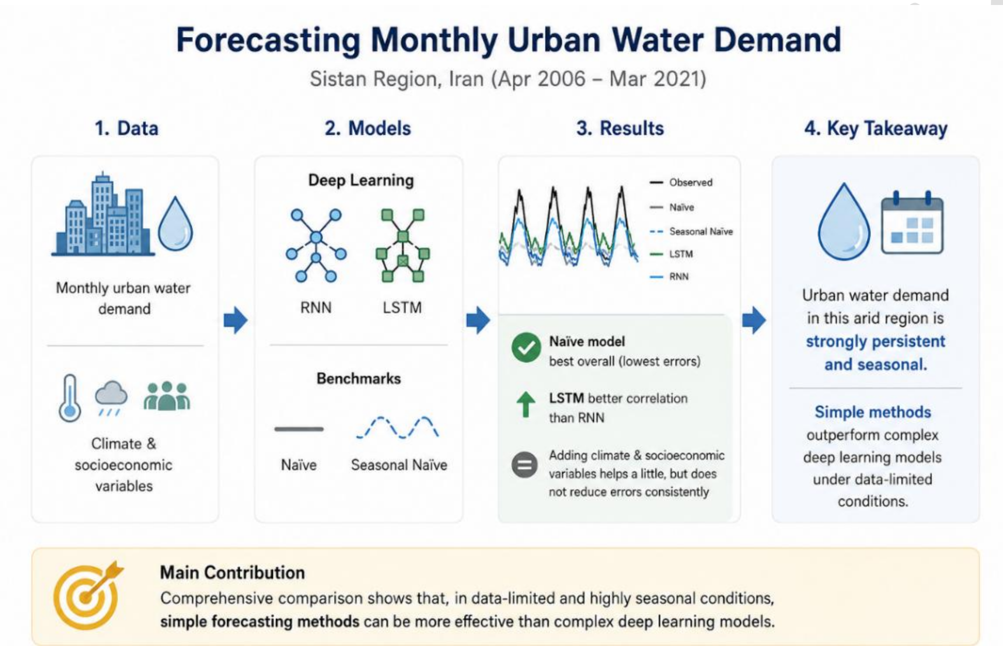
Journal of Applied Research in Water and Wastewater
E-ISSN: 2476-6283
Journal homepage: https://arww.razi.ac.ir

A comparative study of Deep Learning and Benchmark Models for monthly urban water demand forecasting

Zahra Ghaffari Moghadam[*1], **Ali Sardar Shahraki**[2], **Abdulhasseb Azizi**[3], **Somayyeh Mirshekari**[1]

[1]*Agriculture Institute, Research Institute of Zabol, Zabol, Iran.*
[2]Department *of Agricultural Economics, University of Sistan and Baluchestan, Zahedan, Iran.*
[3]*Center for Development Research (ZEF), University of Bonn, Germany.*

GRAPHICAL ABSTRACT



ARTICLE INFO

Article type:
Research Article

Article history:
Received xx Month xxx
Received in revised form xx Month xxx
Accepted xx Month xxx
Available online x Month xx

Keywords:
Water demand forecasting
Deep learning
LSTM neural network
RNN neural network
Climate variability

ABSTRACT

Accurate forecasting of urban water demand is essential for effective water resources management, especially in arid and water-stressed regions. This study evaluates the performance of Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models for forecasting monthly urban potable water demand in the Sistan region of Iran using data from April 2006 to March 2021. Two forecasting approaches are examined: a univariate approach based only on past water demand and a multivariate approach incorporating selected climatic and socioeconomic variables. A one-step-ahead monthly forecasting strategy is applied, with the dataset divided chronologically into training and testing subsets. The deep learning models are compared with two benchmark methods, Naïve and Seasonal Naïve, using MAE, RMSE, MAPE, and Pearson's correlation coefficient. The results show that LSTM provides more stable predictions and higher correlation with observed demand than RNN. However, the simple benchmark models produce lower forecast errors overall, with the Naïve model achieving the best performance. Adding climatic and socioeconomic variables slightly improves correlation in some cases but does not consistently reduce errors. Overall, the findings suggest that urban water demand in the study area is strongly persistent and seasonal, indicating that simple forecasting methods can outperform more complex deep learning models under data-limited conditions.

1. Introduction

Rapid population growth, urban expansion, economic development, and rising living standards have led to a steady increase in global water demand over recent decades (Ercin and Hoekstra, 2014; Sanchez *et al*., 2020). At the same time, climate change has intensified pressure on freshwater resources by altering precipitation patterns, increasing temperatures, and exacerbating the frequency and severity of droughts,

*Corresponding author Email: zahraghaffari@rizabol.ac.ir

Cite this article: Z. Ghaffari Moghadam, A. Sardar Shahraki, A. Azizi, S. Mirshekari, A comparative study of Deep Learning and Benchmark Models for monthly urban water demand forecasting, *Journal of Applied Research in Water and Wastewater*, x (x), 202x, xx-xx. doi: https://doi.org/10.22126/arww.2026.13557.1476

particularly in arid and semi-arid regions (Barbieri *et al*., 2023; Mehrazar *et al*., 2020). These combined pressures have significantly increased global water stress, making efficient planning and sustainable management of urban water systems an urgent priority.

Accurate forecasting of urban water demand plays a critical role in water resources management because it supports informed decision-making related to supply planning, infrastructure development, demand-side management, and drought mitigation strategies (Al-Zahrani and Abo Monasar, 2015). Understanding the temporal dynamics of water consumption and its sensitivity to climatic and socioeconomic drivers is therefore essential for ensuring long-term water security, particularly in regions facing severe hydrological and climatic constraints.

A wide range of approaches has been proposed for urban water demand forecasting. These approaches are generally categorized into three main groups: econometric models, multivariate statistical models, and machine learning-based methods (Panagopoulos, 2014; Daw, Ali and Toriman, 2021; Li *et al*., 2021). Traditional statistical and econometric models are often valued for their transparency and interpretability; however, their ability to capture nonlinear relationships and complex temporal dependencies is often limited. In contrast, machine learning approaches are capable of learning nonlinear patterns directly from data and have therefore been increasingly applied to water demand forecasting problems (Lee and Derrible, 2020; Nunes Carvalho, de Souza Filho and Porto, 2021). Nevertheless, water consumption behavior remains highly complex, as it is influenced by climatic variability, demographic changes, institutional regulations, and human behavior.

In recent years, deep learning models, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, have attracted considerable attention in time-series forecasting because of their ability to model sequential dependencies and nonlinear dynamics. Several studies have reported promising results when applying LSTM-based models to urban water demand prediction across different temporal scales and climatic conditions (Velasco *et al*., 2018; Salloom, Kaynak and He, 2021; Boudhaouia and Wira, 2021; Kim *et al*., 2022). Furthermore, multivariate and hybrid frameworks have been proposed to improve forecasting performance by incorporating climatic and socioeconomic variables into forecasting models (Hu *et al*., 2019; Niknam *et al*., 2023; Pu *et al*., 2023). Recent advances in time-series forecasting have introduced transformer-based models. Liu *et al*. (2025) proposed Timer-XL, a long-context causal Transformer with a TimeAttention mechanism for unified forecasting of univariate, multivariate, and exogenous-variable-informed time series. Li *et al*. (2024) proposed an interpretable hybrid deep learning model, AGRS-LSTM-Transformer, for flood forecasting by combining LSTM, Transformer, and an Adaptive Random Search Algorithm. Their results showed that the hybrid model outperformed LSTM-, Transformer-, BP-, and MLP-based benchmarks in 1–6-hour runoff forecasting, while SHAP-based interpretability analysis helped identify the temporal and spatial contributions of rainfall and runoff inputs. Iwakin and Moazeni (2024) introduced a conformal prediction-based hybrid machine learning model for probabilistic hourly urban water demand forecasting. Their approach combined CNN and BiLSTM to generate both deterministic and uncertainty-aware forecasts, showing improved prediction accuracy and robustness compared with other machine learning models in a real-world case study in Telford, Pennsylvania. Qiao *et al*. (2024) proposed an attention-based spatiotemporal graph convolutional model, FSGCN, for water quality prediction. The model combines temporal attention, graph convolution, and temporal convolution residual mechanisms to capture nonlinear temporal patterns, spatial dependencies among river-network sensors, and long-term forecasting dynamics. Experiments on two real-world datasets showed that FSGCN outperformed several state-of-the-art methods in prediction accuracy.

