A Building-level Digital Twin Approach to improve Consumption Data Imputation in Smart Water Grids

Themistoklis Sarantakos
Industrial Systems Institute, Athena Research Center
Patras, Greece
ORCiD:0000-0002-7517-6997

Dimitrios Amaxilatis
Spark Works Ltd.
Galway, Ireland
d.amaxilatis@sparkworks.net
ORCiD:0000-0001-9938-6211

Georgios Mylonas
Industrial Systems Institute, Athena Research Center
Patras, Greece
ORCiD:0000-0003-2128-720X

Ioannis Chatzigiannakis
Sapienza University of Rome
Rome, Italy
ichatz@diag.uniroma1.it
ORCiD:0000-0001-8955-9270

***Abstract*—Smart Water Grids are quickly becoming an important aspect in smart cities, due to the rising importance of water as a critical resource, as well as the need to improve the efficiency of such networks. In this work, we introduce a novel approach for data imputation in smart water grids based on a digital twin logic, demonstrating that clustering based on water consumption profiles of building-level Digital Twins can substantially outperform building-specific models. We evaluate 7 state-of-the-art imputation algorithms across 5 missing data scenarios, using real-world datasets from institutional and residential/commercial buildings. Our results show that profile-based Digital Twin models can achieve Mean Absolute Error reductions of 40–99% compared to per-building approaches, with the most regular consumption profiles reaching near-perfect imputation accuracy (as low as 0.002). Beyond improved accuracy, this approach reduces operational complexity by replacing numerous building-specific models with a small number of profile-based models, while also enabling seamless integration of new buildings without retraining. Moreover, this can enhance the reliability of water consumption data for critical applications like leak detection, predictive maintenance, and water conservation planning.**

***Index Terms*—Digital Twins, Time Series Imputation, Water Metering, Machine Learning, K-Means Clustering**

I. Introduction

Global water consumption is rising due to population growth and urbanization, while climate change alters distribution and consumption patterns, exacerbating water scarcity challenges. Despite optimizations in Water Distribution Systems (WDS), outdated infrastructure and abnormal demand spikes lead to high leakage rates and intermittent supply. Within this landscape, the digitization of water infrastructure is accelerating globally, driven by growing digital capabilities and data availability. Technologies such as the Internet of Things (IoT) and Machine Learning (ML), can enable more accurate WDS modeling through Digital Twins (DTs), facilitating real-time monitoring, operational testing, and emergency response [1]. Smart water meters can serve as the sensory foundation, providing real-time consumption data when combined to building-level DTs. However, DT effectiveness is always constrained by data quality. Despite advances in metering technology, data gaps from communication failures, sensor malfunctions, and network disruptions remain. Such gaps compromise DT fidelity, undermining the capacity to accurately represent physical counterparts and support critical applications such as leak detection, consumption forecasting, and predictive maintenance.

Traditional data imputation approaches treat each monitored building as an isolated unit, training dedicated models on individual consumption histories. While capturing building-specific patterns, this approach fails to leverage the fact that buildings with similar operational schedules, occupancy patterns, and functional purposes exhibit comparable consumption profiles. By recognizing these similarities, we conceptualize a DT network where individual building DTs are interconnected entities that inform and enhance each other rather than isolated components.

This work introduces a consumption-profile-based clustering approach for data imputation in smart water grids, shifting from building-centric DTs to profile-based representations. Rather than training separate imputation models for each building, we group buildings by consumption characteristics and train shared models that enable cross-building knowledge transfer. This approach aligns with the vision of next-generation intelligent pervasive systems, where DTs interact and collaborate to achieve holistic domain understanding. Our methodology leverages data from real-world deployments: over 50 smart water meters across 22 institutional buildings at a university campus in Greece, and residential/commercial buildings in Alicante, Spain. We compare seven machine learning techniques—k-Nearest Neighbors (kNN), MissForest, SAITS, Transformers, TimesNet, USGAN, and MRNN—under two experimental paradigms: (1) per-building DT models trained on individual building data, and (2) per-profile DT models trained on aggregated

This work was partially supported by the European Union's Horizon Program under the Agile and Cognitive Cloud edge Continuum management (AC3) project (Grant No. 101093129), the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No. 101192750, and by the project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

ISBN 978-3-903176-82-9

data from buildings sharing similar consumption characteristics. Using multi-year datasets with naturally occurring data gaps ranging from 20% to 79%, we demonstrate the applicability of our approach in smart water systems.

