Wastewater flow forecasting model based on the nonlinear autoregressive with exogenous inputs (NARX) neural network

Khalid El Ghazouli a,b,*, Jamal El Khatabi c, Isam Shahrour c and Aziz Soulhi c

a ULR 4515 - LGCE, Laboratoire de Génie Civil et géo-Environnement, Université de Lille, IMT Lille Douai, Université d’Artois, Inreca Hauts-de-France, F-59000 Lille, France
b Laboratoire d’Analyse des Systèmes, Traitement de l’Information et Management Industriel, Université Mohammed V, Rabat, Morocco
c National Higher School of Mines, Agdal Rabat, Morocco
*Corresponding author. E-mail: elghazouli.khalid@gmail.com

KEYWORDS: artificial intelligence, NARX neural network, sewer network, urban drainage system, wastewater flow forecast

ABSTRACT

Wastewater flow forecasts are key components in the short- and long-term management of sewer systems. Forecasting flows in sewer networks constitutes a considerable uncertainty for operators due to the nonlinear relationship between causal variables and wastewater flows. This work aimed to fill the gaps in the wastewater flow forecasting research by proposing a novel wastewater flow forecasting model (WWFFM) based on the nonlinear autoregressive with exogenous inputs neural network, real-time, and forecasted water consumption with an application to the sewer system of Casablanca in Morocco. Furthermore, this research compared the two approaches of the forecasting model. The first approach consists of forecasting wastewater flows on the basis of real-time water consumption and infiltration flows, and the second approach considers the same input in addition to water distribution flow forecasts. The results indicate that both approaches show accurate and similar performances in predicting wastewater flows, while the forecasting horizon does not exceed the watershed lag time. For prediction horizons that exceed the lag time value, the WWFFM with water distribution forecasts provided more reliable forecasts for long-time horizons. The proposed WWFFM could benefit operators by providing valuable input data for predictive models to enhance sewer system efficiency.

INTRODUCTION

Wastewater flow forecasts are key components in the short- and long-term management of sewer systems. In wastewater treatment plants (WWTPs), a wastewater flow forecasting model (WWFFM) could benefit operators by providing valuable input data for predictive models to simulate plant behavior and optimize performances and costs through the control of biological processes (Fernandez et al. 2009). For pumping stations, selecting the best pump scheduling configuration and running the pumps with an appropriate adjustment of rotation speed could help save energy (Wei et al. 2013). These forecasts could also enhance the performance and cost-effectiveness of real-time chemical dosing controllers, thereby preventing hydrogen sulfide formation (Chen et al. 2014).

Several models based on data-driven modeling for forecasting wastewater flows have been developed to address these challenges during the last decade. Wei et al. (2013) developed a multilayer perceptron (MLP) neural network model for the short-term prediction of influent flow rates in WWTPs. This model takes influent flow rate, rainfall rate, and radar reflectivity as inputs and returns an accurate flow forecast with a prediction horizon of up to 180 min. Boyd et al. (2019) proposed a model based on an autoregressive integrated moving average for daily influent flow forecasts tested at five WWTPs across North America and was completed with a multilayer perceptron neural network proposed by Zhang et al. (2019). These models rely only on historical data with no adaptation or external inputs.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/).
external inputs. Although these models are efficient, they remain limited in their approach. In fact, for forecasting wastewater flows, these models only consider sewer flow historical data. Moreover, they do not integrate drinking water consumption, which is the main causal variable that may influence forecasted flows in the case of a water shutdown in a sector or water consumption variation due to a given event.

The current work aimed to fill the gaps in the wastewater flow forecasting research by proposing a novel WWFFM based on the nonlinear autoregressive with exogenous inputs neural network (NARX-NN), real-time, and forecasted water consumption with an application to the sewer system of Casablanca in Morocco.

