ENVIRONMENTAL RESEARCH
INFRASTRUCTURE AND SUSTAINABILITY

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

Combining wavelet-enhanced feature selection and deep learning techniques for multi-step forecasting of urban water demand

To cite this article: Wenjin Hao *et al* 2024 *Environ. Res.: Infrastruct. Sustain.* **4** 035005

ENVIRONMENTAL RESEARCH
INFRASTRUCTURE AND SUSTAINABILITY

OPEN ACCESS

RECEIVED
4 March 2024

REVISED
17 June 2024

ACCEPTED FOR PUBLICATION
2 July 2024

PUBLISHED
19 July 2024

PAPER

Combining wavelet-enhanced feature selection and deep learning techniques for multi-step forecasting of urban water demand

Wenjin Hao[1,*], **Andrea Cominola**[2,3] **and Andrea Castelletti**[1]

[1] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Via Giuseppe Ponzio 34, 20133 Milano, Italy
[2] Chair of Smart Water Networks, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany
[3] Einstein Center Digital Future, Wilhelmstraße 67, 10117 Berlin, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: wenjin.hao@polimi.it

Keywords: urban water demand forecasting, long short-term memory, light gradient boosting machine, attention mechanism, wavelet transforms, feature selection, deep learning

Abstract

Urban water demand (UWD) forecasting is essential for water supply network optimization and management, both in business-as-usual scenarios, as well as under external climate and socio-economic stressors. Different machine learning and deep learning (DL) models have shown promising forecasting skills in various areas of application. However, their potential to forecast multi-step ahead UWD has not been fully explored. Modelling uncertain UWD patterns and accounting for variations in water demand behaviors require techniques that can extract time-varying information and multi-scale changes. In this research, we comparatively investigate different state-of-the-art machine learning- and DL-based predictive models on 1 d- and 7 d-ahead UWD forecasting, using daily demand data from the city of Milan, Italy. The contribution of this paper is two-fold. First, we compare the forecasting performance of different machine learning and DL models on single- and multi-step daily UWD forecasting. These models include an artificial neural network, a support vector regression, a light gradient boosting machine (LightGBM), and long short-term memory networks with and without an attention mechanism (LSTM and AM-LSTM). We benchmark their prediction accuracy against autoregressive time series models. Second, we investigate the potential enhancement in predictive accuracy by incorporating the wavelet transform and feature selection performed by LightGBM into these models. Results show that, overall, wavelet-enhanced feature selection improves the model predictive performance. The hybrid model combining wavelet-enhanced feature selection via LightGBM with LSTM (WT-LightGBM-(AM)-LSTM) can achieve high levels of accuracy with Nash-Sutcliffe Efficiency larger than 0.95 and Kling–Gupta Efficiency higher than 0.93 for both 1 d- and 7 d-ahead UWD forecasts. Furthermore, performance is shown to be robust under the influence of external stressors causing sudden changes in UWD.

1. Introduction

Urban water security is paramount for water utilities and water users alike. Yet, it is hampered by rapidly growing urbanization, population growth, and climate change [1]. Socio-economic and health stressors (e.g. the recent COVID-19 pandemic) can also intensify existing deficits in urban water supply and sanitation services [2, 3]. As the global water supply and demand gap is estimated to reach 40% in 2030 [4], water demand management strategies constitute an important complement to supply-side interventions to achieve water security and efficiency of water use [5]. Recently, advances in water demand monitoring, for example, through smart water metering, have enabled the collection of high resolution data, which can be used to inform water demand management programs [6, 7]. Accurate predictions of urban water demands (UWD) provide a critical input to planning and managing water supply systems, and to formulating demand management programs for both short- and long-term water conservation [8].

Various predictive models have been proposed in the literature on UWD forecasting and have been tested in different cities and case studies around the world [9–12]. Time series models such as simple autoregressive (AR) and AR integrated moving average (ARIMA) have been widely explored for UWD forecasting tasks [13, 14]. This class of models is easy to interpret and implement and can describe linear relationships in UWD and learn time series patterns quickly. However, challenges to their general suitability for UWD forecasting emerge due to non-linearity in UWD data [15]. More recent research has explored how machine learning (ML) models can help improve prediction accuracy due to their ability to approximate nonlinear relationships in observational data. Tree-based models such as random forests (RF) and kernel-based models, including support vector machines (SVM) and support vector regression (SVR), have been employed in the forecasting of UWD [16, 17]. These models have two main advantages, i.e. their relative ease of implementation and minimal requirement for parameter tuning. However, despite these advantages, their structural design is not optimized to capture the complex temporal dependencies and dynamics in time series data. Consequently, these models typically achieve only moderate performance in forecasting one-step-ahead UWD [16, 18]. Ensemble learning models, including RF, have been increasingly used in hydrological modeling due to their efficiency and robustness, which stem from aggregation of multiple base learners [19]. Within this class of models, gradient boosting decision trees (GBDT) have recently outperformed traditional methods like RF in various regression and classification tasks [20]. For instance, in their comparative study of single and ensemble statistical and ML forecasting models for UWD, Shuang and Zhao [21] demonstrated that GBDT outperforms RF and adaptive boosting (AdaBoost) in terms of accuracy and robustness. The application of GBDT to relatively small datasets allows for reasonable tuning and training times. However, a notable limitation of this ensemble learning approach is its significant demand for computational time when applied to large datasets and complex forecasting tasks. This can constrain the efficiency of GBDT, particularly in scenarios where rapid model deployment is critical [22]. An efficient version of GBDT named light gradient boosting machine (LightGBM; [23]) has demonstrated significant capability and time-efficiency for prediction tasks such as forecasting electricity demand and sales [24, 25]. Yet, its application in water demand research remains relatively unexplored. LightGBM can also be regarded as an embedded feature selection model, providing information on the relevance of model predictors and allowing for the reduction and filtering of redundant input variables [26]. As a wide range of factors have so far been considered potential determinants of UWD, weather factors combined with historical UWD are generally considered important for short- to medium-term UWD forecasting [27, 28]. Socio-economic factors and activities such as tourism [29] and holiday periods [30] are also considered influential in particular contexts/seasons. The temporal changes in input variables are highly relevant to UWD, especially with daily and weekly lags. Various time-lag data for each variable are considered as individual input variables of the model, possibly resulting in a high-dimensional input space for the forecasting model [31]. Thus, feature selection that selects the most important predictors can contribute to improving model accuracy while limiting its dimensionality.

In addition to the above tree-based and ensemble learning models, a recent review paper [32] that analyses the most widely adopted neural-based ML models (i.e. artificial neural networks (ANN), feedforward neural networks (FFNN) and multi-layer perceptrons (MLP)) in hydrology and hydraulics emphasizes their suitability to deal with nonlinear processes and observations. Previous studies have acknowledged the advantages of ANN models over traditional regression and time series models for both daily and peak daily UWD forecasting [13, 33]. The ability of ANN to model the non-linear dynamics in UWD patterns, particularly during periods of high demand, represents a key advantage over other methods. ANN-based models also achieve good performance in other forecasting tasks in urban contexts such as electricity consumption [34]. However, the shallow architectures inherent in traditional ANN models present challenges in efficiently processing large-scale data sets. This constraint hinders their ability to identify and isolate useful information or characteristics from large data sets, thereby impeding their overall effectiveness in complex tasks that involve the analysis of extensive datasets [35]. Recent research started to test the potential of deep learning (DL) architectures like recurrent neural networks (RNN), long short-term memory networks (LSTM), convolutional neural networks (CNN), and graph neural networks (GNN) for UWD forecasting [35–38]. Among these, LSTM have shown particularly promising results across case studies and offer relatively straightforward hyperparameter tuning for time series data. This indicates LSTM's superior capability in learning complex structures in UWD time series data. In contrast to shallow neural networks, the LSTM network has sequential memory blocks consisting of multiple layers of interacting neural networks that allow LSTM to preserve useful information from previous steps and make predictions by combining the output of such memory blocks with new data [39]. Their efficacy is further underscored by exceptional performance in recent undertakings involving UWD forecasting [12, 40, 41]. For example, Mu *et al* [42] confirmed the efficacy of LSTM models for short-term demand predictions at sub-daily temporal resolutions: 15 min and hourly. Their study compared the performance of LSTMs with ARIMA, SVM, and

RF models, highlighting the strength of LSTM in managing demand spikes. A pilot study in Cairo [43] utilizing 10 min smart meter data from 2–20 households demonstrated that LSTM surpassed both SVM and RF in overall accuracy. However, the same study also noted a limitation of LSTM in accurately predicting peak demand at the household level. Recent research has explored hybrid LSTM models to enhance the predictive power and versatility of traditional LSTM architectures by incorporating diverse methodologies within a unified framework. Du *et al* [35] highlighted the efficacy of integrating discrete wavelet transform (DWT) and principal component analysis as pre-processing techniques with LSTM. This combination yielded superior performance in daily UWD forecasting when compared to ANN and SVM models. This finding underscores the benefits of leveraging advanced pre-processing methods to enhance LSTM's predictive accuracy. Other studies [44, 45] investigated the combination of CNN and LSTM networks for daily UWD prediction. These hybrid models have demonstrated enhanced performance compared to their standalone counterparts, primarily due to robust feature extraction capabilities of CNN. This synergy effectively harnesses the strengths of both architectures, leading to more accurate predictions. An attention mechanism (AM) can improve the accuracy of DL methods by assigning greater importance to pertinent inputs in various applications, as demonstrated by state-of-the-art research in speech and image recognition and natural language processing [46]. While LSTM coupled with the AM (AM-LSTM) can perform better than simple LSTM [40, 47], this mechanism has not been fully explored for UWD forecasting, especially for multistep forecasting tasks. In summary, existing research highlights the considerable potential of advanced DL models for UWD forecasting, particularly as data acquisition grows and datasets become more complex. However, the majority of these studies have concentrated solely on short-term forecasting, with a forecast horizon limited to a maximum of one day, and have not extensively addressed multi-step forecasting. Additionally, recent advancements in DL models and their hybrid variations have predominantly been evaluated against models with similar structures. This narrow focus results in a lack of comprehensive comparisons with a broader range of baseline and fundamental models, which could offer deeper insights into their relative performance.

Beside choosing a class of models, data processing plays a key role toward accurate UWD forecasting. Wavelet transform (WT), for instance, is recognized as an effective data pre-processing tool to better deal with non-stationarity and decompose multiscale changes in time series data, further improving the predictive capacity of ML and DL models. Coupled models with WT emerged as one group of promising hybrid models in UWD forecasting [35, 48]. The wavelet data-driven foresting framework (WDDFF) was proposed in [49] to overcome the drawbacks of inappropriate applications (i.e. boundary condition (BC)-related issues) and provide a set of best practices (i.e. algorithm selection, filter and decomposition selection, data partitioning, feature selection) for developing an appropriately coupled WT-ML model. However, so far, WDDFF has only been tested with ANN, RF, SVM, least squares regression methods, extreme learning machine, and second-order Volterra series models in UWD forecasting applications. In light of its flexibility, new predictive and feature selection models like LSTM and LightGBM can be combined with WT within the WDDFF and their potential for UWD predictions can be compared to traditional models.

Building on recent advances in data-driven models and their potential for UWD forecasting, we comparatively test the suitability and performance of various standalone ML and DL models. Additionally, we propose hybrid WT-LightGBM-ML/DL models for UWD forecasting over different horizons, ranging from 1 to 7 d-ahead, using daily water demand data from the city of Milan, Italy. The contribution of this study is three-fold. First, we develop a state-of-the-art AM-LSTM model and compare it with other predictive models from various classes, namely AR, SARIMAX, ANN, SVR, LightGBM, and conventional LSTM to predict daily UWDs in Milan (Italy). Second, we introduce wavelet-enhanced feature selection (WT-LightGBM) to explore whether hybrid models integrated with wavelet decomposition and LightGBM as a feature selection mechanism yield more precise multi-step (1 d- and 7 d-ahead) UWD forecasts. Finally, we apply the developed models to forecast UWD in 2020, a period potentially impacted by the pandemic lockdown, testing their predictive robustness against unprecedented changes not observed in pre-pandemic data.

The rest of the paper is organised as follows: section 2 describes the case study we consider, the predictive models, WT technique, and the set up of the detailed experiments that we implement for UWD forecasting. Section 3 reports the main results and discussions. Finally, section 4 summarises the main conclusions of this study.

2. Material and methods

2.1. Case study

This study focuses on forecasting the total daily UWD in Milan, a city in Northern Italy. There are a total of 28 active pump stations that extract groundwater and distribute it to final users through an articulated water

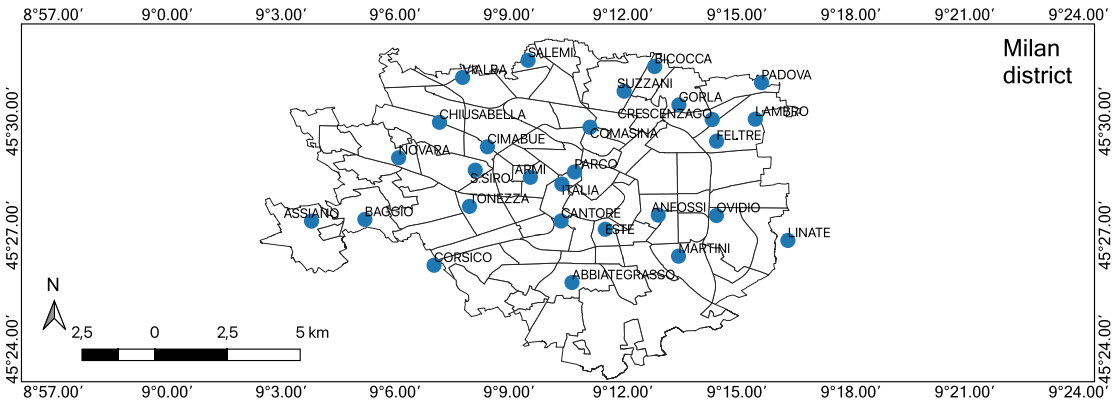

Figure 1. Map with the districts in Milan, with blue circle markers indicating the location of the 28 pump stations supplying water to the city.

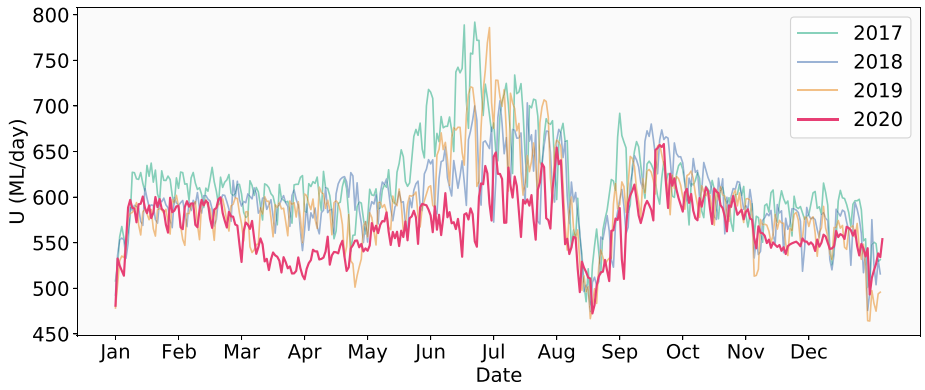

Figure 2. Urban water demand time series in Milan observed for each year between 2017 and 2020. The time series of each year is represented with a different color.

distribution network with a total length of approximately 2228 km. The location of the pump stations and the study area are shown in figure 1. As one of the largest cities in Italy, the average water consumption in Milan had a long decline period from 248 liters per capita per day (lpd) in 2002 to 206 lpd in 2016 [50] and then rebounded to 265 lpd in 2020 [51] possibly due to hot weather and severe heatwaves, which are becoming more and more frequent. The total amount of water pumped into the distribution system is measured with a daily frequency at the input node of the entire system. For this study, time series of daily UWD data at the city level (U) were available from 1 January 2017 to 31 December 2020. These data include in total 1461 records after preliminary pre-processing carried out by the local water utility (MM). Figure 2 illustrates the temporal dynamics of UWD in Milan from 2017 to 2020, represented by different colors for each year. Overall, the UWD showed pronounced seasonal fluctuations, peaking in the summer months, particularly in July, and decreasing towards the winter months. This pattern is typical due to increased usage in warmer weather for activities such as gardening and cooling. Additionally, a notable decreasing trend in water demand was observed in August across the years, which can be attributed to the summer vacation period when a significant portion of the population typically leaves the city. Another observable decrease in water demand occurred towards the end of December, which can be attributed to the Christmas holiday period. The year 2018 followed a similar trend, with peaks also around the summer months but slightly lower than those observed in 2017, suggesting a possible mild reduction in peak summer water usage. The year 2019 displayed a continuation of this trend but with slightly higher variability than the previous years. The peaks in 2019 surpassed those in 2018, indicating a rebound in water usage intensity during the summer. Notably, the fluctuations in 2019 were more frequent, reflecting possibly sporadic rainfall or varying temperature conditions affecting water use. However, a significant change was observed in the year 2020, where the water demand notably flattened compared to the preceding years. The water use in 2020 was markedly lower, particularly during the months from March to May, aligning with the initial COVID-19 lockdown period where restrictions may have led to reduced commercial and recreational water usage. The UWD in summer also had smoothe peaks in 2020 compared to the previous years. The observed mean UWD in 2017–2019 was 601.14 megalitres per day (ML d^{-1}), while it dropped to 567.73 megalitres per day (ML d^{-1}) in 2020. Consequently, 2020 is distinguished as a year with notable shifts in water usage patterns. The

Table 1. Summary of statistics of urban water demand and meteorological data in Milan (Italy) for the period 2017–2020.

