Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-33 (2022) 92–98

Comparing Predictive Machine Learning Models for Short- and Long-Term Urban Water Demand Forecasting in Milan, Italy

Wenjin Hao * **Andrea Cominola** **,*** **Andrea Castelletti** ****

* *Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza L. da Vinci, 32, 20133 Milan, Italy (e-mail: wenjin.hao@polimi.it)*

** *Chair of Smart Water Networks, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany (e-mail: andrea.cominola@tu-berlin.de)*

*** *Einstein Center Digital Future, Berlin, Germany*

**** *Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza L. da Vinci, 32, 20133 Milan, Italy (e-mail: andrea.castelletti@polimi.it)*

Abstract: Urban water demand forecasting is essential for water supply network optimization and management. In this case study, we comparatively investigate different state-of-the-art predictive models on short- (1 day-ahead) and long-term (7 day-ahead) urban water demand (UWD) forecasting for the city of Milan, Italy. The contribution of this paper is two-fold. First, we compare the forecasting performance of different time series and machine learning models on daily UWD. The tested models include Autoregressive Integrated Moving Average (ARIMA) models, Artificial Neural Networks (ANN), Support Vector Regression (SVR), Light Gradient Boosting Machine (LightGBM), and Long Short-Term Memory (LSTM) networks. Second, we investigate whether coupling a Wavelet Data-Driven Forecasting Framework (WDDFF) with these models further improves predictive capacity. Results show that, in general, WDDFF can improve model predictive performance. LSTM coupled wavelet decomposition technique can achieve high levels of accuracy with R^2 larger than 0.9 for both short- and long-term UWD forecasts. LightGBM can efficiently reduce the number of predictors and show the potential to forecast and select important features in the hydrology and water resources field.

Keywords: urban water demand forecasting, Long Short-Term Memory, Light Gradient Boosting Machine, Wavelet Data-Driven Forecasting Framework

1. INTRODUCTION

Urban water security is paramount for water utilities and water users alike. Yet, it is hampered by rapid urbanisation, population growth, and climate change (Hoekstra et al., 2018). As the global water supply and demand gap is estimated to reach 40% in 2030 (UNEP, 2015), water demand-side management strategies have emerged as important complementary measures to supply-side interventions in pursuit of water security (Cominola et al., 2015). Accurately predicting urban water demands (UWD) constitutes a critical input for planning and managing water supply systems and their operations and formulating demand-side management programs (Qi et al., 2018).

A variety of mathematical models have been developed in the literature on UWD forecasting and have been tested in different cities all over the world (e.g., Tiwari and Adamowski, 2013; Guo et al., 2018; Rezaali et al., 2021). Time series models like Autoregressive Integrated Moving Average (ARIMA) have been popular for UWD forecasting tasks (e.g., Adamowski et al., 2012; Bougadis et al., 2005). However, challenges emerge due to non-linearity in hydrological and UWD data (Chang and House-Peters, 2011). A recent review by Zounemat-Kermani et al. (2020) analysed the most widely adopted neural-based machine learning (ML) models in the hydrology and hydraulics fields, emphasising their suitability to deal with non-linear processes and observations. Within this class of models, those mainly used in the literature on UWD forecasting include Artificial Neural Networks (ANN), Feedforward Neural Networks (FFNN), and Multi-Layer Perceptrons (MLP). Only a few researchers also explore the potential of Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) Networks in UWD forecasting (Guo et al., 2018; Mu et al., 2020). Tree-based models (e.g., Random Forests (RF)) and kernel-based models (e.g., Support Vector Machines (SVM) and Regression (SVR)) have also been proven effective to accurately forecast UWD (e.g., Braun et al., 2014; Chen et al., 2017).

In last decades, the use of hybrid models has also signifi-

⋆ 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

10.1016/j.ifacol.2022.11.015

cantly increased. Coupled ML models with wavelet transform (WT) emerged as one of the promising hybrid models in hydrological process forecasting as well as UWD prediction (Nourani et al., 2014). Wavelet transform is regarded as a data pre-processing step to deal with non-stationarity in time series and further improve the predictive capacity of ML models. However, Quilty and Adamowski (2018) pointed out the misuse of wavelet decomposition in some research leading to over-optimal forecasting results in real-world applications. Consequently, a new Wavelet Data-Driven Foresting Framework (WDDFF) has been proposed (Quilty and Adamowski, 2018) to overcome the drawbacks of inappropriate applications (i.e., boundary condition-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.

