DOI
https://doi.org/10.18690/um.fov.2.2025.73

ISBN
978-961-286-963-2

Data-Centric Approach to Short-Term Water Demand Prediction Using Big Data and Deep Learning Techniques

Alkiviadis Tsimpiris,[1] Georgios Myllis,[1]
Vasiliki Vrana[2]

[1] International Hellenic University, Informatics and Telecommunications Engineering, Department of Computer, Serres, Greece
tsimpiris@ihu.gr, georgmyll@ihu.gr
[2] International Hellenic University, Department of Business Administration, Thessaloniki-N.Moudania, Greece
vrana@ihu.gr

This study introduces a data-centric approach to short-term water demand forecasting, utilizing univariate time series data from water reservoir levels in Eastern Thessaloniki. The dataset, collected over 15 months via a SCADA system, includes water level recordings from 21 reservoirs, generating a substantial Big Data resource. Key components of the methodology include data preprocessing, anomaly detection using techniques like the Interquartile Range method and moving standard deviation, and the application of predictive models. Missing data is addressed with LSTM networks optimized via the Optuna framework, enhancing data quality and improving model accuracy. This approach is particularly valuable in regions where reservoirs are the primary water source, and flow meter readings alone cannot determine demand distribution. By integrating deep learning techniques, such as LSTM models, with traditional statistical methods, the study achieves improved accuracy and reliability in water demand predictions, offering a robust framework for efficient water resource management.

Keywords:
water demand forecasting, big data, deep learning, LSTM networks, anomaly detection

University of Maribor Press

1 Introduction

Water demand forecasting is crucial for sustainable urban water management, particularly amid climate change and urbanization (Wilson, 2016). Regulatory frameworks, such as the EU's Water Framework Directive (Directive - 2000/60 - EN - Water Framework Directive - EUR-Lex, n.d.), emphasize efficient water use and waste prevention . Utility managers rely on short-term forecasts to optimize distribution and meet sustainability targets.

Traditional time series models like ARIMA and SARIMA effectively capture linear trends but struggle with non-linearities and sudden fluctuations in water consumption due to weather variations and sensor malfunctions (Wang et al., 2023). Deep learning models, particularly Long Short-Term Memory (LSTM) networks, offer a superior alternative by capturing long-term dependencies and complex temporal relationships (Hochreiter & Schmidhuber, 1997). Optimizing model performance requires a data-centric approach, focusing on preprocessing, cleaning, and augmentation to improve data quality (Wang et al., 2023).

Given real-world data anomalies—such as sudden spikes from sensor malfunctions—proper preprocessing enhances model reliability. Well-preprocessed datasets consistently yield better forecasts than raw data (Shan et al., 2023). This study evaluates imputation methods, including bi-directional LSTM, linear and polynomial interpolation, mean imputation, and K-Nearest Neighbors (KNN), to determine their impact on forecasting accuracy (Liu et al., 2023).

By integrating deep learning, machine learning, and statistical models, this research aims to enhance water demand forecasting, particularly in reservoir-based systems by analyzing tank levels instead of flow measurements. Advanced preprocessing techniques, including anomaly detection, imputation, and feature selection, improve data integrity, ensuring more accurate short-term forecasts and effective water resource management (Wang et al., 2023).