However, the findings reported in the literature are not fully consistent, and the performance of deep learning models relative to simpler forecasting approaches cannot be generalized across all settings. Recent studies indicate that model performance strongly depends on factors such as data availability, temporal resolution, the strength of seasonal patterns, and persistence in the demand series (Zanfei *et al*., 2022; Mu *et al*., 2020). While some studies have shown that RNN- and LSTM-based models can outperform conventional statistical methods by capturing nonlinear relationships and long-term temporal dependencies, these advantages are typically reported when sufficiently large datasets, relatively high-frequency observations, and informative exogenous variables are available. In contrast, other studies have indicated that in relatively short, strongly seasonal, or persistence-dominated time series, simpler models may remain highly competitive. Under such conditions, the additional complexity of deep learning models does not necessarily translate into improved out-of-sample forecasting accuracy.

This inconsistency in previous findings highlights several important limitations in the existing body of knowledge. First, the reported performance of deep learning models appears to be strongly dependent on data characteristics, including sample size, temporal resolution, seasonality strength, and the stability of the demand process. Second, complex models such as RNN and LSTM are more prone to overfitting when the available data are limited, particularly at monthly resolution, which is common in many real-world water management systems. Third, although multivariate forecasting frameworks often include climatic and socioeconomic variables, the practical added value of these variables has not been consistently demonstrated across studies. In some cases, such inputs improve forecasting skill, whereas in others they contribute little to no measurable gain, especially when persistence and seasonal repetition dominate the demand series.

Another important issue is that many previous studies have emphasized the application of advanced forecasting models without systematically benchmarking them against simple baseline methods. This is a critical limitation because methods such as Naïve and Seasonal Naïve forecasting often provide surprisingly strong performance in time series characterized by short-term persistence and regular seasonal structure. Without such benchmark comparisons, it is difficult to determine whether the reported success of a complex model reflects a substantial predictive advantage or performance that could already be achieved by simpler alternatives. Therefore, systematic benchmarking is necessary not only for performance evaluation but also for interpreting the practical usefulness of model complexity in specific forecasting contexts.

Despite the growing literature on urban water demand forecasting, a clear research gap remains regarding the empirical performance of widely used deep learning models for monthly forecasting in data-limited and strongly seasonal contexts, particularly in arid and water-stressed regions. In such environments, water demand may be influenced not only by climatic variability but also by managerial, institutional, and socioeconomic conditions, which further complicates model behavior and forecast interpretation. Accordingly, there is still a need for context-specific evaluation of whether models such as RNN and LSTM provide practical forecasting benefits when applied to such conditions and how their performance compares with that of simple benchmark methods.

To address these research gaps, this study investigates monthly potable water demand forecasting in the Sistan region of Iran using data from April 2006 to March 2021. For this purpose, two widely used deep learning architectures, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM), are evaluated under both univariate (demand only) and multivariate (demand combined with climatic and socioeconomic variables) settings and are systematically benchmarked against simple baseline forecasting methods, namely Naïve and Seasonal Naïve approaches. In this way, the study examines whether increased model complexity yields meaningful forecasting improvements under data-limited and strongly seasonal conditions. Model performance is assessed using several complementary metrics, including MAE, RMSE, MAPE, and Pearson's correlation coefficient.

The novelty of this study lies not in proposing a new forecasting architecture, but in providing a systematic and context-specific comparative evaluation of deep learning and simple benchmark models under data-limited and strongly seasonal conditions. In contrast to many previous studies that primarily emphasize methodological complexity, this research examines whether advanced models such as RNN and LSTM actually offer measurable predictive advantages over simple baseline approaches in a real-world urban water demand system. By focusing on the Sistan region, a water-scarce and climatically vulnerable area, the study provides practical evidence on the applicability and relative performance of these models under challenging hydrological and data conditions.

Accordingly, the main contributions of this study can be summarized as follows: (1) presenting a systematic benchmarking comparison between RNN and LSTM models and simple baseline methods for monthly urban water demand forecasting in a strongly seasonal and data-limited context; (2) evaluating the incremental predictive value of climatic and socioeconomic variables through both univariate and multivariate modelling frameworks; (3) providing clearer insights into the forecasting behavior of deep learning models relative to baseline performance under real-world monthly data conditions; and (4) contributing empirical evidence from the Sistan region, thereby helping to reduce the current regional knowledge gap regarding deep learning applications for urban water demand forecasting in arid and water-stressed environments.

2. Materials and methods

2.1. Case study

This study was conducted to examine the relationship between urban water consumption and climatic and socioeconomic factors and to develop suitable models for forecasting monthly urban water demand in the Sistan region of Iran. In recent years, Sistan has faced widespread droughts, a severe decline in the inflow of the Helmand River, and instability of both surface and groundwater resources, making water resources management one of the major challenges in this region.

Sistan is located in the northern part of Sistan and Baluchestan Province and lies on an extensive plain characterized by an arid and desert climate. The region experiences extremely hot summers, cold winters, and frequent dust and sandstorms. As one of the driest areas in the country, Sistan receives very low annual precipitation approximately 50 mm and exhibits extremely high evaporation rates, ranging between 4,000 and 5,000 mm per year. These climatic characteristics, combined with the region's heavy dependence on inflows from the Helmand River, have intensified the water crisis, groundwater depletion, recurring dust storms, and the deterioration of living conditions in recent years.

Given these conditions, accurate forecasting of urban water consumption is essential for supporting water allocation, infrastructure planning, and drought-risk management. The combination of low precipitation, high evaporation, strong climatic variability, and fluctuations in water availability makes Sistan an important case study for evaluating data-driven water-demand forecasting models.

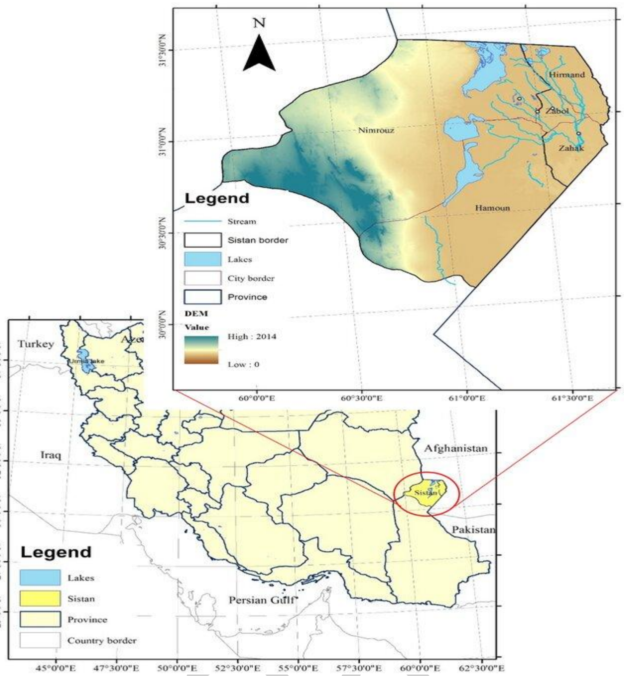


Fig 1. Geographical Location of the Sistan Region.

