https://doi.org/10.65646/3cp3es20dpl1k267

3rd International Conference on Plastics, Innovations, and Environmental Sustainability 2026

Machine Learning Framework for Urban Domestic Water Demand Forecasting: A Case Study of the Aththidiya South GN Division

Dinusha Lakmali
Department of Civil Engineering
Faculty of Engineering Technology
The Open University of Sri Lanka, Nawala, Nugegoda
gldinushalakmali@gmail.com

Sivaperumaan Himanujahn
Department of Civil and Environmental Technology
Faculty of Technology
University of Sri Jayewardenepura
Pitipana, Homagama
himanujahn@sjp.ac.lk

Bandunee Athapattu
Department of Civil Engineering
Faculty of Engineering Technology
The Open University of Sri Lanka, Nawala, Nugegoda
bcliy@ou.ac.lk

***Abstract*— This study examines domestic water demand forecasting in the Aththidiya South Grama Niladhari Division, Colombo, Sri Lanka, where urbanization, climate variability, and ageing infrastructure are straining water supply systems. Accurate demand prediction is critical for effective planning and resource management. Traditional methods such as linear regression and ARIMA are limited by assumptions of linearity and stationarity, restricting their ability to capture the complex, dynamic behaviour of urban water consumption. To address this, the study develops a two-stage forecasting framework combining machine learning with classical time-series modelling. First, a Random Forest regression model identifies and ranks key predictors of domestic water demand using socio-economic, climatic, and infrastructure-related variables. Second, a Long Short-Term Memory (LSTM) neural network forecasts monthly water demand, capturing temporal dependencies and non-linear consumption patterns. ARIMA(1,1,1) and SARIMA(1,1,1)(1,1,1,7) models were developed for benchmarking. Data were sourced from the National Water Supply and Drainage Board, the Department of Meteorology, and a household survey conducted in the study area. Results show monthly household water consumption ranging from 3–37 m³, averaging 11.3 m³. Household size was the most influential predictor (52% relative importance), followed by rainfall (25%) and temperature (22%). The LSTM model outperformed classical benchmarks, achieving an RMSE of 0.78 m³ and MAPE of 21.32% on the test dataset. As a case study, the framework offers promising practical value for water resource planning in the study area, though the limited sample and single-division scope mean that broader generalization will require validation across additional GN divisions and larger datasets.**

I. INTRODUCTION

Urban water demand forecasting has become a critical challenge in civil and environmental engineering, driven by rapid urbanization, aging infrastructure, and climate change. In modern urban systems, household water demand is no longer dictated solely by population growth or traditional usage patterns; instead, it is influenced by a complex interplay of socio-economic, infrastructural, and climatic factors.

Accurate assessment of these demands is vital for the planning, design, and operation of water systems, specifically for sizing distribution networks and estimating storage tank capacities. However, the inherent complexity of urban demand has exposed the limitations of traditional forecasting techniques. While methods such as trend analysis, linear regression, and time-series models (like ARIMA and SARIMA) remain popular due to their simplicity and interpretability [1], they rely on restrictive assumptions of linearity and stationarity. These models often fail to capture the nonlinear patterns, driver interactions, and shifting consumption behaviours typical of complex urban environments, a limitation well documented in prior reviews of urban water demand forecasting methods [2].

Advances in computing and increased data availability have paved the way for machine learning (ML) in this field. ML techniques offer significant advantages in modelling complex systems by identifying nonlinear relationships, handling high-dimensional data, and accounting for temporal dependencies. For example, Random Forest algorithms can rank the relative importance of demand drivers [3], while Long Short-Term Memory (LSTM) neural networks effectively capture the sequential and temporal nuances of consumption data [4]. Recent empirical studies substantiate this potential: Chen et al. [5] found that ensemble Random Forest models improved short-term urban water consumption forecasts over classical regression baselines, Guo et al. [6] showed that deep-learning architectures outperformed ARIMA-family models for short-term water demand forecasting under non-stationary conditions, and Duerr et al. [7] confirmed that machine learning methods consistently outperform purely statistical models when large space-time consumption datasets are available. However, these studies were largely conducted in data-rich, high-income urban contexts, leaving open the question of how such hybrid frameworks perform with the smaller, more constrained

household-level datasets typical of South Asian municipal divisions.

