water

MDPI

Article

Water Flow Modeling and Forecast in a Water Branch of Mexico City through ARIMA and Transfer Function Models for Anomaly Detection

David Barrientos-Torres, Erick Axel Martinez-Ríos, Sergio A. Navarro-Tuch *, Jose Luis Pablos-Hach and Rogelio Bustamante-Bello

Tecnologico de Monterrey, School of Engineering and Sciences, Mexico City 14380, Mexico; a01331999@tec.mx (D.B.-T.); a01331212@tec.mx (E.A.M.-R.); jlpabloshach@yahoo.com (J.L.P.-H.); rbustama@tec.mx (R.B.-B.)

* Correspondence: snavarro.tuch@tec.mx

Abstract: Early identification of anomalies (such as leakages or sensor failures) in urban water distribution systems is critical to mitigating water scarcity in cities and is a challenge in water resource management. Several data-driven methods based on machine learning algorithms have been proposed in the literature for leakage detection in urban water distribution systems. Still, most of them are challenging to implement due to their complexity and requirements of vast amounts of reliable data for proper model generation. In addition, the required infrastructure and instrumentation to collect the data needed to train the models could be unaffordable. This paper presents the use and comparison of Autoregressive Integrated Moving Average models and Transfer Function models generated via the Box–Jenkins approach to modeling the water flow in water distribution systems for anomaly detection. The models were fit using water flow data from tanks operating in a branch of the water distribution system of Mexico City. The results showed that both methods helped select the best model type for each variable in the analyzed water branch, with Seasonal ARIMA models achieving a lower mean absolute percentage error than the fitted Transfer Function models. Furthermore, this methodology can be adjusted to different time windows to generate alerts at different rates and does not require a large sample size. The generated anomaly detection models could improve the efficiency of the water distribution system by detecting anomalies such as wrong measurements and water leakages.

Keywords: anomaly detection; ARIMA; transfer function model; urban water branch; leakage detection

Citation: Barrientos-Torres, D.; Martinez-Ríos, E.A.; Navarro-Tuch, S.A.; Pablos-Hach, J.L.; Bustamante-Bello, R. Water Flow Modeling and Forecast in a Water Branch of Mexico City through ARIMA and Transfer Function Models for Anomaly Detection. *Water* **2023**, *15*, 2792. https://doi.org/10.3390/w15152792

Academic Editors: Kunlun Xin and Hexiang Yan

Received: 8 July 2023
Revised: 26 July 2023
Accepted: 28 July 2023
Published: 2 August 2023

1. Introduction

The growing worldwide population, the increasing living standards, the altered water consumption habits, and the spread of irrigated agriculture are the primary causes of the rising global demand for water. Water scarcity has become a danger to the sustainable development of human society [1]. According to the World Urbanization Prospects published by the United Nations in 2018, almost 90% of Mexico's population is projected to reside in urban areas [2]. Nevertheless, 20 million people in Mexico suffer from severe water scarcity [1]. Even though there is sufficient infrastructure in Mexico, water management could be improved, and the system needs to be appropriately maintained. According to the estimates, the distribution networks lose 40% of their water due to aging pipelines, lengthy periods without sufficient maintenance, poor building and management techniques, and ongoing land subsidence in metropolitan areas [3].

Managing water resources and preventing, identifying, and fixing leaks are essential to reduce city water scarcity. However, this requires an information system. Therefore, real-time monitoring and data collection are crucial to creating trustworthy and practical information systems. Additionally, the data must be accurate and thorough to draw

reliable conclusions and models [4]. Nevertheless, data collection is hampered by several structural and physical restrictions, environmental factors, and human errors. Moreover, the complexity and breadth of the modern city's water system necessitate a significant infrastructure investment for communications, location, and data processing [5]. The above has motivated the development of solutions that help to monitor water distribution networks. Several studies and technologies have been developed for leakage detection in water distribution systems, which can be classified into hardware- and software-based methods. Acoustic monitoring, gas injection, thermography, ground-penetrating radar, and free swimming systems are examples of hardware-based techniques. However, employing these techniques in broad regions can be time-consuming, expensive, and inappropriate for automation or long-term monitoring [6].

Software-based leakage detection techniques can be categorized into model-based and data-driven approaches. Model-based approaches define the link between the variables of the water distribution network in a mathematical model of the water distribution network while considering the network's physical properties [7]. Model-based leak identification techniques do not require previous network data; instead, the leak diagnosis is performed by comparing the model outputs to the measured variables. Its development, however, could be challenging, confined by the accuracy of the mathematical models, and dependent on accurate model calibration [8].

On the other hand, data-driven approaches create data analysis plans using the network's historical data as a resource. In contrast to model-based approaches, data-driven methodologies need to know the network's structural characteristics and historical data. Recently, there has been an interest in employing machine-learning methods because of the robust capacities for pattern recognition and feature identification and the rising development and accessibility of data-collecting technology [8]. For instance, multilayer perceptrons, support vector machines, clustering algorithms, or deep learning algorithms have been proven efficient for solving leak localization problems, as discussed in [6,8].

In this regard, water leakage detection based on machine learning algorithms has been proposed to use different data modalities to train the detection models. These data modalities include flow sensor data, pressure data, vibration data, vibro-acoustic data, acoustic emission data, and satellite data [9–14]. However, satellite and acoustic emission data collection may be unaffordable [15]. Besides, using flow sensors and vibration data for water leakage detection requires the installation of several sensors across the water pipeline or its junctions, which limits their use on large water distribution networks. In addition, some studies that have developed water leakage detection systems have used simulation or laboratory tests under controlled conditions without considering the uncertainty to which data may be susceptible in real scenarios [16–19].

Moreover, the key challenge with applying machine learning techniques is choosing the suitable algorithm, building appropriate feature extractors to learn complicated features, accessing a large amount of data for training the models, and needing efficient signal processing tools [6]. In addition, the black-box nature of deep learning algorithms makes them less interpretable by humans and necessitates specialized computer hardware for their training (e.g., Graphics Processing Units) [20].

Most of the studies that have proposed water leakage detection systems based on data-driven methods either use machine learning techniques (i.e., random forest, support vector machines, Adaboost, XGBoost) or deep learning algorithms (i.e., convolutional neural networks (CNNs)), which often requires high-quality data to train them. However, gathering enough data could be expensive, time-consuming, and unrealistic. Furthermore, black-box classification techniques such as CNNs, random forests, multilayer perceptrons, or XGBoost have many parameters to be adjusted, which limits their interpretability and makes them prone to overfitting [21]. On the other hand, one of the crucial characteristics of Autoregressive Integrated Moving Average (ARIMA) and Transfer Function (TF) models is that they provide a linear estimate of the system to be modeled. Besides, the number of parameters of ARIMA and TF models tends to be lower [22], which is a strong simplification

compared to the large number of parameters that non-linear machine learning and deep learning approaches often require [23].

This study presents a methodology for anomaly detection in water distribution systems by employing water flow data and two classical time series modeling techniques, the ARIMA and TF models, which were fit following the Box–Jenkins methodology [24]. This study modeled water flow data from tanks in a primary network branch of the water distribution system in Mexico City. This branch carries a significant volume of water through tanks and supplies the secondary network. Analyzing this branch is crucial due to the substantial water flows and pipeline sizes involved. A leakage occurring in this system would result in more-significant water losses than other public water network systems of Mexico City.

As previously stated, the studies in the literature have performed simulations, laboratory tests, or placed sensors (e.g., flow and vibrations sensors) across the water pipelines to collect the necessary data to develop water-leakage-detection models [10,25–28]. On the contrary, this study focused on analyzing a branch of the water distribution system of Mexico City, which supplies water to the water pipes, instead of analyzing the water pipes directly through flow or vibration sensors. The above was performed to detect anomalies in the water flow behavior that could indicate the presence of sensor malfunction or water leakages along the analyzed water branch. Thus, the data on the inlet and outlet water flow of the tanks that comprise the analyzed water branch were used to develop the ARIMA and TF models proposed in this work. Such a study has not yet been performed to the authors' knowledge.

The principal contribution of this work was the use and comparison of the TF and ARIMA models generated through the Box–Jenkins methodology applied for anomaly detection in water flow variables of a water branch of the Mexico City water distribution system, which allowed us to:

- Adjust the models and forecasts to different time windows of the water flow consumption in Mexico City.
- Generate anomaly-detection models with incomplete and small datasets by employing the water flow data of a branch of the water distribution system of Mexico City.
- Generate interpretable and sparse models of water flow for anomaly detection in a branch of the water distribution system of Mexico City.
- Perform a comparison of the ARIMA and TF models for modeling the water flow behavior of a branch of the water distribution system of Mexico City.

The rest of this paper is structured as follows. Section 2 presents the literature review on water leakage detection based on machine learning algorithms and an analysis of the state-of-the-art. Section 3 presents the case study and describes the data collection process of the flow data of the water distribution branch analyzed in this work. Moreover, the overall methodology is explained, including the theoretical background of the ARIMA and TF models and the model-generating process. In addition, the proposed anomaly-detection methodology for the analyzed water branch, which integrates the best models of both methods, is explained. Section 4 presents the results of the present study, while Section 5 presents the analysis and discussion of the results. Section 6 shows the main limitations and areas of opportunity of this work. Finally, Section 7 presents the conclusions and future work.

2. Literature Review

This section presents an overview of the studies that have proposed methods for water leakage detection based on machine learning. In addition, an analysis and discussion of the gaps in the current state-of-the-art is shown. Table 1 summarizes the literature on recent approaches towards leakage detection in water distribution systems that used machine learning techniques.

Recent studies have suggested using data-driven methodologies to detect water leakage, primarily relying on machine learning algorithms. Islam et al. [29] presented and

discussed this trend of using machine learning for water leakage detection. For instance, the study of Moulik et al. [17] proposed to detect water leakages and blockages in water pipelines by processing the vibrations of PCV pipes. Moulik's study employed three-axis accelerometers to measure the vibration on the PCV pipes produced by water leakages; then, the vibration data were utilized as the input into a k-means clustering technique to perform the detection. Similarly, Choi et al. [30] utilized sound vibration data from water pipes to detect water leakages by employing the magnitude spectra of the sound vibration data to train a 2D CNN. Likewise, Yu et al. [10] employed vibration data collected from piezoelectric accelerometers placed in the water distribution networks of several cities in China for water leakage detection. Yu's study tested different machine learning algorithms such as support vector machines, decision trees, the SqueezeNet CNN, and K-nearest neighbor, with the SqueezeNet achieving a higher performance when trained with the spectrograms of the Short-Time Fourier Transform of the vibration data.

Fereidooni et al. [9] installed flow sensors in the pipeline network junction to detect water leakages. The flow sensor data were processed using hydraulic equations to generate velocity and head loss features. The trained algorithms were a decision tree, a K-nearest neighbor, a random forest, and a Bayesian network. Satellite data have also been used for water leakage detection. An example of this was presented by Chen et al. [11], who utilized augmented satellite images to detect water leakages in the canal systems in Arizona. The authors employed Landsat 8 satellite images to train a CNN, used as the water-leakage-detection algorithm. Sousa et al. [12] proposed to analyze pressure data measured from pumps in district-metered areas of Stockholm, Sweden. The analyzed area corresponded to a residential area with no water tanks or reservoirs. The detection algorithms involved a comparison of unsupervised learning algorithms, such as k-means clustering, and supervised learning algorithms, such as learning vector quantization algorithms. In [31], it was proposed to detect water leakages by processing acoustic emission signals collected from the water distribution networks of Jiangsu, Zhejiang, and Shanghai. The acoustic emission signals were characterized by computing the main frequency, the spectral roll-off rate, the spectral flatness, and the Mel frequency cepstrum coefficients. Then, the authors trained tree-based algorithms such as decision trees, Adaboost, and random forests, with Adaboost achieving the highest performance. Likewise, Fares et al. [13] utilized acoustic emission signals to detect water leakages in water distribution networks. Fares' study utilized time and frequency domain features to represent the acoustic emission signals and used them as the input to train a support vector machine, an artificial neural network, and deep learning algorithms.

Furthermore, Xue et al. [18] introduced a leakage-fault-detection approach using a hydraulic simulation model encompassing all potential leakage faults. Subsequently, XGBoost was trained, and an alert-triggering algorithm generated a leakage signal associated with the specific pipe's name. Cody and Narasimhan [32] proposed a linear prediction model, specifically an autoregressive moving average model in conjunction with a multivariate Gaussian mixture model to perform semi-supervised leakage detection. This method utilizes data collected with hydrophone sensors and simulated leakages within a water distribution network. Additionally, the authors suggested a coarse-resolution leakage location using the average baseline root mean square of the collected data and a fine location estimation utilizing cross-correlation based on the time series data from linear prediction filter sensors. Taghlabi et al. [19] conducted experiments employing two methods for water leakage detection. Firstly, they simulated artificial leaks using the EPANET code on the MATLAB platform to establish a database of pressure values that describe the network's behavior when leaks are present. Subsequently, these data were utilized to train a random forest algorithm, enabling it to forecast the rate and location of leaks within the network. Secondly, they simulated artificial leaks by manipulating hydrants in different locations, considering two distinct leak sizes, and comparing results.

Similarly, Pérez-Pérez et al. [16] proposed using artificial neural network (ANN) techniques and online measurements of pressure and flow rate to detect and locate water leaks

in pipelines. The friction factor of the pipe was estimated and utilized as an input for computing the leak position. Fabbiano et al. [33] considered that the energy variation transmitted to the pipe walls by the radial component of vibrations induced by fluid turbulence might be related to the flow leak. Hence, Fabbiano measured the radial vibrational status of specific pipes in the network. Finally, Tornyeviadzi et al. [34] proposed a one-dimensional CNN deep autoencoder trained to locate and identify water leaks. This technique uses multivariate time series data to lessen the adverse effects of random noise. The proposed autoencoder's input data involved flow, pressure, and tank-level data.

