Hindawi
Scientific Programming
Volume 2019, Article ID 8319549, 16 pages
https://doi.org/10.1155/2019/8319549

Research Article

A High-Frequency Data-Driven Machine Learning Approach for Demand Forecasting in Smart Cities

Juan Carlos Preciado,[1] **Álvaro E. Prieto**,[1] **Rafael Benitez**,[2] **Roberto Rodríguez-Echeverría**,[1] **and José María Conejero**[1]

[1]*Dept. Ingeniería Sistemas Informáticos y Telemáticos, Universidad de Extremadura, Cáceres, Extremadura, Spain*
[2]*Dept. Matemáticas para la Economía y la Empresa, Universidad de Valencia, Valencia, Spain*

Correspondence should be addressed to Álvaro E. Prieto; aeprieto@unex.es

Received 7 March 2019; Revised 3 May 2019; Accepted 13 May 2019; Published 3 June 2019

Academic Editor: Can Özturan

Different types of sensors along the distribution pipelines are continuously measuring different parameters in Smart WAter Networks (SWAN). The huge amount of data generated contain measurements such as flow or pressure. Applying suitable algorithms to these data can warn about the possibility of leakage within the distribution network as soon as the data are gathered. Currently, the algorithms that deal with this problem are the result of numerous short-term water demand forecasting (WDF) approaches. However, in general, these WDF approaches share two shortcomings. The first one is that they provide low-frequency predictions. That is, most of them only provide predictions with 1-hour time steps, and only a few provide predictions with 15 min time steps. The second one is that most of them require estimating the annual seasonality or taking into account not only data about water demand but also about other factors, such as weather data, that make their use more complicated. To overcome these weaknesses, this work presents an approach to forecast the water demand based on pattern recognition and pattern-similarity techniques. The approach has a twofold contribution. Firstly, the predictions are provided with 1 min time steps within a time lead of 24 hours. Secondly, the laborious estimation of annual seasonality or the addition of other factors, such as weather data, is not needed. The paper also presents the promising results obtained after applying the approach for water demand forecasting to a real project for the detection and location of water leakages.

1. Introduction

The current big data scenario is based on using a large volume of data to get new insights and acquire knowledge that support the daily decision-making process [1]. One of the main sources of these data are IoT (Internet of Things) systems that collect and transfer a great amount of sensor data [2]. The use of these technologies for water management allows gathering data in order to monitor water usage and water waste, what is regarded as one of the application areas of a smart city [3]. In this sense, the application of information and communication technology (ICT) devices to water distribution systems (WDSs) is considered a key subarea of a smart city and introduces the concept of Smart WAter Network (SWAN) [4]. A SWAN consists of a large number of sensors that measure automatically and continously a wide range of parameters present in WDS.

It should be noted that WDSs are big and complex. Only in Europe, there are more than 3.5 million kilometers of pipes [5], and in the United States, around 159 billion liters of water are withdrawn from water sources each day [6]. The management of WDS implies to deal with different issues. One of them is the problem of water pressure that could affect significantly the level of service for the users and where there are novel approaches such as [7] that proposes the division of the network in subregions according to the expected water peak demand.

Another huge problem managing WDS is to deal with water loss. Water loss can be attributed to several causes, including leakage, metering errors, and fraud although

leakage is usually the major cause. It is estimated that the amount of water in the world that is lost is more than 30 percent of production [8].

The data obtained by the sensors that compose a SWAN can be an important turning point to avoid this problem. This is due to the fact that the usual gathered data include flow, pressure, or totalizer measurements. The application of water demand forecasting algorithms over all these data allows detecting leakages at an early stage.

There are several works that present different approaches to try to forecast the water demand applying different techniques. Due to the necessity to detect a water leakage as soon as this problem arises, the more suitable approaches are those with a short-term forecast horizon, that is, how far the prediction about the future demand is able to accurately reach. Thus, a short-term forecast horizon is generally considered for a range between 1 and 48 hours.

The existing short-term water demand forecast approaches can achieve good results. However, in general, they have in common two important limitations that the approach proposed in this work reduces.

The first limitation refers to its frequency, in other words, how many predictions within this horizon the approach is able to provide. The usual time steps of most of the approaches are 1 hour, so that a frequency of 24 predictions per day may be achieved. Only a few approaches provide higher frequency being, at most, one prediction every 15 minutes. Considering that the sooner the prediction is able to detect an anomaly, the better any improvement in the frequency of the predictions could significantly reduce the loss of water. Although the time horizon of our approach is on average (24 hours), we are able to get a time step of one minute, that is, a frequency of 1440 each day, without reducing the accuracy of the prediction. Notice that this is not a trivial contribution because we identified that neural networks approaches were unfeasible with this frequency and more classic methods such as ARIMA and dynamic harmonic regression were even too computationally expensive.

The second limitation concerns the data needed apart from previous water demand. Most of the current approaches need extra data about weather (temperature, rainfall, etc) or demand changes according to factors related to weekly or annual seasonality, being particularly the estimation of the latter, annual seasonality, a very demanding task. Our approach uses previous water demand data just considering weekly seasonality reduction and thus the complexity of its application. Therefore, it avoids the troublesome estimation and inclusion of annual seasonality or the usage of weather data.

Our approach is based on pattern similarity and is inspired by the work of Grzegorz Dudek [9–11] for short-term load forecasting in the daily operation of power systems and energy markets. It has been implemented using the model-driven development (MDD) paradigm [12, 13] and has been tested in one of the partner cities of the European project SmartWater4Europe [5]. The following goodness-of-fit (GoF) parameters have been used to determine the performance of the approach: MAPE (mean average percentage error), RMSE (root mean squared error), and FOB (fraction out of bounds).

It should also be emphasized that this approach not only reduces both aforementioned limitations but also presents the next advantage: (a) it is relatively easy to implement; (b) it is not highly time-consuming; (c) as the historical record increases, the performance improves; and (d) the method is robust enough to deal with minor data issues such as small segments of missing data. The latter avoids that it causes "false alarms".

The rest of the paper is organized as follows. In Section 2, we review previous work on water demand forecasting. Section 3 describes the locations where the data were gathered and the proposed algorithm. In Section 4, we present the results and discussion. Finally, the conclusions and future work are outlined in Section 5.

2. Related Work

Water demand has been a field where quantitative forecasting has been applied profusely because it meets the twofold requirement [14] to use this kind of forecasting: (a) there are historical numerical data about the variable to forecast and (b) it is plausible to presuppose that some features of the patterns recognized in the historical data are recurring.

We found a number of water demand forecasting approaches proposed in the literature. In this sense, there are works published during the 1990s that can be considered as fundamentals in this field such as the ones by Shvartser et al. [15] or Buchberger et al. [16, 17]. Donkor et al. [18] reviewed the literature on urban water demand forecasting published from 2000 to 2010, in order to identify the methods and models that are useful for specific water utility decision-making problems. More recently, Sebri [19] conducted a meta-analysis to estimate in a statistical way how different features of primary studies could influence the correctness of urban water demand forecasts.

In this section, we focus on reviewing the most relevant methods published since 2010 to date (to the best of our knowledge) focused on short-term predictions (1–48 hours) sorted according to the frequency used (from lowest to highest).

To begin with, Adamowski et al. [20] tested if coupled wavelet-neural network models (WA-ANNs) applied to forecast daily urban water demand could provide promising results during the summer months in the city of Montreal, Canada. They used daily total urban water demand, daily total precipitation, and daily maximum temperature, all of them gathered during the summer period to conduct their work. Concretely, they integrated artificial neural networks together with discrete wavelet transforms to elaborate coupled wavelet-neural network models. They stated that their approach provided better results forecasting short-term (24 hours) water demand than other techniques such as artificial neural networks (ANN) alone, autoregressive integrated moving average (ARIMA), multiple linear regression (MLR), or multiple nonlinear regression (MNLR).

However, their approach only provided one prediction for the whole day.

Herrera et al. [21] focused their work on trying to forecast the water demand in the next hour in an urban area of a city in southeastern Spain. Not only did they use previous water demand data but also temperature, wind velocity, atmospheric pressure, and rain data. They concluded that support vector regression (SVR) models were the more adequate ones for this task, and multivariate adaptive regression splines (MARS), projection pursuit regression (PPR), and random forest (RF) could also be used. However, the neural network that they used (feedforward neural networks with one hidden layer in conjunction with the backpropagation learning algorithm) seemed to provide very poor results.

Odan and Reis [22] compared different ANNs to forecast water demand. They used hourly consumption data from the water supply system of Araraquara, São Paulo, Brazil, as well as temperature and relative humidity data. Their estimations were made for the next 24 hours with a frequency of 1 for each hour. Concretely, they analyzed a multilayer perceptron with the backpropagation algorithm (MLP-BP), a dynamic neural network (DAN2), and two hybrid ANNs. The more interesting finding of their work is that the different variants of DAN2 that they used either to forecast the first hour or the whole 24 hours did not need the use of weather variables and achieved better results that the rest ones.

Ji et al. [23] used different factors along with a least-square support vector machine (SVM) to forecast water demand for one day with one-hour frequency. The factors that they have taken into account were flow data, the maximum and the minimum temperature, precipitations, holiday information, and information of incidents. The novelty of this work lies in the adjustment of the hyperparameters of the SVM system by using swarm intelligence via a teaching learning-based optimization algorithm.

Hutton and Kapelan [24] were concerned about the uncertainties that influenced the results of water demand forecasts and proposed an iterative methodology based on probabilistic that tried to decrease the effect of such uncertainties during the development of hourly short-term water demand prediction models. They used static calendar data in addition to water demand data. On the one hand, their approach exposed the unsuitability of simplistic Gaussian residual assumptions in predicting water demand. On the other hand, they concluded that a model whose kurtosis and heteroscedasticity in the residuals are revised iteratively using formal Bayesian likelihood functions allow building better predictive distributions.

