IST-Africa 2024 Conference Proceedings
Miriam Cunningham and Paul Cunningham (Eds)
IST-Africa Institute and IIMC, 2024
ISBN: 978-1-905824-73-1

Ensemble Deep and Machine Learning for Improving Short-Term Water Demand Forecast in Cities

Paul TOTO[1], Michael KIMWELE[2], Richard RIMIRU[3]
Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000, Nairobi, 00200, Kenya
[1]Tel: +254720456694, Email: paultoto90@yahoo.co.uk
[2] Tel: +254721611436, Email: mkimwele@jkuat.ac.ke
[3] Tel: +254729110513, Email: rimiru@jkuat.ac.ke

Abstract: Intelligent water distribution systems are essential to ensure water availability for all city users amidst increasing water scarcity driven by urbanization and climate change. In this study, our goal was to enhance the forecasting accuracy of a water utility using a Bayesian Moving Averaging (BMA) ensemble model. This model combines Random Forest (RF), Extreme Gradient Boosting (XGB), and Multilayer Perceptron (MLP) algorithms. We utilized daily monitoring data from the water utility spanning from 2018 to 2023 to predict short-term water demand, incorporating weather features. Our findings demonstrate that the proposed ensemble model achieved the best performance, yielding the lowest mean absolute percentage error (MAPE) of 15.99% and the highest R squared (R^2) value of 0.98 on the testing set. Additionally, the RF model exhibited better results compared to the other single models. These outcomes underscore the potential of the ensemble approach in short-term water demand forecasting, surpassing more traditional and state-of-the-art methods.

Keywords: Water demand prediction, machine learning, deep learning, smart cities, ensemble model, IoT.

1. Introduction

The development of intelligent urban water supply systems is imperative for the advancement of modern smart cities, especially in light of the looming threat of water scarcity driven by rapid urbanization and climate change. The United Nations Environmental Program has underscored that the global water supply and demand gap is projected to widen to 40% by 2030 [1]. Consequently, effective water demand forecasting emerges as a critical component in providing water utilities with essential data support for urban water resource management, ensuring equitable distribution of urban water resources across time and space.

Accurate short-term water demand forecasting assumes paramount importance as it empowers waterworks to proactively detect potential leakages and minimize energy consumption costs associated with water pump operations, all while ensuring a reliable and sufficient water supply to meet the needs of burgeoning cities. Therefore, the imperative for intelligent water demand forecasting emerges as a central tenet in the future vision of smart cities, facilitated by advancements in the Internet of Things (IoT), big data analytics, and artificial intelligence (AI) to precisely forecast water consumption [2]. This introduction sets the stage for the significance of the study, emphasizing the pressing need for intelligent water demand forecasting in the face of escalating water scarcity challenges. By leveraging cutting-edge technologies and methodologies, such as IoT, big data, and AI, water utilities

can enhance their capacity to forecast water consumption accurately, thereby bolstering urban water resource management and sustainability efforts in the era of smart cities.

The development of urban water demand forecasting models presents complex challenges due to the inherently nonlinear and stochastic nature of water resources. These challenges are exacerbated by the impacts of climate change, including altered precipitation patterns, as well as rapid population growth and urbanization, which intensify the demand for urban water resources [3]. Initially, traditional statistical models such as time-series analysis and regression were employed for water demand forecasting. However, these models often lack the nonlinear capability required to accurately forecast water demand, leading to uncertainties in the operation and management of water supply systems [4]. Moreover, traditional water management methods tend to be time-consuming, expensive, and resistant to investment in cutting-edge technologies and infrastructure upgrades.

Recent advances in computing methods, particularly wireless sensor networks and smart monitoring systems using Internet-of-Things (IoT) technology, have significantly enhanced the availability of big sensor data. These technologies have revolutionized data collection, processing, and analytics, providing scientists with vast amounts of real-time data for analysis. Artificial Intelligence (AI) has played a pivotal role in leveraging these advancements in computing methods. AI offers efficient methods for accurate data collection, advanced processing, and analytics, enabling scientists to extract valuable insights from large and complex datasets [5].

