

A Data-Driven Temporal Framework for Water Consumption Monitoring with Spatial Visualization Using K-Means and STL-LSTM

Salsabila Septi Sukmayanti

Informatics Engineering Study Program, Telkom University, Purwokerto, Indonesia
E-mail: salsabilaseptisukmay@student.telkomuniversity.ac.id
ORCID iD: <https://orcid.org/0009-0008-9645-1311>

Sudianto Sudianto*

Informatics Engineering Study Program, Telkom University, Purwokerto, Indonesia
E-mail: sudianto@telkomuniversity.ac.id
ORCID iD: <https://orcid.org/0000-0001-8610-0587>

*Corresponding Author

Aminatus Sa'adah

Informatics Engineering Study Program, Telkom University, Purwokerto, Indonesia
E-mail: aminatuss@telkomuniversity.ac.id
ORCID iD: <https://orcid.org/0000-0002-3469-7745>

Received: 16 November 2025; Revised: 15 December 2025; Accepted: 02 February 2026; Published: 08 April 2026

Abstract: The water distribution sector in Indonesia still faces challenges in detecting leaks early due to manual data checks that are time-consuming and labor-intensive. PDAM (Regional Water Company) Tirta Wijaya Cilacap, Indonesia, faces similar problems. This study aims to implement a spatial customer prediction model to detect customer water usage and support data-driven operational decision-making. K-Means clustering groups customers by consumption patterns and geographic location, achieving a Silhouette Score of 0.4473 and a Davies–Bouldin Index of 0.7658, which indicates reasonably well-separated clusters in real-world data. In addition, water consumption forecasting was carried out with Seasonal–Trend Decomposition using Loess–Long Short-Term Memory (STL–LSTM) to predict trends and seasonality of water usage for each Customer Connection ID (CCID). The forecasting performance varies across CCIDs; the best case achieves an R^2 of up to 0.95, while low-performing cases are discussed to clarify conditions where STL–LSTM is less reliable. The forecasting and clustering outputs are presented through a spatial visualization (map) of water-consumption categories and model results to support identifying areas that may require closer inspection for potential leakage and waste. This research contributes to strengthening technology-based public infrastructure, in line with SDG 9: Industry, Innovation, and Infrastructure, to promote sustainable water management.

Index Terms: Consumption Pattern; K-Means; STL-LSTM; Spatial Prediction; Water Consumption Forecasting

1. Introduction

Data-driven management and operational decision-making in the water distribution sector are essential to supporting sustainable development, in line with SDG 9: Industry, Innovation, and Infrastructure [1]. One of the biggest challenges in the water distribution sector is the high level of Non-Revenue Water (NRW), which is the volume of water lost before it reaches customers due to leaks, theft, or measurement errors [2]. In Indonesia, high NRW levels—estimated at around 32% on average—directly erode utility revenues and constrain the reliable supply of clean water [3]. Although digitization has transformed various industrial sectors, Indonesia's water distribution system still lags in adopting innovative technology. This condition is evident at PDAM (Regional Water Company) Tirta Wijaya Cilacap, the leading provider of clean water in Cilacap Regency, Indonesia. In Cilacap Regency, PDAM Tirta Wijaya continues to expand its service coverage—reportedly setting a 2024 target of 5,600 new customer connections—and maintains a bulk-water supply partnership with PT Kilang Pertamina Internasional (RU IV), which allegedly contributes up to Rp 22

billion per year [4]. However, operational activities, such as water meter recording, are still carried out manually and are labor-intensive. Due to field constraints, if the customer's house has a closed gate, officers cannot record water usage for that month or for months. As a result, the consumption data produced is static, has high latency, and cannot be used for monitoring [5]. This limitation hinders early detection of anomalies, such as leaks or abnormal usage, so corrective action is taken only after customers receive inflated bills or after distribution disruptions occur.

To address the limitations of manual checking and delayed consumption reporting, water utilities increasingly adopt machine learning to improve metering data quality and support loss monitoring in water distribution systems. Utilities and researchers often use K-means clustering to group customers based on similarities in water consumption patterns [6,7]. However, the K-Means method alone is still static and cannot adaptively capture temporal dynamics [8]. On the other hand, researchers widely apply Long Short-Term Memory (LSTM) networks to time-series data because they can learn long-term temporal dependencies. However, standard LSTM models often struggle to disentangle complex trend and seasonal components without additional structure [9]. Another study reported good performance after applying LSTM to model sequential chatbot data, supporting its use to learn temporal patterns in consumption time series [10]. The latest development introduces the integration of Seasonal-Trend Decomposition using Loess (STL) with LSTM to improve the model's ability to handle seasonal variations and long-term trends [12]. However, previous STL-LSTM studies still focus on temporal patterns without considering spatial variations or customer behavior heterogeneity. On the other hand, clustering-based research highlights the importance of customer grouping for understanding consumption characteristics but has not yet integrated it with time-series forecasting models [6,12].

Based on these limitations, this study proposes an integrated spatio-temporal framework that combines K-Means clustering and Seasonal-Trend Decomposition using Loess with Long Short-Term Memory (STL-LSTM) for monitoring PDAM customer water consumption. Unlike previous studies that treat clustering and forecasting as separate analytical tasks, this framework unifies customer segmentation, time-series forecasting, and spatial visualization into a single operational pipeline. The novelty of this work lies not in proposing a new algorithm, but in demonstrating how the integration of consumption-based clustering and CCID-level STL-LSTM forecasting can generate actionable spatial insights for early anomaly indication and operational decision-making in real-world water utility systems [13,17]. This approach enables utilities to move from reactive billing-based monitoring toward proactive, data-driven consumption management.

Despite extensive studies on water consumption clustering and time-series forecasting, most existing works still treat customer segmentation and demand prediction as separate analytical tasks. Clustering-based approaches effectively capture consumption heterogeneity but lack temporal modeling capabilities, while STL-LSTM and other deep learning models primarily focus on temporal patterns without incorporating spatial customer characteristics. Consequently, there is a lack of an integrated spatio-temporal framework that unifies customer-level clustering, CCID-based forecasting, and spatial visualization into a single operational pipeline for real-world water consumption monitoring.

2. Materials and Method

2.1. Dataset and Study Area

This study uses water usage data from customers of PDAM Tirta Wijaya, Cilacap Regency, Indonesia. The study focuses on customers whose connection points are equipped with geographic coordinates, as shown in Fig.1, which supports spatial mapping. In addition to the geographic coordinates of connection points (*latitude/longitude*), the dataset contains monthly water meter readings that include CCID (*Customer Connection ID*), *meter recording period*, and *water consumption volume* (m^3). Forty-two customers and 5,393 monthly observation records (raw data) from April 2015 to April 2025 make up the historical data used. The study focuses on the georeferenced sub-area (connections with verified latitude/longitude coordinates); the sample size and its relation to the utility-wide coverage are clarified as follows:

- The 42 customers constitute the full set of available connections in the study sub-area (Griya Prima Rinjani II) with verified coordinates and monthly meter readings for the analysis period.
- The “5,600 connections” figure refers to PDAM-wide coverage; thus, the dataset represents the coordinate-verified sub-area ($42/5,600 \approx 0.75\%$) and is not a random sample.
- Findings should be interpreted as a pilot demonstration; broader generalization requires expanded georeferenced coverage across additional service areas.

Despite the limited sample, robustness is evaluated using multiple clustering validity indices (SSE/WCSS, silhouette, and Davies–Bouldin), which provide consistent evidence for the selected clustering structure.

2.2. Research Framework

This study began with the development of a framework for two models, K-Means Clustering and STL-LSTM, to predict and detect anomalies in PDAM customer water consumption. K-Means clustering works by grouping spatial data based on similarities in consumption patterns and identifying Customer Connection IDs (CCIDs), which can then be used for forecasting to produce predictions of future water usage through STL-LSTM modeling based on seasonal

decomposition and trends. The research process flow is shown in Fig. 2.

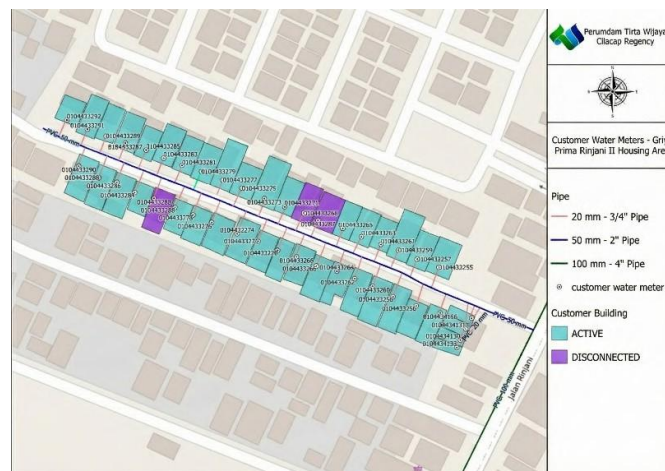


Fig. 1. Spatial layout of customer water meters in housing area.

