Bayanika Journal of Artificial Intelligence and Data Science

journal homepage: https://bayanika.org/index.php/bjaids

Regime-Aware Forecasting of Customer Water Consumption: Segmented Evaluation, Fair Benchmarking, and Significance Testing for Classical and LSTM Models

Research Article

Journal: BJAIDS, Vol. 1, No. 1, pp. 13-24, 2026

Delsi Kariman*
Department of Data Science, Faculty of Science and Technology, Universitas PGRI Sumatera Barat, Padang, 25143, Indonesia
E-mail: delsik@upgrisba.ac.id
ORCID iD: https://orcid.org/0000-0001-6063-9189
*Corresponding Author

Nengsi Syaputri
Department of Data Science, Faculty of Science and Technology, Universitas PGRI Sumatera Barat, Padang, 25143, Indonesia
E-mail: nengsisyaputri2022@gmail.com
ORCID iD: https://orcid.org/0009-0002-4325-8944

Received: 16 June 2026; Revised: 22 June 2026; Accepted: 29 June 2026; Published: 30 June 2026

Abstract: Planning clean-water production and distribution at a Regional Drinking-Water Utility (PDAM) requires reliable consumption forecasts at the per-customer-address level—a series with many small and zero values that conventional percentage metrics assess poorly. Prior water-forecasting studies typically report aggregate accuracy and claim the superiority of complex models without fair baselines, per-demand-pattern evaluation, or significance testing. This study proposes a regime-aware evaluation framework for forecasting customer water consumption at PDAM Tirta Langkisau Batang Kapas (1,407 customers; 47,547 records; January 2023–December 2025). Each customer is classified into one of four demand regimes (smooth, intermittent, erratic, lumpy) using the Syntetos–Boylan scheme, after which nine models are compared fairly: Naive, Seasonal-Naive, Moving Average, Croston, SBA, TSB, and three multi-input LSTM variants (MSE, two-stage, and Tweedie loss). Models are evaluated with scaled metrics (WAPE, MASE, RMSSE) across multiple seeds and tested with the Diebold–Mariano test. The Naive baseline proves very strong (aggregate WAPE 29.90%) and remains superior on the smooth, intermittent, and erratic regimes; intermittent-demand methods (Croston/SBA/TSB) do not help because intermittency is mild (median ADI ≈ 1). The Tweedie-loss LSTM achieves the lowest RMSE (6.54 m³) and is the only model that significantly outperforms Naive on squared error (DM = −5.367; $p < 0.001$), while address embedding provides no measurable advantage. The main contribution is not a new algorithm but a fair, segmented, regime-aware framework combining demand-regime classification, multi-seed scaled benchmarking, and significance testing—making claims of model superiority verifiable and transferable to other utilities.

Keywords: Water consumption forecasting, Demand regime classification, Syntetos–Boylan, Tweedie loss, Diebold–Mariano, WAPE

This work is licensed under the Creative Commons Attribution 4.0 International License.

1. Introduction

Access to clean water underpins sanitation, household activities, and people's quality of life. In Indonesia, the equitable distribution of water supply remains uneven - Statistics Indonesia (BPS) records disparities in access to adequate drinking water across regions amid demand that keeps rising as the population grows [1]. Consequently, Regional Drinking-Water Utilities (PDAM) require accurate, data-driven planning for production and distribution [2].

In practice, each customer's water demand fluctuates, making it difficult for the PDAM to anticipate. Prediction errors trigger an imbalance between production and distribution, ranging from declining water pressure and service disruptions to wasted operational resources [2,3]. A similar problem occurs at PDAM Tirta Langkisau Batang Kapas, whose consumption data over the January 2023–December 2025 period exhibit patterns that change over time with indications of trend and seasonality, yet have not been optimally exploited to build a predictive model.

As time-series data, water consumption demands modeling that respects the chronological order while capturing its trend and seasonal components. Long Short-Term Memory (LSTM) is a Recurrent Neural Network architecture designed to capture long-term dependencies in sequential data and to overcome the vanishing-gradient problem [4,5]. Several studies have applied LSTM to water-demand forecasting and reported high accuracy; for example, Simanjuntak et al. [6] obtained a MAPE of 1.54% and Sari et al. [7] obtained a MAPE of 2.52%, while Agustina et al. [8] employed a backpropagation artificial neural network to predict PDAM water distribution.

However, the very low percentage errors reported in those studies are generally obtained on aggregate consumption series (totals per region or per city) rather than at the per-customer-address level, which contains many small or zero values. Aggregate reporting obscures the heterogeneity in demand patterns across customers, making claims of complex-model superiority difficult to verify. Moreover, three important methodological practices are still rarely applied consistently in PDAM water forecasting: (i) fair testing against simple baselines using scaled metrics; (ii) evaluation disaggregated by demand pattern/regime; and (iii) statistical significance testing of accuracy differences. As a result, it is often unclear whether the reported improvements are genuinely meaningful or merely small numerical differences relative to a strong baseline.