Machine Learning (ML) is a key AI technique that enables computers to learn from data and make predictions. ML models have been extensively employed in various studies for short-term water demand forecasting. These models are capable of predicting complex phenomena, handling enormous datasets of varying sizes, and are resilient to missing data [6]. Deep Learning (DL) has emerged as a powerful AI technique for deciphering complex patterns and making accurate predictions from massive datasets. Therefore, DL models, such as neural networks, excel at learning intricate patterns and relationships within data, making them well-suited for short-term water demand forecasting.

Sahour et al. [7] utilized ML algorithms such as Random Forest (RF) and Extreme Gradient Boosting Machine (XGB) models to forecast water demands of a city. They standardized the range of features in the datasets and compared the performance of these models. RF is implemented using a bagging technique, where the final forecast is the average of all decision trees in the forest. On the other hand, XGB uses the error residuals from previous decision tree models to fit subsequent models. The final forecast in XGB is a weighted sum of all the tree forecasts, with more weight given to models that perform better on the training data.

Vijai and Sivakumar [8] employed a variety of machine learning models, including Artificial Neural Network (ANN), Deep Neural Network (DNN), RF, and Least Square Support Vector Machine (LSSVM), to predict water demand at different time intervals (1, 12, and 24 hours). Their results indicated that the ANN model performed the best across these time intervals. Overall, these studies demonstrate the effectiveness of machine learning algorithms in forecasting water demand. While RF and XGB are powerful ensemble methods, ANN, particularly deep neural networks, have shown promising results in capturing complex patterns and making accurate predictions in water demand forecasting tasks. In addition to traditional machine learning models, DL architectures such as multilayer perceptron (MLP) neural networks have gained popularity and demonstrated superiority in multivariate time series prediction tasks. MLP neural networks have the capability to combine various data sources, including meteorological data, remote sensing data, and historical water demand data, thereby improving forecast accuracy [9].

The application of IoT, ML and DL models in water resource management offers significant benefits, including nonlinear capability and real-time monitoring and analysis of

water resources. These techniques enable data-driven decision-making and optimization of water allocation and demand forecasting processes. By leveraging these techniques, water companies can predict water quality, allocate water resources based on real-time data, mitigate water losses in the distribution system, and identify consumption trends and potential risks in water resource management [10]. Despite these advantages, existing water demand forecasting models have encountered certain obstacles, which our proposed model aims to address while offering additional benefits. One such limitation is the challenge of dealing with missing meteorological data, which is essential for accurate forecasting.

Various ML and DL techniques, including ensemble models, have been developed to predict water demand in the absence of complete meteorological data. However, the complex structures and architectures of many of these models can pose challenges in water demand prediction. Moreover, accessing government data, particularly in African countries, where data from water companies may be incomplete or stored manually, has been a significant obstacle [11]. This limitation hampers the development and implementation of accurate forecasting models. Furthermore, previous studies have primarily relied on data sources proximate to the water utilities' target areas as input for water demand prediction. While this approach provides localized insights, it may overlook broader trends and patterns in water consumption behaviour.

To address these challenges, our proposed model aimed to develop robust ML and DL techniques that can effectively handle missing data, utilize diverse data sources, and provide accurate water demand forecasts. By overcoming these limitations, our model sought to contribute to improved water resource management and sustainability efforts, particularly in regions facing data accessibility challenges like African countries [11]. Our study proposes a heterogeneous Bayesian Model Average (BMA) ensemble model that integrates IoT, ML and DL techniques. By combining various ML and DL algorithms, our heterogeneous ensemble model aims to enhance forecasting accuracy and mitigate the impact of model uncertainty. This integration aims to leverage real-time data acquisition, processing, and analysis capabilities to further improve forecast accuracy and facilitate the achievement of Sustainable Development Goals (SDGs) in the future [12].

As part of our study, we implemented three individual models, namely Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGB), and Random Forest (RF), using the BMA method. These models were specifically chosen for their effectiveness in handling complex data patterns and their ability to provide accurate forecasts. By utilizing the BMA ensemble method with these models, we aim to provide a daily time-scale urban water demand forecast for the Nairobi County Water and Sewage Company (NCWSC) with higher accuracy, thereby surpassing the capabilities of traditional forecasting models.