Statistics	U (ML d^{-1})	Taver (°C)	Tmax (°C)	Pcum (mm d^{-1})	Pmin (mm h^{-1})	Pmax (mm h^{-1})
			2017–2019			
Minimum	464.14	−2.80	0.50	0.00	0.00	0.00
Maximum	791.85	32.00	38.90	79.40	0.20	38.60
Mean	601.14	14.97	20.28	2.57	0.00	0.96
1st Quartile	575.25	7.80	12.30	0.00	0.00	0.00
Median	595.72	14.80	20.20	0.00	0.00	0.00
3rd Quartile	620.51	22.85	28.85	0.60	0.00	0.40
Standard deviation	48.36	8.46	9.27	7.46	0.01	2.80
			2020			
Minimum	472.48	−0.10	2.00	0.00	0.00	0.00
Maximum	658.22	30.10	36.10	77.00	0.00	43.20
Mean	567.73	14.74	20.01	2.80	0.00	1.10
1st Quartile	545.64	7.93	12.73	0.00	0.00	0.00
Median	567.25	14.25	19.70	0.00	0.00	0.00
3rd Quartile	589.47	21.50	27.30	0.60	0.00	0.40
Standard deviation	32.55	7.86	8.51	8.75	0.00	4.09

graphical depiction underscores the importance of considering external factors such as climate and societal changes when analyzing UWD trends.

As previously noted in figure 2, the UWD in Milan exhibits both inter-annual and intra-annual variability, which may be influenced by meteorological fluctuations. Characterized by a humid subtropical climate, Milan experiences hot summers, cold winters, and moderate to heavy rainfall throughout the year. Consequently, we have incorporated potentially relevant daily meteorological variables into our dataset, drawing from data collected at the Lambrate meteorological station. This integration allows for a comprehensive analysis of the impact of weather conditions on water consumption patterns. These include daily average temperature (T_{aver}), daily maximum temperature (T_{max}), daily cumulative precipitation (P_{cum}), maximum hourly cumulative precipitation (P_{max}), and minimum hourly cumulative precipitation (P_{min}). Table 1 shows a summary of the statistics of each variable in our dataset. The dataset reveals that the minimum daily UWD slightly increased from 464.14 megaliters per day (ML d^{-1}) in the pre-pandemic years to 472.48 ML d^{-1} in 2020, while the maximum demand notably decreased from 791.85 ML d^{-1} to 658.22 ML d^{-1}, reflecting a significant shift in peak consumption patterns during the COVID-19 pandemic year. Concurrently, the average daily demand also saw a reduction from 601.14 ML d^{-1} to 567.73 ML d^{-1}, with the standard deviation narrowing from 48.36 ML d^{-1} to 32.55 ML d^{-1}. This reduction in variability of daily consumption aligns with the trends observed in figure 2, suggesting a consistent pattern of decreased water usage. Meteorological observations from the same period revealed only minor changes in average temperatures, yet indicated an increase in minimum temperatures, suggesting warmer conditions during typically colder days. Additionally, there was a significant rise in the intensity of precipitation events, with the maximum hourly precipitation reaching 43.20 mm h^{-1} in 2020, up from 38.60 mm h^{-1} in previous years. This escalation in precipitation intensity, together with the slight increase in baseline temperatures, points to a complex interaction between climatic conditions and UWD patterns. These observations highlight the imperative to incorporate meteorological dynamics into UWD forecasting.

2.2. DL models for UWD forecasting

In our research, the main DL models we implement are LSTM and its state-of-the-art variant with an AM-LSTM. LSTM was first proposed by [52] as an improvement of conventional RNN model to learn long-term dependencies when the prediction is correlated with observations occurred far earlier. LSTM has a sequence of modules instead of simple layers, where three gates control the information passed to the cell states (C_t). These gates are a forget gate (with corresponding forget function f_t), an input gate (with corresponding input function i_t), and an output gate (with corresponding output function o_t). These layers are organized in a forward sequence composed of sigmoid neural network layers (σ) and pointwise multiplication actions ($\otimes$). C_t represents the long-term memory of data information. The fundamental equations of a LSTM are formulated in [53] as follows:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{2}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{3}$$

$$\tilde{C}_t = \tanh\left(W_C x_t + U_C h_{t-1} + b_C\right) \tag{4}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \tag{5}$$

$$h_t = o_t \otimes \tanh\left(C_t\right). \tag{6}$$

Here $x_t \in R^k$ is a k-dimensional vector representing the input time series at time t. In the first forget gate, the forget activation f_t is calculated by combining the current data x_t and the previous output vector h_{t-1} through a σ activation function (equation (1)), and then multiplied with the previous cell state (C_{t-1}), representing how much information is discarded or kept in C_{t-1}. Then, new information is processed and important parts are kept in the current cell state (C_t). The input activation i_t produced by the input gate (equation (2)) is multiplied by the output $\tilde{C}_t$ (equation (4)) from a *tanh* activation function to decide how much the new input is added to C_t. These two results are added together to form the C_t shown in equation (5). Finally, the output of the current module (h_t) is the element-wise product of output activation o_t of the output gate and the current cell state C_t transformed by the function *tanh*. All outputs of the activation functions (f_t, i_t and o_t) are calculated using weights (W_f, W_i, W_o, U_f, U_i, U_o) and biases (b_f, b_i, b_o) determined in the training process. The number of hidden layers and the number of neurons are two important hyperparameters of the LSTM structure. In this study, we develop a DL structure by concatenating LSTM layers and the number of layers is chosen between 3 and 4. The number of neurons in each layer is chosen within a grid in the range 1–100.

The AM, modeled after the selective focus of the human brain, can organise and manage essential information efficiently within the limits of available processing resources [54]. Recent advances in DL methods with an AM have garnered interest in time series forecasting, with few pioneering studies exploring its application in UWD forecasting (e.g. [40, 47]). Evidence suggests that these models (especially AM-LSTM) can outperform benchmark models. In this research, we incorporate the AM as a subsequent layer following the LSTM layers. This configuration is designed to enhance the model ability to dynamically focus on pertinent segments of the input data, thereby improving its processing efficiency and predictive accuracy while not increasing its computational complexity significantly [55]. The mechanism is formulated here in equations (7)–(10). First (equation (7)), the AM calculates the attention weights α_{ts} based on the hidden state at each timestep t, denoted as h_t, which is derived from the final LSTM layer. These weights are computed for all hidden states h_s, each representing individual timesteps within the input sequence. Specifically, the weight for each state h_s is determined as an exponential function of the score function comparing h_t and h_s, normalized by the sum of exponentials of the score function computed between h_t and every other state $h_{s'}$ in the sequence (where s' ranges from 1 to S). Subsequently, these normalized weights quantify the contribution of each h_s towards the composition of the context vector at each output timestep t. The equation for α_{ts} is presented below:

$$\alpha_{ts} = \frac{\exp\left(\text{score}\left(h_t, h_s\right)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}\left(h_t, h_{s'}\right)\right)}. \tag{7}$$

The attention weights α_{ts} are computed using a scoring function that quantifies the importance or relevance of each input sequence element h_s to the output at time t. This scoring function can be of different types, e.g. additive [55] or multiplicative [56]. In our research, the Bahdanau additive AM [55] is applied and is calculated as the following:

$$\text{score}\left(h_t, h_s\right) = v_a^{\top} \tanh\left(W_1 h_t + W_2 h_s\right). \tag{8}$$

Both h_t and h_s are first transformed by their respective weight matrices W_1 and W_2. The transformed states are summed and the tanh function is applied to the sum, which helps to normalize and stabilize the learning process. The result is then reduced to a single scalar through a dot product with transposed v_a to convert the output of the tanh function into a scalar score.

Then the context vector c_t is the weighted sum of all hidden states h_s, where the weights are the attention weights α_{ts}. This vector serves as a summary of the input sequence, emphasizing the parts deemed most relevant by the AM:

$$c_t = \sum_{s=1}^{S} \alpha_{ts} h_s. \tag{9}$$

The final attention vector a_t is derived from the context vector c_t concatenated with the current hidden state h_t and passed through a tanh function. This process allows integration of both the immediate and the broader contextual information:

$$a_t = \tanh\left(W_c\left[c_t; h_t\right]\right). \tag{10}$$

All weights in the above equations are tuned during model training. The Bahdanau AM overcomes the limitations of fixed-length context vectors in traditional encoder–decoder models by allowing dynamic focus on different parts of the input sequence. Further details about the Bahdanau AM can be found in [55]. A critical parameter in the attention layer in this research is the number of neurons in the attention vector a_t, which we fine-tune during hyperparameter optimization within the range 1–200.

2.3. Wavelet transform

WT has been widely applied in signal processing to identify frequencies in data corresponding to information from the time domain and improve the understanding of characteristics of time series data [57]. In general, the original time series data can be decomposed into sub-components, namely wavelet and scaling coefficients, representing useful low- and high-frequency information. DWTs have been regarded an efficient preprocessing tool in hydrology and water resources to improve forecasting capacity with low computational effort [58, 59]. However, as pointed out by [60] and [49], wavelet decomposition has been sometimes misused in research, leading to overoptimal forecasting results in real-world applications. Possible causes include the use of future data, inappropriate selection of the decomposition level and filter, and incorrect partitioning of calibration and validation data, which are related to the BC problems [61]. To address the disadvantages of DWT applications in hydrological foresting, the WDDFF was proposed in [49] which incorporates a couple of best practices to develop an appropriate hybird WT-ML model (i.e. decomposition algorithm selection, calculation of 'boundary-corrected' wavelet and scaling coefficients, calibration and validation on adjusted data, etc). The prerequisite of WDDFF is to use the maximal overlap DWT (MOWDT) or the à trous algorithm (AT) instead of the DWT, since the DWT cannot be adjusted to eliminate BC issues due to its decimation property [60]. In this study, we implement the MOWDT proposed by [62] as the decomposition algorithm. The details of MOWDT are introduced in appendix B. We also utilize four popular wavelet filters for feature extraction: Haar wavelet (*haar*), Daubechies wavelet of order 2 (*db*2), Coiflet wavelet of order 1 (*coif*1), and Symlet wavelet of order 2 (*sym*2). *haar* is known for its simplicity and effectiveness in detecting sudden changes in a signal. *db*2 offers smoothness and fine detail capture with its two vanishing moments. *coif*1 provides high accuracy in signal reconstruction due to its near-symmetric properties and one vanishing moment. Lastly, *sym*2 balances compact support and symmetry with its two vanishing moments. Further details on these wavelet filters can be found in [63].

2.4. Experimental setup

In our study, we first train and test different classes of models to forecast UWD for two lead times, i.e. 1 day and 7 days, based on time series data collected from 2017 to 2019. For SVR, ANN, LightGBM, LSTM and AM-LSTM, we develop both standalone models using original data and hybrid WT-LightGBM-ML/DL models using wavelet-decomposed data. We then re-train the best performing model on an extended pre-pandemic dataset spanning from January 2017 to March 2020. To ascertain its robustness, we then evaluate its performance in predicting UWD post-March 2020, under conditions it did not previously encounter. The complete multi-step daily UWD forecasting framework we developed is depicted in figure 3.

During the feature engineering step, we conduct a preliminary correlation analysis using Spearman's rank correlation coefficient (table 2) to investigate the potential influence of various weather-related variables on UWD. We identify average temperature (T_{aver}), maximum temperature (T_{max}), and cumulative precipitation (P_{cum}) as significantly correlated external variables. For time series-based models (AR, SARIMAX, LSTM, and AM-LSTM), we maintain the data in its original sequential format. To accommodate the absence of time indices in tabular data used in models such as LightGBM, ANN, and SVR, we introduce additional dummy variables for temporal features: months ($Mon_1, Mon_2, \ldots, Mon_{12}$), weekdays/weekends (*Wek*), and holidays (*Holy*). Considering the weekly fluctuation in daily UWD patterns, we include all data from the previous 14 days as input variables. Subsequent to feature engineering, feature selection is performed by LightGBM to identify significant variables in non-wavelet-decomposed and wavelet-decomposed data for further development of the predictive models. Feature importance in LightGBM is computed once for each predictor in the analysis. This method quantifies both the frequency and impact of each feature contribution to the decision-making within the model trees. Other models such as ANN, SVR, and LSTM lack intrinsic mechanisms for direct feature importance evaluation and instead rely on the features identified by LightGBM. Utilizing LightGBM for feature selection not only improves the training efficiency, but also enhances the prediction accuracy of these models, compensating for their inability to independently assess feature relevance. We calculate the feature importance values for each predictor used in LightGBM using the split method. This method represents how many times a feature is used for splitting the data across all trees in the model. Specifically, a feature importance increases with the frequency of its use in splits. This method

Table 2. Spearman's rank correlation coefficient of U and meteorological variables.

	U	P_{cum}	P_{min}	P_{max}	P_{occ}	T_{aver}	T_{min}	T_{max}
U	1	−0.157	−0.024	−0.149	−0.153	0.396	0.369	0.405
P_{cum}	−0.157	1	0.051	0.997	0.969	−0.152	−0.027	−0.243
P_{min}	−0.024	0.051	1	0.044	0.035	−0.022	−0.012	−0.033
P_{max}	−0.150	0.997	0.044	1	0.969	−0.136	−0.015	−0.223
P_{occ}	−0.153	0.969	0.035	0.969	1	−0.170	−0.054	−0.250
T_{aver}	0.396	−0.152	−0.022	−0.136	−0.170	1	0.970	0.979
T_{min}	0.369	−0.027	−0.012	−0.015	−0.054	0.970	1	0.913
T_{max}	0.405	−0.243	−0.033	−0.223	−0.250	0.979	0.913	1

effectively pinpoints the most influential features in the model, emphasizing those that contribute most to data segmentation. In the wavelet decomposition phase, we experiment with various filters (*haar*, *db*2, *coif*1, and *sym*2). Ultimately, we calibrate and evaluate standalone and hybrid WT-LightGBM-ML/DL models using a set of different error and performance metrics, namely, root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), Nash–Sutcliffe efficiency (NSE), and Kling–Gupta efficiency (KGE). Detailed calculations are shown in appendix C.3. The number of autoregressive terms in AR is set to 14, considering weekly patterns in UWD data. The optimal structure of the SARIMAX model is determined automatically by the auto-ARIMA algorithm implemented in the Python library pmdarima [64], which determines the order of p, d, q and P, D, Q by minimizing the Akaike information criterion value. AR and SARIMAX models are trained on the data in 2017–2018 and evaluated on the whole year of 2019, since the model requires the complete seasonality in a year to learn from. To prevent overfitting, we reserve 20% of the observations from the end of the training time series as a validation sample, which is not used in the training process. Employing this method allows maintaining the temporal sequence of the dataset for models analyzing time series data, thereby preventing lookahead bias. For ML/DL models (ANN, SVR, LightGBM, and LSTM), model calibration is performed according to the following empirical procedure:

(i) **Data partitioning**: after boundary condition correction for both WT and non-WT settings, data in 2017–2019 are separated into 80% as training data for calibration with k-fold cross-validation (k = 3) and 20% as test data for evaluation.