Candelieri et al. [25–27] have works that make use of unsupervised (time series clustering) and supervised (support vector machines regression models) machine learning strategies. These strategies were combined in a two-stage framework in order to identify typical urban water demand patterns and successively provide reliable one day forecasts for each hour of the day. They used real data gathered from different sources of Milan (Italy) to check their proposal. Their last work extended the previous ones by allowing also anomaly detection.

Alvisi and Franchini [28] have the goal of estimating the predictive uncertainty in water demand forecasting. To this end, they joined short-term water demand predictions provided by two or more models by means of the model conditional processor (MCP). Then, MCP computed a probability distribution of the real future demand according to the different predictions of each particular model. This probability distribution, together with a predefined hourly pattern based on the season and the day of the week, allows them to estimate the expected hourly water demand for a whole day as well as the associated predictive uncertainty.

Brentan et al. [29] considered that the result of the applying fixed regression structure with time series can be biased and prone to errors. Their proposal tried to reduce both of them when building a short-term (24 hours) hourly water demand forecasting. To do this, firstly, they used support vector regression (SVR) together with calendar data to build a base forecasting, and secondly, they improved this forecasting applying Fourier time series process.

Romano and Kapelan [30] proposed the use of evolutionary artificial neural networks (EANNs) to perform adaptive hourly water demand forecasting for the whole next day. Their goal is to provide near real-time operational management by analyzing water demand time series and weekly seasonality. This approach was tested on a real-life UK case study, and one of its main features was that it did not need too much human intervention.

Gagliardi et al. [31] proposed two models based on homogeneous Markov chain model (HMC) and non-homogeneous Markov chain model (NHMC) to forecast next day hourly water demand. They used water demand data and weekly seasonality; concretely, they differentiated between working and nonworking days. They recommended the use of HMC to do this type of predictions because their results showed that its performance was better than the one obtained using NHMC.

Pacchin et al. [32] proposed a model based on moving windows that predicted the hourly water demand during the next day. This model presented two different features with respect to other similar models. On the one hand, it updated the prediction taking into account the demand data of the previous day. On the other hand, it did not need two much historic data in comparison with other models since it was able to do accurate predictions only using the data of three or four previous weeks. It also should be pointed that they also took into consideration the weekly seasonality.

Arandia et al. [33] proposed a methodology to predict 15 min, hourly, and daily water demand either offline (using historical data) or online (using a real-time feed of data). Their proposal joined seasonal ARIMA (SARIMA) and data assimilation. They also used in their approach weekly seasonality and daily periodicity and concluded than their methodology showed a better performance using weekly seasonality.

Bakker et al. [34] presented a model to forecast 15 min water demand for the next two days. Their model used static calendar data in addition to six years of water demand data gathered from different areas of the Netherlands. According to this work, a frequency of 15 minutes is more suitable than 1-hour frequency when detailed optimization is needed.

As we have seen, a number of approaches have been widely used for forecasting; however, as it is shown in Table 1, the frequency of these approaches is usually around 1 for each hour. Additionally, this table also shows the factors that each proposal needs to work apart from the previous water demand measurements. In most cases, the inclusion of more factors to make the forecast, such as annual calendar data or weather data, can be quite cumbersome. In turn, we propose the application of pattern similarity-based techniques proposed by Dudek [9–11] to the water demand forecasting problem. The main reason for selecting these techniques is their ability to simultaneously cope with the aforementioned difficulties: they remove the need to add weather data or to determine the annual seasonality by constructing the input and output patterns in which the series has been normalized, and at the same time, since the considered signal segments encompass a full day, the frequency of the predictions is 1440 per day.

3. Materials and Methods

This section describes the data sites used (taken from diverse real-world locations with different characteristics) and the preprocessing procedure carried out before starting the data analysis. In addition, we describe some relevant concepts, such as trends and seasonalities, before describing the input/output patterns and the proposed algorithm.

3.1. Data Sites. The algorithm has been tested in different locations of one of the member cities of the European project SmartWater4Europe. Concretely, this city is located in northern Spain, and it has about 180.000 inhabitants, with a population density of 1680 inhabitants per square km. With respect to the climate, it has an average annual precipitation of 546 mm, and it has a range of daily mean temperatures from 3.5°C in winter to 19.5°C in summer.

The data were collected by the company responsible of managing the water distribution of this city. This company has 58507 customers, the length of the distribution network is 467.315 metres, and the mean quantity of supplied water each day per inhabitant is 392,39 litres.

To gather the data, the company used the following:

(i) 14 sectoral sensors spread throughout the network pipes of the 3 sites (see below) that were able to measure flow, pressure, and totalizer each minute. This means that each sensor measures 1440 times a day and the company had been storing 20160 measurements (14×1440) each day for 10 years.

(ii) 1502 intelligent water meters spread throughout industries and homes located in the 3 sites (see below). In this case, each one performed 24 measurements per week.

Concretely, this company measured data of three different areas of this city whose characteristics are as follows.

(i) Site 1: Industrial Area. It is an industrial estate at the outskirts of the city. In this area, there are almost no domestic end users of the water supply.

(ii) Site 2: High-Density Population Area. It is a neighborhood located in the center of the city. It is a zone with high buildings where there are thousands of families.

(iii) Site 3: Low-Density Population Area. It is a suburb of the city. Most homes are either low-rise buildings or single-family homes, so the density of users is very low. It is important to note that the houses of this area have private backyards. It may be assumed that this is a factor which influences the water use pattern of the area.

At each timestamp, the minimum, maximum, and average flows (measured in l/min) were recorded. Table 2 shows, as an example, the first six measurements obtained by a sectoral sensor for the industrial area site. Note that the variable timestamp reflects the local time (CET), and +01 or +02 only reflects the difference from Greenwich Mean Time (GMT) (or coordinated universal time, abbreviated to UTC).

3.2. Proposed Algorithm. Domestic water demand data conform a time series with several seasonalities being the daily, weekly, and annual seasonalities the most important. In addition to these seasonalities, there are usually a long-term trend component and a high-frequency noise term.

As was mentioned before, the signal was sampled at a 1-minute frequency and we were considering a 24 hour forecast horizon. This means that, at any given moment, we need to forecast the next 1440 values of our signal. This rules out the possibility of using, directly, classical time series analysis methods such as ARIMA, exponential smoothing, and Winter–Holts methods. Moreover, direct neural network methods are also not feasible since for these methods the output layers would have 1440 neurons and the input layer would be much bigger, and therefore, the training of such large number of weights would require far more data than what is available.

The main problem here is that, with this high sampling frequency, the number of data needed in order to capture the weekly and annual seasonalities is simply too large. Therefore, we need to devise a method in which the seasonalities can be treated in a different way.

Our approach here is based on the pattern-similarity search proposed by Dudek in [9–11] for forecasting electric load. This method first splits the time series into segments of length equal to the forecast horizon and then maps those segments into two signals x and y—input and output signals—which will be used for a query-predict procedure. Those signals will be somehow normalized and will not be affected by trends and large period seasonalities. They will only contain the information within the forecast horizon (24 hours), and each 24-hour segment will be considered as a measure unit.

Table 1: Related work comparison with respect to frequency, forecast horizon, and other factors or complex estimation needed to apply the approach.

Work	Related work comparison		
	Frequency	Forecast horizon	Other factors
Adamowski et al. [20]	1 for each day	24 hours	Weather data during summer
Herrera et al. [21]	1 for each hour	1 hour	Weather data
Odan and Reis [22]	1 for each hour	24 hours	Weather data
Ji et al. [23]	1 for each hour	24 hours	Weather, holidays, and incident data
Hutton and Kapelan [24]	1 for each hour	24 hours	Annual calendar data
Candelieri et al. [25–27]	1 for each hour	24 hours	Working days and seasons of the year
Alvisi and Franchini [28]	1 for each hour	24 hours	Weekly seasonality and seasons of the year
Brentan et al. [29]	1 for each hour	24 hours	Annual calendar data
Romano and Kapelan [30]	1 for each hour	24 hours	Weekly seasonality
Gagliardi et al. [31]	1 for each hour	24 hours	Weekly seasonality
Pacchin et al. [32]	1 for each hour	24 hours	Weekly seasonality
Arandia et al. [33]	1 for each 15 minutes	24 hours	Daily and weekly seasonality
Bakker et al. [34]	1 for each 15 minutes	48 hours	Annual calendar data
Our proposal	1 for each minute	24 hours	Weekly seasonality

Table 2: Structure of the raw data from the industrial area (first six measurements).

	Industrial area		
Timestamp	Average	Maximum	Minimum
2014-01-01 00:00:00 + 01	3.278	4.031	2.919
2014-01-01 00:01:00 + 01	3.591	5.064	3.049
2014-01-01 00:02:00 + 01	4.875	5.352	4.518
2014-01-01 00:03:00 + 01	4.263	5.074	3.475
2014-01-01 00:04:00 + 01	3.966	5.004	3.406
2014-01-01 00:05:00 + 01	3.771	4.031	3.188
...	...	...	...

In particular, the input and output signals are defined by the following:

$$x_{i,t} = \frac{F_{i,t} - \overline{F}_i}{\sqrt{\sum_{l=1}^{n}\left(F_{i,l} - \overline{F}_i\right)^2}}, \quad t = 1, \ldots, n, \tag{1}$$

$$y_{i,t} = \frac{F_{i,t+\tau} - \overline{F}_i}{\sqrt{\sum_{l=1}^{n}\left(F_{i,l} - \overline{F}_i\right)^2}}, \quad t = 1, \ldots, n, \tag{2}$$

where $x_{i,t}$ and $y_{i,t}$ denote the input and output signals for day i at time t, respectively; $F_{i,t}$ denotes the water flow of the day i at time t; $\overline{F}_i$ denotes the average water flow of day i; τ is the forecast horizon; n is the number of measurements in each day (in our case, both τ and n are 1440).