This work makes 3 contributions: a) we introduce a consumption-profile-based framework that organizes buildings into clusters, enabling cross-building knowledge transfer within smart water grids, b) we systematically compare 7 state-of-the-art imputation techniques under both isolated (per-building) and ecosystem (per-profile) configurations across 5 missing data scenarios, providing insights into model performance across diverse consumption patterns, c) using multi-year datasets with data gaps between 20%-79%, we demonstrate the applicability of our approach in smart water grids.

Our results show that consumption-profile-based models, suitable for DTs, achieve superior imputation performance compared to building-specific ones, with Mean Absolute Error (MAE) reductions of 40–99% depending on consumption profile characteristics. This finding demonstrates that the profile-based DT paradigm, where similar DTs/buildings share knowledge and support mutual enhancement, offers significant advantages over treating each one as an isolated entity.

II. Related Work

With respect to the digital transition of water networks and DTs in WDS, [2] provides an overview of the respective trends in recent years, concluding that DTs will be foundational in the digital transition towards more efficient WDS. A survey of DT applications in the water sector is provided in [3], arguing that DTs can facilitate the digital implementation of sustainable and adaptive WDS. Regarding real-world pilots of such DTs, [4] discussed the application of a DT modelling part of the WDS in the island of Madeira, Portugal. The authors reported that simulation results aiming to minimize water leakage point to potential savings of 15%. In another related study [5], a DT facilitated the implementation of a smart water grid management system in Singapore, in order to enable better anomaly detection and localization in the network, with an accuracy of more than 80% for detecting anomaly events.

As regards data imputation, recent work like [6] provided a review of imputation methods ranging from simple deletion to advanced ML techniques, focusing specifically on WDS. They also introduce a method for selecting the most suitable imputation technique, based on data analysis and missing data characteristics, along with a discussion of each method's strengths and weaknesses. [7] proposed to enhance the Multivariate Imputation by Chained Equations (MICE) method by reshaping sensor data to strengthen the correlation between observed and missing data, thereby improving imputation accuracy. This approach is demonstrated through its application to water quality monitoring data, showing a significant improvement in model accuracy with at least a 23% increase in R^2 values.

Concerning water quality monitoring, [8] provided a review of various approaches to the handling of missing data in near real-time environmental monitoring systems. A missing data imputation system was also presented, with similarities to our work. The authors conclude that the size of the missing data and the method selected greatly affect data imputation performance, with large data gaps dealt with significantly better by neural network-based methods.

As regards the use of specific methods in missing data imputation, [9] examined the effectiveness of the K-Nearest Neighbors method across various scenarios with different missing data mechanisms. The study validates KNN's robustness in maintaining high data integrity and closely matching the accuracy of complete datasets, underscoring its utility in diverse conditions. MissForest, a non-parametric method that leverages the random forest approach for imputing missing values in datasets containing both continuous and categorical variables was introduced in [10]. This method capitalizes on the random forest's ability to handle complex data structures and interactions without a predefined parametric form, demonstrating superior performance over other imputation methods, particularly in complex datasets. A self-attention mechanism was employed in [11] to enhance the imputation of missing values in multivariate time series data, showing superior performance through extensive experiments on real-world datasets. This method's robustness is highlighted by its dynamic adjustment of weights in response to missing information. [12] presented the TrAdaBoost-LSTM model, which combines LSTM neural networks with instance-based transfer learning to address large-scale consecutive missing data, particularly in water quality datasets. This model significantly improves imputation accuracy by leveraging existing complete datasets to enhance missing data prediction. TimesNet was introduced in [13], allowing the effective application of 2D convolutional operations and enhancing performance across tasks like forecasting, classification, and anomaly detection. Lastly, GRU-D [14] incorporated missing data patterns directly into its architecture, enhancing the handling and prediction accuracy of multivariate time series data.