(ii) **Data scaling**: ML/DL models are sensitive to the different scales of various features. To avoid favoring only features that have larger value ranges, a min-max scaler is applied to scale the input-output space to the range 0 to 1:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{11}$$

where X is the vector of original observations and X' is the transformed (scaled) vector. X_{min} and X_{max} are the minimum and maximum values of input variables, respectively.

(iii) **Hyperparameter tuning**: in this study, hyperparameters of LightGBM, ANN and SVR are tuned by applying grid search [65] of decision-theoretic methods. In grid search, a specific parameter space is defined for each ML model according to empirical experience and data set, then the algorithm exhaustively searches over the space and finds the optimal combination of different parameters. Here, a typical k-fold cross-validation step is applied to avoid over-fitting ($k = 3$), and then the best set of parameters is decided by looking at the minimization of MSE on validation data. For both LSTM and AM-LSTM models, a 20% segment of the most recent training data is designated as a validation set to prevent lookahead bias in time series analysis. The bandit-based approach implemented in Keras [66] is adopted for hyperparameter optimization of LSTM and AM-LSTM models, as it improves random search and ensures efficiency. By applying adaptive resource allocation and early-stopping, the algorithm can largely speed up the searching over the parameter space and retain optimal candidates to select the best parameterization. A large number of 1000 epochs ensures thorough learning and adjustment of model weights, optimizing performance by iteratively evaluating both training and validation data.

A comprehensive explanation of all experimental settings is provided in appendix C.

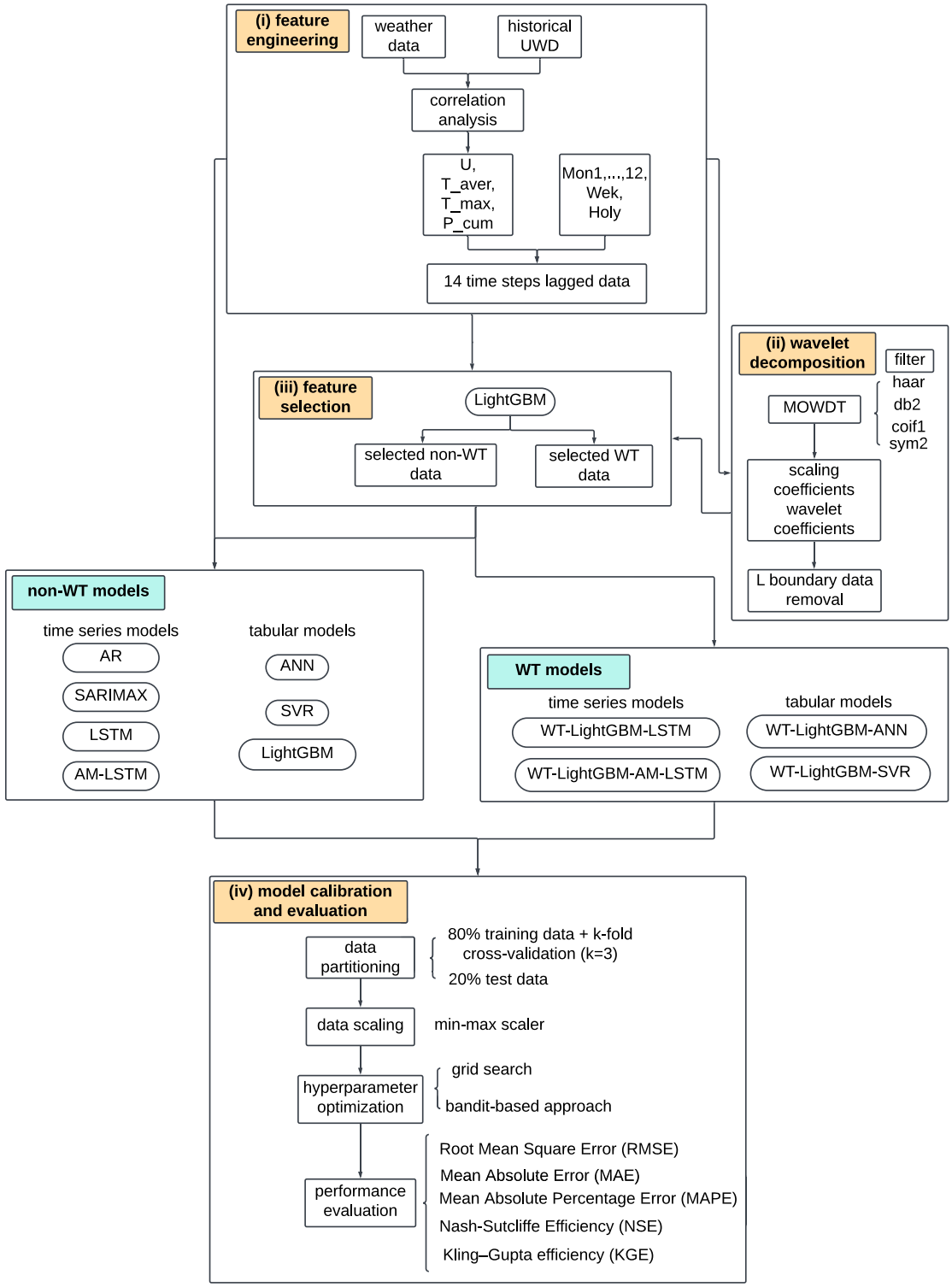

Figure 3. The multi-step daily UWD forecasting framework proposed in this research. Four main steps are implemented: (i) feature engineering, (ii) wavelet decomposition, (iii) feature selection, and (iv) model calibration and evaluation.

3. Results and discussion

3.1. Comparative model performance for 1 d-ahead prediction

The performance of the 1 d-ahead prediction model for the 2019 test phase of all models developed in standalone and WT-LightGBM-ML/DL settings are compared in figure 4. In general, most models can achieve satisfactory results with NSE $>$ 0.8, MAPE $<$ 0.03, and RMSE from 2.6 to 24.9 ML d^{-1} in the test phase. According to similar research on daily UWD forecasting [9, 12, 49], RMSE $<$ 27 ML d^{-1}, MAPE $<$ 0.02, NSE $>$ 0.9 are usually very satisfying performance values for 1 d-ahead prediction. In the non-wavelet

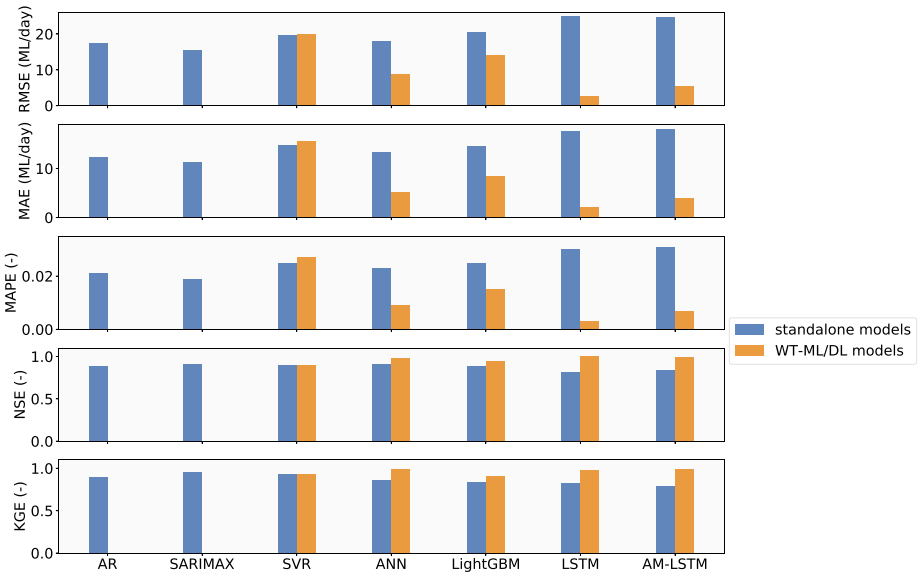

Figure 4. Performance comparison of 1 d-ahead prediction achieved by standalone models (blue bars) and their corresponding best WT-LightGBM-ML/DL models (orange bars) in the test phase. From top to bottom: RMSE, MAE, MAPE, NSE and KGE. From left to right: AR (only trained on non-decomposed data), SARIMAX (only trained on non-decomposed data), SVR, ANN, LightGBM, LSTM, and AM-LSTM.

setting, LightGBM achieves the best performance in the training phase (see appendix 1), but its performance deteriorates in the test phase. Meanwhile, ANN and SARIMAX outperform the other models in the test phase, with NSE of 0.911 and 0.906, and RMSE smaller than 20 ML d^{-1}. These two are typical ML and time series models, which are usually used as benchmark models in comparative research (see, e.g. [9, 13]). In our study, LightGBM is initially trained to make forecasts and identify important features. These selected features are then used to train the ANN, SVR, and (AM)-LSTM models. The high performance of the ANN model in figures 4 and 5, resulting from the use of these selected predictors, illustrates the suitability of LightGBM as a feature selection tool, which enhances model training efficiency and prediction accuracy. Additionally, lower performance is observed when all predictors are used for the ML/DL models in table 1, likely due to collinearity issues in the multivariable dataset [67], which further confirms the necessity of using feature selection techniques. WTs, in general, improve the performance of most types of models for 1 d-ahead predictions, demonstrating the usefulness of extracting the most relevant information for forecasting tasks from the available data. However, SVR models do not enhance their forecasting ability with wavelet decomposition, resulting in similar NSE in the range of 0.85–0.89 and KGE in the range of 0.77–0.93 for SVR and WT-LightGBM-SVR models, respectively (see figure 4). Especially for WT-LightGBM-SVR model forecasting *sym*2-decomposed data, although all four metrics indicate optimal performance, the lowest is KGE, with a value of 0.767. According to [68], KGE cannot be simply compared with NSE in a straightforward manner, as it is a valuable diagnostic tool that helps identify the weaknesses and strengths of the forecasts, while other scores, such as MAE and NSE, are general model performance indicators. It is also worth noting that filter selection is critical to model performance. For tabular data prediction, the best model performance of WT-LightGBM and WT-LightGBM-ANN is achieved using *coif*1-decomposed data and *sym*2-decomposed data, respectively. For time series data prediction, WT-LightGBM-LSTM performs best on *haar*-decomposed data with NSE of 0.998 and RMSE of 2.603 ML d^{-1}, while WT-LightGBM-AM-LSTM achieves the best performance with *sym*2-decomposed data (see appendix 1). For 1 d advance forecasting of UWD, integrating an AM into LSTM yields a performance comparable to that of conventional LSTM. Overall, filters with wider length (*db*2, *sym*2: filter length = 4, *coif*1: filter length = 6) seem appropriate for tabular data prediction and *haar* (filter length = 2) appears to be more suitable for time series data in this study. Similar findings are also discussed in the literature on daily and hourly UWD prediction and monthly groundwater level prediction [11, 49, 69]. These studies show that WTs do not guarantee improved performance of ML models with respect to different filters and decomposition levels. Especially for SVR, similar performance degradation is observed for short-term forecasting [69]. In our research, only a few filters are explored due to the constraint of the length of dataset, in light of the removal of BC-influenced data. However, many previous studies ignored BC issues and tested a wide range of filters and decomposition levels, which usually results in unrealistic high performance [49]. All in all, there is no specific filter that can

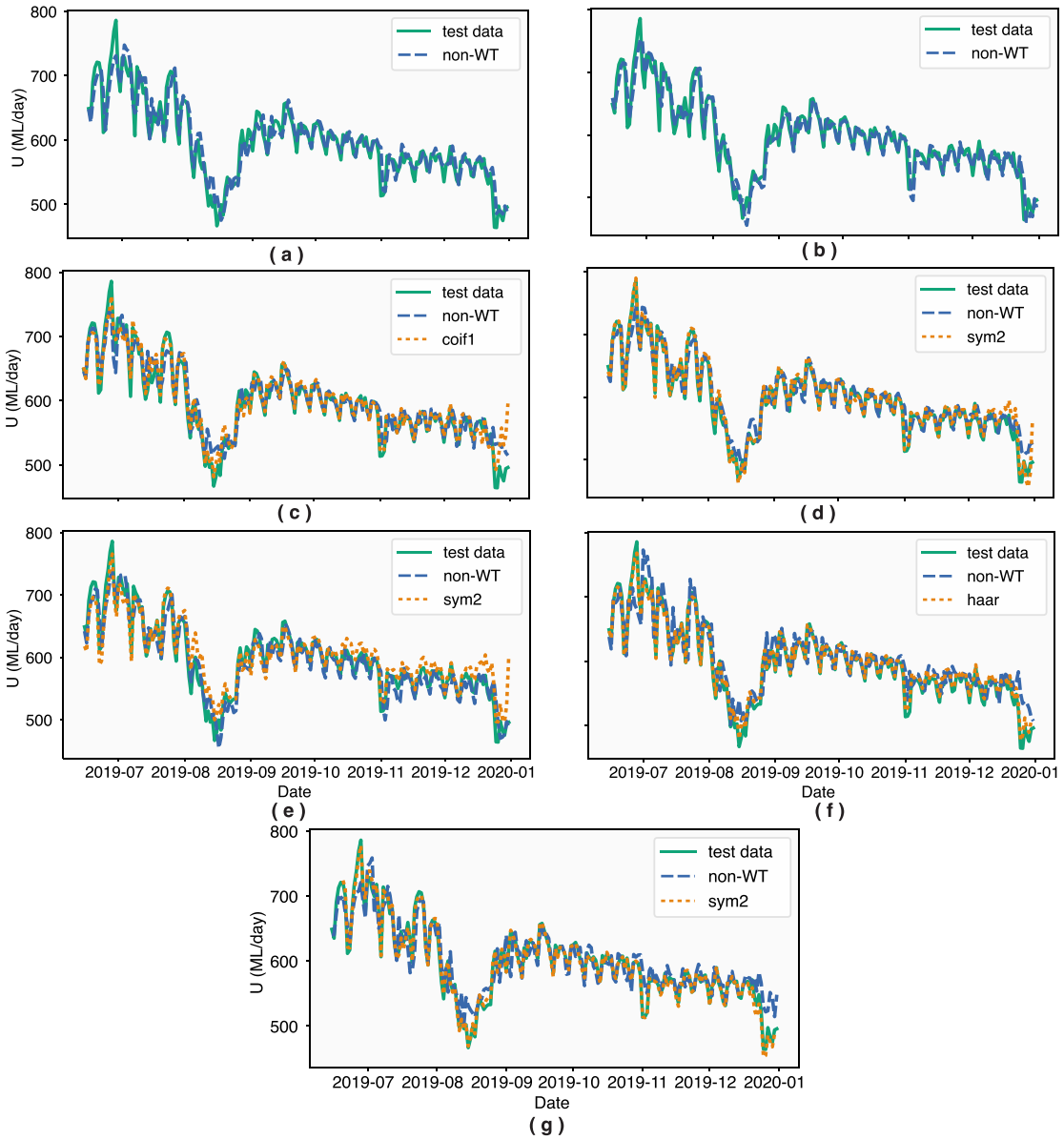

Figure 5. Comparison between predicted and observed daily UWD for 1 d-ahead forecasting during the test period in 2019 for both standalone and WT-LightGBM-ML/DL models. (a): AR; (b): SARIMAX; (c): LightGBM; (d): ANN; (e): SVR; (f): LSTM; (g): AM-LSTM. In each line plot, the green line represents observations, the blue dash line represents model predictions in a non-wavelet setting, and the orange dash line represents model predictions in a wavelet setting with the corresponding filter.

lead to the best performance for different ML models, thus an exploratory analysis of various filters and decomposition levels is suggested for hybrid WT-LightGBM-ML/DL model development.