Accounting for the variety of forecasting models existing in the literature, the main objective of this paper is to compare the forecasting performance of different ML models on UWD forecasting at different temporal scales. The contribution of this study is three-fold. First, we comparatively test different state-of-the-art predictive models, namely, ARIMA, ANN, SVR, LightGBM and LSTM, for predicting UWD in Milan (Italy). Second, we combine the WDDFF to investigate whether using hybrid models coupled with wavelet transform technique enables more accurate forecasting of short- (1 day-ahead) and long-term (7 day-ahead) urban water demands. Finally, we use LightGBM to identify and select relevant predictors of UWD for the same case study.

The rest of the sections are organised as follows: Section 2 describes the case study and introduces methods as well as the experimental settings; Section 3 includes the main results and discussions; finally, Section 4 summarises the main conclusions of this study.

2. MATERIAL AND METHODS

2.1 Data: the Milan Case Study (Italy)

Water supply in Milan is managed by Metropolitana Milanese S.p.A. (MM), the largest water utility in Milan. In total, there are 28 active pump stations that extract groundwater and distribute it to final users through a complex distribution network with a total length of approximately 2,228 km. The total water pumped to the system is measured with a daily frequency at the input node of the whole system. For this study, time series with daily UWD data (U) are available from January 1, 2017 to December 31, 2019 (1095 records). The observed mean UWD in these 3 years is 601.14 megalitres per day (ML/D). Here, we use these data to forecast water demands 1 day- and 7 day-ahead.

Future UWD may depend on historical UWD as well as meteorological variables. Therefore, we also included potentially relevant meteorological data, which are collected at the Lambrate meteorological station by ARPA Lombardia (the Lombardy Regional Environmental Protection Agency). These data include: daily average temperature (T_{aver}), daily maximum temperature (T_{max}), and daily cumulated precipitation (P_{cum}). Table 1 shows a summary of the statistics of each variable in our dataset.

Table 1. Summary statistics of urban water demand and meteorological data.

Statistics	$U(ML/day)$	$T_{aver}(°C)$	$T_{max}(°C)$	$P_{cum}(mm/day)$
Minimum	464.14	-2.80	0.50	0.00
Maximum	791.85	32.00	38.90	79.40
Mean	601.14	14.97	20.28	2.57
1st Quartile	575.25	7.80	12.30	0.00
Median	595.72	14.80	20.20	0.00
3rd Quartile	620.51	22.85	28.85	0.60
Standard Deviation	48.36	8.46	9.27	7.46

2.2 Predictive Models

ARIMA. ARIMA was first developed by Box and Jenkins (1976) and assumes a linear relationship in the data sequence considering random errors. There are three components in an ARIMA model: (i) autoregressive (AR) terms, (ii) moving average (MA) terms, and (iii) integrated (I) terms dealing with non-stationality in data. In this case study, a seasonal version of the ARIMA model was developed to model the influence of external variables and account for seasonality. The resulting model is named Seasonal Autoregressive Integrated Moving Average with eXogenous factors (SARIMAX). Consequently, the seasonal AR, MA, and I terms are added to the model structure, and meteorological variables are also added to the input space to check for weather influences on UWD.

ANN. The multilayer perceptron (MLP), a layered feed-forward neural network, has been proven as a simple and efficient ANN architecture in UWD forecasts (Pacchin et al., 2019). For this reason, we also use it in this case study. An MLP consists of three types of layers that are interconnected with each other in a sequence: input layer (which receives and processes the input data), hidden layer (which transfers and processes intermediate data), and output layer (which predicts the outputs). An important step in constructing accurate ANN models is to identify the optimal number of hidden layers and neurons in each layer via hyperparameter optimisation. In this study, the number of hidden layers was selected in a range of 1-3, and the number of neurons was bounded between 5 and 50. The optimal hyperparameter values were found via k-fold cross-validation (k=3) and grid search over the defined parameter space to minimise the Mean Squared Error (MSE) on the validation set.

SVR. SVM model is a well-known ML algorithm in classification, and the SVR is the extended regression version of SVMs first proposed by Vapnik et al. (1996). An important step is to map the original data to a higher dimensional space by a kernel function ϕ so that a linear regression problem is solved in that space (Bishop, 2006). In order to train a SVR model, the kernel function and several important hyperparameters need to be chosen and tuned appropriately. Polynomial, radial basis function, and sigmoid kernels were tested in this study. Kernel coefficient γ was tuned in a range from 0.0001 to 0.9 and the regularisation parameter C in a range from 0.1 to 1000 by means of the same hyperparameter tuning procedure used for the ANN model.