III. Input Datasets

Tethys: The first dataset we used was provided in the context of the Tethys deployment [15], an Edge-Computing-ready water metering system implemented at Aristotle University of Thessaloniki, Greece. The system monitored water consumption across 22 university buildings using over 50 IoT-enabled smart water meters connected via wM-Bus and LoRaWAN. The dataset comprises hourly aggregated consumption data for 6 years, normalized to address varying reporting intervals across different meter hardware. Missing data ranges from 20% to 79% per building due to wireless communication failures, sensor malfunctions, and network disruptions which are typical challenges in real-world IoT deployments. While cumulative metering tolerates occasional gaps for billing, constructing accurate DTs for advanced analytics requires complete, high-resolution data.

Alicante: Our second dataset, the Alicante Smart Water Management dataset [16], comes from Alicante, Spain. It

originates from a smart water metering deployment and provides complementary data for evaluating our approach in a different urban setting. It contains water consumption measurements from residential and commercial buildings equipped with smart meters. Unlike Tethys, which focuses on university buildings, this dataset encompasses diverse building types with varying occupancy patterns and consumption behaviors, including residential apartments, office buildings, and commercial establishments. This dataset does not include missing data across the monitored buildings.

The inclusion of both datasets allows assessing whether consumption-profile-based DT clustering remains effective across different building typologies (institutional vs. residential/commercial), locations with different climate patterns and water usage norms, and missing data patterns. For both datasets, we apply consistent preprocessing steps: (1) hourly aggregation to normalize temporal resolution, (2) identification and marking of missing values, and (3) extraction of consumption patterns for profile clustering. This enables direct comparison of imputation performance across deployments.The combination of building types provides a diverse testbed for evaluating our approach.

IV. Building Profiling

A fundamental premise of the DT paradigm is that individual twins sharing similar characteristics can benefit from collective knowledge and shared models. In the context of WDS, buildings with similar operational patterns, occupancy behaviors, and functional purposes exhibit comparable consumption profiles. Rather than treating each building's DT in isolation, we propose clustering buildings based on their consumption characteristics to create profile-based DT groups.

A. Feature Extraction

To identify consumption similarities across buildings, we extract a comprehensive set of features from the available hourly consumption data. These features are designed to be scale-independent and capture the shape, temporal patterns, and behavioral characteristics of each building's consumption profile rather than absolute consumption magnitudes. This approach ensures that buildings are clustered by their behavioral patterns rather than their size or overall consumption volume. Our feature set comprises 10 key metrics for each building i, which are extracted from the complete available consumption data, resulting in a feature vector $\mathbf{f}_i \in \mathbb{R}^{10}$:

Variability and Distribution Features: *1. Coefficient of Variation (cv_hourly)* ratio of standard deviation to mean consumption, measuring relative variability independent of scale; *2. Night Valley Depth* ratio of minimum to maximum hourly consumption, indicating the prominence of day/night patterns; *3. Peak-to-Base Ratio* normalized measure of consumption range relative to average consumption.

Temporal Pattern Features *4. Peak Hour* the hour of day when maximum consumption typically occurs; *5. Number of Peaks* count of local maxima in the average daily consumption profile; *6. Bimodal Indicator* flag identifying buildings with distinct morning and evening peaks.

Shape Characteristics: *7. Peak Sharpness* concentration of consumption around peak hours, measuring how pronounced peaks are; *8. Flatness Index* measure of consumption uniformity across hours, with higher values indicating more constant usage; *9. Smoothness Index* quantifies the gradualness of transitions between consumption levels.