On the one hand, the input signal x, also called the *query signal*, represents the normalized pattern for a current day with all its intraday information. On the other hand, the output signal y, also called the *forecast signal*, represents the normalized pattern of the following day (with our particular value of the forecast horizon). The normalization procedure filters all the seasonalities and trends beyond the daily frequency.

Now, the procedure would be as follows: for any given day i_0, we want to estimate the unknown value of the output signal y_{i_0} from the known input signal x_{i_0}. Once we have the estimation $\widehat{y}_{i_0}$, we can predict the values of the water demand for the forecast horizon using equation (2):

$$\widehat{F}_{i_0+1,t} = \widehat{F}_{i_0,t+\tau} = \overline{F}_{i_0} + \widehat{y}_{i_0,t}\sqrt{\sum_{l=1}^{n}\left(F_{i_0,l} - \overline{F}_{i_0}\right)^2}. \tag{3}$$

Therefore, the problem reduces to obtain the forecast for the output signal $\widehat{y}_{i_0}$. To make such forecast, we follow the next procedure (shown in Figure 1):

(1) We select the k nearest neighbors (using the Euclidean distance) of the query pattern x_{i_0} from the data in the history record from days of the same class (same day of the week/holiday) such that the following day is not an atypical day (e.g., holiday).

(2) We compute the estimate $\widehat{y}_{i_0}$ via the following equation:

$$\widehat{y}_{i_0} = \frac{1}{k}\sum_{j \in \Theta\left(x_{i_0}\right)} y_j, \tag{4}$$

where $\Theta(x_{i_0})$ is the set of indices of the k x patterns nearest to the query pattern x_{i_0} obtained in the previous step.

(3) Finally, we transform $\widehat{y}_{i_0}$ to obtain the water flow estimate according to equation (3).

Along with the estimation, we can obtain pointwise confidence bands:

$$I_{i_0,t} = \widehat{F}_{i_0+1,t} \pm T^*_{\alpha/2}\frac{S_{i_0,t}}{\sqrt{k}}\sqrt{\sum_{l=1}^{n}\left(F_{i_0,l} - \overline{F}_{i_0}\right)^2}, \quad t = 1, \ldots, n, \tag{5}$$

where $T^*_{\alpha/2}$ denotes the two-tailed critical value for Student's t-distribution with $k-1$ degrees for a confidence level α and $S^2_{i_0,t}$ is the sample variance of the k output signals used for the computation of $\widehat{y}_{i_0}$.

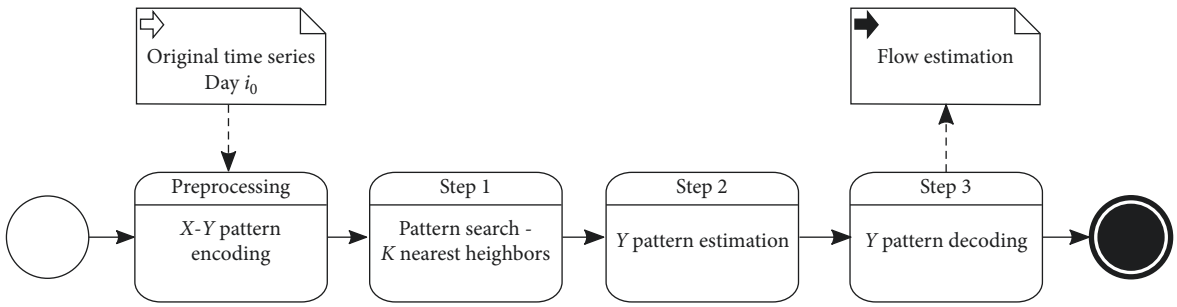

Figure 1: Forecasting procedure.

Once we have an estimation, we need to assess the quality of the forecast in order to validate the prediction model. We have considered three GoF parameters: the mean average percentage error (MAPE), the root mean squared error (RMSE), and the fraction out of bounds (FOB). The first two parameters are well-known error measures. The FOB for estimation at day i_0 is as follows:

$$\text{fob}_{i_0} = \frac{\text{NOB}_{i_0}}{\text{NIB}_{i_0} + \text{NOB}_{i_0}}, \tag{6}$$

where NOB_i and NIB_i are the number of measurements on day i_0 that lie outside and inside the confidence band for the given day i_0, respectively.

The MAPE parameter is very widely used in forecasting practice, but it becomes of little use when the actual values to be forecast are very small (close to zero). The problem is that, in our segments, there is a significant fraction of the day for which the water demand is indeed very small (night ours). The RMSE is an absolute value of the deviance of the forecast from the observed data. However, for small values of the water flow, it is difficult to assess the goodness of fit when the RMSE is small since the measure is not relative to the magnitude of the quantity to be predicted. Finally, the FOB can be regarded as a measure of the deviance of the observed day from what could be considered an average day of the same type. For small values of the FOB, we could say that the observed water flow corresponds to an "average" day, while if the FOB is large, the observed data does not follow the same pattern of other days of the same type in the historical record (and this could be related to either measurement anomalies or even water leaks).

4. Results and Discussion

4.1. Algorithm Parameters. Figure 2 depicts the average water flows vs. the day of the week for the three measurement sites. It is clearly shown that, for the industrial area site, there are at least four different patterns: Monday–Thursday, Friday, Saturday, and Sunday. For the high-density population area, there are two patterns corresponding to the labor days (Mon–Fri) and the weekends (Sat–Sun). Finally, at the low-density population area, there are more irregular patterns. Therefore, we considered the most restrictive pattern distribution (each day of the week to follow a different pattern, low-density population area) with the aim of easing the development of the algorithm.

Moreover, another distinct pattern is shown on holidays (Figure 3). Since the distribution of holidays varies from year to year, the inclusion of a holiday pattern in a model based on periodicities is difficult and cumbersome to implement.

The algorithm was tested for all days from 15 February 2014 until 18 September 2016. We did not start with the first measurements because we needed some weeks of historical data for the k nearest neighbors approach. Since we searched for the five nearest neighbors (see details below), we left a margin of seven weeks of historical data. The parameters considered for all sites were as follows: (a) the number of nearest neighbors: $k = 5$ for all years, (b) the threshold limits: min = 0.05 l/min, max = 100 l/min, and (c) the confidence level for the confidence band estimation: 90% (i.e., $\alpha = 0.1$).

The number of neighbors is an important parameter. If it is too small, the resulting pattern will not be representative of a true pattern for the forecasted day, but if it is too large, then the neighbors might be "far away" from the query pattern, and thus, we would be considering very different days for the estimation of our pattern.

4.2. Results. Figure 4 depicts a general perspective of all three GoF parameters at the three sites. All of them showed small global values: in 75% of the cases, the predictions showed values of FOB, MAPE, and RMSE less than 0.20, 39%, and 7.86 l/min, respectively, for the industrial area site; 0.21, 50%, and 1.80 l/min for the high-density population area site; and 0.25, 41%, and 5.82 l/min for the low-density population area site (Table 3).

Although the errors showed small values in the overall outcomes, a more thorough analysis is needed to determine the causes for the cases in which the parameters took higher values. To this aim, daily values were obtained and represented as scatter plots. For the sake of simplicity, values were categorized in three different levels: good, regular, and bad. The results are shown in Figure 5.

4.3. Discussion. The pattern-similarity forecasting method presented here proved very suitable for obtaining accurate daily predictions for the water flow values. However, we found several cases in which predictions were worse than what was expected or they simply delivered values completely different from the actual measurements.

At the industrial area site, the most difficult days to forecast were Sundays (for example, Figure 6). The main

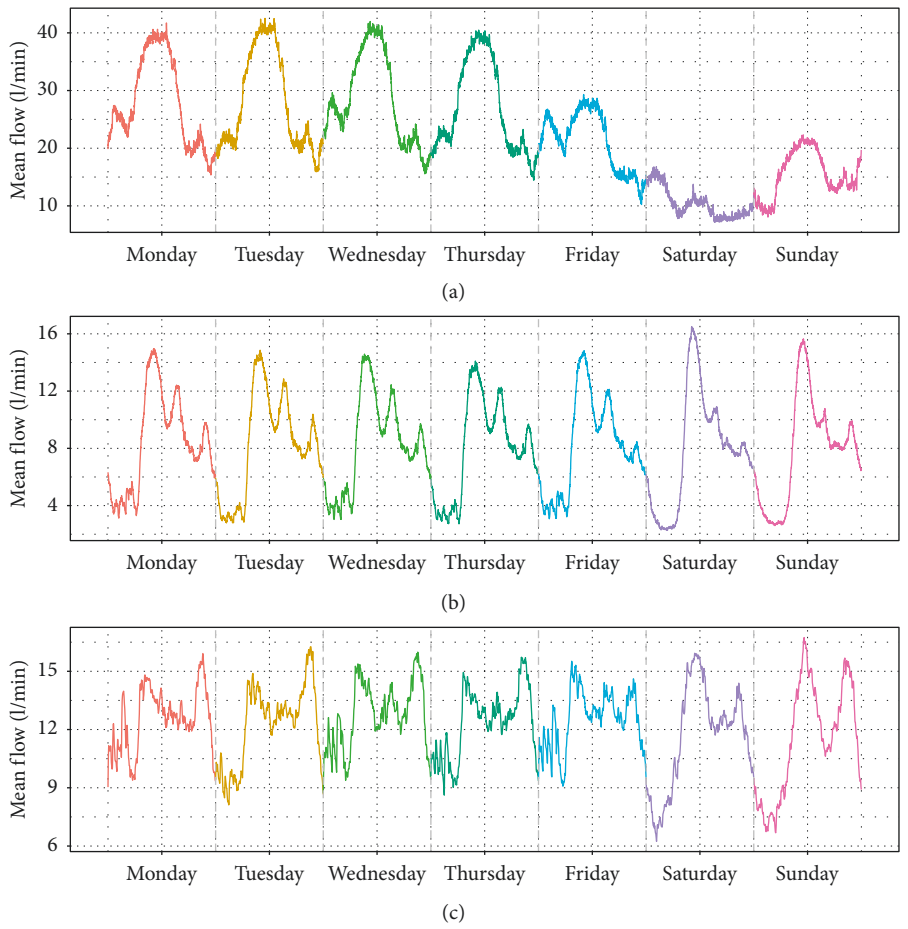

Figure 2: Mean flow vs. day of the week. For each of the three sites, the average for each minute of each weekday for the whole period is plotted. (a) Industrial area; (b) high-density population area; (c) low-density population area.