LSTM. To overcome the limitations of traditional RNN models, which are not able to deal with long-term dependencies when the prediction has the connection with

information far before, LSTM was proposed by Hochreiter and Schmidhuber (1997). LSTM has a sequence of memory modules, where three gates control the information passed to the cell states (C_t). C_t is a key component in LSTM representing the long-term memory of data information. The three gates are the forget gate (producing forget value f_t), input gate (producing input value i_t), and output gate (producing output value o_t) in a forward sequence, which is composed of a sigmoid neural network layer (σ) and a pointwise multiplication action ($\otimes$). The fundamental equations of the LSTM are given by Gers et al. (2000):

$$\begin{aligned} f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\ \tilde{C}_t &= tanh\left(W_C x_t + U_C h_{t-1} + b_C\right) \\ C_t &= f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \\ h_t &= o_t \otimes tanh\left(C_t\right) \end{aligned} \tag{1}$$

where h_t is the output vector, $\tilde{C}_t$ is the candidate vector added to update the new cell state, the weights (W_f, W_i, W_o, U_f, U_i, U_o) and biases (b_f, b_i, b_o) parameters are determined by the training process.

As ANN models, the number of hidden layers and neurons are two important hyperparameters. In this case, the number of hidden layers is chosen between 2 and 3 and the number of neurons is chosen between 1 to 100.

LightGBM. LightGBM was designed by Microsoft Research Asia (Ke et al., 2017) as a computational efficient Gradient Boosting Decision Tree (GBDT) framework. Recently, LightGBM has been used as a predictive model and as an embedded input variable selection (IVS) method (Banga et al., 2021; Effrosynidis and Arampatzis, 2021). In LightGBM, the gain of variance after splitting is measured, and the leaf with the largest value is found to do the split in each tree, resulting leaf-wise growing trees. Gradient-based One-Side Sampling (GOSS) was developed in LightGBM to efficiently reduce the number of data instances for training. In the GOSS technique, 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. As for the remaining data, $b \times 100\%$ number of data 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 follows(Ke et al., 2017):

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

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 a 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 the loss function with respect to the output x_i of the model. The optimal threshold $d^\star$ is decided by optimizing $d_j^* = \mathrm{argmax_d}\, V_j(d)$. Hyperparameter tuning in LightGBM is similar to that in more traditional RF. Some critical parameters are chosen by the grid search algorithm in the following ranges: number of estimators between 2 and 400, maximum depth of tree between 2 and 8, number of leaves between 10 and 160, minimum child samples in leaf between 1 and 20.

2.3 Wavelet Transform and WDDFF

WT has the advantage of dealing with highly non-stationary data by using amplitude decayed wavelets and improve the understanding of characteristics of time series data (Daubechies, 1992). 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. Discrete wavelet transform (DWT) has been widely used in hydrological and water resources problems to improve forecasting capacity with low computational efforts (Graf et al., 2019; Zhou et al., 2020). However, as pointed by Quilty and Adamowski (2018), wavelet decomposition has been misused in some research which leads to over-optimal forecasting results in real-world applications. These problem are mainly related to the boundary condition (BC) issue (Aussem et al., 1998). To address the disadvantages of DWT applications, Quilty and Adamowski (2018) proposed the WDDFF which incorporates a couple of best practices to develop an appropriate hybrid WT-ML model. The prerequisite of WDDFF is using the Maximal Overlap DWT (MOWDT) or à trous algorithm (AT) instead of the DWT, since DWT cannot be adjusted to eliminate the BC issues due to its decimation property (Du et al., 2017). In this case study, the MOWDT introduced by Percival and Walden (2000) is implemented as a decomposition algorithm.