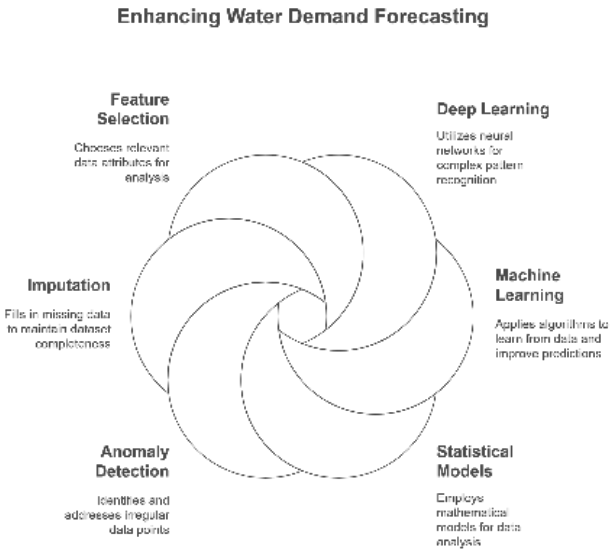


Figure 1: Research aim
Source: Own

2 Materials and Methods

The dataset consists of water level recordings from 21 reservoirs in Eastern Thessaloniki, covering 85 km², provided by EYATH S.A. and collected via a SCADA system at one-minute intervals over 15 months (November 1, 2022 – March 30, 2024), resulting in 13.9 billion high-resolution measurements. While geographically focused, the dataset captures diverse reservoirs in suburban areas with varying population densities, land use, and water consumption behaviors. It includes storage reservoirs, which directly supply residential, commercial, and industrial areas, and intermediary reservoirs, which transfer water without direct end-user supply. To analyze water level dynamics, k-means clustering grouped reservoirs into four categories: (Cluster1) stable water levels with minimal seasonal variation, (Cluster 2) moderate seasonal variation reflecting changing consumption patterns, (Cluster 3) high fluctuations indicating significant usage variability, and (Cluster 4) pronounced seasonal trends, primarily intermediary stations (Figure 2). This clustering enhances understanding of water distribution and consumption patterns, offering insights into operational strategies and optimization (Osman et al., 2018).

Figure 2: Water Level Trends By Cluster (Daily Average)
Source: Own

2.1 Methodology

The study focused on the following research areas, as seen in Figure 3:

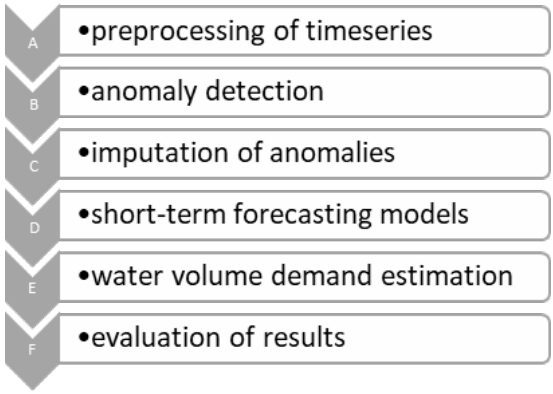


Figure 3: Short-term time series forecasting methodology applied to water tanks level data
Source: Own

Figure 4 presents the workflow of the data formatting at each step of the applied methodology.

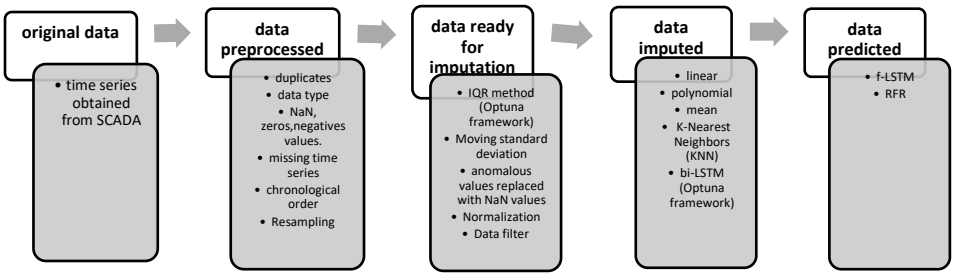


Figure 4: Short-term time series forecasting methodology applied to water tanks level data
Source: Own

2.2 Anomaly Detection

Accurate anomaly detection in time series data is crucial for effective correction and forecasting. Anomalies fall into three categories: contextual anomalies, point anomalies, and collective anomalies (Arslan et al., 2024). This study primarily focuses on detecting point and collective anomalies, as they directly affect data integrity and predictive accuracy, while contextual anomalies have been manually adjusted based on expert knowledge. To identify anomalies in time-series data, point anomalies are detected using the IQR method enhanced with the Optuna framework, while collective anomalies are identified using the moving standard deviation method.

