

Data-Driven Deman Forecasting to Enable Regenerative and Net-Positive Urban Water Systems: A District-Level Case Study

Semra Sıla Ertuğ*, Nihan Özbaltan²

¹Department of Computer Engineering, Izmir Bakırçay University, Turkey

²Department of Computer Engineering, Izmir Bakırçay University, Turkey

*(nihan.ozbaltan@bakircay.edu.tr) Email of the corresponding author

(Received: 18 February 2026, Accepted: 25 February 2026)

(3rd International Conference on Pioneer and Academic Research ICPAR 2026, February 16-17, 2026)

ATIF/REFERENCE: Ertuğ, S. S. & Özbaltan, N. (2026). Data-Driven Deman Forecasting to Enable Regenerative and Net-Positive Urban Water Systems: A District-Level Case Study, *International Journal of Advanced Natural Sciences and Engineering Researches*, 10(2), 216-226.

Abstract – Global efforts to move beyond conventional sustainability towards regenerative and net positive water systems require robust, spatially explicit intelligence on future water demand. This study develops a data-driven framework to support regenerative urban water planning by forecasting district- and customer-type-specific water consumption and identifying behaviourally district consumer clusters. Using multi-year billing records disaggregated by district, customer category, year and month, we first clean and harmonise subscriber labels and compute per-subscriber consumption. To capture seasonal dynamics, month information is encoded via sine-cosine transformations, while right-skewed distributions of total consumption, subscriber counts and per-subscriber use are stabilized. Log-transformed consumption variables and seasonality components are then used in a HDBSCAN-based behavioural clustering, with cluster characteristics visualized via standardized heatmaps. For forecasting, we construct monthly time series for each district–customer combination, assess temporal completeness and retain series with limited gaps, which are filled using time-based interpolation. Lagged consumption variables and rolling statistics are engineered to represent short-, medium- and long-term dependencies. A LightGBM regression model with recursive forecasting is trained on pre-2022 data and validated on 2022–2024, using seasonality, subscriber attributes, lags and moving-window features as predictors. The model generates monthly projections of per-subscriber use and subscriber counts through 2026, which are combined to obtain total water demand at district level and aggregated to annual indicators. The results reveal districts with rising or stabilising demand, spatial variation in model errors and behaviourally distinct consumer clusters, providing actionable intelligence for circular, regenerative and net-positive water and wastewater interventions at local and regional scales.

Keywords – regenerative water systems; circular water use; urban water demand forecasting; machine learning; LightGBM; behavioural clustering; HDBSCAN

I. INTRODUCTION

Global water demand from households, industry and agriculture has been rising steadily over the last decades, driven by population growth, urbanisation, economic development and changing lifestyles. More

recently, emerging pressures such as rising water requirements of renewable energy technologies, the rapid expansion of data centres and AI-enabled digital infrastructures, and shifting geopolitical dynamics have started to reshape both direct and embedded water use at local, regional and global scales. These intertwined trends not only increase total withdrawals, but also amplify pollution loads and strain water and wastewater infrastructures. Up to %50 of water is lost in some parts of Europe, and approximately %26 of this loss is due to various infrastructural problems [1]. As a result, many regions face growing risks in terms of water scarcity, quality degradation and service reliability, which challenge the limits of incremental efficiency improvements and conventional sustainability approaches.

In response, the water sector is increasingly looking beyond traditional “do less harm” paradigms towards circular, regenerative and net-positive system concepts. Regenerative water systems aim to work with, rather than against, natural processes and cycles; they strive to restore and replenish water-related ecosystems, close water and resource loops, and deliver broader social and ecological co-benefits. Net-positive approaches seek to ensure that water and wastewater infrastructures contribute more to the health of catchments and communities than they take away, for instance by enabling water reuse, energy cogeneration and material recovery at multiple scales. Designing and governing such systems, however, requires much more granular and forward-looking intelligence on how water demand will evolve across space, time and user categories.

Urban water utilities and planners need to understand where, when and for which types of customers water demand pressures will intensify in the coming years in order to prioritise regenerative and circular interventions. Decisions about the sizing and siting of decentralised reuse systems, greywater and rainwater harvesting, advanced treatment technologies, and demand management programmes all depend on reliable projections of future consumption. Yet, in many contexts, demand is still analysed using aggregated indicators and relatively simple extrapolation methods that obscure spatial heterogeneity and behavioural diversity among consumers. This can lead to misaligned investments, under- or over-dimensioned infrastructures, and missed opportunities for targeted net-positive strategies. Early studies have mostly used traditional statistical approaches such as time series analysis and linear regression models to predict water demand [2].