Table 1. Recent data-driven approaches for water-leakage-detection technologies.

Project	Year	Country	Methodology	Results
Leakage detection in water distribution systems based on time–frequency convolutional neural network [15]	2021	China	A leakage spectrogram was employed to capture the characteristics of leakage signals, and a time–frequency convolutional neural network (TFCNN) model was compared with other classification models across various signal-to-noise ratio (SNR) conditions.	The TFCNN model demonstrated superior performance with a mean accuracy of 98% across different SNR conditions. Even at a challenging −10 dB SNR, the mean detection accuracy remained high at 90%.
Water Leakage Detection in Hilly Region PVC Pipes using Wireless Sensors and Machine Learning [17]	2020	Taiwan	Wireless sensors were utilized to capture vibrations in PVC pipes during water flow. Machine learning algorithms were applied to these vibration records to identify any disruptions in the regular water flow caused by leakage or blockage.	Analysis of vibration records with the help of K-means algorithm to determine the water level and the leakages, if any.
Application of CNN Models to Detect and Classify Leakages in Water Pipelines Using Magnitude Spectra of Vibration Sound [30]	2023	Korea	CNN model for water leakage detection and classification using sound vibration data from sensors in water pipes.	The proposed CNN model achieved an F1-score of 94.82% and a Matthew's correlation coefficient of 94.47%.
Leak detection in water distribution systems by classifying vibration signals [10]	2023	China	Support vector machine (SVM), decision tree (DT), and K-nearest neighbor (KNN) for leak detection models using signal data from piezoelectric accelerometers in Chinese water distribution systems (WDSs).	SqueezeNet performed best with 95.15% accuracy in leak identification, while KNN excelled among the three classifiers with superior sensitivity and 88.17% accuracy.
A hybrid model-based method for leak detection in large scale water distribution networks [9]	2021	Netherlands	Influential leak detection features using hydraulic equations (Hazen–Williams, Darcy–Weisbach, and pressure drop) and decision tree, KNN, random forest, and Bayesian network used to locate leaks and determine their pressure based on pipeline topology.	Of the models, 80.5% consistently achieved results above 92% in all scenarios. The Naïve Bayesian Model performed the best overall, with a top result of 85.81%.
Augmenting a deep-learning algorithm with canal inspection knowledge for reliable water leak detection from multispectral satellite images [11]	2020	USA	A deep learning approach, combined with canal inspection knowledge, enabled automated and reliable water leak detection of canal sections using Landsat 8 satellite images.	The proposed approach can achieve recall at 86%, precision at 86%, and accuracy at 85%.

Table 1. *Cont.*

Project	Year	Country	Methodology	Results
Leakage detection in water distribution networks using machine-learning strategies [12]	2023	Sweden	Analyzed pressure measurements from pumps in district-metered areas (DMAs) using unsupervised learning (K-means and cluster validation techniques) and supervised learning (learning vector quantization algorithms).	The proposed learning strategies are able to obtain correct classification rates up to 93.98%.
A Tree-Based Machine Learning Method for Pipeline Leakage Detection [31]	2022	China	Distinctive features such as main frequency, spectral roll-off rate, spectral flatness, and 1D Mel frequency cepstrum coefficient (MFCC) using random forest and Adaboost models.	The Adaboost model had the lowest false positive rate of 7.35%. The recall rates of the random forest and Adaboost models were 100% and 99.52%.
Leak detection in real water distribution networks based on acoustic emission and machine learning [13]	2022	China	Acoustic signals in time and frequency domains were used to develop leak-detection models, employing SVM, ANN, and deep learning (DL) techniques.	Demonstrated a largely stable performance and a high accuracy, particularly for new unlabeled cases.
Machine learning-based leakage fault detection for district heating networks [18]	2020	China	Hydraulic simulation model and an XGBoost-based model	85.85% of mean accuracy.
Field implementation of linear prediction for leak-monitoring in water distribution networks [32]	2020	Canada	Linear prediction model for semi-supervised leak detection.	A detection accuracy in most cases of over 70%.
Prelocalization and leak detection in drinking water distribution networks using modeling-based algorithms [19]	2021	Morocco	A simulation of artificial leaks and a random forest machine learning algorithm	Leak position identified within a 100 m radius.
Leak diagnosis in pipelines using a combined artificial neural network approach [16]	2021	Mexico	ANN techniques and online measurements of pressure and flow rate measurements	An average error of 0.629% for leak location.
Smart water grid: A smart methodology to detect leaks in water distribution networks [33]	2020	Italy	Measuring the radial vibrational status of opportune pipes of the network	Radial vibration signals are linearly dependent only on the flow rate variations due to the leakages.
Leakage detection in water distribution networks via 1D CNN deep autoencoder for multivariate SCADA data [34]	2023	Norway	A one-dimensional convolutional neural network deep autoencoder (AE) using multivariate time series data	Identified 16 of the 19 leaks in 2019.

From this literature review, it is possible to observe that machine learning techniques have already been used extensively to perform water leakage detection. To a lesser extent, satellite data have been employed. Nevertheless, satellite data could be difficult to collect and label and may not be useful for detecting leakages inside the water pipelines. On the other hand, sound vibration data may be unaffordable due to the need for specific hardware to sample the vibro-acoustic signals. In the case of vibration data collected from accelerometers, it is necessary to install multiple sensors across the water pipelines, which can be costly and require extensive maintenance. Hence, analyzing the flow behavior of the water network can be considered a cost-effective solution since flow data are frequently monitored in water distribution systems. Nevertheless, similar to using accelerometers placed along the water pipeline to measure the pipe vibration, it is necessary to install multiple flow sensors along the water pipeline.

Moreover, in some of the reviewed works, the leakage-detection algorithm was developed in laboratory conditions, such as the studies of Pérez-Pérez et al. [16], Moulik et al. [17], and Taghlabi et al. [19]. Nevertheless, as mentioned by Shen et al. [31], on-site leakage signals have greater interference and randomness than leakage signals in a laboratory. Hence, there is an opportunity to analyze flow sensor data sampled from real water distribution systems and develop algorithms that can tackle the uncertainty to which

the data are susceptible when developing models for water leakage detection. Furthermore, most of the related works focused on detecting water leakages by directly measuring the vibration or flow sensor data from the water pipeline. Nevertheless, analyzing the sensor data along the complex water pipelines could be inefficient and costly.

In the case of the machine learning techniques that have been used to develop the leakage detection models, it can be appreciated that deep neural networks have been extensively used, mainly variants of CNNs [15,30]. Even so, CNNs required a sizable sample size to avoid overfitting and a lack of interpretability due to the complexity that this type of technique often requires. Other techniques frequently used in water leakage detection are non-linear classification techniques such as decision trees, random forests, support vector machines, Adaboost, and XGBoost [35]. However, these non-linear classification techniques, similar to CNNs, require a large sample size to avoid overfitting and are less interpretable than linear machine learning techniques [36].

Considering the above, there is an opportunity to develop techniques for detecting water leakage from other locations besides measuring water flow data directly from the water pipelines. Furthermore, the gathering of data and required flow sensors could be reduced if the water branch that delivers water to the water pipelines is analyzed, rather than directly measuring the flow or vibration in the water distribution pipelines. Finally, linear machine learning techniques such as the TF and ARIMA models could be tested to avoid using non-linear classification techniques, frequently employed in the literature, as presented in Table 1.

3. Materials and Methods

3.1. Case Study and Dataset

The case study examined in this work consisted of six tanks from the Mexico City water distribution system situated in the Álvaro Obregón delegation and connected by 48 in-diameter pipelines. Figure 1 shows a general schematic of the primary water distribution network in Mexico City, where the main two sources of water (Cutzamala System and Lerma System) feed several branches interconnected in cascade and fed by gravity. This study analyzed the data from Branch C of Figure 1. The branch presented in Figure 1 is instrumented to measure the water flow that is input and output to each water tank. The sensors used to sample the data were ISOMAG electromagnetic flow meters. Moreover, it is essential to highlight that the water distribution system of Mexico City is not instrumented in the sub-branch of the water pipelines that serve to supply water to the users. Due to the above, the case study was limited to analyzing the input and output of each water tank of the analyzed branch to detect anomalies and their behavior that could indicate the presence of leakages or measurement errors in the flow sensors.

The tanks are fed by gravity and connected by a leading pipeline, which separates them by 0.5 to 2 km. The tanks are located in Álvaro Obregón delegation in Mexico City, and each tank also supplies water to the local surrounding areas. Between Tank 1 and Tank 5 exists a difference of 200 m of elevation. Figure 2 shows the geographic location of the water branch analyzed in this study. The blue squares represent the tanks of Branch C presented in Figure 1. The blue line in Figure 2 indicates the connection between the tanks across Mexico City.

The water flow rate of six tanks was measured. A total of 11 water flow variables related to each tank's input and output flow with their corresponding timestamp were recorded every 15 min (i.e., the water flow measurements had a sampling rate of 15 min). The data utilized in this study were collected during the final two weeks of August 2020. The first week was the sampling period, during which the model was developed using the available data. The second week was then designated as the forecasting period, where the model's performance was evaluated by generating forecasts based on the learned patterns from the previous week. Each tank within the system is equipped with flow rate measurements from electromagnetic flowmeters at its entry and one or two of its exits, allowing for monitoring flow distribution within the main pipeline in liters per second (lps).

Furthermore, it is important to note that certain flow rates designated for local consumption or exit flows of the tank are unavailable due to a lack of instrumentation.

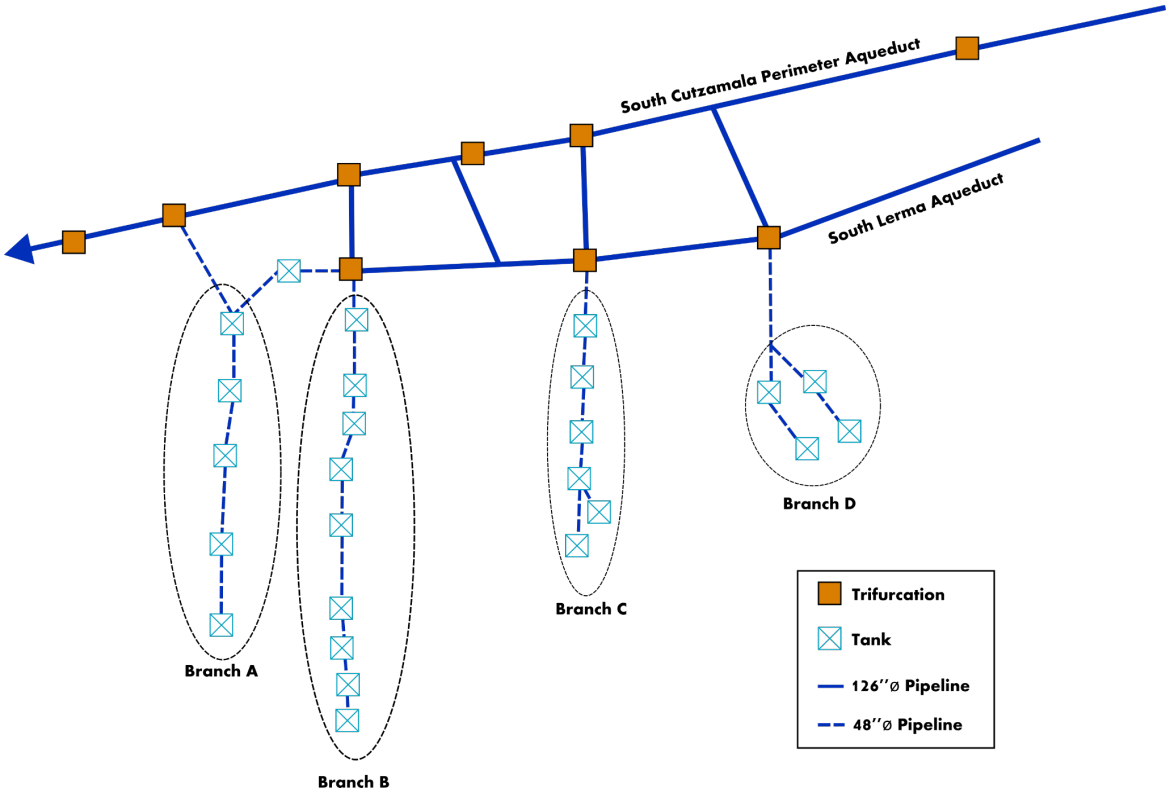

Figure 1. A schematic representation of a section of the primary water distribution system of Mexico City analyzed in this study. Filled orange squares represent trifurcations, while cyan squares represent tanks of various capacities. The blue lines represent pipelines with a diameter exceeding 126 inches, while the blue dotted lines indicate pipelines with a diameter of 48 inches.

A schematic representation of the water distribution branch analyzed in this work is presented in Figure 3, where the blue lines represent the main pipeline and connection with the next tank, and the gray lines correspond to the exit to the local network of the region. Furthermore, Figure 3 shows an example of the average water consumed in a week in percentage; therefore, for the first tank, all the input water corresponds to 100%, while in Tank 5, the water to the next stage of the network corresponds to 32%; this means that the region of this system consumed 68% of the total input water during the analyzed period.

Table 2 shows the variables and summary statistics from the period of water flow analyzed in this study for the tanks shown in Figures 1–3. The summary statistics are minimum, maximum, and mean flow rate presented in lps and the percentage of not available (NA) or empty observations, taking as the total the entire period of each variable every 15 min. The analysis and models presented in this study were implemented in RStudio Version 2022.02.3 + 492 using R Version 4.2.0 on a 64 bit Windows 7 PC with 12 GB of RAM and an AMD A10-5800K processor.