2. Objectives

The main objective of this research was to provide water utility managers with effective tools for predicting short-term water demand to improve water resource planning and management in cities. To achieve this objective, we developed a BMA ensemble model that combines the outputs of the two conventional ML (XGB and RF) and DL time series (MLP) algorithms. By selecting the computationally efficient features, the BMA ensemble model could achieve better regression accuracy with a minimum error. A comparison of the performance of the ensemble water demand forecast with that of the standalone models would provide valuable insights for sustainable water resource management in the study area and beyond.

3. Methodology

3.1 Data

We obtained the daily water consumption data related to smart water flow meters from the database of Nairobi County Water and Sewage Company (NCWSC) with a one-hour interval. The water distribution smart meters are IoT devices that were used by the NCWSC to build a sustainable and advanced consumption data system. The water supply network consists of more than 20 metering stations equipped with wireless connection between the water flow meters using global system for mobile communications (GSM) and general packet radio services (GPRS) protocols. This paper considers the average daily precipitation and temperature as major variables in the consumption of water. A total of 120 recorded values for each variable was used as input data. Data from January 1, 2018, to December 31, 2022, were used for training, while data from January 1, 2023, to December 31, 2023, were used for testing.

4. Technology Description

The framework adopted in this study encompasses the implementation of single-based deep and machine learning models (XGB, RF, and MLP) and the Bayesian Model Average (BMA) ensemble method for forecasting short-term water demands of a city water supply system. The framework as shown in Figure 1 also includes the evaluation of their accuracies and error measurements.

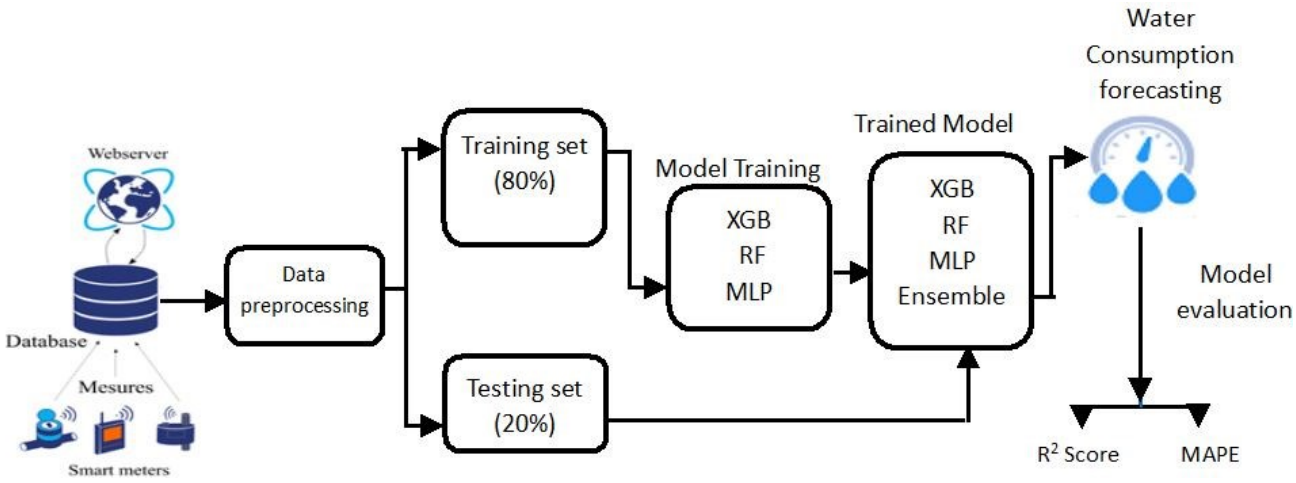

Figure 1: The structure of the proposed model

4.1 Modelling Steps in the Proposed System

1. Data Acquisition: Water demand data and weather information were obtained from the daily monitoring service of the water supply plant and the meteorological department, respectively. We incorporated historical temperature, precipitation, and water demand data within target area.