At the same time, advances in data availability and machine learning offer new possibilities for constructing spatially explicit, high-resolution demand forecasts that can directly inform regenerative water system planning. Algorithms such as Artificial Neural Networks (ANN), support vector machine (SVM), extreme learning machine (ELM), and Random Forest have become widely used in water demand forecasting [3]. Successful water demand management requires accurate representation of user behavior, timing, frequency of use, and the diversity in consumption profiles. Utility billing records, when carefully cleaned, harmonised and combined with appropriate feature engineering, can reveal distinct consumption behaviours and temporal patterns at the level of districts and customer types. Clustering methods can uncover behaviourally similar user groups, while gradient boosting models and related algorithms can capture complex, nonlinear dependencies and seasonality in water use [4], [5]. Integrating such data-driven insights into planning processes can strengthen the alignment between projected demand, circular water options and infrastructure design.

In this study, we develop a data-driven framework to support regenerative and net-positive urban water system planning by forecasting water demand at the district and customer-type level and by identifying behaviourally distinct consumer clusters. Using multi-year billing data disaggregated by district, customer category, year and month, we construct temporally consistent time series, engineer seasonality and memory features, and apply an HDBSCAN-based clustering in combination with a LightGBM regression model and recursive forecasting. The resulting projections of per-subscriber consumption, subscriber numbers and total demand through 2026 are analysed to reveal spatial and behavioural heterogeneity in future water use. We discuss how these insights can inform the prioritisation and tailoring of circular and regenerative water and wastewater interventions at local and regional scales, thereby contributing to the broader transition towards net-positive water systems.

II. MATERIALS AND METHOD

Figure 1 presents an overview of the analytical workflow adopted in this study, from raw data through to regenerative planning insights.

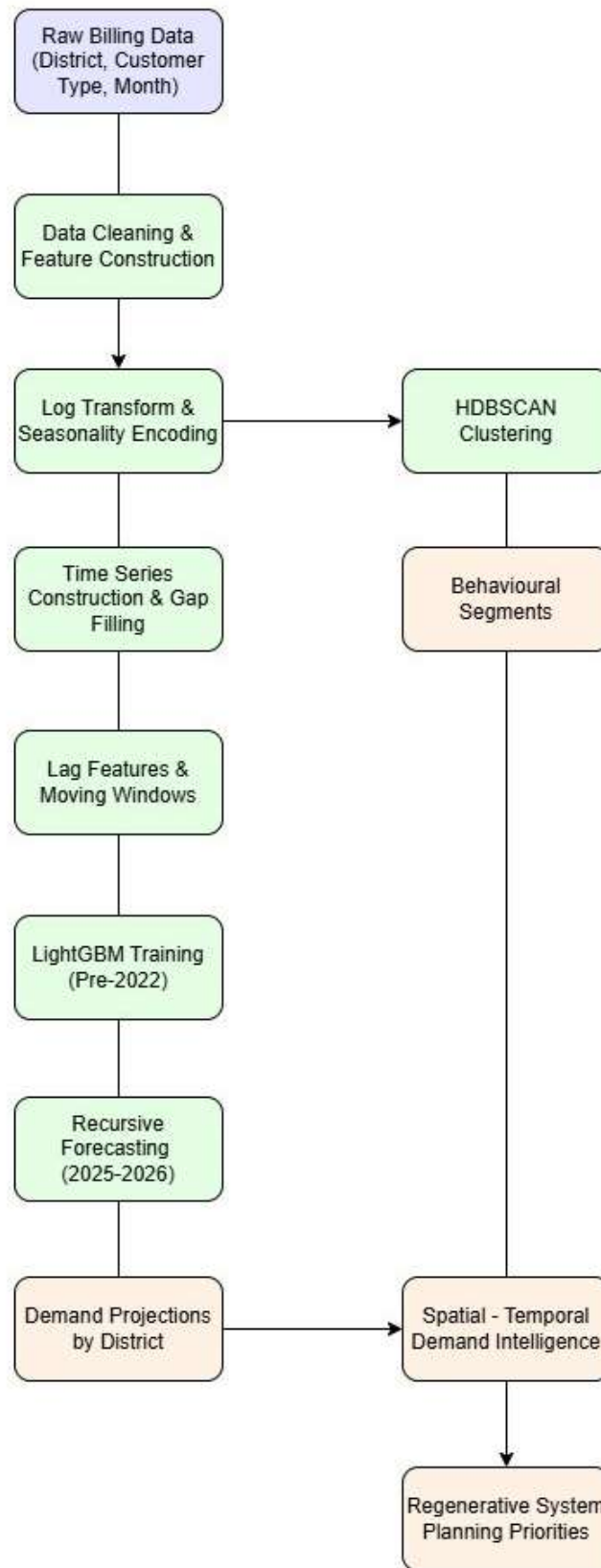


Fig. 1 Analytical workflow from raw billing data to regenerative water system planning insights, showing parallel clustering and forecasting streams that converge to inform spatial prioritisation.

A. Study Context and Data

Raw records were first cleaned to ensure internal consistency and to harmonise categorical labels. Customer type codes were standardised into a set of general categories (e.g., residential, commercial, industrial) to avoid fragmentation due to minor coding differences.

For each observation, per-subscriber consumption was computed by dividing total billed consumption by the corresponding number of subscribers. This indicator is used throughout the study as a central measure of demand intensity.