The first step involved identifying missing values in the raw time series flow sensor data, as seen in Figure 4. This process was repeated for each water tank variable presented in Table 2. Figure 4 illustrates the Tank 4 Inflow variable time series. The second step consisted of filling in the original raw time series data missing values, as seen in Figure 5. The average of the neighborhood values around the missing data points in the time series was computed to fill in the missing values, using observations from an equal number of data points on both sides of a central missing value. The process presented in Figure 5 was

repeated for each water tank variable in Table 2; the same Figure illustrates the filling of missing values for the Tank 4 Inflow variable.

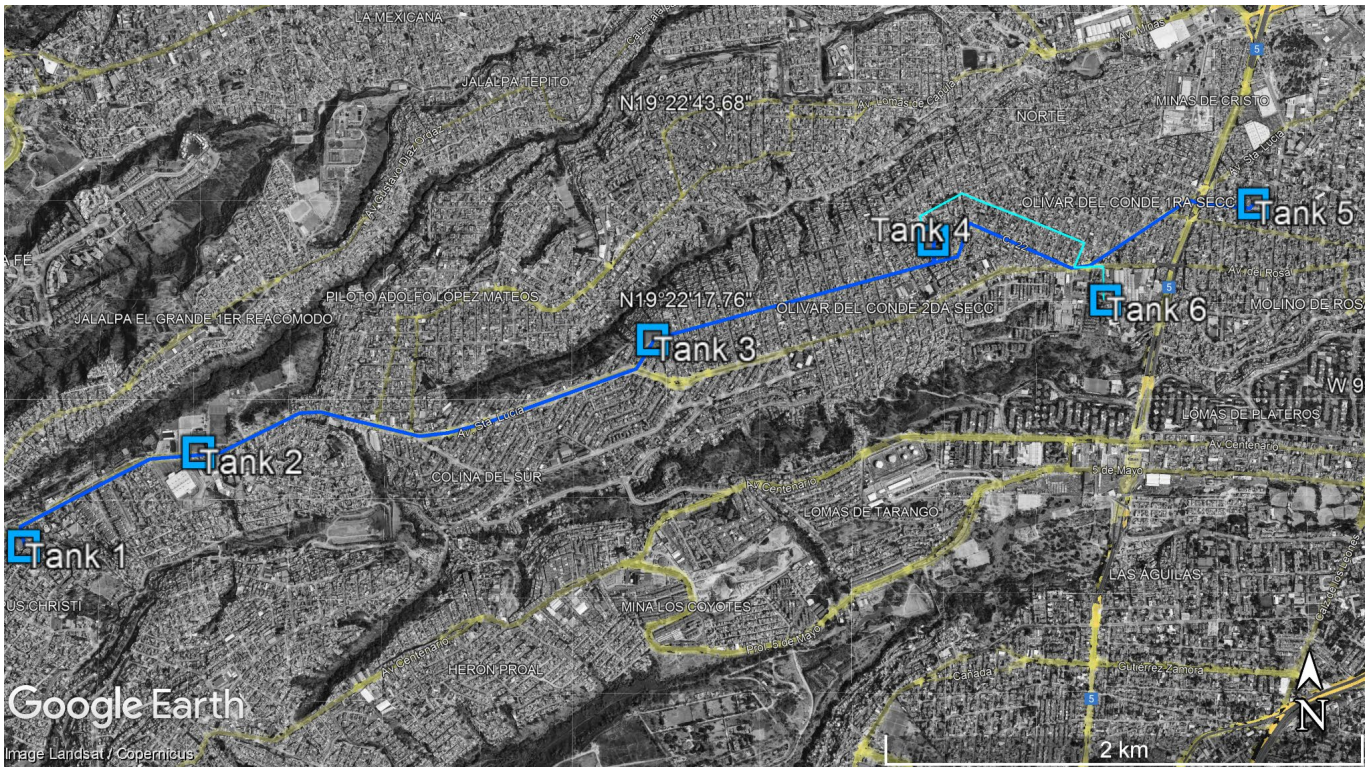

Figure 2. Geographic locations and connections of the water tanks of the water distribution branch of Mexico City analyzed in this study. The blue squares represent the locations of the tanks along the city. The lines on the map represent the connections between the tanks across Mexico City. The blue line represents a pipeline of 48 in in diameter, and the cyan line represents a pipeline of 20 in in diameter.

Table 2. Characteristics of water flow variables in the sampling period for the six tanks of the analyzed water distribution branch. The variable number corresponds to the ones shown in Figure 3.

Variable Number	Variable	Minimum (lps)	Maximum (lps)	Mean (lps)	NA (%)
1	Tank 1 Inflow	0	2587	1837	0%
2	Tank 1 Outflow	0	92.73	65.3	0%
3	Tank 2 Inflow	0	2430	1843	0%
4	Tank 2 Outflow	0	401.4	280.9	0%
5	Tank 3 Inflow	−1.86	2220.15	1554.58	0%
6	Tank 4 Inflow	0	2489.2	994.8	14%
7	Tank 4 Outflow	−107.01	117.41	6.334	8%
8	Tank 5 Inflow	0	2465.8	1088.3	5%
9	Tank 5 Outflow A	−352.5	2912	819.3	5%
10	Tank 5 Outflow B	0	530.3	412.9	0%
11	Tank 6 Inflow	0	101.5	67.167	0%

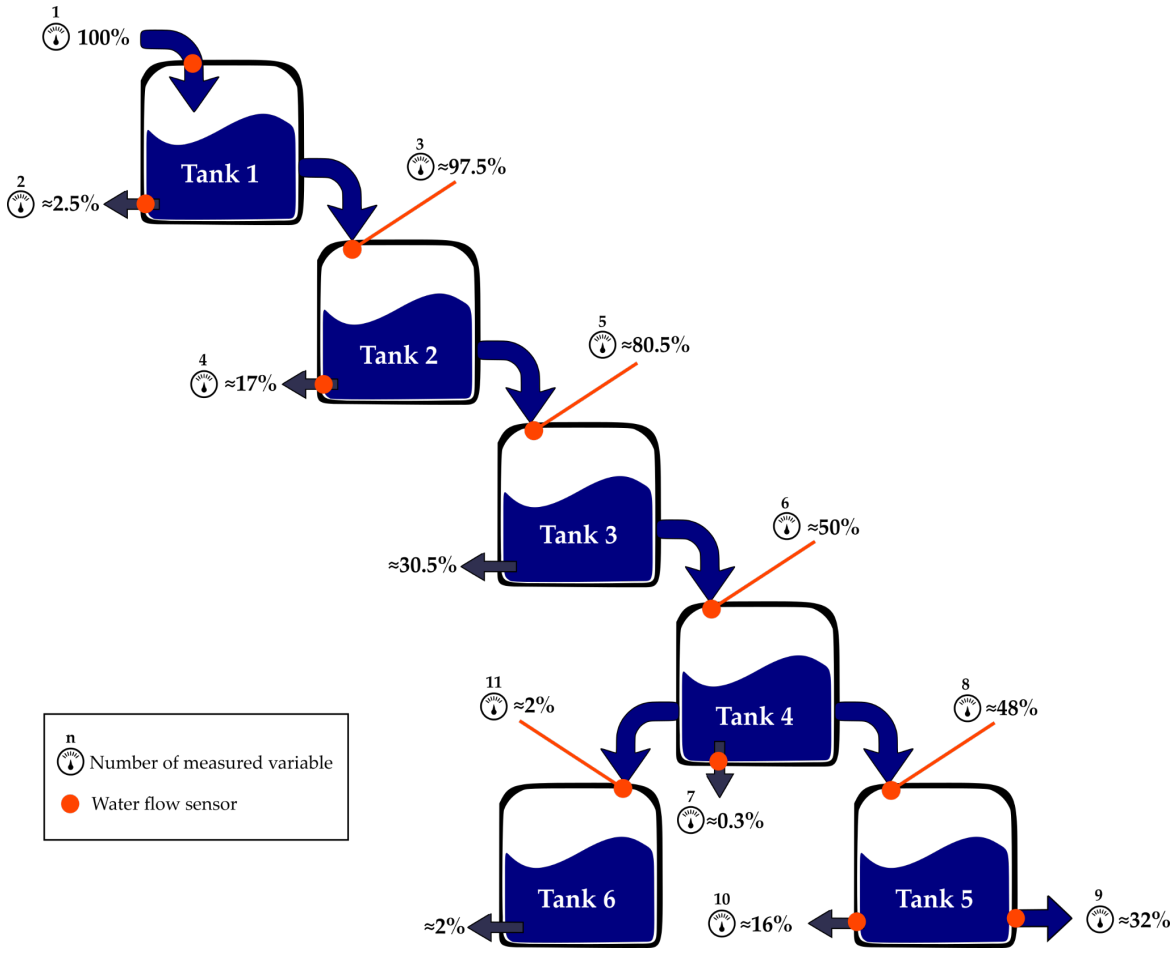

Figure 3. Schematic representation of the water distribution branch analyzed in this study. The case study comprised six tanks connected by gravity. The average water mass of one week is presented as a percentage. The measured variables are indicated with a symbol, and the variable numbers correspond with those shown in Table 2 for each water flow sensor. The orange dots indicate the positions of the water flow sensors along the water distribution branch that were measured.

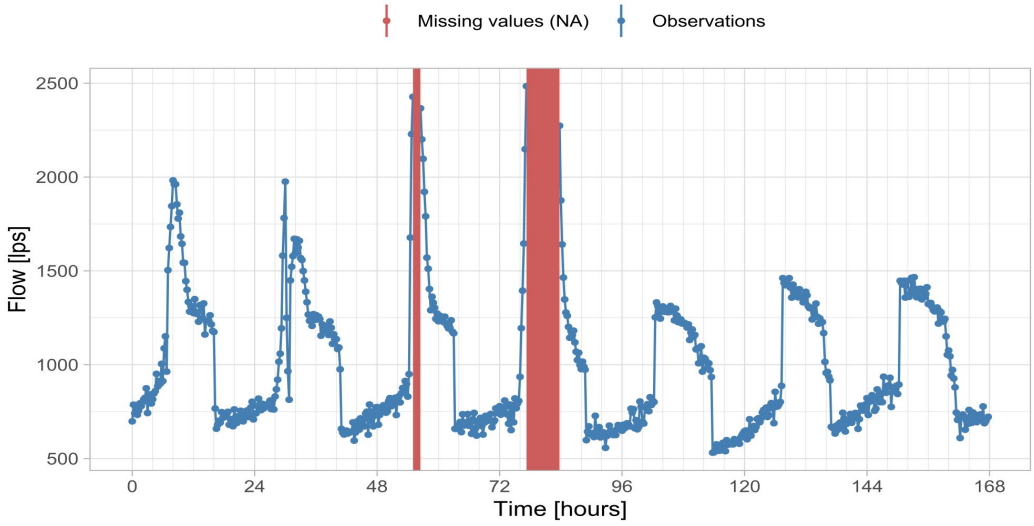

Figure 4. Example of identifying missing values in the raw time series flow sensor raw data for Tank 4 Inflow variable during one week. The dotted blue line represents the observations of the Tank 4 Inflow variable. The red line represents the missing values in the Tank 4 Inflow variable.

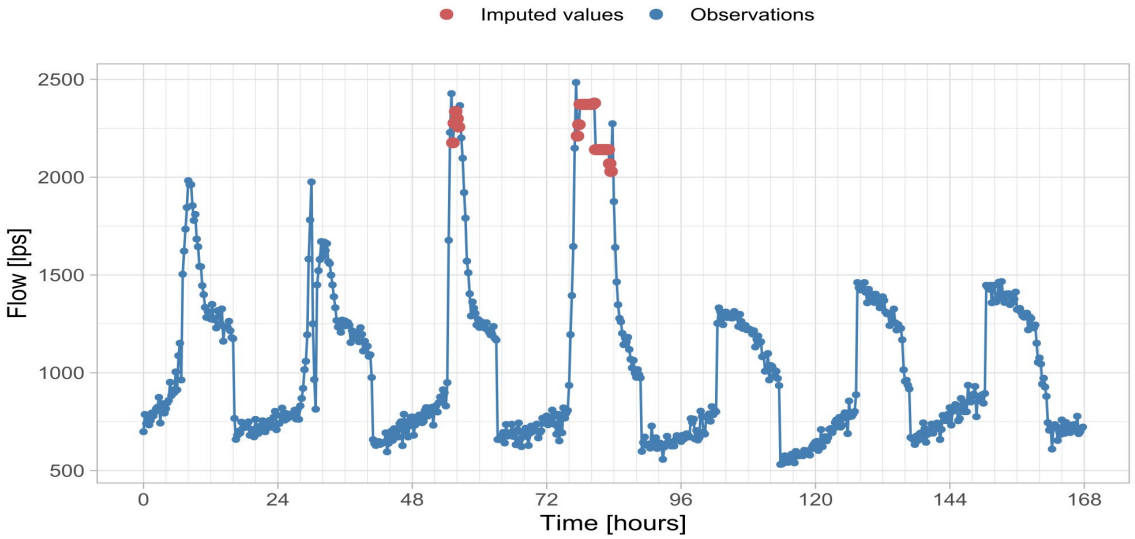

Figure 5. Example of filling the missing values for Tank 4 Inflow variable corresponding to one week. The dotted blue line represents the observations of the Tank 4 Inflow variable. The dotted red line represents the imputed values in the Tank 4 Inflow variable.

3.2. Autoregressive Integrated Moving Average and Seasonal Autoregressive Integrated Moving Average

The overall methods used to model the water flow variables of the water distribution branch analyzed in this study are illustrated in Figure 6. Initially, measurements of each water tank were collected, and then, the data were pre-processed to prepare them for modeling. The best model for each variable was used to forecast over two different periods: one day and one week. In the case of the TF models, time-ahead data were also included as an input variable for estimating the forecasts. Further details on this process are explained in subsequent sections of this research work. The theoretical background of the ARIMA models is described below.

