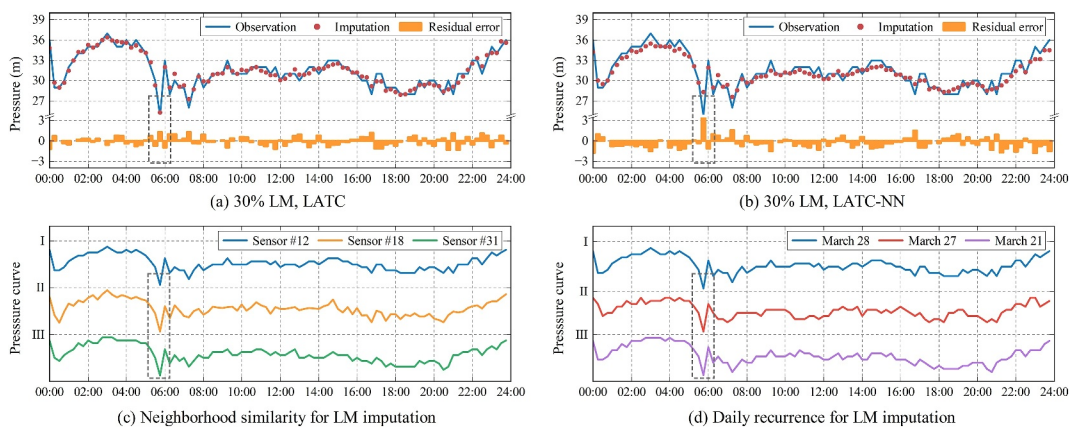

Figure 10. Visualizations of truncation effects on capturing global low-rank structure. (a) and (b) present data from sensor #12 of *Z-city Pres* data set on 28 March 2024. (c) Presents pressure curves of sensor #12 and its two spatially adjacent sensors on 28 March 2024. (d) Presents pressure curves of sensor #12 on 28 March 2024, the previous day, and the corresponding day of the previous week.

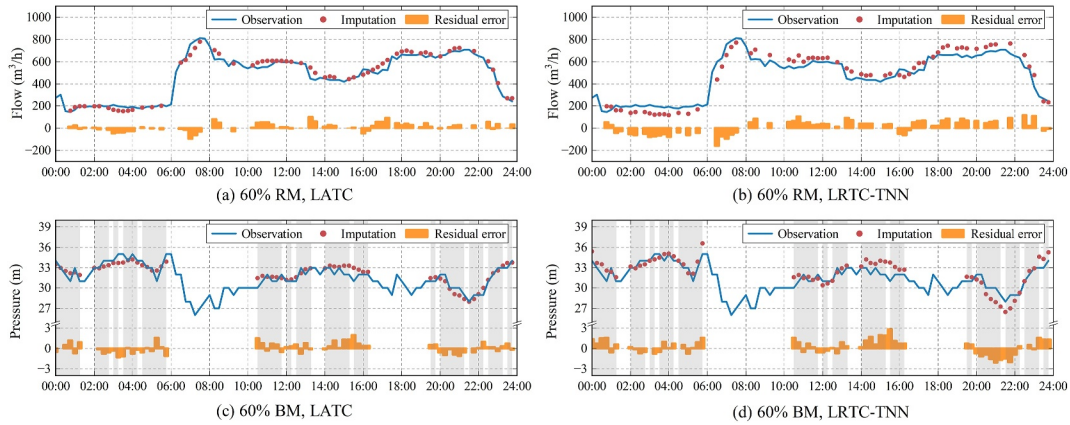

Figure 11. Visualizations of imputation performance comparison between Low-rank Autoregressive Tensor Completion and Low-Rank Tensor Completion with Truncated Nuclear Norm. (a) and (b) present data from sensor #3 of *Z-city Flow* data set on 27 August 2024. (c) and (d) present data from sensor #11 of *Z-city Pres* data set on 11 May 2024.

RMSE on *Z-city Pres* data set. Figure 11 illustrates several examples of imputation on these data sets, suggesting that autoregressive regularization is a highly effective tool for real-world WDN data imputation tasks. In contrast, the imputation performance of LATC slightly surpasses that of LRTC-TNN on both simulated data sets. That is, for simpler data sets, relying solely on low-rank models might be sufficient to achieve reasonable estimation accuracy.