2.2. Dataset and variables

This study uses a monthly dataset of urban water consumption, climatic variables, and socioeconomic indicators for the Sistan region of Iran, covering the period from April 2006 to March 2021. Monthly water consumption data are reported in million cubic meters (Mm^3/month) and were compiled for the entire study period using official records from regional water authorities. Climatic variables were collected from the Sistan meteorological station and include maximum, minimum, and mean air temperature (°C), relative humidity (%), wind speed (m/s), sunshine duration (hours), and evaporation (mm). Climatic observations were aggregated to a monthly scale to ensure consistency with the water consumption data in addition, two socioeconomic variables, namely population and drinking water price, were incorporated to account for non-climatic influences on urban water demand. Population data were obtained from national statistical yearbooks, while drinking water price data were compiled from administrative records. Since these socioeconomic variables were available at an annual scale, they were assigned as constant monthly values within each corresponding year. The dataset exhibits pronounced seasonal variability. Peak water consumption generally occurs during the hot months from July to September, while the lowest consumption values are observed during the colder months from December to February. In addition to this seasonal behavior, an increasing trend in monthly water consumption is observed over the study period, reflecting growing pressure on the region's limited water resources and highlighting the importance of reliable demand forecasting.

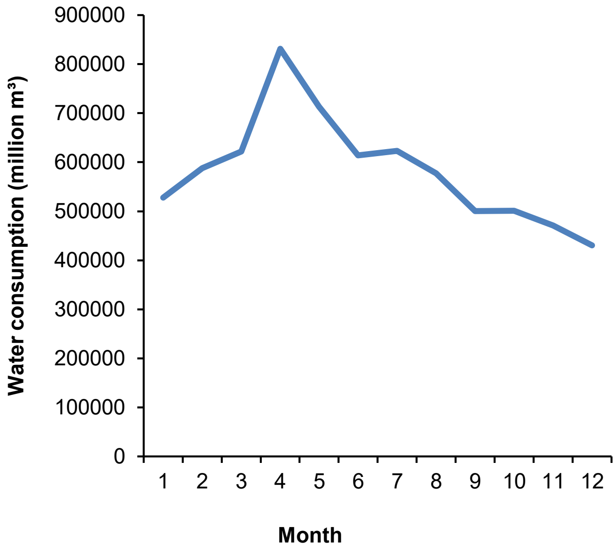


Fig. 2. Monthly Urban Water Consumption in the Sistan Region (2006–2021) (million m³).

Before model development, the dataset was checked for temporal consistency, completeness, and chronological ordering. All variables were aligned at the monthly time step to ensure consistency across the full study period.

2.3. Data pre-processing

Before model development, the dataset was checked for consistency, completeness, and temporal ordering. Since the study uses time-series data, all observations were arranged chronologically. The water-consumption series and all explanatory variables were aligned at the monthly time step.

Missing values were examined before model training. In cases where missing values occurred in climatic variables, they were treated using linear interpolation based on neighboring monthly observations. For annual socioeconomic variables, the corresponding annual value was assigned to all months of that year. No future observations were used during the preprocessing stage in order to avoid information leakage. Outliers and abnormal records were visually inspected using time-series plots. Since extreme values may reflect actual climatic or demand conditions in arid regions, no observation was removed solely on the basis of magnitude unless it was inconsistent with official records. This approach was adopted to preserve the natural variability of water demand in the study region.

To improve numerical stability during neural network training, all input variables were normalized. A min–max scaling procedure was applied using the following equation:

$$x = \frac{\acute{x} - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

where, x is the original value, x′ is the normalized value, and x_{min} and x_{max} are the minimum and maximum values calculated from the training data only. The same scaling parameters were then applied to the validation/test data. This procedure prevents information leakage from the validation/test period into the training process.

2.4. Supervised Time-Series Formulation

The forecasting task was formulated as a one-step-ahead monthly time-series prediction problem. To prepare the data for recurrent neural network modeling, the original time series was transformed into a supervised learning structure using a sliding-window approach. A 12-month lag window was selected as the input sequence. Therefore, the observations from the previous 12 months were used to forecast water demand in the following month. The 12-month window was chosen because the dataset is monthly and urban water consumption in the study area exhibits strong annual seasonality.

Using a 12-month input sequence allows the models to capture one complete seasonal cycle. For the univariate configuration, only past values of water consumption were used as input. For the multivariate configuration, climatic and socioeconomic variables were added using the same lag structure. Therefore, for each prediction step, the model received information from the previous 12 months and generated the water-demand estimate for the next month.

2.5. Data splitting and validation strategy

Since the dataset is time-dependent, a chronological splitting strategy was used instead of random sampling. The first 80% of observations were allocated to model training, while the remaining 20% were reserved for validation/testing. This time-ordered splitting strategy prevents future information from entering the training process and is consistent with standard practice in time-series forecasting. The validation period was used to evaluate the out-of-sample forecasting performance of the models. Because the available dataset is relatively limited, the study focused on a hold-out validation strategy rather than repeated random resampling. This limitation is acknowledged, and future studies may apply rolling-origin or walk-forward validation to further assess the robustness of forecasting performance.

2.6. Model development

In this study, two recurrent neural network-based models were developed and compared: a standard Recurrent Neural Network and a Long Short-Term Memory network. Both models were implemented for univariate and multivariate forecasting settings. The forecasting horizon was one month ahead.

Both the RNN and LSTM models consisted of one recurrent hidden layer with 48 units, followed by a fully connected output layer with one neuron. The output neuron produced the predicted water demand for the next month. A dropout layer with a rate of 0.2 was included to reduce overfitting. The models were trained using the Adam optimization algorithm and the mean squared error loss function. The maximum number of epochs was set to 400, and early stopping was applied based on validation loss with a patience of 20 epochs. The batch size was set to 8.

The selection of model hyperparameters was based on the characteristics of the dataset and empirical testing over a limited set of candidate values. The 12-month input window was selected to represent annual seasonality. The number of hidden units was kept moderate to balance the ability of the models to learn temporal patterns with the risk of overfitting due to the relatively small sample size. The dropout rate of 0.2 was selected as a commonly used regularization value and was also found to provide stable validation performance during preliminary model tuning. A small batch size of 8 was used because of the limited number of monthly observations and to allow more frequent weight updates during training.

To reduce the influence of random initialization, a fixed random seed of 42 was used for all stochastic components involved in model initialization and training. In addition, the same data split, preprocessing steps, input window, and evaluation metrics were applied to all models to ensure comparability.

2.7. Hyperparameter selection and overfitting control

Hyperparameter selection was performed through a limited empirical tuning process. Candidate values were tested for the main model settings, including the input window length, number of hidden units, dropout rate, and batch size. The final configuration was selected based on validation performance and model stability, while also considering the limited size of the dataset. The selected configuration was as follows:

Table 1. Selected hyperparameters and rationale for the Deep Learning Models.