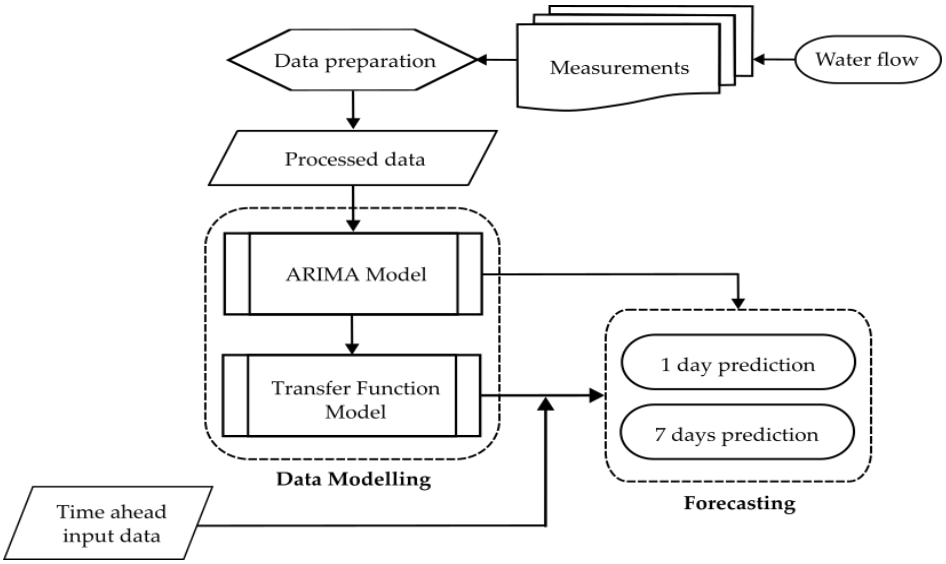

Figure 6. The methodology flowchart utilized to fit the ARIMA and TF models based on the water flow data.

ARIMA models are composed of a dependent variable Y_t, which depends on past values Y and an error term e_t. Besides, these models are characterized by three elements: a moving average component, an autoregressive component, and a differencing (integration)

component. The autoregressive component indicates that Y_t depends on one or multiple lagged values of Y_t. The moving average component shows that Y_t depends on one or multiple lagged values of the error e_t. Finally, the integration or differencing component indicates that the series should be stationary; computing the difference between neighboring observations in the time series accomplishes the above [37].

The notation ARIMA(p,d,q) represents the order of the ARIMA models, where p is the order of the autoregressive component, d is the order of the differencing (integration) component, and q is the order of the moving-average process [38]. The backshift operator (B) can be used to define an ARIMA model as follows:

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)e_t \quad (1)$$

where Y_t is the value of the series observed at time t; B is the backshift operator; ϕ are the autoregressive polynomials; θ is the moving average polynomial; e_t are the error terms of the model. The error terms were assumed to be independent and identically distributed with a normal distribution and zero mean [24].

However, the Seasonal Autoregressive Integrated Moving Average was considered to model the seasonal component of the time series. In this regard, Seasonal ARIMA models were selected since, as shown in Figures 4 and 5, the water flow time series have a seasonal component (i.e., the series exhibits a regular fluctuation), which appears every 96 observations, corresponding to a day of water flow measurements. This seasonal term makes the water flow time series nonstationary; therefore, to consider the seasonal component of the time series and to fit an ARIMA model, the seasonal component needs to be considered for the models [24].

Seasonality implies that Y_t depends on lagged values of Y_t at a regular interval s. Seasonal ARIMA models consider the non-Seasonal ARIMA(p,d,q) and three additional parameters labeled as $(P, D, Q)_m$ to account for the seasonality presented in a time series. The m term refers to the number of time steps corresponding to a single seasonal period. On the other hand, the term P represents the order of the seasonal autoregressive component; the term Q refers to the seasonal moving average component; the term D represents the seasonal differencing component [38]. The mathematical representation of the Seasonal ARIMA models is shown in the next expression:

$$\Phi_P(B^m)\Phi_P(B)(1-B^m)^D(1-B)^d Y_t = \Theta_Q(B^m)\theta_q(B)w_t \quad (2)$$

where Y_t is a seasonal time series; w_t is the Gaussian white noise process; $\Phi_P(B)$ is the non-seasonal autoregressive polynomials; $\theta_q(B)$ represents the non-seasonal moving average polynomial. d is the non-seasonal differencing term. D is the seasonal differencing term. One key aspect is that, when $D = 1$, this is sufficient to ensure stationarity in the time series. $\Phi_P(B^m)$ represents a seasonal autoregressive polynomial; the term $\Theta_Q(B^m)$ is a seasonal moving average polynomial. Finally, B is the backshift operator [38].

In general, the optimal ARIMA model parameters are determined by considering three criteria: (a) using Akaike's information criterion (AIC); (b) examining the auto-correlation function (ACF) to determine the q parameter of the ARIMA model and the number of moving average (MA) coefficients and computing the partial auto-correlation function (PACF) of the residuals to determine the p parameter for the number of autoregressive coefficients; (c) by plotting the series residuals to confirm that the error term is equivalent to white noise. The following sections describe the definitions and procedures to compute the ACF, PACF, and AIC in more detail.

3.3. Auto-Correlation Function and Partial Auto-Correlation Function

Auto-correlation can be defined as the degree of similarity of a time series with a lagged version of itself. Furthermore, the plot of a time series' auto-correlations against lags is known as the auto-correlation plot. Thus, the so-called ACF shows the linear relationship between the observation y_t at time t and the observation at a previous time (y_{t-k}) that

is separated by k lags at time [38]. Taking the above into account, the mathematical representation of the ACF for a time series y_t is shown in the expression below:

$$ACF(y_t, y_{t-k}) = \frac{Covariance(y_t, y_{t-k})}{variance(y_t)} \tag{3}$$

where k is the lag, and it is defined as the difference in time between the observation y_t and the observation y_{t-k}. The term $ACF(y_t, y_{t-k})$ denotes the correlation between the observations y_t and y_{t-k} that are separated by k periods. The ACF serves to know the order of the moving average component of an ARIMA model. Moreover, the ACF also allows analyzing the periodicity and detecting recurrence in a time series [24].

On the other hand, the so-called partial auto-correlation or conditional correlation removes the intermediate observations when computing the correlation between two observations at different lags. In this case, the PACF is conditional on the intermediate observation of the time series, since they are taken out from the covariance computation. For instance, consider the PACF of two observations y_t and y_{t-k} (i.e., $k = 2$) [38]. The above can be expressed as shown below:

$$PACF(y_t, y_{t-2}) = \frac{covariance(y_t, y_{t-2}|y_{t-1})}{\sqrt{variance(y_t|y_{t-1})}\sqrt{variance(y_{t-2}|y_{t-1})}} \tag{4}$$

where the term $PACF(y_t, y_{t-2})$ is the PACF between the observations y_t and y_{y-2}. Notice that the covariance between y_t and y_{y-2} and the variance of y_t and y_{y-2} are conditional on the intermediate observation y_{t-1} since the PACF removes the effect of the intermediate observations [38]. The computation of the PACF serves to know the order of the autoregressive component of an ARIMA model.

Figure 7 illustrates the Tank 4 Inflow ACF and PACF, from which it can be inferred that the series is not stationary. It is important to mention that the auto-correlation and partial auto-correlation functions are dimensionless; the above implies that they are independent of the scale of measurement of the analyzed time series [24].

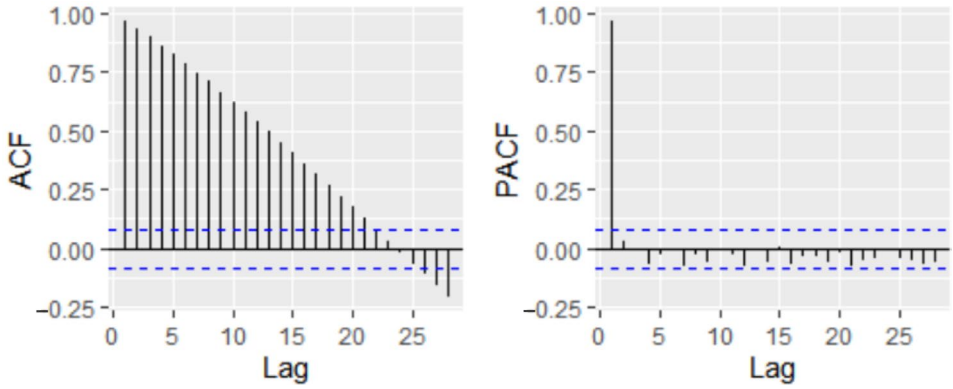

Figure 7. ACF (**left**) and PACF (**right**) plot of Tank 4 Inflow original time series.

Given that the observations were taken every 15 min and matched up with the earlier visual inspection of the residuals, it was determined that the series becomes stationary by differencing at a lag of 96. This corresponds to the 96 observations in a single day. Therefore, according to the ACF and residuals of the ARIMA(1,0,0)(1,1,1)(96) model of Tank 4 Inflow presented in Figure 8, the model is adequate, since the residuals follow a normal distribution. Similarly, the remaining water flow variables of the system depicted in Table 2 were processed, and the ACF, PACF, and residual analyses were repeated.

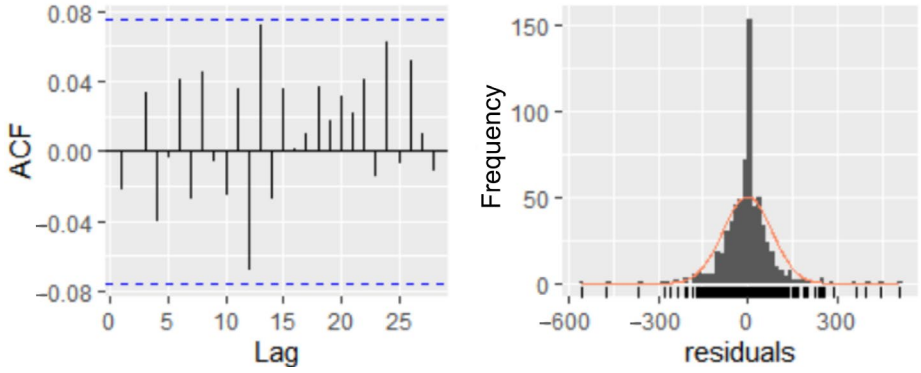

Figure 8. ACF (**left**) and residuals (**right**) of the ARIMA(1,0,0)(1,1,1)(96) model of Tank 4 Inflow.

As previously stated, the ACF and PACF correlogram analysis is required to determine the components of an ARIMA model. If the time series is stationary or not, it can be determined by looking at the residual plots. After a few auto-correlations, the ACF for a stationary time series will zero out. However, the ACF for nonstationary time series will decline slowly or increase positively [24]. Following multiple iterations and the initial analysis, some models were suggested as the best. After identifying the best ARIMA models for the series, the best model was chosen by comparing its residuals and information criteria.

The AIC, mean absolute percentage error (MAPE), and root-mean-squared error (RMSE) criteria were used to measure the performance of each model. The residuals were then examined for the model's diagnosis, and if the model was satisfactory, it could be used to forecast; otherwise, additional models must be tested [24]. Figure 9 shows the methodology to generate an ARIMA model via the Box–Jenkins approach. The following section describes the metrics utilized in this study in further detail.

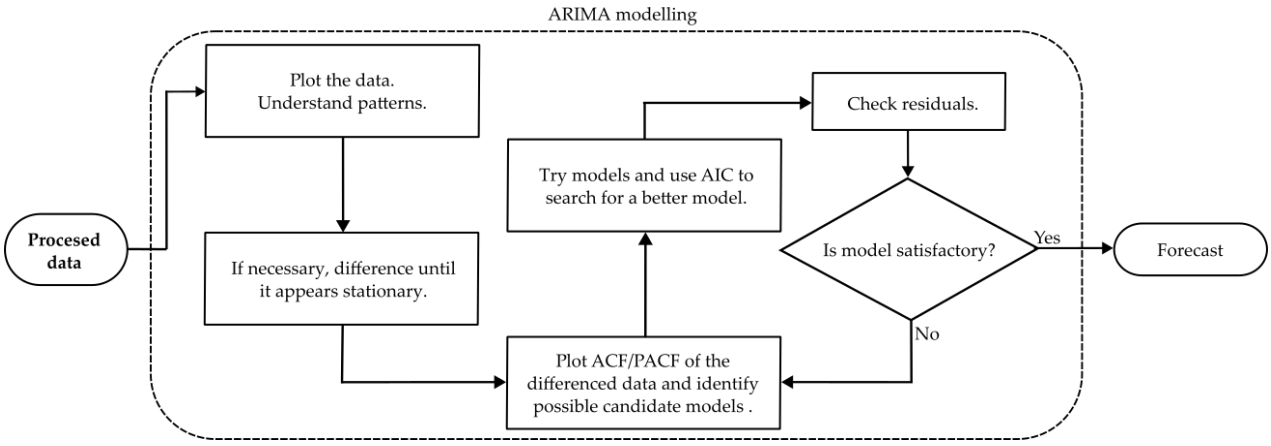

Figure 9. An overview of the generation process of an ARIMA model.

3.4. Evaluation Metrics

The fitting procedure of the resulting ARIMA models was assessed with the aid of the AIC, as shown in Figure 9. The information criterion measures the model's ability to explain the relationship between the variables. A common criterion is to compute the AIC, which is an information criterion that enables the assessment of the quality of the models by rewarding those with minor errors while penalizing those with too many parameters [38]. Thus, this criterion allows the selection of sparse models [39]. The mathematical representation of the AIC is shown in the following expression:

$$AIC = -2logL(\hat{\theta}) + 2K \quad (5)$$

where $logL(\hat{\theta})$ represents the likelihood function and K is the total number of parameters of the model. A lower value of the AIC represents a better model with a higher likelihood value. Compared to other metrics, such as the Bayesian information criterion, the AIC value provides a greater penalty on the number of parameters.