2. Data Pre-processing: The collected data underwent pre-processing to address missing values and outliers. Missing observations, such as those due to hourly leakages or temporary breaks in data recording, were removed from the water demand time series. Input data, including historic water demand data, precipitation, temperature, and day-of-year, were normalized to fall between 0 and 1. Normalization improves the convergence rate of the model and reduces the influence of the absolute scale.

3. Feature Selection: Pertinent features, such as time of day, day of the week, and seasonality, were extracted from the pre-processed data to enhance model accuracy.

4. Data Splitting: The pre-processed and feature-engineered data were divided into single model training and test sets at an 80:20 ratio. The single model test set was further divided into the BMA ensemble model training and test sets.

5. Model Training: The input data were transformed into sequential data and converted into a supervised learning format. Three deep and machine learning models (MLP, RF, and XGB) were defined and fitted to the training data. The predictions of the three models were then ensembled to generate ensemble predictions using the BMA method.

6. Evaluation Metrics: The accuracy of each model in the training/testing stage was evaluated by calculating the difference between the predicted and real values in the datasets. In this study, two sets of statistical measures were employed to investigate the accuracy of the predictive model, including mean absolute percent error (MAPE), and coefficient of determination (R^2). The best model was elected based on the lowest MAPE, and highest R^2. The details of these indices are expressed below:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|yi - \hat{y}|}{yi} \times 100\% \quad \text{----------(1)} \qquad R^2 = 1 - \frac{\sum_{i=1}^{N}(yi - \hat{y})^2}{\sum_{I=1}^{N}(Yi - \bar{y})^2} \quad \text{----------(2)}$$

where y_i and $y^{\wedge}$ represent the actual and predicted values respectively, and N represents the total number of observations.

By following these modelling steps, the study aimed to develop accurate and reliable models for short-term water demand forecasting, contributing to improved water resource management and efficiency in the city water supply system.

5. Model development

In this study, three single-based machine learning models (Extreme Gradient Boosting, Random Forest, and Multi-layer Perceptron) and a Bayesian Model Average (BMA) ensemble model were developed using the Scikit-learn library in Python. Each model was trained using historical water demand data and corresponding weather variables.

1. Random Forest (RF): RF is an ensemble learning method that utilizes multiple decision trees to address classification and regression problems. It aggregates the predictions of individual decision trees to produce robust and accurate forecasts. RF employs a bagging algorithm, which leverages the strength of individual decision trees and combines their outputs to generate reliable predictions [13].

2. Extreme Gradient Boosting (XGB): XGB is an ensemble technique that combines weak predictors to create a strong predictor by correcting the errors of other trees within the model. It supports scalability, cache optimization, and handles missing data efficiently. XGB is known for its high performance in various prediction tasks [14].

3. Multi-layer Perceptron Neural Network (MLP): MLP is a type of deep learning neural network inspired by the human neural system. Data pass through input layers, hidden layers, and output layers, allowing for complex pattern recognition and prediction. MLP has been widely applied in hydrological predictions and water resources management due to its capability to capture nonlinear relationships in data [15].

4. Bayesian Model Average (BMA) Method: BMA is a statistical method that is employed to calculate the parameters of the model and generate forecasts. BMA combines forecasts from variety of models (MLP, RF, and XGB) by weighing them in accordance with their posterior probabilities and offers an improved comprehension of the overall forecasting uncertainty. The probability density function (PDF) of the model can be expressed as [16]:

$$p(y|D) = \sum_{k=1}^{k} p(fk|D) \cdot pk(y|fk, D) \qquad (3)$$

where $p(y|D)$ is the posterior probability of the predicted sequence f_k, which reflects the degree of coincidence between f_k and the observed water demand volume.

6. Results

In this study, the ensemble model combined the prediction results of single-based models (XGB, RF, and MLP) to enhance time series prediction accuracy. Model performance was evaluated using MAPE and adjusted R^2 values, detailed in Table 1 and Figure 2, respectively.