To enhance the IQR measure, Optuna is employed for automated outlier detection optimization by selecting the best lower and upper bounds within predefined quantile ranges (0.01–0.25 for lower, 0.75–0.99 for upper). The objective function minimizes the negative sum of detected outliers, optimizing the inclusion of valid values while excluding anomalies (Niknam et al., 2022). The optimization process, performed using the Tree-structured Parzen Estimator (TPE) algorithm , models good and bad parameter distributions by computing probability densities $l(x)$ and $g(x)$, guiding the search towards optimal bounds via the ratio $l(x)/g(x)$. The densities, estimated using kernel density estimators (KDEs), refine the decision space iteratively (James et al., 2023). The function seeks to minimize detected outliers through the equation (1)

$$f(x,l,u)=1 \text{ if } x<lx<\text{l or } x>u, \text{otherwise } 0 \quad (1)$$

and evaluates the bounds by maximizing equation (2).

$$P(l,u)=-\sum x\in Df(x,l,u) \quad (2)$$

The final optimization goal (Equation 3) maximizes

$$\max_{l,u} \mathrm{P(l,u)} \quad (3)$$

ensuring optimal bounds that improve data quality and reduce false anomaly detection, providing a robust framework for outlier detection and dataset integrity enhancement .

2.3 Data Imputation

Various imputation techniques were employed to restore missing data, including linear and polynomial interpolation, which assume simple relationships between data points but may not handle complex patterns, mean imputation, which replaces missing values with the average but struggles with non-random gaps, and K-Nearest Neighbors (KNN) imputation, which leverages nearby data points but deteriorates with high missing data percentages. Additionally, bi-directional LSTM (bi-LSTM) imputation, leveraging both past and future temporal dependencies, provides a more robust approach for time-series data.

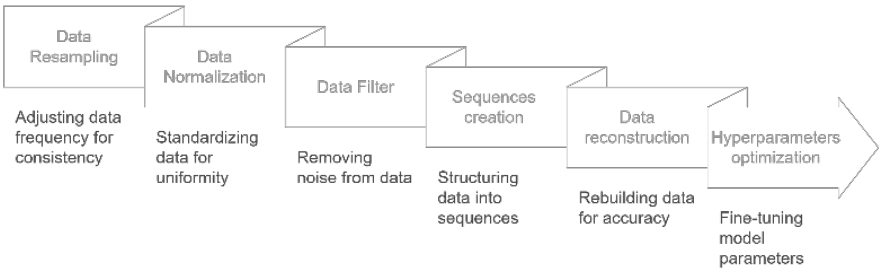


Figure 5 Process of bi-LSTM data imputation method
Source: Own

Bi-LSTM Imputation: This advanced method utilizes a bi-directional Long Short-Term Memory network to learn from data sequences in both forward and backward directions, providing a comprehensive understanding of the temporal dynamics. It is particularly effective for time-series data where capturing temporal dependencies is crucial. Figure 5 illustrates the process followed to impute data for NaN (Not a Number) values.

Data reconstruction: An enhanced approach is employed, interpolating normalized sequence values trained on a bi-LSTM model to reconstruct missing data. This method ensures adequate sequence length for each NaN value, preventing short or insufficient sequences and avoiding issues related to time-dependent indices. By carefully calibrating missing value predictions to prevent overlap with existing data, the approach preserves the temporal integrity of the time series, allowing the model to capture event sequences accurately—critical for time-series forecasting. Unlike traditional methods that disrupt temporal coherence and fail to account for non-linear dependencies, bi-LSTM-based imputation maintains the natural flow of time-series data by learning from both past and future values. This innovative strategy enhances the model's ability to generalize from incomplete datasets, leading to more reliable predictions, particularly in datasets with anomalies and irregular consumption patterns.