On the other hand, for this work, the MAPE and RMSE were used to measure the error and to have a numerical comparison of the effectiveness of the proposed models after selecting the best through the AIC. The RMSE and MAPE were used to compare the accuracy of the model's forecast to the actual values, with a lower value indicating a better fit [24,38,40]. The equations of these indicators are shown as follows:

$$MAPE = \frac{1}{N_f} \sum_{i=1}^{N_f} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N_f} \sum_{i=1}^{N_f} (y_i - \hat{y}_i)^2} \quad (7)$$

where y_i is the observed value and $\hat{y}_i$ the predicted value at time i; N_f is the number of forecast time steps.

A week was chosen to evaluate the Seasonal ARIMA models. The models were tested within different time frames and assessed on various dates. Two data transformations were considered to transform a nonstationary time series into a stationary series and use the Box–Jenkins methodology: first, differencing, and second, differencing with a transformation using the natural logarithm. By calculating the difference between two consecutive observations, differencing makes a nonstationary time series stationary. The time series' variance can be stabilized using the natural logarithm. Some preliminary models that follow the patterns and methodology of Box–Jenkins can be provided after comparing and analyzing the resulting ACF and PACF of the water flow time series. The AIC was calculated for each fitted model to choose the optimal [41].

Moreover, to evaluate the forecast of each model, one day (equivalent to 96 observations) and one week (corresponding to 672 observations) were examined, with a confidence interval of 95%. Figure 10 shows an example of the forecast of one week in the future for the Seasonal ARIMA models for the Tank 4 Inflow time series. The forecast's confidence interval was calculated at 95% by obtaining the standard errors of the estimates as described in the Box–Jenkins methodology [24].

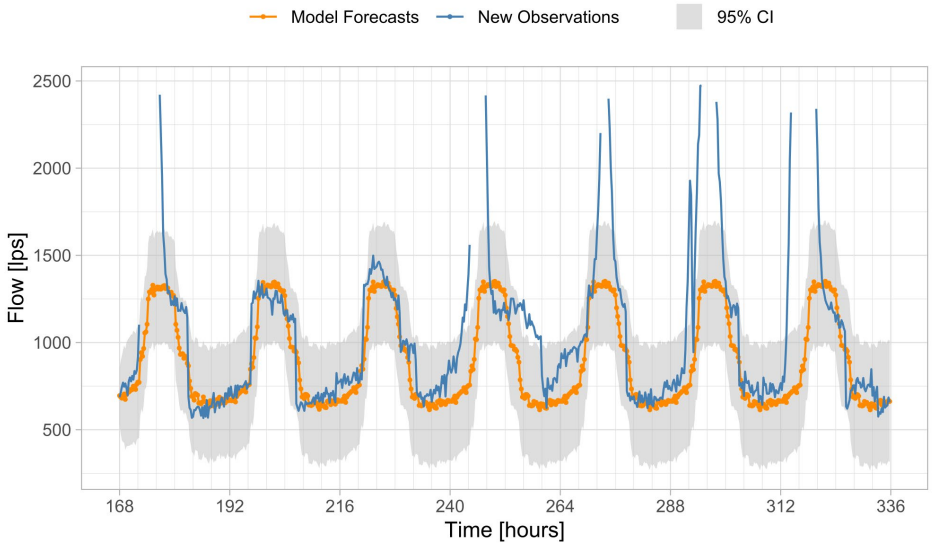

Figure 10. Forecasts of 1 week of the model ARIMA(1,0,0)(1,1,1)(96) of Tank 4 Inflow with the corresponding model forecast in orange (in blue are the actual observations), and the grey zone indicates the 95% confidence interval of the model.

3.5. Transfer Function Models

This section presents the theoretical background and methodology for developing TF models based on the Box–Jenkins approach. TFs are models that combine a causal approach and a time series approach. The time series X_t affects the time series Y_t through a TF, which spreads the impact X_t via some period in the future. The resultant TF model connects the output series (Y_t), the input series (X_t), and a noise term (N_t). The addition of a noise term is considered since, in practice, the response of a system could be affected by disturbances and noise induced by the environment, which corrupts the system's output by an amount N_t. Hence, a TF is equivalent to a response function. The mathematical representation of TF models can be written in terms of the backward operator B, as shown in Equation (8) [24].

$$Y_t = \delta^{-1}(B)\,\omega(B)B^b X_t + N_t \tag{8}$$

where Y_t is the output of the system at time t; X_t is the input of the system at time t; B is the backshift operator; $\omega(B)$ is an sth-order polynomial operator; δ^{-1}(B) is an rth-order polynomial operator; B^b is a bth order dead time operator, which indicates the number of periods before any effect is discernible; finally, N_t is the amount of noise to which the system is susceptible. The terms (b, s, r) are integers greater than or equal to zero. The term $\omega(B)$ controls the effect of current and previous input values in the system's response. On the other hand, the term δ^{-1}(B) controls the effect of previous output values in the system's response [42].

A TF estimation of the system based on the Box–Jenkins methodology was developed, and it was motivated by the correlation between the system variables displayed in Figure 3, specifically the input and output flow of each water tank. Figure 11 illustrates the overall procedure for predicting using TFs, based on the Box–Jenkins approach for fitting, and validating TF models. The definition of the input and output ARIMA models that were used, the prewhitening of both series (i.e., the method of removing the impact of serial correlation on trend analysis), and the calculation of cross-correlation for the identification of the pre-estimates and final parameters of the model or tentative models are the key steps in this process. The models' diagnostics were then determined, and if the model was sufficient, it could be utilized for the forecast; if not, a different model should be suggested [24].

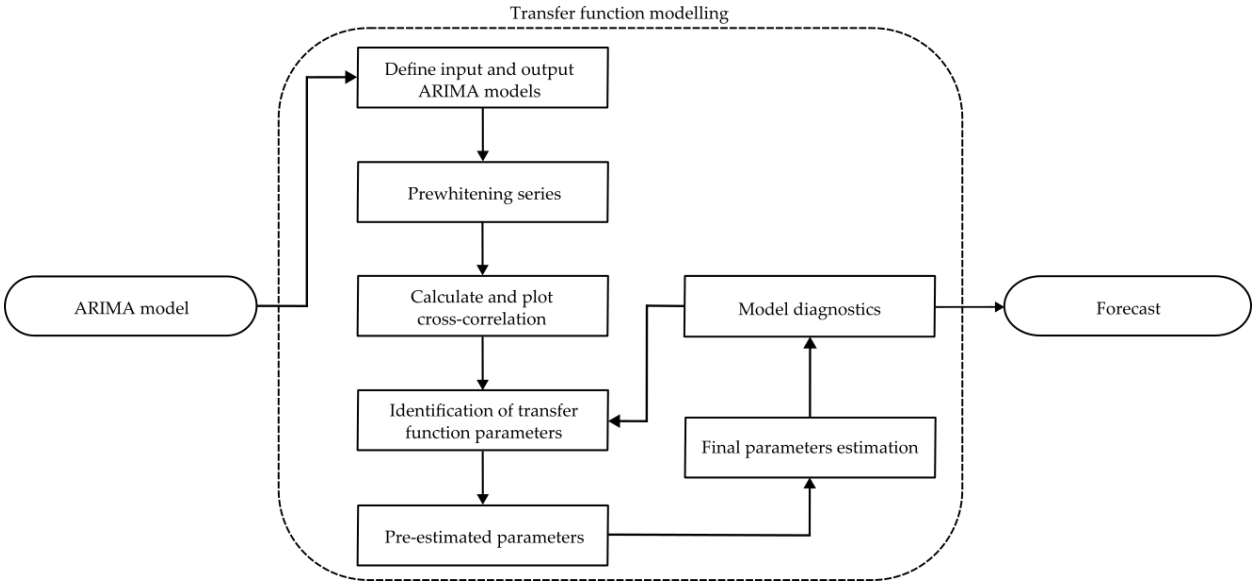

Figure 11. Transfer Functions' modeling and forecasting process.

The dependent (output) and independent (input) variables were modeled via ARIMA models. The ARIMA models used for the TF were the same as those developed for

each water flow variable, as explained in Section 3.2. Then, the input and output series generated by the fitted ARIMA models were prewhitened. Consequently, the series were cross-correlated to find the relationship between the lags, or the effect of X_t over Y_t. The cross-correlation function is represented as shown in Equation (9) [24].

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \tag{9}$$

where $\rho_{xy}(k)$ is the cross-correlation function of a stationary bivariate process; $\gamma_{xy}(k)$ is the cross-covariance coefficients between the series x_t and y_t at lags $k = \pm 0, \pm 1, \pm 2, \ldots$; σ_x is the standard deviation of the x series; σ_y is the standard deviation of the y series. The cross-covariance function $\gamma_{xy}(k)$ of a stationary bivariate process is defined as shown in Equation (10) for lags $k = \pm 0, \pm 1, \pm 2, \ldots$.

$$\gamma_{xy}(k) = E\left[(x_{t-k} - \mu_x)(y_t - \mu_y)\right] = E\left[(y_t - \mu_y)\left[(x_{t-k} - \mu_x)\right] = \gamma_{yx}(-k)\right. \tag{10}$$

The importance of the cross-correlation function of the prewhitening input and output series is that it provides an estimate of the impulse response of the system. This impulse response estimate serves to know the order of the s and r polynomials, as well as the order of the dead time operator (B^b) that should be used to fit the TF model. Similar to other areas such as signal processing and system analysis, the impulse response is used for the graphical or mathematical representation of the output of a system or a model in response to a brief input signal or impulse. The impulse response provides valuable insights into the system's behavior, including its frequency response, stability, and the effect of the input signal on the output [43].

The plot pattern of the cross-correlation function determines the values of b, r, and s, which, according to the Box and Jenkins [24] notation, are the parameters for a TF (b, s, r) model. The parameters b and s determine the number of lagged terms of x that entered into the TF. The value of b is determined by the first lag significantly different from zero in the cross-correlation plot. The s term is established by how long x influences y after the first significant lag. The r value represents how long the output series (y_t) is connected with the prior value of the output series. The value of r can be set by analyzing the plot of auto-correlation or determined by the plot pattern of lag $(b + s)$; if it has an exponential decay, then $r = 1$ could provide an appropriate approximation of the TF, and if it has a sine wave plot pattern, then $r = 2$ could provide an approximation of the TF [24].

The input Tank 1 Inflow and the Tank 2 Inflow series are cross-correlated in Figure 12, demonstrating that the fifth lag is the most-significant. Nevertheless, Lag 0 is the first latency that deviates sufficiently from zero. It is also clear that the fifth lag (from Lag 0 to Lag 5) is the number of delays between the previous significant lag and the current lag. Finally, the plot development appears to follow a sine wave. Consequently, a TF with parameters (2,3,0) could be a first model proposal.

The fourth step used the fitted TF models to forecast one day and week in the future. The forecast's confidence interval was also calculated, as can be seen in Figure 13.

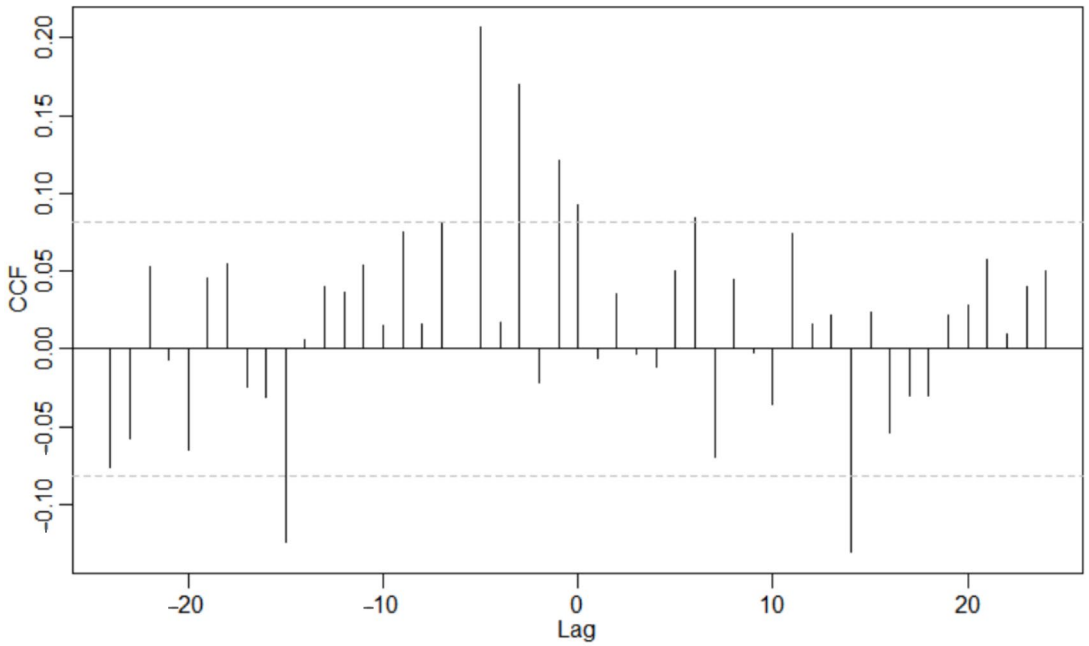

Figure 12. Cross-correlation of prewhitened input and output for Tank 2.

