https://doi.org/10.71573/frypvh19

Machine-learning forecast model for predicting annual water consumption in budget estimation for urban drainage system management

Michael Trojer [1], *Martin Oberascher* [2] https://orcid.org/0000-0003-3968-4684, *Bernhard Zit* [1] & *Robert Sitzenfrei* [2,*] https://orcid.org/0000-0003-1093-6040

[1]*Innsbrucker Kommunalbetriebe (IKB), Salurner Straße 11, 6020 Innsbruck, Austria*
[2]*Unit of Environmental Engineering, Department of Infrastructure Engineering, Faculty of Engineering Sciences, Universität Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria*
**Corresponding author email: robert.sitzenfrei@uibk.ac.at*

Abstract

Typically, the usable budget for operating the urban drainage network is calculated at the end of the year based on the billed drinking water consumption at customer sites. To estimate the available budget in advance, the network operator uses a simple forecast of annual water consumption, calculated as the average of the past four years. To improve this process, different machine learning based forecasting models were developed with the aim of predicting annual water consumption on a quarterly basis. These models integrate not only historical data but also actual weather conditions and system state measurements with higher temporal resolution. The results showed that Support Vector Machine achieved the highest accuracy across the quarterly forecasting time points, followed by Linear Regression. Consequently, Linear Regression combined with Bayesian statistics was selected as the forecasting model, as it provides the network operator with an uncertainty assessment for the predicted values.

Highlights

- A machine-learning based forecasting model for budget estimation was developed.
- The best results are achieved incorporating data of the last three annual water consumption.
- Bayesian Linear Regression includes an uncertainty assessment of the forecasting value.

Introduction

Urban drainage networks (UDNs) are critical infrastructure systems in urban areas, designed to ensure the environmentally friendly disposal of wastewater into water bodies and the effective management of stormwater runoff. In the case study under investigation, the network operator allocates an operational budget for the UDN based on the billed drinking water consumption at customer sites, the exact value of which is determined in January of the following year. Until that time point, operations rely solely on a forecast of the annual drinking water consumption made at the beginning of each year. The network operator currently uses a simple forecasting model, calculated as the average consumption over the previous four years.

However, in recent years, this approach has often led to significant discrepancies between the forecasted and actual assessed values, with deviations reaching high six-figure euro amounts in the annual operating result. Consequently, economic considerations on the part of the network operator have highlighted the need for an improved forecasting model.

In this context, recent advances in the field of machine learning (ML) offer promising opportunities. ML is increasingly applied in various areas of UDN operations, such as fault detection, prediction of system states, and optimisation of operations (Ahmed et al., 2024; Fu et al., 2022; Huang et al., 2021). Moreover, ML methods are frequently used for forecasting water consumption (Niknam et al., 2022). The aim of this work is to develop an ML-based forecasting model that predicts annual water consumption on a quarterly basis, integrating actual weather data and measurements from within the network.

Methodology

The case study is the UDN of the city of Innsbruck, located in Austria. Innsbruck has approximately 130'000 inhabitants, and the total annual billed water consumption has decreased from 10.3 million m³ in 2003 to 8.2 million m³ in 2023.
The aim of this work is to predict the annual billed water consumption on a quarterly basis using ML methods. The following subsections provide a brief overview of the available datasets and the ML methods that were tested.

Available data sets

The following data sources were incorporated into the analysis:

- Annual total billed water consumption in the case study since 2003
- Daily outflow measurements from the elevation tank of the water distribution network
- Daily inflow measurements to the wastewater treatment plant

In the first step, the annual total billed water consumption was normalised by dividing it by the number of inhabitants per year, thereby eliminating the influence of population growth. This dataset was further enhanced by integrating weather data, such as rain and temperature, from a publicly available platform (https://data.hub.geosphere.at/ with station 39, accessed on 10.01.2025)

These data served as input features for the ML models. Up to six different dataset settings were tested, depending on the forecast point in time (either at the beginning of the year or during the year) as outlined in Table 1.

Table 1. Tested data sets for year n.