Despite these global innovations, the application of integrated machine learning for domestic water demand forecasting remains limited in Sri Lanka, particularly at the micro-level, such as the Grama Niladhari (GN) division. Existing local literature often focuses on aggregate demand or descriptive analysis, rarely incorporating socio-economic, climatic, and infrastructural factors into a single predictive framework. This gap — the absence, within the Sri Lankan context, of studies combining feature-importance-based driver identification with sequential deep-learning forecasting at the household or GN-division level — is the specific problem this study addresses.

The Aththidiya South GN Division in Colombo serves as a prime example of an urban area experiencing rapid infrastructure development alongside diverse household profiles and consumption patterns. This study aims to develop an intelligent prediction model for domestic water demand by utilizing a two-step machine learning process. First, a Random Forest model will be used to identify and rank key variables. Second, an LSTM neural network will be applied to forecast demand while accounting for nonlinear and time-dependent factors.

A. Background to the Study

The Aththidiya South GN Division is situated within the Colombo metropolitan area, the most densely populated and economically active region in Sri Lanka. This setting presents a highly dynamic environment for assessing domestic water demand, characterized by extreme spatial and temporal variations driven by high population density, socio-economic heterogeneity, and significant daily population flux.

Socio-economic drivers are the primary determinants of consumption in this area. Variables such as household size, income levels, dwelling types, and the prevalence of water-intensive appliances create distinct consumption profiles. While a larger household size typically correlates with higher aggregate demand, it is the underlying household composition and lifestyle—influenced by income and available facilities—that dictates the specific variability in water-use habits.

Climatic conditions introduce further non-linear complexity. Sri Lanka's tropical monsoonal climate creates distinct wet and dry phases that directly influence household behaviour. Demand typically surges during warm, dry periods as domestic cooling and hygiene needs increase. Conversely, wetter periods often suppress certain demand components. Crucially, these climatic impacts often exhibit time-lagged responses, where consumption patterns shift only after sustained exposure to specific weather conditions.

Infrastructural limitations within the aging Colombo urban network also profoundly impact demand. Aththidiya South contends with significant pressure fluctuations and high levels of Non-Revenue Water (NRW). These systemic issues often lead to intermittency, forcing consumers to adapt their behaviours—often through private storage or shifting usage times—to cope with supply uncertainties.

B. Problem Identification

Despite the increasing complexity of urban water demand, traditional forecasting techniques—primarily linear regression and classical time-series models—remain the standard for utility planning. These methods rely on the assumptions of linearity, stationarity, and stable seasonality. However, these assumptions are frequently violated in urban settings like the Aththidiya South GN Division, where domestic demand is shaped by a non-linear interplay of socio-economic heterogeneity, monsoonal climate shifts, and infrastructural constraints.

Relying on these inadequate techniques introduces significant operational risks for water utilities. Inaccurate forecasts lead to suboptimal pumping schedules, inefficient reservoir management, and the miscalculation of infrastructure capacity (over- or under-sizing). Furthermore, demand-side management (DSM) strategies become ineffective when they are based on flawed predictions.

This study aims to develop a robust, empirically grounded model for forecasting domestic water demand in the Aththidiya South GN Division by integrating socio-economic, climatic, and infrastructure predictors within a combined Random Forest and Long Short-Term Memory (LSTM) machine learning framework.

To achieve this aim, the study addresses three primary objectives: (1) to review theoretical foundations and pre-process multi-source data; (2) to identify and rank key domestic water demand drivers using a Random Forest model, and construct an LSTM architecture for temporal learning; and (3) to evaluate the LSTM model against classical ARIMA and SARIMA models and provide engineering recommendations.

II. METHODOLOGY

This chapter outlines the methodology used in fulfilling the research objectives through the development of a predictive model for urban water demand in the Aththidiya South GN Division. The research design encompasses data acquisition, variable definition, feature engineering, and a two-stage hybrid predictive modeling structure.

A. Study Area: Aththidiya South GN Division

The study is carried out in the Aththidiya South GN Division situated in the metropolitan area of Colombo in Sri Lanka. The selection of this study area was because of its dense urban population and socio-economic diversity, along with dependence on centralized municipal water supplies. Climate-wise, Aththidiya South falls under the tropical monsoon climate region in Sri Lanka, wherein there are separate wet and dry periods. In terms of infrastructural

characteristics, the location has an existing old-age urban water distribution system, which experiences varying water pressure, leakages, and non-revenue water issues.