Hyperparameter	Selected value	Rationale
Input window length	12 months	Captures one full annual seasonal cycle
Hidden units	48	Provides moderate model capacity while limiting overfitting risk
Dropout rate	0.2	Regularization to reduce overfitting
Batch size	8	Suitable for small monthly dataset
Optimizer	Adam	Efficient gradient-based optimization
Maximum epochs	400	Allows sufficient training iterations
Early stopping patience	20	Stops training when validation performance no longer improves

Overfitting was controlled through several mechanisms. First, the model architecture was intentionally kept simple by using only one recurrent hidden layer. Second, dropout regularization was applied during training. Third, early stopping was used to stop training when validation loss did not improve. Fourth, the number of hidden units was kept moderate relative to the dataset size. Finally, performance was assessed on a chronologically separated validation/test period to evaluate out-of-sample forecasting ability.

2.3. Recurrent Neural Network (RNN) model

Recurrent Neural Networks (RNNs) are deep-learning architectures specifically designed for sequential and time-series data. Unlike feed-forward neural networks, RNNs include recurrent connections that allow information from previous time steps to influence the current state and output. Recurrent Neural Networks (RNNs), also known as recursive neural networks, are one of the key deep-learning architectures for analyzing sequential and time-series data. Unlike feed-forward networks, in which information flows only in one direction from input to output, RNNs contain internal feedback loops whereby the output of the previous step is fed back into the network along with the new input. This recurrent structure allows the network to maintain an internal memory and temporarily store information from previous inputs; therefore, RNNs can identify temporal patterns in sequential data and incorporate historical dependencies into the prediction process (Selvin *et al*., 2017). At each time step, the hidden state of a standard RNN can be expressed as:

$$h_t = \phi(w_x x_t + w_h h_{t-1} + b_h) \quad (2)$$

where, x_t is the input vector at time t, h_t is the hidden state, $h_{t-1}$is the previous hidden state, w_x and w_h are weight matrices, b_h is the bias vector, and ϕ is a nonlinear activation function. The output of the network is calculated as:

$$\hat{y}_i = W_y h_t + b_y \quad (3)$$

where, $\hat{y}_i$ is the predicted value, W_y is the output weight matrix, and b_y is the output bias. Although RNNs are effective for capturing short-term temporal dependencies, they may suffer from vanishing-gradient problems, which can limit their ability to learn long-term dependencies. This limitation has motivated the development of advanced recurrent architectures such as LSTM networks.

2.4. Long Short-Term Memory (LSTM) Model

Long Short-Term Memory (LSTM) networks are one of the most widely used types of recurrent neural networks for analyzing and forecasting time-series data. Their major advantage lies in their ability to preserve and transmit important information over long periods, thereby overcoming common RNN problems such as vanishing and exploding gradients (Du *et al*., 2021). In the LSTM architecture, a long-term information transfer path known as the *cell state* acts as the main memory of the network, while the flow of information into, within, and out of the memory is regulated by nonlinear gates.

Each memory block contains three main gates: the input gate, the forget gate, and the output gate. The inputs at each time step consist of the input vector x_t and the previous output h_{t-1}. First, these two vectors are processed by the forget gate to determine which information from the previous state should be retained or discarded:

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (4)$$

Next, the input gate determines the amount of new information that should enter the cell state. This gate consists of a sigmoid layer that regulates input weights (i_t) and a tanh layer that generates the candidate cell state ($\acute{c}_t$):

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\acute{c}_t = \tanh(W_c . [h_{t-1}, x_t] + b_c) \quad (6)$$

The new cell state is then computed by combining the previous cell state with the new candidate information:

$$c_t = f_t \odot c_{t-1} + i_t \odot \acute{c}_t \quad (7)$$

Finally, the output gate generates the new output h_t using the cell state and the gate activation:

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \odot \tanh(c_t) \tag{9}$$

here, W_f ,$W_c W_o W_i$ are weight matrices and b_o, b_c, b_f, b_i are bias components optimized during the training process. The goal of the network is to minimize the loss function. RNN and LSTM were selected as representative recurrent forecasting models to evaluate their practical usefulness for monthly potable water demand prediction, rather than to propose a new neural network architecture.

2.5. Evaluation metrics

Each model is evaluated using several performance metrics, and model comparison is conducted to identify the best long-term water-demand forecasting model for the study area. Detailed descriptions of each metric are presented below. Mean Absolute Percentage Error (MAPE) This metric evaluates the predictive accuracy of the model based on the mean absolute deviation between observed and predicted values, standardized by the observed values:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{10}$$

Mean Absolute Error (MAE)
MAE measures the average magnitude of the prediction errors in a dataset, regardless of their direction. It represents the mean of absolute differences between predicted and actual values, with all individual errors weighted equally (Kang *et al.*, 2015):

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{11}$$

Root Mean Square Error (RMSE)
RMSE provides information about the short-term performance of the model and compares the actual and estimated values term by term:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{12}$$

where, n is the number of observations$\hat{y}_i$is the predicted value, and $y_i$is the actual value (Kang *et al.*, 2015).
Pearson's Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{13}$$

where, y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of observed values, $\bar{\hat{y}}$ is the mean of predicted values, and n is the number of observations. Higher correlation values indicate stronger agreement between observed and predicted temporal patterns.

To assess whether differences in forecast accuracy were statistically significant, we applied the Diebold–Mariano (DM) test with the Harvey–Leybourne–Newbold small-sample correction, using squared forecast errors (MSE loss) and a one-step-ahead forecast horizon (h=1).

2.11. Computational environment

All preprocessing, model development, training, and evaluation procedures were implemented in Python using commonly adopted scientific computing libraries. Neural network models were developed using TensorFlow/Keras, while NumPy and Pandas were used for numerical analysis and data handling. Additional preprocessing and evaluation support were provided through Scikit-learn. All models were trained under the same computational settings and data partitions to ensure fair comparison and reproducibility.

2.12. Methodological limitations

The present study focuses on deterministic one-step-ahead forecasting of monthly urban water demand. Therefore, prediction intervals and formal uncertainty quantification were not included. This is acknowledged as a limitation of the present work, since uncertainty information can be valuable for water-resources planning under risk. In addition, formal statistical significance testing between competing forecasting models was not conducted. Because the available dataset consists of a single monthly time series with a limited number of observations, model comparison was based primarily on standard error metrics and out-of-sample validation performance. Future studies may extend the analysis using probabilistic forecasting frameworks, sensitivity analysis, and statistical comparison tests such as the Diebold–Mariano test.

3. Results and discussion

In this section, the performance of different models used for monthly water consumption forecasting is evaluated to determine which approach most accurately represents the observed consumption patterns. Given that urban water demand is influenced by climatic conditions, behavioral factors, and strong seasonal effects, comparing simple time-series benchmarks based solely on historical persistence with more complex deep-learning architectures provides valuable insight into the inherent predictability of the series. Accordingly, model performance is assessed using multiple complementary evaluation metrics, including MAE, RMSE, MAPE, and the Pearson correlation coefficient, for both deep-learning models (RNN and LSTM) and baseline approaches (Naive and Seasonal Naive).

Table 2. Evaluation results of water consumption prediction models using the metrics Pearson_r, MAE, RMSE, and MAPE.

Model	MAPE_%	RMSE (billion m3)	MAE (billion m3)	Pearson_r
RNN_UV	21	0.185	0.158	-0.78
RNN_MV	20	0.191	0.157	0.63
LSTM_UV	22	0.191	0.162	0.71
LSTM_MV	20	0.186	0.153	0.76
Naive	11	0.10	0.075	0.74
SeasonalNaive	14	0.11	0.088	0.67

Table 3. Summary of Diebold–Mariano test results for benchmark versus deep learning models.