Temporal Dynamics: *10. Trend Reversals* number of directional changes in hourly consumption.

B. Clustering Methodology

We apply K-means clustering to identify natural groupings of buildings with similar consumption profiles. K-means is well-suited for this task as it efficiently partitions buildings into compact, well-separated clusters based on Euclidean distance. To identify the appropriate number of clusters k, we evaluate multiple configurations using two complementary metrics. **Silhouette Score** measures how well each building fits within its assigned cluster relative to other clusters. Scores range from -1 to 1, with values above 0.5 indicating good cluster separation. **Elbow Method** examines the within-cluster sum of squares (inertia) as a function of k, identifying the point where additional clusters provide diminishing returns. The optimal number of clusters is determined by maximizing the silhouette score while considering the elbow point in the inertia curve. This approach balances statistical optimality with interpretability. Our analysis revealed that $k = 3$ clusters provide the best trade-off between cluster cohesion and separation. Figure 1 presents both the silhouette scores and elbow curve across different values of k, illustrating the selection process.

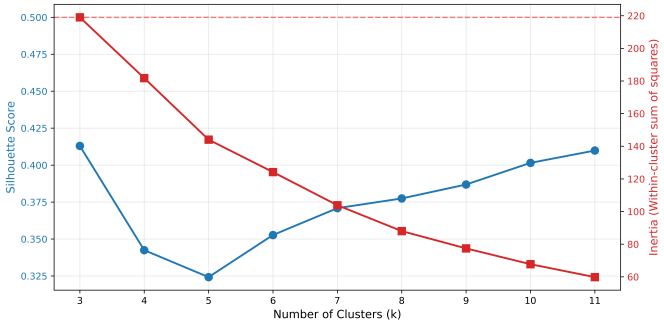

Fig. 1. Selection of optimal cluster count (k) using the Silhouette Score and Inertia-Elbow Method. The blue line indicates cluster cohesion (peaking at $k = 3$), while the red line tracks the within-cluster sum of squares, showing an "elbow" at $k = 3$.

Once the optimal clustering is identified, we validate and interpret each cluster by its consumption patterns:

Variability Level: Low CV (< 0.3) indicates stable consumption; moderate CV (0.3–0.6) suggests scheduled variations; high CV (> 0.6) implies erratic patterns.

Temporal Profile: Peak hour identifies dominant usage periods (6AM–10AM, 11AM–16PM, 17PM–21PM, 22PM–5AM).

Operational Pattern: Night valley depth reveals 24-hour operations (> 0.5) versus clear day/night cycles (< 0.2).

Usage Type: Bimodal patterns suggest residential-like

behavior with morning and evening peaks; unimodal patterns indicate institutional or commercial usage.

Table I summarizes the identified consumption profiles that represent our DT structure, including the number of buildings per cluster, dominant characteristics, and interpreted operational patterns for both Tethys and Alicante datasets. Each cluster C_j represents a consumption profile group containing buildings $\{b_1, b_2, \ldots, b_n\} \in C_j$ that share similar behavioral characteristics. For data imputation purposes, we train a single shared model for each cluster using the combined training data from all member buildings.

To assign a new building to an existing cluster, we:
a) Extract the 10 consumption features from available data.
b) Use trained K-means model to predict cluster membership.
c) Apply the corresponding profile-based imputation model.

V. Imputation Pipeline

Our imputation methodology leverages a DT approach defined by consumption profile clustering. The pipeline consists of three main steps: data preprocessing, model training, and data imputation. Here we describe the key aspects of this pipeline, with particular emphasis on how profile-based clustering modifies the training strategy.

A. Data Preprocessing

We preprocess the consumption data through three stages:

1) **Missing Value Identification**: Mark all missing data points as NumPy `nan` values.
2) **Temporal Segmentation**: Partition data into 24-hour vectors to exploit daily periodicity patterns, where consumption typically peaks during working hours and decreases during nighttime.
3) **Dataset Splitting**: Separate complete days (*non-missing-days*) and days with gaps (*missing-days*). Complete days are split in 80% training and 20% validation sets.