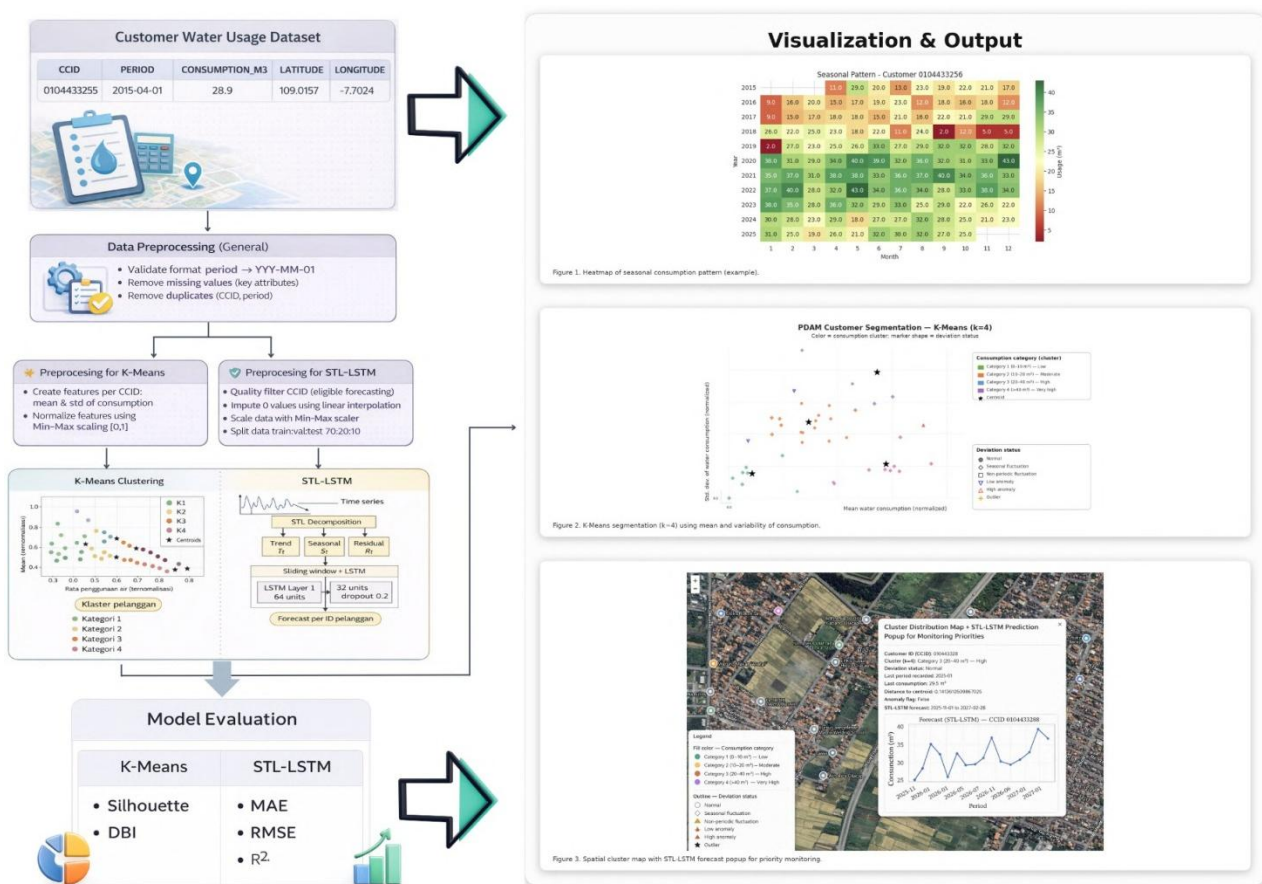


Fig. 2. Research methodology workflow.

2.3. Data Preprocessing

The initial dataset before preprocessing consisted of 5,393 monthly observation records. During the preprocessing stage, missing values in key attributes (*CCID*, *period*, *consumption*, *latitude*, and *longitude*) were addressed through data cleaning, and duplicate records were identified and eliminated based on the (*CCID*, *period*) pair so that each customer had only one observation per month. After this cleaning process, the number of data points decreased to 4,899. In addition, the period attribute was standardized to a date format (YYYY-MM-01), and the attribute structure was adjusted as shown in Fig.3. The figure presents an excerpt of the cleaned dataset before further preprocessing for K-Means clustering and STL-LSTM modelling.

	CCID	period	consumption_m3	latitude	longitude
0	0104433255	2015-04-01	0	109.015754230564596	-7.702386699931322
1	0104433255	2015-05-01	0	109.015754230564596	-7.702386699931322
2	0104433255	2015-06-01	0	109.015754230564596	-7.702386699931322
3	0104433255	2015-07-01	0	109.015754230564596	-7.702386699931322
4	0104433255	2015-08-01	0	109.015754230564596	-7.702386699931322
...	...	...	...	...	...
4894	0104434133	2024-12-01	8	109.015826101800400	-7.702701764027637
4895	0104434133	2025-01-01	6	109.015826101800400	-7.702701764027637
4896	0104434133	2025-02-01	7	109.015826101800400	-7.702701764027637
4897	0104434133	2025-03-01	6	109.015826101800400	-7.702701764027637
4898	0104434133	2025-04-01	8	109.015826101800400	-7.702701764027637

4899 rows x 5 columns

Fig. 3. Excerpt of the cleaned dataset after preprocessing.

Latitude/longitude coordinates are used only for mapping and spatial visualization, so they are not used in K-Means or STL-LSTM modeling. For K-Means, the features are normalized using Min-Max scaling to ensure a uniform variable scaling. Next, each customer's time series data passes through a quality filter to ensure the series is sufficiently long and consistent, thereby stabilizing the STL decomposition and LSTM training.

2.4. K-Means Clustering

In the K-Means Clustering modeling stage, customers are grouped into (k) different clusters by minimizing the variance within each cluster. The clustering process flow for customer water usage using K-Means in this study is shown in Fig. 4.

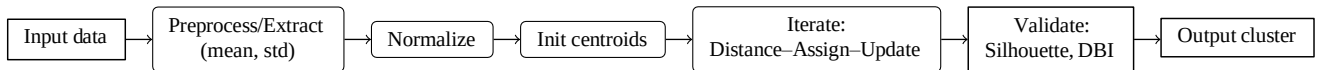


Fig. 4. Compact K-Means clustering workflow for water consumption data.

This process is carried out by calculating the distance between each data point and the cluster center, then updating the cluster center position based on the average of the data points in that cluster [7], which is formulated as follows [14]:

$$J = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

Where C_i represents the set of customer data in cluster i , x is the water consumption feature vector, and μ_i is the cluster center. The consumption statistics used in the K-Means feature include the mean, standard deviation, maximum, and minimum. All features are normalized using Min-Max scaling to balance the contribution of each feature in the Euclidean distance calculation, given differences in range and variability across consumption features. The cluster centroid is computed as the mean of all data points in the cluster, as follows [15]:

$$C_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

Where $|C_j|$ denotes the number of customers in cluster j .

2.5. STL-LSTM (Seasonal-Trend Decomposition using Loess – Long Short-Term Memory)

STL-LSTM (Seasonal and Trend decomposition using Loess – Long Short-Term Memory) modeling is used to separate trend, seasonal, and residual components. Meanwhile, Long Short-Term Memory (LSTM) is used to study long-term nonlinear temporal dependencies [16,17]. Separating trend and seasonal components before LSTM training can improve the stability and accuracy of forecasting compared to classical time-series models and pure LSTM, in the context of water demand [9,12,13].

The initial STL-LSTM modeling will select CCIDs that are feasible to predict through a quality filter to ensure that the time series has sufficient length and consistency. At this stage, values of 0, representing invalid records/no readings, are handled via linear interpolation. The data is then normalized using Min-Max scaling to be within the range [0, 1] before LSTM training. Next, the dataset is divided into training, validation, and test sets in the ratio 70%:20%:10% in a chronological order. In the next stage, each CCID time series is decomposed using STL to obtain trend, seasonal, and residual components. These three components are then arranged in a sliding window format as input to the LSTM model to generate water usage predictions for the next period. The general flow of the STL-LSTM process is shown in Fig. 5.