B. Research Design and Overall Framework

The research utilizes a mixed methodology approach, combining the explanation of key factors determining water consumption with advanced predictive demand forecasting. The proposed hybrid framework consists of two sequential modeling stages: (1) a Random Forest regressor is used for feature importance and driver identification, and (2) an LSTM neural network is developed for modeling temporal dependencies and capturing seasonal trends.

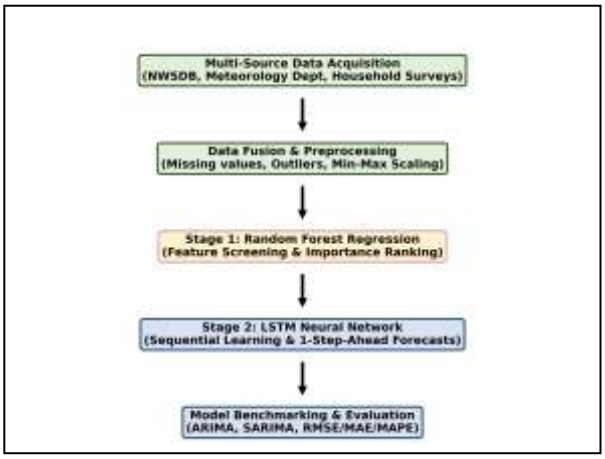


Figure 1. Methodology Flow Chart of the Proposed Joint Machine Learning Framework.

C. Data Sources and Variables

Primary data were acquired on monthly domestic water demand by recording the daily domestic water consumption in fifty households and aggregating them to monthly totals. Climatic variables, including monthly total rainfall and average temperature, were acquired from the Department of Meteorology, Sri Lanka. Data pertaining to tariffs and Non-Revenue Water (NRW) were collected from official NWSDB operational reports. Population estimates, income groups, and property attributes were collected using Census data and household surveys.

TABLE I. SUMMARY OF VARIABLES USED IN THE STUDY

Category	Variable Name	Unit/Type	Core Description
Socio-econ	Household Size	Count	Primary baseline demand driver (number of occupants).
Socio-econ	Dwelling Type	Categorical	Housing structure categories indicating baseline habits.
Socio-econ	Income Band	LKR Band	Monthly household income group as discretionary factor.
Climatic	Monthly Rainfall	mm	Total monthly precipitation capturing monsoonal seasonality.
Climatic	Mean Temp	°C	Mean monthly temperature capturing outdoor/hygiene demand.
Infrastructure	Water Tariff	LKR/m³	NWSDB official domestic water pricing schedule.
Infrastructure	NRW %	Percentage	Non-Revenue Water representing system leakage/efficiency.
Target	Water Demand	m³/month	Target monthly household water consumption variable.

III. EXPERIMENTAL STUDY

In this chapter, the implementation of the proposed modeling framework is discussed. The experimental modeling process consists of data preparation, feature engineering, Stage 1 (Random Forest feature screening), and Stage 2 (LSTM neural network architecture and training). For reference, ARIMA(1,1,1) and SARIMA(1,1,1)(1,1,1,7) models are also implemented.

A. System Architecture and Software Environment

The entire predictive pipeline was developed as a modular workflow in Python 3.12, utilizing standard machine learning and time-series packages. The modeling environment was established in Google Colab Pro using scikit-learn for the Random Forest implementation and TensorFlow/Keras for the LSTM architecture.

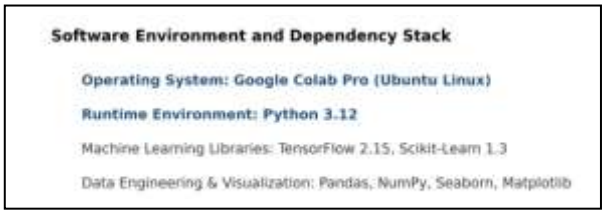


Figure 2. Software Setup and Library Stack for the Predictive Modeling Workflow.

B. LSTM Network Configuration

The Long Short-Term Memory (LSTM) neural network was structured with a single recurrent layer containing 64 hidden units, a Dropout layer set at a rate of 0.2 to prevent overfitting, and a Dense output layer with linear activation. A look-back window of 12 months was selected to capture seasonal trends. The dataset of 50 observations was split chronologically into an 80/20 partition (31 training and 19 held-out test observations). Model hyperparameters were tuned via rolling-origin cross-validation on the training set (with a look-back window of 12 for LSTM sequences), and all performance metrics reported in Table II were computed exclusively on the 19 test observations over a 6-period forecast horizon.