This helps to accurately evaluate imputation quality by testing on known values that are temporarily hidden.

B. Model Training

We evaluate seven state-of-the-art imputation algorithms: K-Nearest Neighbors (KNN - A), MissForest (B), Self-Attention-based Imputation for Time Series (SAITS - C), Transformer (D), TimesNet (E), Unsupervised GAN (US-GAN - F), and Multi-Directional Recurrent Neural Network (MRNN - G).

The critical distinction in this work lies in how we construct training datasets. We compare two approaches:

- **Per-Building Models (Baseline)**: Each building's DT uses a dedicated imputation model trained exclusively on that building's historical data.
- **Per-Profile Models (Proposed)**: Buildings within the same consumption profile cluster C_j (identified in Section IV-B) share a single imputation model trained on the aggregated data from all cluster members. This approach embodies the DT paradigm where similar twins collaborate and share knowledge.

For the per-profile approach, the training dataset for cluster C_j is constructed by concatenating the training data from all buildings $\{b_1, b_2, \ldots, b_n\} \in C_j$. This significantly increases the training sample size and exposes the model to diverse manifestations of the same underlying consumption pattern. All neural network models use the Adam optimizer with a learning rate of 0.001, trained for 100 epochs with early stopping based on validation loss. The best-performing model checkpoint is retained for evaluation.

Parameter	C_1	C_2	C_3
cv_hourly	2.23 ± 0.73	0.54 ± 0.36	0.69 ± 0.17
peak_hour	3.8 ± 4.8	14.7 ± 7.3	14.5 ± 0.7
night_valley_depth	0.028 ± 0.023	0.285 ± 0.236	0.103 ± 0.028
peak_sharpness	18.37 ± 9.71	4.57 ± 6.64	1.81 ± 0.52
number_of_peaks	1.6 ± 0.7	7.5 ± 2.5	4.5 ± 2.1
peak_to_base_ratio	10.66 ± 3.84	1.86 ± 1.15	2.42 ± 0.38
flatness_index	0.79 ± 0.02	0.72 ± 0.07	0.71 ± 0.02
trend_reversals	12.7 ± 2.6	19.4 ± 4.1	14.5 ± 6.4
Bimodal buildings	10/21 (47.6%)	23/24 (95.8%)	2/2 (100.0%)

TABLE I
SUMMARY OF THE CONSUMPTION PROFILE CLUSTERS IDENTIFIED IN THE DT.

VI. Evaluation

A. Experimental Setup

To comprehensively evaluate the effectiveness of profile-based DT models compared to building-specific approaches, we design a systematic evaluation framework that tests imputation performance under varying missing data scenarios representative of real-world deployment conditions.

Missing Data Scenarios: We evaluate each imputation model across five distinct scenarios, systematically varying both the number and distribution of missing values within 24-hour consumption vectors. The number of missing values ranges from 1, 2, or 3 missing hourly measurements per 24-hour vector. For the distribution patterns, we consider two cases: random missing values are randomly distributed across the 24-hour period, simulating sporadic sensor failures or intermittent transmission errors, while sequential missing values occur consecutively, simulating sustained communication outages or extended sensor malfunctions.

This yields 5 evaluation configurations: single (R1), 2 random (R2), 2 sequential (S2), 3 random (R3), and 3 sequential (S3) missing values. Sequential missing values present a more challenging imputation task, as they eliminate temporal context from contiguous periods, whereas randomly distributed gaps preserve surrounding information for the models to exploit. For each configuration, we randomly select 20% of the complete 24-hour validation vectors and hide the specified number of values according to the chosen pattern. The same hidden positions are used across all models to ensure fair comparison. We assess quality using Mean Absolute Error (MAE), where lower values indicate better accuracy.