Another challenge is structural: at the per-address level, many targets are small or zero, so the standard squared loss (MSE) and percentage metrics become less representative. Classical intermittent-demand forecasting approaches (Croston, SBA, TSB) and reformulated zero-handling (two-stage models and Tweedie loss) are potentially relevant but have not been tested systematically in this context. Furthermore, the spatial-novelty claim of address embedding in a multi-input LSTM architecture - including the one proposed by the authors in a preliminary study on the same dataset [9]- has never been validated through a controlled ablation.

Based on these gaps, this study proposes a regime-aware evaluation framework to predict PDAM customers' water consumption. Specifically, the study seeks to answer the main question: do deep-learning models (LSTM) genuinely outperform simple baselines for predicting water consumption at the PDAM per-customer-address level when evaluated fairly, segmented by demand regime, and tested for statistical significance? This main question is broken down into four specific questions: (i) how are customer demand regimes distributed under the Syntetos–Boylan scheme, and what are the implications for model selection; (ii) can the simple baseline (Naive) be significantly beaten by classical intermittent-demand methods and LSTM variants under scaled metrics (WAPE, MASE, RMSSE) and the Diebold–Mariano test; (iii) does reformulated zero-handling via a two-stage LSTM and Tweedie loss yield measurable improvement over the standard MSE-LSTM; and (iv) does the address representation (embedding) contribute meaningfully to accuracy when tested through a controlled ablation? To answer these questions, the contributions of this study are as follows:

(1) Per-customer demand-regime classification using the Syntetos–Boylan scheme (ADI vs CV^2) into four categories (smooth, intermittent, erratic, lumpy) as the basis for segmented evaluation.

(2) A fair and rigorous benchmark of nine models (three baselines, three intermittent-demand methods, and three LSTM variants) using scaled metrics (WAPE, MASE, RMSSE) in a multi-seed setting, together with the Diebold–Mariano significance test.

(3) Reformulated zero-handling via a two-stage LSTM (zero/non-zero classification followed by magnitude regression) and an LSTM with Tweedie loss, compared with the standard MSE-LSTM.

(4) A controlled ablation of the address representation (full embedding vs one-hot vs no embedding) to test the spatial-novelty claim empirically.

2. Related Work

This section reviews relevant studies and positions the present work within the literature, emphasizing methodological trends and the limitations of prior studies.

2.1. Recent Development in the Research Area

Deep-learning approaches, particularly LSTM [4], are widely used for water-demand forecasting because they can model non-linear patterns and temporal dependencies [6,7,10], with hybrid variants such as Wavelet–CNN–LSTM reported to improve accuracy on urban water-demand series [11]. However, at the per-customer-address level, Pesantez et al. [12] show that water-consumption data contain many zero values and high variability (a median of zero), making forecasting far harder than for aggregate series - the very context this study focuses on. In the intermittent-demand setting, classical methods remain the reference: Syntetos and Boylan [13] propose a bias correction (the Syntetos–Boylan Approximation, SBA); while Teunter, Syntetos, and Babai [14] develop the TSB method, which updates the demand probability every period and is thus more responsive to demand obsolescence. Demand-pattern classification based on ADI and CV^2 has also been used for dynamic model selection: Erjiang et al. [15] group retail products into the four Syntetos–Boylan regimes and then adaptively select or weight forecasting models per regime, demonstrating an advantage over both baselines and forecasting-competition winners - underscoring that model selection should be conditioned on the demand regime.

On the evaluation side, the M5 competition underscored the importance of scaled metrics normalized against the naive error, such as RMSSE and WRMSSE [16]. Hyndman and Koehler [17] introduced MASE as a metric that is reliable across series and robust to zero values. To statistically compare the accuracy of two forecasters, the Diebold–Mariano test [18] is the standard. For non-negative targets with many zeros and a right tail, the Tweedie family of distributions within exponential-dispersion models [19,20] provides a suitable loss function because it bridges the point-mass component at zero and the positive continuous component within a single framework.

2.2. Research Gap

Most water-forecasting studies at the PDAM scale report aggregate accuracy and rarely (i) test simple baselines fairly with scaled metrics, (ii) disaggregate the evaluation by demand regime to reveal where complex models are genuinely useful, or (iii) apply statistical significance testing. Moreover, claims about the benefits of spatial representations, such as address embeddings, are generally not validated through controlled ablation studies, making it hard to separate the feature's real contribution from training variation. Regime-aware model-selection approaches have indeed been applied to retail forecasting [15], but they remain limited to classical models and have not been tested in the water-utility context, which demands per-address zero handling, deep-learning variants, and statistical significance testing. This study closes that gap with a transparent regime-aware evaluation framework, including informative reporting of negative results [13,14].

3. Methodology

This section, in turn, describes the problem formulation, the regime-classification scheme, and the proposed models and procedures.