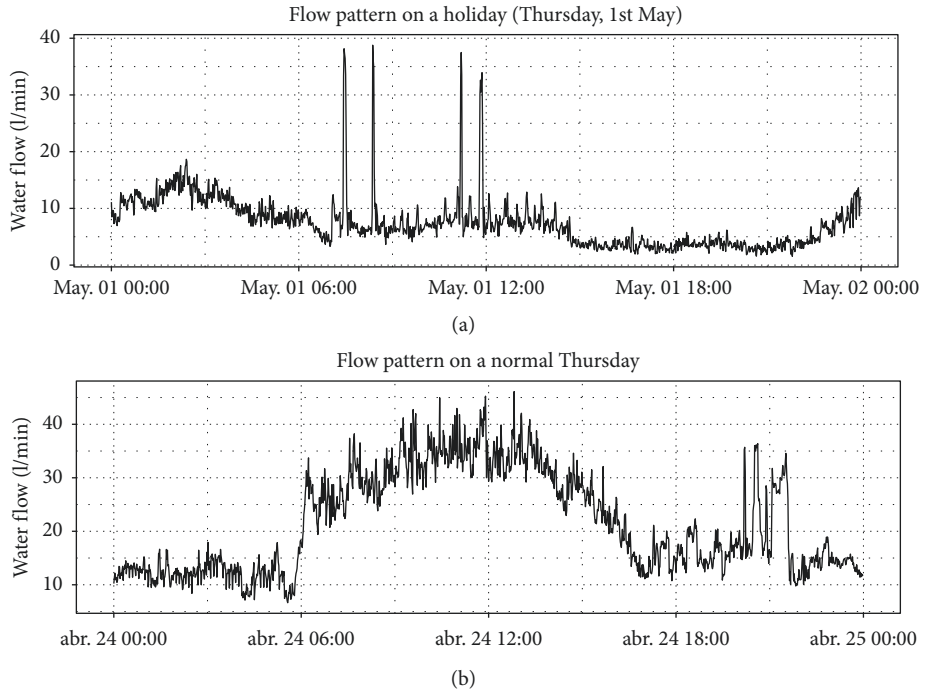

Figure 3: A typical holiday pattern (a) compared to a typical pattern for the same weekday on a nonholiday day (b).

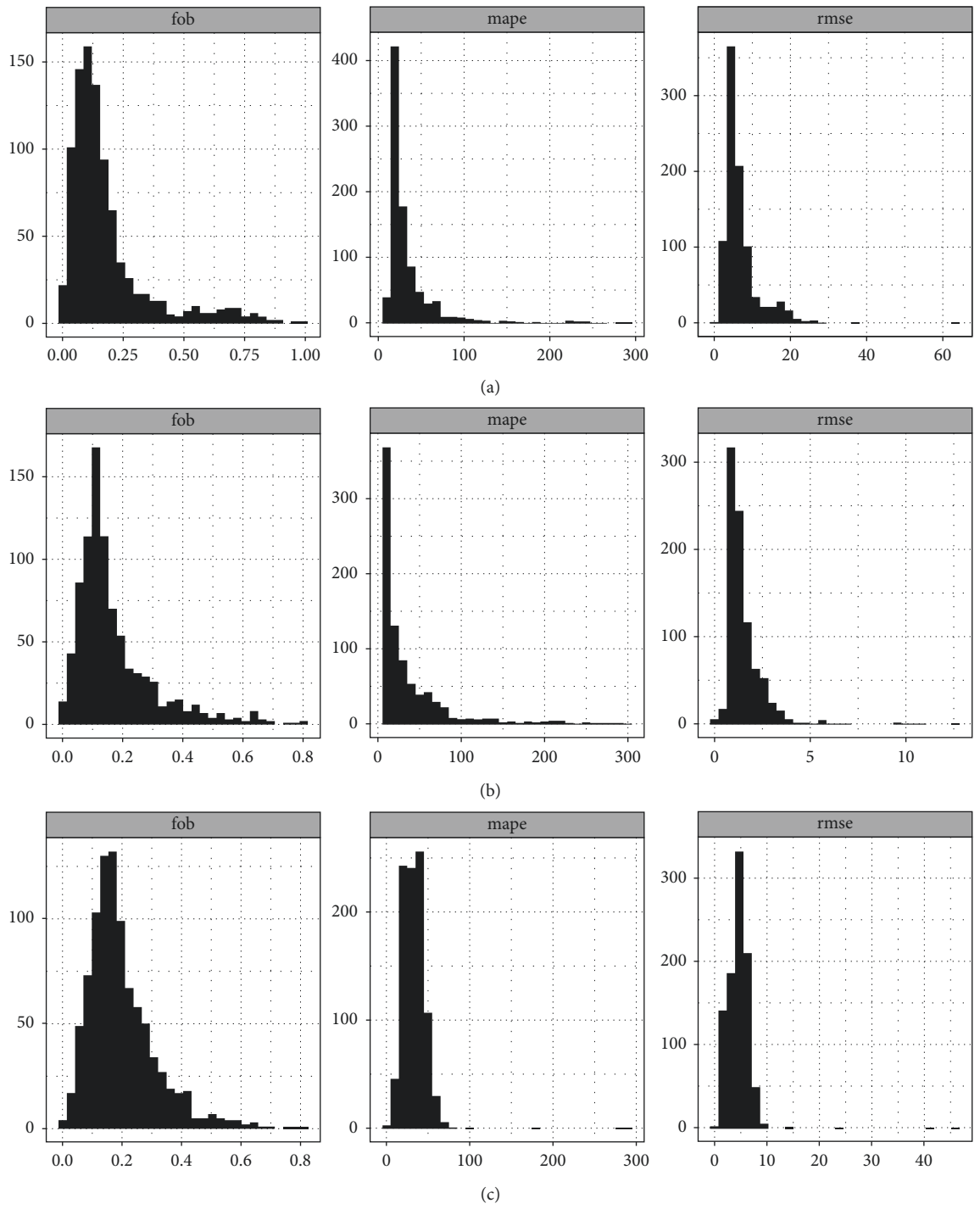

Figure 4: Distributions for the goodness of fit parameters values at the industrial area (a), the high-density population area (b), and the low-density population area (c) sites over the entire period of study, 2014–2016.

reason for this difficulty was that there were several different patterns for Sundays, but on the other hand, all Saturdays showed almost the same pattern. For any given Sunday, we took the preceding Saturday X pattern and we looked for the k nearest neighbors for this Saturday pattern. Since most of that patterns were very similar, regardless of the corresponding Y pattern (for next Sunday), there were cases in which the k Y patterns were almost random and did not reflect the characteristic Y pattern for our day. In other terms, a given X pattern had very different possible Y patterns linked to it.

Other difficulties arose when dealing with anomalous days. The most common anomalous days were holidays. We had an issue with the forecasting procedure, when the day we wished to forecast, the following day, or the preceding day was a holiday. Moreover, even if the prediction day was not a

Table 3: 75% quantiles for the goodness-of-fit parameters at the three sites over the entire period.

SITE	GOF		
	FOB (ratio)	MAPE (%)	RMSE (l/min)
Industrial area	0.20	39	7.86
High-density population area	0.21	50	1.80
Low-density population area	0.25	41	5.82