5. Discussion

LATC suggests a flexible and general method for the efficient and accurate imputation of missing values in various WDN data sets. Experimental results on real-world data sets show that LATC not only surpasses state-of-the-art baseline methods in common missing scenarios, but also outperforms existing approaches in data sets characterized by high missing rates and MM patterns. Compared with previous studies on WDN data imputation, LATC achieves significant advancements in both methodology and practical applicability. For instance, Osman et al. (2018) relied on manual model selection based on missing data characteristics, resulting in high computational costs and poor scalability. LATC addresses this challenge with an efficient and unified framework capable of automatically adapting to diverse spatiotemporal missing patterns. Furthermore, when applying methods such as Chu et al. (2021) or those evaluated by Zanfei, Menapace, et al. (2022)—including KNN and Missforest—they struggle with structured missing patterns like long-range gaps and spatially continuous blocks due to limited spatiotemporal modeling capabilities. Consequently, their imputation performance often degrades severely in complex real-world WDN scenarios, failing to meet the practical requirements of SWN applications. In contrast, LATC explicitly models both global low-rank structures and local temporal dynamics, enabling robust and reliable imputation across diverse and severe missing conditions, which is essential for real-world SWN operations.

As analyzed in Section 4.2, LATC's success is mainly attributed to two factors. The first lies in its ability to characterize global redundancy inherent in spatiotemporal WDN data. In particular, the TNN component acts as a "feature selector", preserving the most prominent features that contribute significantly to data generation while minimizing the impact of noise or inessential patterns. Second, unlike conventional low-rank prior-based methods, LATC integrates autoregressive regularization, making it more effective for handling temporal dependencies in real-world WDN data sets. As shown in Figure 11, the LRTC-TNN model appears to overfit short-term fluctuations due to the lack of consideration for local correlations in sensor time series. Conversely, by incorporating an autoregressive regularization scheme, LATC greatly enhances the rationality and credibility of the imputation results, which is especially valuable for decision-making applications. The autoregressive processes impose continuity constraints on the temporal dimension, allowing the imputed values to reflect the authentic temporal dynamics and reducing the interpretive difficulty posed by discontinuous "jumps." Thus, although LATC is a data-driven imputation model based on optimization, it retains good interpretability for spatiotemporal WDN data imputation.

In developing an approach for imputing missing sensor data, the starting point of this study is the unique nature of the WDN data itself, particularly the spatiotemporal redundancy arising from the network topology, sensor placement, residential water consumption patterns, and so on. This redundancy is commonly observed in diverse WDNs, regardless of their size or components. Therefore, the proposed approach can be readily generalized to the vast majority of WDNs. However, this study does not sufficiently account for the correlation between LATC and the WDN states when modeling spatiotemporal data. If sensor data is missing during abnormal events (e.g., pipe bursts) or changes in component states (e.g., valve openings or closings), the imputed data may fail to capture the true dynamics of the pipe network. This problem can be addressed by incorporating the component state information associated with the sensor as a constraint when solving Equation 10.

Moreover, a limitation of this study is the lack of an appropriate water quality data set for experiments. Indeed, with the increasing emphasis on water quality monitoring at both the network and residential taps, parameters such as residual chlorine and ORP are being collected and analyzed in real-time (Zeng et al., 2018). Since large-scale spatiotemporal water quality data also exhibit daily similarity (Martinez Paz et al., 2022), an in-depth exploration of LATC's applicability to such data could provide valuable insights.

Overall, LATC shows promising application prospects for data quality enhancement in SWNs, and this study is expected to serve as a starting point for further exploration in this field. Specifically, the increasing popularity of smart meters (Gurung et al., 2014; Pesantez et al., 2020) and monitoring equipment is bringing new problems in acquiring, transmitting, and storing large-scale data (Mei et al., 2022). Asynchronous uploading (T. Yu et al., 2022) and compressed sensing (Wei et al., 2019) are considered effective methods to reduce the operating costs of sensor systems. Reconstructing dynamics of the entire WDN from data partially collected by these two methods can be approached as a problem of missing value imputation, aligning precisely with the application scenario of this study.

6. Conclusion

Spatiotemporal WDN data offer unprecedented opportunities for advancing SWNs. However, missing values hinder the full potential of these networks. This study treats the missing value imputation task as the low-rank completion of a third-order tensor (sensors $\times$ intervals $\times$ days) from the viewpoint of high-dimensional data analysis. By leveraging the inherent redundancies and temporal dependencies within spatiotemporal WDN data, a general LATC method is developed to achieve accurate and efficient data imputation. Extensive experiments on large-scale WDN data sets demonstrate that LATC performs significantly better than some state-of-the-art baseline models. To our knowledge, this study is the first to provide comprehensive theoretical and experimental insights into spatiotemporal data imputation for real-world WDNs with complex missing patterns.