MATERIALS AND METHODS

The WWFFM aims at predicting instantaneous dry weather flows at specific points of watersheds. Dry weather flow usually corresponds to flows with no rainfall influence or at a maximum rainfall intensity of 0.3 mm and without inflows (Staufer et al. 2012). Given that the wastewater flow production function is nonlinear and depends on the spatial and temporal variations of water consumption through watersheds, using a model that can handle nonlinear problems for forecasting purposes is important. The proposed WWFFM is based on the NARX that has shown its efficiency through various nonlinear times-series forecasting applications (AbouRjeily et al. 2017; Koschwitz et al. 2018; Wunsch et al. 2018; Marcjasz et al. 2019; Di Nunno et al. 2021). The WWFFM considers real-time water consumption and previous infiltration flow records as inputs and predicted wastewater flows with forecast horizons that vary from 30 to 240 min as outputs. These periods offer a sufficient lead time to real-time and predictive control models to process and apply optimal control strategies.

The proposed architecture of the network includes two layers, namely, a hidden layer and an output layer (Figure 1). The inputs were weighted with appropriate weights (w), and the sum of the weighted inputs and biases forms the input to the transfer function. A nonlinear transfer function, the tan-sigmoid function bounded between –1 and 1 and described by Equation (1), was used in the hidden layer. An unbounded linear transfer function depicted by Equation (2) was employed in the output layer due to its ability to extrapolate to a certain extent beyond the training data range (Solomatine & Khada 2003):

\[
tansig(x) = \frac{1}{1 + e^{-2x}} - 1
\]

\[
purelin(x) = x
\]

The NARX-NN is considered a black box containing the information to be learned. In the beginning, the neural network architecture is composed of layers and nodes without any information or knowledge of the simulated phenomenon. During the learning stage, the weights and biases were adjusted according to an optimization algorithm to minimize the error of the neural network output and measured data. In addition, the Levenberg–Marquardt back-propagation function was utilized to train the artificial neural network, as it demonstrated its ability to speed up the convergence rate of neural networks with MLP architectures (Hagan & Menhaj 1994). The Levenberg–Marquardt algorithm described by Equation (3) combines the gradient descent method that updates the parameters in the steepest descent direction to reduce the sum of the squared quadratic errors. Additionally, the Gauss–Newton method reduces the sum of squared errors, assuming that the least-squares

![Figure 1](http://iwaponline.com/h2open/article-pdf/doi/10.2166/h2oj.2021.107/946636/h2oj2021107.pdf)

**Figure 1** | Neural network architecture.
function is quadratic in the parameters and finding the minimum of this quadratic:

$$\Delta \omega = [J^T(\omega)J(\omega) + \lambda I]^{-1}J^T(\omega)e(\omega)$$  \hspace{1cm} (3)$$

where $\omega$ is the weight vector, $J$ is the Jacobian matrix, $J^T$ is the transpose matrix of $J$, $\lambda$ is a learning parameter, $I$ is the identity matrix, and $e$ is the vector of the network error.

The early stopping method for improving generalization was used, and the divide block method was employed to split the dataset into three subsets. The first subset representing 70% of the data is the training set, which was utilized to compute the gradient and update the network weights and biases to find the model parameters. The second subset is the validation set (15%). The error in the validation set was monitored during the training process to avoid the increase of errors in the validation set and overfitting. When a validation error increases for a specified number of iterations (six iterations in our case), the training is stopped, and the weights and biases at the minimum of the validation error are returned. Furthermore, the total number of allowed epochs was set to 1,000. The remaining 15% of the dataset was employed as a test set to assess the generalization error in the final model.

The NARX trained in its open-loop form (Figure 2(a)) also called series-parallel architecture, given by Equation (4), efficiently predicts a time-series value for a one-time step ahead. In the open-loop form, the predicted value $\hat{y}(t)$ of the target time series $y(t)$ is predicted from the past values of $u(t)$ and the past measured values of $y(t)$ with the appropriate tapped delay line:

$$\hat{y}(t+1) = f(y(t), y(t-1), \ldots, y(t-n_y), u(t+1), u(t), u(t-1), \ldots, u(t-n_u))$$  \hspace{1cm} (4)$$