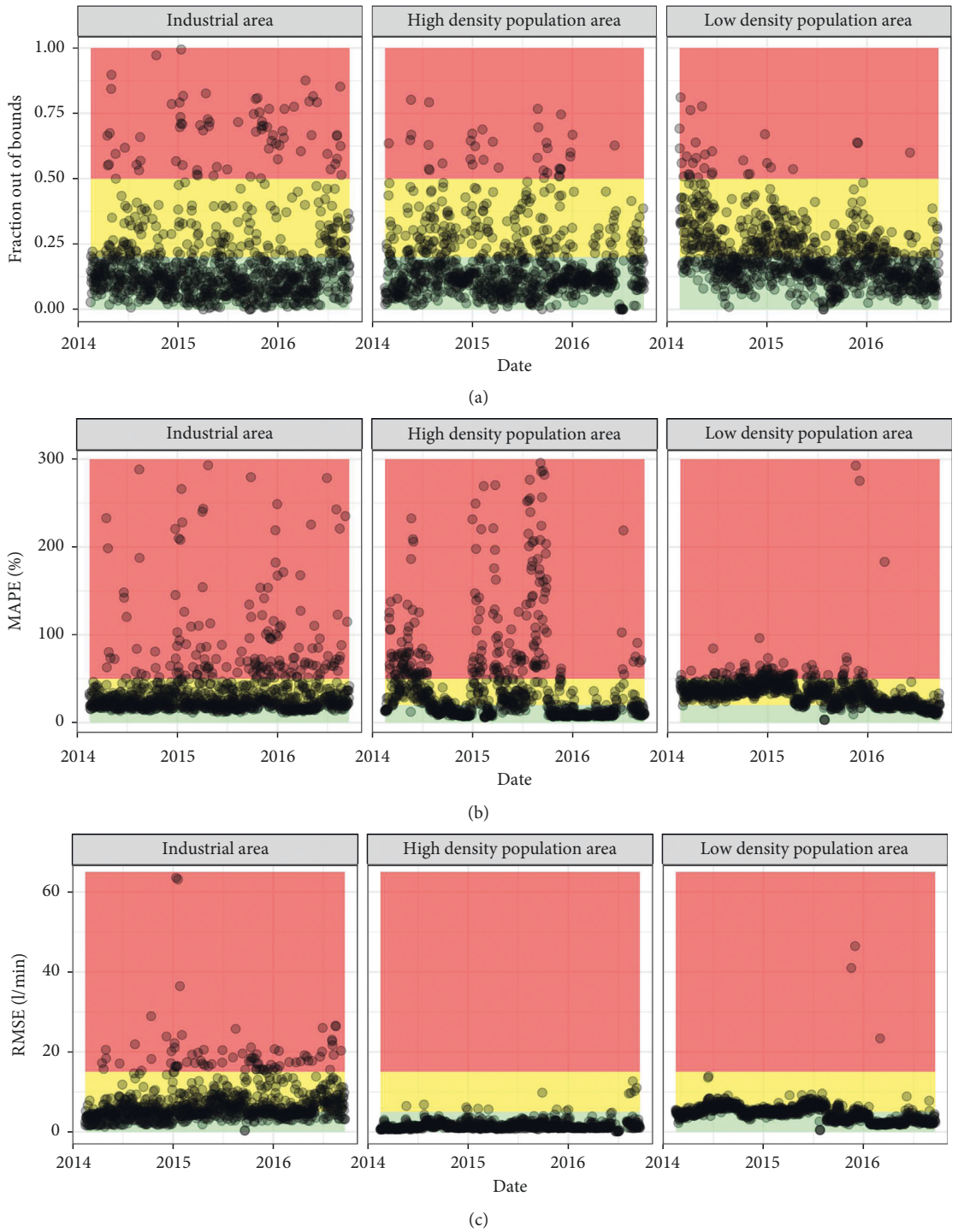

Figure 5: Daily values of the (a) FOB (fraction out of bounds), (b) MAPE (mean average percentage error), and (c) RMSE (root mean square error) for the industrial area, the high-density population area, and the low-density population area sites. The values are gathered into three different categories: good (green shading) for values lower than 20%, regular (green shading) for values between 20 and 50%, and bad (red shading) for values greater than 50%.

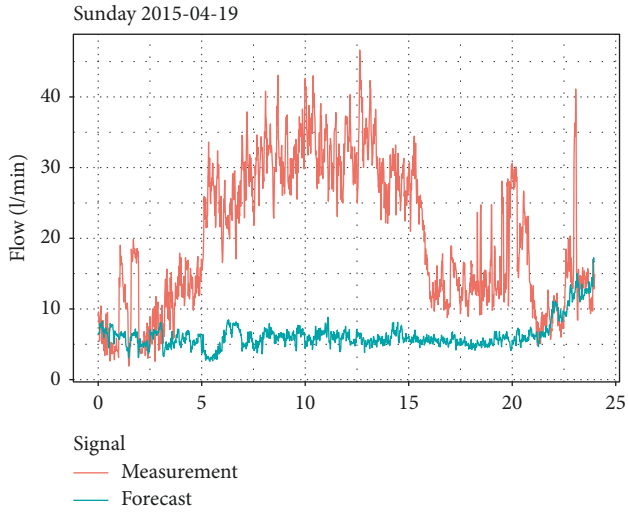

Figure 6: A wrongly forecasted signal. In this case, the forecasted day was on Sunday, 19 April 2015, at the industrial area site. For this site, Saturdays are very similar and Sundays differ depending on other factors (e.g., the time of the year). This led to an improper forecasting since the (k) Saturdays closest to 18 April 2015 do not need to be followed by Sundays similar to the forecasted day.

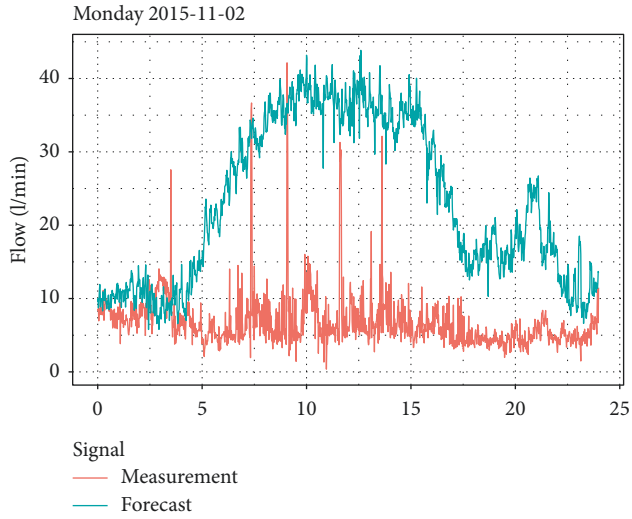

Figure 7: Forecasting at the industrial area on Monday, 2 November 2015. This day in particular was a holiday. In this case, since the day before was a Sunday, the predicted values were the average of (k) normal Mondays, which, obviously had a very different behaviour.

holiday (nor the preceding or following day), but one of the k nearest neighbors of the X pattern for the query day was followed by an anomalous day, the results were distorted by it (Figure 7).

Finally, the better the measurements are, the more accurate the forecasts will be. If a particular day has gaps in the measurements, although it is feasible to cope with the NAs (days with nonavailable data), they will introduce errors and mispredictions into the algorithm (Figure 8). Days with missing data should be flagged as anomalous and not considered in the forecasting procedure.

The high-density population area is the site where predictions were most accurate. This is partly because in this site which is an urban neighborhood in the center of the city, there are a large number of inhabitants living in residential buildings. This large number of people using the water supply at once regularizes the water flow time series. This is seen in Figure 2 (middle plot). Two patterns (labor days and weekends) can be observed, and they even look similar. Since the signal is so regular, the statistical forecasting procedure is more reliable, and thus, the GoF parameters showed very good results (with the exception of the MAPE which, as we stated above, was not considered due to its flaws when near-zero values are measured).

In the high-density population area, one of the causes of errors in the forecasting was misbehavior of the time series, probably because of malfunctions in the measurement device. For example, on 24 July 2014, there was, at around 10 a.m., a jump in the signal. After the jump, the time series continued to follow the normal pattern but maintained the offset (Figure 9).

Another type of misbehavior seen in the time series was an increase in the random fluctuations of the signal. For example, on 18 November 2015, a high-frequency random

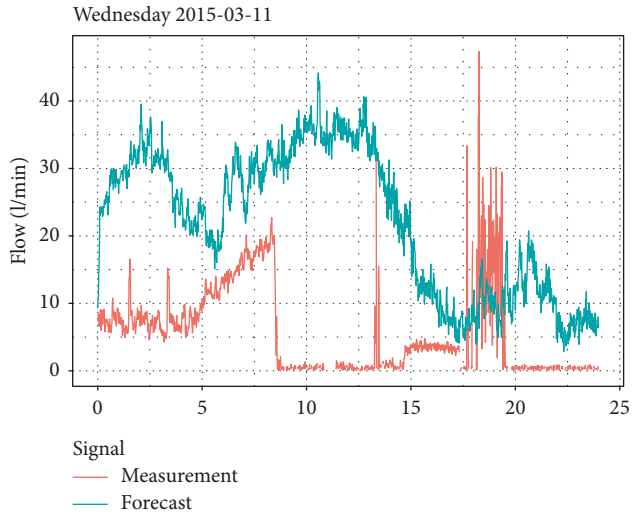

Figure 8: Forecasting at the industrial area on Wednesday, 11 March 2015. The effect of missing data in both the prediction and query days is visible in this figure. When there are missing data in the query day, the neighbors are improperly obtained since only the remaining data are considered, and thus, one might obtain a misleading neighbor which would yield erroneous forecasting results. When the data are missing in the forecasting day, there cannot accurate predictions.

component appeared and overlapped the signal (Figure 10). This random noise made the FOB value increase to 0.51 (red flag). Nonetheless, the original signal was still well predicted since both the forecasted and the measured values followed the same pattern, and this is why the RMSE value stayed small (RMSE = 2.8 l/min) although the FOB was high.

The low-density population area site was the most difficult to forecast. The reason for this was because this site is a

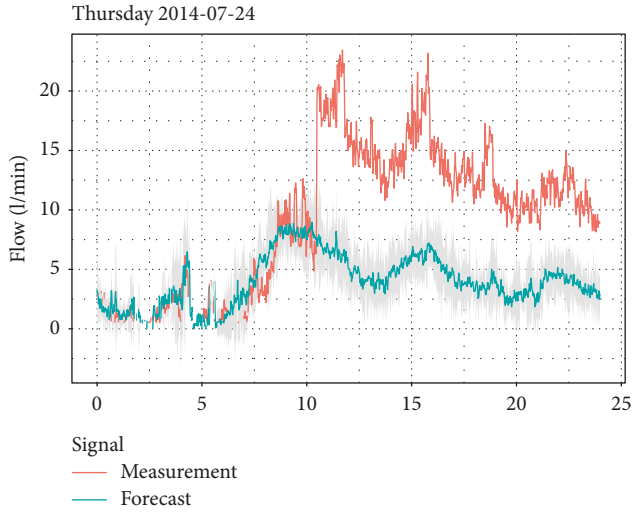

Figure 9: Forecasting at the high-density population area on Thursday, 24 July 2014. The water flow jumped at 10:28 a.m. from 9.39 l/min to 19.03 l/min. This offset of around 10 l/min was maintained after the jump. Therefore, although the RMSE was not very high (RMSE = 6.81 l/min), the FOB put that day in red (FOB = 0.63). The pointwise 95% confidence band of the estimation is depicted in grey.