3.1. Problem Formulation

Let each customer have a monthly water-usage series u_t (m³). With a window size W = 12, input–target pairs are formed in which the 12-month history is used to predict the following month:

$$\hat{y}_i = f\left(u_{t-12,t-1}; s_t, a; \theta\right), \tag{1}$$

where $u_{t-12,t-1}$ is the most recent 12-month sequence, s_t is the seasonal feature, a is the customer-address index, f is the forecasting model, and θ is the model parameters. The seasonal feature is encoded cyclically through a sine–cosine transformation of the month, $s_t \left[sin\left(\frac{2\pi m}{12}\right), cos\left(\frac{2\pi m}{12}\right)\right]$ so that the annual pattern is learned consistently.

3.2. Demand-Regime Classification (Syntetos-Boylan)

For each customer, the Average Demand Interval (ADI) - the ratio of the number of periods to the number of months with non-zero usage - and the squared coefficient of variation (CV^2) of the non-zero demand sizes are computed:

$$ADI = \frac{n}{k}, CV^2 = \left(\frac{\sigma_{nz}}{\mu_{nz}}\right)^2, \tag{2}$$

where n is the number of periods, k the number of non-zero months, and σ_{nz} and μ_{nz} the standard deviation and mean of non-zero usage. Based on the Syntetos–Boylan thresholds (ADI = 1.32; CV^2 = 0.49), customers are categorized into four regimes as shown in Table 1.

Table 1. Demand-regime categorization under the Syntetos–Boylan scheme

Regime	ADI	CV^2
Smooth	< 1.32	< 0.49
Erratic	< 1.32	≥ 0.49
Intermittent	≥ 1.32	< 0.49
Lumpy	≥ 1.32	≥ 0.49

3.3. Comparison Models and Proposed Algorithm

Nine models are compared. Three baselines: Naive (the last month's value), Seasonal-Naive (the value 12 months earlier), and Moving Average (the 12-month mean). Three intermittent-demand methods: Croston [10], SBA [13], and TSB [14], applied per window over the 12-month history. Three multi-input LSTM variants share a common base architecture - following the model design from the authors' preliminary study [9] - that combines the usage sequence, seasonal features, and address embedding, followed by two LSTM layers (64 and 32 units) with 0.2 dropout and a dense layer:

(1) LSTM-MSE: a linear output with Mean Squared Error loss in the normalized log1p space (replicating the standard approach).

(2) Two-stage LSTM: $\hat{y} = P(y > 0) \times E[y|y > 0]$; the classifier is trained on all windows (binary cross-entropy), while the magnitude regressor is trained only on windows with non-zero targets.

(3) LSTM-Tweedie: a softplus output in the original m³ space that minimizes the Tweedie deviance.

The Tweedie loss function with power $p \in (1,2)$ is formulated as:

$$L_{Tweedie} = \frac{1}{N}\sum\left[\frac{-y.\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right], \tag{3}$$

where μ is the non-negative prediction and p = 1.5. All LSTM models are trained with the Adam optimizer [21], EarlyStopping, and ReduceLROnPlateau. The overall procedure is summarized in Algorithm 1.

Algorithm 1: Regime-aware benchmark for water-consumption prediction

Input: Per-customer usage series u_t, window size W = 12, set of seeds S
Output: Aggregate and per-regime metrics, Diebold–Mariano test results

1 Clean the data: remove negative values, winsorize at the 99th percentile, normalize addresses;
2 For each customer: compute ADI and CV², assign the regime label (Table 1);
3 Form windows (12 → t+1) with regime labels; split chronologically at the time cut-off (train through July 2025, test August–December 2025);
4 Compute the MASE/RMSSE scale from the one-step naive error on the training data;
5 Train/apply the baselines, intermittent methods, and LSTM variants (multi-seed over S);
6 Report aggregates and per-regime WAPE/MASE/RMSSE; run the Diebold–Mariano test vs Naive;
7 return the metrics and significance-test results

4. Experiments

4.1. Dataset and Preprocessing

This study uses secondary data on the monthly water consumption of PDAM Tirta Langkisau Batang Kapas customers from January 2023 to December 2025. After month parsing and cleaning, 47,547 rows were obtained, covering 1,407 customers over 36 months. One row with a negative value (a recording error) was removed, and usage was winsorized at the 99th percentile (upper bound ≈ 68 m³) for consistency and comparability across metrics. The proportion of zero values in the target reaches 14.3%, confirming the need for appropriate zero handling. Addresses were normalized (standardizing spelling and abbreviations) into 46 clean, unique addresses used as the spatial feature.

Each window carries its customer's regime label, allowing the evaluation to be broken down by segment. The split is chronological (train through July 2025, test August–December 2025), yielding 26,188 training windows and 5,212 test windows (31,400 in total). The distribution of all windows (training and test) by regime is: Smooth 20,761, Intermittent 4,964, Erratic 3,683, Lumpy 1,800, and all-zero 192 (the proportion of customers is shown in Figure 1). Most customers are classified as smooth, while a median ADI ≈ 1 indicates mild intermittency rather than sparsity.