Once the training process is over, the NARX is turned to its closed-loop form (Figure 2(b)), which is called the parallel architecture given by Equation (5) to perform multistep-ahead time-series forecasting. The closed-loop form takes the past and present values of $x(t)$ and $y(t)$ previously predicted values as inputs:

$$\hat{y}(t+1) = f(\hat{y}(t), \hat{y}(t-1), \ldots, \hat{y}(t-n_y), u(t+1), u(t), u(t-1), \ldots, u(t-n_u))$$  \hspace{1cm} (5)$$

Two statistical metrics were used in this study to assess the efficiency of the model. The Nash–Sutcliffe efficiency (NSE) given by Equation (6), where a value is closed to 1, represents a perfect fit between the observed and forecasted data. And the root-mean-square error (RMSE) is given by Equation (7), where low RMSEs are preferred for model validation:

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n} (Q_i^o - Q_i^f)^2}{\sum_{i=1}^{n} (Q_i^o - Q_i^o)^2}$$  \hspace{1cm} (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Q_i^o - Q_i^f)^2}{n}}$$  \hspace{1cm} (7)$$

Figure 2 | (a) Series-parallel architecture and (b) parallel architecture.
where $Q_i^o$ is the observed flow at time step $i$, $Q_i^f$ is the forecasted flow at time step $I$, $Q_{to}$ is the mean observed value, and $n$ is the number of observations.

In the present work, two approaches of the forecasting model were compared (Figure 3):

- The first approach consists of forecasting wastewater flows on the basis of real-time water distribution flows for eight district metering areas (DMAs) and infiltration flows.
- The second approach comprises forecasting wastewater flows according to infiltration flow, water demand flow, and short-term water demand forecasts for the eight DMAs. The water consumption forecasting model is based on a feed-forward back-propagation neural network. The input dataset is composed of historical temperature, water consumption, and days of specification data.

The water consumption forecasting model is based on a feed-forward back-propagation neural network that has shown its efficiency in forecasting water consumption on the campus of Lille University (Farah et al. 2019). The input dataset comprises historical temperature, water consumption, and days of specification data. The model gives as output, and water demand forecasts are used as inputs for the WWFFM.

In the model, days of specifications are represented as vectors containing information about the following:

- Day of the week (i.e., Monday to Sunday, where values range from 1 to 7).
- Holidays and special days (New Year’s Day and religious celebrations such as Aid El-Adha) are represented with a vector where the values are either 0 or 1.
- Special consumption periods as Ramadan, where consumption patterns differ from normal consumption ones. The vector values are either 0 or 1, where 1 corresponds to the Ramadan period.
- The daily time is represented with 288 5-min timesteps, where values range between 1 and 288.

### EXPERIMENTAL DATA

#### Site description

The data were collected from a watershed of 3,315 ha, which covers the townships of the Eastern part of Casablanca (Figure 4). The urbanization of the area is fairly heterogeneous and comprises industrial and residential areas. The urban drainage system (UDS) is a combined system in the historical part of the townships with a separate sewer system in the new urbanized areas.

#### Data collection and processing

The area is equipped with a monitoring system based on quantitative sensors that measure sewer flows at the watershed outlet and water consumption at the eight DMAs. The monitoring system of the DMAs is composed of insertion and electromagnetic flowmeters that conduct measurements at a 5-min time step. The UDS is

![Figure 3](https://iwaponline.com/h2open/article-pdf/doi/10.2166/h2oj.2021.107/946636/h2oj2021107.pdf)
equipped with a depth meter to measure the water level and a flow meter to measure the discharge at the watershed outlet. The measurement for the UDS is conducted at a 15-min time step.

In the framework of the current study, wastewater flow ($Q_w$), precipitation ($P$), water consumption ($W_c$), and temperature ($T$) data were collected for 3 years between March 2014 and July 2017.

The mean dry weather flow rate pattern presented in Figure 5 shows that wastewater flows vary between 390 L/s for the minimum night flow (MNF) and 900 L/s for the peak flow that occurs around 12:00 pm. Figure 6 illustrates the diurnal patterns for days of the week, average diurnal, seasonal patterns, and special diurnal patterns for specific periods. For normal days, the flow rates of water consumption vary from 270 to 1,100 L/s.

Figure 4 | Sewer system and DMAs of the studied area.