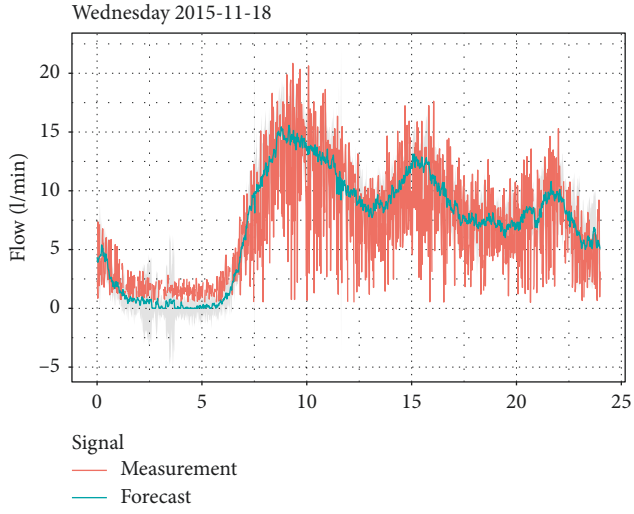

Figure 10: Forecasting at the high-density population area on Wednesday, 18 November 2015. The water flow measurements show a very large random component that increases the variance by a great amount. However, the general pattern of the series was well predicted by the forecast. Therefore, the RMSE remained low (RMSE = 2.8 l/min), even when the FOB was high (FOB = 0.51).

residential suburb of the city, where there are lots of single-family homes with gardens. Moreover, the zone is not very big, so the number of end users is very low compared to that of the high-density population area (hundreds of end users vs. thousands of end users). Therefore, the weight of each domestic user is very high and so is the variance in the signals. This led to a highly irregular time series. For example, Figure 11 shows the same week of July (8th to 14th) for 2014, 2015, and 2016. No regular pattern can be easily foreseen.

In this site, the overall results were fairly good. However, the main characteristic of these signals, the sudden peaks in the water flow, remained largely unpredictable since their distribution is, to a large extent, random.

For example, Figure 12 shows the forecast for two days. The first one corresponds to 12 March 2014 and was flagged as "green" in the FOP plot (Figure 5), while the second, which is for 13 March 2014, was flagged as "red" in the same FOP plot. It looks like the forecasts were more or less equally good, but in the first case, the peaks happened to occur inside the confidence bands, while in the second case, they fell outside those bands. This seems to be the reason why there were differences in the FOB for almost the same type of prediction even though the RMSE was, in both cases, very small (RMSE = 4.21 l/min, RMSE = 4.43 l/min).

As we have seen, an advantage of the pattern-similarity algorithm is that there is no need to estimate the annual seasonality since the procedure of normalizing the signal (obtaining the X–Y patterns) deseasonalizes the time series.

The pattern-similarity algorithm is easy to implement, and it is not very time-consuming. The part of the algorithm that takes the most time is the filtering of the training data, which are all days in the historical record of the same day of the week as the query day, such that neither them nor the next day are holidays (or anomalous days). The filtering can be accelerated by using database processing language tools (such as SQL).

In addition, if we have a large enough historical record with good data quality at our disposal, the patterns obtained as the forecast of the water flow on a given day are a good prediction of what that day should be like (Figure 13).

Even when the day to be forecasted shows minor data issues (small segments of missing data), the method is robust enough to deal with them. This keeps the algorithm from making false alarms (Figure 14).

Finally, as the historical record increases, the performance of the algorithm will improve. The method is based on obtaining statistical knowledge from previous data in order to determine the most similar situation to the one ahead, from the past data. Therefore, as the historical database gets larger, the forecast will get more accurate.

4.4. Comparison with Previous Work. In this section, we compare our approach (STPS, short-term pattern similarity) with another two similar ones: $\alpha\beta$-WDF ($\alpha\beta$ water demand forecast that was recently published by Pacchin et al. [32]) and GRNN (generalized regression neural network that was published by Dudek [11]).

The $\alpha\beta$-WDF approach is based on a moving window in which average parameters are obtained for similar days (same day of the week) from one, two, and three weeks earlier (a moving window of 3 weeks). $\alpha\beta$-WDF and STPS work in a very similar way; they both obtain average patterns for the same day of the week as the query day. However, since $\alpha\beta$-WDF takes into account three weeks prior to the time at which the forecast is made, STPS selects the k nearest

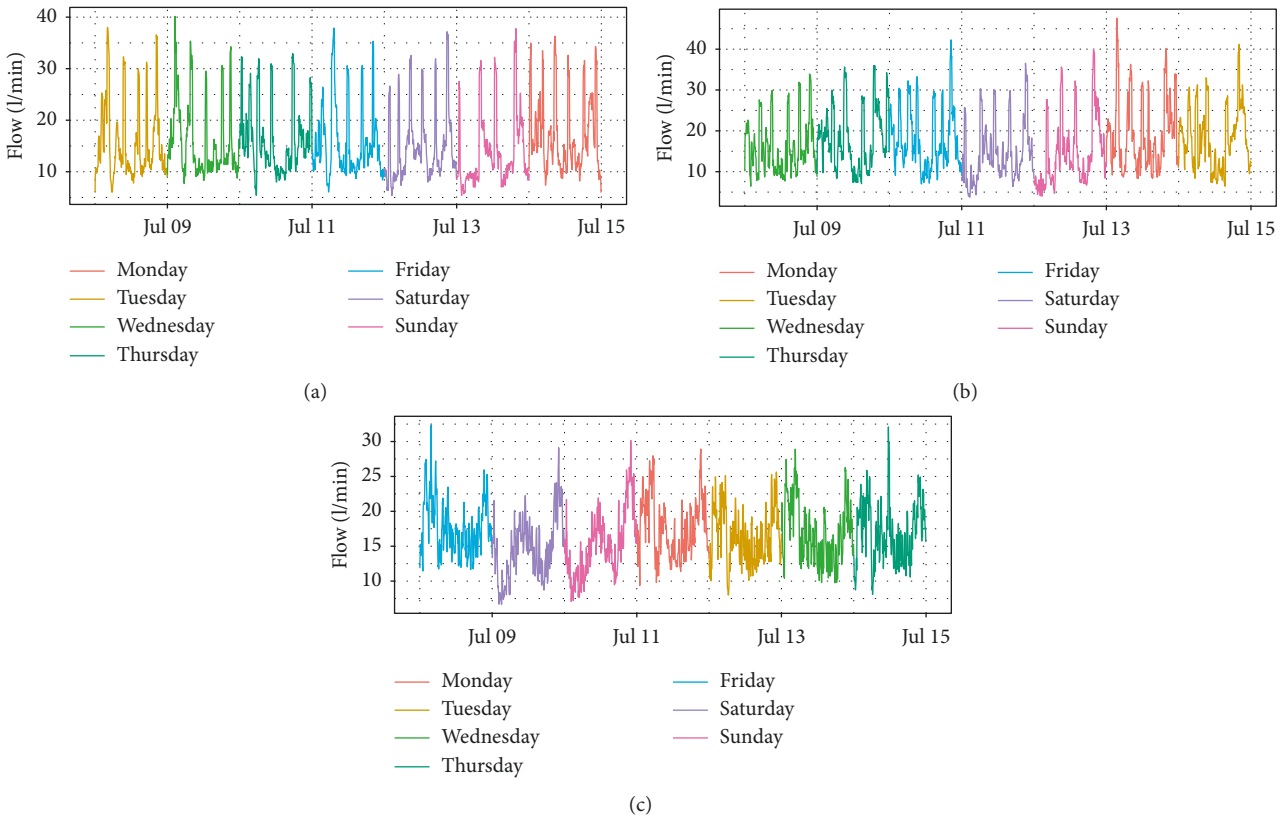

Figure 11: Comparison of the same week of July over the three sampled years (a) 2014, (b) 2015, and (c) 2016 at the low-density population area site.

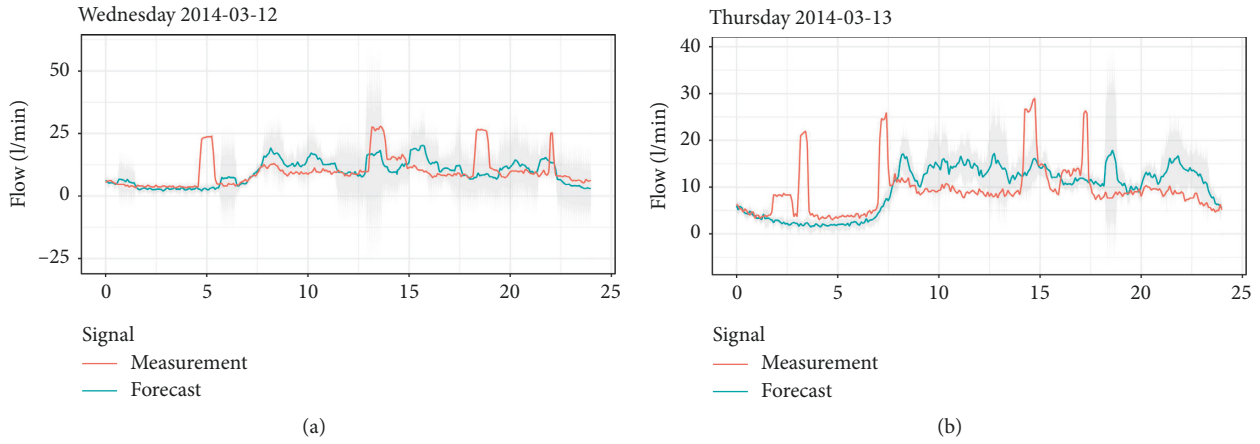

Figure 12: Comparison of two forecasts at the low-density population area site. The first one (a) was flagged as "green" in the FOB plot (FOB = 0.18), while the second (b) was flagged as "red" (FOB = 0.58). The pointwise 95% confidence band of the estimation is depicted in grey.

neighbors of the query pattern from the data in the history record.