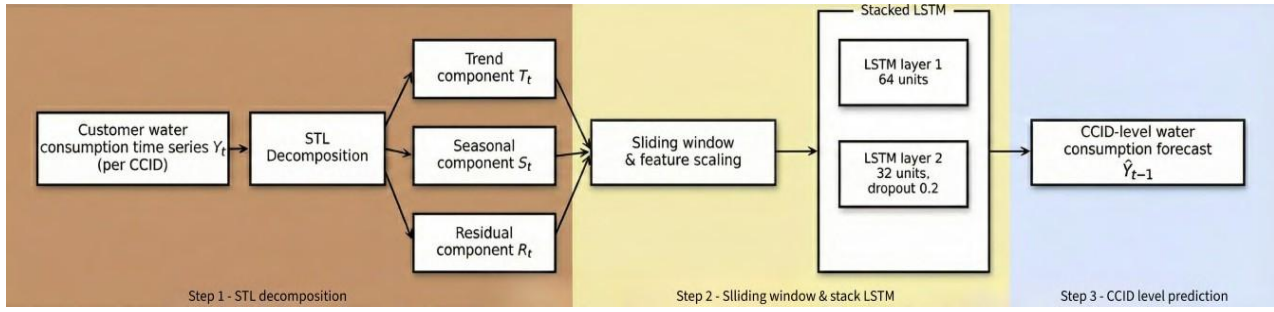


Fig. 5. Proposed STL-LSTM architecture for CCID-level water consumption forecasting.

The stacked LSTM model uses a series of components resulting from STL decomposition as input. To reduce overfitting, the network uses 64 units in the first LSTM layer and 32 units in the second, with a dropout rate of 0.2. The output of the second layer is then mapped through a fully connected layer to generate water usage predictions for the next period for each CCID. The optimizer settings, data partitioning, and evaluation metrics (RMSE, MAE, and R^2) are set in the training configuration.

2.5.1 STL Decomposition

With three STL additive components that separate the consumption time series Y_t belonging to each CCID, as shown in this formulation:

$$Y_t = T_t + S_t + R_t \quad (3)$$

Where T_t expressing long-term trend components, S_t periodically recurring seasonal components, and R_t residual components containing random fluctuations and local deviations. Component estimation is performed using Locally Estimated Scatterplot Smoothing (LOESS), so that trends and seasonal patterns can gradually change in line with data dynamics [18]. In the context of PDAM water consumption, STL decomposition yields three components: the trend component (T_t) which reveals long-term changes due to customer growth or rate changes, seasonal components (S_t) which captures recurring patterns such as increased consumption during the dry season, and residual components (R_t) which isolates unexpected deviations as potential anomaly indicators. This decomposition simplifies LSTM learning by separating complex patterns into components that are easier to model, while facilitating the interpretability of prediction result by attributing each component [12,13].

2.5.2 Long-Short Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are specialized recurrent neural networks that overcome the vanishing gradient problem and can capture long-term temporal dependencies in sequential data [19,20]. LSTM uses cell state C_t and three gate mechanisms, namely forget gate, input gate, and output gate, which regulate the flow of information in memory. Given the time series input x_t and the previous hidden state h_{t-1} , the gate value at time step t is defined as follows:

Input Gate (i_t) controls new information stored in the cell state:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (4)$$

Forget Gate (f_t) determines whether information from the previous cell state is retained or forgotten:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (5)$$

Output Gate (o_t) and Hidden State (h_t):

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (6)$$

Memory cell state updates combine old memories with new information candidates:

$$g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (8)$$

The hidden state calculation produces an output representation at time step t :

$$h_t = o_t \odot \tanh(C_t) \quad (9)$$

Where σ is the sigmoid function that maps values into the range $[0, 1]$, $\tanh$ is the hyperbolic tangent function, and $\odot$ is element-wise multiplication that computes the product of corresponding entries. During training, the model learns W and b by updating them to minimize the loss [19,21].

2.6. Evaluation Metrics

2.6.1 Clustering Evaluation (K-Means)

Optimization of cluster parameters (k) using the Elbow Method as an initial heuristic by analyzing the elbow point on the Within-Cluster Sum of Squares (WCSS) curve to help determine the best cluster (k) candidate [22]. The results were then re-evaluated using the Silhouette Score to measure cluster cohesion and separation, as well as the Davies-Bouldin Index (DBI), which compares intra-cluster dispersion with the distance between cluster centers. A higher Silhouette value and a lower DBI value each indicate better cluster quality. The Silhouette value is formulated as follows [23]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (10)$$

Where $a(i)$ is the average distance of point i to all points in the same cluster. Meanwhile, $b(i)$ is the average distance between points i to the closest other cluster. The Silhouette score $s(i)$ is in the range $[-1, 1]$. Values above 0.5 generally indicate well-structured clusters. Values between 0.25 and 0.50 represent fairly well-formed clusters. Cluster validity is further assessed using the Davies-Bouldin Index (DBI), which is formulated as follows [23]:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)} \right) \quad (11)$$

Where σ_i expressing the intra-cluster dispersion of i and $d(\mu_i, \mu_j)$ is the distance between the center of cluster i and cluster j .

2.6.2 Forecasting Evaluation (STL-LSTM)

The performance of the STL-LSTM model was evaluated using three commonly used time series forecasting metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination R^2 [24]. In the implementation, these three metrics were calculated separately for each customer (CCID) with different data conditions. The three metrics are defined as follows:

1. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (12)$$

2. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2} \quad (13)$$

3. Coefficient of Determination R^2

$$R^2 = 1 - \frac{n(RMSE)^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2} \quad (14)$$

3. Result and Discussion

3.1. Result of Clustering (K-Means)

The results of K-Means Clustering have grouped water consumption patterns from 42 CCIDs. The clustering features include average monthly consumption, standard deviation, maximum usage, minimum usage, and total consumption during the observation period. Before modeling, the analysis normalized all features to the range $[0, 1]$ using a Min-Max scaler so that each dimension had a comparable scale. The workflow excluded customer geographic coordinates (latitude, longitude) from the clustering feature set and then combined them with cluster labels to provide spatial visualization on the water distribution map.

The determination of the cluster value (k) was explored over the range 2–9. For each value of k , three metrics were calculated using Inertia/SSE (WCSS), forming an elbow curve that showed an elbow point at $k = 4$ and entered the diminishing returns zone after $k > 4$ in Fig.6 [25]. Furthermore, a silhouette score close to 1 indicates that the object is better at separating clusters, whereas a smaller DBI suggests that the clusters are more distinct [26,27]. Table 1 presents the SSE, silhouette, and DBI values for each k .

