water

MDPI

Article

A Water Demand Forecasting Model Based on Generative Adversarial Networks and Multivariate Feature Fusion

Changchun Yang [1], Jiayang Meng [1], Banteng Liu [2], Zhangquan Wang [2] and Ke Wang [2,*]

[1] School of Microelectronics and Control Engineering, Changzhou University, Changzhou 213164, China; ycc@cczu.edu.cn (C.Y.); s22060809021@smail.cczu.edu.cn (J.M.)
[2] College of Information Science and Technology, Zhejiang Shuren University, Hangzhou 310015, China; 600932@zjsru.edu.cn (B.L.); 600389@zjsru.edu.cn (Z.W.)
* Correspondence: wangke1992@zju.edu.cn

Abstract: Accurate long-term water demand forecasting is beneficial to the sustainable development and management of cities. However, the randomness and nonlinear nature of water demand bring great challenges to accurate long-term water demand forecasting. For accurate long-term water demand forecasting, the models currently in use demand the input of extensive datasets, leading to increased costs for data gathering and higher barriers to entry for predictive projects. This situation underscores the pressing need for an effective forecasting method that can operate with a smaller dataset, making long-term water demand predictions more feasible and economically sensible. This study proposes a framework to delineate and analyze long-term water demand patterns. A forecasting model based on generative adversarial networks and multivariate feature fusion (the water demand forecast-mixer, WDF-mixer) is designed to generate synthetic data, and a gradient constraint is introduced to overcome the problem of overfitting. A multi-feature fusion method based on temporal and channel features is then derived, where a multi-layer perceptron is used to capture temporal dependencies and non-negative matrix decomposition is applied to obtain channel dependencies. After that, an attention layer receives all those features associated with the water demand forecasting, guiding the model to focus on important features and representing correlations across them. Finally, a fully connected network is constructed to improve the modeling efficiency and output the forecasting results. This approach was applied to real-world datasets. Our experimental results on four water demand datasets show that the proposed WDF-mixer model can achieve high forecasting accuracy and robustness. In comparison to the suboptimal models, the method introduced in this study demonstrated a notable enhancement, with a 62.61% reduction in the MSE, a 46.85% decrease in the MAE, and a 69.15% improve in the R^2 score. This research could support decision makers in reducing uncertainty and increasing the quality of water resource planning and management.

Keywords: smart water management; water demand forecasting; time series forecasting; long-term forecasting; multi-feature fusion; data enhancement

Citation: Yang, C.; Meng, J.; Liu, B.; Wang, Z.; Wang, K. A Water Demand Forecasting Model Based on Generative Adversarial Networks and Multivariate Feature Fusion. *Water* **2024**, *16*, 1731. https://doi.org/10.3390/w16121731

Academic Editor: Stefano Alvisi

Received: 28 May 2024
Revised: 14 June 2024
Accepted: 15 June 2024
Published: 19 June 2024

1. Introduction

With the development of the social economy, the contradiction between supply of and demand for urban water resources is increasing. The forecasting of water demand is of profound significance in urban water management planning and urban comprehensive development planning [1,2]. The study of water demand forecasting is an important component of and a fundamental prerequisite to ensuring the optimal operation of water supply networks [3,4]. By accurately forecasting changes in water demand at the water supply network's nodes, we can better plan and manage the operation of urban water supply and drainage pipe networks, improve the safety and reliability of the pipe network, reduce leakage and waste, and ensure the stability of water supply.

The most common methods for water demand forecasting include regression analysis, traditional time series analysis, and artificial neural networks. Regression analysis methods

forecast future water demand by establishing a regression model between water demand and relevant influencing factors [5,6]. However, due to the highly nonlinear nature of water demand sequences, it is difficult to accurately model water demand based on linear regression models. The traditional time series analysis method selects an appropriate model to fit the change pattern of historical data in order to describe the changing trend of the data [7–10]. Although this method is easy to implement, its models are not easily interpretable and are highly dependent on historical data. If the sample size of the historical data is small or the data quality is poor, the forecasting accuracy will experience a cliff-like decline. Artificial neural networks learn the characteristics and patterns of changes in historical water demand through connections and weights between neurons, and they then forecast future water demand based on the learned models [11–14]. Compared to other methods, artificial neural networks have stronger nonlinear modeling abilities and adaptability, enabling them to process more complex time series data. However, when the dataset has data quality issues or is a small size, artificial neural networks are prone to overfitting, leading to a decrease in forecasting accuracy. The forecasting of urban water demand is influenced by many factors, including environmental factors, historical water demand, and the water demand of the surrounding nodes. Traditional methods for water demand forecasting rely solely on historical water usage data. The quantity and quality of this data can greatly affect the forecasting accuracy. Forecasting accuracy often decreases rapidly as the time range increases, making long-term water demand forecasting difficult to achieve.

The limitations of traditional forecasting methods have led scholars to conduct research on multivariate water demand forecasting methods. Guo [15] proposed a new hybrid model for urban water demand forecasting, which employed the random forest method for data dimensionality reduction to eliminate unimportant influencing factors, utilized discrete wavelet transform to decompose the original sequence into several subsequences with different characteristics, and applied a time convolutional network to model multiple sub-sequences to generate forecasting results. The model demonstrated excellent predictive performance on a real dataset provided by a water plant in Suzhou, China, proving the reliability of using multivariate sequences for forecasting. Compared to the methods mentioned above, Transformer-based models that capture long-term dependencies have gradually become the mainstream model in the field of water demand forecasting. Many Transformer variant architectures have been designed for long-term time series forecasting modeling tasks. Zhou [16] proposed the Informer model, which first applies the Transformer architecture to the field of long-term time series forecasting and introduces a sparse self-attention mechanism to reduce the model's time complexity. Extensive experiments on multiple large-scale datasets confirmed the reliability of the Transformer architecture in the field of long-term time series forecasting. Wu [17] addressed the issue of low information utilization of the sparse self-attention mechanism by proposing a new decomposition architecture with an auto-correlation mechanism in the Autoformer model, progressively decomposing complex time series. The Autoformer model achieved better forecasting results than the Informer on multiple datasets in the five categories of time series forecasting tasks. Zhou [18] combined the Transformer with a seasonal trend decomposition method, further improving the long-term forecasting performance of the Transformer. The proposed FEDformer demonstrated strong predictive performance on large datasets, further confirming the reliability of the Transformer in the field of long-term time series forecasting. However, the Transformer-based model heavily relies on positional encoding to ensure the sequence of attention scores. In the process of water demand forecasting, the sorting information preserved by positional encoding introduces noise, which is difficult to eliminate at the model's output end. This assumption has been confirmed in Zeng's study [19], where he achieved better forecasting results than most Transformer variant models, using a simple DLiner model.