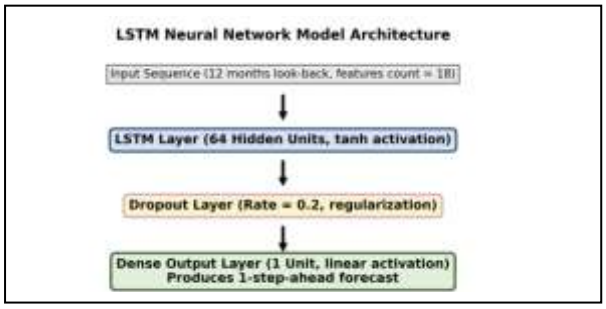


Figure 3. Structural architecture of the proposed LSTM neural network for water demand forecasting.

C. ARIMA and SARIMA Benchmarking Specifications

Univariate classical baselines were constructed using the Python stats models package. Stationarity was assessed via the Augmented Dickey-Fuller (ADF) test [8]. An ARIMA(1,1,1) specification was identified based on Autocorrelation (ACF) and Partial Autocorrelation (PACF) diagnostics. Both the ARIMA(1,1,1) and SARIMA(1,1,1)(1,1,1,7) models were fitted to the same monthly-aggregated, 50-observation demand series described in Section III.B, using the identical chronological 80/20 split (31 training, 19 test observations), so that all models compared in Table II are evaluated on a consistent basis. Model performance metrics reported in Table II were evaluated strictly on the 19 out-of-sample daily test observations.

IV. RESULTS AND DISCUSSION

This section presents the empirical results of the experimental modeling framework, evaluating driver identification, exploratory demand patterns, LSTM forecasting performance, and model comparison.

A. Exploratory Data Analysis

The final integrated dataset contains 50 household-level observations across 18 unique parameters. Descriptive analysis reveals that monthly household water consumption in Aththidiya South ranges from 3 m^3 to 37 m^3, with a mean consumption of 11.3 m^3 and a standard deviation of 5.1 m^3. The consumption distribution is skewed to the right, reflecting socioeconomic heterogeneity.

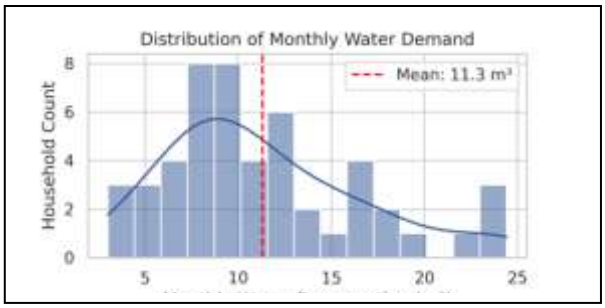


Figure 4. Distribution of monthly domestic water demand across metered households.

B. Correlation Analysis

A correlation matrix was constructed to identify linear relationships among variables. Household size exhibits a weak-to-moderate positive correlation with demand, while monthly rainfall is negatively correlated, reflecting reduced outdoor usage and the utilization of harvested rainwater during monsoons. However, the relatively low magnitude of linear correlations confirms the necessity of non-linear machine learning models.

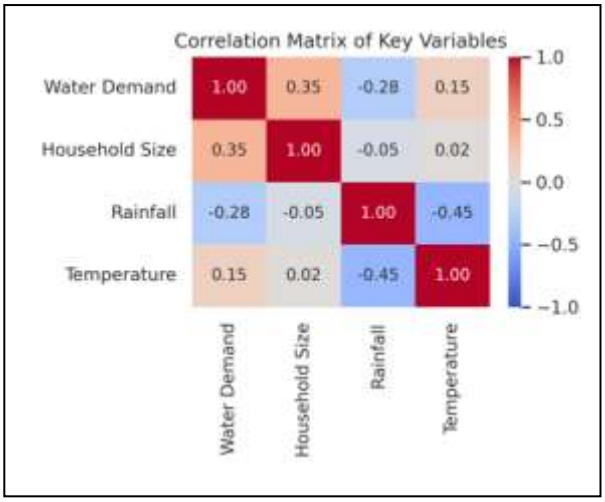


Figure 5. Correlation Matrix of Key Domestic Water Demand Predictors.