B. Results on Per-Building Models

Table II presents the MAE for all 7 algorithms across the 5 missing data scenarios on both datasets using per-building models, trained on its own consumption data. Transformer (D) consistently performs better across both datasets, with MAE values ranging from 0.308 to 0.328 on Tethys and 0.153 to 0.214 on Alicante. In contrast, MRNN (G) consistently exhibits the worst performance, indicating that this architecture struggles when trained on limited per-building data. The remaining algorithms (KNN (A), MissForest (B), SAITS (C), TimesNet (E), USGAN (F)) show intermediate performance. Notably, Alicante exhibits lower absolute MAE values across all algorithms, suggesting more predictable consumption patterns in residential/commercial buildings.

The impact of missing data scenarios on performance is modest. Single missing values (R1) establish baseline performance, while multiple random gaps (R2, R3) show only minor degradation (typically <0.02 MAE increase). Sequential gaps (S2, S3) occasionally show slightly higher MAE, particularly for S3 on Alicante where consecutive missing values eliminate more temporal context. This limited degradation suggests that per-building models leverage strong daily consumption patterns effectively, though the relatively high MAE values indicate substantial room for improvement.

Algorithm	R1	R2	R3	S2	S3
Tethys					
A (KNN)	0.345	0.350	0.342	0.335	0.354
B (MissForest)	0.358	0.380	0.366	0.375	0.417
C (SAITS)	0.319	0.331	0.334	0.323	0.329
D (Transformer)	0.308	0.323	0.326	0.313	0.328
E (TimesNet)	0.528	0.547	0.530	0.573	0.570
F (USGAN)	0.332	0.345	0.352	0.354	0.374
G (MRNN)	0.616	0.590	0.601	0.605	0.614
Alicante					
A (KNN)	0.270	0.300	0.234	0.281	0.290
B (MissForest)	0.192	0.228	0.181	0.310	0.375
C (SAITS)	0.194	0.209	0.185	0.246	0.220
D (Transformer)	0.163	0.214	0.153	0.201	0.198
E (TimesNet)	0.277	0.434	0.390	0.434	0.509
F (USGAN)	0.210	0.227	0.186	0.268	0.280
G (MRNN)	0.730	0.721	0.689	0.718	0.716

TABLE II
MAE FOR ALL ALGORITHMS IN ALL 5 SCENARIOS ACROSS TETHYS AND ALICANTE DATASETS PER BUILDING MODELS.

C. Results on Per-Profile Models

Table III presents imputation performance using profile-based DT models, where buildings within the same consumption cluster (C_0, C_1, C_2) share a common imputation model trained on aggregated data from all members, embodying the DT paradigm through cross-building knowledge transfer.

The 3 consumption profile clusters exhibit dramatically different imputation characteristics. Cluster C_0 achieves exceptional accuracy, representing an order of magnitude improvement over per-building models, with MissForest (B) and KNN (A) achieving MAE as low as 0.002 for single missing values. This cluster likely represents buildings with highly regular, predictable consumption patterns where shared profile models excel. Cluster C_1 shows intermediate performance, where Transformer (D) consistently performs best. Cluster C_2 exhibits the most challenging imputation task, though still showing considerable improvement over per-building models, with Transformer (D) achieving MAE of 0.043–0.100. The higher variability in C_2 may indicate more diverse or irregular consumption patterns within the group.

The ranking remains consistent across clusters, with Transformer (D) demonstrating superior performance in nearly all scenarios, followed by KNN (A) and MissForest (B). MRNN (G) continues to underperform, though the gap narrows substantially in C_0. Missing data scenarios show patterns similar to per-building models, with sequential gaps (S2, S3) proving more challenging than random gaps, but absolute MAE values are substantially lower across all scenarios.