Sheng's thorough analysis of current Transformer-based models indicates that the incorporation of plug-and-play lightweight attention modules, coupled with reliable ensem-

ble strategies and forward-looking interpretability methods, can optimize the performance of runoff forecasting models [20]. Sheng's application of this lightweight attention in Res-GRU, which dynamically adjusts feature emphasis, has proven to refine predictive accuracy and affirm the method's validity [21]. Zhang's integration with a temporal convolutional network has demonstrated its robustness over various forecasting horizons [22]. Geng's AGON, with its attention-based gating for noise reduction and LSTM integration, has shown improved predictive outcomes across domains [23]. While these methods offer significant advancements in short-term forecasting, they also highlight the need for extensive datasets, which can result in higher data acquisition costs.

To address the problem of reduced water demand forecasting performance on small-scale datasets, this paper proposes a water demand forecasting method based on generative adversarial networks (GANs) and multi-feature fusion. The main contributions are as follows: (1) A data augmentation method based on GANs is introduced for multivariate time series, including a gradient penalty to constrain and enhance stability. (2) A multi-feature fusion method is applied as an improved feature extraction method based on temporal and channel features. (3) A self-attention mechanism adaptively allocates feature weights, improving long-term water demand forecasting performance. (4) Sample data of four types are modeled and compared with a variety of comparison models to verify the superiority of the proposed method.

2. Water Demand Forecasting Model Based on WDF-Mixer

This paper proposes a long-term water demand forecasting framework based on the WDF-mixer model, as shown in Figure 1. Firstly, to address the overfitting problem due to the impact of insufficient data, data augmentation is considered, to generate synthetic data that resemble and complement the original train set. Assuming that the input data are X_h, the process can be described as

$$X_h^{\text{aug}} = \delta(X_h) \tag{1}$$

where $\delta(\cdot)$ represents the data augmentation method and X_h^{aug} is the time series after passing through the data augmentation layer.

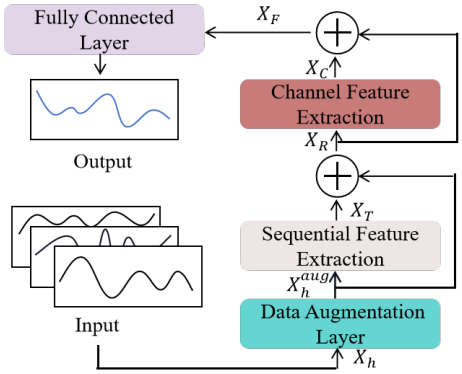

Figure 1. The overall architecture diagram of the WDF-mixer.

Additionally, temporal and channel feature extraction modules are constructed to exploit the correlation structure between the multiple water demand sequences, therefore modeling dynamic long-term dependencies in the time series. Among these, the process of temporal feature extraction can be simply summarized by the following formula:

$$X_T = \xi\left(X_h^{\text{aug}}\right) \tag{2}$$

where $\xi(\cdot)$ represents the methods for extracting temporal features and X_T represents the extracted temporal feature tensor.

Then, the input of the channel feature extraction X_R is obtained by superimposing the time series and the temporal feature tensor through a residual connection layer:

$$X_R = X_h^{\text{aug}} + X_T \tag{3}$$

The residual connection layer helps in training deep networks and alleviating the problem of information loss caused by feature extraction. Similarly, the process of channel feature extraction can be summarized as follows:

$$X_C = \mu(X_R) \tag{4}$$

where $\mu(\cdot)$ represents the methods for extracting channel features and X_C represents the extracted channel feature tensor.

Finally, combining X_R with the channel feature tensor X_C results in a multi-feature fusion vector X_F, which serves as the input of the fully connected network, establishing the mapping relationship between the multi-features and water demand:

$$\begin{aligned} X_F &= X_R + X_C \\ Y &= FC(X_F) \end{aligned} \tag{5}$$

2.1. Data Augmentation Layer Based on WGAN-GP

The performance of most deep learning models is usually conditional on the size of the dataset available for training. To deal with this issue, data augmentation techniques can be employed to learn from the series toward the creation of multiple new ones with the same underlying patterns but different randomness, with promising results [24–27]. Hence, data augmentation can be helpful to make sure that the trained models well-generalize beyond the training data and are capable of providing accurate long-term forecasts.

The data augmentation strategy used in this paper is shown in Figure 2. The Wasserstein generative adversarial network (WGAN), based on the Wasserstein distance, is a common data augmentation method. This method avoids the instability and mode collapse issues in GAN training by optimizing the distance between distributions. Additionally, to prevent the problems of gradient explosion and vanishing during training, a gradient penalty term (GP) is introduced on top of WGAN to constrain the gradient norm of the discriminator [28,29].

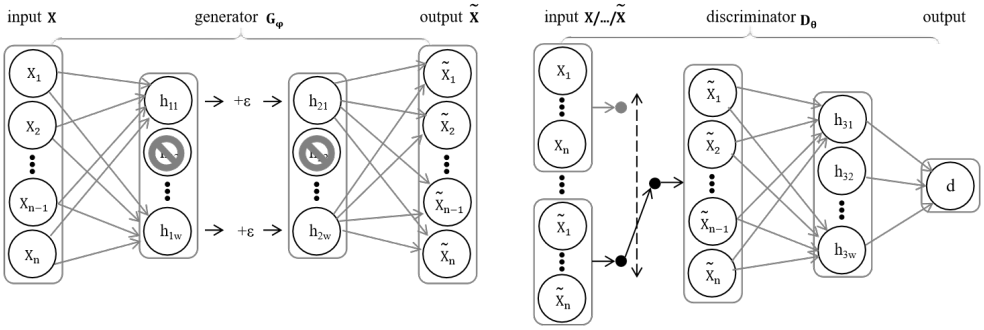

Figure 2. Schematic diagram of WGAN-GP data enhancement layer network structure.

For a water demand dataset X, suppose its probability distribution is $P_r(X)$ and that z is a latent space variable with probability distribution $P_z(z)$. First, parameter z is obtained, using samples from the latent space. Then, real data X and simulated data $\widetilde{X} = G_\varphi(z)$ are used to train a discriminator D_θ. The discriminator D_θ measures the difference between the data distribution generated by the generator G_φ and the real data distribution by calculating the Wasserstein distance between the two probability distributions, while the generator optimizes the generated data based on the Wasserstein distance to make it closer to the real data distribution. The Wasserstein distance is defined as

$$W(P_r, P_z) = \inf_{\gamma \sim \prod(P_r, P_z)} E_{(x,z)-\gamma}[\|x - z\|] \tag{6}$$

where $\prod(P_r, P_z)$ is the set of all possible joint distributions formed by combining $P_r(X)$ and $P_z(z)$ and (x, z) is a sample from this set. By optimizing the Wasserstein distance, $P_r(X)$ and $P_z(z)$ become closer, thereby providing a stable gradient for the generator:

$$W(P_r, P_z) = \frac{1}{K} \sup_{\|f\|_L \leq K} E_{x \sim P_r}[f(x)] - E_{x \sim P_z}[f(x)] \tag{7}$$