Figure 5 | Diurnal pattern of the mean dry wastewater flow rate.
with an average flow rate of 650 L/s and can reach a value of 1,600 L/s during the Aid El-Adha celebration. Furthermore, Figure 6(a) and 6(b) displays the similar variations of the diurnal patterns for each day of the week and each season, with a rise of the MNF in summer of approximately 70 L/s and the peak flow of nearly 150 L/s. For all the consumption patterns, the peak flow is recorded between 11:00 am and 12:00 pm and decreases to reach the MNF between 2:00 am and 4:00 pm. However, the water consumption diurnal pattern trend changes during Ramadan, where we observe an increase in water consumption during the night with a peak flow around 4:00 am.

**Figure 6 |** Diurnal patterns of water consumption flow rate for the seasons of the year (a), the days of the week (b), and special periods (c).
before the beginning of the fast and an MNF that shifts to 6:00 am. We can also observe a fast drop and variation in water consumption roughly 7:00 pm, which corresponds to the fast break time.

However, given that the main sewer system was combined, the first step consisted of identifying rainy days on the basis of the rainfall records of the rain gauges and removing the corresponding data to keep only dry weather flows in the dataset.

For model predictive control systems and forecasting models, missing data constitute a major issue that does not fulfill the requirements of algorithms (Yuri et al. 2016). These problems could result from several factors, such as a power outage or a communication failure between the remote terminal units and the SCADA system (Walski et al. 2003). Many filling methods were proposed and could be found in the literature (Li et al. 2006; Qin et al. 2009; Fan et al. 2012), such as artificial filling, average value filling, special value filling, and regression. The reconstitution of the missing values of the dataset was performed through a linear interpolation.

In addition to missing values, data from field measurements usually include noise (Ruiz et al. 2016) that can affect the efficiency of machine learning algorithms (Lucas 2010; Munawar et al. 2011). The LOESS nonparametric regression method proposed by Cleveland (1979) and further developed by Cleveland et al. (1988), Cleveland & Grosse (1991), and Cleveland et al. (1992) was employed to smooth the collected data (Figure 7).

Dry weather flows in sewer networks consist of strict wastewater flows and infiltration flows (Figure 8). The origin of infiltration water or ‘parasite water’ commonly corresponds to diffuse groundwater infiltration or seawater. This water enters the network through leaky joints, cracks, and defective manholes. Therefore, considering infiltration rate variation as an input for our model and decomposing the hydrogram components into strict wastewater and infiltration are essential. Many studies have developed and applied methods for the

Figure 7 | Smoothed data with the LOESS method.

Figure 8 | Flow components in sewer networks.
quantification and detection of infiltration water and could be found in the literature (Ertl et al. 2002; Weiss et al. 2002; Mitchell et al. 2006; Ertl et al. 2008; Staufer et al. 2012; WSAA 2013; USEPA 2014; Water NZ 2015; Hey et al. 2016). There are two common methods for quantifying the base infiltration flow (BIF), namely, the flow rate method based on daily flow monitoring and the tracer method based on natural tracers or pollutant load mass balance (Hey et al. 2016). The infiltration rate was determined on the basis of the flow rate method according to the following equation:

\[
BIF = \frac{\text{MWF}}{\text{MNF}(1 - RL)RC}
\]

where MWF is the average MNF on the last three dry weather days, MNF is the minimum water consumption flow, RL is the real loss percentage where values range between 23 and 25%, and RC is a restitution coefficient equal to 80% and corresponding to the fraction of consumed water released back to the sewer network.

**Data analysis**

The visualization of the total distributed water and the wastewater flows (Figure 9) shows that the maximum lag time between the peaks of these two variables is around 80 min. Additional lag time analysis was performed using the cross-correlation analysis between distribution water and wastewater flows (Figure 10). The analysis results show a high correlation between these two variables because the lag is less than 80 min. Above this value, the correlation starts decreasing under 80%, exhibiting a weaker relation between both variables. Thus, the lag value for the NARX model is considered to be 80 min, corresponding to 16-time step delays for the NARX.