As Dudek states [35], GRNN is a method equivalent to the STPS in terms of data preprocessing. It is also based on the X-Y pattern similarity that eliminates the seasonal components and also considers the nearest k neighbors. However, the prediction is made using a neural network of a single neuron in the intermediate layer using radial basis functions as weights. In this way, the result of the prediction is also an average of certain Y patterns, but with weights given by a Gaussian kernel (for further details, please refer [35]).

Other recently developed well-known approaches are based on ANNs, e.g., feedforward neural networks or long

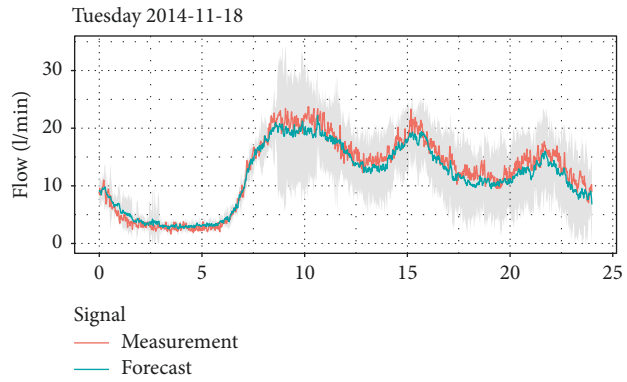

Figure 13: Example of a good forecast at the high-density population area site. The pointwise 95% confidence band of the estimation is depicted in grey.

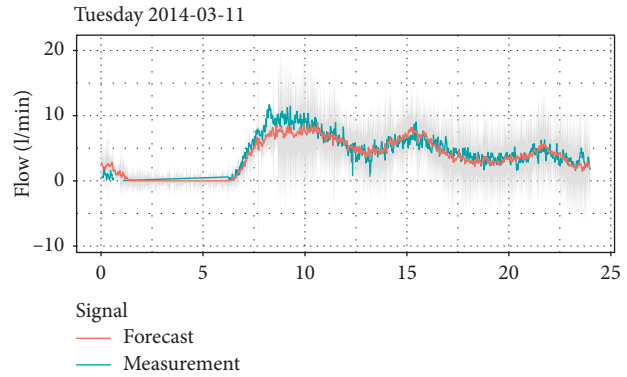

Figure 14: Example of a good forecast with missing data at the high-density population area site. There were missing/wrong data from approximately 1:00 a.m. until 6:00 a.m. Nevertheless, the overall outcome of the forecast was very accurate. The pointwise 95% confidence band of the estimation is depicted in grey.

short-term memory networks. Unfortunately, we could not apply these approaches to our study. From an engineering perspective, a 24 h forecast is commonly considered to be a short-term forecast; however, for data series with a high frequency, the number of predicted values is huge (1440 values). In our study, a method based on ANNs would require an output layer with 1440 nodes as well as an elevated number of nodes within both the hidden and input layers; therefore, the required data to train the ANN would be huge.

In the comparison study, we considered data from the year 2015 (in which the historical data is complete), and we computed the two error parameters proposed by Pacchin et al. [32]: MAPE and RMSE. Note that, since the number of data predicted for each day was high (1440 values) and because during an important fraction of the day, the water consumption values were very low (close to zero), and the value of the MAPE could become very high; therefore, for these special cases, we considered the RMSE to present a more adequate value to determine the goodness of the prediction.

In Figure 15, we present the measured values and the predicted values of the three approaches for a randomly-

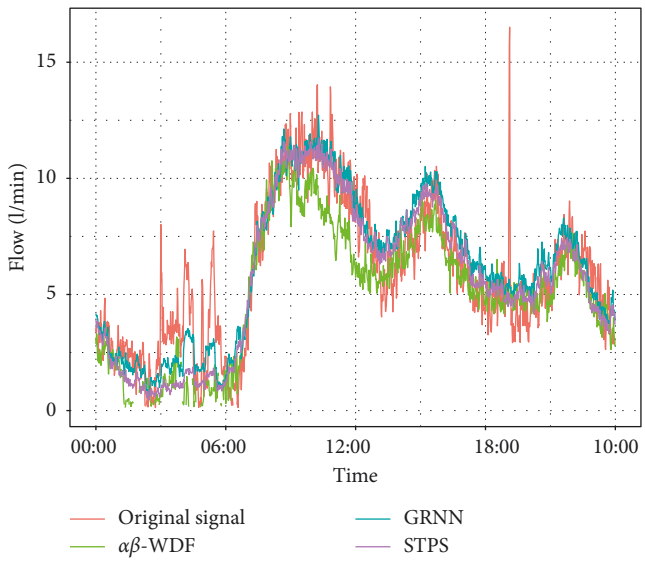

Figure 15: Forecasting at the high-density population area for Wednesday, 1 June 2015. The red line denotes the measured values, the green line denotes the predicted values by the $\alpha\beta$-WDF approach, the blue line denotes the prediction of GRNN, and the violet line denotes the predicted values by our approach (STPS).

picked day: 1 June 2015 at the high-density population area. Figure 16 illustrates the goodness of the methods under study in the three scenarios (industrial area, high-density population area, and low-density population area). It can be observed that our approach (STPS) and GRNN present better results than $\alpha\beta$-WDF for the three scenarios.

5. Conclusions and Future Works

This paper has presented an approach based on pattern-similarity techniques to forecast water demand. This work faces two important challenges that have been traditionally neglected in previous approaches, namely, a high frequency of predictions (based on measurements in terms of minutes) and the need for external data such as annual seasonality or weather that increments the complexity of the approaches. In that sense, on the one hand, the approach presented here is based on 1 min steps predictions, and, on the other hand, it does not require estimating annual seasonality since it determines this seasonality by constructing the X and Y patterns in which the series has been normalized.

In order to validate the approach, the study was applied over three different sites of a city in northern Spain. The results obtained provided interesting insights, such as the best predictions obtained in high-density population areas, the difficulties for identifying patterns for Sundays in industrial areas, or the higher random behaviour in low-density areas.

Additionally, our pattern-similarity approach (STPS) was also compared to other similar techniques that have been previously used for water forecasting, i.e., $\alpha\beta$-WDF and GRNN. The results obtained evidenced that $\alpha\beta$-WDF was the approach with worst results whilst GRNN and STPS behave similarly. This similar behaviour is normal since, in both cases, the estimation is obtained by an average

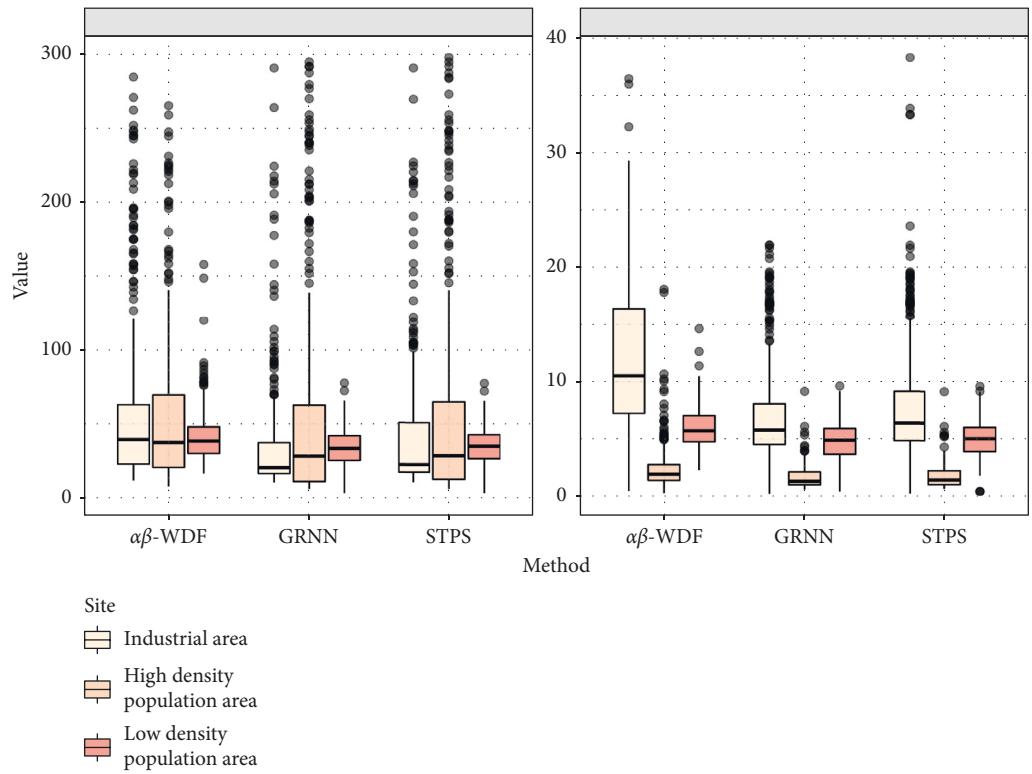

Figure 16: Goodness comparison among the $\alpha\beta$-WDF, GRNN, and STPS approaches at the three measured sites: (a) MAPE (%); (b) RMSE (l/min).

of some past values. However, while GRNN uses a weighted average where the weights are obtained by the Gaussian radial basis, in STPS, we use a simple average what is easier to compute. Thus, the main difference relies on the fact that, in GRNN, the abovementioned average is computed from some fixed previous days (one, two, or even three weeks before), while the STPS averages the last k nearest neighbors. Therefore, in cases in which the past weeks were, by any chance, nontypical (e.g., Christmas or Easter week), our method is providing better results due to its higher flexibility because in such cases it will look for similar days in the whole recorded history, whereas GRNN will be using only the past few weeks.