To explicitly capture intra-annual seasonality, the calendar month was encoded using sine and cosine transformations, i.e., $\sin(2\pi m/12)$ and $\cos(2\pi m/12)$, where m denotes the month. The distributions of total consumption, subscriber counts and per-subscriber consumption exhibited strong right skewness with a small number of high-valued observations. To stabilise variance and reduce the influence of extreme values in regression-based models, logarithmic transformations were applied to these variables. All log transformations were carried out after checking for non-positive values and, where necessary, applying appropriate offsets.

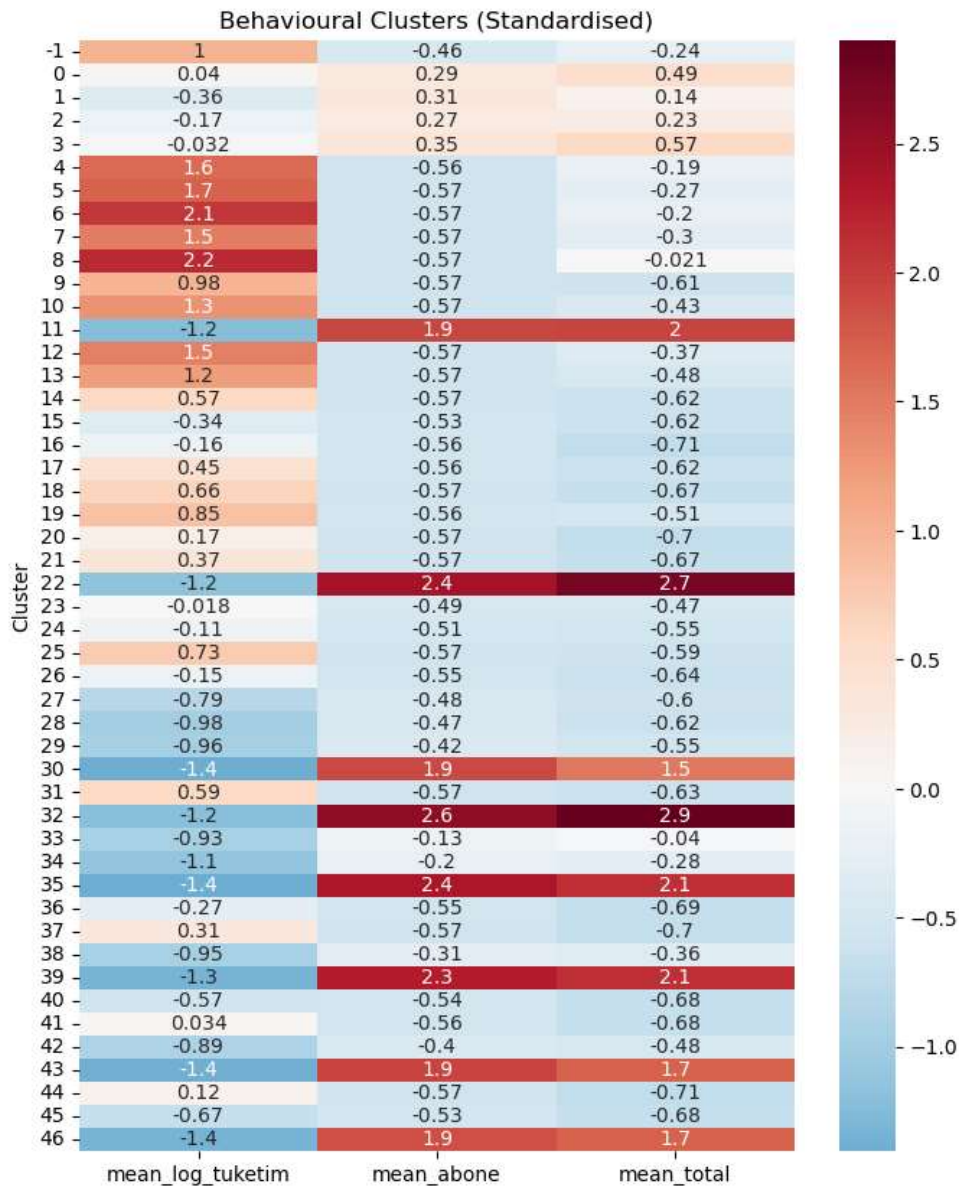


Fig. 2 Standardised feature profiles for the main behavioural clusters of district–customer-type combinations.

B. Behavioural Clustering of Consumption Patterns

To identify groups of customers with similar consumption behaviours, an unsupervised clustering analysis was conducted using the HDBSCAN algorithm. The clustering feature space consisted of log-transformed consumption indicators and the sine–cosine seasonality components, summarised over a suitable reference period for each district–customer-type combination. HDBSCAN was chosen because it can discover clusters of varying density, handle noise points and does not require the number of clusters to be specified a priori.

Cluster-specific summaries were computed for the selected features and standardised to facilitate comparison. These summaries were visualised as heatmaps to highlight characteristic profiles of each behavioural cluster, such as high or low average consumption, strong or weak seasonality and distinct variability patterns. The resulting clusters provide a behavioural segmentation of water users that is later related to projected demand trajectories and used to discuss implications for regenerative and circular interventions.