**Figure 9** | Plot of water consumption and wastewater flow peaks.

**Figure 10** | Cross-correlation analysis.
RESULTS

During the training stage, the NARX neural network minimizes the error between the model results and the real observed data. A different number of neurons were tested and, after several trials, the best training, testing, and validation results were obtained with a hidden layer with 10 neurons allowing the reduction of the mean squared error (MSE) that decreases from $10^5$ at the beginning of the training stage to 0.17 after 302 iterations. Tables 1 and 2 present the performance statistics of the NARX-NN architectures. The presented results show that increasing the number of neurons increases the efficiency of the model. However, increasing the number of neurons to more than 10 results in poor performances in multistep ahead forecasts. Figure 11 shows the performance of the trained ANN in the training, validation, and testing sets. In addition, Figure 12 highlights that the efficiency of the trained network presented by high regression values ($R$) of 0.999 is presented for the training, validation, and testing parts.

Once the model had been trained, further validation of the accuracy of the WWFFM was performed through multistep ahead predictions for 5 days, from September 8, 2016 to September 12, 2016, with hidden data not used during the training process. Figure 13 exhibits the water consumption of the eight DMAs and BIF employed for forecasting wastewater flows for a 5-day period. During this period, high water consumption was recorded on September 12 and corresponded to Aid El-Adha celebration day. The predictions of the WWFFM were

| Table 1 | RMSE for the different number of neurons |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $\mathbf{Q_t}$  | 1 neuron | 5 neurons | 10 neurons | 15 neurons | 1 neuron | 5 neurons | 10 neurons | 15 neurons |
| $Q_{t+6}$ | 0.397 | 0.6181 | 1.828 | 1.495 | 6.311 | 5.809 | 5.603 | 4.727 |
| $Q_{t+9}$ | 4.119 | 3.823 | 4.828 | 1.762 | 2.287 | 2.180 | 1.729 | 6.413 |
| $Q_{t+12}$ | 8.081 | 7.676 | 7.238 | 14.678 | 24.108 | 17.659 | 16.922 | 29.529 |
| $Q_{t+15}$ | 24.108 | 17.659 | 16.922 | 29.529 | 35.447 | 25.646 | 17.868 | 29.528 |

| Table 2 | NSE for the different number of neurons |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $\mathbf{Q_t}$  | 1 neuron | 5 neurons | 10 neurons | 15 neurons | 1 neuron | 5 neurons | 10 neurons | 15 neurons |
| $Q_{t+6}$ | 0.9998 | 0.9999 | 0.9995 | 0.9997 | 0.9934 | 0.9944 | 0.9948 | 0.9963 |
| $Q_{t+9}$ | 0.9918 | 0.9933 | 0.9957 | 0.9924 | 0.9989 | 0.9990 | 0.9994 | 0.9924 |
| $Q_{t+12}$ | 0.9892 | 0.9866 | 0.9806 | 0.9916 | 0.9852 | 0.9866 | 0.9756 | 0.9511 |
| $Q_{t+15}$ | 0.8348 | 0.9113 | 0.9756 | 0.7521 | 0.6881 | 0.8367 | 0.9207 | 0.7835 |

Figure 11 | Performance evaluation of the trained neural network.
Figure 12 | Regression results of the trained NARX-NN.

Figure 13 | Plot of water consumption of the eight DMAs and BIF.
conducted for different horizons $Q_{t+k}$. Where $Q_t$ designates the wastewater flow at timestep $t$, while $Q_{t+k}$ stands for the wastewater flow at timestep $t + k$ $(k = 6, 9, 12, 15, 18, 24, \text{and } 48)$ with a 5-min time step.

Tables 3 and 4 present the performance statistics of the WWFFM without water demand forecasts and the WWFFM with water demand forecasts, respectively. Figure 14(a)–14(g) depicts the predicted and observed flows for both approaches.