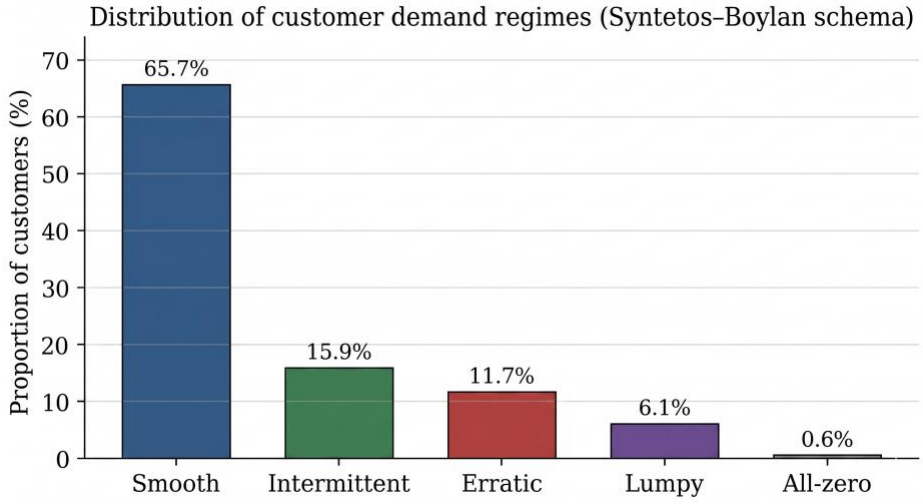

Figure 1. Distribution of customer demand regimes under the Syntetos–Boylan scheme.

4.2. Evaluation Metrics

WAPE is used as the primary metric because it is robust to zero values at the per-address level. MASE and RMSSE are normalized by the one-step naive error on the training data (the M5 convention), so that a value < 1 indicates performance better than persistence [16], [17]. MAE, RMSE, and MAPE* (with flooring ε = 1 m³) are reported as complements. Let a be the actual value and p the prediction:

$$WAPE = \frac{\sum|a - p|}{\sum|a|} \times 100\%, \tag{4}$$

$$MASE = \frac{MAE}{MAE_{naive(train)}}, RMSSE = \frac{RMSE}{RMSE_{naive(train)}}, \quad (5)$$

The significance of accuracy differences relative to the Naive baseline is tested with the Diebold–Mariano statistic [18] over the squared-error difference $d_t = e^2_{model,t} - e^2_{naive,t}$:

$$DM = \frac{\bar{d}}{\sqrt{\frac{var(d)}{n}}}, \quad (6)$$

where $\bar{d}$ is the mean and $var(d)$ the variance of d_t. A significant DM < 0 ($p < 0.05$) indicates that the model is better than Naive.

4.3. Experimental Setup

All LSTM variants are trained with a window size of 12, a batch size of 128, for up to 120 epochs with EarlyStopping (patience 10) and ReduceLROnPlateau. Evaluation is multi-seed (42, 7, 123); results are reported as mean ± standard deviation, and the cross-seed ensemble mean is used for metric reporting. Targets for the LSTM-MSE and two-stage models are log1p-transformed and then scaled with a MinMaxScaler fitted only on the training data to prevent information leakage, whereas LSTM-Tweedie operates in the original m³ space through a softplus output.

5. Results and Discussion

5.1. Aggregate Comparison

Table 2 presents the aggregate performance of all models on the test data. The Naive baseline shows the lowest WAPE (29.90%) together with the best MAE (3.924 m³) and MASE (0.678), confirming that persistence is a very strong baseline on this data. The intermittent-demand methods (Croston, SBA, TSB) are in fact worse than Naive on almost all metrics; this is consistent with the finding that the intermittency is mild (median ADI ≈ 1), so the sparse-demand assumption does not hold. This is an informative negative result: it strengthens the case for not choosing the Croston approach in similar cases.

Among the LSTM variants, LSTM-Tweedie achieves the lowest RMSE (6.54 m³) and the best RMSSE (0.644), outperforming both the Naive baseline (RMSE 7.153 m³) and other deep learning models. Because RMSE penalizes extreme errors more heavily, this result shows that the Tweedie loss effectively suppresses the large errors common at small/bursty values. Conversely, on WAPE and MAE, Naive remains unbeaten, so the advantage of the complex models depends on which type of error is prioritized. This trade-off is visualized in Figure 2.

Table 2. Aggregate performance on the test data (WAPE, MAE, RMSE, MASE, RMSSE, MAPE*)

Model	WAPE (%)	MAE (m³)	RMSE (m³)	MASE	RMSSE	MAPE* (%)
Naive	29.90	3.924	7.153	0.678	0.705	52.97
SeasNaive	65.31	8.571	13.694	1.481	1.349	220.24
MovAvg	39.55	5.190	7.991	0.897	0.787	123.15
Croston	46.37	6.085	9.231	1.051	0.909	205.97
SBA	45.21	5.934	8.980	1.025	0.885	195.63
TSB	40.64	5.334	8.226	0.921	0.810	119.78
LSTM-MSE	34.38	4.511	7.618	0.779	0.750	51.88
LSTM-TwoStage	32.77	4.300	7.282	0.743	0.717	59.56
LSTM-Tweedie	31.69	4.159	6.540	0.719	0.644	84.23

Note: MAPE* is computed with flooring ε = 1 m³. The best value per metric is achieved by Naive (WAPE, MAE, MASE), LSTM-Tweedie (RMSE, RMSSE), and LSTM-MSE (MAPE*).