A comparison of the observed and predicted UWDs of the best hybrid models of WT and non-WT is illustrated in figure 5. Except for the SVR model, the UWD time series predicted by the WT-LightGBM, WT-ANN, WT-LSTM and WT-AM-LSTM models are closer to the real data (see figures 5(c), (d), (f) and (g)). Small differences can be observed as the WT-LightGBM model overestimates low UWD and slightly underestimates peak UWD, whereas WT-ANN slightly overestimates medium UWD levels. In contrast, WT-LSTM and WT-AM-LSTM demonstrate near-perfect prediction accuracy. Importantly, simpler models such as AR and SARIMAX also produce comparable and satisfactory results in 1 d-ahead forecasting. This suggests the merit of initially employing a straightforward time series model for short-term UWD forecasting, which can serve as a valuable benchmark for further comparative analysis.

Overall, all models developed within this study for 1 d-ahead UWD forecasting exhibit satisfactory accuracy. The SARIMAX model, distinguished by its explicit integration of trends and seasonality, offers a more straightforward and interpretable formulation. Its structural design facilitates easier understanding and

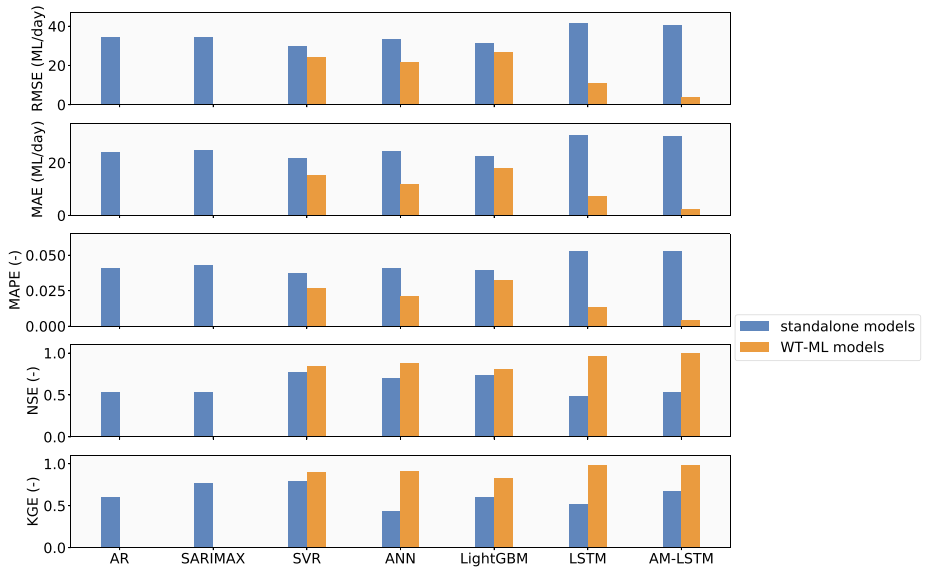

Figure 6. Performance comparison of 7 d-ahead prediction between standalone models (blue bars) and their corresponding best WT-LightGBM-ML/DL models (orange bars) in the test phase. From top to bottom: RMSE, MAE, MAPE, NSE and KGE. From left to right: AR (only trained on non-decomposed data), SARIMAX (only trained on non-decomposed data), SVR, ANN, LightGBM, LSTM, and AM-LSTM.

application, making it an alternative for scenarios requiring quick deployment and straightforward training processes. Conversely, while ML and DL models, particularly the (AM)-LSTM, demonstrate robust predictive abilities, their intricate data handling and training requirements may not provide substantial benefits over SARIMAX for such a limited forecasting horizon. For practical applications on forecasting a single time step ahead, SARIMAX often emerges as the more pragmatic and straightforward alternative in real-world settings.

3.2. Comparative model performance for 7 d-ahead prediction

Figure 6 displays the performance of various standalone and hybrid WT-LightGBM-ML/DL models during the test phase, evaluated across different metrics for 7 d-ahead prediction. In general, the predictive performance of all models deteriorates compared to the 1 d-ahead forecasting results. AR and SARIMAX achieve the worst performance in the test phase for most performance metrics, except KGE. In the literature, AR and ARIMA models are usually developed for 1 lead time prediction and can achieve satisfactory results, as shown in section 3.1. A few studies also explored the possibility of predicting sequential data and obtained a similar conclusion that ARIMA models have limited capacity for sequence prediction tasks [9, 42]. Meanwhile, some ML models using tabular data (SVR and LightGBM) achieve comparable and relatively satisfying results with NSE around 0.75 and RMSE around 30 ML d^{-1} in the test phase of the non-wavelet setting. The performance of these models improves with wavelet decomposition to NSE in the range 0.8–0.87 and RMSE reduced to 21–26 ML d^{-1}. Similarly to the 1 d-ahead predicting, there is not a specific filter fit for all models, as the best performance is achieved by applying various filters for different models (see table 1 in appendix E). However, all models using *coif*1-decomposed data achieve overarchingly satisfactory performance. The potential reason is that *coif*1 has the longest length (6) among all filters. A comparable research [49] also predicts 7 d lead time UWD and finds predictive model uses data decomposed by a filter of length 14 (the longest length) can obtain the best performance. Long filter length combined with higher decomposition level result in smooth scaling coefficients which record interannual, annual, and intraannual changes that represent the main patterns of UWD free of short-term disturbances [62]. Meanwhile, long length filters can capture well the typical periodic weekly patterns of UWD. The best performance on four evaluation metrics is still achieved by the AM-LSTM model that predicts the *haar*-decomposed data: NSE = 0.996, RMSE = 3.894 ML d^{-1}, MAE = 2.543 ML d^{-1}, MAPE = 0.004, KGE = 0.985. Even if *haar* has a shorter length, the special data construction for the LSTM model considers the time consistency well to capture periodicity in the data. Lastly, by analyzing the sensitivity of different thresholds applied to select informative input variables, the most useful features emerge within a small range (top 20), as summarized in table C1. This result can be related to the findings in the following section 3.3 that applying wavelet decomposition will improve the importance values of a few critical features which are the most influential ones for model performance.

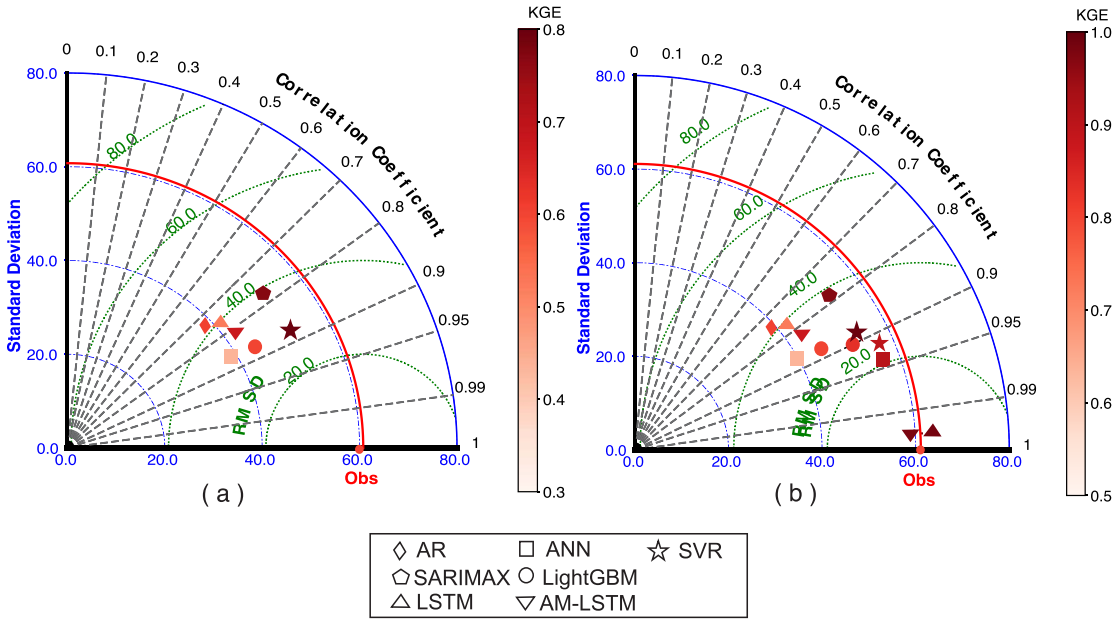

Figure 7. Taylor diagrams comparing the following 7 d-ahead UWD forecasting performance for: (a) AR, SARIMAX, ANN, SVR, LightGBM, LSTM and AM-LSTM; (b) WT-LightGBM-ANN, WT-LightGBM-SVR, WT-LightGBM, WT-LightGBM-LSTM and WT-LightGBM-AM-LSTM. These diagrams illustrate the statistical performance of each model in terms of their correlation, standard deviation, and root-mean-square deviation from observed data. The colour gradient from light red to dark red represents the range of KGE values, with darker shades indicating higher values and lighter shades representing lower values. Models closer to the observation (Obs) point indicate higher accuracy.

Additionally, we visualize two Taylor diagrams for standalone and optimal WT-LightGBM-ML/DL models in figure 7. Taylor diagrams offer graphical representations that facilitate a comprehensive statistical evaluation of 7 d-ahead forecasting performance of non-WT and WT models. Taylor diagrams combine multiple aspects of model accuracy by simultaneously showing three key metrics: the correlation coefficient, the mean square difference (RMSD), and the standard deviation [70]. The performance of the model is plotted in a two-dimensional space where the radial distance from the origin represents the standard deviation of the model results, the angular displacement indicates the correlation, and the distance from the reference point (typically representing the observed data) reflects the RMSD. In our plots, the KGE values are incorporated as the fourth metric, with darker red denoting higher values and lighter red indicating lower values. This format allows for a quick visual assessment of how closely model outputs match observations. Higher correlation, smaller RMSD, and matching standard deviations with observations generally indicate better model performance. As a result, in the non-wavelet setting (figure 7(a)), all models exhibit similar outcomes, with ML/DL models showing slightly superior performance in terms of correlation and RMSD. Notably, SARIMAX also demonstrates relatively higher KGE value. In contrast, the application of wavelet-enhanced feature selection notably improves the performance of neural-based models, particularly LSTM and AM-LSTM, which are positioned close to the observation point. This finding aligns with recent research indicating that the WT tends to enhance neural-based models more effectively [71].

A comparison of predicted and observed values in the test phase is shown in figure 8. The prediction results of AR and SARIMAX have a hysteresis effect shown in figures 8(a) and (b) that leads to degraded performance. For (c)–(e), similar trends can be observed that LightGBM, ANN and SVR overestimate low UWD especially at the end of December for both WT and non-WT settings. The improvement of performance is more evident for LSTM and AM-LSTM by applying the wavelet decomposition technique. The standalone LSTM and AM-LSTM models cannot provide an accurate prediction in test data with the lowest NSE of 0.487, even though they performed well in training data. Specifically, this model overestimates low UWD during summer/Christmas holidays in August/December and underestimates peak UWD in July (shown in (f) and (g)). The possible reason is that the information on holiday is not considered in standalone models, influencing the 7 d-ahead forecasting. By applying wavelet decomposition, low-frequency information representing the main patterns of time series data become more precise and relevant high-frequency information representing weekly patterns/monthly patterns are preserved. Also, time information is still embedded in the decomposed data. As a result, WT-LSTM and WT-AM-LSTM can model the transformed time series data better than other ML models modelling transformed tabular data. Figure 9 further underscores the temporal performance variations illustrated by the weekly RMSE values of

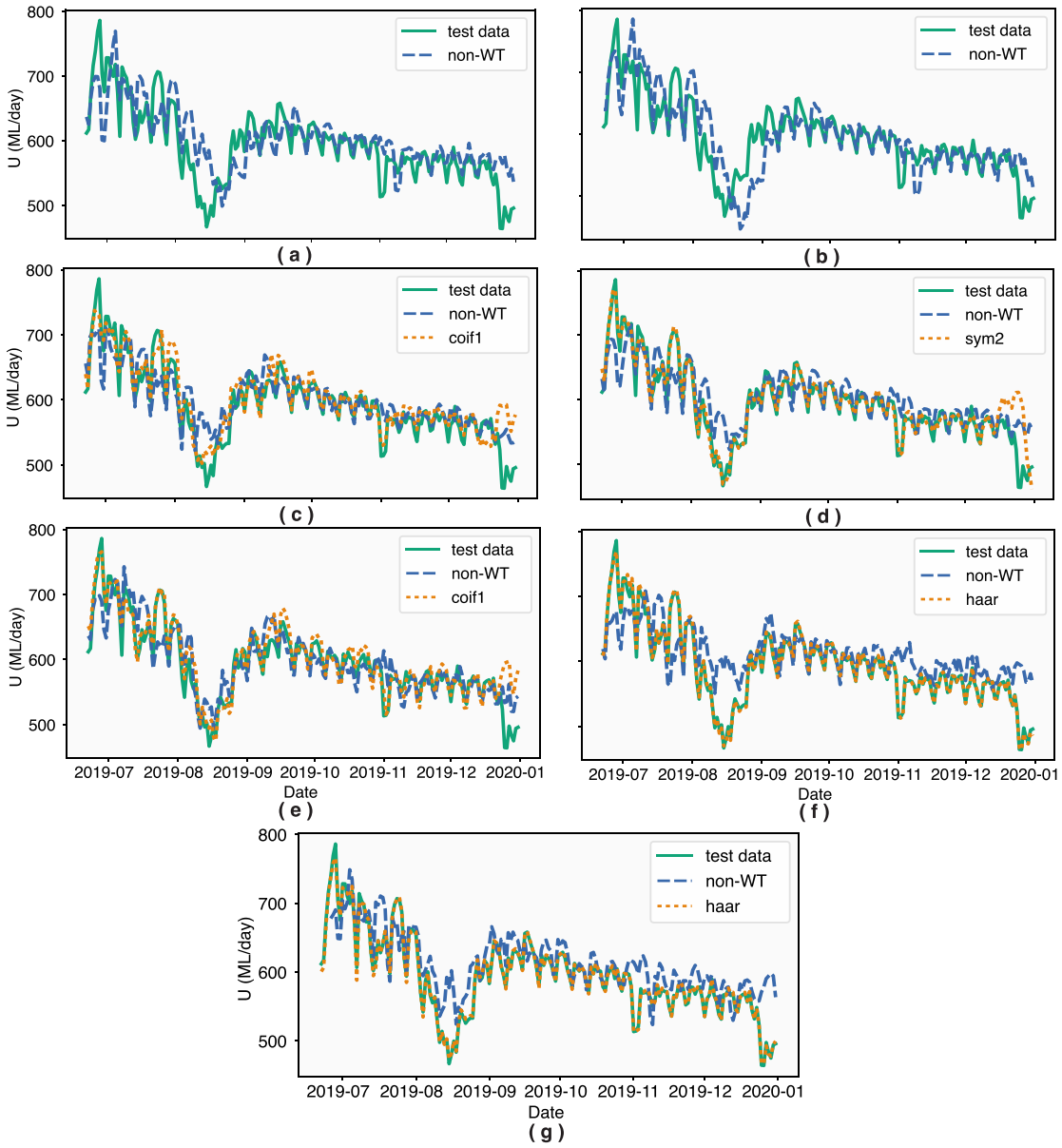

Figure 8. Comparison between predicted and observed daily UWD for 7 d-ahead forecasting during test period in 2019 of standalone and WT-LightGBM-ML/DL models. (a): AR; (b): SARIMAX; (c): LightGBM; (d): ANN; (e): SVR; (f): LSTM; (g): AM-LSTM. In each line plot, the green line represents observations, the blue dash line represents model predictions in a non-wavelet setting, and the orange dash line represents model predictions in a wavelet setting with the corresponding filter.