Time	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Forecast • 01.01.	• AWD_{n-1}	• AWD_{n-1} • AWD_{n-2}	• AWD_{n-1} • AWD_{n-2} • AWD_{n-3}	• AWD_{n-1} • AWD_{n-2} • AWD_{n-3} • AWD_{n-4}	• AWD_{n-1} • AWD_{n-2} • AWD_{n-3} • AWD_{n-4} • AWD_{n-5}	
Forecast • 01.04. • 01.07. • 01.10.	• AWD_{n-1} • $T_{Forecast}$ • $R_{Forecast}$	• AWD_{n-1} • $T_{Forecast}$ • $R_{Forecast}$ • $ET_{Forecast}$	• AWD_{n-1} • $T_{Forecast}$ • $R_{Forecast}$ • $WWTP_{Forecast}$	• AWD_{n-1} • $T_{Forecast}$ • $R_{Forecast}$ • $ET_{Forecast}$ • $WWTP_{Forecast}$	• $T_{Forecast}$ • $R_{Forecast}$ • $ET_{Forecast}$ • $WWTP_{Forecast}$	• AWD_{n-1} • AWD_{n-2} • AWD_{n-3} • $T_{Forecast}$ • $R_{Forecast}$

AWD_n = annual water consumption per inhabitant for year n, $T_{Forecast}$ = mean temperature from 01.01. until forecast time, $R_{Forecast}$ = rain sum from 01.01. until forecast time, $ET_{Forecast}$ = sum of outflow at elevation tank from 01.01. until forecast time, $WWTP_{Forecast}$ = sum of inflow at wastewater treatment plant from 01.01. until forecast time,

Implementation ML-based forecasting model

In this work, the performance of four standard ML models - Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) - was evaluated. For each model, dataset, and forecast time point, hyperparameters were optimised using a grid search based on parameter ranges defined in Maußner et al. (2024). Due to the limited number of data points (a maximum of 20 years), cross-validation was employed.

The best performing ML model and dataset for each forecast time point were identified using the Mean Squared Error (MSE) as the evaluation metric. The selected ML model was then trained on the complete dataset and used to predict the annual water consumption for the network operator.
All ML models were implemented in Python (version 3.9.13; Python Software Foundation, https://www.python.org/, accessed on 10.01.2025) using the scikit-learn library (Pedregosa et al., 2011).

Results and discussion

ML model selection

The evaluation revealed that SVM consistently delivered the best performance across all four forecasting time points, achieving an average MSE of 0.0156. For the initial forecast time point of January 1st, the most effective input features were the annual water consumption per inhabitant over the last three years. For later forecast time points throughout the year, these input features were further enhanced with precipitation and temperature data. Interestingly, the inclusion of measurement data from the network decreased the model performance across all ML models. This can be attributed to following factors: (1) outflow measurements from the elevation tank also include water losses (e.g., due to leakages) and unmeasured consumption (e.g., for irrigation or firefighting) and thus do not accurately reflect consumption pattern at the customer sites and (2) inflow measurements to the wastewater treatment plant include not only wastewater from the case study but also groundwater infiltration during dry periods and wastewater from neighbouring municipalities, which have a different population growth than the case study itself.
Surprisingly, LR also performed well, achieving an average MSE of 0.0254 across all forecast time points. As a result, LR was selected as the implemented ML model and extended to Bayesian Linear Regression (BLR). BLR provides an uncertainty bandwidth associated with each point prediction, providing decision-makers with additional insights. Moreover, the uncertainty bandwith increases when fewer data points are available, therefore being beneficial useful for small datasets (Li, 2025; Ray, 2019).

ML-based forecasting model

Figure 1 presents an overview of the implemented ML-based forecasting model for the annual water consumption at the forecast time point of October 1st for the year 2023: (a), (b), and (c) display all input features, with those selected by the ML model highlighted in colour; and (d) illustrates the predicted values over the course of the year. As shown, the predicted values deviates by a maximum of only 0.9% from the target value, representing a considerably better forecast than the previous approach based the average value over the past four years. Furthermore, the extension of the LR to BLR enables the inclusion of prediction uncertainties in form of error bars, rather than single point values. This enhancement helps to address annual fluctuations in both water consumption and weather conditions.