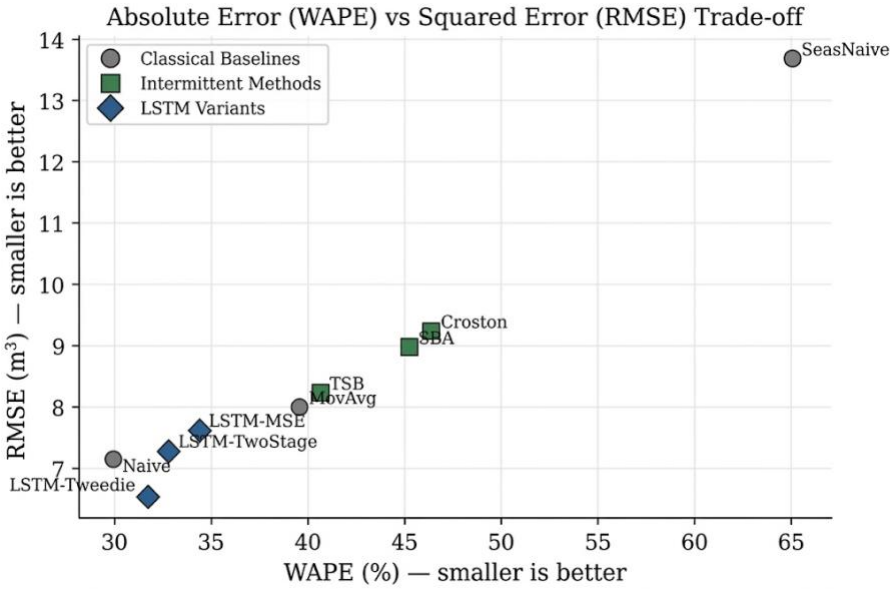

Figure 2. Trade-off between absolute error (WAPE) and squared error (RMSE) across models.

The consistency of the results is reinforced by the multi-seed evaluation: LSTM-MSE obtains a WAPE of 34.50 ± 1.32; the two-stage LSTM 33.66 ± 2.53; and LSTM-Tweedie 31.81 ± 0.43. The smallest standard deviation, for LSTM-Tweedie, indicates good training stability in addition to its superior squared-error accuracy.

5.2. Segmented Evaluation by Regime

Table 3 and Figure 3 present the WAPE per regime. This breakdown changes the interpretation of the aggregate results. In the smooth regime, which covers the majority of customers (65.7%), Naive (WAPE 25.56%) remains the most accurate, followed by LSTM-Tweedie (26.18%) and LSTM-none. In the intermittent and erratic regimes, Naive is again superior (56.05% and 43.97%), whereas the Croston/SBA methods balloon to above 160% due to mismatched assumptions. The advantage of the deep-learning models only emerges in the most difficult lumpy regime: LSTM-MSE (78.99%) slightly outperforms Naive (82.82%).

Thus, the real challenge lies not in the smooth regime - already handled well by persistence - but in the lumpy and erratic regimes, which are low-volume yet highly variable. Aggregate evaluation alone would mask this fact and risks overstating the benefit of complex models.

Table 3. WAPE (%) per demand regime on the test data

Model	Smooth	Intermittent	Erratic	Lumpy
Naive	25.56	56.05	43.97	82.82
SeasNaive	50.36	171.22	114.68	194.32
MovAvg	31.14	96.83	68.46	112.04
Croston	32.88	167.17	75.26	178.10
SBA	32.22	161.95	72.94	172.00
TSB	32.18	92.59	71.35	121.09
LSTM-MSE	29.94	67.99	46.65	78.99
LSTM-TwoStage	28.19	67.55	44.71	83.43
LSTM-Tweedie	26.18	74.10	45.90	90.98