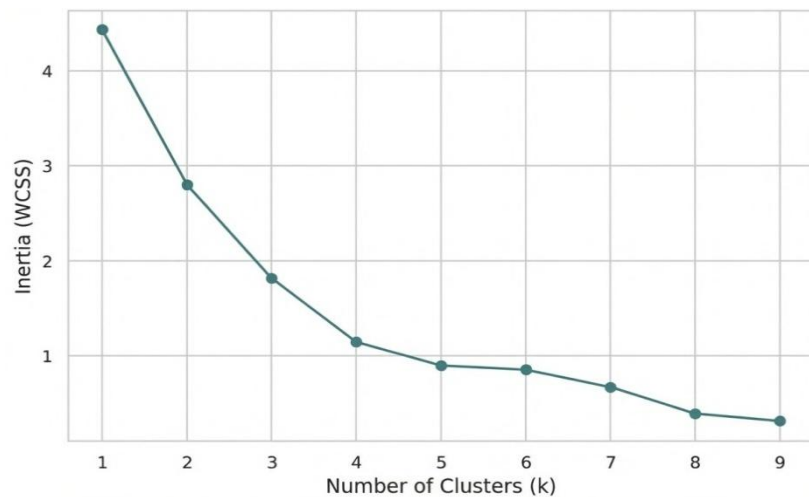
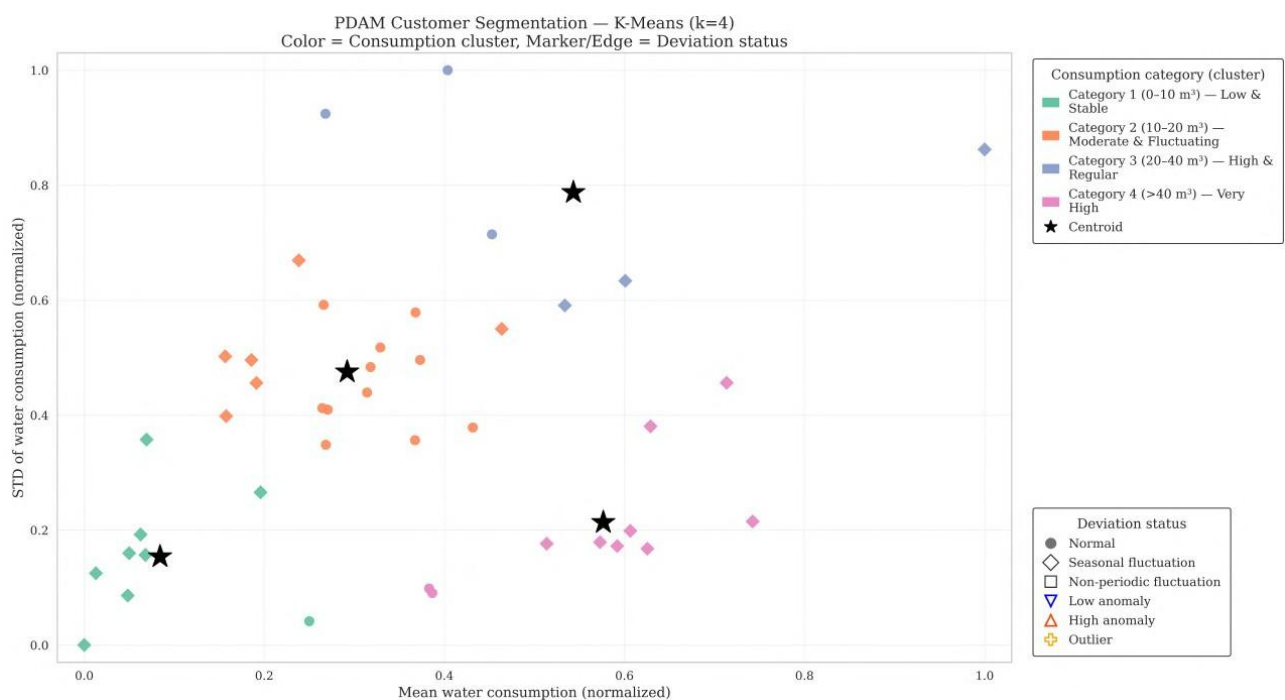


Fig. 6. Elbow method curve.

Table 1. K-Means evaluation results.

k	Inertia (SSE)	Silhouette	Davies–Bouldin
2	2.7786	0.3121	1.1248
3	1.7793	0.3757	0.8371
4	1.1162	0.4473	0.7658
5	0.8615	0.4291	0.6561
6	0.6289	0.4628	0.5842
7	0.5057	0.4326	0.5914
8	0.3912	0.4185	0.6375
9	0.3102	0.4402	0.6007

Based on the evaluation results, this study selects $k = 4$ as the most representative number of clusters to capture variation in water-usage patterns among the 42 customers. Although $k = 6$ was also quite good, determining the value of k that was too large resulted in several clusters with very few customers, making the separation too fragmented and less practical for information. Fig.7 shows the distribution of 42 customers in the average feature space, along with the standard deviation of monthly water consumption. The color of the dots indicates the membership of the consumption cluster (category 1–4), while the black star symbols mark the position of each cluster's centroid.

Fig. 7. K-Means clustering of customers ($k = 4$) with centroids.

Although Figure 7 presents the clustering in the normalized feature space, Table 2 provides CCID-level traceability by listing each customer's assigned cluster and deviation status, along with the mean and standard deviation on the original scale (m^3).

Table 2. Cluster, status, and consumption statistics.

CCID	Cluster	Deviation status	Mean (m^3)	Std (m^3)
0104433285	C1	Normal	10	3
0104433255	C1	Seasonal fluctuation	1	2
0104433257	C1	Seasonal fluctuation	8	7
0104433268	C1	Seasonal fluctuation	3	6
0104433272	C1	Seasonal fluctuation	3	5
0104433276	C1	Seasonal fluctuation	1	4
0104433277	C1	Seasonal fluctuation	3	9
0104434126	C1	Seasonal fluctuation	3	3
0104434132	C1	Seasonal fluctuation	3	5
0104433258	C2	Normal	11	9
0104433259	C2	Normal	18	9
0104433261	C2	Normal	13	11
0104433263	C2	Normal	11	10
0104433266	C2	Normal	14	12
0104433278	C2	Normal	11	13
0104433280	C2	Normal	11	10
0104433282	C2	Normal	13	10
0104433284	C2	Normal	15	13
0104433287	C2	Normal	15	12
0104433288	C2	Normal	15	9
0104433260	C2	Seasonal fluctuation	8	12
0104433264	C2	Seasonal fluctuation	7	12
0104433270	C2	Seasonal fluctuation	19	13
0104433292	C2	Seasonal fluctuation	10	15
0104434130	C2	Seasonal fluctuation	8	11
0104434133	C2	Seasonal fluctuation	7	10
0104433262	C3	Normal	18	16
0104433273	C3	Normal	11	20
0104434131	C3	Normal	16	22
0104433275	C3	Seasonal fluctuation	22	13
0104433286	C3	Seasonal fluctuation	40	19
0104433291	C3	Seasonal fluctuation	24	14
0104433281	C4	Normal	16	4
0104433289	C4	Normal	16	4
0104433256	C4	Seasonal fluctuation	25	9
0104433265	C4	Seasonal fluctuation	29	11
0104433269	C4	Seasonal fluctuation	24	6
0104433271	C4	Seasonal fluctuation	24	5
0104433274	C4	Seasonal fluctuation	25	5
0104433279	C4	Seasonal fluctuation	23	5
0104433283	C4	Seasonal fluctuation	30	6
0104433290	C4	Seasonal fluctuation	21	5

3.2. Quality Filter for STL-LSTM

A CCID was retained for STL-LSTM only if all thresholds below were satisfied:

- **Minimum length:** $L_{\min} = 60$ active monthly observations ($L \geq 60$).
- **Missingness:** missing-month proportion $m \leq M_{\max} = 20\%$ over the trimmed active period.
- **Zero dominance:** zero-consumption proportion $z \leq Z_{\max} = 30\%$, computed over non-missing months.
- **Longest zero run:** longest consecutive zero run $r \leq R_{\max} = 12$ months (missing months are not counted as zeros).

The quality filter was rule-based (threshold screening) and applied deterministically. Table 3 reports the sequential exclusion counts; 42 CCIDs were screened, and 21 CCIDs remained eligible for STL–LSTM forecasting.

Table 3. Quality-filter screening summary for STL–LSTM forecasting.

Criterion (applied sequentially)	Count
Initial CCIDs	42
Insufficient length ($L < 60$)	4
Too many missing months ($m > 20\%$)	1
Zero dominance ($z > 30\%$)	9
Longest zero run ($r > 12$ months)	7
Eligible CCIDs after quality filter	21

3.3. Result of Prediction (STL-LSTM)

For each CCID, this study decomposes the monthly consumption series Y_t using the STL procedure with a 12-month seasonal period. Fig.8 illustrates the decomposition results for one representative customer.

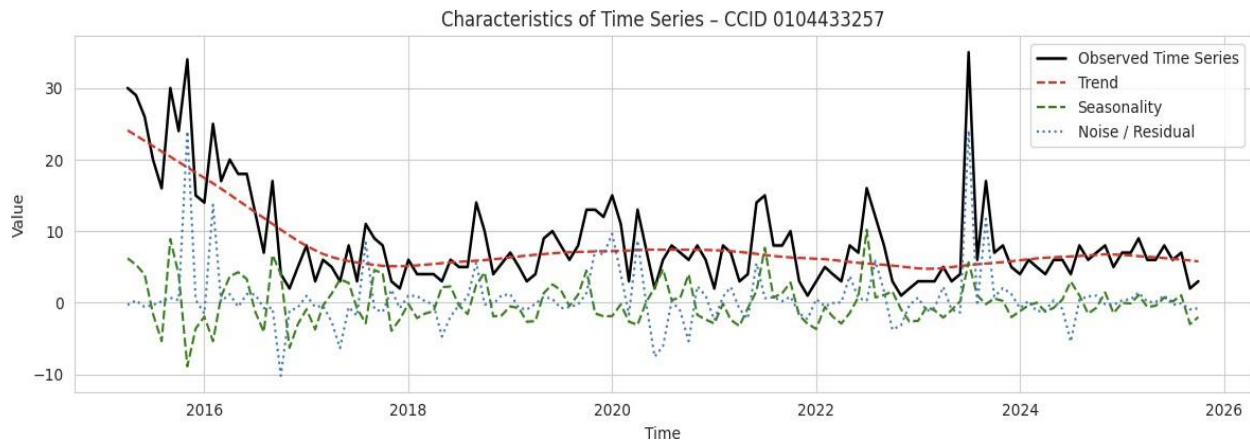


Fig. 8. Decomposition of monthly water consumption.