Comparison	DM statistic	p-value	Result
LSTM-MV vs Naïve	3.6145	0.0009	Naïve significantly better
LSTM-UV vs Naïve	3.6226	0.0009	Naïve significantly better
RNN-MV vs Naïve	3.5529	0.0011	Naïve significantly better
RNN-UV vs Naïve	3.4518	0.0015	Naïve significantly better
LSTM-MV vs Seasonal Naïve	3.0865	0.0039	Seasonal Naïve significantly better
LSTM-UV vs Seasonal Naïve	3.1275	0.0035	Seasonal Naïve significantly better
RNN-MV vs Seasonal Naïve	3.0634	0.0042	Seasonal Naïve significantly better
RNN-UV vs Seasonal Naïve	2.9447	0.0057	Seasonal Naïve significantly better
Naïve vs Seasonal Naïve	-0.6248	0.5361	No significant difference

The results reveal notable differences in predictive behavior among the evaluated models. Overall, the benchmark persistence-based approaches achieved lower prediction errors than the deep-learning models under the characteristics of the present dataset. In particular, the Naive model produced the lowest error values across all error-based metrics (MAE = 0.075, RMSE = 0.10, MAPE = 11%), indicating strong short-term persistence and regular seasonal structure in the monthly water consumption series. The Seasonal Naive model also demonstrated competitive performance and ranked among the best-performing approaches. Among the deep-learning architectures, LSTM generally performed better than RNN across most evaluation metrics. In addition, the multivariate LSTM configuration (LSTM_MV) showed modest improvements compared with the univariate version (LSTM_UV), reflected by a higher Pearson correlation coefficient and slightly lower prediction errors. This suggests that climatic and socioeconomic variables contain supplementary predictive information that can partially improve temporal pattern representation within the LSTM framework, although the magnitude of improvement remains limited. The RNN models exhibited comparatively lower predictive capability. Although the multivariate configuration improved upon the univariate RNN model, its overall performance remained less competitive than both the LSTM and benchmark persistence-based approaches.

These findings may indicate that the temporal structure of the dataset is relatively stable and strongly seasonal, reducing the practical advantage of more complex recurrent architectures under limited data conditions. From a trend representation perspective, the LSTM_MV model achieved the highest Pearson correlation coefficient (r = 0.76), suggesting a stronger ability to capture the general temporal evolution of water consumption. Nevertheless, its error-based metrics remained inferior to those of the Naive benchmark. This observation highlights that stronger correlation with observed trends does not necessarily imply lower numerical forecasting error, particularly in monthly series characterized by seasonality and persistence.

Overall, the findings emphasize the importance of systematic benchmarking in water demand forecasting studies. The results suggest that, in strongly seasonal and data-constrained environments, simple baseline approaches can remain highly effective, while more advanced deep-learning models may provide added value primarily in representing broader temporal dynamics rather than minimizing short-term prediction errors.

Table 3 reports the summary results of the Diebold–Mariano test for pairwise comparisons between the benchmark models and the deep learning models. The results show a clear and consistent pattern: both benchmark models, Naïve and Seasonal Naïve, significantly outperformed all deep learning specifications in terms of forecast accuracy.

Specifically, the Naïve model was significantly more accurate than LSTM-MV, LSTM-UV, RNN-MV, and RNN-UV, with p-values ranging from 0.0009 to 0.0015. Likewise, the Seasonal Naïve model also significantly outperformed all four deep learning models, with p-values between 0.0035 and 0.0057. These findings indicate that the superior performance of the benchmark models is not limited to descriptive error measures such as RMSE, but is also statistically supported by pairwise predictive-accuracy tests.

At the same time, the comparison between Naïve and Seasonal Naïve did not reveal a statistically significant difference (p=0.5361) (p = 0.5361) (p=0.5361). This suggests that although the Naïve model achieved the best overall error performance, its advantage over the seasonal benchmark is not large enough to be distinguished statistically at conventional significance levels.

Overall, the findings imply that the added complexity of deep learning architectures did not translate into superior predictive performance for this dataset. A plausible explanation is that the water-demand series in the study area is relatively stable and dominated by strong short-term persistence and recurring temporal structure, conditions under which simple benchmark models can perform very effectively. From an applied perspective, this result is particularly important because it suggests that, for operational water-demand forecasting in the Sistan region, parsimonious benchmark models may provide forecasts that are not only easier to implement and interpret, but also statistically more reliable than more complex neural-network alternatives.

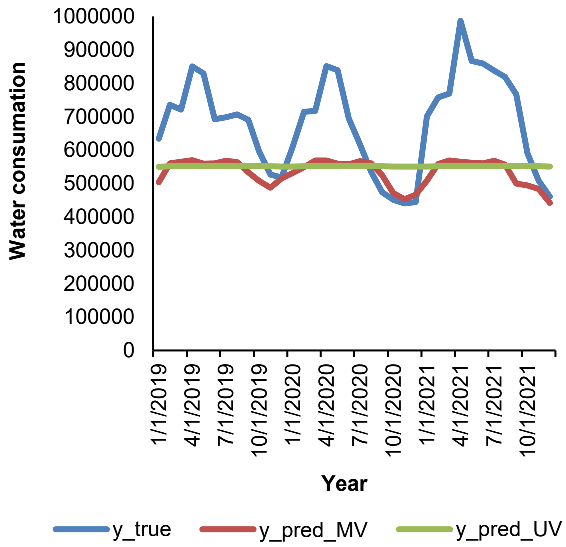


Fig 3. Actual and Validation Forecasts of the Univariate and Multivariate LSTM Models.

Fig. 3 presents a comparison between observed monthly water consumption and forecasts generated by the univariate (UV-LSTM) and multivariate (MV-LSTM) models over the validation period. Both models are able to reproduce the general seasonal structure of the series, indicating that the LSTM architecture captures the dominant temporal patterns in urban water demand.

The UV-LSTM produces smooth predictions fluctuating around a nearly constant mean level. Although the model preserves the overall direction of long-term changes, it substantially underestimates the magnitude of seasonal peaks and troughs. This behavior reflects the model's reliance solely on historical consumption data, which limits its ability to respond to abrupt increases or decreases in demand. The MV-LSTM shows more dynamic behavior and tracks observed variability more closely in several months. Incorporating climatic and explanatory variables improves the model's responsiveness to seasonal and inter-annual fluctuations, resulting in forecasts that are visually closer to the observed values than those of the univariate model. However, pronounced peaks and sharp declines remain underestimated, suggesting that water consumption dynamics are influenced by additional behavioral, institutional, and household-level factors that are not captured in the available dataset.

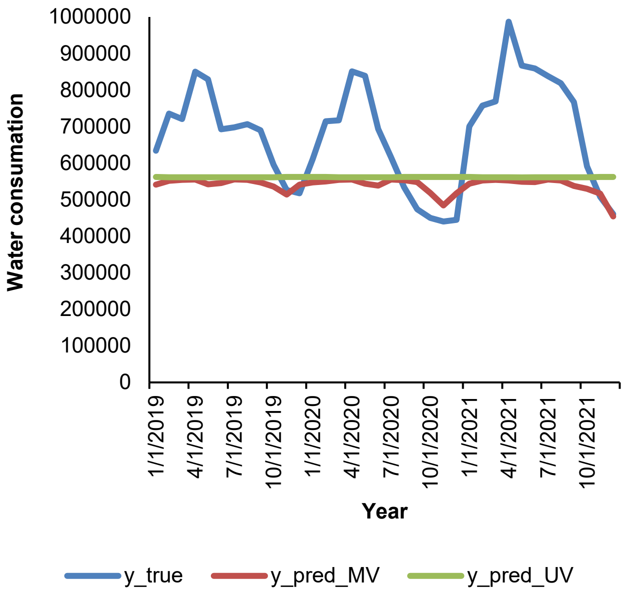


Fig. 4. Actual and Validation Forecasts of the Univariate and Multivariate LSTM Models RNN Models.