C. Time Series Construction and Handling of Missing Data

For the forecasting component, monthly time series were constructed for each district-customer-type combination. The data were sorted by district, customer category, year and month, and a monthly date variable was created to ensure a consistent temporal index. For each series, the earliest and latest observed months were identified and used to generate a complete monthly date range. This complete range was then compared with the actually observed months to detect missing periods.

For each series, the number of missing months and the specific missing positions were recorded. Series were then filtered into two subsets: an analysis set with at most two missing months and an experimental set with at most three missing months. Only the analysis set was used for the main forecasting experiments. Within this set, short gaps were filled using time-based interpolation on the monthly time index, after reindexing each series to its full monthly range. This procedure ensured temporal continuity while avoiding the inclusion of highly fragmented or unreliable series.

D. Lagged Features and Moving-Window Statistics

To encode temporal dependencies in consumption, lagged features and moving-window statistics were engineered for each series in the analysis set. Specifically, lagged values of log-transformed per-subscriber consumption at 1-, 3-, 6- and 12-month horizons were generated, capturing short-, medium- and annual-scale memory effects. In addition, rolling means and rolling standard deviations were computed over selected window lengths; for example, a three-month rolling mean was used to represent short-term local trends, whereas a six-month rolling standard deviation captured medium-term variability and volatility in consumption. The creation of these lagged and window-based features inevitably led to missing values at the beginning of each series. Observations affected by these initial feature gaps were removed so that the final modelling dataset contained only fully specified feature vectors. The engineered features were combined with the seasonality encodings, customer category information and district identifiers to form the predictor set used in the forecasting models.

E. Model Training, Validation and Recursive Forecasting

A gradient boosting decision tree model based on the LightGBM library was employed to forecast log-transformed per-subscriber consumption. The full dataset was split into a training period and a validation period in a temporally consistent manner: observations up to (but excluding) 2022 were used for model training, while data from 2022 to 2024 served as a hold-out validation set. This split reflects the operational requirement of using only historical information to inform forecasts of future demand. The input features to the LightGBM model comprised the sine–cosine seasonality components, customer-type indicators, district identifiers, lagged consumption values and moving-window statistics. Model hyperparameters were selected based on a combination of heuristic choices informed by prior work and limited tuning on the training set, while avoiding excessive complexity given the size and structure of the data. Because future periods lack observed consumption values, a recursive multi-step forecasting

strategy was adopted. Starting from the last available observation in 2024, the model was used to predict the next month's log consumption for each district–customer- type series. These predictions were then exponentiated back to the original scale where needed and fed back into the feature generation process as pseudo-observed lag values for subsequent forecast steps. This procedure was repeated iteratively to generate monthly forecasts up to the end of 2026. In parallel, a similar LightGBM-based approach was used to forecast the number of subscribers, allowing total demand to be derived by combining per-subscriber consumption and subscriber count projections.

F. Performance Evaluation and Aggregation of Results

Model performance was evaluated on the 2022–2024 validation period using error metrics computed in the logarithmic domain, such as the root mean squared error (RMSE) of log-transformed consumption. To explore spatial variability in model accuracy, validation errors were summarised by district, providing a diagnostic view of where forecasts were more or less reliable. For interpretation and planning purposes, monthly forecasts of per-subscriber consumption and subscriber numbers were combined to obtain total water demand at the district–customer-type and district-wide levels. These monthly values were then aggregated to yearly indicators for 2025 and 2026, including total annual consumption, average per-subscriber use and average subscriber counts. Year-on-year percentage changes and absolute differences between 2025 and 2026 were calculated to identify districts with particularly strong expected increases or decreases in demand. These patterns, together with the behavioural clusters, were used to discuss priorities and opportunities for implementing circular, regenerative and net-positive water and wastewater interventions in different parts of the urban area.

III. RESULTS

A. Behavioural Clusters of Water Consumption

The HDBSCAN-based clustering revealed several behaviourally distinct groups of district–customer-type combinations. Figure 2 summarises the standardised feature profiles of the main clusters. Broadly, we observed:

- clusters characterised by high average per-subscriber consumption and relatively weak seasonality,
- clusters with moderate consumption but pronounced seasonal peaks, and
- clusters with low and relatively stable consumption throughout the year.

High-consumption clusters were predominantly associated with specific commercial and mixed-use customer types in a subset of districts, whereas low-consumption clusters were more common among residential subscribers in other areas. Clusters with strong seasonal signatures tended to correspond to districts where climatic or tourism-related effects are expected to drive temporary demand spikes. The presence of noise points in the HDBSCAN output indicates that a small fraction of series did not conform well to any dominant behavioural pattern, which is consistent with the heterogeneous nature of urban water use. These behavioural segments provide a first lens on where regenerative and circular measures might need to be tailored differently. For example, high-consumption clusters could be prioritised for water reuse and efficiency interventions, while strongly seasonal clusters may require flexible, seasonally adaptive solutions.