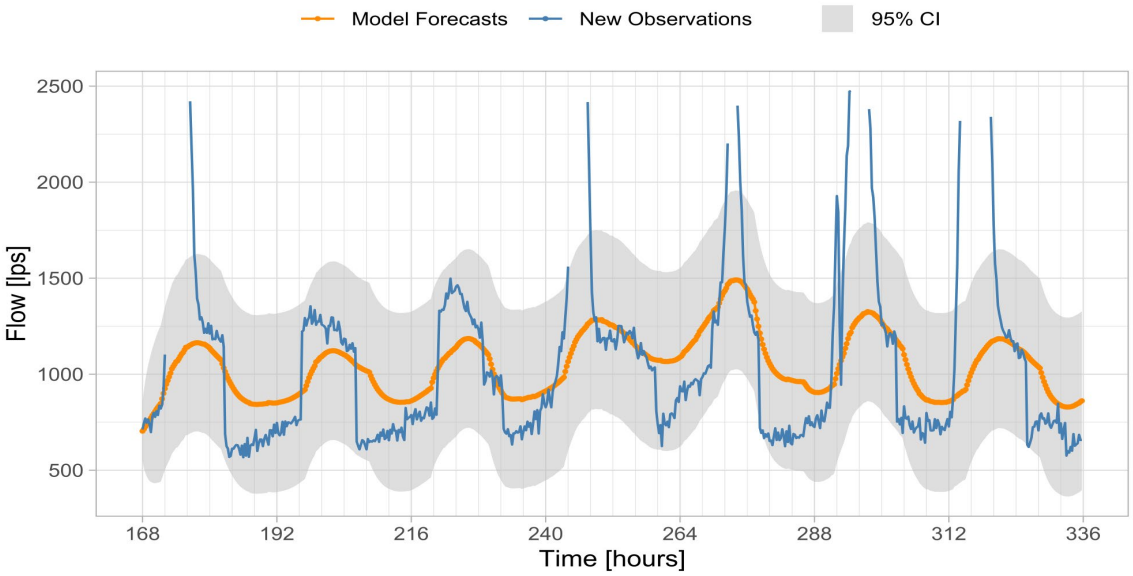

Figure 13. Forecasts of 1 week of the model TFM(0,1,0) of Tank 4 Inflow taking into consideration Tank 3 Inflow data as input data of the model. The orange line represents the TF model forecast; the blue line represents the actual observations; the gray area represents the 95% confidence interval of the fitted TF model.

3.6. Anomaly Detection in Water Distribution Branches Methodology

After generating and using the models for forecasting, the anomaly detection procedure involved comparing the observed values with the 95% confidence interval of the model's forecast. This work assumed that an anomaly presented in the measured water flow's water branch is outside the model forecast's confidence interval. Figure 14 shows the general methodology for the data evaluation for anomaly detection in water distribution branches.

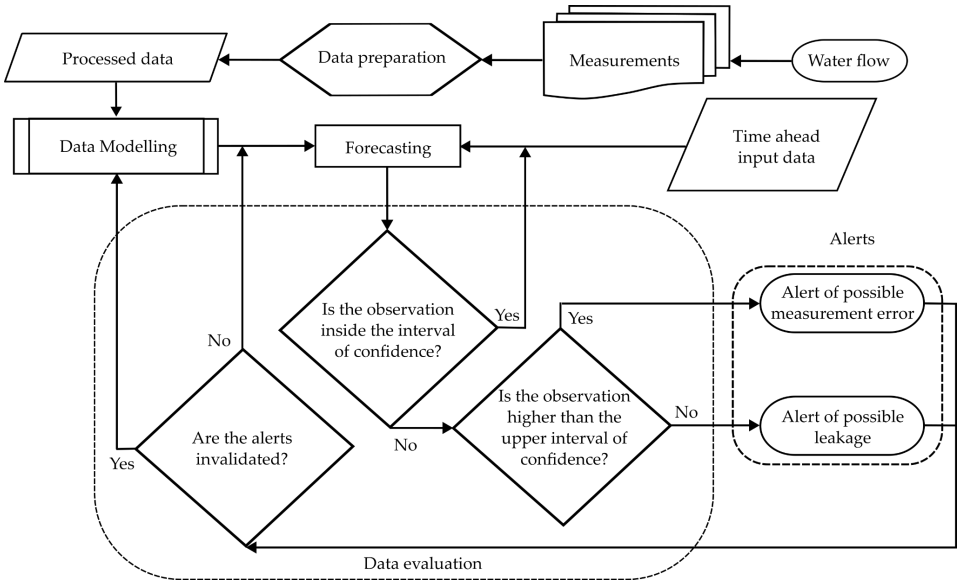

Figure 14. Data evaluation flowchart to detect the anomalies in the water tanks of the analyzed water distribution branch by employing the fitted Seasonal ARIMA and TF models' 95% confidence intervals.

The forecast values, the confidence interval, and the following day and week observations are needed by the methodology shown in Figure 14 before any other feedback processes can begin. First, it determines whether the observed values for each model are within the confidence interval of the forecast; if they are, it goes back to the forecast stage and compares the subsequent observations with the subsequent prediction. In cases with missing data points, it is plausible that they are due to various factors, such as interruptions in sensing or communication caused by issues with the energy supply at the station, sensor malfunctions, intermittent data transmission problems, or failures in the database. When these data points are missing, the corresponding alert is labeled as "not available" to indicate the absence of data.

On the other hand, if an observation exists and falls outside the confidence interval, it can indicate two main possibilities. Firstly, it could suggest a potential measurement error where the sensor may have malfunctioned and provided an incorrect reading. Alternatively, it may indicate a genuine change in the water system's behavior, potentially caused by external factors such as water leakages, variations in water demand, hydraulic system issues, water availability, or operational actions. This methodology suggests two key alerts: potential measurement error and potential water leakage to facilitate the detecting and understanding of different types of anomalies. When new observations significantly exceed the confidence interval's upper limit, it indicates a potential measurement error, which is more likely than a possible water leakage, since having a higher water flow rate than what the source can supply is not feasible. However, it is also possible that the sensor briefly malfunctioned if the alarm is not persistent. Conversely, suppose the new observation of water inflow falls below the confidence interval's lower limit. In that case, it suggests a potential water loss, indicating the possibility of leaks occurring between the water tanks. This inference is drawn from the observed water inflow being below the expected range, but a misread by the sensor cannot be completely ruled out as a possibility. Incorporating these alerts into the methodology makes identifying and categorizing anomalies easier, leading to improved system operation and early detection of potential issues.

The appearance of the warnings is next examined; if they are persistent and invalidated by the user or another qualified individual, the model must be redesigned because the old model does not account for the new observations. Additionally, the distribution of the series could have been altered, necessitating a return to the model-generation stage.

4. Results

The errors from the Seasonal ARIMA models were calculated to select the best model for each variable. On average, it required 8 to 12 iterations to generate different models and compare the AICs between them to find the best fitted model for each water flow variable. A Seasonal ARIMA model was generated for each water tank's input and outflows. Table 3 shows the Seasonal ARIMA models selected for each water tank in Figure 2 and the resulting AIC, RMSE, and MAPE.

Table 3. Seasonal ARIMA models' AIC, MAPE, and RMSEs from each water flow variable. The variable number corresponds to the ones shown in Figure 3 and Table 2.

Variable Number	Variable	Model Notation	AIC	RMSE	MAPE
1	Tank 1 Inflow	ARIMA(2,0,1)(0,1,1)(96)	6011.62	34.87778	1.412087
2	Tank 1 Outflow	ARIMA(0,1,1)(0,1,1)(96)	1882.58	1.045328	1.112624
3	Tank 2 Inflow	ARIMA(2,0,1)(0,1,1)(96)	5998.8	34.47143	1.378066
4	Tank 2 Outflow	ARIMA(0,1,0)(0,1,1)(96)	3322.98	3.854425	1.069508
5	Tank 3 Inflow	ARIMA(2,0,0)(0,1,1) (96)	5744.72	29.29716	1.069776
6	Tank 4 Inflow	ARIMA(1,0,0)(1,1,1)(96)	6727.91	75.56354	5.415305
7	Tank 4 Outflow	ARIMA(1,0,1)(0,1,1)(96)	4901.18	16.79202	33.91554
8	Tank 5 Inflow	ARIMA(1,1,2)(0,1,1)(96)	7497.96	135.589	5.748487
9	Tank 5 Outflow A	ARIMA(1,0,0)(0,1,1)(96)	7435.93	155.8952	22.79048
10	Tank 5 Outflow B	ARIMA(1,1,0)(0,1,1)(96)	3129.15	2.869073	0.3859113
11	Tank 6 Inflow	ARIMA(0,1,1)(0,1,1)(96)	1296.07	0.5830809	0.3625255

Then the models were utilized to forecast one day and one week ahead, and the obtained forecasts were compared with the actual observations to calculate the forecasting MAPE and RMSE. The obtained RMSEs and MAPEs that each model obtained for one-day and one-week forecasts are presented in Table 4.

Table 4. Seasonal ARIMA models' MAPEs and RMSEs from each model forecast for one day and one week in the future. The variable number corresponds to the ones shown in Figure 3 and Table 2.

Variable Number	Variable	Model Notation	1-Day Forecast		1-Week Forecast	
			RMSE	MAPE	RMSE	MAPE
1	Tank 1 Inflow	ARIMA(2,0,1)(0,1,1)(96)	70.33606	3.221687	156.852	6.182667
2	Tank 1 Outflow	ARIMA(0,1,1)(0,1,1)(96)	5.972818	7.192756	6.454227	8.924136
3	Tank 2 Inflow	ARIMA(2,0,1)(0,1,1)(96)	83.17052	3.757622	142.41759	5.610582
4	Tank 2 Outflow	ARIMA(0,1,0)(0,1,1)(96)	14.04879	4.574004	28.658838	8.957665
5	Tank 3 Inflow	ARIMA(2,0,0)(0,1,1)(96)	86.98132	4.963585	169.63034	8.577175
6	Tank 4 Inflow	ARIMA(1,0,0)(1,1,1)(96)	221.12301	11.697919	272.26303	14.447808
7	Tank 4 Outflow	ARIMA(1,0,1)(0,1,1)(96)	18.52472	352.6942	17.1954	77.69167
8	Tank 5 Inflow	ARIMA(1,1,2)(0,1,1)(96)	303.4553	23.445315	384.0901	23.786044
9	Tank 5 Outflow A	ARIMA(1,0,0)(0,1,1)(96)	367.1641	-	556.4129	-
10	Tank 5 Outflow B	ARIMA(1,1,0)(0,1,1)(96)	15.568536	3.3782927	28.982522	5.397012
11	Tank 6 Inflow	ARIMA(0,1,1)(0,1,1)(96)	0.6684805	0.5289368	1.5103029	1.220953

On the other hand, Table 5 presents the AICs, MAPEs, and RMSEs of the best fitted TF models for each variable. Only the possible and correlated water flow variables were used to generate the TF models based on the system presented in Figure 3. In addition, Table 5 shows the order of the TF models' polynomials and dead time operator of the obtained TF models with the corresponding RMSE and MAPE values computed with the data interval used for generating each model.

Table 5. Input and output variables were determined for the Transfer Function model and AIC, MAPE, and RMSE of each of the best-selected models for each water flow variable. The variable number corresponds to the ones shown in Figure 3 and Table 2.

Input Variable	Output Variable	TF Model(b,s,r)	AIC	RMSE	MAPE
Tank 1 Inflow (Variable 1)	Tank 2 Inflow (Variable 3)	TFM(4,2,3)	6815.29	38.47823	1.64836
Tank 2 Inflow (Variable 3)	Tank 3 Inflow (Variable 5)	TFM(2,4,2)	5536.81	29.46901	0.99219
Tank 1 Inflow (Variable 1)	Tank 3 Inflow (Variable 5)	TFM(6,1,2)	5548.24	29.76044	0.94921
Tank 3 Inflow (Variable 5)	Tank 4 Inflow (Variable 6)	TFM(0,1,0)	7755.84	77.31749	5.27644
Tank 4 Inflow (Variable 6)	Tank 5 Inflow (Variable 8)	TFM(1,3,0)	8598.87	144.66703	4.90388
Tank 4 Inflow (Variable 6)	Tank 6 Inflow (Variable 11)	TFM(0,2,0)	1227.6	0.69927	0.44741
Tank 1 Inflow (Variable 1)	Tank 5 Outflow B (Variable 10)	TFM(0,2,2)	3511.45	3.28181	0.47553

Furthermore, Table 6 presents the MAPEs and RMSEs of the one-day and one-week forecasts of the fitted TF models. The models with the lowest MAPE were chosen and utilized in the data-evaluation stage. The asterisk indicates the variable and error values lower than those obtained for the Seasonal ARIMA model.

Table 6. Input variable for Transfer Function model and the best model selected for the variable, AIC, MAPE, and RMSE of each selected model forecasts one day and one week after. The asterisk denotes the variable and error values that are lower than those obtained from the Seasonal ARIMA model corresponding to the same output variable. The variable number corresponds to the ones shown in Figure 3 and Table 2.

Input Variable	Output Variable	TF Model (b,s,r)	1-Day Forecast		1-Week Forecast	
			RMSE	MAPE	RMSE	MAPE
Tank 1 Inflow (Variable 1)	Tank 2 Inflow * (Variable 3)	TFM(4,2,3)	43.727 *	1.863451 *	41.114 *	1.806004 *
Tank 2 Inflow (Variable 3)	Tank 3 Inflow * (Variable 5)	TFM(2,4,2)	61.848 *	3.317456 *	52.368 *	2.8926 *
Tank 1 Inflow (Variable 1)	Tank 3 Inflow (Variable 5)	TFM(6,1,2)	114.887	6.828815	141.018	7.95437
Tank 3 Inflow (Variable 5)	Tank 4 Inflow (Variable 6)	TFM(0,1,0)	212.258	18.27315	259.44	21.63761
Tank 4 Inflow (Variable 6)	Tank 5 Inflow (Variable 8)	TFM(1,3,0)	373.345	47.85059	745.2495	66.45463
Tank 4 Inflow (Variable 6)	Tank 6 Inflow (Variable 11)	TFM(0,2,0)	2.402	2.195143	2.256	1.982629
Tank 1 Inflow (Variable 1)	Tank 5 Outflow B (Variable 10)	TFM(0,2,2)	10.218	2.023649	36.372	7.027344

Based on the results presented in Tables 4 and 6, the best models for each water tank inflow and outflow were selected based on the MAPE values; these models were used to develop the anomaly-detection methodology presented in Section 3.6. The shorthand notation for the model and the corresponding mathematical models, including the parameters and coefficients, are shown in Table 7.