The results of the univariate and multivariate RNN models show that both models capture the general trend of water consumption over the validation period, but their accuracy in reproducing actual values especially during months with sharp consumption fluctuations differs. As illustrated in the table and comparative plots, both RNN models underestimate water consumption across a significant portion of the validation period. This indicates that RNN tends to capture the "mean pattern" of the series while struggling with extreme peaks or sudden drops.

The univariate model (RNN-UV), despite having lower error in certain months, performs poorly in terms of Pearson correlation. The negative correlation value suggests that the model fails to track simultaneous increases and decreases in consumption and instead produces nearly constant predictions around an average level. This behavior is visually evident in the flattened, low-variance prediction line in the plot.

The multivariate RNN model (RNN-MV), while having MAE and RMSE values similar to the univariate model, achieves better correlation and partially captures relative changes in consumption. The presence of additional explanatory variables increases the model's sensitivity to month-to-month variations, improving its alignment with actual consumption in several periods. Nevertheless, the predicted range remains narrower than the true range, and the model struggles to reproduce high peaks or sharp drops.

This behavior reflects the structural limitations of RNNs: these models tend to smooth the output and revert predictions toward the mean when faced with large fluctuations. This is particularly problematic in months with unusually high consumption (e.g., spring peaks) or periods with substantial declines. Consequently, although adding explanatory variables improves the model's responsiveness, RNN remains inferior to more advanced architectures such as LSTM in capturing complex temporal dependencies.

In summary, RNN models adequately reproduce the overall trend of water consumption but lack precision in forecasting fluctuations and consumption peaks. Therefore, for applications requiring higher accuracy in variability and extreme values, deeper architectures or hybrid approaches offer better alternatives.

The results of the present study indicate that the LSTM model did not outperform simpler benchmark models for forecasting water demand. A similar observation has been reported in several forecasting studies where model performance strongly depends on the characteristics of the data. For instance, Makridakis, Spiliotis and Assimakopoulos, (2018), in the large-scale M4 forecasting competition, demonstrated that more complex models do not necessarily yield superior forecasting accuracy compared with simpler statistical approaches. Likewise, Hewamalage, Bergmeir and Bandara, (2021) reported in their comprehensive review of deep learning methods for time-series forecasting that classical models such as ARIMA, ETS, and naïve benchmarks can often provide competitive performance, particularly when the underlying time series exhibits relatively stable patterns. In hydrological applications, the effectiveness of LSTM models has also been shown to depend on data characteristics and the availability of informative inputs. Kratzert *et al.* (2019) demonstrated the potential of LSTM models for hydrological prediction when sufficient data and relevant predictors are available, while Gauch *et al.* (2021) emphasized that the predictive skill of LSTM models varies with temporal scale and dataset properties. Furthermore, in contrast to the amoxicillin demand forecasting study where LSTM combined with temperature information outperformed seasonal naïve and statistical baselines, the findings of the present study suggest that for water-demand data with a relatively stable temporal structure and limited complexity, simpler benchmark models may provide more reliable forecasting performance.

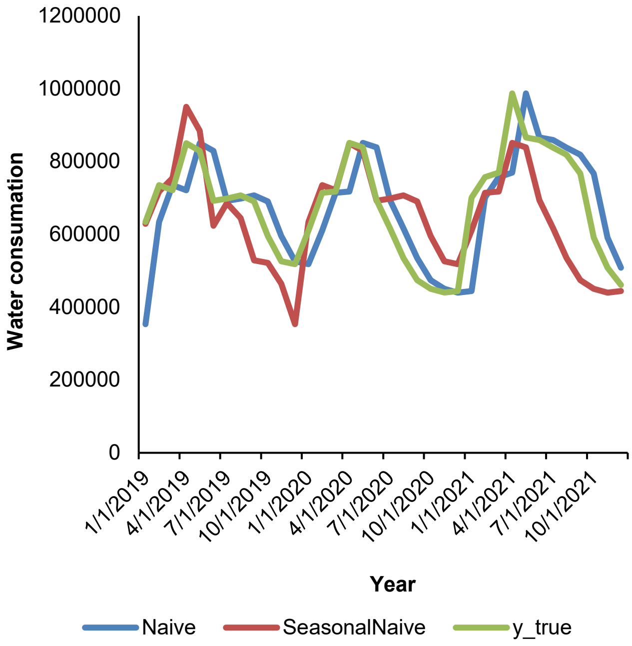


Fig. 5. Actual and validation forecasts for the Naïve and Seasonal Naïve Models.

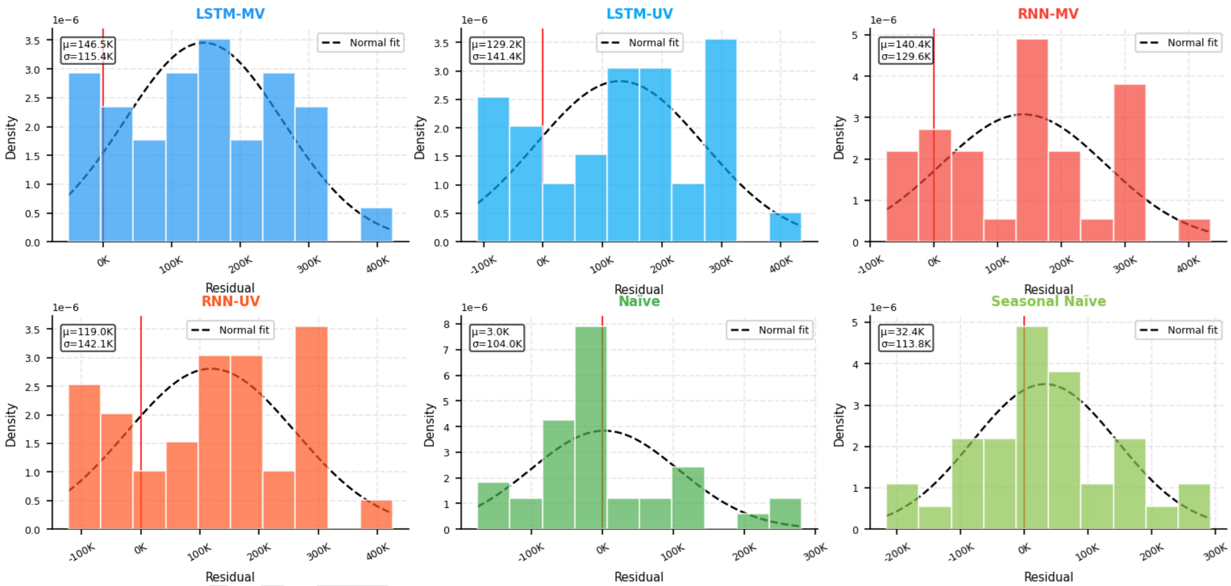


Fig. 6. Histogram of Residuals with normal fit.

