



Proceeding Paper

Harnessing the Power of Random Forest for Precise Short-Term Water Demand Forecasting in Italian Water Districts [†]

Adam Kulaczkowski ^{1,2,*}  and Juneseok Lee ¹ 

¹ Civil and Environmental Engineering, Manhattan College, Riverdale, NY 10463, USA; juneseok.lee@manhattan.edu

² CDM Smith, Melville, NY 11747, USA

* Correspondence: akulaczkowski01@manhattan.edu

[†] 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: Water demand forecasting is essential for ensuring a reliable water supply in any water utility. It involves making accurate predictions for both short- and long-term water needs. Many traditional time series forecasting methods are presently used; however, recent machine learning techniques have grown in popularity for their robustness and accuracy. Random forest is an emerging machine learning algorithm which was used to forecast short-term water demand for ten district metered areas in Italy. Our predictions on test datasets using the trained model yielded correlations as high as 0.98. Important explanatory variables affecting model performance included consumption patterns represented by the seven-day water demand lag. In this paper, we present a reliable application of the random forest algorithm for short-term water demand forecasting.

Keywords: artificial intelligence; machine learning; random forest; water demand; forecast



Citation: Kulaczkowski, A.; Lee, J. Harnessing the Power of Random Forest for Precise Short-Term Water Demand Forecasting in Italian Water Districts. *Eng. Proc.* **2024**, *69*, 81. <https://doi.org/10.3390/engproc2024069081>

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

Published: 6 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/>).

1. Introduction

Water demand forecasting plays a vital role in the planning, operations, and management of physical assets for water utilities. Optimization of near-future dispatch is one of the most crucial practices of drinking water utilities, which are often resource-limited. We need to make data-informed decisions that consider long-term operation. Traditional time series forecasting methods, such as Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA), have been used for decades to forecast water demand using time series historical data [1]. In the past decade, however, artificial intelligence (AI) has had a rapidly growing presence in the water sector, especially machine learning (ML) techniques. ML techniques have the advantage of being able to forecast nonlinear relationships between response variables in the presence of noisy data. This is especially important in recent times with the increasing use of smart water metering. Since not all data are applicable or valuable, ML models have offered a great advantage over traditional forecasting methods in extracting valuable information for powerful predictive analytics. In this vein, several ML techniques, such as artificial neural networks (ANNs), support vector regression (SVR), and random forest (RF), have been studied extensively to accurately forecast future trends [1].

2. Water Demand Forecasting Using Random Forest

RF emerges as a compelling machine learning algorithm for water demand forecasting due to its ability to capture complex, non-linear relationships between independent and dependent variables and to handle noisy data [1]. RF provides a valuable measure of feature importance, shedding light on which variables exert the most influence on predicting water demand [1]. This information aids in understanding the underlying factors driving water demand and informs data-driven decision-making processes.

Several studies have underscored the efficacy of RF in water demand forecasting. Tyrallis et al. (2019) showcased RF's versatility across applications, including demand forecasting [2]. Researchers have highlighted the comparative advantage of RF in high-temporal-resolution and short-term water demand forecasting [3].

Comparative analyses have often shown RF to outperform alternative ML algorithms. Villarin et al. (2019) demonstrated RF's superior predictive capacity compared to other ML models [4]. Xenochristou et al. (2018) illustrated RF's potential in short-term demand forecasting, considering a multitude of factors including consumption patterns, household characteristics, socio-economic indicators, and climatic variables [5]. Smol A.K. et al. (2020) reported RF's superior accuracy in water use prediction compared to classical time series-based methods and other ML algorithms [6]. IwA.K.in and Moazeni (2023) highlighted RF as a reliable model for near-real-time forecasting, essential for water resource management decisions [7]. However, De Souza Groppo et al. (2019) emphasized the necessity of evaluating model performance case by case, recognizing that no single model universally excels in all scenarios [8].

In alignment with the findings of these studies, RF emerges as a robust and effective algorithm for water demand forecasting, often surpassing alternative ML approaches in terms of accuracy and efficiency. The collective evidence underscores the potential of ML, particularly RF, in addressing challenges in short-term water demand forecasting, affirming the rationale for its utilization in this study.

3. Materials and Methods

This study sought to apply the RF ML model for short-term prediction of urban water demand for ten (10) district metered areas (DMAs), part of a Water Distribution Network (WDN) in the northeast of Italy. These DMAs varied considerably in size, average water demand, and usage characteristics [9].