The analysis of the error statistical results in Figure 14 demonstrates that the WWFFM model with both approaches shows good performances in forecasting dry weather flow as long as the lag time remains less than 80 min. The forecast results are highly accurate, with an RMSE ranging between 3.3 and 16.16 and an NSE ranging between 0.995 and 0.999. Nonetheless, for prediction horizons exceeding 80 min, the WWFFM without water distribution forecasts has a poor performance that decreases with the increase of the forecasting horizon that fails to predict peak, especially for September 12, where the NARX-NN overestimates the peak flow of more than 550 L/s. Conversely, the WWFFM with water distribution forecasts enables the forecast of long-time horizons with a slight variation of the RMSEs over the different forecasting horizons ranging between 3.5 and 12.

**DISCUSSION**

The current study explored a new approach for predicting instantaneous dry weather flows in the UDS on the basis of the NARX-NN and drinking water consumption, and such an approach was tested on a part of the sewer system of Casablanca, which comprises approximately five million people. The construction of the model required essential steps to reconstitute data through linear interpolation because most modeling techniques cannot deal with missing values and cast out the whole instance value if one of the variable values is missing. In addition, the LOESS nonparametric regression method was used to smooth the data lying far from the bulk of the data range, and a cross-correlation analysis was also conducted to assess the suitable lagged information of the model.

The findings of this study validate that both tested approaches of the WWFFM display accurate results and similar performances in predicting dry weather flows with low RMSEs less than 16.16 and high NSEs as long as the forecasting horizon does not exceed 80 min. Nonetheless, the results further confirm that for prediction horizons that exceed 80 min, the WWFFM without water distribution forecasts presents poor performances that decrease with the increase of the forecasting horizon due to the lack of appropriate causal input variables, thereby making it unsuitable for long-time horizon forecasts for model predictive system use. Conversely, the WWFFM with water distribution forecasts is continuously updated with appropriate lagged input data, thereby enabling it to perform highly accurate forecasts for long-time horizons though representing all the flow ranges. The findings also highlight the importance of the WWFFM that could benefit operators and water engineers, thereby providing valuable input data for predictive model control to enhance the efficiency of sewer systems.

To our knowledge, this is the first study that has explored this new approach of forecasting dry weather flows on the basis of real-time water consumption and the BIF, which thus improves the knowledge of and complements previous research works in forecasting dry weather flows. The currently known models proposed in the literature

**Table 3 | Performance statistics of the WWFFM without water demand forecasts**

|          | $Q_{t+6}$ | $Q_{t+9}$ | $Q_{t+12}$ | $Q_{t+15}$ | $Q_{t+18}$ | $Q_{t+24}$ | $Q_{t+48}$ |
|----------|-----------|-----------|------------|------------|------------|------------|------------|
| RMSE (m$^3$ s$^{-1}$) | 3.300     | 5.492     | 10.383     | 16.166     | 18.918     | 43.487     | 82.855     |
| NSE      | 0.999     | 0.999     | 0.998      | 0.995      | 0.993      | 0.967      | 0.881      |

**Table 4 | Performance statistics of the WWFFM with water consumption forecast**

|          | $Q_{t+6}$ | $Q_{t+9}$ | $Q_{t+12}$ | $Q_{t+15}$ | $Q_{t+18}$ | $Q_{t+24}$ | $Q_{t+48}$ |
|----------|-----------|-----------|------------|------------|------------|------------|------------|
| RMSE (m$^3$ s$^{-1}$) | 3.504     | 4.367     | 4.135      | 4.7915     | 11.711     | 11.888     | 12.017     |
| NSE      | 0.999     | 0.999     | 0.999      | 0.999      | 0.997      | 0.997      | 0.997      |
(Wei et al. 2013; Boyd et al. 2019; Zhang et al. 2019) rely only on historical data with no external inputs. Additionally, they do not integrate drinking water consumption, which is the main causal variable that may influence forecasted flows in case of a water shutdown in a sector or water consumption variation due to a given event.

Figure 14 | Prediction of (a) \( Q_t + 6 \), (b) \( Q_t + 9 \), (c) \( Q_t + 12 \), (d) \( Q_t + 15 \), (e) \( Q_t + 18 \), (f) \( Q_t + 24 \), and (g) \( Q_t + 48 \) using the NARX-NN.