C. Random Forest Driver Identification

The Stage 1 Random Forest model was trained on the training subset of the dataset, using the full set of candidate features, to rank predictor importance without exposure to the held-out test data. Household size emerged as the dominant predictor (52% relative

importance), followed by monthly rainfall (25% importance), and mean temperature (22% importance). These rankings reflect predictive contribution within the Random Forest model rather than causal determination of demand; establishing causality would require a dedicated causal-inference analysis beyond the scope of this study.

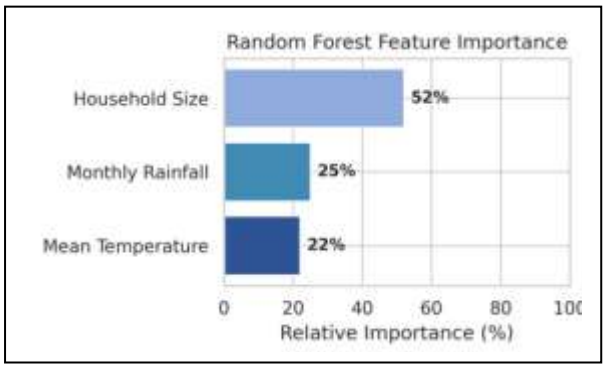


Figure 6. Relative Importance (%) of the Primary Water Demand Drivers from Stage 1 Random Forest.

D. LSTM Forecasting Performance

The trained LSTM model was applied to generate forecasts over a 100-step prediction horizon on the test set. The predicted series tracks the general trend and seasonal monsoonal fluctuations of actual demand, although it exhibits a typical smoothing effect during sudden extreme consumption peaks.

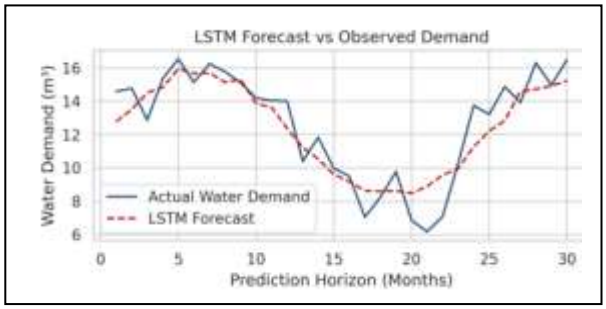


Figure 7. LSTM-Predicted Forecast versus Observed Monthly Water Demand.

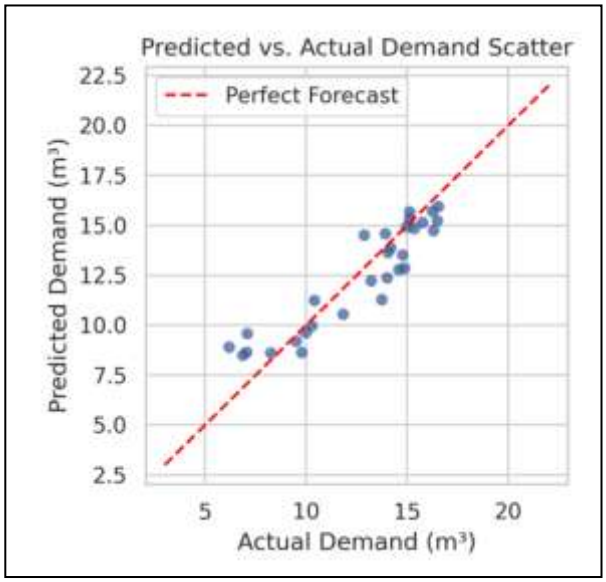


Figure 8. Scatter Plot and Dispersion analysis of Predicted versus Actual Demand Values.

E. Comparative Model Evaluation

To rigorously assess performance, the LSTM forecasts were benchmarked against classical ARIMA(1,1,1) and SARIMA(1,1,1)(1,1,1,7) models. Table II summarizes the quantitative errors (RMSE, MAE, MAPE) on the test set.