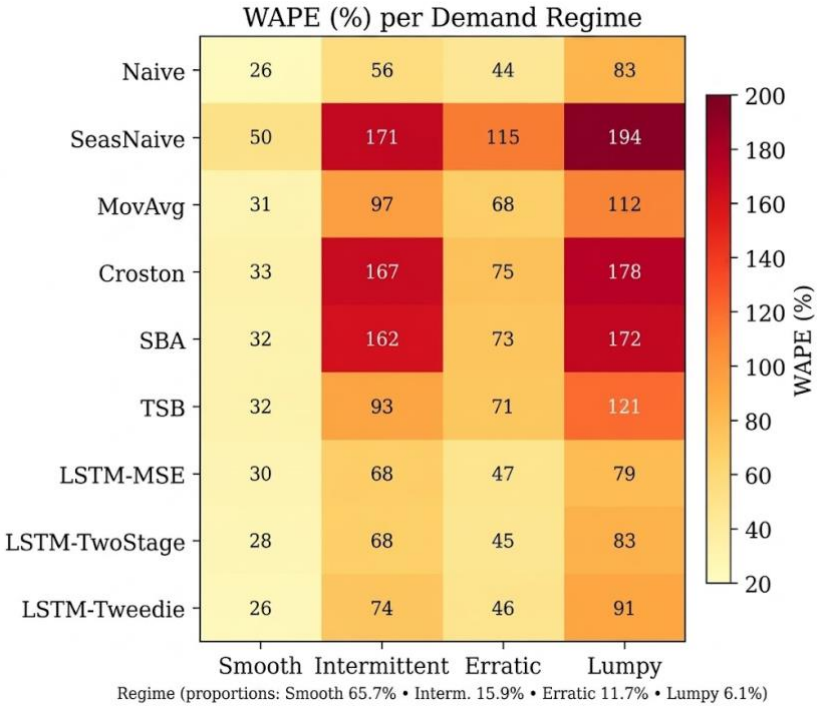

Figure 3. Heatmap of WAPE per regime; darker shades indicate larger errors.

5.3. Diebold-Mariano Significance Test

Table 4 summarizes the Diebold–Mariano test against Naive based on squared error. LSTM-Tweedie is the only model that significantly outperforms Naive (DM = −5.367; $p < 0.001$). Conversely, LSTM-MSE is in fact significantly worse than Naive (DM = +2.883; $p = 0.004$), while the two-stage LSTM shows no significant difference ($p = 0.369$). At the regime level for the two-stage LSTM, Naive is significantly better in the smooth regime (DM = +2.920; $p = 0.004$), whereas in the erratic and lumpy regimes the differences are not significant - confirming that the advantage of complex models in the difficult regimes is not yet statistically strong and requires more data or further tuning.

Table 4. Diebold–Mariano test against the Naive baseline (DM < 0 and $p < 0.05$ means better than Naive)

Comparison	Scope	DM statistic	p-value	Remark
LSTM-MSE vs Naive	Aggregate	+2.883	0.0039	Naive better (significant)
LSTM-TwoStage vs Naive	Aggregate	+0.898	0.3690	Not significant
LSTM-Tweedie vs Naive	Aggregate	−5.367	<0.0001	LSTM-Tweedie better (significant)
LSTM-TwoStage vs Naive	Smooth	+2.920	0.0035	Naive better (significant)
LSTM-TwoStage vs Naive	Intermittent	+0.479	0.6320	Not significant
LSTM-TwoStage vs Naive	Erratic	−1.038	0.2992	Not significant
LSTM-TwoStage vs Naive	Lumpy	−1.334	0.1821	Not significant

5.4. Ablation of the Address Representation

Table 5 and Figure 4 present the ablation of the address representation. Surprisingly, the no-embedding variant (LSTM-none) obtains the best WAPE (32.16%) and a MASE of 0.729, slightly outperforming full embedding (33.65%; MASE 0.763), while one-hot is the worst (42.55%; MASE 0.965). This pattern is consistent across all regimes. This is an important negative result: on data with only 46 unique addresses and a dominance of smooth patterns, address embedding provides no measurable advantage and even risks adding parameters without benefit. The spatial-novelty claim for the multi-input architecture is therefore not empirically supported in this context, and the decision to include spatial features should rest on ablation validation rather than assumption.

Table 5. Ablation of the address representation (aggregate WAPE and MASE, plus WAPE per regime)

Model	WAPE (%)	MASE	Smooth	Intermittent	Erratic	Lumpy
LSTM-embed	33.65	0.763	29.12	67.81	46.36	78.76
LSTM-onehot	42.55	0.965	38.76	71.70	52.61	82.20
LSTM-none	32.16	0.729	27.51	66.42	45.71	78.42

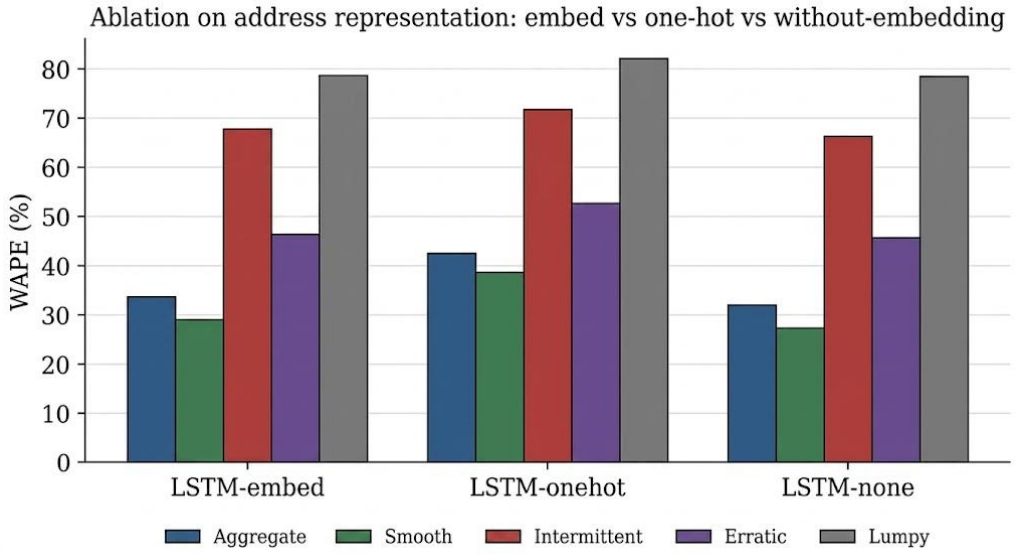

Figure 4. Ablation of the address representation: embedding vs one-hot vs no embedding.