As future work, we intend to handle some weaknesses identified in the current method. Firstly, predictions success is reduced when anomalous days are considered. Anomalous days refer to two different situations: holidays and days with a behaviour different from what is considered usual. The former may be solved by constructing training sets from which to obtain the nearest neighbors since historical record contains enough data. To tackle the complexity of the latter, once these types of anomalous days have been identified, they could be just removed from the possible training subsets in the forecasting of other days.

Secondly, in order to improve the results for data sites where apparently there is not regularity, such as the low-density population area in our study, shorter prediction horizons could be considered, e.g., 4–6 hours. However, this is an issue that remains currently untested.

Finally, another interesting line of further work is the application of the proposed approach for water distribution in different cities.

Data Availability

The water consumption data of the Spanish city of Burgos, used to support the findings of this study, have not been made available because restrictions apply to the availability of these data, which were used under license for the current study and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the company Acciona Agua concretely through its subsidiary Aguas de Burgos.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors wish to acknowledge the collaborative funding support from (i) Ministerio de Ciencia, Innovación y Universidades (MCIU), Agencia Estatal de Investigación (AEI), and European Regional Development Fund (ERDF) Project (RTI2018-098652-B-I00); (ii) POCTEP 4IE Project (0045-4IE-4-P); and (iii) Consejería de Economía e Infraestructuras/Junta de Extremadura (Spain), European Regional Development Fund (ERDF) Projects (IB16055 and GR18112).

References

[1] R. Espinosa, L. Garriga, J. J. Zubcoff, and J. N. Mazón, "Linked open data mining for democratization of big data," in *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pp. 17–19, Los Alamitos, CA, USA, 2014.

[2] A. Taivalsaari and T. Mikkonen, "A taxonomy of IoT client architectures," *IEEE Software*, vol. 35, no. 3, pp. 83–88, 2018.

[3] A. Degbelo, C. Granell, S. Trilles, D. Bhattacharya, S. Casteleyn, and C. Kray, "Opening up smart cities: citizen-centric challenges and opportunities from GIScience," *ISPRS International Journal of Geo-Information*, vol. 5, no. 2, p. 16, 2016.

[4] A. Di Nardo, M. Di Natale, G. F. Santonastaso, and S. Venticinque, "An automated tool for smart water network partitioning," *Water Resources Management*, vol. 27, no. 13, pp. 4493–4508, 2013.

[5] W-SMART Association, *Background and FP7 Goals*, W-SMART Association, Paris, France, 2018.

[6] Alliance for Water Efficiency, *Water Loss Control—Efficiency in the Water Utility Sector*, Alliance for Water Efficiency, Chicago, IL, USA, 2018, http://www.allianceforwaterefficiency.org/Water_Loss_Control_Introduction.aspx.

[7] A. D. Nardo, M. D. Natale, R. Gargano, C. Giudicianni, R. Greco, and G. F. Santonastaso, "Performance of partitioned water distribution networks under spatial-temporal variability of water demand," *Environmental Modelling & Software*, vol. 101, pp. 128–136, 2018.

[8] Center for Neighborhood Technology (CNT), *The Case for Fixing the Leaks: Protecting People and Saving Water while Supporting Economic Growth in the Great Lakes Region*, Center for Neighborhood Technology (CNT), Chicago, Illinois, USA, 2013.

[9] G. Dudek, "Pattern similarity-based methods for short-term load forecasting—part 1: Principles," *Applied Soft Computing*, vol. 37, pp. 277–287, 2015.

[10] G. Dudek, "Pattern similarity-based methods for short-term load forecasting—part 2: Models," *Applied Soft Computing*, vol. 36, pp. 422–441, 2015.

[11] G. Dudek, "Pattern-based local linear regression models for short-term load forecasting," *Electric Power Systems Research*, vol. 130, pp. 139–147, 2016.

[12] S. Meliá, J. Gómez, S. Pérez, and O. Díaz, "A model-driven development for GWT-based rich internet applications with OOH4RIA," in *Proceedings of the Eighth International Conference on Web Engineering, ICWE*, vol. 14–18, pp. 13–23, Yorktown Heights, New York, USA, July 2008.

[13] Y. Martínez, C. Cachero, and S. Meliá, "MDD vs. traditional software development: a practitioner's subjective perspective," *Information and Software Technology*, vol. 55, no. 2, pp. 189–200, 2013.

[14] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, Melbourne, Australia, 2nd edition, 2018.

[15] L. Shvartser, U. Shamir, and M. Feldman, "Forecasting hourly water demands by pattern recognition approach," *Journal of Water Resources Planning and Management*, vol. 119, no. 6, pp. 611–627, 1993.

[16] S. G. Buchberger and L. Wu, "Model for instantaneous residential water demands," *Journal of Hydraulic Engineering*, vol. 121, no. 3, pp. 232–246, 1995.

[17] S. G. Buchberger and G. J. Wells, "Intensity, duration, and frequency of residential water demands," *Journal of Water Resources Planning and Management*, vol. 122, no. 1, pp. 11–19, 1996.

[18] E. A. Donkor, T. A. Mazzuchi, R. Soyer, and J. Alan Roberson, "Urban water demand forecasting: review of methods and models," *Journal of Water Resources Planning and Management*, vol. 140, no. 2, pp. 146–159, 2014.

[19] M. Sebri, "Forecasting urban water demand: a meta-regression analysis," *Journal of Environmental Management*, vol. 183, pp. 777–785, 2016.

[20] J. Adamowski, H. Fung Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva, "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*, vol. 48, no. 1, 2012.

[21] M. Herrera, L. Torgo, J. Izquierdo, and R. Pérez-García, "Predictive models for forecasting hourly urban water demand," *Journal of Hydrology*, vol. 387, no. 1-2, pp. 141–150, 2010.

[22] F. K. Odan and L. F. R. Reis, "Hybrid water demand forecasting model associating artificial neural network with fourier series," *Journal of Water Resources Planning and Management*, vol. 138, no. 3, pp. 245–256, 2012.

[23] G. Ji, J. Wang, Y. Ge, and H. Liu, "Urban water demand forecasting by LS-SVM with tuning based on elitist teaching-learning-based optimization," in *Proceedings of the Control and Decision Conference (2014 CCDC), the 26th Chinese*, pp. 3997–4002, IEEE, Changsha, China, May 2014.

[24] C. J. Hutton and Z. Kapelan, "A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting," *Environmental Modelling & Software*, vol. 66, pp. 87–97, 2015.

[25] A. Candelieri and F. Archetti, "Identifying typical urban water demand patterns for a reliable short-term forecasting—the icewater project approach," *Procedia Engineering*, vol. 89, pp. 1004–1012, 2014.

[26] A. Candelieri, D. Soldi, and F. Archetti, "Short-term forecasting of hourly water consumption by using automatic metering readers data," *Procedia Engineering*, vol. 119, pp. 844–853, 2015.

[27] A. Candelieri, "Clustering and support vector regression for water demand forecasting and anomaly detection," *Water*, vol. 9, no. 3, p. 224, 2017.

[28] S. Alvisi and M. Franchini, "Assessment of predictive uncertainty within the framework of water demand forecasting using the model conditional processor (MCP)," *Urban Water Journal*, vol. 14, no. 1, pp. 1–10, 2017.

[29] B. M. Brentan, E. Luvizotto Jr., M. Herrera, J. Izquierdo, and R. Pérez-García, "Hybrid regression model for near real-time urban water demand forecasting," *Journal of Computational and Applied Mathematics*, vol. 309, pp. 532–541, 2017.

[30] M. Romano and Z. Kapelan, "Adaptive water demand forecasting for near real-time management of smart water distribution systems," *Environmental Modelling & Software*, vol. 60, pp. 265–276, 2014.

[31] F. Gagliardi, S. Alvisi, Z. Kapelan, and M. Franchini, "A probabilistic short-term water demand forecasting model based on the Markov chain," *Water*, vol. 9, no. 7, p. 507, 2017.

[32] E. Pacchin, S. Alvisi, and M. Franchini, "A short-term water demand forecasting model using a moving window on previously observed data," *Water*, vol. 9, no. 3, p. 172, 2017.

[33] E. Arandia, A. Ba, B. Eck, and S. McKenna, "Tailoring seasonal time series models to forecast short-term water demand," *Journal of Water Resources Planning and Management*, vol. 142, no. 3, article 04015067, 2016.

[34] M. Bakker, J. H. G. Vreeburg, K. M. van Schagen, and L. C. Rietveld, "A fully adaptive forecasting model for short-term drinking water demand," *Environmental Modelling & Software*, vol. 48, pp. 141–151, 2013.

[35] G. Dudek, "Neural networks for pattern-based short-term load forecasting: a comparative study," *Neurocomputing*, vol. 205, pp. 64–74, 2016.