MODWT decomposes the original time series $X(t = 0, 1, ..., N-1)$ 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. The decomposition equations are given by Percival and Walden (2000):

$$\tilde{W}_{j,t} = \sum_{l=0}^{L_j - 1} \tilde{h}_{j,l} X_{t-l \bmod N} \tag{3}$$

$$\tilde{V}_{j,t} = \sum_{l=0}^{L_j - 1} \tilde{g}_{j,l} X_{t-l \bmod N} \tag{4}$$

where L is the length of filter, $j = 1, 2, ..., J$ represents the decomposition level, *mod* refers to the modulo operator. For more detailed definition and formulas of MODWT we refer the reader to Percival and Walden (2000). According to the best practices in the WDDFF, two important steps should be considered:

(1) Choose wavelet filters and decomposition levels: *db2* filter was explored here regarding the length of time series and the decomposition level. The decomposition level of $J = 4$ was chosen to keep enough data for training various models and the frequency range.
(2) Remove 'boundary-influenced' records to avoid boundary condition-related issues: the number of removed records in the input-output space is determined by

$L_j = (2^j - 1)(L - 1) + 1$ where j represent the decomposition level, and L is the length of a wavelet filter (Quilty and Adamowski, 2018). Accordingly, a total of 46 (4.2% of total data) records were removed from the beginning of original data and wavelet and scaling coefficients.

After selecting the appropriate filter and decomposition level, a direct approach is implemented where only input variables are decomposed and used to predict the original target variable. Then an IVS algorithm is run to select favorable wavelet and scaling coefficients to the predictive models. LightGBM is used to determine informative features here for both non-wavelet and wavelet settings. As the feature selection is an embedded function of the learning algorithm, LightGBM can learn for prediction and obtain importance values for input variables during the training period. Finally, ANN, SVR and LSTM models are calibrated and evaluated based on features selected by LightGBM in both settings.

2.4 Model Performance Metrics

Four wide-adopted evaluation metrics were used in this study: Root Mean Square Error ($RMSE$), Mean Absolute Error (MAE), Mean Absolute Percentage Error ($MAPE$), and Coefficient of Determination (R^2). They are formulated as follows:

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \\ \mathrm{MAE} &= \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \\ \mathrm{MAPE} &= \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \\ \mathrm{R}^2 &= 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \end{aligned} \tag{5}$$

where $(\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ are predicted data, $(y_1, y_2, ..., y_n)$ are observed data, $\bar{y}$ is the mean of all observed data. Values of $RMSE$, MAE, and $MAPE$ close to zero, and R^2 values approaching 1 indicate good model performance.

2.5 Experiment Settings

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 five models used in this study, SARIMAX and LSTM use time-series data, while LightGBM, ANN, and SVR process tabular data. Available data were split to use 80% for training and 20% for test in model development.

Future UWD is usually determined by historical UWD, meteorological, and temporal variables. Therefore, the initial input space of non-wavelet decomposed tabular data includes numerical variables of U, T_{aver}, T_{max}, P_{cum}, and a binary variable $Holy$ indicating holiday occurrence ($Holy = 1$ indicates holidays; $Holy = 0$ represents non-holiday days), a weekend indicator Wek ($Wek = 1$ represents weekends; $Wek = 0$ represents weekdays) and binary month indicators $Mon_1, Mon_2, ..., Mon_{12}$ ($Mon_{1,...,12} = 1$ represents the day is in that month; $Mon_{1,...,12} = 0$ represents the day is not in that month). Each predictor (U, T_{aver}, T_{max}, P_{cum}, Wek, $Holy$, $Mon_1, Mon_2, ..., Mon_{12}$) was time-lagged up to 14 days (t, $t-1$,..., $t-13$) considering the close relationship between successive data and weekly patterns. The resulting dataset after data processing consisted of 252 variables. The inputs for time series data are only historical UWD (U) and meteorological variables (T_{aver}, T_{max}, and P_{cum}) with the assumption that all time information is embedded. The target variables are the UWD at time $t + 1$ and $t + 7$ for 1 day- and 7 day-ahead prediction respectively for all models.

Once the input space is defined, all five predictive models are run with and without wavelet analysis. In both settings, only the top 20 features with high importance values are selected by LightGBM and used to make predictions for models using tabular data.

3. RESULTS

3.1 Model Performance for 1 day-ahead Prediction

Table 2. Model performance metrics computed for 1 day-ahead forecasting.

Models	$RMSE(ML)$	$MAE(ML)$	$MAPE(\%)$	R^2
Training Phase				
SARIMAX	19.285	13.716	2.3	0.828
SVR	18.101	14.025	2.3	0.837
ANN	15.126	10.990	1.8	0.886
LightGBM	8.629	6.516	1.1	0.963
LSTM	10.373	6.274	1.1	0.947
WA-SVR	16.435	13.895	2.3	0.873
WA-ANN	**2.556**	**1.927**	**0.3**	**0.997**
WA-LightGBM	2.627	2.032	**0.3**	**0.997**
WA-LSTM	4.666	3.158	0.5	0.990
Test Phase				
SARIMAX	15.460	11.284	1.9	0.906
SVR	19.546	14.747	2.5	0.896
ANN	18.058	13.273	2.3	0.911
LightGBM	20.524	14.510	2.5	0.885
LSTM	24.942	17.630	3.0	0.810
WA-SVR	20.721	14.193	2.4	0.886
WA-ANN	10.913	**4.874**	**0.9**	0.968
WA-LightGBM	16.637	12.038	2.1	0.926
WA-LSTM	**9.596**	6.300	1.0	**0.974**