(continued.)
The limitation of the proposed WWFFM model lies in its use of real-time data, which can pose a problem in the event of data unavailability due to a sensor failure or a communication problem. Therefore, ensuring the good maintenance of the flow meters and continuous data transmissions for the needs of the NARX-NN is essential. Moreover, defining strategies for filling in data in case of communication failures would be interesting. In the meantime, the proposed model only integrates the forecasts of wastewater flows, and it is planned in the perspective of future works to develop the model by integrating the forecasts of combined sewer flows considering the fraction of stormwater flows.

CONCLUSION

The present work aims to fill the gaps in the wastewater flow forecasting research across the world by proposing a novel WWFFM based on the NARX. The proposed model considers real-time and forecasted water consumption as the main causal variable input of wastewater flow production. This study differs from the approaches presented through the literature that remain limited considering the only sewer flow historical data and that would fail to forecast sewer flows in the case of a water shutdown in a sector or water consumption variation due to a given event. This research compares the two approaches of the forecasting model. The first approach consists of forecasting wastewater flows on the basis of real-time water consumption and infiltration flows, and the second
The approach considers the same input in addition to the water distribution flow forecasts. Consequently, both approaches display accurate results and similar performances in predicting wastewater flows, while the forecasting horizon does not exceed 80 min. Nonetheless, for prediction horizons that exceed 80 min, the WWFFM without water distribution forecasts presents poor performances that decrease with the increase of the forecasting horizon. Conversely, the WWFFM with water distribution forecasts is continuously updated with the appropriate lagged input data, thereby making it able to perform highly accurate forecasts for long-time horizons. Hence, the WWFFM developed in this study could benefit operators and water engineers, providing valuable input data for predictive model control and thus enhancing UDS efficiency.

**DATA AVAILABILITY STATEMENT**

Data cannot be made publicly available; readers should contact the corresponding author for details.

**REFERENCES**

Abou Rjeily, Y., Abbas, O., Sadek, M., Shahrour, I. & Hage Chehade, F. 2017 Flood forecasting within urban drainage systems using NARX neural network. *Water Science and Technology* 76 (9), 2401–2412.

Boyd, G., Na, D., Li, Z., Snowling, S., Zhang, Q. & Zhou, P. 2019 Influent forecasting for wastewater treatment plants in North America. *Sustainability* 11 (6), 1764.

Chen, J., Ganighe, R., Liu, Y. & Yuan, Z. 2014 Real-time multi-step prediction of sewer flow for online chemical dosing control. *ASCE Journal of Environmental Engineering*.

Cleveland, W. S. 1979 Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74 (368), 829–836.

Cleveland, W. S. & Grosse, E. 1991 Computational methods for local regression. *Statistics and Computing* 1, 47–62.

Cleveland, W. S., Devlin, S. J. & Grosse, E. 1988 Regression by local fitting. *Journal of Econometrics* 37, 87–114.

Cleveland, W. S., Grosse, E. & Ming-Jen, S. 1992 *A Package of C and Fortran Routines for Fitting Local Regression Models*. Unpublished paper.

Di Nunno, F., Granata, F., Gargano, R. & Marinis, G. 2021 Prediction of spring flows using nonlinear autoregressive exogenous (NARX) neural network models. *Environmental Monitoring and Assessment* 195, 350.

Ertl, T., Dlauhy, F. & Haberl, L. 2002 Investigations of the amount of infiltration inflow in to a sewage system. In: *Proceedings of the 3rd ‘Sewer Processes and Networks’ International Conference*, Paris, France, pp. 15–17.

Ertl, T., Spazierer, G. & Wildt, S. 2008 Estimating groundwater infiltration into sewerages by using the moving minimum method – a survey in Austria. In: *11th International Conference on Urban Drainage*, Edinburgh, Scotland, UK.

Fan, B., Zhang, G. & Li, H. 2012 Multiple models fusion for pattern classification on noise data. In: *2012 International Conference on System Science and Engineering (ICSSE)*, pp. 64–68.

Farah, E., Abdallah, A. & Shahrour, I. 2019 Prediction of water consumption using artificial neural networks modelling (ANN). *MATEC Web of Conferences* 295, 01004.