Since the $\inf_{\gamma \sim \prod(P_r, P_z)}$ in Equation (6) cannot be solved computationally, it is transformed into Equation (7). The $\|f\|_L \leq K$ indicates that there exists a constant $K \geq 0$, such that for any two numerical values x_1 and x_2 in the domain, the following holds:

$$|f(x_1) - f(x_2)| \leq K|x_1 - x_2| \tag{8}$$

where $f(x)$ satisfies K-Lipschitz. Therefore, the loss function of the discriminator for WGAN can be defined as

$$\text{Loss}_{WGAN}(D) = E_{z \sim P_z}[D(G(z))] - E_{x \sim P_r}[D(x)] \tag{9}$$

and the loss function of the generator G_φ is defined as

$$\text{Loss}_{WGAN}(G) = -E_{z \sim P_z}[D(G(z))] \tag{10}$$

WGAN-GP introduces a gradient penalty term into WGAN to constrain the gradient norm of the discriminator, ensuring that the equation satisfies the Lipschitz continuity:

$$\text{Loss}_{WGAN-GP}(D) = E_{z \sim P_z}[D(G(z))] - E_{x \sim P_r}[D(x)] + \lambda E_{\hat{x} \sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1] \tag{11}$$

Compared to Equation (9), the gradient penalty $\lambda E_{\hat{x} \sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1]$ helps the discriminator ensure that the L2 norm of the gradient varies in a stable manner about a fixed value. This avoids the concentration of parameters ω at the edges of the interval and stabilizes the gradients.

2.2. Temporal Feature Extraction Based on Temporal Factorization

Time series datasets may have only a limited number of time series that are valuable for reaching their full potential. This low-rank property results in significant computational cost increases during feature extraction, being prone to overfitting. As shown in Figure 3, to address the low-rank characteristics of time series, the temporal feature extraction module is designed to extract time varying features and capture temporal dependencies. The non-linear and non-stationary water demand series can be regarded as a quasi-periodic signal, which can be contaminated by various noise signals. There is compelling evidence that the performance of forecasting models can be improved by using signal decomposition techniques to produce cleaner signals as model inputs. Therefore, in a down-sampling layer, the augmented sample data X_h^{aug} is decomposed into sub-sequences $X_{h,i}$:

$$\begin{aligned} X_h^{\text{aug}} &= \text{generator}(\text{norm}(X_h)) \\ X_{h,1}, \ldots, X_{h,s} &= \text{sampled}\left(X_h^{\text{aug}}\right) \end{aligned} \tag{12}$$

where $X_{h,i}$ represents the i-th down-sampling sub-sequence and s denotes the total number of down-sampling sub-sequences.

For a down-sampling sub-sequence, feature extraction can be performed using a multi-layer perceptron (MLP), and the features can be re-merged in the original sequence order:

$$\begin{aligned} X_{h,i}^{M} &= \mathrm{MLP}(X_{h,i}), i \in [1, s] \\ X_{h}^{M} &= \mathrm{merge}\left(X_{h,1}^{M}, \ldots, X_{h,s}^{M}\right) \end{aligned} \tag{13}$$

Additionally, to filter out non-important features from the extracted features, the WDF-mixer model adds a simple attention gating unit after each MLP block. This unit can amplify important features and diminish non-important features based on their feature weights:

$$\begin{aligned} X_T &= W_{AT} X_h^M \\ W_{AT} &= \mathrm{softmax}\left(\mathrm{artention}\left(X_h^M\right)\right) = \mathrm{softmax}\left(\left(\frac{Q_t K_t^T}{\sqrt{d_t}}\right) V_t\right) \end{aligned} \tag{14}$$

where $Q_t = K_t = V_t = X_h^M$, d_t is the scaling factor, which is equal to the X_h^M dimension, and where W_{AT} represents the attention weight matrix, which can be obtained from the temporal feature sequence extracted by the MLP. By effectively guiding the model to focus on important features, the attention weights enhance the model's learning ability.

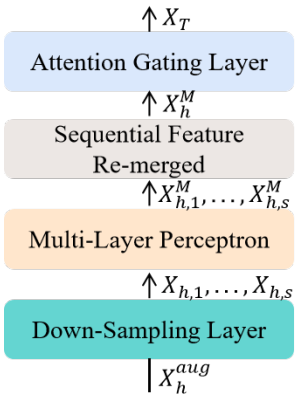

Figure 3. Module structure diagram of sequential feature extraction.

2.3. Channel Feature Extraction Based on Sparse Representation

Channel features refer to the correlations between different variables or input dimensions in multivariate time series data. The module of channel feature extraction is shown in Figure 4, consisting of a "Hamburger" structure and an attention gate layer. The "Hamburger" structure consists of two "bread" parts and one "ham" part [30]. The "bread" parts use simple linear layers for feature mapping. The lower "bread" X_l can project the input data into a higher-dimensional space, helping to introduce more detailed features. The upper "bread" X_u is used for dimensionality reduction and capturing the "essence" of the data, which are therefore less prone to overfitting. Additionally, the "ham" part uses non-negative matrix factorization (NMF) for channel feature extraction. In neural networks, NMF represents the original data matrix as the product of basis feature vectors and weight coefficient matrices, thereby extracting and representing features of the data, which can help the network learn latent patterns or feature representations in the data. The process of the channel feature extraction of input data X_R through the "Hamburger" structure can be described by the following formula:

$$\begin{aligned} X_l &= W_l X_R + b_l \\ M(X_l) &= \mathrm{NMF}(X_l) = \bar{X} + E = DC + E \\ X_u &= W_u M(X_l) + b_u \end{aligned} \tag{15}$$

Similar to the temporal feature extraction module, the channel feature extraction module also utilizes an attention gating layer to filter out non-important features, guiding the model to focus on important features and thus enhancing the model's ability to learn features between channels:

$$\begin{aligned} X_C &= W_{AC} X_u \\ W_{AC} &= \text{softmax}(\text{attention}(X_u)) = \text{softmax}\left(\left(\frac{Q_c K_c^T}{\sqrt{d_c}}\right) V_c\right) \end{aligned} \tag{16}$$

where $Q_c = K_c = V_c = X_u$. d_c is the scaling factor, which is equal to the X_u dimension.

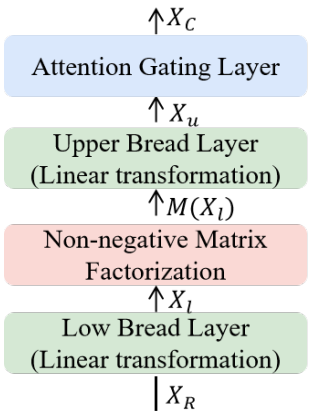

Figure 4. Module structure diagram of channel feature extraction.

3. Experiments and Analysis

To evaluate the effectiveness of the proposed method, experiments were conducted on a real water demand dataset. Section 3.1 introduces the relevant experimental settings. Section 3.2 presents long-term forecasting comparative experiments to verify the forecasting accuracy and stability of the model. Section 3.3 conducts ablation experiments to validate the importance of the model components and the effectiveness of multi-feature fusion.