SARIMAX, SVR, ANN, and LightGBM models used original selected tabular data for calibration and validation, while LSTM used original time series data for calibration and validation; WA-SVR, WA-ANN, and WA-LightGBM used selected wavelet-decomposed tabular data for calibration and validation, while WA-LSTM used selected wavelet-decomposed time series data for calibration and validation. The bold values indicate performance achieved by the best model in training and test phase specifically.

The results in Table 2 show that all models can achieve satisfying results with R^2 larger than 0.8 and $RMSE$ less than 20 ML/day in the test phase. According to several researches about daily UWD forecasting (e.g., Tiwari and Adamowski, 2013; Quilty and Adamowski, 2018), $RMSE$ lower than 27 ML/day and R^2 larger than 0.9 are usually very satisfying performance for 1 day-ahead prediction. In the non-wavelet setting, ANN and SARIMAX outperform the other models resulting R^2 of 0.911 and 0.906. These two are typical machine learning and time series models, which are usually used as benchmark models in comparative research (e.g., Adamowski et al., 2012; Tiwari and Adamowski, 2013). LightGBM achieves the best

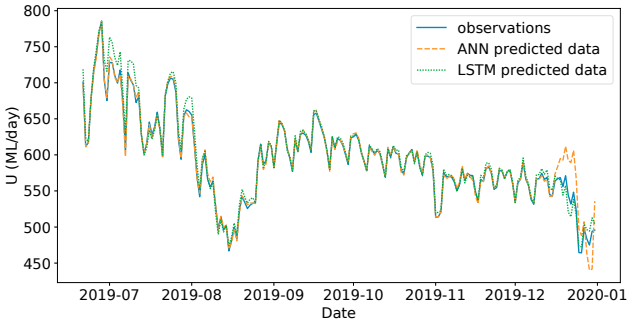

Fig. 1. 1 day-ahead prediction results achieved by the WA-ANN and WA-LSTM models in model testing.

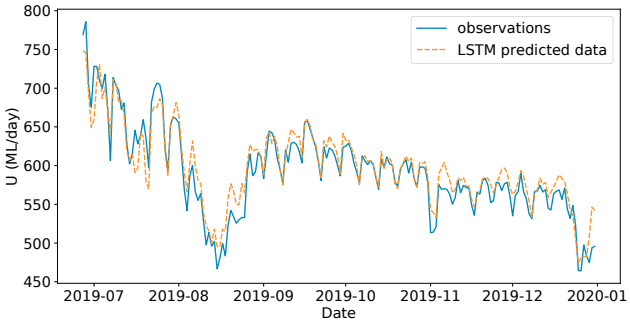

Fig. 2. 7 days-ahead prediction results achieved by the WA-LSTM models in model testing.

performance in the training phase, but its performance deteriorates in the test phase in both wavelet and non-wavelet settings. However, LightGBM is suitable for selecting appropriate inputs for machine learning models using tabular data, which improves model training efficiency and prediction interpretability. Wavelet decomposition, in general, improves the performance of all types of models for 1 day-ahead predictions, demonstrating the usefulness of extracting major information in data for forecasting tasks. In the wavelet setting, WA-ANN and WA-LSTM are the most suitable hybrid models (see prediction results in Fig. 1). The major difference between these two models is observed in December 2019. The WA-ANN model overestimates the UWD in early December and underestimates the UWD later, while WA-LSTM can capture the trend well.

Table 3. Model performance metrics computed for 7 days-ahead forecasting.