Table 7. Mathematical models of the best ARIMA(p,d,q)(P,D,Q)(s) and TF(b, s, r) models for each variable in the analyzed water distribution branch, where: y_t is the output variable at time t, x_t is the input variable at time t, B is the backshift operator defined as $B^j Y_t = Y_{t-j}$, and ε_t is the error term at time t, assumed to be normally distributed with mean 0 and constant variance [24]. The variable number corresponds to the ones shown in Figure 3 and Table 2.

Variable Number	Variable	Model Notation	Written Model Including Parameters
1	Tank 1 Inflow	ARIMA(2,0,1)(0,1,1)(96)	$(1-1.1144B+0.1173B^2)(1-B^{96})y_t = (1-0.8911B)(1-0.9995B^{96})\varepsilon_t$
2	Tank 1 Outflow	ARIMA(0,1,1)(0,1,1)(96)	$(1-B)(1-B^{96})y_t = (1-0.2119B)(1-0.8294B^{96})\varepsilon_t$
3	Tank 2 Inflow	TFM(4,2,3)	$y_t = \frac{(1+0.39B-2B^2+0.68B^3)}{(1-1.87B+0.88B^2)}(1-B)B^4 x_t$
4	Tank 2 Outflow	ARIMA(0,1,0)(0,1,1)(96)	$(1-B)(1-B^{96})y_t = (1-0.6201B^{96})\varepsilon_t$
5	Tank 3 Inflow	TFM(2,4,2)	$y_t = \frac{(1-0.76B-0.94B^2)}{(1-0.8B-0.34B^2-0.81B^3+0.98B^4)}(1-B)B^2 x_t$
6	Tank 4 Inflow	ARIMA(1,0,0)(1,1,1)(96)	$(1-0.8729B)(1-0.1742B^{96})(1-B^{96})y_t = (1-0.902B^{96})\varepsilon_t$
7	Tank 4 Outflow	ARIMA(1,0,1)(0,1,1)(96)	$(1+0.3205B)(1-B^{96})y_t = (1+0.5807B)(1-0.9996B^{96})\varepsilon_t$
8	Tank 5 Inflow	ARIMA(1,1,2)(0,1,1)(96)	$(1-0.8741B)(1-B)(1-B^{96})y_t = (1-0.984B-0.016B^2)(1-0.8242B^{96})\varepsilon_t$
9	Tank 5 Outflow A	ARIMA(1,0,0)(0,1,1)(96)	$(1-0.8737B)(1-B^{96})y_t = (1-0.3874B^{96})\varepsilon_t$
10	Tank 5 Outflow B	ARIMA(1,1,0)(0,1,1)(96)	$(1-0.2493B)(1-B)(1-B^{96})y_t = (1-0.9992B^{96})\varepsilon_t$
11	Tank 6 Inflow	ARIMA(0,1,1)(0,1,1)(96)	$(1-B)(1-B^{96})y_t = (1-0.6446B)(1-0.9999B^{96})\varepsilon_t$

In the last stage, the new data were assessed, and any anomalies were found using the forecasting values that were produced. First, the top models for each variable were chosen from the previous step. These models match those displayed in Table 7. Then, the models were used to forecast over the short and medium term (i.e., one day and one week). Finally, the limits for assessing whether a new observation was an anomaly were the models forecast upper and lower 95% confidence intervals. Three categories, possible leakage, possible measurement mistake, and not available (NA) datapoint, were used to group the notifications.

Figure 15 presents the forecasting of the fitted models presented in Table 7 for each analyzed water flow variable. The orange line represents the model's forecast. The blue lines represent the new observations. The gray zone represents each model's 95% confidence interval. Table 8 presents the alerts generated for each variable, providing information on the final model utilized, the three types of potential alerts, and the total count of alerts. The results shown in Table 8 are based on the methodology presented in Figure 14.

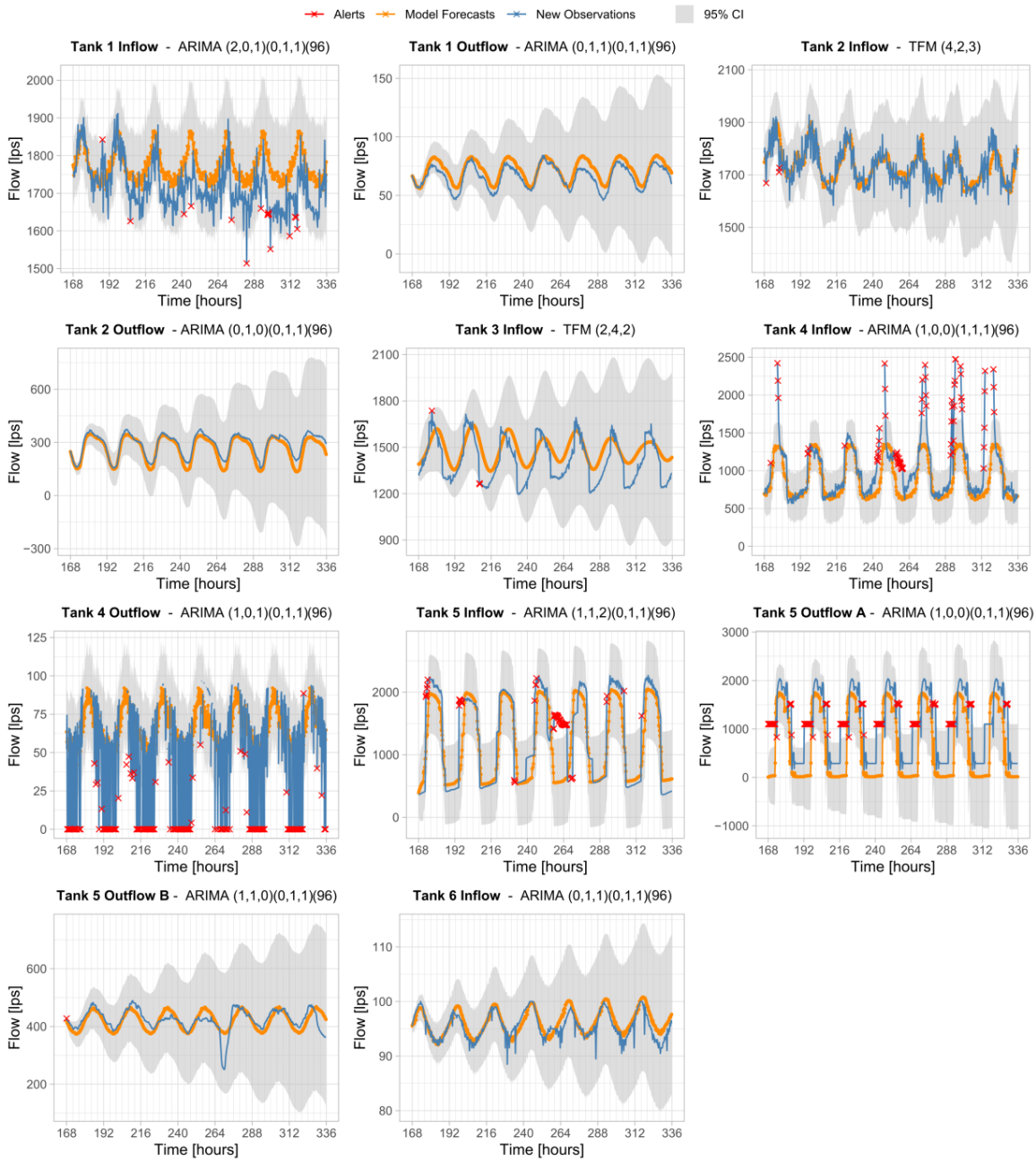

Figure 15. Water flow forecasts and alerts of all the selected models. The first row shows the model for Variables 1 to 3. The second row shows the model for Variables 4 to 6. The third row shows the models for Variables 7 to 9. The fourth row shows the models for Variables 10 and 11. The orange line represents the model's forecast values. The blue line represents the actual or new observations of the analyzed week. The gray zone represents the 95% confidence interval of each model. The red dots represent the observations that fall out of each model's 95% confidence interval gray area. These red dots are considered anomalies in the water flow behavior of the analyzed water distribution branch of Mexico City. The variable number corresponds to the ones shown in Figure 3 and Table 2.

Table 8. The count of three types of alerts and the total amount of alerts generated by each model in two forecast periods for the whole branch. The variable number corresponds to the ones shown in Figure 3 and Table 2.

n	Variable	Model	1 Day (96 Observations)				1 Week (672 Observations)			
			Possible Leakage	Possible Measurement Error	NA	Total Alerts	Possible Leakage	Possible Measurement Error	NA	Total Alerts
1	Tank 1 Inflow	ARIMA(2,0,1)(0,1,1)(96)	0	1	0	1	15	1	0	16
2	Tank 1 Outflow	ARIMA(0,1,1)(0,1,1)(96)	0	0	0	0	0	0	0	0
3	Tank 2 Inflow	TFM(4,2,3)	3	0	0	3	3	0	0	3
4	Tank 2 Outflow	ARIMA(0,1,0)(0,1,1)(96)	0	0	0	0	0	0	0	0
5	Tank 3 Inflow	TFM(2,4,2)	0	1	0	1	2	1	0	3
6	Tank 4 Inflow	ARIMA(1,0,0)(1,1,1)(96)	0	4	17	21	0	70	69	139
7	Tank 4 Outflow	ARIMA(1,0,1)(0,1,1)(96)	0	21	25	46	135	1	190	326
8	Tank 5 Inflow	ARIMA(1,1,2)(0,1,1)(96)	0	6	0	6	7	60	2	69
9	Tank 5 Outflow A	ARIMA(1,0,0)(0,1,1)(96)	1	26	0	27	3	144	0	147
10	Tank 5 Outflow B	ARIMA(1,1,0)(0,1,1)(96)	1	0	0	1	1	0	0	1
11	Tank 6 Inflow	ARIMA(0,1,1)(0,1,1)(96)	0	0	0	0	0	0	0	0
		Total of the branch	5	59	42	106	166	277	261	704

5. Discussion

The AICs, RMSEs, and MAPEs shown in Table 3 were computed by comparing the fitted models with the data observations (one week) used to generate the Seasonal ARIMA models. On the other hand, Table 4 shows the RMSEs and MAPEs obtained by comparing the fitted model with the observations used for forecasting. By comparing both tables, it can be appreciated that the errors were lower when comparing the model with the observations used for generating the models than the error obtained when comparing with the observations used for forecasting. Despite the above, the difference between the fit and prediction errors was low for most models. The smaller error obtained with the fit data compared to the prediction data suggested a slight overfit.

In addition, it can be appreciated that the order of the seasonal components of the fitted models was the same and that all of them required a first-order seasonal difference and moving average component. On the other hand, only the Tank 4 Inflow sensor data modeling required a first-order seasonal autoregressive component. In the case of the non-seasonal part of the fitted ARIMA models, it can be appreciated that certain heterogeneity existed in the order and the components that each of the input and output water flow variables required. The above could be attributed to the difference in the dynamics of the studied water tanks initially presented in Figure 2.

Moreover, based on the results presented in Table 4, it can be observed that the fitted Seasonal ARIMA models had a lower MAPE and RMSE for a one-day forecast than a one-week one. Hence, the models were better for forecasting in short periods. The above was useful to define the remodeling period and forecast for future use of the methodology. The only exception in which RMSE and MAPE were lower for the one-week forecast was the Seasonal ARIMA model fit for the Tank 4 Outflow. The above variable presented for both forecasts' periods a greater MAPE and RMSE compared to the rest of the water tanks. The new observed values for the next day were very extreme and differed from those used for the model generation. Nevertheless, the new observations were closer to the previous data during the week. This could mean that the sensor of this variable had issues and, in some periods, was not working accurately. Furthermore, Tank 5 Outflow did not contain a MAPE calculation because the nature of the variable data did not allow it; as can be seen in Table 2, the variable had minimum negative flow and positive maximum values, but some of the observed values were zero, so the division by zero in Formula (6) gets undefined.

In the case of the fitted TF models shown in Table 5, it can be appreciated that the order of the polynomials and the dead time operator that provided the lowest AIC value were different for each tank. Moreover, all TF models were influenced by past input values since they all had an sth-order polynomial. However, not all models were influenced by

the past values of the output since the rth-order polynomial equaled zero, as in the models for Tanks 4 to 6, as shown in Table 5. On the other hand, by comparing the errors of the TF models presented in Table 5 with the ones shown in Table 6, it can be appreciated that the errors of Table 5 were lower. The above was expected since the errors presented in Table 5 were computed with the same data used to fit the model. Despite the above, the errors were similar. Like the Seasonal ARIMA models, the obtained MAPE and RMSE values of the TF models, when used for forecasting, were generally lower for one-day forecasting than one-week forecasting, except for the first two models (see Table 6). Furthermore, only the Tank 3 Inflow TF model obtained an AIC value lower than its corresponding Seasonal ARIMA models (5536.81 and 5744.72, respectively). The above suggested a lower error while fitting and generating sparse models. However, even though the AIC values of the Seasonal ARIMA models were lower than the TF models, the first could provide an overfitted model when selected through the AIC [44].