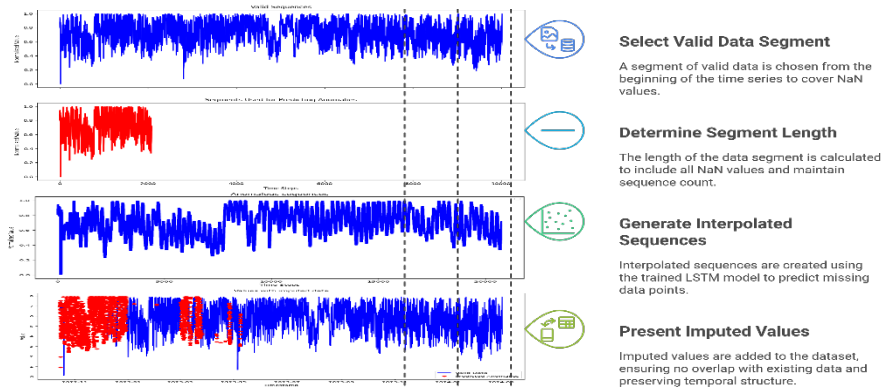


Figure 6 Time series Data imputation process

Source: Own

In Figure 6, the data prepared for imputation includes only valid observations used for training and validating the model. A segment of valid data is selected from the beginning of the time series, with its length carefully determined to cover the total NaN values in the anomaly set while ensuring an appropriate number of sequences. Subsequently, the interpolated sequences for the NaN values are generated, utilizing the trained LSTM model to predict missing data points. Finally, the imputed values are presented, ensuring no overlap with existing valid observations and preserving the temporal structure and continuity within the time series. This approach enhances the accuracy of imputation while maintaining the integrity of the dataset for subsequent forecasting analysis.

Bi-LSTM Hyperparameters Optimization: The Optuna framework is employed for selecting the optimal hyperparameters, which are parameters that define the bi-LSTM structure and are learned during training. These hyperparameters include sequence length, LSTM units, activation function, optimizer, learning rate, additional dense layers, dropout rate, batch size, and the appropriate objective function that best fits training the bi-LSTM model. Sampling techniques such as Random sampling and Tree Parzen Estimator (TPE) were also used. The range of hyperparameters considered is as follows:

- Sequence length: sampled as an integer between 3 and 20,
- LSTM units: sampled as an integer between 20 and 100,
- Activation function: sampled from ['relu', 'tanh', 'sigmoid'],
- Optimizer: sampled from ['adam', 'rmsprop', 'sgd'],
- Learning rate: logarithmically sampled between 0.0001 and 0.01,
- Additional dense layers: sampled as a boolean,
- Dropout rate: sampled between 0.0009 and 0.4274,
- Batch size: sampled from a predefined list of values [16, 32, 64].

3 Results

3.1 Data Imputation performance

The effectiveness of these methods was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) to evaluate their impact on the forecasting accuracy

of f-LSTM, Random Forest models, ensuring high-quality data reconstruction for improved predictive performance (Figure 7).