The graph shows an increasing trend component, a stable annual seasonal component, and a residual that fluctuates around zero. The decomposition successfully isolates short-term variations that cannot be explained by trends and seasonality, treating them as noise, and then passes them as input to the LSTM stage. Based on the STL decomposition results, the annual seasonal component St and the residual component Rt are recombined in the form of a residual series that is free from long-term and seasonal patterns. This residual series is then normalized with Min–Max scaling and arranged into target input data pairs using a 12-month sliding window scheme. Each window contains 12 past residual values as input and one residual value for the next month as the target. This process is performed separately for each CCID, yielding a set of time-series samples ready for LSTM model training.

Based on the scenario comparison in Table 4, Scenario 3 was selected for the STL–LSTM model. The architecture comprises two stacked LSTM layers with 64 and 32 units, followed by a dropout layer (rate = 0.2) and a fully connected output layer. Training uses a 12-month window, a batch size of 32, up to 200 epochs, and the AdamW optimizer with a learning rate of 0.001, resulting in the lowest validation loss and the most stable training on representative CCIDs.

Table 4. Training configurations of the STL–LSTM model.

Scenario	Epoch	Window Size	Batch Size	Optimizer	Learning Rate	Units	Dropout	Number of Samples
1	200	12	32	Adam	0.001	64, 32	0.2	1299
2	200	12	32	RMSprop	0.001	64, 32	0.2	1299
3	200	12	32	AdamW	0.001	64, 32	0.2	1299

As part of the hyperparameter tuning, this study evaluates several training scenarios with different optimizers and configuration settings for each CCID. Table 5 reports the best epoch, training loss, and validation loss for each scenario.

Although the differences in validation loss are relatively small, Scenario 3 yielded the lowest validation loss and the most stable training curve, so this study selected it as the final STL-LSTM configuration.

Table 5. Training result.

Scenario	Best Epoch	Train Loss	Val Loss
1	26	0.12974	0.08356
2	23	0.13504	0.08308
3	37	0.12804	0.08120

Fig.9 presents the training curve for Scenario 3 on the representative CCID. The model shows stable convergence, where both training and validation losses decrease smoothly and remain close throughout training. This behavior indicates no significant overfitting and supports the selection of Scenario 3 as the final STL-LSTM configuration.

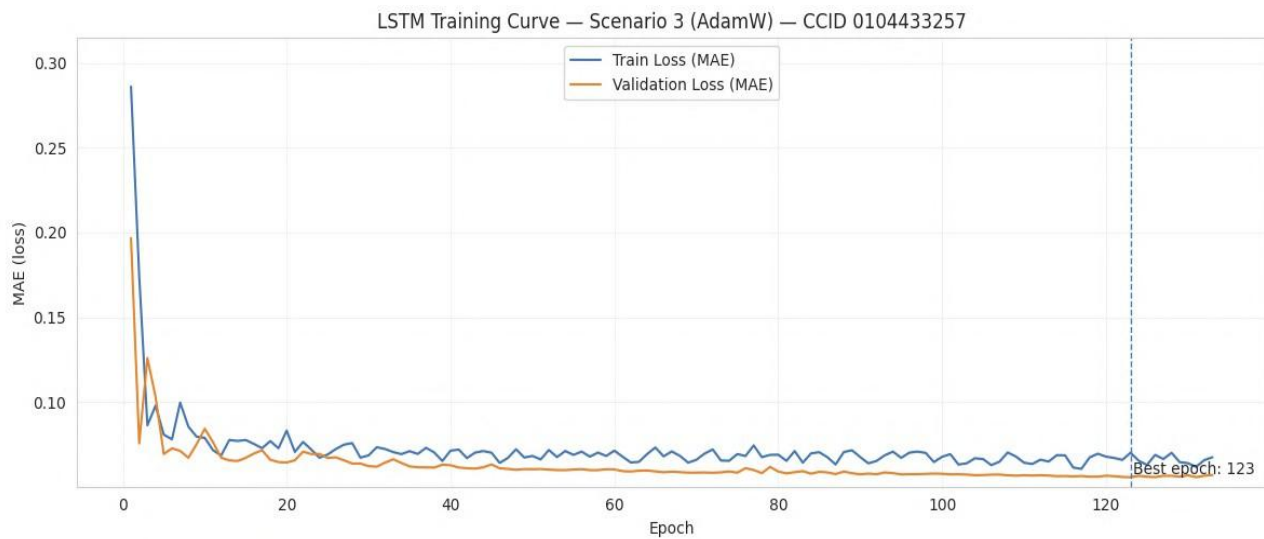


Fig. 9. Training and validation loss curves for Scenario 3.

Following the clustering stage described in Subsection 3.1, this study applied STL-LSTM forecasting only to CCIDs that satisfied the quality filter. Table 6 reports the test-set MAE, RMSE, and R^2 for each retained CCID, indicating that forecasting performance varies across customers. Although several CCIDs achieve high accuracy (up to $R^2 = 0.954629$), some CCIDs show low or negative predictive skill. In this study, low-performance cases are defined as $R^2 < 0.2$ and are examined further. To investigate these failures, Table 7 summarizes their K-Means cluster assignments (cf. Fig.7) and key consumption statistics (mean, standard deviation, minimum, maximum, and zero-rate), which helps distinguish whether the low performance is mainly driven by consumption-pattern characteristics, data quality, or model limitations.

Table 6 indicates that the lowest- R^2 cases are associated with distinct consumption profiles. CCID 0104433270 shows a relatively high zero-rate (21.26%), suggesting intermittent zero consumption or potential reading/occupancy issues that can degrade STL decomposition and residual learning. CCID 0104433262 also exhibits the largest variability (Std = 16.23) and an extreme maximum (190 m³), indicating sporadic spikes that are difficult to extrapolate and may dominate the error. In contrast, CCIDs 0104433279 and 0104433271 have near-zero zero-rates but still low R^2 , implying that poor performance is not explained by zeros alone and may reflect weak seasonality/irregular residual dynamics or clusterspecific patterns. Overall, these failures suggest that STL-LSTM is less reliable for customers with frequent zeros, extreme spikes, or highly irregular residual behavior, motivating additional data cleaning and/or alternative models for these profiles.

Table 6. Test-set performance of the STL-LSTM model.

CCID	MAE	RMSE	R^2
0104433256	0.783062	0.920287	0.954629
0104433257	1.186399	1.382498	0.497761
0104433259	1.019550	1.275342	0.921615
0104433261	1.340166	1.774605	0.870098
0104433262	4.799582	8.110213	-0.410316
0104433263	2.087359	2.518240	-0.630680
0104433265	1.468298	1.793799	0.940680
0104433270	2.927331	3.725546	-0.831967
0104433271	2.088144	2.293490	0.143152
0104433274	2.151842	2.387697	0.724140
0104433275	4.043959	5.034766	0.903195
0104433279	1.440849	1.750388	0.055256
0104433281	1.926179	2.208288	0.462723
0104433283	1.819048	2.149893	0.695250
0104433284	4.278550	9.768570	0.556278
0104433285	1.260754	1.382202	0.343415
0104433286	0.992688	1.188430	0.737912
0104433288	0.687971	0.949103	0.913523
0104433289	1.169857	2.193494	0.521517
0104433290	1.052440	1.367261	0.747710
0104433292	0.719889	0.770339	0.899586

Table 7. Failure cases: cluster membership and consumption statistics for low- R^2 CCIDs.

CCID	Cluster	Mean	Std	Min	Max	Zero%	R^2
0104433262	C3	18.834646	16.225966	8.0	190.0	0.000000	-0.410316
0104433263	C2	11.582677	9.799126	0.0	38.0	7.874016	-0.630680
0104433270	C2	17.637795	13.063744	0.0	50.0	21.259843	-0.831967
0104433279	C4	23.220472	5.185317	0.0	36.0	0.787402	0.055256
0104433271	C4	23.771654	4.971648	8.0	38.0	0.000000	0.143152

Baseline comparisons are reported in Table 8. Across 21 eligible CCIDs, STL-LSTM achieves the lowest forecasting errors (RMSE and MAE; mean/median) and the highest predictive accuracy (median R^2) among all evaluated methods. In contrast, Holt-Winters, Naive(t-1), SeasonalNaive(t-12), and TrainMean show substantially larger errors and predominantly negative R^2 , indicating limited predictive skill under the same evaluation setup. These results support the added value of combining STL decomposition with LSTM for modeling water-consumption time series.