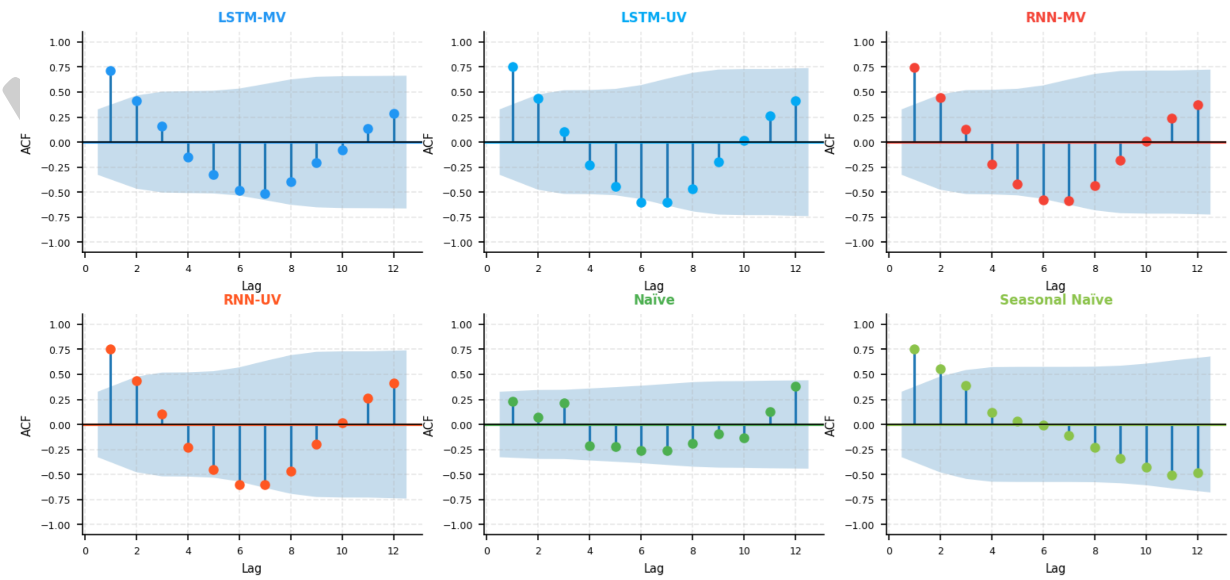


Fig. 7. Autocorrelation Function (ACF) of Residuals.

Fig. 6 shows that all DL models have high positive mean residuals (μ≈119K-146K), indicating a systematic Underprediction Bias the model consistently underestimates actual values. The error distribution in these models is right-skewed and fits poorly to a normal curve. In contrast, the Naïve model with μ≈3K is nearly unbiased, and its distribution is closer to normal. Seasonal Naïve with μ≈32K also outperforms the DL models in terms of bias. And Figure 7 shows that Deep Learning models (LSTM and RNN) exhibit high autocorrelation at lag 1 (~0.75), indicating that consecutive errors are not independent. A notable negative spike is also visible at lag 6, suggesting a semi-annual seasonal pattern that the DL models have not fully captured. In contrast, Naïve and Seasonal Naïve models show weaker autocorrelation, though Seasonal Naïve still exhibits significant correlation at lags 1 and 2. These patterns confirm a violation of the i.i.d. error assumption.

4. Conclusions

This study evaluated the performance of recurrent deep learning models, including RNN and LSTM, for forecasting monthly urban potable water consumption in the Sistan region. The results showed that although these models were able to capture the general temporal pattern of water demand, they did not consistently outperform simpler benchmark methods. In particular, the Naïve and Seasonal Naïve models achieved the best overall performance in terms of error-based metrics, reflecting the strong persistence and annual seasonality of the monthly consumption series. Among the deep learning models, LSTM performed more robustly than the standard RNN and produced smoother and more consistent forecasts. However, its advantage was more evident in representing the general trend and seasonal behavior than in reducing forecasting errors. Both RNN and LSTM tended to underestimate abrupt peaks and troughs, suggesting that recurrent models may require longer time series and richer explanatory information to capture short-term fluctuations more accurately. The comparison between univariate and multivariate models indicated that adding climatic and socioeconomic variables did not lead to consistent improvements in forecasting accuracy. This finding suggests that the available explanatory variables did not fully represent the main drivers of potable water consumption in the study area. In a water-stressed region such as Sistan, urban water demand may also be influenced by operational, institutional, supply-side, and policy-related factors that were not included in the dataset. A key limitation of this study is the use of monthly data over a 16-year period, which provided 192 observations. This sample size is relatively limited for data-intensive deep learning architectures such as LSTM. To reduce the risk of overfitting, model complexity was deliberately constrained, and regularization techniques were applied during training. In addition, the statistical significance of performance differences was assessed using the Diebold–Mariano test with small-sample correction. Nevertheless, the limited sample size restricts the generalizability of the findings, and the results should be interpreted within the specific data structure, validation framework, and regional context of this study. Overall, the findings emphasize that greater model complexity does not necessarily guarantee better predictive performance, especially when the available time series is short, highly seasonal, and affected by external management constraints. For data-limited regions, simple and transparent benchmark models can provide competitive and practical forecasting performance. At the same time, advanced deep learning models may become more effective when longer historical records, higher-resolution data, more informative explanatory variables, and detailed operational data are available. The practical implication of this study is that water demand forecasting in arid and data-limited regions should be based on systematic benchmark comparison rather than an assumption that advanced models are inherently superior. Reliable forecasts can support urban water planning, demand management, drought preparedness, and supply–demand balancing in the Sistan region. Future research should extend the temporal coverage of the dataset, examine daily or weekly consumption data, incorporate operational and policy-related variables, assess possible structural changes in water demand, and compare a broader range of statistical, machine-learning, hybrid, and deep learning models.

Conflict of Interest

The authors declare no competing interests and non-financial competing interests.

Acknowledgment

The authors would like to acknowledge the Research Institute of Zabol, Zabol, Iran, for its financial support of this research under Grant No. PRRIOZ-1403–0702-1.

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

Zahra Ghaffari Moghadam: Conceptualization, methodology, data curation, formal analysis, visualization, writing-original draft, writing-review and editing, and project administration.
Ali Sardar Shahraki: Supervision, conceptualization, methodology, validation, interpretation of results, and writing-review and editing.
Abdulhasseb Azizi: Model implementation, formal analysis, validation, interpretation of results, and writing-review and editing.
Somayyeh Mirshekari: Data collection, data preparation, literature review, data curation, visualization

References

Al-Zahrani, M.A. and Abo-Monasar, A. (2015) 'Urban residential water demand prediction based on artificial neural networks and time series models', *Water Resources Management*, 29(10), pp. 3651–3662. doi: https://doi.org/10.1007/s11269-015-1021-z

Barbieri, M. *et al.* (2023) 'Climate change and its effect on groundwater quality', *Environmental Geochemistry and Health*, 45(4), pp. 1133–1144. doi: https://doi.org/10.1007/s10653-021-01140-5

Boudhaouia, A. and Wira, P. (2021) 'A real-time data analysis platform for short-term water consumption forecasting with machine learning', *Forecasting*, 3(4), pp. 682–694. doi: https://doi.org/10.3390/forecast3040042

Choi, J. and Kim, J. (2018) 'Analysis of water consumption data from smart water meter using machine learning and deep learning algorithms', *Journal of the Institute of Electronics and Information Engineers*, 55, pp. 31–39. doi: https://doi.org/10.5573/ieie.2018.55.7.31