5.5. Discussion and Implications

Synthesizing the findings yields several practical implications. First, in services dominated by smooth-pattern customers, the Naive baseline is already adequate and cost-efficient for routine planning; investment in complex models should be focused on the lumpy/erratic segments - consistent with Erjiang et al. [15], who likewise find that the choice of forecasting model should be conditioned on the demand regime rather than applied uniformly. Second, if the priority is to suppress extreme errors (important for capacity planning and anomaly detection), the LSTM with Tweedie loss is the most appropriate choice because it excels on RMSE and is statistically significant; this aligns with Smyth and Jørgensen [20] and Jørgensen [19], whose Tweedie and exponential-dispersion framework is well suited to non-negative targets with a point mass at zero and a positive right tail. Third, classical intermittent-demand methods are not recommended for data with mild intermittency such as this, consistent with Syntetos and Boylan [13] and Teunter et al. [14], whose methods are designed for genuinely sparse demand (high ADI) and therefore offer little advantage when the intermittency is mild (median ADI ≈ 1). Fourth, the addition of spatial features must be validated through ablation before being claimed as a novelty.

This study has limitations that constrain the generalization of its findings: the number of seeds is still limited (three), the observation period is three years, the scope is a single PDAM, and there are only 46 unique addresses. In particular, the strength of the Naive baseline very likely depends on the characteristics of this data - mild intermittency (median ADI ≈ 1) and a dominance of smooth-pattern customers; in utilities or regions with sparser and burstier demand, the relative advantage of complex models could be larger, so these findings should be re-tested before being generalized. The Tweedie loss excels on squared error but not on MAPE*, so interpretation must consider the metric relevant to the operational objective.

6. Conclusion

This study proposes a regime-aware evaluation framework to predict water consumption among PDAM Tirta Langkisau Batang Kapas customers. Using the Syntetos–Boylan classification, a fair benchmark of nine models with scaled metrics, segmented per-regime evaluation, and the Diebold–Mariano test, it yields a more honest picture than aggregate reporting alone. The Naive baseline proves very strong and remains superior across the

smooth, intermittent, and erratic regimes, whereas classical intermittent-demand methods are ineffective because the intermittency is mild (median ADI ≈ 1). The LSTM with Tweedie loss attains the lowest RMSE (6.54 m^3) and is the only model to significantly outperform Naive on squared error (DM = −5.367; $p < 0.001$), with the best cross-seed stability. The two-stage LSTM, by contrast, shows no statistically significant improvement over Naive ($p = 0.369$), so among the reformulated zero-handling approaches only the Tweedie loss proves beneficial. This does not mean complex models are worthless; rather, model selection should be driven by customer segment and operational objective instead of a single model: Naive is adequate and cost-efficient for routine planning on smooth-pattern customers, whereas the LSTM with Tweedie loss becomes valuable when the priority is suppressing extreme errors - for example, peak-capacity planning and anomaly detection - and in the most challenging lumpy and erratic segments. The ablation shows that address embedding provides no measurable advantage on this data.

The main contribution of this study is methodological-evaluative rather than algorithmic: a transparent regime-aware evaluation framework - binding Syntetos - Boylan classification, multi-seed scaled benchmarking, and the Diebold–Mariano significance test into a single standard procedure - that makes model-superiority claims verifiable and, because it does not depend on the specific characteristics of water data, can be reapplied to demand forecasting for other utilities (electricity, gas) and retail. Informative reporting of negative results strengthens the transparency of this framework. Future work is recommended to increase the number of seeds to 5–10 and add a Friedman–Nemenyi test across customers, to develop a probabilistic pathway (P10–P90 prediction intervals) by leveraging the softplus/Tweedie output foundation, and to broaden the coverage of customers, regions, and observation period so that the results are more generalizable.

Declarations

Author Contributions Statement

Delsi Kariman: conceptualization, methodology, software, formal analysis, investigation, writing - original draft, and writing - review and editing. **Nengsi Syaputri**: provision and curation of data (acquisition of the water-consumption data). All authors have read and approved the final version of the manuscript and take full responsibility for its content.

Conflict of Interest Statement

The authors declare that there is no conflict of interest related to the publication of this article.

Funding Declaration

This research received no external funding.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgment

The authors thank PDAM Tirta Langkisau Batang Kapas for providing the customer water-consumption data.

Declaration of Generative AI in Scholarly Writing

In preparing this manuscript, the authors used generative AI tools solely for language editing and proofreading. The authors reviewed and edited all content and take full responsibility for the published article.