Models	$RMSE(ML)$	$MAE(ML)$	$MAPE(\%)$	R^2
Training Phase				
SARIMAX	32.947	23.663	4.0	0.492
SVR	21.228	16.792	2.8	0.776
ANN	21.841	16.060	2.7	0.763
LightGBM	14.042	10.586	1.8	0.902
LSTM	12.841	7.058	1.2	0.919
WA-SVR	13.110	10.195	1.7	0.919
WA-ANN	11.652	8.818	1.5	0.936
WA-LightGBM	**7.209**	5.487	**0.9**	**0.975**
WA-LSTM	7.701	**5.396**	**0.9**	0.972
Test Phase				
SARIMAX	34.391	24.969	4.3	0.526
SVR	29.561	21.784	3.7	0.763
ANN	30.672	22.292	3.8	0.745
LightGBM	31.326	22.699	3.9	0.734
LSTM	45.802	34.152	5.9	0.439
WA-SVR	31.854	16.080	2.9	0.730
WA-ANN	28.362	14.829	2.6	0.786
WA-LightGBM	29.538	20.737	3.6	0.773
WA-LSTM	**18.374**	**13.777**	**2.4**	**0.900**

SARIMAX, SVR, ANN, and LightGBM models used original selected tabular data for calibration and validation, while LSTM used original time series data for calibration and validation; WA-SVR, WA-ANN, and WA-LightGBM used selected wavelet-decomposed tabular data for calibration and validation, while WA-LSTM used selected wavelet-decomposed time series data for calibration and validation. The bold values indicate performance achieved by the best model in training and test phase specifically.

3.2 Model Performance for 7 days-ahead Prediction

The results of the 7 days-ahead forecasting are summarised in Table 3. In general, the predictive performance of all models deteriorates compared to 1-day forecasting results. Similar results are also illustrated in Quilty and Adamowski (2018), where from 1 day- to 7 day-ahead prediction, R^2 drops from 0.916 to 0.735, $RMSE$ increases from 27.358 ML/day to 46.840 ML/day. SARIMAX had the worst performance in both training and test phases. In many pieces of research, ARIMA models are usually developed for 1 lead time prediction and can achieve satisfying results, as in Section 3.1. A few researchers also explore the possibility of predicting successive data and obtained a similar conclusion that ARIMA models have limited capacity for multi-step prediction tasks (Tiwari and Adamowski, 2013; Mu et al., 2020).

In turn, ML models using tabular data (ANN, SVR, and LightGBM) achieve comparable and relatively satisfying results with R^2 around 0.75 in the test phase of the non-wavelet setting. Applying the wavelet technique only slightly improved the performance of these models in the test phase. The best performance across all four evaluation metrics is still achieved by the LSTM model coupled with the wavelet decomposition technique. The comparison of predicted values and observed values in the test phase is shown in Fig. 2. In contrast, the traditional LSTM model cannot provide accurate prediction on test data, with the lowest R^2 of 0.439 even though it performed well on training data. Specifically, this model cannot predict the low UWD during summer holidays in August and Christmas holidays in December. A possible reason is that the information on holiday is not well captured in the traditional LSTM model due to the short length of the time series data, influencing the 7 days-ahead forecasting. By applying wavelet decomposition, low-frequency information indicating the main patterns of time series data is better represented and relevant high-frequency information representing weekly/monthly patterns are preserved. Also, time information is still embedded in the decomposed data. As a result, WA-LSTM can model the transformed time series data better than other machine learning models relying on transformed tabular data.

3.3 IVS Results

The top 20 important features selected by LightGBM in the input spaces used for 1 day-ahead and 7 day-ahear forecasts are listed in Fig. 3 and 4. The importance value

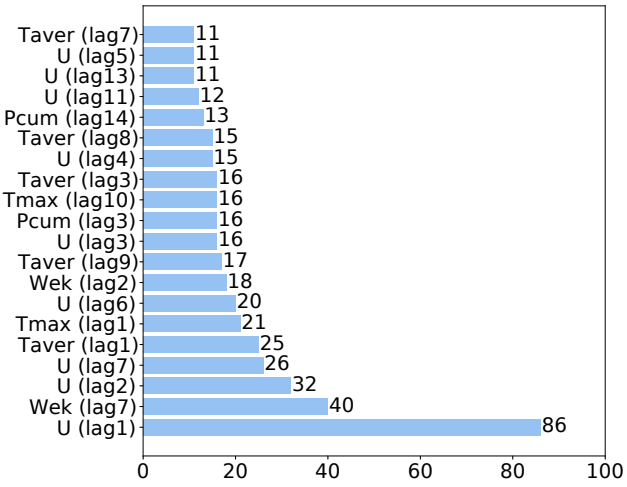

Fig. 3. Feature importance of the top 20 variables ranked for 1 day-ahead prediction.