B. Model Calibration and Validation Performance

The LightGBM model for log-transformed per-subscriber consumption achieved satisfactory accuracy on the 2022–2024 validation period. At the aggregate level, the root mean squared error (RMSE) in the logarithmic domain remained within a moderate range across most districts and customer categories, as discussed in the spatial patterns section below. Overall, districts with more complete and less noisy time series exhibited lower validation errors, reflecting more predictable consumption dynamics. In contrast, a few districts showed comparatively higher RMSE values, which can be linked to stronger year-to-year variability, structural changes in subscriber composition or data quality issues. Despite these differences,

the model was generally able to reproduce the main temporal patterns and magnitudes of per-subscriber consumption in the validation period. Importantly, the use of logarithmic transformation constrained the influence of extreme observations and improved the stability of the forecasts, while the combination of lagged features and moving-window statistics allowed the model to capture both short-term fluctuations and longer-term trends.

C. Projected Per-Subscriber Consumption and Subscriber Numbers

Using the recursive forecasting scheme, monthly projections of per-subscriber consumption and subscriber counts were generated for each district–customer-type series up to December 2026. Figure 3 shows an example for a selected district and residential customer type, where the model closely follows the historical trajectory over the validation period and then extends it into the forecast horizon.

Across the study area, the projections suggest that per-subscriber consumption will remain broadly stable or exhibit only modest changes in many districts, while a subset of districts is expected to experience more noticeable increases or decreases. Subscriber numbers are projected to grow in several rapidly developing districts, contributing to rising total demand even where per-subscriber use is relatively stable. In other districts, subscriber counts are expected to plateau, leading to more moderate growth in total demand. These patterns underscore the importance of considering both intensity (per-subscriber use) and scale (number of subscribers) when planning future water and wastewater infrastructures, particularly in the context of regenerative system design where local balances between withdrawals, returns and reuse are critical.

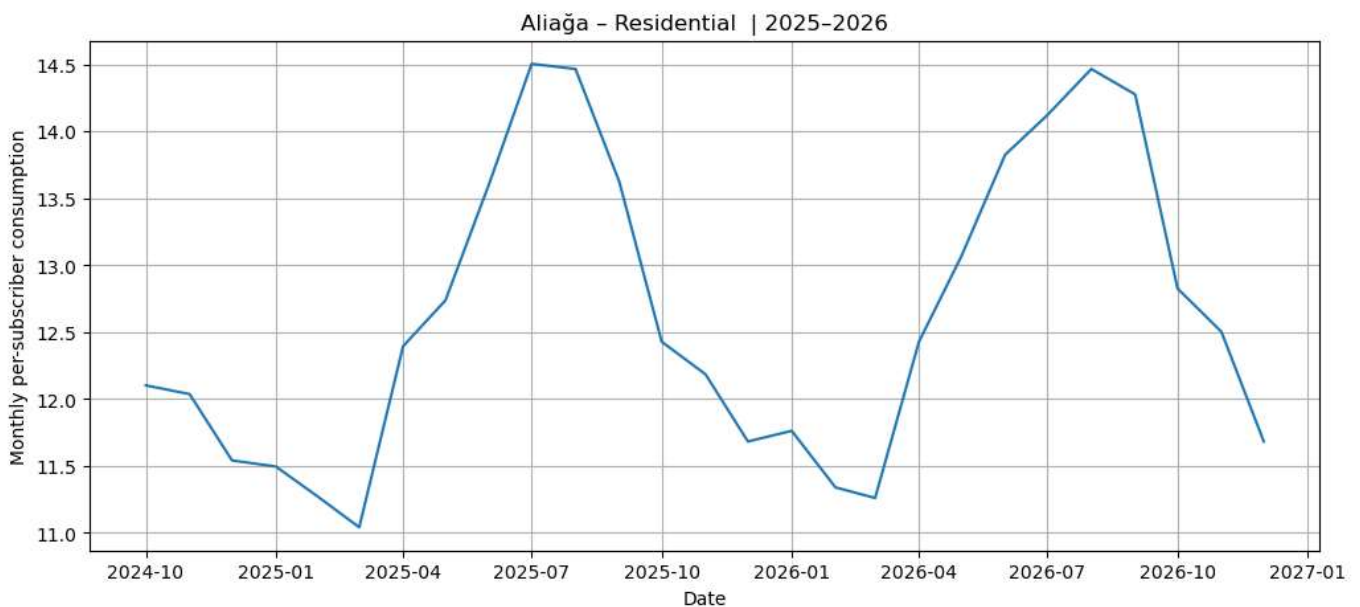


Fig. 3 Monthly per-subscriber consumption forecast for Aliaga residential customers (2025–2026), demonstrating the model’s ability to capture seasonal patterns and project future demand.