all models for 7 d-ahead forecasting during the test period in 2019. The hybrid models (WT-LightGBM, WT-ANN, WT-SVR, WT-LSTM, WT-AM-LSTM) consistently exhibit lower RMSE values compared to their non-wavelet counterparts and baseline models throughout the entire period. In the weeks 29 December 2019 and 5 January 2020, the performance of WT-LightGBM, WT-ANN, and WT-SVR is slightly inferior to that of their standalone counterparts. According to figure 2, the water demand during the corresponding two-week periods in 2017 and 2018 was higher than in 2019 and 2020. The WT-enhanced ML models especially WT-LightGBM and WT-SVR, trained on these earlier patterns, may have extrapolated similar high values to the later period, potentially explaining the observed decrease in model performance during these specific dates. WT-LSTM and WT-AM-LSTM still frequently demonstrate superior predictive performance, as evidenced by their lower RMSE (below 10 ML d^{-1}) compared to other models. This superior performance is particularly evident during the weeks in August, characterized by a decrease in demand due to the summer holiday. During this period, the standalone models exhibit high RMSE (values around 60–80 ML d^{-1}), whereas the hybrid neural-based models (WT-ANN, WT-LSTM, and WT-AM-LSTM) maintain stable and strong performance with RMSE values below 20 ML d^{-1}, highlighting their robustness in handling significant demand fluctuations. Additionally, the baseline AR and ARIMAX models exhibit prediction lags, as shown in figure 8, and underperform in 7 d-ahead UWD forecasting due to their inherent reliance on

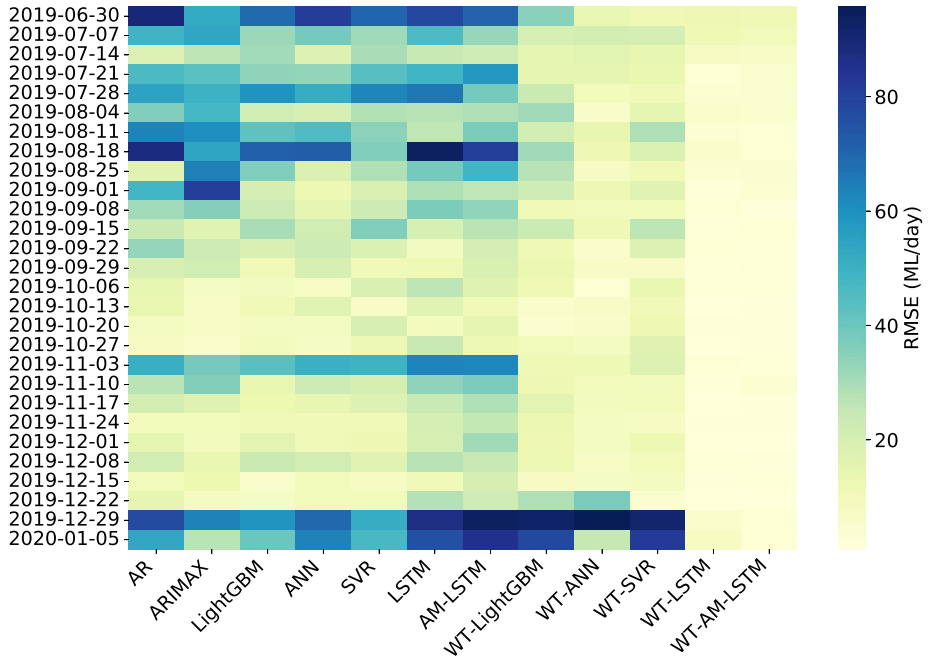

Figure 9. Heatmap of weekly RMSE (ML d^{-1}) for different models for 7 d-ahead forecasting during test period in 2019. Each row represents a week, and each column represents a model. The models include AR, ARIMAX, LightGBM, ANN, SVR, LSTM, AM-LSTM, and wavelet-transformed variants of these models (WT-LightGBM, WT-ANN, WT-SVR, WT-LSTM, WT-AM-LSTM). The color bar on the right indicates the RMSE values, with lighter colors representing lower RMSE values and darker colors representing higher RMSE values.

linear relationships and short-term dependencies. These models struggle to capture the complex, non-linear patterns and dynamic seasonal trends present in UWD data, leading to delayed and less accurate predictions.

The robustness of hybrid neural-based models can be attributed to several factors. Firstly, wavelet transformation decomposes the original time series into different frequency components, enabling the models to capture both short-term fluctuations and long-term trends more effectively. This decomposition allows the models to isolate and process the seasonal and irregular variations separately, which becomes particularly useful during periods of significant demand shifts such as summer holidays. Secondly, LSTM and AM-LSTM are inherently capable of learning complex long-term temporal dependencies and non-linear relationships within time series data. This capability is crucial for 7 d-ahead forecasting, where capturing the underlying demand trend is essential for accurate predictions. However, in the last two weeks of December, even the WT-ANN model struggles to accurately predict water demand, whereas the hybrid WT-LSTM and WT-AM-LSTM models continue to perform well. This discrepancy can be attributed to the nature of the demand patterns during this period. The drop in demand at the end of December is more abrupt compared to the steady and predictable decline observed in August. The consistent decrease in August is easier for models to capture, while the sharp fluctuations in December present a greater challenge. The shallow structure of the ANN model is less capable of capturing the complexity of such abrupt changes. In contrast, the deep LSTM model is better equipped to handle these sudden variations in data. This highlights the LSTM ability, enhanced by wavelet transformation, to learn long-term dependencies and non-linear relationships, enabling it to adapt more effectively to the dynamic nature of water demand influenced by external stressors.

Italy was the first European country to declare lockdown measures to contain the spread of COVID-19 in Europe starting from 9 March 2020 to 18 May 2020. Large-scale behaviour changes can lead to UWD changes, as demonstrated in a recent work analyzing water use in California during COVID-19 [72]. Their results show that UWD decreased during the lockdown in response to the stay-at-home orders. According to the exploratory analysis of UWD shown in table 1, the average UWD in Milan in 2020 also decreased compared to previous years and had smoother summer peaks. During the lockdown period, a decreasing trend is also observed compared with same period in 2017–2019 (figure 2). To evaluate the robustness of the developed forecasting models on unseen patterns, we re-train the best models (WT-SVR, WT-ANN, WT-LightGBM, WT-LSTM and WT-AM-LSTM) for the 7 d-ahead lead time forecasting on pre-pandemic data (1 January 2020–08 March 2020) and test them on data during pandemic period (09 March 2020–31 December 2020). Model performance values are summarized in table 3 and the comparison between predicted values and actual observations are plotted in figure 10. According to the performances in table 3, all

Table 3. 7 d ahead forecasting results before and after the COVID-19 lockdown for different types of predictive models. The best performance achieved by different models and filters during the training and test phases is highlighted in bold.

Models	Filter	RMSE (ML d^{-1})	MAE (ML d^{-1})	MAPE (—)	NSE (—)	KGE (—)
			Training phase			
WT-SVR	*coif*1	14.574	11.437	0.019	0.912	0.954
WT-ANN	*sym*2	10.625	7.935	0.013	0.953	0.962
WT-LightGBM	**coif1**	**1.113**	**0.843**	**0.001**	**0.999**	**0.996**
WT-LSTM	*haar*	7.727	5.777	0.010	0.974	0.900
WT-AM-LSTM	*haar*	2.839	2.427	0.004	0.985	0.948
			Test phase			
WT-SVR	*coif*1	19.798	15.275	0.027	0.662	0.876
WT-ANN	*sym*2	16.561	8.732	0.016	0.763	**0.882**
WT-LightGBM	*coif*1	25.063	19.774	0.036	0.458	0.772
WT-LSTM	*haar*	7.003	5.459	0.010	0.958	0.828
WT-AM-LSTM	**haar**	**3.036**	**2.290**	**0.004**	**0.992**	**0.947**

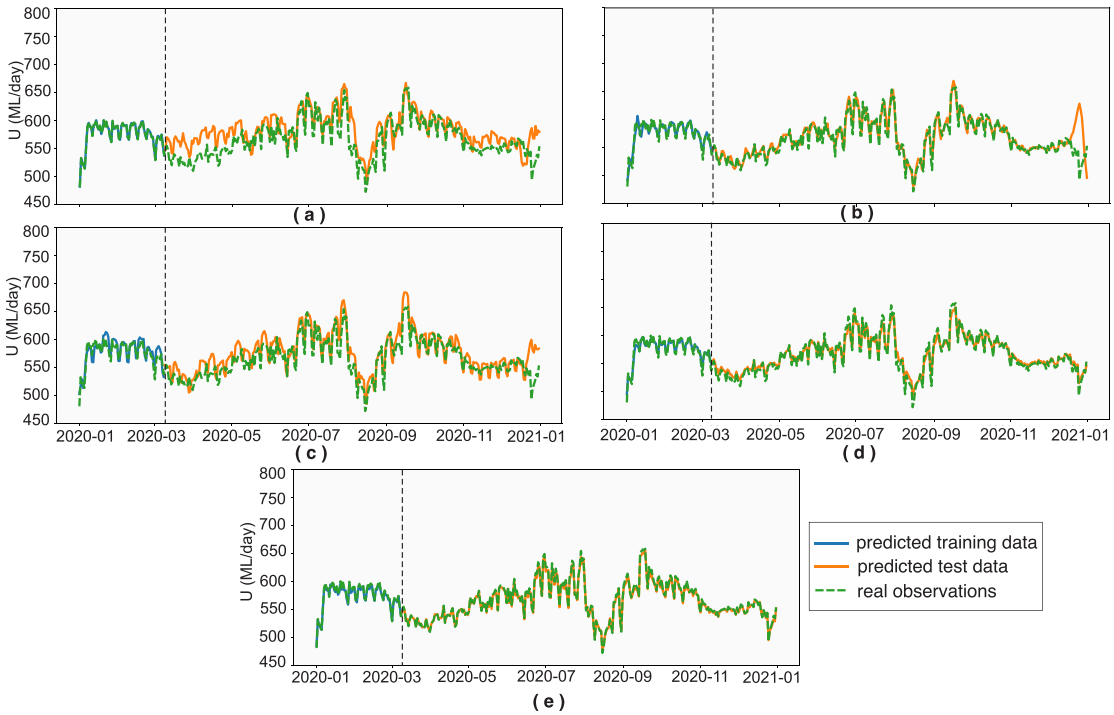

Figure 10. Predicted UWD vs. observed UWD before and after the COVID-19 lockdown (dashline: 9 March 2020). From (a) to (e): WT-LightGBM, WT-LightGBM-ANN, WT-LightGBM-SVR, WT-LightGBM-LSTM, and WT-LightGBM-AM-LSTM. In each line plot, the green dashed line represents observations, the blue line represents model predictions in training phase, and the orange line represents model predictions in test phase.

best models except AM-LSTM have performance degradation to varying degrees in predicting UWD during the pandemic period. One unanticipated finding is that LightGBM has an obvious drop in performance from the training to test phase: RMSE from 1.113 ML d^{-1} to 25.063 ML d^{-1}, MAE from 0.843 ML d^{-1} to 19.774 ML d^{-1}, MAPE from 0.001 to 0.036, NSE from 0.999 to 0.458 and KGE from 0.996 to 0.772. LightGBM largely overestimates the UWD during the lockdown period (March–May) (figure 10(a)). A similar trend is observed for WT-SVR which predicts higher values for March to May and peaks in September (see values in figure 10(c)). In contrast to LightGBM and SVR, ANN is more robust to predict reduced UWD and smoother summer peaks in 2020. Yet, it performs opposite at the end of December 2020 where a peak is predicted rather the real low values resulting in an overall performance deterioration. WT-LSTM shows only a marginal decline in forecasting performance, with overall predictions closely aligning with actual observations, as illustrated in figure 10(d). Remarkably, AM-LSTM maintains its high accuracy in predictions. In conclusion, neural-based ANN and LSTM model structures predicting time series data

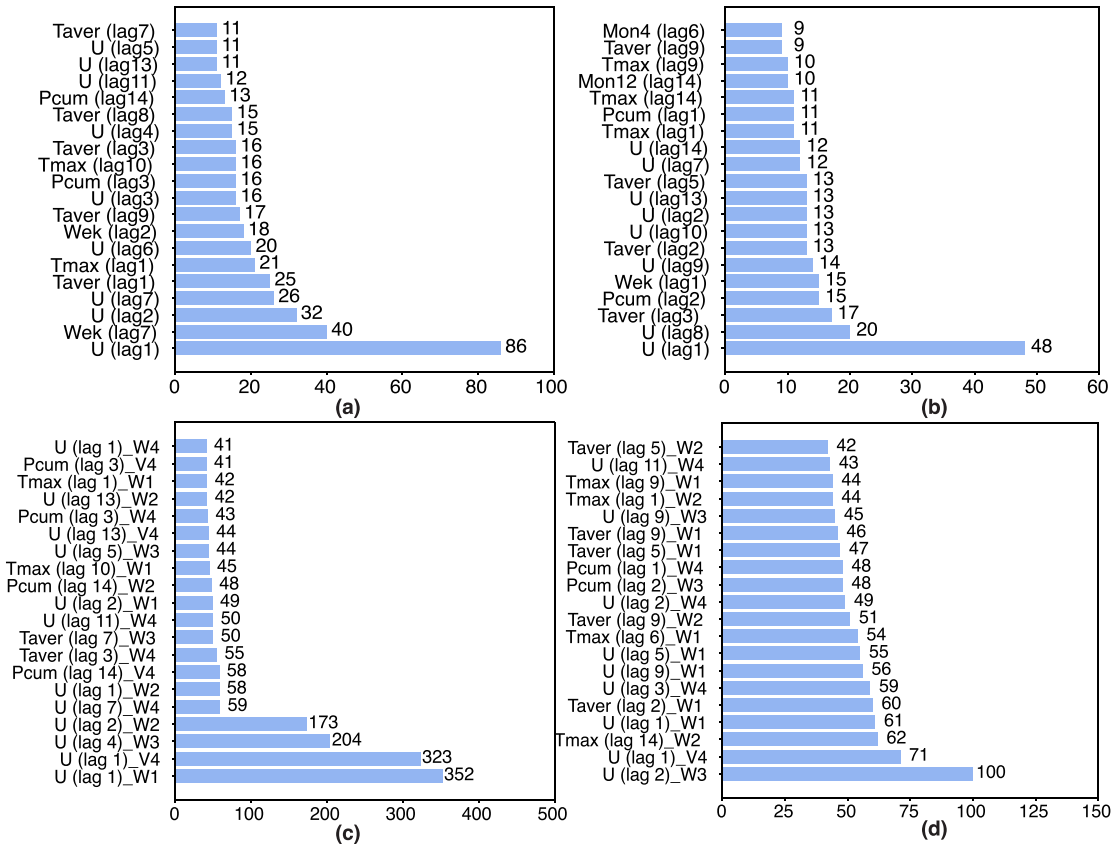

Figure 11. Selected features and corresponding feature importance values for (a): 1 d-ahead prediction in the non-WT setting; (b) 7 d-ahead lead time in the non-WT setting; (c) 1 d-ahead prediction in the WT setting; (d) 7 d-ahead lead time in the WT setting. The importance values are obtained by the default setting of split check in LightGBM, which represents how many times a feature is used for splitting a node across the trees. W and V represent wavelet coefficient and scaling coeffient achieved by WT decomposition at different levels.

outperform the other models when UWD has unexpected changes due to, e.g. socio-economic uncertainty or enforced water demand changes.