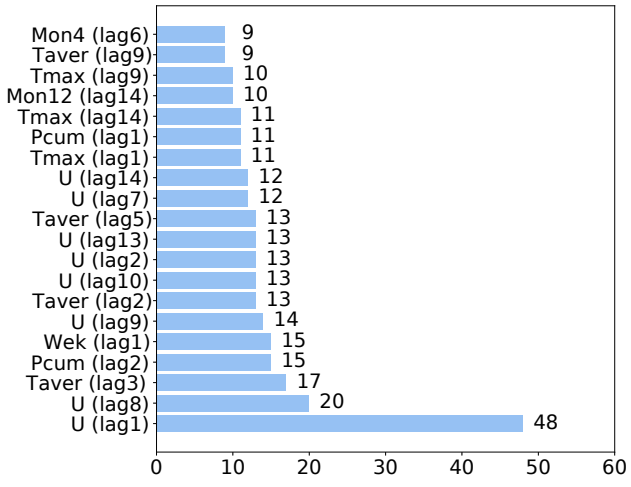

Fig. 4. Feature importance of the top 20 variables ranked for 7 day-ahead prediction.

represents the number of times that one feature is used in a model. In general, variables related to lagged historical UWD (U) and temperature (T_{aver}, T_{max}) are usually ranked as most important ones. For both 1 day and 7 day lead time input spaces, one time lag UWD ($U(lag1)$) is always the most influential feature, and the weekday/weekend indicator of 7 days lead time ($Wek(lag7)$) is also identified as one of the top 5 important ones. Differences are also observed between these two data sets. For 1 day lead time data selection, UWD of lead time less than 7 days are more critical, while UWD of 8, 9, 10, 13, and 14 lead times become more significant in 7 day-ahead data selection. Also, the cumulated precipitation (P_{cum}) emerges as relevant for long-term prediction.

In the wavelet setting, IVS selection is based on all wavelet coefficients and scaling coefficients of each selected variable in the non-wavelet setting. It is not surprising that all scaling coefficients of important variables (e.g., $U(lag1)$, $T_{aver}(lag1)$, $T_{max}(lag1)$) are selected, which are main patterns of original data representing low frequency information. Additionally, all level decomposed wavelet coefficients of UWD are kept for prediction, while only high-level wavelet coefficients (level 3 and level 4) are kept for meteorological variables. According to Bašta (2014), decomposition levels 1 to 4 have corresponding frequency ranges from 2^{-1} to 2^{-5} indicating changes in intervals of 1 day to 1 week. In this case study, the selection results show that changes 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 days.

4. CONCLUSIONS

Five different models, including ARIMA, ANN, SVR, LightGBM, and LSTM, are developed and compared to predict UWD 1 day- and 7 day-ahead using UWD data collected in Milan, Italy, in the period 2017-2019. Among them, ML models are also distinguished between those using tabular data (SVR, ANN, LightGBM) and those using time series data (LSTM). Moreover, the best practices of the wavelet decomposition technique from a recent proposed Wavelet Data-Driven Forecasting Framework (WDDFF; Quilty and Adamowski, 2018) are adopted to improve model performance, where the input variable selection is conducted by LightGBM. All wavelet-setting models (WA-SVR, WA-ANN, WA-LightGBM, WA-LSTM) are compared with non-wavelet-setting models (SVR, ANN, LightGBM, LSTM) based on wide-adopted evaluation metrics ($RMSE$, MAE, $MAPE$, and R^2). The numerical results indicate the following key insights:

(1) For 1 day-ahead UWD prediction, ANN model is robust and reliable in both non-wavelet and wavelet-settingz, while for 7 day-ahead prediction WA-LSTM had significant performance over all other models with an R^2 of 0.9.
(2) Historical UWD is the main predictor of future water demands over the time scale from daily to weekly, but weekly patterns of meteorological variables also contribute in a non-negligible way to accurate prediction.
(3) The wavelet decomposition technique can improve the overall machine learning model performance but it mostly improves the LSTM model using time series data.
(4) LightGBM is suitable for conducting IVS in the WDDFF to improve the model performance and has relatively high predictive capacity following ANN and LSTM.

Future studies may focus on testing the developed methodology on other UWD forecasting cases and evaluating the possibility of forecasting UWD under varying social and environmental uncertainties (e.g., COVID-19 Pandemic, climate change).

ACKNOWLEDGEMENTS

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.

REFERENCES

Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., and Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research*, 48(1).