Table 8. Comparison of STL-LSTM with baseline forecasting methods across 21 CCIDs.

Method	RMSE Mean	RMSE Med	MAE Mean	MAE Med	R^2 Mean	R^2 Med	N
HoltWinters	7.984	4.871	6.899	4.034	-2.002	-1.209	21
Naive(t-1)	5.260	4.103	4.012	2.833	-0.504	-0.507	21
STL-LSTM	2.625	1.820	1.885	1.479	0.458	0.677	21
SeasonalNaive(t-12)	10.347	5.454	7.873	4.917	-6.498	-1.674	21
TrainMean	9.669	5.605	8.653	4.828	-13.413	-1.658	21

To provide a visual illustration, Fig.10 shows the historical fit and 12-month projection results for one selected CCID, while Fig.11 summarizes the corresponding monthly forecast pattern. The numerical values of the forecasting results are listed in Table 9.

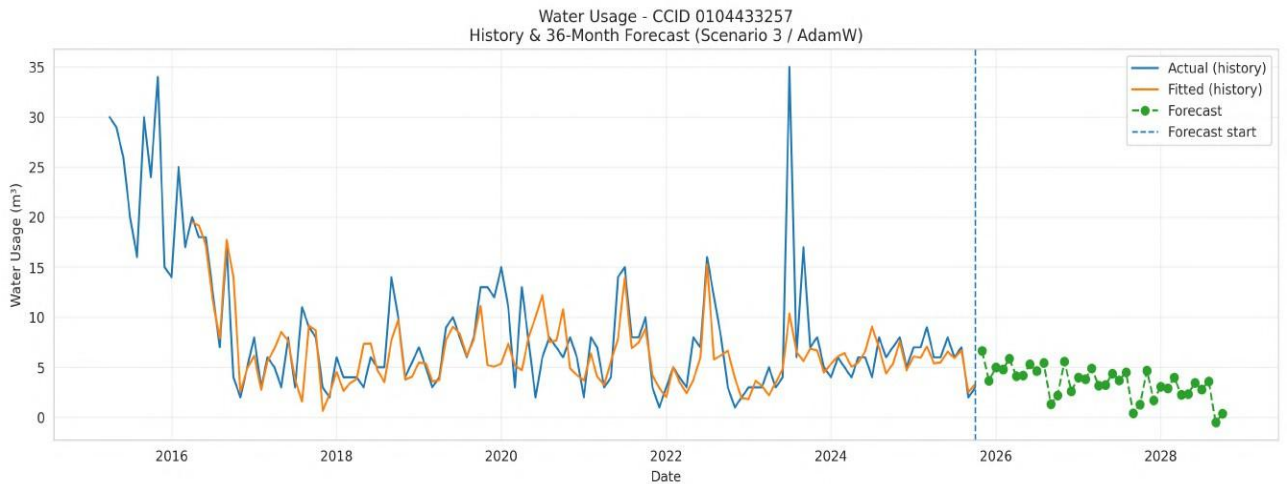


Fig. 10. CCID-level water consumption modeling using STL–LSTM, showing historical patterns and future forecasts to support consumption monitoring.

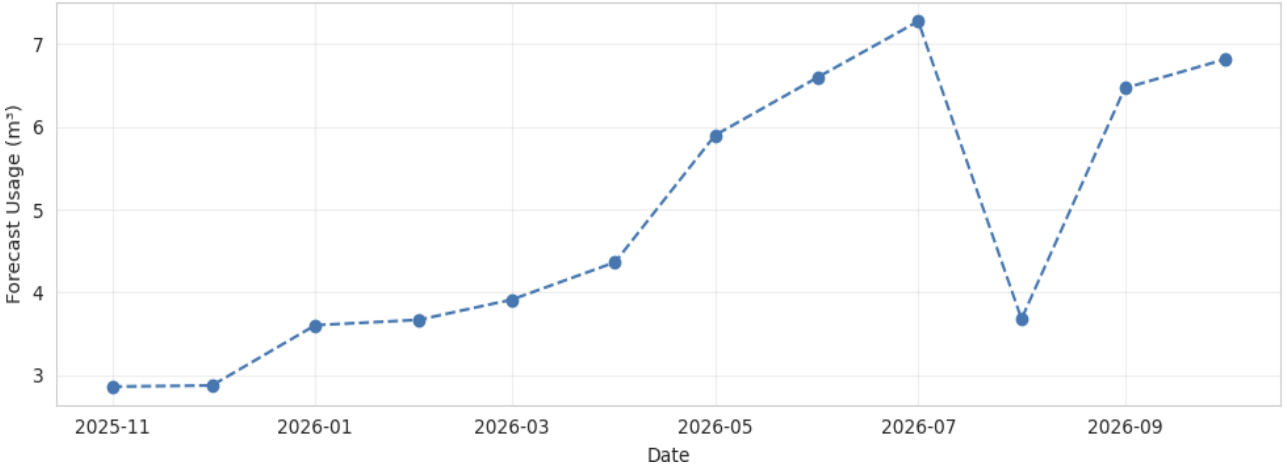


Fig. 11. Monthly forecast pattern produced by the STL–LSTM model for CCID 0104433257.

Table 9. Forecasted monthly water usage (STL–LSTM) for CCID 0104433257.

Period	Forecast usage (m^3)
2025-11	2.86
2025-12	2.88
2026-01	3.60
2026-02	3.67
2026-03	3.91
2026-04	4.36
2026-05	5.90
2026-06	6.60
2026-07	7.27
2026-08	3.68
2026-09	6.47
2026-10	6.81

Based on the results of K-Means clustering and STL–LSTM modeling, spatial predictions are presented in an interactive Folium map, as shown in Fig.12. Each point represents a customer (CCID) with colors indicating consumption categories and anomaly status. At the same time, a pop-up displays a summary of customer attributes and the STL–LSTM projection curve for the next 12 months. The integration of spatial information and time series projections provides a visual basis for supporting data-driven decision-making, for example, by determining inspection priorities for customers with unusual usage patterns or those showing sharp upward trends. Additionally, the projection

results can serve as an operational reference when meter readings cannot be taken due to field issues, such as limited customer access or locked gates, enabling monitoring and operational decisions to be made based on data.

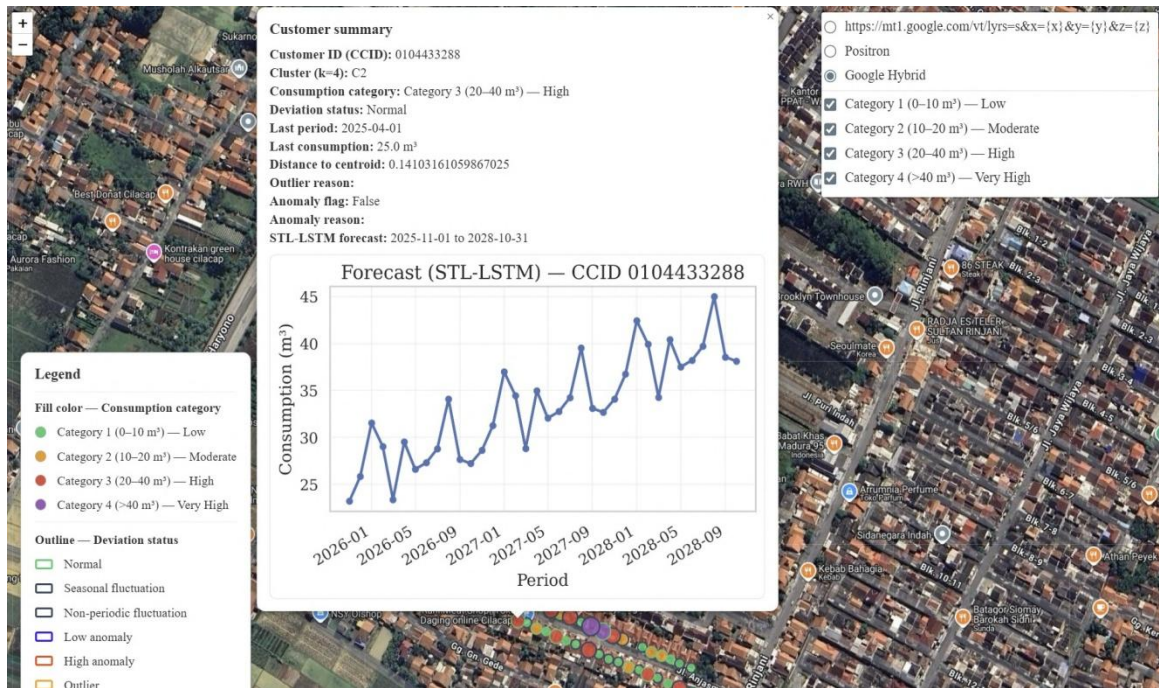


Fig. 12. Monthly forecast pattern produced by the STL–LSTM model for CCID 0104433257.