Daw, M.M., Ali, E.R. and Toriman, M.E. (2021) 'Determinants of urban residential water demand in Libya', *International Journal of Innovation and Sustainable Development*, 15(3), pp. 261–279. doi: https://doi.org/10.1504/IJISD.2021.115963

Du, B. et al. (2021) 'Deep learning with long short-term memory neural networks combining wavelet transform and principal component analysis for daily urban water demand forecasting', *Expert Systems with Applications*, 171, p. 114571. doi: https://doi.org/10.1016/j.eswa.2021.114571

Ercin, A.E. and Hoekstra, A.Y. (2014) 'Water footprint scenarios for 2050: A global analysis', *Environment International*, 64, pp. 71–82. doi: https://doi.org/10.1016/j.envint.2013.11.019

Gauch, M. *et al.* (2021) 'Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network', *Hydrology and Earth System Sciences*, 25, pp. 2045–2062. doi: https://doi.org/10.5194/hess-25-2045-2021

Hewamalage, H., Bergmeir, C. and Bandara, K. (2021) 'Recurrent neural networks for time series forecasting: Current status and future directions', *International Journal of Forecasting*, 37(1), pp. 388–427. doi: https://doi.org/10.1016/j.ijforecast.2020.06.008

Hu, P. *et al.* (2019) 'A hybrid model based on CNN and Bi-LSTM for urban water demand prediction', in *2019 IEEE Congress on Evolutionary Computation (CEC)*. Wellington, New Zealand, 10–13 June 2019. IEEE, pp. 1088–1094. doi: https://doi.org/10.1109/CEC.2019.8790060

Iwakin, O. and Moazeni, F. (2024) 'Improving urban water demand forecast using conformal prediction-based hybrid machine learning models', *Journal of Water Process Engineering*, 58, pp. 104721. doi: https://doi.org/10.1016/j.jwpe.2023.104721.

Kang, H.S. *et al.* (2015) 'Optimization of pumping schedule based on water demand forecasting using a combined model of autoregressive integrated moving average and exponential smoothing', *Water Science and Technology: Water Supply*, 15(1), pp. 188–195. doi: https://doi.org/10.2166/ws.2014.104

Kim, J. *et al.* (2022) 'Development of a deep learning-based prediction model for water consumption at the household level', *Water*, 14(9), pp. 1512. doi: https://doi.org/10.3390/w14091512

Kratzert, F. *et al.* (2019) 'Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks', *Hydrology and Earth System Sciences*, 23(12), pp. 5089–5110. doi: https://doi.org/10.5194/hess-23-5089-2019

Li, C. *et al.* (2021) 'Determinants of agricultural water demand in China', *Journal of Cleaner Production*, 288, pp. 125508. doi: https://doi.org/10.1016/j.jclepro.2020.125508

Li, W. *et al.* (2024) 'An interpretable hybrid deep learning model for flood forecasting based on Transformer and LSTM', *Journal of Hydrology: Regional Studies*, 54, pp. 101873. doi: https://doi.org/10.1016/j.ejrh.2024.101873

Lim, B. and Zohren, S. (2021) 'Time-series forecasting with deep learning: A survey', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), pp. 20200209. doi: https://doi.org/10.1098/rsta.2020.0209

Liu, Y. *et al.* (2025) 'Timer-XL: Long-context transformers for unified time series forecasting', in *International Conference on Learning Representations*. pp. 1–25. Available at: https://github.com/thuml/Timer-XL (Accessed date: 5 May 2025).

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2018) 'The M4 Competition: Results, findings, conclusion and way forward', *International Journal of Forecasting*, 34(4), pp. 802–808. doi: https://doi.org/10.1016/j.ijforecast.2018.06.001

Mehrazar, A. *et al.* (2020) 'Adaptation of water resources system to water scarcity and climate change in the suburb area of megacities', *Water Resources Management*, 34(12), pp. 3855–3877. doi: https://doi.org/10.1007/s11269-020-02648-8

Mu, L. *et al.* (2020) 'Hourly and daily urban water demand predictions using a long short-term memory-based model', *Journal of Water Resources Planning and Management*, 146(9), pp. 05020017. doi: https://doi.org/10.1061/(ASCE)WR.1943-5452.0001276

Niknam, A. *et al.* (2023) 'Developing an LSTM model to forecast the monthly water consumption according to the effects of the climatic factors in Yazd, Iran', *Journal of Engineering Research*, 11(1), pp. 100028. doi: https://doi.org/10.1016/j.jer.2023.100028

Nunes Carvalho, T.M., de Souza Filho, F.D.A. and Porto, V.C. (2021) 'Urban water demand modeling using machine learning techniques: Case study of Fortaleza, Brazil', *Journal of Water Resources Planning and Management*, 147(1), pp. 05020026. doi: https://doi.org/10.1061/(ASCE)WR.1943-5452.0001310

Panagopoulos, G.P. (2014) 'Assessing the impacts of socio-economic and hydrological factors on urban water demand: A multivariate statistical approach', *Journal of Hydrology*, 518, pp. 42–48. doi: https://doi.org/10.1016/j.jhydrol.2013.10.036

Pu, Z. *et al.* (2023) 'A hybrid Wavelet-CNN-LSTM deep learning model for short-term urban water demand forecasting', *Frontiers of Environmental Science & Engineering*, 17(2), pp. 22. doi: https://doi.org/10.1007/s11783-023-1622-3

Qiao, J. *et al.* (2024) 'Attention-based spatiotemporal graph fusion convolution networks for water quality prediction', *IEEE Transactions on Automation Science and Engineering*, 22, pp. 1-10. https://doi.org/10.1109/TASE.2023.3285253

Salehi, M. (2022) 'Global water shortage and potable water safety: Today's concern and tomorrow's crisis', *Environment International*, 158, p. 106936. doi: https://doi.org/10.1016/j.envint.2021.106936

Salloom, T., Kaynak, O. and He, W. (2021) 'A novel deep neural network architecture for real-time water demand forecasting', *Journal of Hydrology*, 599, pp. 126353. doi: https://doi.org/10.1016/j.jhydrol.2021.126353

Sanchez, G.M. *et al.* (2020) 'Forecasting water demand across a rapidly urbanizing region', *Science of the Total Environment*, 730, p. 139050. doi: https://doi.org/10.1016/j.scitotenv.2020.139050

Selvin, S. *et al.* (2017) 'Stock price prediction using LSTM, RNN and CNN-sliding window model', in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Udupi, India, 13–16 September 2017. pp. 1643–1647. doi: https://doi.org/10.1109/ICACCI.2017.8126078

Velasco, L.C.P. *et al.* (2018) 'Performance analysis of artificial neural networks training algorithms and transfer functions for medium-term water consumption forecasting', *International Journal of Advanced Computer Science and Applications*, 9(4), pp. 109–116. doi: https://doi.org/10.14569/IJACSA.2018.090419

Xiang, Z., Yan, J. and Demir, I. (2020) 'A rainfall–runoff model with LSTM-based sequence-to-sequence learning', *Water Resources Research*, 56(1), pp. e2019WR025326. doi: https://doi.org/10.1029/2019WR025326

Zanfei, A. *et al.* (2022) 'Graph convolutional recurrent neural networks for water demand forecasting', *Water Resources Research*, 58(7), p. e2022WR032299. doi: https://doi.org/10.1029/2022WR032299.

Zanfei, A. *et al.* (2022) 'A short-term water demand forecasting model using multivariate long short-term memory with meteorological data', *Journal of Hydroinformatics*, 24(5), pp. 1053–1065. doi: https://doi.org/10.2166/hydro.2022.055