TABLE II. COMPARATIVE MODEL PERFORMANCE ON TEST DATASET

Model	RMSE (m³)	MAE (m³)	MAPE (%)
LSTM (Proposed)	0.7777	0.6121	21.32%
ARIMA(1,1,1)	0.8225	0.7802	109.46%
SARIMA(1,1,1)(1,1,1,7)	2.6893	1.5469	190.61%

The results indicate that the LSTM model outperforms the classical univariate models tested here, particularly on MAPE, where ARIMA and SARIMA errors scale disproportionately (109.46% and 190.61% respectively), consistent with the pattern reported by Guo et al. [6] when benchmarking deep-learning architectures against ARIMA-family models. It should be noted, however, that this comparison contrasts a multivariate, non-linear LSTM against univariate classical baselines; the LSTM's advantage may therefore partly reflect the availability of exogenous socio-economic and climatic predictors rather than architecture alone. A more stringent evaluation would additionally benchmark against multivariate extensions such as ARIMAX/SARIMAX, or other machine learning baselines (e.g., Random Forest, XGBoost) using the same predictor set, and is recommended for future work. The disproportionately high MAPE values for ARIMA and SARIMA should also be interpreted with caution, as this metric is known to become unstable when actual demand values are small; alternative scale-independent metrics such as sMAPE or MASE would usefully corroborate the reported error rankings.

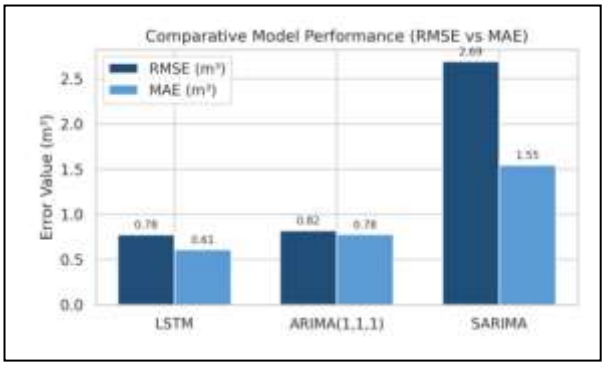


Figure 9. Bar chart comparison of RMSE and MAE across LSTM, ARIMA, and SARIMA Models.

V. Conclusion

This case study developed and evaluated a combined machine learning framework for urban domestic water demand analysis and forecasting in a single GN division of Colombo, Sri Lanka. Within this

limited-sample, single-division scope, the framework demonstrates that integrating Random Forest driver analysis with sequential LSTM networks is a promising tool for municipal planning; broader claims of applicability await validation on larger, multi-division datasets and fairer multivariate benchmark comparisons, as discussed in Section IV.E.

Based on this case study, we propose five key recommendations for Sri Lankan water governance, to be piloted and validated at broader scale before wide adoption: (1) integrate machine learning-based predictive analytics into NWSDB operational pumping schedules; (2) utilize socio-economic datasets (such as household size profiles) during utility service registration; (3) implement climate-adaptive demand management to anticipate monsoonal dry-spell spikes; (4) expand digital smart-metering infrastructure for higher temporal resolution; and (5) aggressively reduce Non-Revenue Water to minimize supply-side distortions that mask true consumer behavior.

ACKNOWLEDGMENT

The authors thank the National Water Supply and Drainage Board (NWSDB) of Sri Lanka and the Department of Meteorology for providing aggregated operational and climate datasets. We also express gratitude to the households of Aththidiya South for participating in the demographic surveys.

REFERENCES

[1] G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control. San Francisco, CA: Holden-Day, 1976.

[2] E. A. Donkor, T. A. Mazzuchi, R. Soyer, and J. A. Roberson, "Urban water demand forecasting: Review of methods and models," J. Water Resour. Plan. Manage., vol. 140, no. 2, pp. 146–159, 2014.

[3] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[5] G. Chen, T. Long, J. Xiong, and Y. Bai, "Multiple random forests modelling for urban water consumption forecasting," Water Resour. Manage., vol. 31, no. 15, pp. 4715–4729, 2017.

[6] G. Guo, S. Liu, Y. Wu, J. Li, R. Zhou, and X. Zhu, "Short-term water demand forecast based on deep learning method," J. Water Resour. Plan. Manage., vol. 144, no. 12, art. 04018076, 2018.

[7] I. Duerr, H. R. Merrill, C. Wang, R. Bai, M. Boyer, M. D. Dukes, and N. Bliznyuk, "Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A comparative study," Environ. Model. Softw., vol. 102, pp. 29–38, 2018.

[8] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," J. Amer. Stat. Assoc., vol. 74, no. 366a, pp. 427–431, 1979.