3.1. Experimental Setup

3.1.1. Dataset

The experimental data in this paper were derived from the historical water demand dataset of a real-world water distribution system. As shown in Table 1, they are from four different types of areas, including commercial centers, universities, large malls, and residential areas, named as Company, School, Mall, and Apartment, respectively. Each dataset consists of seven channels of data, with each channel representing the flow data collected from flow sensors at different locations. The sampling interval is 5 min, and the time span is 1 year, totaling 105,120 water demand data points. In Figure 5, the waveform characteristics of the four datasets are distinctly represented. The Company dataset primarily exhibits a moderate oscillation with sporadic pulse-like waveforms, signaling abrupt alterations. The School dataset demonstrates a higher amplitude of fluctuation, accompanied by notable nonlinear attributes. The Mall dataset exhibits strong periodicity with waveforms akin to a square wave, yet it shows some fluctuation at the peak magnitudes. The Apartment dataset has a general periodic pattern, but it is subject to sharp variations within a single cycle's duration. One month of data, totaling 8640 data points, was extracted for experimentation to validate the model's performance under small sample sizes. The datasets were partitioned into three distinct subsets: a training set, a verification

set, and a test set, following the ratio of 7:1:2, respectively. The linear interpolation method was used to process missing data:

$$W(X) = \frac{(X - X_1)(Y_2 - Y_1)}{(X_2 - X_1)} + Y_1 \tag{17}$$

where X is the target value, $W(X)$ represents the target water demand, and (X1, Y1) and (X2, Y2) are adjacent data points of the target value.

Table 1. Overview of the experimental datasets.

The Name of the Dataset	Data Sources	Number of Channels	Total Amount of Data
Company	The city's commercial center.	7	105,120
School	Universities in the city.	7	105,120
Mall	The city malls.	7	105,120
Apartment	Residential areas of the city.	7	105,120

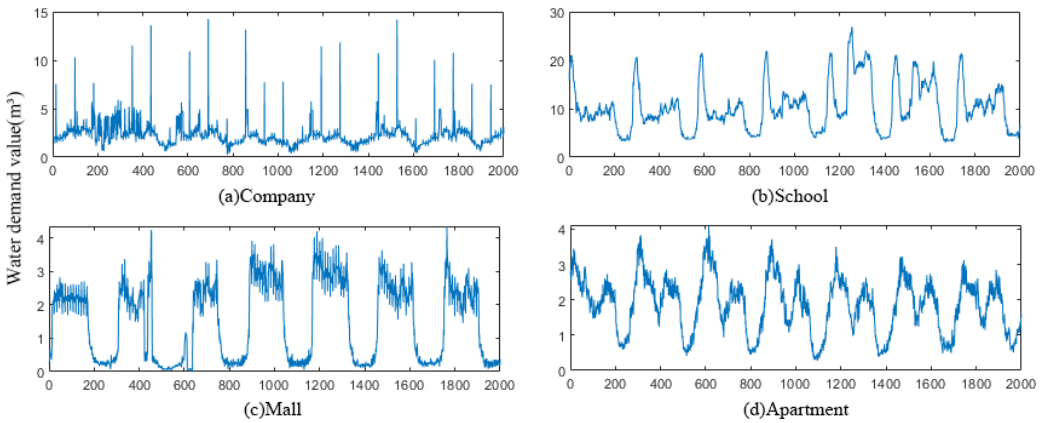

Figure 5. The water demand across four different datasets, with 2000 samples per dataset (the x-axis corresponds to the sample point number and the y-axis to the water demand value).

3.1.2. Evaluation Metrics

In this study, we employed three distinct error assessment metrics—mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2)—to evaluate the long-term forecasting accuracy of our model. The mathematical formulations of these metrics are delineated as follows:

$$\begin{aligned} \text{MSE} &= \frac{1}{N}\sum_{i=1}^{N}(\text{Truth}_i - \text{Pred}_i)^2 \\ \text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|\,\text{Truth}_i - \text{Pred}_i\,| \\ R^2 &= 1 - \frac{\sum_{i=1}^{N}(\text{Truth}_i - \text{Pred}_i)^2}{\sum_{i=1}^{N}\left(\text{Truth}_i - \overline{\text{Truth}}\right)^2} \end{aligned} \tag{18}$$

where N represents the length of the forecasting sequence, Truth_i represents the actual water demand at time point i, $\overline{\text{Truth}}$ represents the average of the actual water demand, and Pred_i represents the model's forecast water demand at time point i. The smaller the value of the evaluation metric, the higher the accuracy of the model's forecasts.

3.2. Results and Discussion

The model proposed in this paper was compared with long-term forecasting models, including Informer, Autoformer, FEDformer, DLiner, and SCINet [31]. In the fine-tuning processes of all the models, the Adam optimizer was employed. As shown in Table 2, the performance was compared across four types of datasets, with each consisting of 8640 water demand data points. The forecasting lengths of these models ranged from 48 to 720.

Table 2. Time series error scores for predicting water demand in four distinct datasets, all initialized with an input length of 96 and extending to output lengths of 48, 72, 96, 192, 336, and 720. The best-performing results are displayed in bold, and near-optimal results are underscored.