Several research directions remain for future exploration. First, the number of sensors in the real-world data sets is still insufficient. Therefore, it is highly desirable to test whether LATC remains accurate and efficient in WDNs with denser sensor placement. Second, tensor completion techniques can separate data into low-rank and sparse components. This capability could be extended to denoising and anomaly detection, addressing noise and outliers commonly found in WDN sensor data (Lu et al., 2020). By addressing these challenges, LATC has the potential to enhance data quality and reliability in SWNs, supporting a wide range of real-time applications and improving overall water distribution management.

Appendix A: Computational Efficiency Analysis

To further assess the computational efficiency of the proposed LATC model, the running time is compared against several baseline methods under eight missing scenarios on four data sets, as shown in Table A1. The reported results are the average running time over 10 trials under the optimal parameter settings for each model. All experiments are conducted on a desktop equipped with an AMD Ryzen 7 5700G CPU and 32 GB RAM, using Python 3.11 as the programming environment. The comparison yields the following major findings regarding computational efficiency and the method's usability for real-time applications.

Firstly, in terms of computational efficiency, HaLRTC achieves the fastest runtime across all scenarios due to its simple convex optimization framework without temporal modeling. However, this simplicity results in poor imputation accuracy, especially under structured missing scenarios and on large-scale data sets, as shown in

Table A1
Running Time Comparison (in Seconds) Under Eight Missing Scenarios on Four Data Sets

Data set	Scenario	LATC	HaLRTC	TRMF	LAMC	KNN	Missforest
Z-city Flow	30% RM	481.87	**28.48**	2262.54	557.95	577.32	2849.52
	60% RM	558.55	**31.58**	2109.49	500.52	774.86	1,619.66
	30% LM	474.08	**29.44**	1,978.15	461.77	519.84	2281.71
	30% BM	459.59	**30.48**	1,313.26	646.57	/	/
	60% BM	534.36	**37.48**	1,249.26	666.65	/	/
	30% MM	397.52	**30.56**	2289.20	634.55	/	/
	50% MM	442.00	**32.97**	2147.28	607.41	/	/
	70% MM	573.97	**38.38**	2044.86	627.66	/	/
Sim-net Flow	30% RM	171.19	**3.69**	244.49	222.20	25.75	1,385.99
	60% RM	171.86	**8.14**	364.55	140.06	38.97	906.55
	30% LM	188.37	**7.11**	397.50	213.68	23.37	769.18
	30% BM	187.99	**7.31**	350.91	258.44	/	/
	60% BM	183.89	**8.27**	335.66	256.76	/	/
	30% MM	170.21	**7.47**	344.05	219.97	/	/
	50% MM	165.71	**7.85**	225.67	229.89	/	/
	70% MM	178.69	**8.73**	215.61	209.45	/	/
Z-city Pres	30% RM	187.63	**6.26**	1,008.27	201.11	191.00	494.96
	60% RM	215.15	**6.89**	982.26	177.22	219.83	488.96
	30% LM	195.22	**6.11**	1,032.68	162.31	148.30	497.08
	30% BM	195.17	**6.75**	626.14	273.94	/	/
	60% BM	208.43	**8.30**	597.75	273.81	/	/
	30% MM	197.48	**6.29**	962.61	267.45	/	/
	50% MM	204.45	**6.97**	926.48	243.57	/	/
	70% MM	228.13	**7.85**	934.47	247.32	/	/
Sim-net Pres	30% RM	68.94	**2.78**	404.28	71.91	10.52	263.46
	60% RM	75.16	**3.22**	349.28	86.59	16.70	236.54
	30% LM	69.68	**2.96**	344.62	68.98	10.79	228.18
	30% BM	72.57	**2.44**	207.83	101.74	/	/
	60% BM	73.81	**2.94**	302.24	118.86	/	/
	30% MM	67.35	**2.92**	223.07	94.09	/	/
	50% MM	69.91	**3.30**	210.17	101.56	/	/
	70% MM	76.61	**3.36**	286.86	101.53	/	/

Note. Values highlighted in boldface represent the best performance within each respective row, while values with underlines indicate the second-best performance. The runtime of KNN and Missforest is marked as "/" under BM (Block Missing) and MM (Mixed Missing) scenarios because these methods fail to produce valid results due to their inability to handle spatially continuous or mixed missing patterns.