3.4. Practical Deployment Considerations

To support practical deployment of the proposed clustering–forecasting pipeline as an operational monitoring tool, this study outlines a workflow for utility screening and follow-up using the spatial dashboard (Folium map) and CCID-level forecasts.

Operational workflow. (1) *Daily/weekly screening*: operators open the dashboard to view customer points colored by consumption cluster and recent deviation status; a pop-up provides CCID attributes, historical time series, and the 12-month STL–LSTM forecast. (2) *Prioritization*: inspection candidates are ranked using forecast-based risk indicators (defined below) so that field visits focus on customers with the highest likelihood of abnormal behavior. (3) *Verification and action*: prioritized cases are checked against meter-reading logs and customer context; confirmed issues trigger follow-up actions (re-reading, meter replacement, leakage inspection, or customer outreach). (4) *Feedback loop*: outcomes from field verification are recorded to refine thresholds and improve subsequent screening.

Proposed anomaly-flagging thresholds. In this study, operational screening uses criteria aligned with the practices applied by PDAM Tirta Wijaya. A CCID is treated as *high-priority* if at least one of the following rules holds over the most recent 3 months (or the next 3-month forecast window): (i) *Zero-consumption persistence*: zero-rate $z \geq 30\%$ over the last 3 months (or ≥ 2 zero months in the last 3 months), indicating potential occupancy or meter-reading issues; (ii) *Abrupt change*: the absolute month-to-month change exceeds 2σ of the CCID’s historical month-to-month differences; (iii) *Forecast deviation*: the forecasted mean for the next 3 months lies outside the historical interquartile range (IQR) by more than $1.5 \times \text{IQR}$; (iv) *Model reliability warning*: CCIDs identified as low- R^2 failure profiles (e.g., frequent zeros or extreme spikes) are flagged for manual review rather than automated decision-making.

These thresholds are intended as initial operational heuristics and can be adjusted to align with each utility’s policies, service context, and historical incident patterns.

Cost–benefit and NRW implications (conceptual). The dashboard supports allocating limited inspection resources by prioritizing customers with higher expected operational impact. A simple expected-benefit proxy can be computed as:

$$\text{PriorityScore}_i = w_1 \text{Risk}_i + w_2 \text{Impact}_i - w_3 \text{Cost}_i \quad (15)$$

Where Risk_i is derived from the anomaly thresholds, Impact_i can be approximated by the CCID’s typical consumption level (or tariff-weighted usage), and Cost_i reflects travel/inspection effort. In practice, utilities can map confirmed anomalies to potential NRW drivers (e.g., leakage, meter under-registration, or illegal connections), supporting a more targeted inspection plan.

4. Conclusion

This study proposes a data-driven temporal framework for water-consumption monitoring with spatial visualization, leveraging K-Means clustering and STL-LSTM forecasting. Applied to PDAM customer data, customers were segmented into four consumption categories, providing an interpretable overview of usage patterns; clustering quality was reflected in a Silhouette Score of 0.4473 and a Davies–Bouldin Index of 0.7658. Of the 42 customers included in the clustering stage, only 21 CCIDs met the predefined data quality thresholds and were therefore eligible for forecasting. The STL–LSTM approach delivered the most competitive predictive performance, achieving R^2 values up to 0.954629. Finally, combining clustering and forecasting output into a map-based interface supports operational monitoring, such as prioritizing field checks for locations with atypical patterns. Future work includes expanding the data set’s coverage to broader service areas, aligning alert thresholds and follow-up rules with operational criteria, and improving robustness by benchmarking additional forecasting and anomaly-detection approaches for irregular consumption profiles.

Author Contributions Statement

Salsabila Septi Sukmayanti – Conceptualization, methodology, data curation, software implementation, and writing draft paper.

Sudianto Sudianto – Conceptualization, Supervision, validation, and writing (review and editing of the manuscript).

Aminatus Sa’adah – Formal analysis, provided guidance on K-Means clustering analysis and manuscript review.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest.

Funding Declaration

This research received no external funding.

Data Availability Statement

The dataset used in this study is not publicly available due to institutional restrictions. However, the data processing procedures are described in detail in this paper.

Ethical Declarations

This study does not involve human participants or animals.

Acknowledgment

The authors would like to thank PDAM Tirta Wijaya Cilacap for providing access to the data used in this study.

Declaration of Generative AI in Scholarly Writing

The authors used AI-assisted tools to support language refinement and clarity improvement during the writing process. All scientific content remains the responsibility of the authors.

Abbreviations

The following abbreviations are used in this manuscript:

CCID - Customer Connection ID

PDAM - Regional Water Supply Company

STL - Seasonal-Trend Decomposition using Loess

LSTM - Long Short-Term Memory

K-Means - K-Means Clustering

NRW - Non-Revenue Water

MAE - Mean Absolute Error

RMSE - Root Mean Squared Error
 MAPE - Mean Absolute Percentage Error
 MinMax - Min-Max Normalization
 Train-Test Split - Training and Testing Data Split

Appendix

To support reproducibility, this appendix provides the complete implementation details used in the study, including preprocessing, STL configuration, LSTM architecture and training setup, and the evaluation protocol. The code and processed datasets will be made publicly available in an online repository (URL to be provided) to support reproducibility.

A. Preprocessing and Data Split

Preprocessing pipeline.

1. **ID standardization:** CCID converted to string and leading zeros preserved.
2. **Coordinate join:** CCID joined with georeferenced latitude/longitude.
3. **Quality filtering:** thresholds applied as listed in Table 10.
4. **Feature extraction (clustering):** per-CCID mean and standard deviation computed from non-missing observations.
5. **Normalization:** features scaled using *[StandardScaler/MinMaxScaler]* fitted on the training set.

Table 10. Preprocessing thresholds and forecasting split settings.

Parameter	Value
Quality Filter Thresholds	
Minimum active length	60 months
Maximum missingness ratio	20%
Maximum zero-dominance ratio	30%
Forecasting Split Protocol	
Forecast horizon	12 months
Train period	2015–2023
Validation period	2022–2024
Test period	2024–2025
Random seed	42

B. STL Decomposition Parameters

The STL decomposition settings used in this study are summarized in Table 11.

Table 11. STL decomposition settings used in this study (per customer).

Parameter	Value
STL Configuration	
Implementation	statsmodels STL
Input frequency	Monthly (MS)
Missing handling	Linear interpolation
Seasonal period (p)	12 months
Robust weighting	True
Seasonal smoother length (n_s)	7
Trend smoother length (n_t)	Auto
Low-pass length (n_l)	Auto
Degrees	1
Jump parameters	1
Fit iterations	inner=2, outer=15
Output components	trend, seasonal, residual
LSTM input component	residual

C. LSTM Hyperparameters and Training Details

The LSTM architecture and training configuration used in this study are summarized in Table 12.

Table 12. LSTM architecture and training settings used in this study (per customer).

Parameter	Value
Model Architecture	
Input window length (w)	12 months
LSTM layer 1 Dropout LSTM layer 2	64 units
Dropout Output layer	0.2
Activation	32 units
	0.2
	Dense(1) tanh / sigmoid
Training Configuration	
Target series	Residual component
Scaling	Min-Max normalization
Loss function	MAE
Optimizer	AdamW
Learning rate	0.001
Batch size	32
Max epochs	200
Validation strategy	10% split
Early stopping	patience = 10
Learning rate schedule	none
Regularization	none

D. Evaluation Protocol

Evaluation is reported only for the held-out test period. For each CCID, MAE, RMSE, and R^2 (Section 2.6) are computed at the same test timestamps and then summarized by averaging across customers. In the STL–LSTM pipeline, the LSTM is trained with a 12-month input window to predict the STL residual component; forecasts are reconstructed in the original consumption scale by adding the corresponding STL trend and seasonal components to the predicted residual. All baseline methods are evaluated using the identical train/validation/test split and the same test timestamps, so that each CCID is assessed under the same evaluation setting.