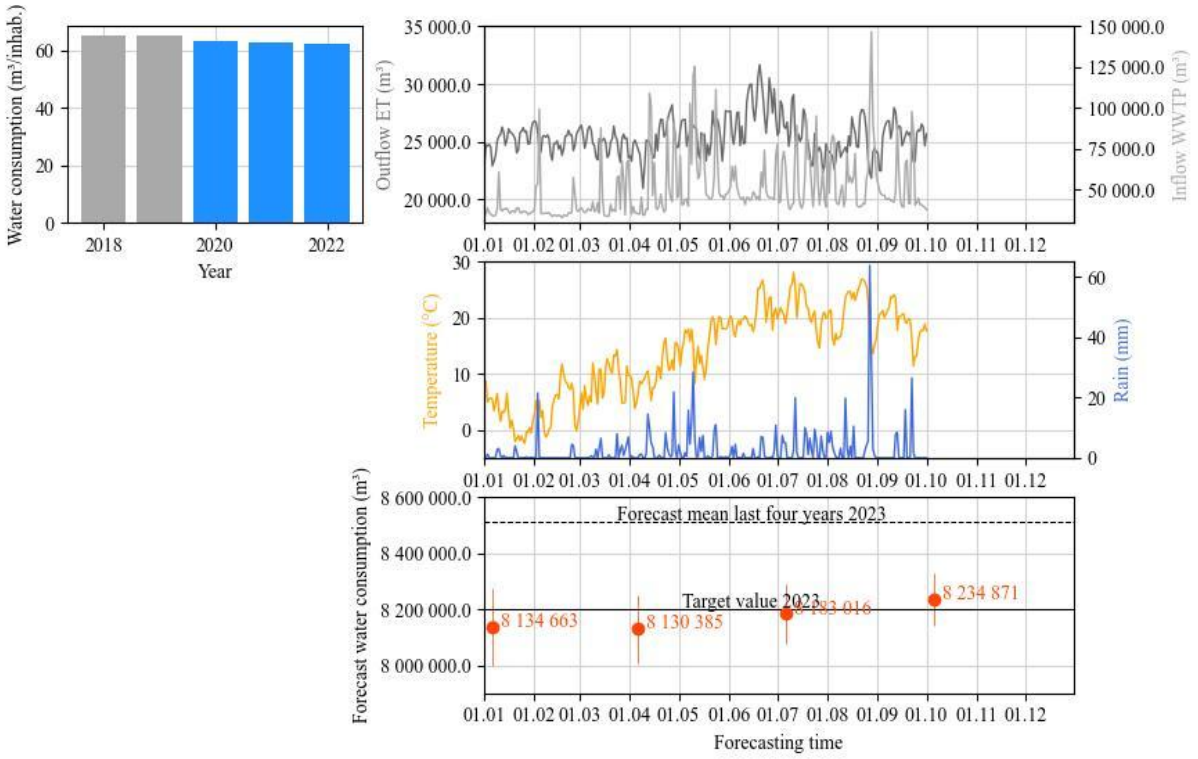

Figure 1. Overview of the developed ML-based forecasting approach, showing the input features (utilised ones are coloured) for (a) historical annual water consumption per inhabitant, (b) measurement data of the networks and (c) weather data as well as (d) prediction values including uncertainty assessment of the annual water consumption for all forecast time points with BLR.

Conclusions and future work

The budget for the urban drainage network in the case study is calculated at the end of each year based on the billed drinking water consumption at customer sites. To support the network operator in planning the activities within the expected budget, a machine learning-based forecasting model was developed. This model predicts the annual total water consumption on a quarterly basis using a range of different input features. The results show that Linear Regression achieved very high accuracy. Moreover, by extending it to Bayesian Linear Regression, the model also provides uncertainty assessments, offering additional information for decision-making.

Future work will focus on the annual evaluation of the forecasting model, including retraining it each year with updated billed drinking water consumption data. It is also planned to test more spatially resolved measurement data (e.g., inflow measurements in individual zones) to further enhance prediction accuracy.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve the readability and language of the work by fine-tuning some of the grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Ahmed, A. A., Sayed, S., Abdoulhalik, A., Moutari, S., and Oyedele, L. (2024). "Applications of machine learning to water resources management: A review of present status and future opportunities." Journal of Cleaner Production. https://doi.org/10.1016/j.jclepro.2024.140715.

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

Huang, R., Ma, C., Ma, J., Huangfu, X., and He, Q. (2021). "Machine learning in natural and engineered water systems." Water Res 205: 117666. https://doi.org/10.1016/j.watres.2021.117666.

Li, C., 2025. Bayesian Methods in Machine Learning Applications and Challenges. Economics and Management Innovation. 2(2), 15-28. https://doi.org/10.71222/j5gxe564.

Maussner, C., Oberascher, M., Autengruber, A., Kahl, A., and Sitzenfrei, R. (2025). "Explainable artificial intelligence for reliable water demand forecasting to increase trust in predictions." Water Res 268(Pt B): 122779. 10.1016/j.watres.2024.122779.

Niknam, A., Zare, H. K., Hosseininasab, H., Mostafaeipour, A., and Herrera, M. (2022). "A Critical Review of Short-Term Water Demand Forecasting Tools—What Method Should I Use?" Sustainability 14(9). https://doi.org/10.3390/su14095412.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). "Scikit-learn: Machine learning in Python." J Mach Learn Res 12: 2825-2830. https://doi.org/10.48550/arXiv.1201.0490.

Ray, S., 2019. A Quick Review of Machine Learning Algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). https://doi.org/10.1109/COMITCon.2019.8862451.