References

[1] S. Sulfiati, A. Rizal, and M. Riswanto, “Analisis Kebutuhan dan Ketersediaan Air Bersih di Universitas Muhammadiyah Palu,” *Jurnal kolaborasi Sains*, vol. 7, no. 1, 2024.

[2] N. Nurdin, N. Suarna, and W. Prihartono, "Algoritma Regresi Linier Sederhana Untuk Prediksi Penggunaan Volume Air Berdasarkan Jenis Pelanggan PDAM," *Jurnal Kecerdasan Buatan dan Teknologi Informasi*, vol. 4, no. 1, pp. 43–52, Jan. 2025, doi: 10.69916/jkbti.v4i1.187.

[3] N. A. Syarifuddin, T. Wahyuni, M. Faisal, M. Syafaat, and A. M. Syamsuri, "Prediksi Pemakaian Air Bulanan di PDAM Kecamatan Tamalate Menggunakan Metode Autoregressive Integrated Moving Average (ARIMA)," *Jurnal Informatika Progres*, vol. 17, no. 2, 2025.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[5] S. H. Noh, "Analysis of gradient vanishing of RNNs and performance comparison," *Information (Switzerland)*, vol. 12, no. 11, Nov. 2021, doi: 10.3390/info12110442.

[6] D. F. Simanjuntak, N. H. Djanggu, and R. Budiman, "Implementasi Long Short-Term Memory Dalam Peramalan Permintaan Air Bersih Di Kota Pontianak," *INTEGRATE: Industrial Engineering and Management System*, vol. 9, no. 2, 2025.

[7] D. P. Sari, L. Karlitasari, and F. D. Wihartiko, "Clean Water Demand Prediction Model Using The Long Short Term Memory (LSTM) Method," *Komputasi: Jurnal Ilmiah Ilmu Komputer dan Matematika*, vol. 20, no. 2, pp. 160–168, Jul. 2023, doi: 10.33751/komputasi.v20i2.8060.

[8] D. Agustina, Moh. Hafiyusholeh, A. Fanani, and D. Prasetijo, "Prediksi Distribusi Air Perusahaan Daerah Air Minum (PDAM) Tirta Dharma Kota Pasuruan Menggunakan Metode Jaringan Syaraf Tiruan Backpropagation," *PROCESSOR: Jurnal Ilmiah Sistem Informasi, Teknologi Informasi dan Sistem Komputer*, vol. 18, no. 1, Apr. 2023, doi: 10.33998/processor.2023.18.1.697.

[9] N. Syaputri, D. Kariman, and I. Fadhli, "Prediksi Konsumsi Air Pelanggan PDAM Tirta Langkisau Menggunakan Long Short-Term Memory," *manuscript submitted for publication*, 2026.

[10] A. Boudhaouia and P. Wira, "A Real-Time Data Analysis Platform for Short-Term Water Consumption Forecasting with Machine Learning," *Forecasting*, vol. 3, no. 4, pp. 682–694, 2021, doi: 10.3390/forecast3040042.

[11] Z. Pu *et al.*, "A hybrid Wavelet-CNN-LSTM deep learning model for short-term urban water demand forecasting," *Front. Environ. Sci. Eng.*, vol. 17, no. 2, p. 22, 2022, doi: 10.1007/s11783-023-1622-3.

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

[13] A. A. Syntetos and J. E. Boylan, "The accuracy of intermittent demand estimates," *Int. J. Forecast.*, vol. 21, no. 2, pp. 303–314, 2005, doi: https://doi.org/10.1016/j.ijforecast.2004.10.001.

[14] R. H. Teunter, A. A. Syntetos, and M. Zied Babai, "Intermittent demand: Linking forecasting to inventory obsolescence," *Eur. J. Oper. Res.*, vol. 214, no. 3, pp. 606–615, 2011, doi: https://doi.org/10.1016/j.ejor.2011.05.018.

[15] E. Erjiang, M. Yu, X. Tian, and Y. Tao, "Dynamic Model Selection Based on Demand Pattern Classification in Retail Sales Forecasting," *Mathematics*, vol. 10, no. 17, 2022, doi: 10.3390/math10173179.

[16] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecast.*, vol. 38, no. 4, pp. 1346–1364, 2022, doi: https://doi.org/10.1016/j.ijforecast.2021.11.013.

[17] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006, doi: https://doi.org/10.1016/j.ijforecast.2006.03.001.

[18] F. X. Diebold and R. S. Mariano, “Comparing Predictive Accuracy,” *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995, doi: 10.1080/07350015.1995.10524599.

[19] B. Jørgensen, “Exponential Dispersion Models,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 49, no. 2, pp. 127–145, 1987, doi: https://doi.org/10.1111/j.2517-6161.1987.tb01685.x.

[20] G. K. Smyth and B. Jørgensen, “Fitting Tweedie’s Compound Poisson Model to Insurance Claims Data: Dispersion Modelling,” *ASTIN Bulletin*, vol. 32, no. 1, pp. 143–157, 2002, doi: DOI: 10.2143/AST.32.1.1020.

[21] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *CoRR*, vol. abs/1412.6980, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106 Accessed: June 14, 2026