Models		Informer			Autoformer			FEDformer		
Metric		MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2
Company	48	0.7739	0.5298	0.3472	0.6325	0.4534	0.1331	0.6434	0.4485	0.4022
	72	0.6987	0.4827	0.3078	0.7060	0.5024	0.3765	0.9338	0.6252	0.4085
	96	0.7599	0.5248	0.3088	0.6896	0.4963	0.2843	0.5118	0.3872	0.2571
	192	0.7483	0.5232	0.0643	0.7660	0.5647	0.2838	0.6102	0.4515	0.4025
	336	0.7466	0.5389	0.2752	0.7563	0.5521	0.1461	1.2126	0.7135	0.4127
	720	0.7713	0.5335	−0.0006	0.7601	0.5632	0.2957	0.9654	0.6474	−0.0401
School	48	0.5416	0.5991	0.5569	1.0745	0.7847	0.0101	1.0266	0.7859	0.1506
	72	0.6090	0.5926	0.3478	1.2706	0.8463	−0.1795	1.2234	0.8215	0.0198
	96	0.6479	0.6387	0.5732	1.1509	0.8133	0.0567	1.3061	0.8922	−0.0949
	192	0.9056	0.7198	0.2810	1.2605	0.8732	−0.1783	1.4936	0.9700	−0.0840
	336	0.6603	0.6060	0.4273	1.1084	0.8037	0.0703	1.1498	0.8241	0.0122
	720	0.9077	0.7147	0.1281	1.0709	0.8081	0.0245	2.4551	1.2135	0.0367
Mall	48	0.6060	0.5694	0.7100	1.3458	0.9247	−0.0620	2.8219	1.1730	0.0193
	72	0.8361	0.7704	0.5244	1.5640	1.0359	0.0322	2.0716	1.1201	−0.3236
	96	0.7327	0.6614	0.4732	1.9132	1.1360	−0.2640	1.2737	0.9279	−0.0661
	192	0.6998	0.6662	0.5222	1.6717	1.1238	−0.3695	1.6539	1.0669	−0.0341
	336	0.7266	0.7387	0.1852	1.3890	0.9653	−0.2076	1.1364	0.8731	0.0684
	720	0.8182	0.6806	0.2306	1.1664	0.8849	−0.3694	1.2084	0.9022	−0.1449
Apartment	48	0.7917	0.6339	0.5966	2.4858	1.2347	−0.1841	3.3957	1.3073	0.1156
	72	1.0033	0.7727	0.6712	2.4219	1.2265	−0.1487	3.2344	1.3523	0.0530
	96	1.0495	0.8105	0.5759	2.3767	1.2201	−0.2571	4.0719	1.5199	−0.2556
	192	1.0101	0.7500	0.6602	2.6900	1.3247	−0.2848	3.4608	1.4361	−0.0409
	336	0.9650	0.7366	0.5740	2.1608	1.1653	−0.0947	4.1460	1.3630	0.1820
	720	1.5341	1.0018	0.4750	2.3209	1.2129	−0.0387	4.4506	1.5262	−0.0281
Models		**DLiner**			**SCINet**			**WDF-Mixer**		
Metric		**MSE**	**MAE**	R^2	**MSE**	**MAE**	R^2	**MSE**	**MAE**	R^2
Company	48	0.5864	0.4551	0.2015	0.5789	0.4539	0.2279	**0.4528**	**0.3526**	**0.4575**
	72	0.6043	0.4659	0.1748	0.6426	0.4913	0.1341	**0.4519**	**0.3604**	**0.4582**
	96	0.6512	0.5041	0.1033	0.7002	0.5324	0.0680	**0.4341**	**0.3451**	**0.4613**
	192	0.6158	0.4722	0.1526	0.6873	0.5248	0.0501	**0.4651**	**0.3544**	**0.4449**
	336	0.6019	0.4536	0.1916	0.6362	0.4779	0.1468	**0.4743**	**0.3506**	**0.4400**
	720	0.6588	0.4968	0.1188	0.6834	0.5023	0.1041	**0.4999**	**0.3609**	**0.4103**
School	48	0.5437	0.5720	0.3692	0.5253	0.5278	0.3905	**0.1099**	**0.2384**	**0.8826**
	72	0.6106	0.5937	0.2915	0.5705	0.5744	0.3380	**0.1041**	**0.2384**	**0.8884**
	96	0.6515	0.6147	0.2481	0.5785	0.5712	0.3324	**0.1133**	**0.2494**	**0.8781**
	192	0.6670	0.6101	0.2630	0.6620	0.6096	0.2685	**0.1415**	**0.2748**	**0.8496**
	336	0.4977	0.5004	0.4482	0.5174	0.5170	0.2463	**0.1709**	**0.2962**	**0.8203**
	720	0.5275	0.5314	0.4071	0.5477	0.5439	0.3844	**0.2377**	**0.3616**	**0.7490**
Mall	48	0.4188	0.4764	0.4660	0.6336	0.5112	0.1921	**0.0876**	**0.1984**	**0.8999**
	72	0.4828	0.5339	0.3619	0.7400	0.5815	0.0219	**0.1029**	**0.2157**	**0.8815**
	96	0.5076	0.5578	0.3179	0.7864	0.6148	−0.0567	**0.1268**	**0.2416**	**0.8532**
	192	0.5447	0.5866	0.2500	0.8927	0.6788	−0.2292	**0.1549**	**0.2696**	**0.8175**
	336	0.5331	0.5691	0.2514	0.7062	0.6011	0.0085	**0.1621**	**0.2771**	**0.8042**
	720	0.5809	0.5968	0.1581	0.9193	0.6869	−0.3324	**0.2271**	**0.3268**	**0.7161**
Apartment	48	1.0060	0.6605	0.3289	1.1648	0.7041	0.2230	**0.1870**	**0.2780**	**0.8458**
	72	1.1499	0.7483	0.2185	1.3434	0.7764	0.0869	**0.2046**	**0.2900**	**0.8299**
	96	1.1389	0.7495	0.2143	1.3491	0.7898	0.0693	**0.2158**	**0.2991**	**0.8194**
	192	1.1564	0.7710	0.1970	1.3570	0.8224	0.0577	**0.2363**	**0.3144**	**0.8024**
	336	1.0138	0.6926	0.3116	1.3656	0.7752	0.0727	**0.2514**	**0.3184**	**0.7920**
	720	1.2561	0.7916	0.1707	1.6307	0.8850	−0.0765	**0.2793**	**0.3342**	**0.7721**

The experimental results presented in Table 2 illustrate that the proposed WDF-mixer model achieved the optimal predictive performance on the small-scale datasets, whereas DLinear and SCInet exhibited suboptimal predictive performance at varying forecasting lengths. Despite DLiner and SCINet's relatively accurate predictions for the majority of data points, indicated by lower MSE and MAE scores, the impact of a small number of data points with larger errors is evident in the diminished R^2 score. SCInet utilized a recursive downsampling technique to decompose sequences for interactive learning, thereby enhancing the time series data and yielding satisfactory predictive outcomes. Meanwhile, DLinear decomposed time series into trend and residual components, employing two simplistic linear networks for modeling, also achieving satisfactory predictive performance. A comparative analysis of the methodological principles between the SCInet and DLinear models allowed us to derive the following conclusions [32,33]: (1) Decomposing sequences to address the low-rank characteristics of time series data can, to a certain extent, extract crucial features, facilitating easier modeling and representation of the data. (2) Due to the inherent low-rank nature of time series, decomposed sequences require only simple feature extraction methods to extract crucial information, resulting in favorable forecasting outcomes. These experimental observations further underscore the scientific rigor and effectiveness of the proposed time feature extraction module within the WDF-mixer framework. Notably, when contrasted with previous state-of-the-art outcomes, WDF-mixer achieved an average mean squared error (MSE) reduction of 25.02% in the Company dataset, 73.41% in the School dataset, 72.48% in the Mall dataset, and 79.53% in the Apartment dataset.

Figure 6 gives the metrics of the six forecasting models relative to the different samples. Along with the increase in forecasting length, the accuracy of the comparative models decreased. This result of the forecasting performance measure can be attributed to the change in correlation between the predictive sequence and the input sequence. Moreover, some characteristics did not only exist in one scale, but also in other scales. Hence, the forecasting performance might be poor if the features of the other scales were not considered. In contrast, the error curve of the WDF-mixer model consistently remained at the lowest position, which demonstrates the superiority of the model in long-term water demand forecasting. These results sufficiently illustrate that the WDF-mixer model has an advantage over comparative models in that it can represent long-range dependencies between observations. In addition, the data augmentation method, WGAN-GP, mitigates the impact of overfitting in water demand forecasting tasks that involve nonstationary series.

To assess the performance stability of all the algorithms, 50 independent experiments on four types of datasets were considered, and the results are shown in Figure 7. Compared to the other methods, the median and the interquartile of the WDF-mixer model remained at the lowest position in the graph, indicating high accuracy and robustness. These improvements achieved by the WDF-mixer model can be attributed to the multi-feature fusion. As such, diversity features can be utilized to improve predictive performance.