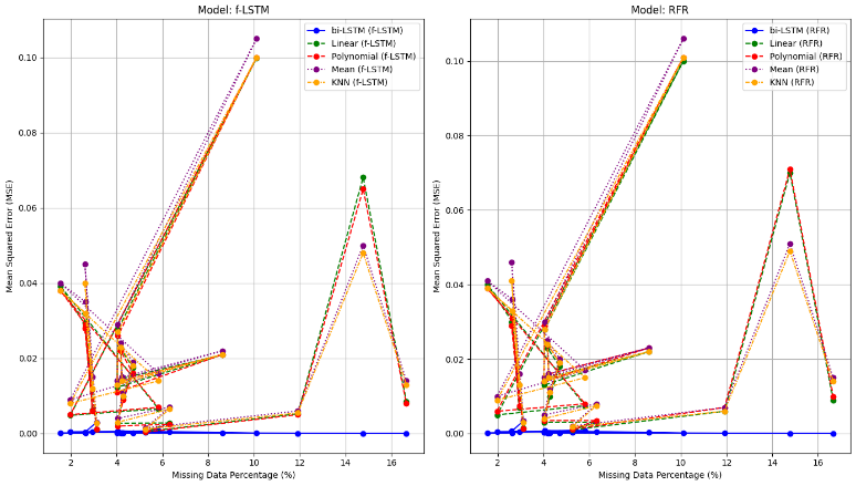


Figure 7: Sensitivity Analysis of Different Imputation Methods per model
Source: Own

Table 1: Analysis of forecasting models with different imputation methods

Model	MSE_ bi-LSTM	MSE_ Linear	MSE_ Polynomial	MSE_ Mean	MSE_ KNN
f-LSTM	0.0003	0.0242	0.0048	0.0206	0.0209
RFR	0.0029	0.0010	0.0015	0.0027	0.0027

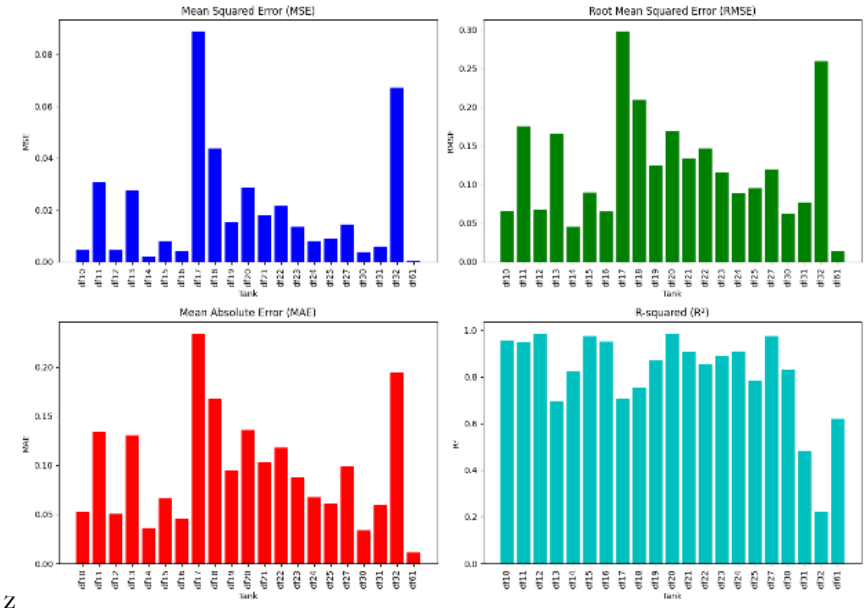
Source: Own

In Figure 8 presented the visualizations for the bi-LSTM imputation accuracy performance metrics (MSE, RMSE, MAE, and R^2) for each tank.

3.2 Water Demand models Forecasting Performance

The Random Forest Regressor (RFR) method predicts water demand using an ensemble of decision trees, where each tree independently partitions the feature space and generates a prediction . Key parameters include n_estimators, determining the number of trees, and random_state, ensuring reproducibility. Each tree is trained on a random subset of data (bootstrap sampling), enhancing accuracy and reducing variance and overfitting. Mean Squared Error (MSE) is used to optimize tree splits

for maximum node purity (Breiman, 2001). The dataset undergoes preprocessing, where date-time columns are transformed into time labels, and lag features are introduced to capture temporal dependencies. The data is split into training (70%), validation (15%), and test sets (15%). The RFR is trained on the training set, validated on unseen data, and tested for predictive accuracy. The final prediction $\hat{y}(X, D)$ for an input X after training on dataset D is computed as the average prediction across all trees, reducing variability and improving forecasting reliability.



z

Figure 8: bi-LSTM imputation accuracy performance metrics for each tank
Source: Own