Tables 2 and 3. In contrast, LATC achieves a better trade-off between efficiency and accuracy, delivering consistently superior imputation performance with only a moderate increase in runtime.

Compared to TRMF and Missforest—both burdened by high computational costs from iterative matrix factorization or tree-based models—LATC reduces runtime by 70%–80% on average while delivering comparable or better accuracy, especially on large-scale data sets like *Z-city Flow* and *Z-city Pres*. LATC also outperforms LAMC, which integrates autoregressive regularization, by offering similar or better efficiency and noticeably higher accuracy under BM and MM scenarios. KNN performs adequately under simple RM and LM scenarios on small data sets (e.g., *Sim-net Flow* and *Sim-net Pres*) but suffers from poor scalability. It completely fails under BM and MM scenarios on larger data sets due to its inability to handle structured and high-dimensional missing

patterns. In contrast, LATC maintains reliable performance across all data sets and missing patterns, underscoring its robustness for real-time applications in SWNs.

Overall, LATC offers a superior balance between accuracy, computational efficiency, and robustness, even under the most challenging BM and MM scenarios, making it well-suited for practical large-scale WDN data imputation tasks.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The four data sets used for the experiments in this study are available at Zenodo (Xu, 2025b). Additionally, the experimental code of this study is publicly available at Zenodo (Xu, 2025a).

Acknowledgments
This work was supported by the Funding for Excellent Contract Research Faculty (Number 113000*194232503/013), the National Natural Science Foundation of China (Number 52200119), the National Key Research and Development Program of China (Number 2023YFC3208204), and by the Israeli Water Authority (Number 2033800).

References

Asif, M. T., Mitrovic, N., Dauwels, J., & Jaillet, P. (2016). Matrix and tensor based methods for missing data estimation in large traffic networks. *IEEE Transactions on Intelligent Transportation Systems*, *17*(7), 1816–1825. https://doi.org/10.1109/TITS.2015.2507259

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(S3), 74. https://doi.org/10.1186/s12911-016-0318-z

Boyle, C., Ryan, G., Bhandari, P., Law, K. M. Y., Gong, J., & Creighton, D. (2022). Digital transformation in water organizations. *Journal of Water Resources Planning and Management*, *148*(7), 03122001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001555

Chen, C., He, B., Ye, Y., & Yuan, X. (2016). The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, *155*(1–2), 57–79. https://doi.org/10.1007/s10107-014-0826-5

Chen, X., Lei, M., Saunier, N., & Sun, L. (2022). Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 12301–12310. https://doi.org/10.1109/TITS.2021.3113608

Chen, X., & Sun, L. (2021). Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 4659–4673. https://doi.org/10.1109/TPAMI.2021.3066551

Chu, S., Zhang, T., Shao, Y., Yu, T., & Yao, H. (2020). Numerical approach for water distribution system model calibration through incorporation of multiple stochastic prior distributions. *Science of The Total Environment*, *708*, 134565. https://doi.org/10.1016/j.scitotenv.2019.134565

Chu, S., Zhang, T., Xu, C., Yu, T., & Shao, Y. (2021). Dealing with data missing and outlier to calibrate nodal water demands in water distribution systems. *Water Resources Management*, *35*(9), 2863–2878. https://doi.org/10.1007/s11269-021-02873-9

Eggimann, S., Mutzner, L., Wani, O., Schneider, M. Y., Spuhler, D., Moy De Vitry, M., et al. (2017). The potential of knowing more: A review of data-driven urban water management. *Environmental Science & Technology*, *51*(5), 2538–2553. https://doi.org/10.1021/acs.est.6b04267

Fu, G., Jin, Y., Sun, S., Yuan, Z., & Butler, D. (2022). The role of deep learning in urban water management: A critical review. *Water Research*, *223*, 118973. https://doi.org/10.1016/j.watres.2022.118973

Gurung, T. R., Stewart, R. A., Sharma, A. K., & Beal, C. D. (2014). Smart meters for enhanced water supply network modelling and infrastructure planning. *Resources, Conservation and Recycling*, *90*, 34–50. https://doi.org/10.1016/j.resconrec.2014.06.005

Hajgató, G., Paál, G., & Gyires-Tóth, B. (2020). Deep reinforcement learning for real-time optimization of pumps in water distribution systems. *Journal of Water Resources Planning and Management*, *146*(11), 04020079. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001287