3.3. Feature selection results for 1 d-ahead and 7 d-ahead forecasting

Figure 11 shows the importance rank of input variables given by LightGBM as the model predicts the UWD for each lead time under both settings. Decomposition results are achieved by using filter *coif1*. Based on the previous model performance results, usually the top 20 variables are the most important ones. Accordingly, only 20 variables are presented in the figure. In the non-WT setting (figures 11(a) and (b)), historical U (especially 1 time-lagged U) is ranked as most important one and the weekday/weekend indicator of 7 d lead time ($Wek(lag7)$) is also identified as one of the top 5 important ones for both lead times. However, for 1 d-ahead prediction, several time-lagged U are ranked as more important variables, while for 7 d-ahead prediction, meteorological variables ($(T_{\text{aver}}, P_{\text{cum}})$ are emphasized. Differences are also observed that UWD of lead time less than 7 d are more critical for short-term prediction, while UWD of 8, 9, 10, 13, and 14 lead times become more significant in 7 d-ahead data selection.

In the wavelet setting (figures 11(c) and (d)), feature selection is based on all wavelet coefficients and scaling coefficients of each selected variable in the non-wavelet setting. After decomposition, some sub-components including wavelet and scaling coefficients of historical UWD become significantly important for 1 d-ahead prediction. The most critical ones are the wavelet coefficient at level 1 and scaling coefficient at level 4 of 1 time-lagged U, followed by wavelet coefficients of 2 and 3 time-lagged U ($U(lag4)\_W3, U(lag2)\_W2$). Compared to the non-WT results, precipitation and temperature periodicity becomes more important as well. However, for 7 d-ahead lead time prediction, the difference between importance values of variables is not very noticeable representing the almost equal importance. Yet, these sub-components are still emphasized with higher values compared with the non-WT setting result. This observation could be the explanation of why the model performance improvement of short-term prediction is higher than that of long-term prediction, as LightGBM has limited capacity to recognize more critical variables for 7 day-ahead prediction. This could be the disadvantage of ML models using tabular data: for

long-term prediction the dependence of time is more relevant, but tabular data can not fully represent the real temporal characteristics even after a feature engineering step. Similar research [35] also concludes that hybrid LSTM outperforms hybrid SVM. According to [73], decomposition levels 1–4 have corresponding frequency ranges from 2^{-1} to 2^{-5} indicating data variations in intervals of 1 day to 1 week. Therefore, the selection results show that variations from daily to weekly of UWD and weekly changes in meteorological variables are highly relevant to forecast UWD at lead times of 1 day and 7 d in this case study.

4. Conclusions

UWD forecasting is vital for short-term water distribution network operations and long-term planning and management of urban water infrastructure. However, developing accurate predictive models is often challenging due to the non-linearity in UWD patterns, the uncertain relationship between UWD and its drivers, and changing behavior across multiple temporal and spatial scales. Building on recent advances in ML and DL models, in this study we compared seven different time series and ML/DL classes of models, i.e. AR, SARIMAX, ANN, SVR, LightGBM, LSTM and AM-LSTM tasked to predict 1 d- and 7- day-ahead UWD on data collected in Milan (Italy) in the period 2017–2020. Some of these models are ML models that rely on tabular data (SVR, ANN, LightGBM), while others are DL models that rely on time series data (LSTM and AM-LSTM). AR and SARIMAX models are utilized as benchmarks for comparative analysis. We propose a wavelet-enhanced feature selection approach that combines a WT with LightGBM to select most useful predictors, adopting best practices from the recently introduced WDDFF. This study compares all hybrid models using this feature selection technique (WT-LightGBM-SVR, WT-LightGBM-ANN, WT-LightGBM, WT-LightGBM-LSTM, WT-LightGBM-AM-LSTM) with standalone models (SVR, ANN, LightGBM, LSTM, AM-LSTM), using widely accepted quantitative model performance evaluation metrics, namely RMSE, MAE, MAPE, NSE, and KGE. After model training and testing on data from 2017 to 2019, we subsequently re-evaluate the best models using UWD data from 2020. This is needed to assess their performance under previously unseen demand patterns, in this case determined by the COVID-19 pandemic.

Our numerical results indicate the following key insights. First, for 1 d-ahead UWD prediction, ANNs are a robust and reliable choice, providing accurate forecasts in both standalone and wavelet-enhanced setting. Additionally, the baseline models AR and SARIMAX show potentials in short-term, 1-step-ahead forecasting with comparable performance, making them viable alternatives for rapid response, short-term UWD forecasting scenarios. Second, for 7 d-ahead predictions, WT-LightGBM-LSTM and WT-LightGBM-AM-LSTM significantly outperform all other models, achieving NSE and KGE values higher than 0.9. Furthermore, for 7 d-ahead forecasting, the detailed analysis of temporal performance variations highlights the robustness of the hybrid neural-based models (WT-LightGBM-ANN, WT-LightGBM-LSTM, and WT-LightGBM-AM-LSTM), particularly during the demand drop observed in the August summer holiday. Historical UWD is confirmed as the main predictor of future water demands for both 1 d-ahead and 7 d-ahead predictions, but the weekly patterns of meteorological variables also contribute in a non-negligible way to accurate predictions. Finally, when tasked to forecasting UWD under unseen demand patterns in 2020, potentially influenced by COVID-19 lockdowns, our study finds LightGBM and SVR have limited capacity to forecast the decreasing UWD trend during the lockdown period, while ANN, LSTM and AM-LSTM can better predict the trend and smoothed summer peaks after the lockdown. Especially the DL models are robust to achieve satisfying performance under this situation. The hybrid AM-LSTM models demonstrate superior performance in multi-step predictions, particularly during significant demand changes caused by the summer/Christmas holidays in Milan and unforeseen patterns due to the pandemic. This superiority is attributed to their deep neural network structure, which effectively learns long-term dependencies in time series data, further enhanced by wavelet transformation and AMs, outperforming traditional LSTM and ANN models with shallower structures.

In summary, our findings indicate that integrating wavelet decomposition within the WDDFF, combined with LightGBM for feature selection, generally enhances the performance of ML/DL models for UWD forecasting, particularly for multi-step long-term forecasting and predicting complex demand dynamics due to external stressors. There is not a universal filter that is beneficial for all models, but results show that wider filters are more suitable for tabular data prediction and time series data prediction is not limited by the filter length. Furthermore, LightGBM is suitable for conducting feature selection integrated into the WDDFF to improve the model performance and also has a relatively high predictive capacity following ANN and LSTM. The broader impact of this research lies in its potential to significantly enhance both management and planning of urban water infrastructure, supply, and demands. Accurate demand forecasts are a key input to water supply systems. By relying on more accurate and reliable short- to medium-term forecasts, water infrastructure operators and suppliers can improve the efficiency of their supply operations, thereby potentially achieving substantial energy saving and reducing supply-related greenhouse gas emissions, while

guaranteeing a reliable and safe service. Likewise, urban planners and policymakers can make better-informed decisions in pursuit of water security and mitigate the effects of hydroclimatic extremes such as water scarcity. These improvements together enhance the resilience of urban water systems and contribute to the overall sustainability and efficiency of urban infrastructure.

Finally, several limitations and opportunities for future research should be acknowledged. First, although the hybrid framework applied to ML/DL models demonstrates its robustness for multistep forecasting and modelling demand changes due to external stressors, in this study we test all models on data from a single city, which may limit the generalizability of our results. Nevertheless, our modelling framework can be easily transferred and applied to other case studies from different contexts. Further research can thus consolidate our results by testing across diverse case studies. Second, the impact of short- and medium-term weather variations has not been deeply explored in our study. Our study assumes perfect predictions of future short- and medium-term weather conditions, yet weather forecasts inaccuracies could have an impact on our modeling results. Future research should address this limitation by incorporating the uncertainty of weather forecasts into the framework and evaluating their potential influence on demand forecasting. Such investigations can further assess the reliability of forecasting applications in real-world scenarios. Moreover, this research only investigates deterministic forecasting. Future research could explore uncertainty estimation and probabilistic forecasting using a bootstrap-based approach. The bootstrap method, which involves resampling the data to create multiple synthetic datasets, can be used to estimate the variability and confidence intervals of predictions. By generating probabilistic forecasts, this approach can further confirm the robustness of the proposed framework and provide a comprehensive understanding of the potential range of future water demand scenarios. This investigation will enhance the reliability of predictions, enabling urban planners and policymakers to better account for uncertainty and make more robust decisions.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This project acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Innovative Training Network NEWAVE — Grant Agreement No. 861509. Thanks to Mr. Fabio Marelli from Metropolitana Milanese (MM) S.p.A. for providing us pumped water volume data for different pumping stations in Milan.

Appendix A. Model formulations

A.1. ARIMA model

ARIMA was first developed by [74] and it assumes a linear relationship in the data sequence considering random errors. There are three components in an ARIMA model: (i) AR terms, (ii) moving average (MA) terms, and (iii) integrated (I) terms. The AR terms show that a specific number of previous, lagged data are used to predict the present value in a linear function. The MA terms account for past errors in predictions compared to actual observations. When the time series data is non-stationary, the integrated differences (I) should be added to remove trends. The order of an ARIMA model is expressed via the three parameters (p, d, q), where p, d and q are the numbers of AR, I, and MA terms.

In this study, we develop a simplified version of the ARIMA model, incorporating only AR terms as an AR model, and a seasonal ARIMA variant, known as seasonal ARIMA with eXogenous factors (SARIMAX) model, to assess the impact of external variables and seasonality on UWD forecasting. For SARIMAX model, the seasonal AR, MA and I terms are added to the model structure, with P, D and Q representing the dimensionality of these terms, respectively. With the structure of $(p, d, q) \times (P, D, Q)$, data and errors at time lags that are multiples of S (the time span of the seasonality) are also modelled. External meteorological variables are added to the input space to check for weather influences on UWD.

A.2. ANN model

ANN is one of the most popular and widely used ML models in many fields, including hydrology and water resources management [75–77]. The MLP, a layered FFNN, has been proven as a simple and efficient ANN architecture in UWD forecasts [78–80], thus we also consider it in our comparative analysis. A MLP consists

of three types of layers, sequentially interconnected: input layer (receives and processes input data), hidden layer (processes and transfers intermediate data) and output layer (predicts outputs). Each layer has a certain number of neurons which contain information transferred from the previous layer. Neurons can process information via an activation function (e.g. relu, tanh, sigmoid functions) and pass it to the next layer. The performance of ANN depends on two important hyperparameters: number of hidden layers and number of neurons in each layer. These parameters should be selected carefully to ensure the optimal model performance. In this study, the number of hidden layers is selected in a range from 1 to 3, while the number of neurons is selected between 5 and 50. Optimal values for these two parameters are found by hyperparameter tuning.

A.3. SVR model

SVM are well-known ML algorithms in classification problems. SVR is an extended version of SVMs for regression problems, first proposed by [81]. The key difference of SVR from other regression models is that SVR identifies a small number of data points so-called support vectors to represent the boundaries of all samples. These support vectors determine the margins of an appropriate line (or hyperplane in higher dimensions) to fit the data and the sum of errors within the margins is minimized [42]. SVR models map the original data into a higher dimensional space by a kernel function ϕ so that a linear regression problem can be solved in that space [82].

In order to train an accurate SVR model, the kernel function and several important hyperparameters need to be chosen and tuned appropriately. Polynomial, radial basis function and sigmoid kernels are commonly considered in SVR applications and, consistently, in this study. Hyperparameter tuning includes: kernel coefficient γ, which controls the shape of hyperplane separation (value in the range from 0.0001 to 0.9), and the regularization parameter C, which decides the sparsity of the model (value in the range from 0.1 to 1000).

A.4. LightGBM model

LightGBM was designed by Microsoft Research Asia [23] as a computationally efficient GBDT framework. Recently, LightGBM has been used as a predictive model and as an embedded feature selection method [83, 84]. Before LightGBM is introduced, eXtreme gradient boosting (XGBoost, [85]) was one of the most used GBDT algorithms on different regression and classification tasks. One key difference between LightGBM and XGBoost is in the construction of trees. Trees grow depth-wise in XGBoost, while in LightGBM, trees grow leaf-wise. In a simple decision tree model, nodes are split with the most informative feature, resulting in the largest information gain. For LightGBM, the gain of variance after splitting is measured, and the leaf with the largest value is selected to split each tree. The gradient-based one-side sampling (GOSS) algorithm is developed in LightGBM to efficiently reduce the number of data instances for training, thus speeding up the training process. In GOSS, all data are first sorted in a descending order based on their absolute value of gradients and the top $a \times 100\%$ data are selected to form a subset A. Out of the remaining data, $b \times 100\%$ are randomly sampled as subset B. In this way, the model focuses more on the data that cause more loss while not changing the data distribution much with the reduced data space $A \cup B$. The variance gain $V_j(d)$ of the splitting feature j at point d is defined as follow [23]:

$$V_j(d) = \frac{1}{k}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{k_l^j(d)} + \frac{\left(\sum_{x_i \in A_h} g_i + \frac{1-a}{b}\sum_{x_i \in B_h} g_i\right)^2}{k_h^j(d)}\right) \tag{A.1}$$

where k is the total number of samples in $A \cup B$, $k_h^j(d)$ and $k_l^j(d)$ represent the number of samples with value of feature j higher than or less than the threshold d. Subsets are defined as:
$A_l = \{x_i \in A : x_{ij} \leqslant d\}, B_l = \{x_i \in B : x_{ij} \leqslant d\}, A_h = \{x_i \in A : x_{ij} > d\}, B_h = \{x_i \in B : x_{ij} > d\}$, and g_i is the negative gradient of instance x_i. The optimal threshold $d^\star$ is decided by optimizing $d_j^* = \text{argmax}_\text{d}\, V_j(d)$.

Hyperparameters tuning in LightGBM is similar to the one for RF. Some critical parameters and corresponding ranges are: number of estimators (2–400), maximum depth of tree (2–8), number of leaves (10–160), and minimum child samples in leaf (1–20).

Appendix B. WT: MODWT

MODWT decomposes the original time series $X(t = 0, 1, \ldots, N-1)$ into wavelet coefficients ($\tilde{W}_{j,t}$) and scaling coefficients ($\tilde{V}_{j,t}$) by applying the wavelet filter ($\tilde{h}_{j,l}$) and scaling filter ($\tilde{g}_{j,l}$) respectively. These filters

are modified by re-normalization of the DWT filters ($\tilde{h}_{j,l} = \frac{h_{j,l}}{2^{\frac{j}{2}}}, \tilde{g}_{j,l} = \frac{g_{j,l}}{2^{\frac{j}{2}}}$). The decomposition equations are given by [62]:

$$\tilde{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N} \tag{B.1}$$

$$\tilde{V}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{g}_{j,l} X_{t-l \bmod N} \tag{B.2}$$

where L is the length of filter, $j = 1, 2, \ldots, J$ represents the decomposition level, *mod* refers to the modulo operator.

Appendix C. Detailed experimental settings

C.1. Feature engineering and selection

At the city scale, future UWD is usually determined by historical UWD, meteorological, and temporal variables. The UWD forecasting models presented in the previous section can be divided into two types: those using time series data versus those using tabular data. Temporal features are treated as independent input variables in tabular data, while time information is embedded with sequential indexes of time series data. Among the seven models used in this study, AR, SARIMAX, LSTM and AM-LSTM use time-series data, while LightGBM, ANN and SVR models process tabular data.

Firstly, we conducted an exploratory analysis of the correlations between UWD and meteorological variables by calculating Spearman's rank correlation coefficients (table 2). T_{aver} and $T_{\max}$ have relatively higher coefficients with U and they are selected as initial input variables. Even if the effect of precipitation on UWD is not significant in some research [86, 87] and has lower coefficients here, we still include P_{cum} to explore its potential influence in our case.

Fourteen consecutive data of U, T_{aver}, $T_{\max}$ and P_{cum} form the input matrix for LSTM and AM-LSTM considering the weekly patterns in UWD, while the amount of previous observations is selected automatically for SARIMAX. For models using tabular data, additional variables are needed to encode time information. Specifically, one-hot encoded variables *Holy*, indicating holiday occurrence ($Holy = 1$ indicates holidays; $Holy = 0$ represents non-holiday days), *Wek*, as a type of day indicator ($Wek = 1$ represents weekends; $Wek = 0$ represents weekdays), and $Mon_1, Mon_2, \ldots, Mon_{12}$ as month indicator ($Mon_{1,\ldots,12} = 1$ represents the day is in that month; $Mon_{1,\ldots,12} = 0$ represents the day is not in that month) are constructed for each day of the original time series. Each predictor (U, T_{aver}, $T_{\max}$, P_{cum}, *Wek*, *Holy*, $Mon_1, Mon_2, \ldots, Mon_{12}$) is time-lagged by up to 14 days ($t, t-1, \ldots, t-13$). The resulting input space after data pre-processing consists of 252 variables. The target variables are the UWD at time $t+1$ and $t+7$ for 1 d- and 7 d-ahead prediction, respectively, for all models.