D. District-Level Total Demand and Temporal Trends

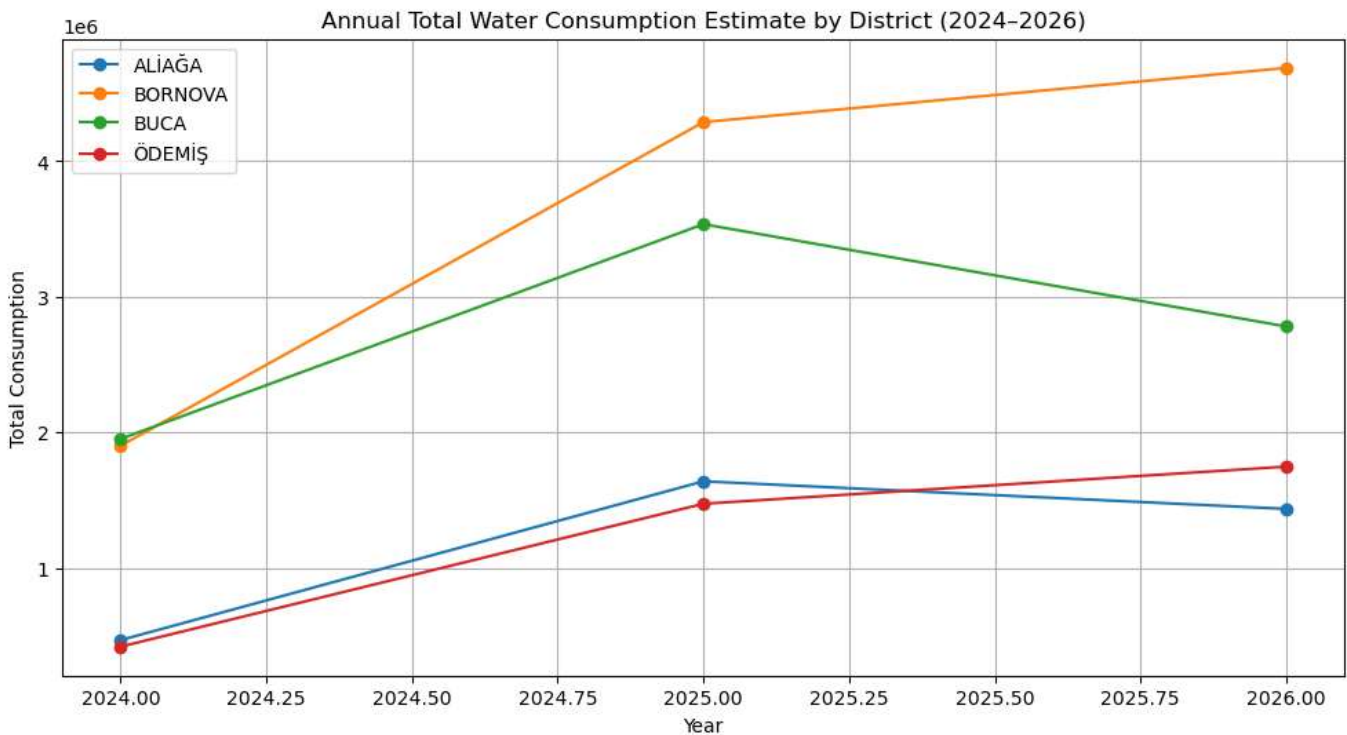


Fig. 4 Projected annual total water consumption for selected districts (2024–2026), illustrating divergent demand trajectories across the study area.

By combining the projected per-subscriber consumption and subscriber counts, we derived monthly and annual total water demand for each district. Figure 4 presents the estimated annual total consumption for 2024, 2025 and 2026 for a subset of representative districts. The results indicate that:

- some districts are expected to show significant increases in total annual demand between 2025 and 2026, driven by both rising subscriber numbers and slightly higher per-subscriber consumption;
- other districts exhibit near-stable total demand, where modest changes in per-subscriber use are offset by limited growth in subscriber counts; and
- a small number of districts may experience stagnating or slightly decreasing total demand, suggesting potential opportunities to consolidate or repurpose existing infrastructure capacity.

Year-on-year percentage changes and absolute differences between 2025 and 2026 highlight those districts where demand pressures are likely to be most acute. These high-growth areas are prime candidates for advanced regenerative and circular solutions, such as decentralised reuse systems, greywater recycling and integrated water–energy–resource recovery schemes.

E. Spatial Patterns in Forecast Errors and Planning Implications

District-level summaries of validation errors provide additional insights for planning. Spatial analysis of log-RMSE values across the study area reveals that districts with relatively low forecast errors can be considered more reliable in terms of modelled demand trajectories, whereas districts with higher errors should be treated with additional caution and may require supplementary scenario analysis or data improvement efforts. When overlaid with the behavioural clusters and projected demand growth rates, these spatial patterns delineate different types of planning contexts:

- High-growth, high-confidence districts, where both demand projections and behavioural profiles are robust, offer strong candidates for piloting net-positive, regenerative interventions at scale.
- High-growth, lower-confidence districts call for parallel efforts to improve data quality and monitoring while beginning to explore regenerative options.