Algorithm	R1	R2	R3	S2	S3
Cluster C_0					
A (KNN)	.004	.005	.012	.020	.011
B (MissForest)	.002	.013	.031	.050	.052
C (SAITS)	.019	.016	.025	.033	.023
D (Transformer)	.008	.008	.015	.023	.016
E (TimesNet)	.061	.055	.065	.075	.068
F (USGAN)	.009	.015	.031	.073	.126
G (MRNN)	.091	.081	.094	.105	.087
Cluster C_1					
A (KNN)	.175	.155	.161	.179	.190
B (MissForest)	.176	.163	.177	.192	.188
C (SAITS)	.221	.202	.208	.205	.221
D (Transformer)	.165	.150	.151	.152	.163
E (TimesNet)	.284	.273	.271	.279	.283
F (USGAN)	.255	.254	.257	.264	.280
G (MRNN)	.317	.307	.305	.312	.319
Cluster C_2					
A (KNN)	.093	.163	.228	.199	.196
B (MissForest)	.0885	.0677	.083	.120	.156
C (SAITS)	.120	.101	.102	.115	.139
D (Transformer)	.0427	.0859	.063	.089	.100
E (TimesNet)	.304	.249	.263	.275	.481
F (USGAN)	.055	.074	.081	.112	.123
G (MRNN)	.534	.462	.459	.468	.524

TABLE III
MAE FOR ALL ALGORITHMS IN ALL 5 SCENARIOS ACROSS 3 DT CLUSTER MODELS.

D. Comparative Analysis

Comparing Tables II and III reveals substantial advantages of the DT approach. Profile-based models achieve dramatic MAE reductions: Cluster C_0 shows 95–99% reduction compared to per-building models (e.g., 0.002 vs. 0.308 for Transformer on R1), representing near-perfect imputation for buildings with regular patterns. Cluster C_1 achieves 40–50% reduction (e.g., 0.165 vs. 0.308), while Cluster C_2 shows 50–86% reduction depending on scenario (e.g., 0.043 vs. 0.308). These improvements demonstrate that leveraging consumption profile similarities through the DT paradigm substantially enhances accuracy through cross-building knowledge transfer and increased training data volume.

The consistent pattern of superior profile-based performance across both Tethys and Alicante datasets validates the generalizability of the approach. While Alicante shows lower baseline MAE compared to Tethys, suggesting more predictable residential/commercial consumption, both datasets benefit substantially from profile-based clustering. The Transformer model (D) maintains its performance advantage in both training paradigms, though the relative gap narrows in profile-based models where simpler methods like KNN (A) and MissForest (B) also achieve excellent results, particularly in C_0. This indicates that with sufficient high-quality training data from similar consumption profiles, even simpler algorithms can achieve near-optimal performance. Conversely, MRNN (G) remains the worst-performing algorithm in both paradigms, indicating fundamental architectural limitations rather than data scarcity issues.

The DT approach offers significant operational advantages beyond improved accuracy: three profile-based models replace 22+ building-specific models, reducing training overhead; buildings with extensive missing data ($>$70%) benefit from shared knowledge where per-building models have insufficient training data; new buildings can be assigned to existing profiles without retraining, enabling plug-and-play deployment; and profile-based models are less susceptible to anomalous patterns in individual buildings. These results demonstrate that organizing building-level DTs into consumption-profile-based architecture substantially improves data imputation performance while reducing operational complexity, aligning with the vision of DTs where interconnected twins collaborate to achieve superior outcomes.

VII. Conclusions

This work demonstrates that organizing building-level DTs into consumption-profile-based ecosystems can outperform conventional building-specific approaches for data imputation in smart water grids. By clustering buildings based on behavioral characteristics rather than treating them in isolation, cross-building knowledge transfer improves imputation accuracy, while reducing operational complexity. Our evaluation across 7 imputation algorithms and five missing data scenarios shows that profile-based models achieve MAE reductions of 40–99% compared to per-building approaches, with the most regular consumption profiles reaching near-perfect accuracy (MAE as low as 0.002). The Transformer architecture consistently delivered the best performance, though simpler algorithms like KNN and MissForest also achieved excellent results with sufficient profile-based training data. Validation across institutional and residential/commercial buildings confirms the approach's generalizability across different geographical contexts and building typologies.