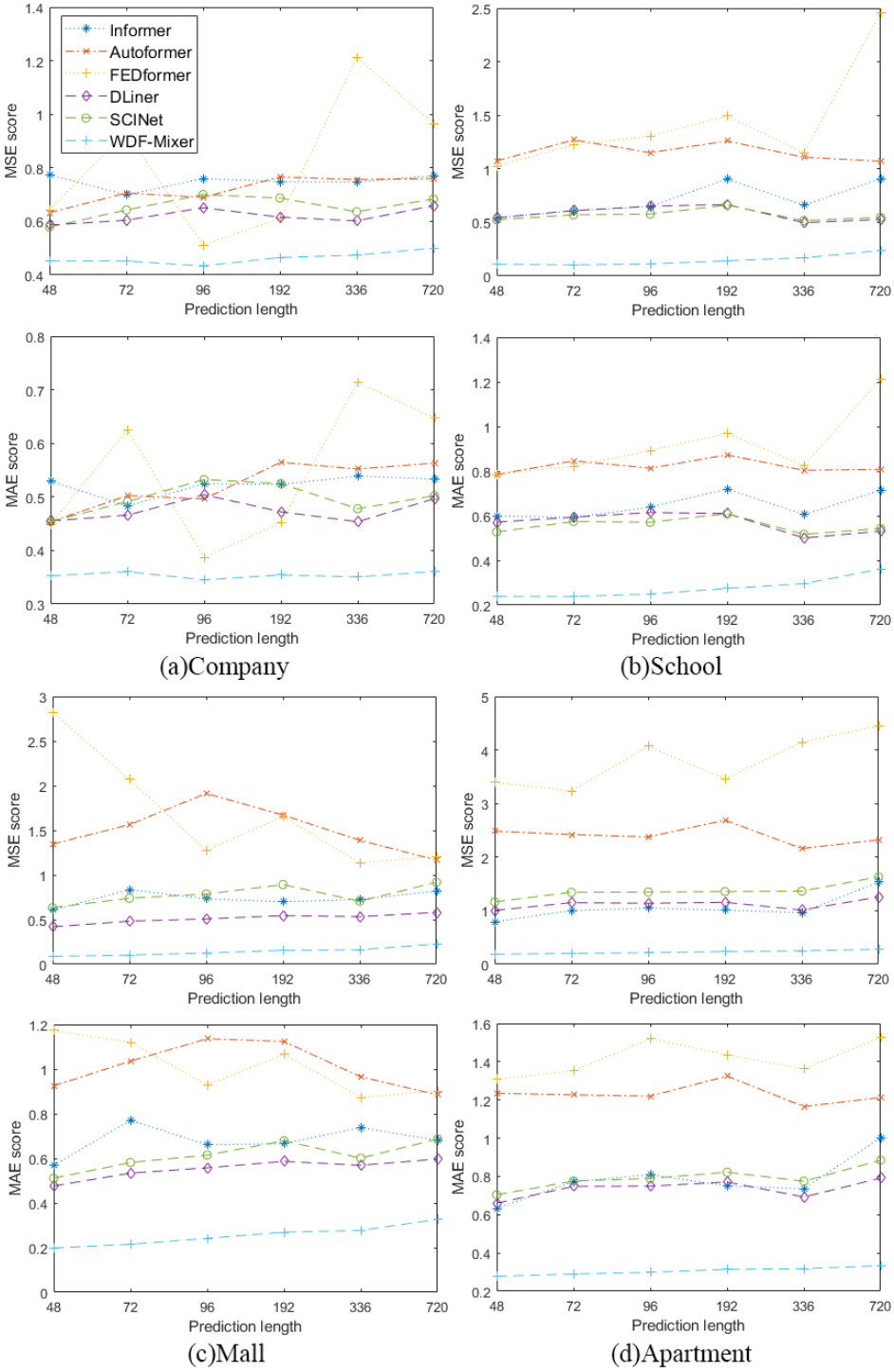

Figure 6. Error score polygons for different water demand datasets, illustrating the performance of various models over a range of forecasting lengths.

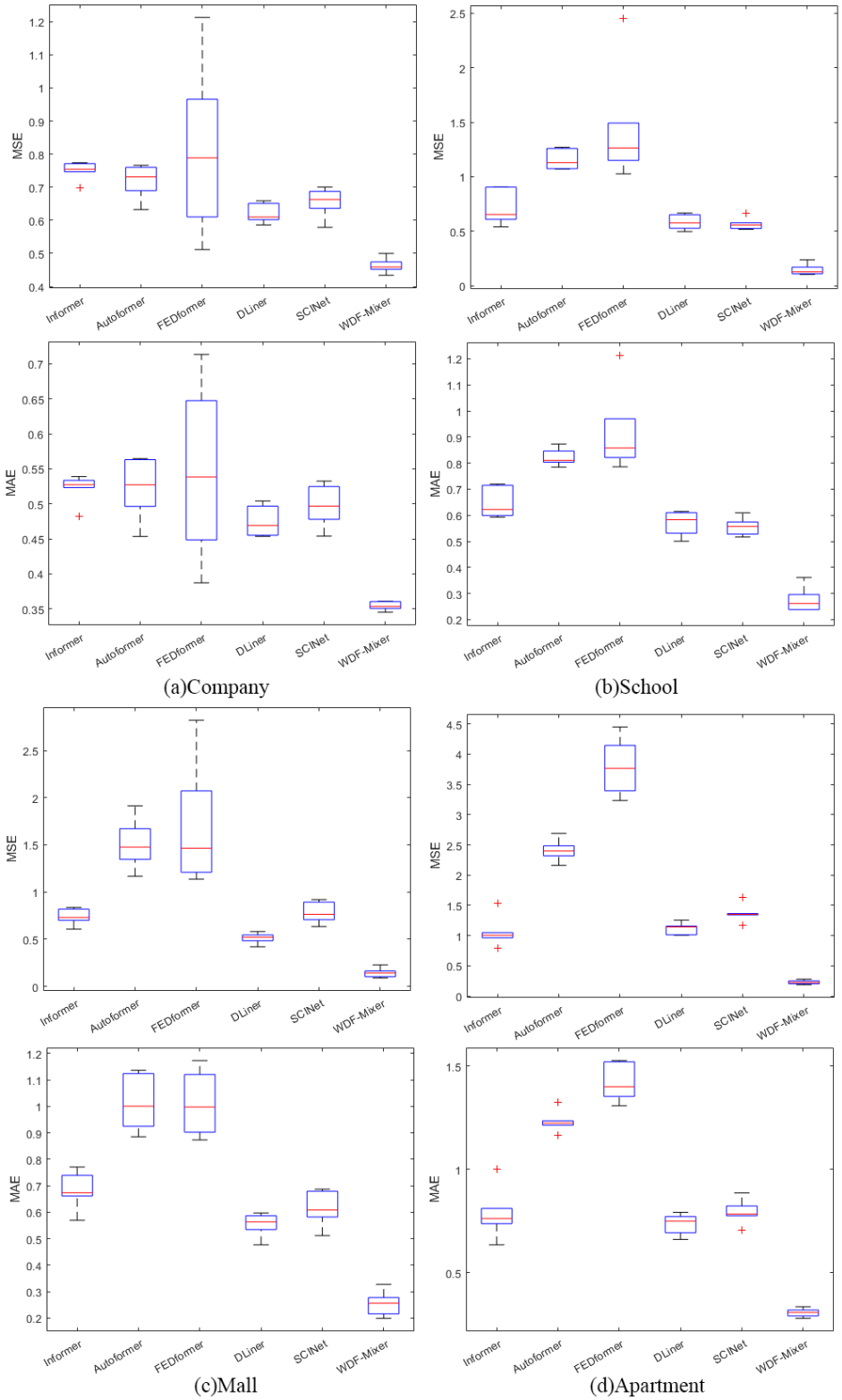

Figure 7. Box-line diagram of forecasting errors of different models for four types of water demand datasets. ("+" represents as the outlier).

