

Predicting Net Inflow for 10 DMAs in North-East Italy [†]

Kristina Arsova ¹, Claudia Quintiliani ^{1,*}, Dennis Schol ¹ and Maaïke Walraad ¹

Brabant Water N.V., 5223 MA 's-Hertogenbosch, The Netherlands; kristina.arsova@brabantwater.nl (K.A.); dennis.schol@brabantwater.nl (D.S.); maaïke.walraad@brabantwater.nl (M.W.)

* Correspondence: claudia.quintiliani@brabantwater.nl

[†] Presented at the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

Abstract: This paper introduces a two-step methodology for short-term water demand forecasting. In the first step, a pre-processing analysis of the inflow input data is conducted to evaluate completeness and quality, ensuring optimal data integrity. Subsequently, in the second step, a robust machine-learning algorithm is employed to predict the water demand patterns. The methodology is applied across 10 District Metering Areas (DMAs) in the north-east of Italy, each characterized by unique demographic features. Accordingly, tailored features are carefully selected for inclusion in the water demand forecast for each DMA.

Keywords: water demand forecast; water shortage; machine learning; Bayesian optimization for parameter tuning; robust model; cross fold validation

1. Introduction

The accurate forecasting of short-term water demand is crucial for the efficient monitoring and operation of Water Distribution Networks (WDNs). This process provides essential support for managing the system's components, including pump settings and valve manipulation, during both normal and crisis situations. There already exists a substantial corpus of research pertaining to short-, medium-, and long-term water demand prediction [1–3]. Two technical challenges need to be overcome to predict residential water demand: choosing (1) the right data and features and (2) the most appropriate modeling technique for high prediction accuracy [4]. The outcomes of the research underscore the absence of a universally effective method applicable to all circumstances for predicting water demand patterns accurately. Additionally, a multitude of factors, such as population size, presence of industries, socio-economic factors (e.g., income), and the deployment of water-saving measures (e.g., rainwater tanks) [5], contribute to the complexity of urban water demand prediction.

This manuscript presents an approach to short-term water demand forecasting as part of the Battle of Water Demand Forecasting (BWDF), for which the whole problem description is provided in [6]. The paper is structured as follows: firstly, a presentation of the method is given, followed by the exposition of the main results pertaining to all the DMAs for a specific forecasted week. Lastly, conclusions are drawn, and future research to improve the proposed method is outlined.

2. Materials and Methods

The challenge presented in the BWDF entails forecasting water demand for an actual case study located in the north-east of Italy. This case study comprises ten DMAs, each showing variations in size (quantified by the number of supplied customers) and average water demand.

The data supplied for the BWDF includes the historical net-inflow time series, spanning from 1 January 2021 to 31 March 2023, at an hourly frequency for each DMA. Additionally, weather data for the same period, along with a calendar detailing national



Citation: Arsova, K.; Quintiliani, C.; Schol, D.; Walraad, M. Predicting Net Inflow for 10 DMAs in North-East Italy. *Eng. Proc.* **2024**, *69*, 178. <https://doi.org/10.3390/engproc2024069178>

Academic Editors: Stefano Alvisi, Marco Franchini, Valentina Marsili and Filippo Mazzoni

Published: 27 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

holidays, is provided. The time series data for the water demand forecast is provided at four intervals, aligning with the requirement for forecasting across four evaluation weeks (from W1 to W4).

The paper's methodology comprises two distinct steps: (1) a pre-processing analysis to evaluate data quality and completeness, and (2) the implementation of a robust machine-learning algorithm, XGBoost, for water demand forecasting. XGBoost was chosen because it is particularly effective at handling large data sets with many features. Moreover, XGBoost has been shown to outperform other popular machine learning algorithms, such as neural networks [7].

2.1. Data Pre-Processing

The objective of the data pre-processing step is to detect and mitigate outliers, identify data gaps, and perform imputation based on trend analysis, including temporal patterns such as hourly variations, day-of-week effects, and seasonal fluctuations.

Data pre-processing consists of two steps. Firstly, missing values are imputed. If $\mu_{h,d,s,i}$ and $\sigma_{h,d,s,i}$ are the empirical mean and standard deviation of all data with the same hour of the day h , day of the week d , and season s in DMA i , respectively, then, if for a certain time t with a certain combination of hour h , day d , and season s , the consumption data are missing, the consumption is taken to be $\mu_{h,d,s,i}$.

Secondly, outliers are detected and taken out using the Z-score approach; if for a datapoint (t,x) in DMA i where the time t has a certain combination of hour h , day d , and season s , and the rescaled variable $|x - \mu_{h,d,s,i}| / \sigma_{h,d,s,i}$ is larger than 3, the datapoint (t,x) is replaced with $(t, \mu_{h,d,s,i})$.