Beyond improved accuracy, our approach offers significant operational advantages: 3 models replace 22+ building-specific models in our deployment, new buildings integrate seamlessly through profile assignment without retraining, and buildings with extensive missing data benefit from shared knowledge where isolated models fail. These enhancements enable more effective decision-making in water conservation, leak detection, and infrastructure maintenance, ultimately improving operational efficiency and sustainability. Our future work will focus on real-time deployment in operational smart water metering environments, dynamic cluster assignment mechanisms that adapt to evolving consumption patterns, and extending the Digital Twin Ecosystem paradigm to other smart city domains.

References

[1] J. E. Pesantez, F. Alghamdi, S. Sabu, G. Mahinthakumar, and E. Z. Berglund, "Using a digital twin to explore water infrastructure impacts during the covid-19 pandemic," *Sustainable Cities and Society*, vol. 77, p. 103520, 2022.

[2] O. Giustolisi, "Digital transition, digital twin and digital water: history, concepts and overview for the application to aqueducts," *Digital Water*, vol. 1, no. 1, p. 2313975, 2023.

[3] P. Ghorbani Bam, N. Rezaei, A. Roubanis, D. Austin, E. Austin, B. Tarroja, I. Takacs, K. Villez, and D. Rosso, "Digital twin applications in the water sector: A review," *Water*, vol. 17, no. 20, 2025. [Online]. Available: https://www.mdpi.com/2073-4441/17/20/2957

[4] H. M. Ramos, M. C. Morani, A. Carravetta, O. Fecarrotta, K. Adeyeye, P. A. López-Jiménez, and M. Pérez-Sánchez, "New challenges towards smart systems' efficiency by digital twin in water distribution networks," *Water*, vol. 14, no. 8, 2022. [Online]. Available: https://www.mdpi.com/2073-4441/14/8/1304

[5] Z. Y. Wu, A. Chew, X. Meng, J. Cai, J. Pok, R. Kalfarisi, K. C. Lai, S. F. Hew, and J. J. Wong, "High fidelity digital twin-based anomaly detection and localization for smart water grid operation management," *Sustainable Cities and Society*, vol. 91, p. 104446, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670723000574

[6] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, "A survey on data imputation techniques: Water distribution system as a use case," *IEEE Access*, vol. 6, pp. 63 279–63 291, 2018.

[7] R. Wu, S. D. Hamshaw, L. Yang, D. W. Kincaid, R. Etheridge, and A. Ghasemkhani, "Data imputation for multivariate time series sensor data with large gaps of missing data," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10 671–10 683, 2022.

[8] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," *Future Generation Computer Systems*, vol. 128, pp. 63–72, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21003794

[9] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-nearest neighbor (k-nn) based missing data imputation," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, 2019, pp. 83–88.

[10] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[11] W. Du, D. Côté, and Y. Liu, "Saits: Self-attention-based imputation for time series," *Expert Systems with Applications*, vol. 219, p. 119619, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423001203

[12] Z. Chen, H. Xu, P. Jiang, S. Yu, G. Lin, I. Bychkov, A. Hmelnov, G. Ruzhnikov, N. Zhu, and Z. Liu, "A transfer learning-based lstm strategy for imputing large-scale consecutive missing data and its application in a water quality prediction system," *Journal of Hydrology*, vol. 602, p. 126573, 2021.

[13] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *The eleventh international conference on learning representations*, 2022.

[14] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[15] D. Amaxilatis, I. Chatzigiannakis, C. Tselios, N. Tsironis, N. Niakas, and S. Papadogeorgos, "A smart water metering deployment based on the fog computing paradigm," *Applied Sciences*, vol. 10, no. 6, 2020.

[16] Hellenic Academic Libraries Link, "Alicante water consumption dataset," https://hardmin.heal-link.gr/en/dataset/700cde6f-ef69-41f0-8fec-9c3f74ae38d4, 2024, accessed: December 2025.