Once the input space is defined, all seven predictive models are run with and without wavelet decomposition for 1 d- and 7 d-ahead predictions. For each lead time, training data (80% of total data), including all predictors and the target variable, are first used as inputs to the LightGBM model. A typical k-fold cross-validation step is applied to avoid over-fitting with k = 3. Then, the model is evaluated on test data (20% of total data) to assess the predictive capacity. Secondly, the feature importance values of all predictors are obtained automatically by LightGBM according to the default setting of the split check, which represents how many times a feature is used for splitting a node across the trees. Therefore, the predictors can be sorted in descending order based on their importance scores (predictors with high importance values are considered as informative and important). Finally, various thresholds (top 20, 0.1, 0.25, 0.5, 0.75, 0.9 quantiles) are set to select different numbers of important features to find the best group of features for prediction and reduce the input variable space. Specifically, each quantile of the sorted importance scores is calculated at that threshold, and all variables with importance scores above the threshold are selected. Then, all selected features associated with each threshold are used as input variables for ANN and SVR models to develop the best model for each lead time. For the LSTM and AM-LSTM models using time series data, all 4 variables (U, T_{aver}, $T_{\max}$ and P_{cum}) are used in non-WT setting. Differently, in the WT setting, the relevant sub-components of these variables are chosen as input variables by manually checking the top 20 features selected in the aforementioned step.

C.2. Wavelet decomposition with WDDFF

The wavelet decomposition procedure follows the set of best practices in the WDDFF. Notably, the selection of filter and decomposition levels is not arbitrary, but rather, it considers the size of the decomposed data set

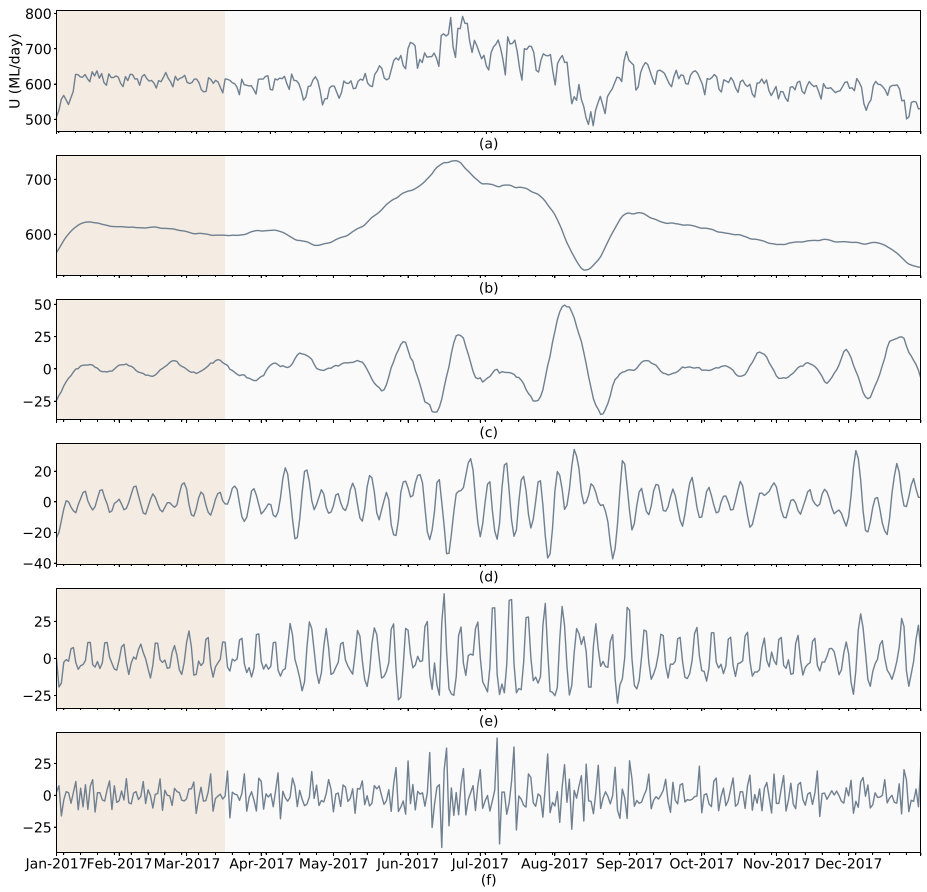

Figure C1. Decomposition results of data in 2017 by applying *coif*1 filter and decomposition level = 4. The shaded areas show the removed data ($n = 76$) due to boundary issues. From top to bottom: (a) original data, (b) scaling coefficients at level 4, (c)–(f) wavelet coefficients at level 4 to 1.

and temporal characteristics of the data. We conduct an exploratory experiment to investigate the combinations of different filters and decomposition levels. According to the results, the *haar*, *db2*, *coif1*, *sym2* filters are suitable for this case study, with the filter lengths from 2 to 6 that approximately cover one week length. The maximum decomposition level (J) is determined as 4 to ensure the number of calibration and validation data is enough for ML model development and also allow us to explore the periodicity up to weekly scale according to the corresponding frequency time-scale range for daily data in [73]. Additionally, to remove the negative effect of BCs, a number of influenced wavelet and scaling coefficients should be removed from the beginning, and the number (L_J) is calculated as the following equation [49]:

$$L_J = \left(2^J - 1\right)(L - 1) + 1 \tag{C.1}$$

where J and L are the decomposition level and filter length. Consequently, a total number of 76 coefficients are influenced and removed. It is also necessary to remove the same number of observations in the target variables and in the input-output space of the non-WT setting. Finally, the single method described in WDDFF is applied as a WT-based forecasting method, where the sub-components of wavelet decomposed variables are used as inputs and the original target variables as output for ML models. A decomposition example of UWD by *coif*1 filter at a decomposition level of 4 is shown in figure C1.

To evaluate the influence of the number of selected variables for tabular data prediction, different selection thresholds are applied for WT-LightGBM, WT-LightGBM-ANN and WT-LightGBM-SVR in the 7 d-ahead lead time forecasting. The selection is performed in two steps. The first step is to select base variables for wavelet decomposition. These base variables are decomposed by filters, resulting in five times more sub-component variables. Due to the constraint of data length, only small thresholds (top 20, 10%, 25% quantiles) are chosen to determine the number of base variables by LightGBM. After the wavelet decomposition, various thresholds (top 20, 10%, 25%, 50%, 75%, 90%, 100% quantiles, i.e. until all sub-component variables are used) are tested to select the final features for WT-LightGBM-ANN and WT-LightGBM-SVR models. The input variables of WT-LightGBM-LSTM and WT-LightGBM-AM-LSTM are selected using the same procedure for 1 d-ahead prediction using only the top 20 as the threshold,

Table C1. Summary of applied filters, cut-off importance thresholds and number of selected features for the best WT-LightGBM-ML models.

Models	Filter	1st threshold	2nd threshold	number of selected features
WT-LightGBM-SVR	coif1	top 20	top 20	20
WT-LightGBM-NN	sym2	top 20	top 20	20
WT-LightGBM	coif1	10%	100%	116
WT-LightGBM-LSTM	haar	top 20	top 20	7
WT-LightGBM-AM-LSTM	haar	top 20	top 20	7

regarding the limited numbers of total variables for time series data prediction. The summary results of the best models associated with thresholds are illustrated in table 1.

C.3. Model performance assessment

Five model performance metrics widely adopted in numerical forecasting are used in this study: RMSE, MAE, MAPE, NSE and KGE. They are formulated as follows:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{C.2}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{C.3}$$

$$\mathrm{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{C.4}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{C.5}$$

where $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ are predicted data, $(y_1, y_2, \ldots, y_n)$ are observed data, $\bar{y}$ is the mean of all observed data. Low values of RMSE, MAE, MAPE and NSE values approaching 1 indicate good model performance.

In addition to the above, the KGE proposed by [88], is being increasingly used in hydrological model performance evaluation. As an improvement of NSE, KGE is a decomposition result of NSE, where correlation, mean bias and relative variability between observations and model forecasts are calculated. The formulation of KGE is shown as the following:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{C.6}$$

where r is the linear correlation coefficient between observations and forecasts, $\alpha = \sigma_{\mathrm{forcast}}/\sigma_{\mathrm{obs}}$ reflects the variability error since it represents the ratio between the standard deviation in forecasts $\sigma_{\mathrm{forcast}}$ and the standard deviation in observations σ_{obs}, and $\beta = \mu_{\mathrm{forecasts}}/\mu_{\mathrm{obs}}$ is a bias term computed as a ratio between the average of forecasts $\sigma_{\mathrm{forecasts}}$ and average of observations σ_{obs}. The ideal value of KGE equal to 1 represents perfect agreement between forecasts and observations. However, the threshold between satisfactory and unsatisfactory performance is not explicitly stated as 0 like NSE, and various researchers treat positive KGE values as indicative of 'good' model accuracy, whereas negative KGE values are considered 'bad' [89, 90].

Appendix D. Comparison of model performances for 1 d-ahead prediction

Table D1. 1 day lead time forecasting results. The best performance of all models across performance metrics and filters during the training and test phases is indicated in bold.

Models	Filter	RMSE (ML d^{-1})	MAE (ML d^{-1})	MAPE (—)	NSE (—)	KGE (—)
			Training Phase			
AR		18.845	13.854	0.023	0.835	0.872
SARIMAX		19.285	13.716	0.023	0.828	0.912
SVR		18.101	14.025	0.023	0.837	0.900
ANN		15.126	10.990	0.018	0.886	0.909
LightGBM		8.629	6.516	0.011	0.963	0.946
LSTM		10.373	6.274	0.011	0.947	0.963
AM-LSTM		8.950	4.973	0.009	0.960	0.951
WT-SVR	*haar*	10.092	7.771	0.013	0.950	0.987
WT-ANN	*haar*	1.515	1.120	0.002	0.999	0.976
WT-LightGBM	*haar*	2.164	1.729	0.003	0.998	0.965
WT-LSTM	*haar*	1.860	1.411	0.002	0.998	0.967
WT-AM-LSTM	*haar*	6.772	6.283	0.102	0.978	0.950
WT-SVR	*db*2	16.435	13.895	0.023	0.873	0.901
WT-ANN	*db*2	2.556	1.927	0.003	0.997	0.984
WT-LightGBM	*db*2	2.627	2.032	0.003	0.997	0.986
WT-LSTM	*db*2	4.666	3.158	0.005	0.990	0.879
WT-AM-LSTM	*db*2	4.793	3.771	0.006	0.989	0.935
WT-SVR	*coif*1	11.996	9.183	0.015	0.935	0.967
WT-ANN	*coif*1	2.198	1.664	0.003	0.998	0.973
WT-LightGBM	*coif*1	0.731	0.585	0.001	0.999	0.971
WT-LSTM	*coif*1	2.802	2.219	0.003	0.996	0.984
WT-AM-LSTM	*coif*1	4.051	2.867	0.005	0.993	0.943
WT-SVR	*sym*2	14.371	11.820	0.020	0.903	0.943
WT-ANN	*sym*2	3.920	3.043	0.005	0.993	0.986
WT-LightGBM	*sym*2	2.645	2.037	0.003	0.997	0.964
WT-LSTM	*sym*2	3.544	2.728	0.005	0.994	0.973
WT-AM-LSTM	*sym*2	4.066	3.557	0.006	0.992	0.984
			Test Phase			
AR		17.546	12.366	0.021	0.879	0.894
SARIMAX		15.460	11.284	0.019	0.906	0.950
SVR		19.546	14.747	0.025	0.896	0.927
ANN		18.058	13.273	0.023	0.911	0.857
LightGBM		20.524	14.510	0.025	0.885	0.828
LSTM		24.942	17.630	0.030	0.810	0.824
AM-LSTM		24.641	18.079	0.030	0.839	0.781
WT-SVR	*haar*	23.469	12.524	0.022	0.854	0.777
WT-ANN	*haar*	28.890	7.466	0.015	0.777	0.864
WT-LightGBM	*haar*	22.683	11.143	0.020	0.862	0.819
WT-LSTM	**haar**	**2.603**	**2.053**	**0.003**	**0.998**	**0.975**
WT-AM-LSTM	*haar*	6.717	5.755	0.009	0.988	0.941
WT-SVR	**db2**	20.721	14.193	0.024	0.886	**0.926**
WT-ANN	*db*2	10.913	4.874	0.009	0.968	0.981
WT-LightGBM	*db*2	16.637	12.038	0.021	0.926	0.840
WT-LSTM	*db*2	9.596	6.300	0.010	0.974	0.944
WT-AM-LSTM	*db*2	6.950	5.008	0.008	0.987	0.928
WT-SVR	*coif*1	23.697	12.940	0.023	0.850	0.820
WT-ANN	*coif*1	11.226	3.436	0.006	0.966	0.976
WT-LightGBM	**coif1**	**14.213**	**8.461**	**0.015**	**0.946**	**0.903**
WT-LSTM	*coif*1	3.919	3.068	0.005	0.996	0.965
WT-AM-LSTM	*coif*1	5.862	4.098	0.007	0.990	0.948
WT-SVR	**sym2**	**20.046**	**15.494**	**0.027**	**0.893**	0.767
WT-ANN	**sym2**	**8.691**	**5.218**	**0.009**	**0.980**	**0.988**
WT-LightGBM	*sym*2	16.632	12.035	0.021	0.927	0.839
WT-LSTM	*sym*2	5.822	4.310	0.007	0.991	0.970
WT-AM-LSTM	**sym2**	**5.426**	**3.982**	**0.006**	**0.992**	**0.982**

Appendix E. Comparison of model performances for 7 d-ahead prediction

Table E1. 7 day lead time forecasting results. The best performance of all models across performance metrics and filters during the training and test phases is indicated in bold.

Models	Filter	1st threshold	2nd threshold	RMSE (ML d^{-1})	MAE (ML d^{-1})	MAPE (—)	NSE (—)	KGE (—)	variable number
				Training Phase					
AR				20.661	14.473	0.024	0.800	0.870	4
SARIMAX				18.036	13.215	0.022	0.846	0.850	4
SVR		top 20		21.228	16.792	0.028	0.776	0.863	20
ANN		top 20		21.841	16.060	0.027	0.763	0.701	20
LightGBM		100%		14.042	10.586	0.018	0.902	0.821	252
LSTM				12.841	7.058	0.012	0.919	0.926	4
AM-LSTM				16.365	11.099	0.019	0.874	0.884	4
WT-SVR	*haar*	10%	10%	10.627	8.481	0.014	0.945	0.973	12
WT-ANN	*haar*	10%	10%	6.665	5.043	0.008	0.978	0.982	12
WT-LSTM	*haar*	top 20	top 20	2.299	1.651	0.003	0.998	0.964	7
WT-AM-LSTM	*haar*	top 20	top 20	2.506	1.839	0.003	0.997	0.992	7
WT-LightGBM	*haar*	25%	100%	5.461	4.171	0.007	0.985	0.987	264
WT-SVR	*db2*	25%	75%	16.826	13.786	0.023	0.867	0.892	215
WT-ANN	*db2*	25%	25%	6.623	4.939	0.008	0.979	0.996	67
WT-LSTM	*db2*	top 20	top 20	7.701	5.396	0.009	0.972	0.908	10
WT-AM-LSTM	*db2*	top 20	top 20	5.189	3.659	0.006	0.987	0.983	10
WT-LightGBM	*db2*	25%	100%	4.996	3.898	0.007	0.988	0.877	264
WT-SVR	*coif1*	top 20	top 20	14.929	11.809	0.020	0.899	0.937	20
WT-ANN	*coif1*	top 20	top 20	12.474	9.585	0.016	0.930	0.927	20
WT-LSTM	*coif1*	top 20	top 20	5.518	4.306	0.007	0.986	0.988	10
WT-AM-LSTM	*coif1*	top 20	top 20	5.592	4.287	0.007	0.986	0.960	10
WT-LightGBM	*coif1*	10%	100%	0.646	0.486	0.001	0.999	0.975	116
WT-SVR	*sym2*	25%	75%	16.789	13.708	0.023	0.867	0.847	217
WT-ANN	*sym2*	top 20	top 20	10.616	7.874	0.013	0.947	0.968	20
WT-LSTM	*sym2*	top 20	top 20	6.621	5.115	0.008	0.979	0.985	11
WT-AM-LSTM	*sym2*	top 20	top 20	6.067	4.407	0.007	0.983	0.939	11
WT-LightGBM	*sym2*	25%	100%	4.964	3.869	0.006	0.988	0.897	264

(Continued.)