Table 1: Results of the models for Short-term Water Demand Forecasting

Data	Training set		Testing set	
Evaluation metrics	**MAPE (%)**	**R^2**	**MAPE (%)**	**R^2**
XGB	**23.29**	**0.89**	**22.25**	**0.88**
R F	**21.01**	**0.96**	**17.89**	**0.89**
MLP	**18.07**	**0.98**	**30.14**	**0.97**
BMA	**18.05**	**0.99**	**15.99**	**0.98**

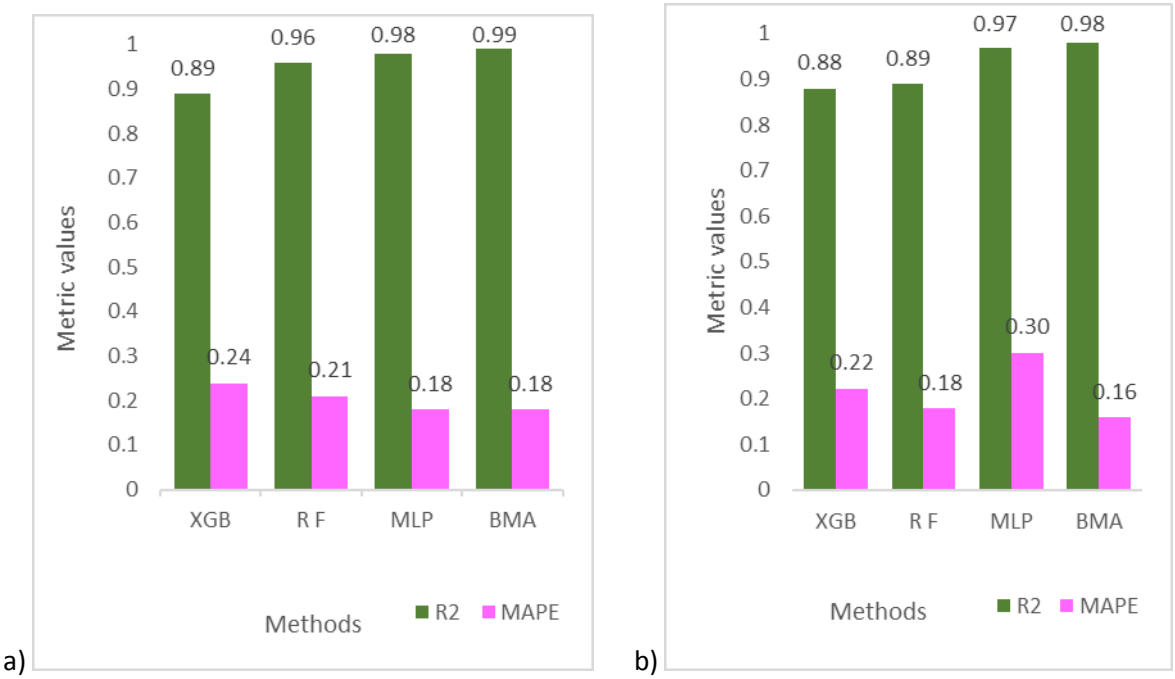

Figure 2: Comparison of performance of XGB, RF, MLP and BMA ensemble models using a) training set and b) test set for a 1-day ahead forecast.

Here is the summary of our findings:

1. Single-based Learning Models: The MLP achieved the highest R^2 values of 0.98 for the training set and 0.97 for the testing set. It had the lowest MAPE value of 18.07% for the training set and 30.14% for the testing set. This model also outperformed the other two machine learning models (RF and XGB) in terms of R^2 values but had slightly higher MAPE on the testing set. The RF produced a low MAPE value of 17.89% on the testing set, which was slightly lower than MLP. It had a high R^2 value of 0.96 for the training set and 0.89 for the testing set. The XGB performed poorly in terms of R^2 values on both training and testing sets. It however provided the fastest training time but had inferior predictive performance compared to MLP and RF.

2. BMA (Bayesian Model Averaging) Ensemble Model: BMA ensemble model outperformed all other models, achieving the lowest MAPE values and the highest R^2 value on the entire dataset during training and validation phases. It was considered the best model overall based on the given information.