Fernandez, F. J., Seco, A., Ferrer, F. & Rodrigo, M. A. 2009 Use of neurofuzzy networks to improve wastewater flow-rate forecasting. *Environmental Modelling & Software* 24 (6), 686–693.

Gavin, B., Dain, N., Zhong, L., Spencer, S., Qianjian, Z. & Pengxiao, Z. 2019 Influent forecasting for wastewater treatment plants in North America. *Sustainability* 11 (6), 1764. https://doi.org/10.3390/su11061764.

Hagan, M. T. & Menhaj, M. B. 1994 Training feed-forward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks* 5, 989–993.

Hey, G., Jonsson, K. & Mattsson, A. 2016 *The Impact of Infiltration and Inflow on Waste Water Treatment Plants: A Case Study in Sweden*.

Koschwitz, D., Frisch, J. & Van Treec, J. C. 2018 Data-driven heating and cooling load predictions for non-residential buildings based on support vector machine regression and NARX recurrent neural network: a comparative study on district scale. *Energy* 165, 134–142.

Li, J., Li, P. & Shu, K. 2006 RMINE: a rough set based data mining prototype for the reasoning of incomplete data in condition-based fault diagnosis. *Journal of Intelligent Manufacturing* 1, 163–176.

Lucas, A. 2010 Corporate data quality management: from theory to practice. *Industrial Electronics and Applications*.

Marcjasz, G., Uniejewski, B. & Weron, R. 2019 On the importance of the long-term seasonal component in day-ahead electricity forecasting with NARX neural networks. *International Journal of Forecasting* 35 (4), 1520–1532.

Mitchell, P., Stevens, P. & Nazaroff, A. 2006 Determining base infiltration in sewers: a comparison of empirical methods and verification results. pp. 1–13.

Munawar, M., Salim, N. & Ibrahim, R. 2011 Towards data quality into the data warehouse development. *International Journal of Electrical Power and Energy Systems*.

Qin, Y., Zhang, S. & Zhu, X. 2009 Pop algorithm: kernel-based imputation to treat missing data in knowledge discovery from databases. *Expert Systems with Applications* 2, 2794–2804.

Ruiz, L. G. B., Cuellar, M. P., Calvo-Flores, M. D. & Jiménez, M. D. C. P. 2016 An application of non-linear autoregressive neural networks to predict energy consumption in public buildings. *Energies* 9, 684.
Solomatine, D. P. & Khada, N. D. 2003 Model trees as an alternative to neural networks in rainfall–runoff modelling. *Hydrological Sciences Journal* 48 (3), 399–411.

Staufer, P., Scheidegger, A. & Rieckermann, J. 2012 Assessing the performance of sewer rehabilitation on the reduction of infiltration and inflow. *Water Research* 46 (16), 5185–5196.

Walski, T., Chase, D. V., Savic, D. A., Grayman, W. M., Beckwith, S. & Edmunds, K. 2003 *Advanced Water Distribution Modeling and Management*, pp. 244–248.

Water NZ 2015 *Infiltration and Inflow Control Manual*, pp. 1–2.

Wei, X., Kusiak, A. & Sadat, H. R. 2013 *Prediction of Influent Flow Rate: Data-Mining Approach*.

Weiss, G., Brombach, H. & Haller, B. 2002 Infiltration and inflow in combined sewer systems: long-term analysis. *Water Science & Technology* 45, 227–230.

WSAA 2013 *Good Practice Guidelines for Management of Wastewater System Inflow and Infiltration*, pp. 1–2.

Wunsch, A., Liesch, T. & Broda, S. 2018 Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). *Journal of Hydrology* 567, 743–758.

Yuri, A. W., Shadr, X. Y. & Steven, X. D. 2016 Quantisation and data quality: implications for system identification. *Journal of Process Control* 40, 13–23.

Zhang, Q., Li, Z., Snowling, S., Siam, A. & El-Dakhakhni, W. 2019 Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. *Water Science and Technology* 80 (2), 243–253.

First received 19 June 2021; accepted in revised form 1 October 2021. Available online 12 October 2021