- Stable or declining demand districts may provide opportunities to reconfigure existing infrastructures, enhance ecological performance and test smaller-scale circular solutions.

Overall, the results show that combining behavioural clustering with spatially explicit, machine-learning-based forecasts yields actionable intelligence on where, when and for which types of customers regenerative and circular water system interventions are likely to be most impactful.

IV. DISCUSSION

The results of this study highlight the value of combining behavioural clustering with spatially explicit, machine-learning-based forecasting to support regenerative and net-positive urban water system planning. By segmenting district–customer-type combinations into distinct behavioural clusters and projecting future per-subscriber consumption, subscriber numbers and total demand, we obtain a nuanced picture of where and how water demand pressures are likely to evolve. Figure 5, shows the percentage change in total water consumption by region between 2025 and 2026, highlighting the areas with the highest expected increase or decrease in demand.

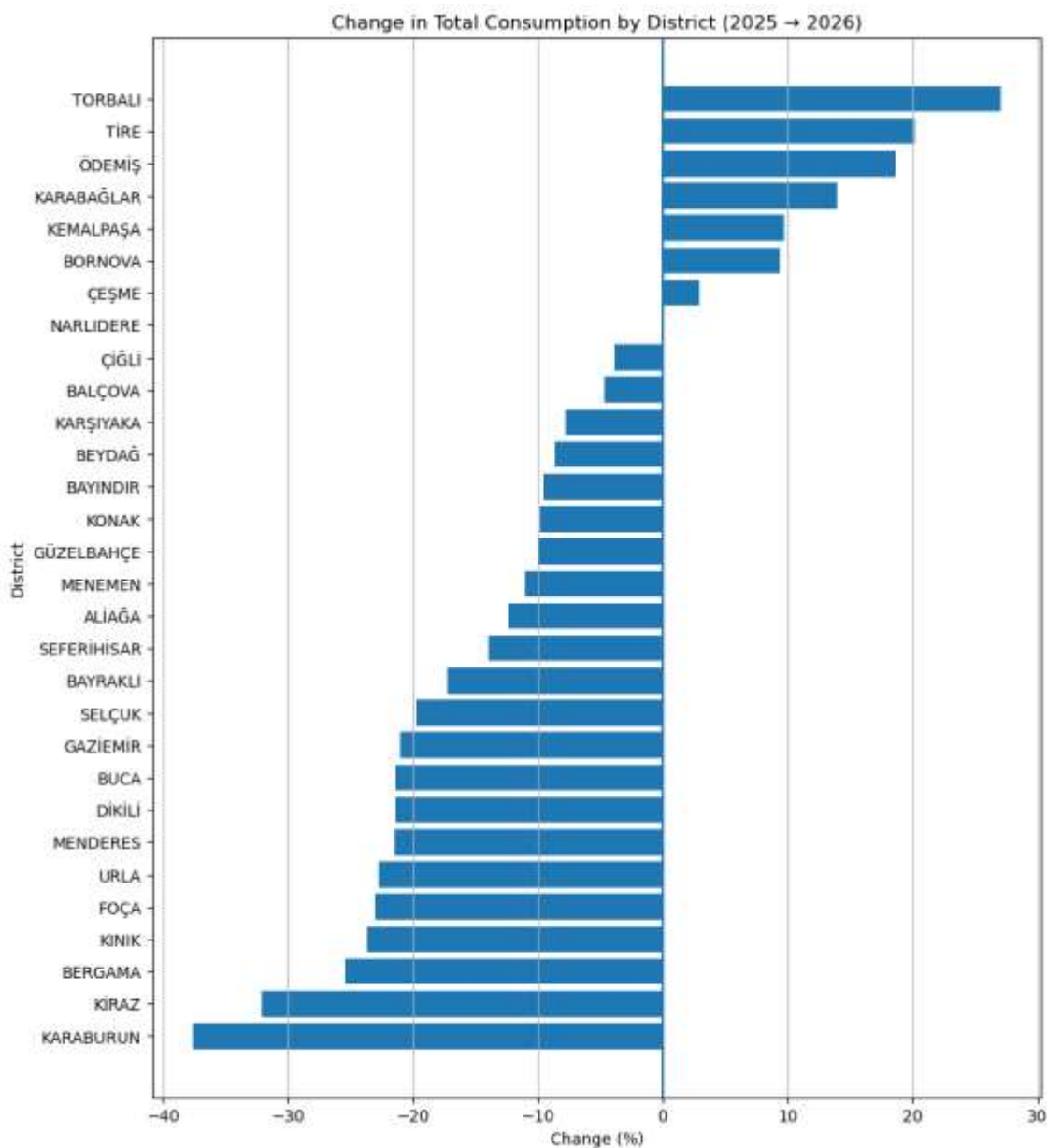


Fig. 5 Percentage change in total water consumption by district between 2025 and 2026, highlighting areas with the highest expected growth or decline in demand.