References

- [1] Satija, K. and Chhatlani, M.: Infrastructure and industrialization: An SDG-9 study of Gujarat. *International Journal of Science and Social Science Research* 3(1), 310–315 (2025)
- [2] J. C. Carrasco-Jiménez, F. Baldaro, and F. Cucchietti, “Detection of Anomalous Patterns in Water Consumption: An Overview of Approaches,” in *Intelligent Systems and Applications*, vol. 1250, K. Arai, S. Kapoor, and R. Bhatia, Eds., *Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2021, pp. 19–33, doi: 10.1007/978-3-030-55180-3_2.
- [3] F. Ali, D. L. Lestari, M. D. Putri, and K. N. Azmi, “Strategy Analysis for the Fulfilment of Clean Water Needs Through Piped-Water Service in Metropolitan City during the COVID-19 Pandemic,” *International Journal of Technology*, vol. 15, no. 5, p. 1237, Sep. 2024, doi: 10.14716/ijtech.v15i5.6160.
- [4] “Tirta Wijaya – Siap melayani kebutuhan air bersih masyarakat Cilacap,” Accessed: Dec. 12, 2025. [Online]. Available: <https://www.pdamcilacap.co.id/>
- [5] G. Rajan and S. Li, “A Systematic Literature Review on Flow Data-Based Techniques for Automated Leak Management in Water Distribution Systems,” *Smart Cities*, vol. 8, no. 3, p. 78, Apr. 2025, doi: 10.3390/smartcities8030078.
- [6] H. Guo, X. Liu, and Q. Zhang, “Identifying daily water consumption patterns based on K-means Clustering, Agglomerative Hierarchical Clustering, and Spectral Clustering algorithms,” *AQUA – Water Infrastructure, Ecosystems and Society*, vol. 73, no. 5, pp. 870–887, May 2024, doi: 10.2166/aqua.2024.294.
- [7] F. A. H. Irfansyah, *Deteksi Anomali Pada Segmentasi Konsumsi Air Pelanggan PDAM dengan Metode K-Means Dalam bentuk buku karya ilmiah*. Universitas Telkom, S1 Teknologi Informasi (Kampus Surabaya), 2024. Accessed: Oct. 30, 2025. [Online]. Available: <https://openlibrary.telkomuniversity.ac.id/pustaka/213437/deteksi-anomali-pada-segmentasi-konsumsi-air-pelanggan-pdam-dengan-metode-k-means-dalam-bentuk-buku-karya-ilmiah.html>
- [8] K.-M. Koo, K.-H. Han, K.-S. Jun, G. Lee, J.-S. Kim, and K.-T. Yum, “Performance Assessment for Short-Term Water Demand Forecasting Models on Distinctive Water Uses in Korea,” *Sustainability*, vol. 13, no. 11, p. 6056, May 2021, doi: 10.3390/su13116056.
- [9] J. Liu, “Forecasting Short-Term Water Demands with an Ensemble deep learning Model for a Water Supply System,” unpublished manuscript (details to be completed).
- [10] M. R. Jhaerol and S. Sudianto, “Implementation of Chatbot for Merdeka Belajar Kampus Merdeka Program using Long Short-Term Memory,” *J. Nas. Pendidik. Tek. Inform. (JANAPATI)*, vol. 12, no. 2, pp. 253–262, Jul. 2023, doi: 10.23887/janapati.v12i2.58794.
- [11] M. S. Rahim, K. A. Nguyen, R. A. Stewart, T. Ahmed, D. Giurco, and M. Blumenstein, “A clustering solution for analyzing residential water consumption patterns,” *Knowledge-Based Systems*, vol. 233, p. 107522, Dec. 2021, doi: 10.1016/j.knsys.2021.107522.
- [12] D. Chen, J. Zhang, and S. Jiang, “Forecasting the Short-Term Metro Ridership With Seasonal and Trend Decomposition Using Loess and LSTM Neural Networks,” *IEEE Access*, vol. 8, pp. 91181–91187, 2020, doi: 10.1109/ACCESS.2020.2995044.

- [13] N.-H. Duong, M.-T. Nguyen, T.-H. Nguyen, and T.-P. Tran, "Application of Seasonal Trend Decomposition using Loess and Long Short-Term Memory in Peak Load Forecasting Model in Tien Giang," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11628–11634, Oct. 2023, doi: 10.48084/etasr.6181.
- [14] S. Gadal, R. Mokhtar, M. Abdelhaq, R. Alsaqour, E. S. Ali, and R. Saeed, "Machine Learning-Based Anomaly Detection Using K-Mean Array and Sequential Minimal Optimization," *Electronics*, vol. 11, no. 14, p. 2158, Jul. 2022, doi: 10.3390/electronics11142158.
- [15] J. Li, H. Izakian, W. Pedrycz, and I. Jamal, "Clustering-based anomaly detection in multivariate time series data," *Applied Soft Computing*, vol. 100, p. 106919, Mar. 2021, doi: 10.1016/j.asoc.2020.106919.
- [16] J. P. Llerena Can˜a, J. Garcıa Herrero, and J. M. Molina Lo'pez, "Forecasting Nonlinear Systems with LSTM: Analysis and Comparison with EKF," *Sensors*, vol. 21, no. 5, p. 1805, Mar. 2021, doi: 10.3390/s21051805.
- [17] Q. Yu, B. A. Tolson, H. Shen, M. Han, J. Mai, and J. Lin, "Enhancing long short-term memory (LSTM)-based streamflow prediction with a spatially distributed approach," *Hydrology and Earth System Sciences*, vol. 28, no. 9, pp. 2107–2122, May 2024, doi: 10.5194/hess-28-2107-2024.
- [18] "A hybrid approach combining the multi-dimensional time series k-means algorithm and long short-term memory networks to predict the monthly water demand according to the uncertainty in the dataset," doi: 10.1007/s12145023-00976-y. (Online source details to be completed.)
- [19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [20] S. M. Al-Selwi et al., "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *Journal of King Saud University Computer and Information Sciences*, vol. 36, no. 5, p. 102068, Jun. 2024, doi: 10.1016/j.jksuci.2024.102068.
- [21] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information*, vol. 15, no. 9, p. 517, Aug. 2024, doi: 10.3390/info15090517.
- [22] V. Permadi, S. Tahalea, and R. Agusdin, "K-MEANS AND ELBOW METHOD FOR CLUSTER ANALYSIS OF ELEMENTARY SCHOOL DATA," *PROGRES Pendidik*, vol. 4, pp. 50–57, Jan. 2023, doi: 10.29303/prospek.v4i1.328.
- [23] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, "The Silhouette coefficient and the DaviesBouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters," *PeerJ Computer Science*, vol. 11, p. e3309, Nov. 2025, doi: 10.7717/peerj-cs.3309.
- [24] Y. Zhou, X. He, S. Wang, S. Hu, X. Sun, and J. Huang, "Hybrid GNSS time-series prediction method based on ensemble empirical mode decomposition with long short-term memory," *Discover Applied Sciences*, vol. 7, no. 1, p. 61, Jan. 2025, doi: 10.1007/s42452-024-06409-9.
- [25] Z. Aziz and R. Bestak, "Insight into Anomaly Detection and Prediction and Mobile Network Security Enhancement Leveraging K-Means Clustering on Call Detail Records," *Sensors*, vol. 24, no. 6, p. 1716, Jan. 2024, doi: 10.3390/s24061716.
- [26] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," *Heliyon*, vol. 11, no. 2, 2025, doi: 10.1016/j.heliyon.2025.e41953.
- [27] B. A. Hassan, N. B. Tayfor, A. A. Hassan, A. M. Ahmed, T. A. Rashid, and N. N. Abdalla, "From A-to-Z review of clustering validation indices," *Neurocomputing*, vol. 601, p. 128198, Oct. 2024, doi: 10.1016/j.neucom.2024.128198.

Authors' Profiles



Salsabila Septi Sukmayanti is a final-year undergraduate student in Informatics Engineering at Telkom University, Purwokerto, Indonesia. She is currently completing her final project on clustering and time-series forecasting, supported by geospatial visualization. Her research interests include machine learning, anomaly detection, and spatial data analytics.



Sudianto Sudianto received a master's degree in computer science from IPB University, Indonesia. His research interests include Machine Learning and Remote Sensing. He is currently a Lecturer at the School of Computing, Department of Informatics from Telkom University, Purwokerto, Indonesia.



Aminatus Sa'adah is a lecturer at Telkom University, specializing in Applied Mathematics, focusing on mathematical modeling, optimization, and control theory. She completed her Master's degree in Mathematics at the Bandung Institute of Technology (2022) and earned her Bachelor's degree in Mathematics from Airlangga University (2018).

How to cite this paper: Salsabila Septi Sukmayanti, Sudianto Sudianto, Aminatus Sa'adah, "A Data-Driven Temporal Framework for Water Consumption Monitoring with Spatial Visualization Using K-Means and STL-LSTM", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.2, pp. 208-225, 2026. DOI:10.5815/ijigsp.2026.02.11