Based on our analysis, the BMA ensemble model seems to be the most suitable for predicting water demand one day ahead, followed by the Random Forest model. While the

MLP model had the highest R^2 values, its MAPE on the testing set was higher than that of the RF model. It's essential to consider both accuracy metrics and computational efficiency when selecting the final model for deployment.

7. Business Benefits, Drawbacks and Recommendations

The integration of artificial intelligence (AI) techniques, particularly deep learning (DL) and machine learning (ML), presents significant opportunities for improving the accuracy and utility of models used to predict short-term water demand. Unlike traditional statistical methods, which can be cumbersome and time-consuming, DL and ML models can generate reliable results quickly, even with limited data. These techniques have already demonstrated their utility in various hydrogeological studies, including water quality monitoring, flood mapping, and groundwater potential zoning.

In our study, we evaluated the effectiveness of DL and ML models such as extreme gradient boosting (XGB), random forest (RF), and multi-layer perceptron (MLP) for predicting short-term water demand. While each model individually exhibited reliability, combining them into an ensemble model using BMA significantly enhanced predictive accuracy. This improvement facilitates better integration and visualization of data, enabling more accurate forecasting of trends and identification of hidden relationships. Additionally, the utilization of smart water flow meters based on IoT technology offers tangible benefits by reducing operational costs through decreased manual checks and enabling real-time monitoring of water volume. These advancements ultimately lead to more efficient resource allocation and reduced operational costs for water utility companies.

Despite the promising benefits, our study also identified several limitations that need to be addressed. Factors such as irregular water use patterns, the impact of the COVID-19 pandemic on consumption behavior, and differences between usage types pose challenges to prediction accuracy. These factors can result in inaccuracies in short-term demand forecasts, impacting resource management decisions. Moreover, MLP deep learning models require significant computational resources and lengthy training processes, especially when dealing with large datasets or multiple variables. This can hinder scalability and performance, particularly in scenarios requiring real-time operation.

To overcome these challenges and further enhance predictive accuracy, future research should explore a wider range of forecasting models and ensemble techniques. Incorporating data from multiple water utilities can improve the generalizability of prediction models and enhance their accuracy across different geographical regions and usage patterns. Additionally, further refinement of ensemble models is recommended to optimize hyperparameters and address specific scenarios like irregular water use patterns and external factors impacting consumption behavior.

8. Conclusions

In conclusion, our study aimed to develop an efficient Bayesian model averaging (BMA) ensemble model to predict short-term water demand in urban water utilities. By integrating extreme gradient boosting (XGB), random forest (RF), and multi-layer perceptron (MLP) models, we sought to enhance predictive accuracy. Leveraging daily water consumption data and weather variables, our model underwent rigorous data pre-processing techniques. Evaluation using mean absolute percentage error (MAPE) and R-squared (R^2) indicated the superior predictive power of the BMA model, followed by RF, MLP, and XGB.

The integration of IoT smart flow meters with machine learning and deep learning techniques presents a more effective approach to monitoring water demand changes and identifying inefficiencies in water infrastructure. Empowering water utility managers to mitigate energy and operational costs associated with managing water supply systems, this approach enhances overall efficiency. Our ongoing research recommends a phased

approach for implementing the ensemble framework, beginning with off-line testing before transitioning to real-time operation. Additionally, extending similar dashboards and AI integration to other sectors such as energy and agriculture promises to further augment the benefits of AI beyond water management, contributing to broader sustainability endeavors.

Declaration of Use of Content generated by Artificial Intelligence (AI) (including but not limited to Generative-AI) in the paper

The authors acknowledge the use of ChatGPT for editing and grammar enhancement in the paper entitled "***Ensemble Deep and Machine Learning for Improving Short-Term Water Demand Forecast in Cities***".

References

[1] UNEP, "Options for Decoupling Economic Growth from Water use and Water Pollution: Report of the International Resource Panel Working Group on Sustainable Water Management," 2016.