3.3. Ablation Study

In order to gain a better understanding of the proposed model's behavior, this paper performed ablation studies on the WDF-mixer model, which mainly consisted of three

parts: (1) the predictive performance of WDF-mixer variants; (2) the predictive effectiveness of multi-feature fusion; (3) the effectiveness of the data augmentation layer in small sample forecasting.

The variants of the WDF-mixer were named as "WDF-Mixer (enhanced functionality)". The enhanced functionality represented either a single-channel feature extraction module (C), a data augmentation layer (E) or a combination of any of those terms. The following variants were compared:

WDF-mixer (none): Using only a temporal feature extraction module for forecasting.

WDF-mixer (C): Using a combination of temporal feature extraction module and channel feature extraction module forecasting.

WDF-mixer (E): Using a data augmentation layer for data expansion and a temporal feature extraction module for forecasting.

WDF-mixer (C,E): Using a data augmentation layer for data expansion and multi-dimensional feature fusion for forecasting.

Table 3 summarizes the performance of the four model variants on the real dataset. It is worth noting that in the first three datasets the WDF-mixer (C) model, which used multi-feature fusion, slightly outperformed the WDF-mixer (none) model. This was due to the Apartment's nonlinearity being weaker than the others. The WDF-mixer (none) model already exhibited overfitting during forecasting. Therefore, when using the WDF-mixer (C) model with stronger nonlinear modeling capabilities, its forecasting error did not decrease but instead increased, exacerbating the overfitting problem.

Table 3. MSE score table of ablation results (the input length was 96 and the output length was 48, 72, 96, 192, 336, 720, respectively. The best results are highlighted in bold, and the suboptimal results are highlighted with an underscore).

Dataset	FL	WDF-mixer (None)	WDF-mixer (C)	WDF-mixer (E)	WDF-mixer (C,E)
Company	48	0.6387	0.5818	0.4880	**0.4528**
	72	0.6501	0.6533	0.5017	**0.4519**
	96	0.6315	0.5993	0.5154	**0.4341**
	192	0.6509	0.6305	0.5400	**0.4651**
	336	0.6180	0.6180	0.5452	**0.4743**
	720	0.6841	0.6790	0.5889	**0.4999**
School	48	0.2882	0.3098	0.1671	**0.1099**
	72	0.3809	0.3465	0.1546	**0.1041**
	96	0.3725	0.3409	0.1619	**0.1133**
	192	0.4144	0.3426	0.2096	**0.1415**
	336	0.3293	0.3253	0.2176	**0.1709**
	720	0.4439	0.4147	0.3032	**0.2377**
Mall	48	1.1114	0.7198	0.1981	**0.0876**
	72	1.0280	0.9236	0.2031	**0.1029**
	96	1.5054	1.3083	0.2349	**0.1268**
	192	1.1978	1.0126	0.2460	**0.1549**
	336	0.9921	1.1655	0.2238	**0.1621**
	720	1.0310	0.9160	0.2996	**0.2271**
Apartment	48	2.9785	3.4051	0.2318	**0.1870**
	72	3.1928	3.5035	0.2721	**0.2046**
	96	3.4504	3.6843	0.2874	**0.2158**
	192	2.9378	3.5552	0.3153	**0.2363**
	336	2.4787	2.2405	0.3039	**0.2514**
	720	3.5042	4.1476	0.3421	**0.2793**
Improvement over WDF-mixer			2.48%	58.24%	67.14%

Similarly, experiments were conducted on the Apartment dataset using the WDF-mixer (E) model. According to the experimental results, by adding a data augmentation layer to expand the sample data, the forecasting accuracy improved, indicating that data augmentation can effectively alleviate the model's overfitting phenomenon. Furthermore,

the WDF-mixer (C,E) model, after data augmentation, performed the best when using multi-feature fusion for feature extraction. It can be seen that the model's data augmentation layer generated more meaningful sequence samples, enriching the features of the samples and significantly improving the forecasting accuracy.

4. Conclusions

Complex water demand patterns are formed during the process of urban development and are mainly affected by human activities. This study proposed a framework to delineate and analyze long-term water demand patterns and dynamics by using multi-feature fusion. The stability and superiority of the proposed model were verified through the experimental analysis of the proposed framework in real-world datasets.

In this study, a WGAN-BP method was used to generate synthetic data that resembled and complemented the original samples. And a gradient constraint was introduced to WGAN-BP, ensuring that the distribution of each parameter was uniform. Then, the obtained samples were decomposed into sub-sequences. For each sub-sequence, multiple MLP models were used for temporal feature extraction. In addition, a channel feature extraction module, based on the "Hamburger" structure, was used to avoid the risk of overfitting. Finally, an attention layer was used to capture depth characteristics and to obtain the temporal dependencies of the water demand.

This study presented the analysis of long-term water demand forecasting, taking advantage of big data sources from a smart water distribution system. This work should be helpful for the sustainable development and management of cities. In our future work, a more efficient forecasting method will be researched, and model interpretability could be considered.