The f-LSTM-based water demand prediction method incorporates temporal features such as year, month, day, and hour to capture seasonal and temporal trends. These features are added as new columns, enabling the model to analyze long-term variations, seasonal patterns, and intra-day fluctuations, improving predictive accuracy (Niknam et al., 2022). The input and output data are normalized to the range (0,1) for optimized training, and the dataset is split into training (70%), validation (15%), and test sets (15%). The prediction model consists of an LSTM layer with 44 units, a ReLU activation function, a dropout layer (0.2) to prevent overfitting, and a Dense output layer for the final prediction. The model is trained

using the Adam optimizer and MSE loss function, with evaluation conducted on the test set. This approach ensures optimal model performance, effectively leveraging temporal dependencies and seasonal variations to enhance forecasting accuracy.

After the application of the bi-LSTM imputation method to the data, the f-LSTM model's performance was compared with RFR model for time series forecasting. Figure 8 presents the comparison focused on the same evaluation metrics (MSE, RMSE, MAE, and R2) to test sets.

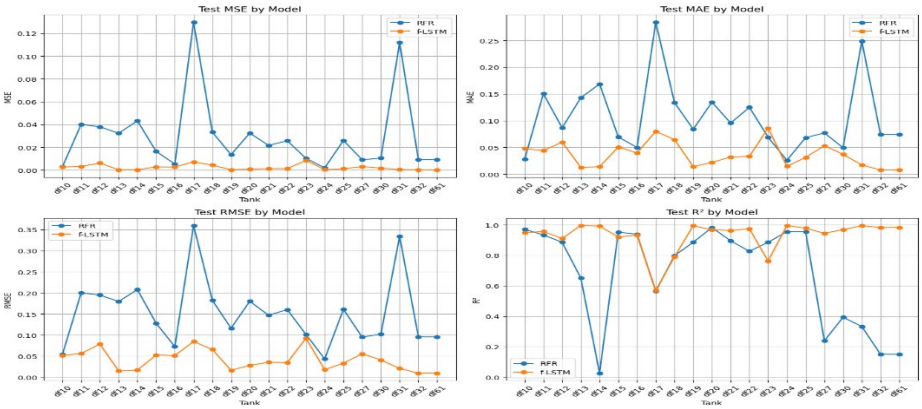


Figure 9: Forecasting performance of test data by model and tank
Source: Own

Table 3: Model comparison forecasting average performance in test sets

Model	Avg_Test_MSE	Avg_Test_MAE	Avg_Test_RMSE	Avg_Test_R2
RFR	0.0228	0.0960	0.1340	0.7199
F-LSTM	0.0002	0.0124	0.0150	0.9921

Source: Own

Table 4: Performance f-LSTM metrics for different imputation methods per cluster

Tank	Model	Missing %	MSE_ bi-STM	MSE_ Linear	MSE_ Polynomial	MSE_ Mean	MSE_ KNN
df10	f-LSTM	3.12	0.0030	0.0011	0.0012	0.0030	0.0028
df11	f-LSTM	2.61	0.0003	0.0294	0.0280	0.0450	0.0400
df31	f-LSTM	11.91	0.0001	0.0052	0.0050	0.0060	0.0055
df32	f-LSTM	14.77	0.0001	0.0681	0.0650	0.0500	0.0480