2.2. Forecast Model

In developing the XGBoost model, two distinct functions are used to generate foundational features, such as the day of the month and hour, alongside a separate function dedicated to creating lagged and rolling mean features. These functions are applied to both the training and testing datasets independently to avoid unintentionally transferring information from the test set to the training set.

In contrast to the traditional 70/30 test/train set split, in this research, 10-fold cross-validation is employed to ensure a comprehensive evaluation over extended time intervals. This approach effectively minimizes the risk of seasonal bias, enhancing the model's predictive accuracy across different seasons. Monitoring the mean squared error (MSE) across the cross-validation folds has provided valuable insights into the model's progressive improvement.

For feature selection, the "hyperopt" library of Python is used, which implements a Bayesian optimization technique for hyperparameter tuning. This process is customized to identify the most significant features for each DMA, focusing on minimizing the mean absolute error (MAE). Following hyperparameter tuning, features with an importance score exceeding 10, based on the "gain" importance type, are incorporated into the model. These features, along with the basic features generated by the initial function (e.g., special days such as national and school holidays, local event days, hour of the day, etc.), form the final set of predictors for the XGBoost model. Weather data were intentionally not included in the model because inaccurate weather forecasting quality may introduce additional error in the water demand prediction.

To assess the model's performance, a range of estimators is used, including MAE (24 h) representing the MAE of the initial 24 h, MAE (144 h) signifying the MAE of the subsequent 144 h, and MaxAE (24 h) denoting the maximum absolute error of the first 24 h. Each metric adheres to the indicators outlined in the BWDF guidelines [6].

3. Results

Since real measurements are available only for W1, the performance of the model is showcased in this paper exclusively for that week.

Table 1 shows the results obtained for all the DMAs in the forecasting of W1, summarized in terms of various performance indicators. Analysis of the results table reveals that the features that are chosen exhibit high R^2 values for several DMAs, such as A, D, E, G, H, I, and J, indicating that these features effectively capture a substantial portion of the signal variance. Regarding the mean absolute percentage error (MAPE) scores, DMA J and E demonstrate the best performance, while DMA B and C perform less favorably. Notably, DMA E exhibits the highest MaxAE, whereas DMA F demonstrates the lowest.

Table 1. XGBoost performance indicators per DMA for week W1.

Metric	DMA A	DMA B	DMA C	DMA D	DMA E	DMA F	DMA G	DMA H	DMA I	DMA J
MSE	2.34	6.93	1.65	17.53	25.87	1.14	11.26	4.82	6.38	2.57
R^2	0.82	0.42	0.53	0.87	0.95	0.78	0.82	0.95	0.84	0.88
R^2 (24 h)	0.84	0.75	0.83	0.93	0.96	0.88	0.90	0.98	0.79	0.87
MAE (24 h)	1.40	0.89	0.48	2.62	3.21	0.91	1.22	1.28	2.23	1.45
MAE (144 h)	1.32	2.50	1.09	3.91	4.41	0.81	2.65	2.01	2.15	1.31
MaxAE	4.31	5.94	3.01	8.62	12.13	2.18	8.94	5.49	5.57	3.91
MaxAE(24 h)	2.26	3.31	1.72	6.49	10.58	1.90	4.31	3.25	4.60	3.91
MAPE	12.48	19.65	16.25	11.97	5.36	13.57	8.59	9.44	10.11	4.76
MAPE (24 h)	13.55	7.99	7.97	8.57	3.93	14.58	4.36	5.97	9.96	5.01

Figure 1 shows that the model performs better for the commercial/industrial DMA J compared to the residential DMA B, a trend observed across all DMAs. In general, the model is successful in capturing the seasonality of the time series, but it tends to underperform in predicting peak demands.