[2] J. Stańczyk, J. Kajewska-Szkudlarek, P. Lipiński and P. Rychlikowski, "Improving short-term water demand forecasting using evolutionary algorithms," *Scientific Reports*, vol. 12, p. 13522, 2022, doi: https://doi.org/10.1038/s41598-022-17177-0

[3] A. D. Martinho, H. S. Hippert and L. Goliatt, "Short-term streamflow modeling using data-intelligence evolutionary machine learning models," *Scientific Reports*, vol. p. 13:13824, 2023, doi: https://doi.org/10.1038/s41598-023-41113-5

[4] 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," *Water*, vol. 12, p. 1885, 2020, doi: https://doi.org/10.3390/w12071885

[5] S. R. Krishnan, M. K. Nallakaruppan, R. Chengoden, S. Koppu, M. Iyapparaja, J. Sadhasivam and S. Sethuraman, "Smart Water Resource Management Using Artificial Intelligence: A Review," *Sustainability*, vol. 14, p. 13384, 2022, doi: https://doi.org/10.3390/su142013384

[6] A. A. Nayan, M. G. Kibria, M. O. Rahman and J. Saha, "River water quality analysis and prediction using GBM," *In Proceedings of the 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 28–29 November*, pp. 219–224, 2020.

[7] H. Sahour, V. Gholami, J. Torkaman, M. Vazifedan and S. Saeedi, "Random Forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings," *Environmental Earth Science*, vol. 80, pp. 1–14, 2021.

[8] P. Vijai and P. B. Sivakumar, "Performance comparison of techniques for water demand forecasting," *Procedia Computer Science*, vol. 143, pp. 258–266, 2018, doi: https://doi.org/10.1016/j.procs.2018.10.394.

[9] N. C. T. Maria, S. F. F. Assis and P. V. Costa, "Urban Water Demand Modeling Using Machine Learning Techniques: Case Study of Fortaleza, Brazil*," Journal of Water Resources Planning and Management*, vol. 147, no. 1, p. 05020026, 2021.

[10] H. Kamyab, T. Khademi, S. Chelliapan, M. S. Kamarposhti, S. Rezania, M. Yusuf, and Farajnezhad et al., "The latest innovative avenues for the utilization of artificial Intelligence and big data analytics in water resource management," *Results in Engineering*, vol. 20, p. 101566, 2023, doi: https://doi.org/10.1016/j.rineng.2023.101566.

[11] N. Mutono, J. Wright, H. Mutembei and S. M. Thumbi, "Spatio-temporal patterns of domestic water distribution, consumption and sufficiency: Neighbourhood inequalities in Nairobi, Kenya," *Habitat International*, vol. 119, p. 102476, 2022, https://doi.org/10.1016/j.habitatint.2021.102476

[12] H. Jenny, E.G Alonso, Y. Wang and R. Minguez, "Using Artificial Intelligence for Smart Water Management Systems," Asian Development Bank: Mandaluyong, Philippines, 2020.

[13] U. Akbulut, M.A. Cifci and Z. Aslan, "Hybrid Modeling for Stream Flow Estimation: Integrating Machine Learning and Federated Learning," *Applied Sciences*, vol. 13, p. 10203. https://doi.org/10.3390/app131810203.

[14] B. Ibrahim, L. Rabelo, E. Gutierrez-Franco and N. C. Buritica, "Machine Learning for Short-Term Load Forecasting in Smart Grids," *Energies*, vol. 15, p. 8079, 2022, https://doi.org/10.3390/en15218079.

[15] S. Difi, Y. Elmeddahi, A. Hebal, V. P. Singh, S. Heddam, S. Kim and O. Kisi, "Monthly streamflow prediction using hybrid extreme learning machine optimized by bat algorithm: a case study of Cheliff watershed, Algeria," *Hydrological Sciences Journal*, vol. 68, no. 2, pp. 189-208, 2023, doi: https://doi.org/10.1080/02626667.2022.2149334.

[16] G. G. Garner and A. M. Thompson, "Ensemble statistical post-processing of the National Air Quality Forecast Capability: Enhancing ozone forecasts in Baltimore, Maryland," *Atmospheric Environment*, Vol. 81, pp. 517–522, 2013.