Table E1. (Continued.)

Models	Filter	1st threshold	2nd threshold	RMSE (ML d^{-1})	MAE (ML d^{-1})	MAPE (—)	NSE (—)	KGE (—)	variable number
				Test Phase					
AR				34.235	24.113	0.041	0.530	0.603	4
SARIMAX				34.391	24.969	0.043	0.526	0.767	4
SVR		top 20		29.561	21.784	0.037	0.763	0.796	20
ANN		top 20		33.590	24.303	0.041	0.694	0.436	20
LightGBM		100%		31.326	22.699	0.039	0.734	0.600	252
LSTM				41.577	30.548	0.053	0.487	0.524	4
AM-LSTM				40.437	30.186	0.053	0.528	0.676	4
WT-SVR	*haar*	10%	10%	27.277	14.502	0.026	0.801	0.893	12
WT-ANN	*haar*	10%	10%	24.444	10.627	0.019	0.840	0.859	12
WT-LSTM	**haar**	**top 20**	**top 20**	**10.728**	**7.404**	**0.013**	**0.960**	0.921	**7**
WT-AM-LSTM	**haar**	**top 20**	**top 20**	**3.894**	**2.543**	**0.004**	**0.996**	**0.985**	**7**
WT-LightGBM	**haar**	25%	100%	27.031	14.940	0.027	0.805	**0.831**	264
WT-SVR	*db*2	25%	75%	24.827	17.422	0.030	0.836	0.823	215
WT-ANN	*db*2	25%	25%	25.304	11.063	0.020	0.830	0.898	67
WT-LSTM	*db*2	top 20	top 20	18.374	13.777	0.024	0.900	0.834	10
WT-AM-LSTM	*db*2	top 20	top 20	15.450	12.098	0.021	0.931	0.949	10
WT-LightGBM	*db*2	25%	100%	27.697	19.305	0.033	0.796	0.644	264
WT-SVR	**coif1**	**top 20**	**top 20**	**24.194**	**15.224**	**0.027**	**0.843**	**0.898**	**20**
WT-ANN	*coif*1	top 20	top 20	21.857	13.154	0.023	0.872	0.871	20
WT-LSTM	**coif1**	top 20	top 20	11.134	7.963	0.014	0.957	**0.979**	10
WT-AM-LSTM	*coif*1	top 20	top 20	9.512	7.448	0.013	0.968	0.970	10
WT-LightGBM	**coif1**	**10%**	**100%**	**26.696**	**18.081**	**0.032**	**0.809**	0.796	**116**
WT-SVR	*sym*2	25%	75%	24.419	17.207	0.030	0.842	0.817	217
WT-ANN	**sym2**	**top 20**	**top 20**	**21.695**	**11.944**	**0.021**	**0.875**	**0.906**	**20**
WT-LSTM	*sym*2	top 20	top 20	15.669	11.713	0.020	0.929	0.936	11
WT-AM-LSTM	*sym*2	top 20	top 20	15.988	12.440	0.021	0.926	0.928	11
WT-LightGBM	*sym*2	25%	100%	27.910	19.695	0.034	0.793	0.637	264

ORCID iDs

Wenjin Hao https://orcid.org/0000-0002-4190-7894
Andrea Cominola https://orcid.org/0000-0002-4031-4704

References

[1] Hoekstra A Y, Buurman J and van Ginkel K C H 2018 *Environ. Res. Lett.* **13** 053002
[2] Gross M-P, Ajami N K and Cominola A 2023 *Environ. Res. Lett.* **18** 094067
[3] Feizizadeh B, Omarzadeh D, Ronagh Z, Sharifi A, Blaschke T and Lakes T 2021 *Sci. Total Environ.* **790** 148272
[4] United Nations Environment Programme (UNEP) 2015 *Options for Decoupling Economic Growth From Water use and Water Pollution: A Report of the Water Working Group of the International Resource Panel* (United Nations)
[5] Cominola A, Giuliani M, Piga D, Castelletti A and Rizzoli A 2015 *Environ. Modelling Softw.* **72** 198–214
[6] Heydari Z, Cominola A and Stillwell A S 2022 *Environ. Res. Infrastruct. Sustain.* **2** 045004
[7] Pesantez J E, Berglund E Z and Kaza N 2020 *Environ. Modelling Softw.* **125** 104633
[8] Cominola A, Giuliani M, Castelletti A, Fraternali P, Gonzalez S L H, Herrero J C G, Novak J and Rizzoli A E 2021 *npj Clean Water* **4** 1–10
[9] Tiwari M K and Adamowski J 2013 *Water Resour. Res.* **49** 6486–507
[10] Guo G, Liu S, Wu Y, Li J, Zhou R and Zhu X 2018 *J. Water Resour. Plan. Manage.* **144** 04018076
[11] Rezaali M, Quilty J and Karimi A 2021 *J. Hydrol.* **600** 126358
[12] Liu J, Zhou X L, Zhang L Q and Xu Y P 2023 *Water Resour. Manage.* **37** 1–22
[13] Adamowski J, Fung Chan H, Prasher S O, Ozga-Zielinski B and Sliusarieva A 2012 *Water Resour. Res.* **48** W01528
[14] Bougadis J, Adamowski K and Diduch R 2005 *Hydrol. Process.* **19** 137–48
[15] Donkor E A, Mazzuchi T A, Soyer R and Alan Roberson J 2014 *J. Water Resour. Plan. Manage.* **140** 146–59
[16] Chen G, Long T, Xiong J and Bai Y 2017 *Water Resour. Manage.* **31** 4715–29
[17] Liu J-Q, Cheng W-P and Zhang T-Q 2013 *J. Water Resour. Plan. Manage.* **139** 23–33
[18] Braun M, Bernard T, Piller O and Sedehizade F 2014 *Proc. Eng.* **89** 926–33
[19] Zounemat-Kermani M, Batelaan O, Fadaee M and Hinkelmann R 2021 *J. Hydrol.* **598** 126266
[20] Piryonesi S M and El-Diraby T E 2020 *J. Infrastruct. Syst.* **26** 04019036
[21] Shuang Q and Zhao R T 2021 *Water* **13** 310
[22] Yang Y, Lv H and Chen N 2023 *Artif. Intell. Rev.* **56** 5545–89
[23] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T Y 2017 *Advances in Neural Information Processing Systems* vol 9
[24] Wang Z, Hong T, Li H and Ann Piette M 2021 *Adv. Appl. Energy* **2** 100025
[25] Deng T, Zhao Y, Wang S and Yu H 2021 Sales forecasting based on LightGBM *2021 IEEE Int. Conf. on Consumer Electronics and Computer Engineering (ICCECE)* pp 383–6
[26] Hua Y 2020 An efficient traffic classification scheme using embedded feature selection and LightGBM *2020 Information Communication Technologies Conf. (ICTC)* pp 125–30
[27] Fiorillo D, Kapelan Z, Xenochristou M, De Paola F and Giugni M 2021 *Water Resour. Manage.* **35** 1449–62
[28] Zubaidi S L, Kumar P, Al-Bugharbee H, Ahmed A N, Ridha H M, Mo K H and El-Shafie A 2023 *Appl. Water Sci.* **13** 184
[29] Mazzoni F, Marsili V, Alvisi S and Franchini M 2022 *Environ. Res. Infrastruct. Sustain.* **2** 025005
[30] Xenochristou M, Hutton C, Hofman J and Kapelan Z 2021 *J. Water Resour. Plan. Manage.* **147** 04021004
[31] Coelho B and Campos A G A 2019 *Int. J. Water* **13** 173
[32] Zounemat-Kermani M, Matta E, Cominola A, Xia X, Zhang Q, Liang Q and Hinkelmann R 2020 *J. Hydrol.* **588** 125085
[33] Adamowski J F 2008 *J. Water Resour. Plan. Manage.* **134** 119–28
[34] Pesantez J E, Li B, Lee C, Zhao Z, Butala M and Stillwell A S 2023 *Energy* **283** 129142
[35] Du B, Zhou Q, Guo J, Guo S and Wang L 2021 *Expert Syst. Appl.* **171** 114571
[36] Zanfei A, Brentan B M, Menapace A, Righetti M and Herrera M 2022 *Water Resour. Res.* **58** e2022WR032299
[37] Namdari H, Haghighi A and Ashrafi S M 2023 *Stoch. Environ. Res. Risk Assess.* 1–16
[38] Jia Z, Li H, Yan J, Sun J, Han C and Qu J 2023 *Appl. Sci.* **13** 10014
[39] ElSaid A, Jamiy F E, Higgins J, Wild B and Desell T 2018 *Appl. Soft Comput.* **73** 969–91
[40] Ghannam S and Hussain F 2023 *Knowl.-Based Syst.* **275** 110660
[41] Wang K, Ye Z, Wang Z, Liu B and Feng T 2023 *Sustainability* **15** 3628
[42] Mu L, Zheng F, Tao R, Zhang Q and Kapelan Z 2020 *J. Water Resour. Plann. Manage.* **146** 11
[43] Nasser A A, Rashad M Z and Hussein S E 2020 *IEEE Access* **8** 147647–61
[44] Hu P, Tong J, Wang J, Yang Y and de Oliveira Turci L 2019 A hybrid model based on CNN and Bi-LSTM for urban water demand prediction *2019 IEEE Congress on Evolutionary Computation (CEC)* (IEEE) pp 1088–94
[45] Sahoo B B, Panigrahi B, Nanda T, Tiwari M K and Sankalp S 2023 *SN Comput. Sci.* **4** 752
[46] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 *Advances in Neural Information Processing Systems* vol 30 (available at: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
[47] Zhou S, Guo S, Du B, Huang S and Guo J 2022 *Sustainability* **14** 11086
[48] Guo J, Sun H and Du B 2022 *Water Resour. Manage.* **36** 3385–400
[49] Quilty J and Adamowski J 2018 *J. Hydrol.* **563** 336–53
[50] Baldino N and David S I P 2018 *Investig. Geogr.* **1** 9–21
[51] Istat 2022 Data tables urban environment (available at: www.istat.it/it/archivio/264816) (Accessed 16 August 2022)
[52] Hochreiter S and Schmidhuber J 1997 *Neural Comput.* **9** 1735–80
[53] Gers F A, Schmidhuber J and Cummins F 2000 *Neural Comput.* **12** 2451–71
[54] Niu Z, Zhong G and Yu H 2021 *Neurocomputing* **452** 48–62
[55] Bahdanau D, Cho K and Bengio Y 2014 arXiv:1409.0473
[56] Luong M T, Pham H and Manning C D 2015 arXiv:1508.04025
[57] Daubechies I 1992 *Ten Lectures on Wavelets* (*CBMS-NSF Regional Conference Series in Applied Mathematics* vol 61) (Society for Industrial and Applied Mathematics)

[58] Graf R, Zhu S and Sivakumar B 2019 *J. Hydrol.* **578** 124115
[59] Zhou F, Liu B and Duan K 2020 *J. Hydrol.* **588** 125127
[60] Du K, Zhao Y and Lei J 2017 *J. Hydrol.* **552** 44–51
[61] Aussem A, Campbell J and Murtagh F 1998 *J. Comput. Intell. Financ.* **6** 5–12 cited By :148 (available at: www.scopus.com)
[62] Percival D B and Walden A T 2000 *Wavelet Methods for Time Series Analysis* vol 4 (Cambridge University Press)
[63] Akujuobi C M 2022 *Wavelets and Wavelet Transform Systems and Their Applications* (Springer)
[64] Smith T G *et al* 2017 pmdarima: ARIMA estimators for Python (available at: www.alkaline-ml.com/pmdarima)
[65] Bergstra J, Bardenet R, Bengio Y and Kégl B 2011 *Advances in Neural Information Processing Systems* vol 24 (available at: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf)
[66] Chollet F *et al* 2015 Keras (available at: https://keras.io)
[67] Srisa-An C 2021 Guideline of collinearity-avoidable regression models on time-series analysis *2021 2nd Int. Conf. on Big Data Analytics and Practices (IBDAP)* (IEEE) pp 28–32
[68] Knoben W J M, Freer J E and Woods R A 2019 *Hydrol. Earth Syst. Sci.* **23** 4323–31
[69] Rahman A T M S, Hosono T, Quilty J M, Das J and Basak A 2020 *Adv. Water Resour.* **141** 103595
[70] Taylor K E 2001 *J. Geophys. Res. Atmos.* **106** 7183–92
[71] Cano A, Arévalo P, Benavides D and Jurado F 2024 *Int. J. Electr. Power Energy Syst.* **155** 109616
[72] Li D, Engel R A, Ma X, Porse E, Kaplan J D, Margulis S A and Lettenmaier D P 2021 *Environ. Sci. Technol. Lett.* **8** 431–6
[73] Bašta M 2014 *Acta Oeconomica Pragensia* **22** 48–70
[74] Box G E and Jenkins G M 1976 *Time Series Analysis. Forecasting and Control* (*Holden-Day Series in Time Series Analysis*) (Holden-Day, 1970)
[75] Jain A, Varshney A K and Joshi U C 2001 *Water Resour. Manage.* **15** 23
[76] Chang L-C, Shen H-Y, Wang Y-F, Huang J-Y and Lin Y-T 2010 *J. Hydrol.* **385** 257–68
[77] Khan M Y A, Tian F, Hasan F and Chakrapani G J 2019 *Int. J. Sediment Res.* **34** 95–107
[78] Msiza I S, Nelwamondo F V and Marwala T 2007 Artificial neural networks and support vector machines for water demand time series forecasting *2007 IEEE Int. Conf. on Systems, Man and Cybernetics* (IEEE) pp 638–43
[79] Babel M S and Shinde V R 2011 *Water Resour. Manage.* **25** 1653–76
[80] Pacchin E, Gagliardi F, Alvisi S and Franchini M 2019 *Water Resour. Manage.* **33** 1481–97
[81] Vapnik V, Golowich S E and Smola A J 1996 *Advances in Neural Information Processing Systems* vol 7
[82] Bishop C M 2006 *Pattern Recognition and Machine Learning* (*Information Science and Statistics*) (Springer)
[83] Banga A, Ahuja R and Sharma S C 2021 *Int. J. Syst. Assur. Eng. Manage.* **14** 732–45
[84] Effrosynidis D and Arampatzis A 2021 *Ecol. Inform.* **61** 101224
[85] Chen T and Guestrin C 2016 *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 785–94
[86] Beal C D and Stewart R A 2014 *J. Water Resour. Plan. Manage.* **140** 04014008
[87] Xenochristou M, Kapelan Z and Hutton C 2020 *J. Water Resour. Plann. Manage.* **146** 12
[88] Gupta H V, Kling H, Yilmaz K K and Martinez G F 2009 *J. Hydrol.* **377** 80–91
[89] Sutanudjaja E H *et al* 2018 *Geosci. Model Dev.* **11** 2429–53
[90] Siqueira V A, Paiva R C D, Fleischmann A S, Fan F M, Ruhoff A L, Pontes P R M, Paris A, Calmant S and Collischonn W 2018 *Hydrol. Earth Syst. Sci.* **22** 4815–42