Aussem, A., Campbell, J., and Murtagh, F. (1998). Wavelet-based feature extraction and decomposition strategies for financial forecasting. *Journal of Computational Intelligence in Finance*, 6(2), 5–12.

Banga, A., Ahuja, R., and Sharma, S.C. (2021). Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities. *International Journal of System Assurance Engineering and Management*.

Bašta, M. (2014). Additive Decomposition and Boundary Conditions in Wavelet-Based Forecasting Approaches. *Acta Oeconomica Pragensia*, 22(2), 48–70.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.

Bougadis, J., Adamowski, K., and Diduch, R. (2005). Short-term municipal water demand forecasting. *Hydrological Processes*, 19(1), 137–148.

Box, G.E. and Jenkins, G.M. (1976). Time series analysis. forecasting and control. *Holden-Day Series in Time Series Analysis*.

Braun, M., Bernard, T., Piller, O., and Sedehizade, F. (2014). 24-Hours Demand Forecasting Based on SARIMA and Support Vector Machines. *Procedia Engineering*, 89, 926–933.

Chang, H. and House-Peters, L. (2011). Urban Water Demand Modeling: Review of Concepts, Methods, and Organizing Principles. *Water Resources Research*, 16.

Chen, G., Long, T., Xiong, J., and Bai, Y. (2017). Multiple Random Forests Modelling for Urban Water Consumption Forecasting. *Water Resources Management*, 31(15), 4715–4729.

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., and Rizzoli, A.E. (2015). Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environmental Modelling & Software*, 72, 198–214.

Daubechies, I. (1992). *Ten lectures on wavelets*. Number 61 in CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pa.

Du, K., Zhao, Y., and Lei, J. (2017). The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. *Journal of Hydrology*, 552, 44–51.

Effrosynidis, D. and Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. *Ecological Informatics*, 61, 101224.

Gers, F.A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10), 2451–2471.

Graf, R., Zhu, S., and Sivakumar, B. (2019). Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach. *Journal of Hydrology*, 578, 124115.

Guo, G., Liu, S., Wu, Y., Li, J., Zhou, R., and Zhu, X. (2018). Short-Term Water Demand Forecast Based on Deep Learning Method. *Journal of Water Resources Planning and Management*, 144(12), 04018076.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Hoekstra, A.Y., Buurman, J., and van Ginkel, K.C.H. (2018). Urban water security: A review. *Environmental Research Letters*, 13(5), 053002.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems*, 30.

Mu, L., Zheng, F., Tao, R., Zhang, Q., and Kapelan, Z. (2020). Hourly and Daily Urban Water Demand Predictions Using a Long Short-Term Memory Based Model. *J. Water Resour. Plann. Manage.*, 11.

Nourani, V., Hosseini Baghanam, A., Adamowski, J., and Kisi, O. (2014). Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. *Journal of Hydrology*, 514, 358–377.

Pacchin, E., Gagliardi, F., Alvisi, S., and Franchini, M. (2019). A Comparison of Short-Term Water Demand Forecasting Models. *Water Resources Management*, 33(4), 1481–1497.

Percival, D.B. and Walden, A.T. (2000). *Wavelet methods for time series analysis*, volume 4. Cambridge university press.

Qi, Z., Zheng, F., Guo, D., Zhang, T., Shao, Y., Yu, T., Zhang, K., and Maier, H.R. (2018). A Comprehensive Framework to Evaluate Hydraulic and Water Quality Impacts of Pipe Breaks on Water Distribution Systems. *Water Resources Research*, 54(10), 8174–8195.

Quilty, J. and Adamowski, J. (2018). Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. *Journal of Hydrology*, 563, 336–353.

Rezaali, M., Quilty, J., and Karimi, A. (2021). Probabilistic urban water demand forecasting using wavelet-based machine learning models. *Journal of Hydrology*, 600, 126358.

Tiwari, M.K. and Adamowski, J. (2013). Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models: Ensemble Water Demand Forecasting. *Water Resources Research*, 49(10), 6486–6507.

UNEP, E. (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.

Vapnik, V., Golowich, S.E., and Smola, A.J. (1996). Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. *Advances in neural information processing systems*, 9.

Zhou, F., Liu, B., and Duan, K. (2020). Coupling wavelet transform and artificial neural network for forecasting estuarine salinity. *Journal of Hydrology*, 588, 125127.

Zounemat-Kermani, M., Matta, E., Cominola, A., Xia, X., Zhang, Q., Liang, Q., and Hinkelmann, R. (2020). Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects. *Journal of Hydrology*, 588, 125085.