This level of granularity goes beyond traditional, city-wide averages and provides a more suitable basis for prioritising circular and regenerative interventions. From a regenerative systems perspective, the emergence of high-consumption clusters and high-growth districts points to locations where the potential impacts of water reuse, greywater recycling, rainwater harvesting and integrated water–energy–resource recovery schemes would be particularly significant. In these areas, the projected increases in total demand suggest that simply improving efficiency may not be sufficient to achieve net-positive outcomes; instead, regenerative solutions that actively restore and replenish water resources are needed. Conversely, districts with stable or slightly declining demand may offer opportunities to repurpose existing infrastructure capacity and strengthen ecological performance, for instance by reallocating treatment capacity to support higher levels of reuse or environmental flows. The spatial patterns in forecast errors also have important implications. Districts with relatively low validation errors provide a higher degree of confidence in model-based projections, which can support more decisive planning actions. In contrast, districts with higher errors warrant a more cautious interpretation of the forecasts and may require complementary approaches such as scenario analysis, expert judgement or targeted data improvements. Recognising these differences is crucial to avoid over-reliance on model outputs in contexts where underlying data or system dynamics are particularly uncertain. This work also illustrates both the strengths and limitations of data-driven approaches for supporting regenerative water system design. On the one hand, the use of advanced feature engineering and gradient boosting models enables the capture of complex temporal dependencies and heterogeneity across districts and customer types. On the other hand, the analysis is constrained by the quality and scope of available billing data and does not explicitly account for future changes in climate, tariffs, technological adoption or broader socio-economic drivers. Moreover, while the framework points to where regenerative and circular measures might be most impactful, it does not model the performance of specific technologies or governance arrangements. Future research could extend this framework in several directions. Integrating climate projections, socio-economic scenarios and land-use change into the demand forecasting process would provide a more comprehensive basis for long-term regenerative planning. Coupling demand projections with models of supply, wastewater generation, reuse potential and energy and material recovery could yield fully integrated net-positive system assessments. Finally, participatory approaches involving utilities, municipalities and communities could help align data-driven insights with local priorities, capacities and governance structures, thereby increasing the likelihood that regenerative and circular solutions are both technically robust and socially acceptable.

V. CONCLUSION

This study developed a data-driven framework to inform regenerative and net-positive urban water system planning by combining behavioural clustering of water consumption with spatially explicit demand forecasting. Using multi-year billing records disaggregated by district, customer category, year and month, we constructed temporally consistent time series, engineered seasonality and memory features, and applied an HDBSCAN clustering together with a LightGBM regression model and recursive forecasting. The resulting projections of per-subscriber consumption, subscriber numbers and total demand through 2026 were analysed across districts and customer types.

The main findings can be summarised as follows. First, the behavioural clustering revealed distinct groups of district–customer-type combinations, including high-consumption, strongly seasonal and low-stable clusters, which provide a useful segmentation of water users for targeting regenerative and circular measures. Second, the forecasting models achieved reasonable accuracy on a temporally held-out validation period and captured key temporal patterns in per-subscriber consumption across most districts, while also highlighting areas where data quality or system variability limit predictive performance. Third, the combined projections of per-subscriber use and subscriber numbers indicated that some districts are likely to experience substantial growth in total demand, whereas others may remain stable or even show slight declines, underscoring the importance of considering both intensity and scale when planning future infrastructures.

Taken together, these results demonstrate that integrating behavioural clustering with machine-learning-based demand forecasting can yield actionable intelligence on where, when and for which types

of customers regenerative and circular water system interventions are likely to be most needed and effective. While the framework does not replace detailed technological or governance design, it provides a critical layer of spatially and temporally resolved demand intelligence that can guide the prioritisation, sizing and siting of net-positive solutions. As utilities and cities seek to move beyond conventional sustainability towards genuinely regenerative water systems, such data-driven approaches can play an important role in aligning infrastructure investments and management strategies with evolving patterns of water use and the broader goals of resilience, equity and ecosystem restoration.

REFERENCES

- [1] A. A. Ahmed, S. Sayed, A. Abdoulhalik, S. Moutari and L. Oyedele, “Applications of machine learning to water resources management: A review of present status and future opportunities,” *Journal of Cleaner Production*, vol. 441, 2024.
- [2] L. A. House-Peters and H. Chang, “Urban water demand modeling: Review of concepts, methods, and organizing principles,” *Water Resources Research*, vol. 47, 2011.
- [3] S. L. Zubaidi, S. Ortega-Martorell, H. Al-Bugharbee, I. Olier, K. S. Hashim, S. K. Gharghan, P. Kot and R. Al-Khaddar, “Urban Water Demand Prediction for a City That Suffers from Climate Change and Population Growth: Gauteng Province Case Study,” vol. 12, no. 7, 2020.
- [4] A. Cominola, M. Giuliani, D. Piga, A. Castelletti and A.-E. Rizzoli, “Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review,” *Environmental Modelling & Software*, vol. 72, p. 198–214, 2015.
- [5] A. Candelieri, “Clustering and support vector regression for water demand forecasting and anomaly detection,” *Water*, vol. 9, no. 3, 2017.