Håstad, J. (1990). Tensor rank is NP-complete. *Journal of Algorithms*, *11*(4), 644–654. https://doi.org/10.1016/0196-6774(90)90014-6

Hu, Y., Zhang, D., Ye, J., Li, X., & He, X. (2013). Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(9), 2117–2130. https://doi.org/10.1109/TPAMI.2012.271

Hu, Z., Chen, W., Wang, H., Tian, P., & Shen, D. (2022). Integrated data-driven framework for anomaly detection and early warning in water distribution system. *Journal of Cleaner Production*, *373*, 133977. https://doi.org/10.1016/j.jclepro.2022.133977

Huang, L.-T., So, H. C., Chen, Y., & Wang, W.-Q. (2014). Truncated nuclear norm minimization for tensor completion. In *2014 IEEE 8th sensor array and Multichannel Signal processing workshop (SAM)* (pp. 417–420). https://doi.org/10.1109/SAM.2014.6882431

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500. https://doi.org/10.1137/07070111X

Liu, J., Musialski, P., Wonka, P., & Ye, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 208–220. https://doi.org/10.1109/TPAMI.2012.39

Lu, C., Feng, J., Chen, Y., Liu, W., Lin, Z., & Yan, S. (2020). Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 925–938. https://doi.org/10.1109/TPAMI.2019.2891760

Lu, C., Peng, X., & Wei, Y. (2019). Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 5989–5997). https://doi.org/10.1109/CVPR.2019.00615

Martinez Paz, E. F., Tobias, M., Escobar, E., Raskin, L., Roberts, E. F. S., Wigginton, K. R., & Kerkez, B. (2022). Wireless sensors for measuring drinking water quality in building plumbing: Deployments and insights from continuous and intermittent water supply systems. *ACS ES&T Engineering*, 2(3), 423–433. https://doi.org/10.1021/acsestengg.1c00259

Mei, L., Zhou, J., Li, S., Cai, M., & Li, T. (2022). Leak identification based on CS-ResNet under different leakage apertures for water-supply pipeline. *IEEE Access*, *10*, 57783–57795. https://doi.org/10.1109/ACCESS.2022.3177595

Nie, T., Qin, G., & Sun, J. (2022). Truncated tensor schatten p -norm based approach for spatiotemporal traffic data imputation with complicated missing patterns. *Transportation Research Part C: Emerging Technologies*, *141*, 103737. https://doi.org/10.1016/j.trc.2022.103737

Oberascher, M., Rauch, W., & Sitzenfrei, R. (2022). Towards a smart water city: A comprehensive review of applications, data requirements, and communication technologies for integrated management. *Sustainable Cities and Society*, *76*, 103442. https://doi.org/10.1016/j.scs.2021.103442

Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access*, *6*, 63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Pesantez, J. E., Berglund, E. Z., & Kaza, N. (2020). Smart meters data for modeling and forecasting water demand at the user-level. *Environmental Modelling & Software*, *125*, 104633. https://doi.org/10.1016/j.envsoft.2020.104633

Rubin, D. B. (1975). Inference and missing data. *ETS Research Bulletin Series*, *1975*(1), i–19. https://doi.org/10.1002/j.2333-8504.1975.tb01053.x

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salloom, T., Kaynak, O., & He, W. (2021). A novel deep neural network architecture for real-time water demand forecasting. *Journal of Hydrology*, *599*, 126353. https://doi.org/10.1016/j.jhydrol.2021.126353

Salomons, E., & Housh, M. (2020). A practical optimization scheme for real-Time operation of water distribution systems. *Journal of Water Resources Planning and Management*, *146*(4), 04020016. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001188

Sanchez, G. M., Terando, A., Smith, J. W., García, A. M., Wagner, C. R., & Meentemeyer, R. K. (2020). Forecasting water demand across a rapidly urbanizing region. *Science of The Total Environment*, *730*, 139050. https://doi.org/10.1016/j.scitotenv.2020.139050

Sivagurunathan, V., Elsawah, S., & Khan, S. J. (2022). Scenarios for urban water management futures: A systematic review. *Water Research*, *211*, 118079. https://doi.org/10.1016/j.watres.2022.118079