Author Contributions: Conceptualization, C.Y.; funding acquisition, B.L. and Z.W.; methodology, J.M. and K.W.; software, J.M.; formal analysis, C.Y., B.L., Z.W. and K.W.; writing—original draft preparation, C.Y., J.M. and K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Zhejiang Provincial Natural Science Foundation of China No. LQ23F030002, the "Ling Yan" Research and Development Project of Science and Technology Department of Zhejiang Province of China under Grant Nos. 2022C03122, 2023C03161, and Public Welfare Technology Application and Research Projects of Zhejiang Province of China under Grant No. LGF22F020006.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to thank the editor and the reviewers for their kind help in improving this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fu, G.; Jin, Y.; Sun, S.; Yuan, Z.; Butler, D. The role of deep learning in urban water management: A critical review. *Water Res.* **2022**, *223*, 118973. [CrossRef]
2. Chen, L.; Yan, H.; Yan, J.; Wang, J.; Tao, T.; Xin, K.; Li, S.; Pu, Z.; Qiu, J. Short-term water demand forecast based on automatic feature extraction by one-dimensional convolution. *J. Hydrol.* **2022**, *606*, 127440. [CrossRef]
3. Zubaidi, S.L.; Hashim, K.; Ethaib, S.; Al-Bdairi, N.S.S.; Al-Bugharbee, H.; Gharghan, S.K. A novel methodology to predict monthly municipal water demand based on weather variables scenario. *J. King Saud-Univ.-Eng. Sci.* **2022**, *34*, 163–169. [CrossRef]
4. Du, B.; Huang, S.; Guo, J.; Tang, H.; Wang, L.; Zhou, S. Interval forecasting for urban water demand using PSO optimized KDE distribution and LSTM neural networks. *Appl. Soft Comput.* **2022**, *122*, 108875. [CrossRef]
5. Mokhtar, A.; Elbeltagi, A.; Gyasi-Agyei, Y.; Al-Ansari, N.; Abdel-Fattah, M.K. Prediction of irrigation water quality indices based on machine learning and regression models. *Appl. Water Sci.* **2022**, *12*, 76. [CrossRef]
6. Stańczyk, J.; Kajewska-Szkudlarek, J.; Lipiński, P.; Rychlikowski, P. Improving short-term water demand forecasting using evolutionary algorithms. *Sci. Rep.* **2022**, *12*, 13522. [CrossRef]
7. Pandey, P.; Bokde, N.D.; Dongre, S.; Gupta, R. Hybrid models for water demand forecasting. *J. Water Resour. Plan. Manag.* **2021**, *147*, 04020106. [CrossRef]

8. Niknam, A.; Zare, H.K.; Hosseininasab, H.; Mostafaeipour, A.; Herrera, M. A critical review of short-term water demand forecasting tools—What method should I use? *Sustainability* **2022**, *14*, 5412. [CrossRef]
9. Liu, X.; Zhang, Y.; Zhang, Q. Comparison of EEMD-ARIMA, EEMD-BP and EEMD-SVM algorithms for predicting the hourly urban water consumption. *J. Hydroinform.* **2022**, *24*, 535–558. [CrossRef]
10. Li, H.; Wang, X.; Guo, H. Uncertain time series forecasting method for the water demand prediction in Beijing. *Water Supply* **2022**, *22*, 3254–3270. [CrossRef]
11. Zubaidi, S.L.; Al-Bdairi, N.S.S.; Ortega-Martorell, S.; Ridha, H.M.; Al-Ansari, N.; Al-Bugharbee, H.; Hashim, K.; Gharghan, S.K. Assessing the benefits of nature-inspired algorithms for the parameterization of ANN in the prediction of water demand. *J. Water Resour. Plan. Manag.* **2023**, *149*, 04022075. [CrossRef]
12. Huang, H.; Lin, Z.; Liu, S.; Zhang, Z. A neural network approach for short-term water demand forecasting based on a sparse autoencoder. *J. Hydroinform.* **2023**, *25*, 70–84. [CrossRef]
13. Zanfei, A.; Brentan, B.M.; Menapace, A.; Righetti, M.; Herrera, M. Graph convolutional recurrent neural networks for water demand forecasting. *Water Resour. Res.* **2022**, *58*, e2022WR032299. [CrossRef]
14. Rustam, F.; Ishaq, A.; Kokab, S.T.; de la Torre Diez, I.; Mazón, J.L.V.; Rodríguez, C.L.; Ashraf, I. An artificial neural network model for water quality and water consumption prediction. *Water* **2022**, *14*, 3359. [CrossRef]
15. Guo, J.; Sun, H.; Du, B. Multivariable time series forecasting for urban water demand based on temporal convolutional network combining random forest feature selection and discrete wavelet transform. *Water Resour. Manag.* **2022**, *36*, 3385–3400. [CrossRef]
16. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 11106–11115.
17. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.
18. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Proceedings of the International Conference on Machine Learning. PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 27268–27286.
19. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 11121–11128.
20. Sheng, Z.; Wen, S.; Feng, Z.k.; Gong, J.; Shi, K.; Guo, Z.; Yang, Y.; Huang, T. A survey on data-driven runoff forecasting models based on neural networks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 1083–1097. [CrossRef]
21. Sheng, Z.; Wen, S.; Feng, Z.k.; Shi, K.; Huang, T. A Novel Residual Gated Recurrent Unit Framework for Runoff Forecasting. *IEEE Internet Things J.* **2023**, *10*, 12736–12748. [CrossRef]
22. Zhang, C.; Sheng, Z.; Zhang, C.; Wen, S. Multi-lead-time short-term runoff forecasting based on Ensemble Attention Temporal Convolutional Network. *Expert Syst. Appl.* **2024**, *243*, 122935. [CrossRef]
23. Geng, X.; He, X.; Xu, L.; Yu, J. Attention-based gating optimization network for multivariate time series prediction. *Appl. Soft Comput.* **2022**, *126*, 109275. [CrossRef]
24. Iglesias, G.; Talavera, E.; González-Prieto, Á.; Mozo, A.; Gómez-Canaval, S. Data augmentation techniques in time series domain: A survey and taxonomy. *Neural Comput. Appl.* **2023**, *35*, 10123–10145. [CrossRef]
25. Wang, H.; Zhang, Y.; Liang, J.; Liu, L. DAFA-BiLSTM: Deep autoregression feature augmented bidirectional LSTM network for time series prediction. *Neural Netw.* **2023**, *157*, 240–256. [CrossRef]
26. Shangguan, A.; Xie, G.; Fei, R.; Mu, L.; Hei, X. Train wheel degradation generation and prediction based on the time series generation adversarial network. *Reliab. Eng. Syst. Saf.* **2023**, *229*, 108816. [CrossRef]
27. Luleci, F.; Catbas, F.N.; Avci, O. Generative adversarial networks for labeled acceleration data augmentation for structural damage detection. *J. Civ. Struct. Health Monit.* **2023**, *13*, 181–198. [CrossRef]
28. Pérez, J.; Arroba, P.; Moya, J.M. Data augmentation through multivariate scenario forecasting in Data Centers using Generative Adversarial Networks. *Appl. Intell.* **2023**, *53*, 1469–1486. [CrossRef]
29. Demir, S.; Mincev, K.; Kok, K.; Paterakis, N.G. Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. *Appl. Energy* **2021**, *304*, 117695. [CrossRef]
30. Geng, Z.; Guo, M.H.; Chen, H.; Li, X.; Wei, K.; Lin, Z. Is attention better than matrix decomposition? *arXiv* **2021**, arXiv:2109.04553.
31. Liu, M.; Zeng, A.; Chen, M.; Xu, Z.; Lai, Q.; Ma, L.; Xu, Q. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 5816–5828.
32. Chen, Z.; Ma, M.; Li, T.; Wang, H.; Li, C. Long sequence time-series forecasting with deep learning: A survey. *Inf. Fusion* **2023**, *97*, 101819. [CrossRef]
33. Sousa, J.; Henriques, R. Intersecting reinforcement learning and deep factor methods for optimizing locality and globality in forecasting: A review. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108082. [CrossRef]