Various factors potentially influence demand data at each DMA, and employing hourly net in-flow and weather data offered a detailed perspective. Data processing involved the incorporation of additional explanatory variables, specifically the 7-, 14-, and 21-day lag, along with the monthly average hourly demand, aimed at addressing hypothetically recurring consumption patterns [1,5]. Lastly, a binary variable was added to indicate whether the demand fell on a calendar holiday. Each DMA was broken into a separate dataset with the above-mentioned explanatory variables.

The full inflow dataset utilized in this study comprised data from January 1st, 2021 to March 5th, 2023. With these original datasets, there were numerous missing values present in the inflow data, which could have resulted from data collection malfunction or other collection/transmission issues [9]. To enhance the reliability of predictions, RF for nonparametric missing value imputations (missForest) was employed ahead of the RF forecast model [10]. The algorithm utilizes RF, trained on the observed values within the dataset, to predict and fill in the missing values. Employing this iterative algorithm allowed the entire dataset to be used rather than the deletion of rows where demand data were not present, preserving on average 8% of the data. Iterations were performed for each DMA until the stopping criterion was met [10]. Following the successful application of missForest, the completed dataset was used to train the RF model.

4. Results and Discussion

After the submission of the competition's week 4 (W4) forecast for the week of 6 March–12 March 2023, the previous week was used as the forecast period and a test dataset to evaluate model performance. The training dataset was compiled until the final week. Table 1 highlights metrics and statistics derived from the trial DMA forecasts.

Figure 1 was generated for each DMA to compare the predicted demand with the recorded demand, facilitating the identification of scenarios where the training dataset led to underpredictions or overpredictions in the test dataset. As shown for DMA I (with an

R^2 decrease from 0.96 to 0.74 between the training and test datasets), the model slightly overpredicted then underpredicted for recorded demands below and above 30 L/s.

Table 1. RF training and testing dataset results for each DMA.

Metric	DMA A ¹	B	C	D	E	F	G	H	I	J
Var. Explained (%)	80.87	87.48	87.75	85.46	97.21	77	93.93	94.81	81.23	86.53
R^2	0.96 (0.74)	0.98 (0.80)	0.97 (0.95)	0.97 (0.84)	0.99 (0.98)	0.95 (0.58)	0.99 (0.95)	0.99 (0.97)	0.96 (0.74)	0.97 (0.79)
RMSE	0.57 (1.11)	0.27 (0.44)	0.23 (0.2)	1.24 (3.15)	1.12 (1.72)	0.43 (1.07)	0.62 (0.96)	0.65 (1.2)	0.69 (1.48)	0.73 (1.56)
MAPE	0.05 (0.16)	0.02 (0.04)	0.03 (0.05)	0.03 (0.08)	0.01 (0.01)	0.04 (0.07)	0.02 (0.03)	0.02 (0.03)	0.02 (0.05)	0.02 (0.06)

¹ First value applies to the training dataset; value in parenthesis is for testing dataset.

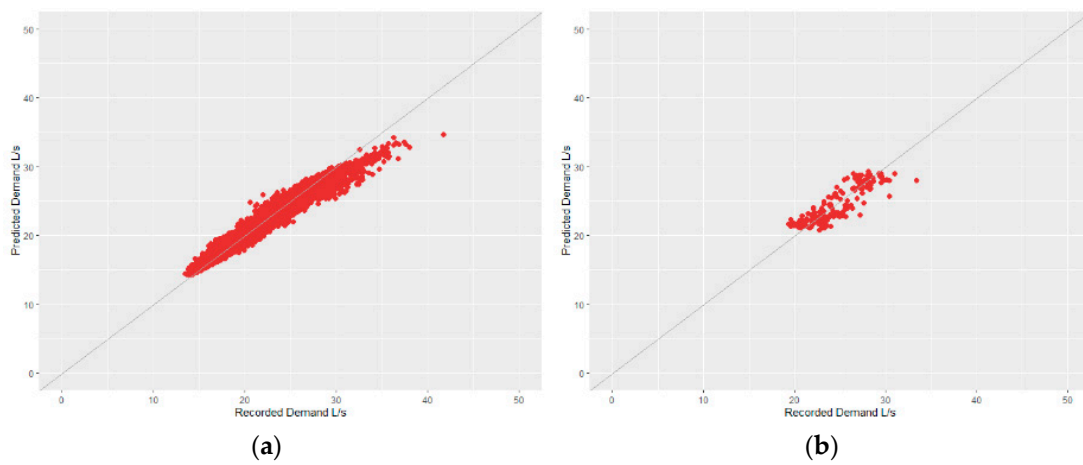


Figure 1. Predicted vs. recorded demand for DMA I (in L/s). (a) Training results, (b) testing results.