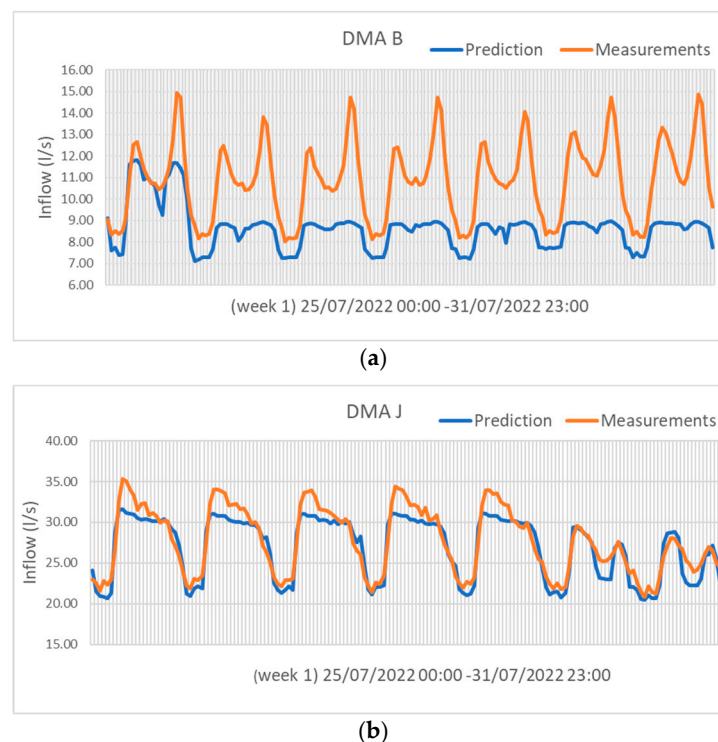


Figure 1. Comparison of predicted and measured inflow time series for evaluation W1 of (a) DMA B and (b) DMA J.

4. Conclusions

This paper introduces a two-step methodology for short-term water demand forecasting, emphasizing the importance of tailored approaches in addressing the complex

dynamics of urban water demand. Tailored features are optimally selected to enhance prediction accuracy across ten distinct DMAs in the north-east of Italy. The results, showcased for the first prediction week, W1, reveal promising performance metrics, with notable variations observed among different DMAs. These findings underscore the efficacy of the proposed methodology while providing valuable insights for future refinement and application. Moving forward, efforts will focus on further enhancing the model's performance through the exploration of a combination of diverse learning algorithms and the potential incorporation of additional features such as high-quality predictions of weather data. Building upon the robustness and generalized approach explained in the paper, the model can be reused for various applications beyond the initial scope, allowing us to deploy the enhanced method within the ongoing development of the Digital Twin of the WDN in the province of North Brabant (The Netherlands), facilitating the monitoring and decision-making processes in day-to-day system operations.

Author Contributions: Conceptualization and methodology, K.A., C.Q., D.S. and M.W.; formal analysis, K.A., D.S. and M.W.; writing—original draft preparation, C.Q. and K.A.; writing—review and editing, D.S. and M.W.; visualization, K.A.; supervision, C.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting this article are openly available on <https://wdsa-ccwi2024.it/battle-of-water-networks> accessed on 23 October 2023 or will be made available by the authors on request.

Acknowledgments: We would like to thank Brabant Water NV for facilitating this research.

Conflicts of Interest: All authors Kristina Arsova, Claudia Quintiliani, Dennis Schol and Maaïke Walraad were employed by the company Brabant Water N.V. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Antunes, A.; Andrade-Campos, A.; Sardinha-Lourenço, A.; Oliveira, M.S. Short-term water demand forecasting using machine learning techniques. *J. Hydroinform.* **2018**, *20*, 1343–1366. [CrossRef]
2. Zubaidi, S.L.; Ortega-Martorell, S.; Kot, P.; Alkhaddar, R.M.; Abdellatif, M.; Gharghan, S.K.; Ahmed, M.S.; Hashim, K. A Method for Predicting Long-Term Municipal Water Demands Under Climate Change. *Water Resour. Manag.* **2020**, *34*, 1265–1279. [CrossRef]
3. Tiwari, M.K.; Adamowski, J.F. Medium-Term Urban Water Demand Forecasting with Limited Data Using an Ensemble Wavelet-Bootstrap Machine-Learning Approach. *J. Water Resour. Plan. Manag.* **2015**, *141*, 04014053. [CrossRef]
4. Lee, D.; Derrible, S. Predicting Residential Water Demand with Machine-Based Statistical Learning. *J. Water Resour. Plan. Manag.* **2019**, *146*, 04019067. [CrossRef]
5. Haque, M.M.; Egodawatta, P.; Rahman, A.; Goonetilleke, A. Assessing the significance of climate and community factors on urban water demand. *Int. J. Sustain. Built Environ.* **2015**, *4*, 222–230. [CrossRef]
6. Alvisi, S.; Franchini, M.; Marsili, V.; Mazzoni, F.; Salomons, E. Battle of Water Demand Forecasting (BWDF) Instructions. 2024. Available online: https://wdsa-ccwi2024.it/wp-content/uploads/2024/06/book_wdsa_ccwi_rev6.pdf (accessed on 19 January 2024).
7. Chen, T.; Guestrin, C. A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.