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Takeuchi, K., Kashima, H., & Ueda, N. (2017). Autoregressive tensor factorization for spatio-temporal predictions. In V. Raghavan, S. Aluru, G. Karypis, L. Miele, & X. Wu (Eds.), *2017 17TH IEEE INTERNATIONAL CONFERENCE ON DATA MINING (ICDM)* (pp. 1105–1110). IEEE. https://doi.org/10.1109/ICDM.2017.146

Wang, J.-L., Huang, T.-Z., Zhao, X.-L., Jiang, T.-X., & Ng, M. K. (2021). Multi-dimensional visual data completion via low-rank tensor representation under coupled transform. *IEEE Transactions on Image Processing*, *30*, 3581–3596. https://doi.org/10.1109/TIP.2021.3062995

Wei, Z., Pagani, A., & Guo, W. (2019). Monitoring networked infrastructure with minimum data via sequential graph fourier transforms. In *2019 IEEE international smart Cities Conference (ISC2)* (pp. 703–708). https://doi.org/10.1109/ISC246665.2019.9071735

Wu, Y., Wang, X., Liu, S., Yu, X., & Wu, X. (2023). A weighting strategy to improve water demand forecasting performance based on spatial correlation between multiple sensors. *Sustainable Cities and Society*, *93*, 104545. https://doi.org/10.1016/j.scs.2023.104545

Xu, A. (2025a). *Ang-xu/WDS-data-imputation: V1.0* (version v1.0) [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.15039291

Xu, A. (2025b). *Ang-xu/WDS-data-imputation-PaperData: V1.0* (version v1.0) [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.15039293

Xue, S., Qiu, W., Liu, F., & Jin, X. (2018). Low-rank tensor completion by truncated nuclear norm regularization. In *2018 24th international conference on pattern recognition (ICPR)* (pp. 2600–2605). https://doi.org/10.1109/ICPR.2018.8546008

Yu, H.-F., Rao, N., & Dhillon, I. S. (2016). Temporal regularized matrix factorization for high-dimensional time series prediction. *Advances in Neural Information Processing Systems*, *29*(NIPS 2016), 29. https://webofscience.clarivate.cn/wos/alldb/full-record/WOS:000458973701017

Yu, T., Lin, B., Long, Z., Shao, Y., Lima Neto, I. E., & Chu, S. (2022). Asynchronous sensor networks for Nodal water demand estimation in water distribution systems based on sensor grouping analysis. *Journal of Cleaner Production*, *365*, 132676. https://doi.org/10.1016/j.jclepro.2022.132676

Yuan, M., & Zhang, C.-H. (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, *16*(4), 1031–1068. https://doi.org/10.1007/s10208-015-9269-5

Zanfei, A., Brentan, B. M., Menapace, A., Righetti, M., & Herrera, M. (2022). Graph convolutional recurrent neural networks for water demand forecasting. *Water Resources Research*, *58*(7), e2022WR032299. https://doi.org/10.1029/2022WR032299

Zanfei, A., Menapace, A., Brentan, B. M., & Righetti, M. (2022). How does missing data imputation affect the forecasting of urban water demand? *Journal of Water Resources Planning and Management*, *148*(11), 04022060. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001624

Zeng, D., Zhang, S., Gu, L., Yu, S., & Fu, Z. (2018). Quality-of-sensing aware budget constrained contaminant detection sensor deployment in water distribution system. *Journal of Network and Computer Applications*, *103*, 274–279. https://doi.org/10.1016/j.jnca.2017.10.018

Zhang, Y., Tu, Z., Lu, J., Xu, C., & Shen, L. (2024). Fusion of low-rankness and smoothness under learnable nonlinear transformation for tensor completion. *Knowledge-Based Systems*, *296*, 111917. https://doi.org/10.1016/j.knosys.2024.111917

Zhou, X., Liu, S., Xu, W., Xin, K., Wu, Y., & Meng, F. (2022). Bridging hydraulics and graph signal processing: A new perspective to estimate water distribution network pressures. *Water Research*, *217*, 118416. https://doi.org/10.1016/j.watres.2022.118416

Zhou, X., Tang, Z., Xu, W., Meng, F., Chu, X., Xin, K., & Fu, G. (2019). Deep learning identifies accurate burst locations in water distribution networks. *Water Research*, *166*, 115058. https://doi.org/10.1016/j.watres.2019.115058

Zhou, X., Zhang, J., Guo, S., Liu, S., & Xin, K. (2023). A convenient and stable graph-based pressure estimation methodology for water distribution networks: Development and field validation. *Water Research*, *233*, 119747. https://doi.org/10.1016/j.watres.2023.119747