For all the DMAs forecasted, training dataset R^2 values ranged from 0.95 to 0.99, with test datasets having R^2 values from 0.58 to 0.98. Low R^2 values may have been due to DMA-specific explanatory variables that were not adequately accounted for in the model input. Feature importance figures were created for each DMA to highlight explanatory variables that contributed to model prediction. An example figure for DMA J is presented in Figure 2.

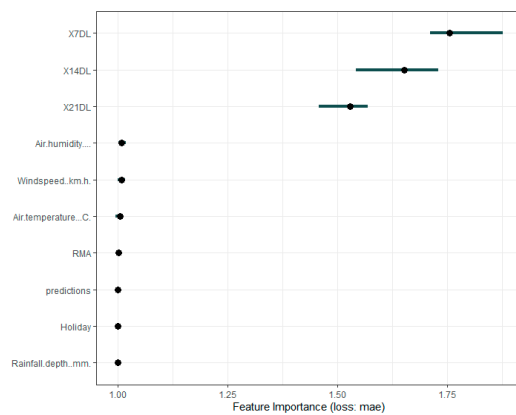


Figure 2. Feature importance for DMA J (commercial/industrial district close to the port).

For all DMAs, the 7-day demand lag proved to be the most important explanatory variable in model training. Subsequent explanatory variables often included the 14- and 21-day demand lag, and in some cases, the air temperature.

5. Conclusions

The application of a robust tool for short- and long-term water forecasting has continued to grow in importance for water utility planning, operations, and management of precious water resources. RF has been cited as a very versatile and accurate ML algorithm for water demand prediction, with capabilities of iterative improvement to offer powerful predictive analytic capabilities. In its application for water demand forecasting across ten DMAs, our model provided consistently reliable predictions and insight into the significance of each explanatory variable. This information can enhance the model's forecasting performance and guide utilities on which variables are crucial to monitor. As an established algorithm with relatively simple implementation, RF offers comprehensive capabilities in addressing challenges for short-term water demand forecasting.

Author Contributions: Conceptualization, A.K. and J.L.; methodology, A.K. and J.L.; software, A.K.; validation, A.K. and J.L.; formal analysis, A.K.; investigation, A.K.; resources, J.L.; data curation, A.K.; writing—original draft preparation, A.K.; writing—review and editing, A.K. and J.L.; visualization, A.K.; supervision, J.L.; project administration, J.L. 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 to study.

Informed Consent Statement: Not applicable to study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Acknowledgments: The authors would like to acknowledge the research resource support provided by Manhattan College.

Conflicts of Interest: Author Adam Kulaczkowski has been involved in the company CDM Smith. The remaining 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. Lee, J.; Keck, J. *Embracing Analytics in the Drinking Water Industry*; IWA Publishing: London, UK, 2022.
2. Tyrallis, H.; Papacharalampous, G.; Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **2019**, *11*, 910. [CrossRef]
3. Liu, G.; Savic, D.; Fu, G. Short-term water demand forecasting using data-centric machine learning approaches. *J. Hydroinform.* **2023**, *25*, 895–911. [CrossRef]
4. Villarin, M.C.; Rodriguez-Galiano, V.F. Machine learning for modeling water demand. *J. Water Resour. Plan. Manag.* **2019**, *145*, 04019017. [CrossRef]
5. Xenochristou, M.; Kapelan, Z.; Hutton, C.; Hofman, J. Smart water demand forecasting: Learning from the data. *EPiC Ser. Eng* **2018**, *3*, 2351–2358.
6. Smolak, K.; Kasieczka, B.; Fialkiewicz, W.; Rohm, W.; Siła-Nowicka, K.; Kopańczyk, K. Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models. *Urban Water J.* **2020**, *17*, 32–42. [CrossRef]
7. Iwakin, O.M.; Moazeni, F. Short-Term Water Demand Prediction Using Machine Learning Techniques—A Case Study of Telford Borough in Pennsylvania. In *World Environmental and Water Resources Congress*; American Society of Civil Engineers: Reston, VA, USA, 2023; pp. 1027–1036.
8. de Souza Groppo, G.; Costa, M.A.; Libânio, M. Predicting water demand: A review of the methods employed and future possibilities. *Water Suppl.* **2019**, *19*, 2179–2198. [CrossRef]
9. WDSA-CCWI. Battle of Water Networks—Battle of Water Demand Forecasting Instructions. Available online: https://wdsa-ccwi2024.it/wp-content/uploads/2023/09/BWDF_Instructions_2023_09_15.pdf (accessed on 22 March 2024).
10. Stekhoven, D.J.; Stekhoven, M.D.J. missForest: Nonparametric missing value imputation using random forest. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef] [PubMed]

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.