Source: Own

4 Conclusions

This study aimed to improve the accuracy and reliability of short-term water demand forecasting by integrating deep learning and machine learning models with traditional statistical approaches. Advanced imputation methods, particularly the bi-directional Long Short-Term Memory (bi-LSTM) network, significantly enhanced forecasting performance, especially in cases with high percentages of missing data. Cluster analysis further revealed that water level dynamics influenced model effectiveness and hyperparameter optimization via Optuna. Simpler models performed well in stable water levels (Cluster 1), while reservoirs with moderate to high seasonal variations (Clusters 2 and 3) required advanced hyperparameters, such as increased LSTM units and varied learning rates, to improve accuracy. Highly dynamic reservoirs (Cluster 4) demonstrated that only deep learning models like forecasting LSTM (f-LSTM) consistently captured complexities, whereas traditional methods struggled. Bi-LSTM consistently outperformed traditional imputation methods, such as linear, polynomial, mean, and K-Nearest Neighbors (KNN) imputation, achieving lower Mean Squared Error (MSE) and higher R^2 values, indicating its superior ability to handle complex temporal dependencies. It reduced forecasting errors by 15–20% in datasets with over 10% missing data, though increased prediction errors were observed in datasets with extreme irregularities or sudden spikes, with MSE values reaching 0.045 in such cases compared to 0.0025 in more stable datasets. Model effectiveness was found to be context-dependent, with f-LSTM demonstrating the highest robustness and accuracy, particularly in scenarios with non-linear dependencies and significant data irregularities, achieving an MSE as low as 0.0026, significantly outperforming traditional machine learning models. The Random Forest Regressor (RFR) effectively handled noisy and imprecise data, with MSE values between 0.015 and 0.050 in high-variability datasets, highlighting its strength in short-term forecasting but limitations in capturing long-term dependencies. Overall, f-LSTM excelled in modeling complex temporal relationships, while RFR offered strong noise tolerance, making both models highly effective in real-world water demand forecasting. Future research should focus on validating these models across diverse operational conditions, refining hyperparameter optimization, incorporating contextual data, and scaling them to larger datasets. Further advancements in real-time data integration, anomaly detection, and imputation techniques will enhance forecasting accuracy. Additionally, practical implementation should prioritize integrating these models

into operational water management systems, developing user-friendly interfaces for utility managers, and aligning model outputs with regulatory frameworks to ensure real-world applicability and adoption.

Acknowledgment

This research work was supported by the "Applied Informatics" Post Graduate Program of Computer, Informatics and Telecommunications Engineering Department, International Hellenic University, Greece and thank EYATH S.A. for their valuable assistance and provision of data.

References

Arslan, F., Javaid, A., Danish Zaheer Awan, M., & Ebad-ur-Rehman. (2024). Anomaly Detection in Time Series: Current Focus and Future Challenges. In V. Krishna Parimala (Ed.), Artificial Intelligence (Vol. 24). IntechOpen. https://doi.org/10.5772/intechopen.111886

Breiman, L. (2001). [No title found]. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Directive—2000/60—EN - Water Framework Directive—EUR-Lex. (n.d.). Retrieved January 30, 2025, from https://eur-lex.europa.eu/eli/dir/2000/60/oj/eng

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning: With Applications in Python. Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0

Liu, G., Savic, D., & Fu, G. (2023). Short-term water demand forecasting using data-centric machine learning approaches. Journal of Hydroinformatics, 25(3), 895–911. https://doi.org/10.2166/hydro.2023.163

Niknam, A., Zare, H. K., Hosseininasab, H., Mostafaeipour, A., & Herrera, M. (2022). A Critical Review of Short-Term Water Demand Forecasting Tools—What Method Should I Use? Sustainability, 14(9), 5412. https://doi.org/10.3390/su14095412

Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A Survey on Data Imputation Techniques: Water Distribution System as a Use Case. IEEE Access, 6, 63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Shan, S., Ni, H., Chen, G., Lin, X., & Li, J. (2023). A Machine Learning Framework for Enhancing Short-Term Water Demand Forecasting Using Attention-BiLSTM Networks Integrated with XGBoost Residual Correction. Water, 15(20), 3605. https://doi.org/10.3390/w15203605

Wang, K., Ye, Z., Wang, Z., Liu, B., & Feng, T. (2023). MACLA-LSTM: A Novel Approach for Forecasting Water Demand. Sustainability, 15(4), 3628. https://doi.org/10.3390/su15043628

Wilson, G. T. (2016). Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. Journal of Time Series Analysis, 37(5), 709–711. https://doi.org/10.1111/jtsa.12194