Based on the results reported in Table 7, it can be appreciated that the water flow data can be modeled better by a Seasonal ARIMA according to the reported AIC, RMSE, and MAPE values. Moreover, as shown in Table 7, the TF models fitted through the Box–Jenkins methodology were less helpful in modeling the analyzed water branch's flow than the ARIMA models. In addition, it can be appreciated that, in general, the generated models had a low number of coefficients, which could facilitate a physical implementation of the proposed system since the computation that this type of model requires could be lower due to their low complexity.

Figure 15 shows the forecasting values of each final model presented in Table 7. The graphs present a one-week forecast, with the initial portion of the forecast (from Hour 168 to 172) representing the first forecast day, which served as the basis for the results reported in Table 7. Based on the forecasting values presented in Figure 15, the different dynamical behaviors of each flow variable of the analyzed branch can be appreciated visually. This visual representation enhances the understanding of and facilitates the alert process for users, enabling them to discern patterns and trends more effectively. In addition, it can be observed that the forecast values and observed values (actual measured water flow) were similar for most of the models. Furthermore, it is evident that the 95% confidence interval demonstrated a dynamic behavior across each model and did not remain constant throughout the analyzed forecasting week. In some cases, the gradual increase in the 95% confidence interval size could be attributed to either the absence of the moving average component in the model or the inherent increase in uncertainty as time progressed. These factors contributed to the widening of the confidence interval, indicating the challenges in accurately forecasting values over an extended period. In most cases, the selected model behaved similarly to the actual observations.

Finally, from Table 8, it can be seen that the variables with the worst (bigger) MAPE and RMSE errors generated the most alerts, as was the case for Tank 4 Outflow, Tank 4 Inflow, and Tank 5 Inflow, whose MAPEs were more prominent (more than 10%) from the rest of models and Tank 5 Outflow A had the biggest RMSE in both forecast periods (one day and one week). Tank 1 Inflow alerts could be considered a particular case because there were no alerts in one day, but there were many possible measurement errors in one week. The above could be due to the sensor malfunction over a long period or a change in data behavior, which might require remodeling. Although Figure 15 shows slight anomalies in the long-term behavior of Tank 3 Inflow, Tank 6 Inflow, and Tank 5 Outflow B, they were not detected, presented in Table 8, because they still fell within the 95% confidence interval. Therefore, the detection of alerts depends on the system's tuning, such as setting a smaller confidence interval or a shorter forecast period (e.g., 1 d, 12 h, etc.). Additionally, more potential measurement errors and not available observations were identified than possible leakages in both forecast periods, indicating the need to verify the calibration and proper functioning of sensors. This information is valuable in justifying investment in equipment maintenance and highlighting the affected areas.

The resulting ARIMA and TF models could be considered less-complex models than the ones produced by other machine learning algorithms proposed in the literature, such as XGBoost [18], ANNs [16], and CNNs [34], due to their reliance on a solid mathematical and statistical background with well-defined interpretations and the lower number of parameters that they have, as shown in Table 7. In addition, contrary to the studies presented in Section 2, Table 7 shows the coefficients and the mathematical representations of the fitted ARIMA and TF models. Moreover, the steps for modeling the ARIMA and TF models are well-defined and based on the specific assumptions of data stationarity, linearity, and the independence of residuals, providing a more-transparent framework and guidance in the modeling process. Finally, the models are suitable for forecasting based on historical patterns of water flow variables, making them a practical tool for anomaly detection in water distribution systems.

A comparison of the methodology presented in this study with the related literature presented initially in Section 2 in terms of the input data, machine learning algorithm, and analyzed system is presented in Table 9. In this regard, it can be observed that heterogeneity existed between the related research and the present study. Authors have proposed to detect leakages from pressure measurements, water flow measurements, acoustic emission, and vibration data; on the other hand, this study based its analysis exclusively on water flow data. Another aspect that can be observed is the frequent use of CNNs to perform leakage detection—other approaches involved tree-based techniques such as decision trees, random forest, XGBoost, and Adaboost. Otherwise, the present study focused on using Seasonal ARIMA and TF models that produce less-complex models than CNNs and tree-based classifiers. Finally, the analyzed systems varied from study to study, with water distribution networks being frequently analyzed. Public datasets and simulation tests of water pipelines have also been considered. In the case of the present work, the water flow data came from a branch of the water distribution system of Mexico City that supplies water to the sub-branch of water pipelines (see Figures 1 and 2).

Table 9. Comparison of the related research approaches with the methodology presented in this study.

Author	Input Data	Machine Learning Algorithm	Analyzed System
Guo et al. [15]	Piezoelectric accelerometers	Time–frequency CNN	Pipe networks from the city of Cheng Du, China
Moulik et al. [17]	3-axis accelerometer	K-means clustering	Laboratory pipeline system prototype
Choi et al. [30]	Sound vibration data magnitude spectra	CNN	AI Hub dataset composed of water leakage data from neighborhoods in Gwangju, Korea
Yu et al. [10]	Piezoelectric accelerometer data	SqueezeNet CNN	Pipe networks from cities in China including Shaoxing, Guangzhou, Sanya, Dalian, and Kunming
Fereidooni et al. [9]	Water flow sensor	Comparison between random forest, decision tree, Bayesian network, and K-nearest neighbor	Vitens company dataset of the water distribution networks of Leeuwarden City in Netherland
Chen et al. [11]	Landsat 8 satellite images	CNN	Water canal systems in Arizona
Sousa et al. [12]	Pressure measurements from pumps	K-means clustering and learning vector quantization	District-metered areas in Stockholm, Sweden
Shen et al. [31]	Acoustic emission signals	Adaboost	Water distribution networks of Jiangsu, Zhejiang, and Shanghai
Fares et al. [13]	Acoustic emission signals	Comparison between support vector machine, artificial neural network, and deep learning techniques.	Water distribution networks from Hong Kong

Table 9. *Cont.*

Author	Input Data	Machine Learning Algorithm	Analyzed System
Xue et al. [18]	Flowmeters and pressure sensors	XGBoost	Hydraulic simulation model
Taghlabi et al. [19]	Pressure values	Random forest	EPANET-Matlab simulation
Pérez-Pérez et al. [16]	Flow and pressure measurements	Artificial neural network	Laboratory water pipeline
Tornyeviadzi et al. [34]	Multivariate time series SCADA data	1D deep CNN autoencoder	L-TOWN water distribution network dataset
This study	Water flow sensor data	Modeling of water flow data via Seasonal ARIMA model	Input and output water flow of a water distribution branch of Mexico City
This study	Water flow sensor data	Modeling of water flow data via Transfer Function model	Input and output water flow of a water distribution branch of Mexico City

Nevertheless, it is difficult to perform a homogenous comparison of the results presented in this study with the related research. This is because the systems analyzed to develop the water-leakage-detection algorithms and the data collected varied from study to study. As previously stated, the authors proposed performing water leakage detection through simulation, laboratory tests, and data collected from water distribution systems. Nonetheless, each analyzed system had different data distributions, which impacted the type of algorithms that best describe the data. Furthermore, most of the works in the literature presented in Table 9, Section 2, and the Introduction Section employed supervised learning techniques to train the detection algorithms. On the other hand, this study tackled the problem from an unsupervised point of view since access to labeled data that classify the anomalies in water leakages or measurement errors were not available when developing this work. However, the above points out an area of opportunity that needs to be addressed in future work.

6. Limitations of the Study

This work was constrained by the branch's existing infrastructure and data availability, limiting it to a single case study focused solely on one operational variable, water flow. In future work, it is proposed to test the methodology in other cases of study and with other operational variables such as pressure, tank level, and more flow points. As shown in Figure 3, some tanks have other water exits, the flow rate of which was not available in this dataset, but are valuable variables that could help to generate a better model and understanding of water usage in the system. The use of additional variables could be performed with the help of multimodal techniques that consider flow time series data and pressure data measured in each of the tanks of the water branch. The above could be assessed through multimodal machine learning techniques such as model-agnostic (i.e., the fusion was carried out before applying the machine learning technique) and model-based (i.e., the fusion of the modalities was performed while generating the model) methods [45].

Furthermore, the methodology could not be validated with real leakages because a report of the actual leakages (detected or repaired) was unavailable when the models and the study were developed. The above implies another limitation, such as the need to develop a physical implementation of the proposed algorithms to validate anomalies and select and design an appropriate hardware platform in which the proposed algorithm can be embedded and executed. In addition, one of the crucial challenges of data-driven models such as ARIMA models is that, if the dynamics of the system changes, the fitted ARIMA models may not work as expected since the distribution of the data used for fitting could have changed. A potential solution to this problem is to fine-tune the models over time to keep their parameters updated in case of a change in the dynamics of the water distribution branch analyzed in this study.

Furthermore, due to the limited and incomplete dataset, this study focused on employing linear models such as the ARIMA and TF models; however, there might be non-linear

dynamics in the water distribution systems that the proposed methods could not capture. Hence, deep learning algorithms such as recurrent neural networks or long short-term memory neural networks could be compared with the proposed ARIMA and TF models in terms of performance. However, deep learning solutions often require a sizable sample size to be trained adequately and avoid overfitting problems. The above could be mitigated using transfer learning techniques in combination with deep learning solutions.

In addition, another potential disadvantage of the fitted ARIMA models is the need to filter the time series through differentiating; despite being essential to produce a stationary time series, it can also have certain biases related to the dynamics of the analyzed system since differentiating acts as a high-pass filter on the time series. The above could be mitigated with hybrid techniques combining nonstationary time series techniques such as wavelet analysis and ARIMA models. For instance, Nury et al. [40] proposed a wavelet-ARIMA model for temperature prediction in Bangladesh to account for the nonstationary behavior of the analyzed temperature time series data.

Additionally, due to the complexity of the water branch analyzed in this study, eleven flow measurements (i.e., considering the input and output flows of each tank) led to adjusting eleven models for the case of the fitted ARIMA models. The above is a potential drawback of the proposed anomaly detection system since it requires at least two models to detect anomalies for a single water tank. Thus, validating the proposed models could be a time-consuming task. Moreover, since the methodology used in this study depends extensively on data, its implementation is limited to water flow data availability. Hence, the approach presented in this work could be combined with hardware approaches to reduce data dependency.

Moreover, another approach that could have been developed is to generate a TF of the system shown in Figure 2 by considering the water tank's dynamics, as in the work of Li et al. [46]. The above could reduce the need for data to generate the TFs. However, certain variables, such as the height of the fluid present in the tanks, the area of the tanks, and the pressure, need to be considered to generate an adequate model that represents the system's dynamics and, consequently, the water flow behavior. Moreover, dynamical models based on linear differential equations do not consider the disturbances and noise the system is susceptible to. Hence, future work could compare estimating a TF based on the Box–Jenkins methodology and a TF obtained by considering the linear differential equations that describe the system dynamics.

Another opportunity that could be tackled is the need for developing a publicly available dataset from which the proposed water leakage detection models can be compared homogeneously. In the present study, sensor flow data were considered to analyze a water branch of the water distribution systems of Mexico City. However, other studies described in Section 2 used other types of data. They analyzed water distribution systems of other regions whose results cannot be compared to those presented in this work since the data modalities and distributions are different. The analyzed systems differed even among works that performed similar research in Mexico, such as Pérez-Pérez et al. [16] and the present study.

7. Conclusions

This work proposed using Seasonal ARIMA and TF models fit through the Box–Jenkins approach to model the flow data of a branch of the water distribution system of Mexico City for anomaly detection in water distribution branches. The results of this study showed that ARIMA models can describe and forecast the flow variables of the analyzed water branch with low error in terms of the MAPE. The generated TF models can also explain the linear branch system's relationship between tanks according to the reported RMSE and MAPE values. Still, in most cases, the ARIMA models achieved a higher performance in terms of the MAPE.

The models proposed in this study have the potential to make significant contributions to reducing water losses and improving the efficiency of the distribution system. These

improvements were achieved by utilizing the existing instrumentation and infrastructure of Mexico City's water distribution system, along with a clear and understandable methodology, visually representing anomalies to aid in the alert process for users. These models can facilitate early detection and localization of potential issues, enabling prompt actions and interventions for more-effective water distribution network management. Additionally, by identifying the specific sensor that triggered the alert, the search for potential issues can be narrowed down to a specific zone, enabling faster localization of the failure and more-efficient troubleshooting. The actions for each alert type (leakage or measurement error) depend on whether the water flow variable type is an inflow or outflow. For example, if less water than expected is arriving into a tank, it can be assumed that there is a leakage in the pipeline before the tank's entry. In such a case, physical inspection would be required.

On the other hand, if more water than expected is arriving, testing the sensor and verifying its calibration are recommended. In the case of an output water flow variable, if less water than expected is outgoing from a tank, it is possible that the leakage occurred inside the tank, such as an overflow or an unauthorized intake. If more water than expected is outgoing from the tank outflow, it is possible that there is a leakage in the pipe ahead, which should be verified.

The variables with greater errors were the ones with the most alarms. Therefore, the corresponding authorities should review and provide maintenance to these variables' sensors and communications systems. In addition, the generated alerts should be reviewed and validated by the operators responsible for the system to determine if new modeling is required or if the alerts are correct.

The current methodology's future work will improve it into an integrated support system with close collaboration between water service providers in Mexico City and action-based research. Implementation requires capable personnel with access to tank instrumentation to monitor and validate alerts constantly. Optimizing this methodology involves developing an online monitoring and detection system that reduces false alerts, detects leaks in real-time, and even remodels the system automatically. The platform should also provide dynamic data exchange and customer information for building positive relationships and pro-environmental attitudes.

Author Contributions: Conceptualization, D.B.-T. and R.B.-B.; methodology, D.B.-T. and J.L.P.-H.; software, D.B.-T.; validation, D.B.-T., R.B.-B. and J.L.P.-H.; formal analysis, D.B.-T.; investigation, D.B.-T.; resources, D.B.-T.; data curation, D.B.-T.; writing—original draft preparation, D.B.-T. and E.A.M.-R.; writing—review and editing, D.B.-T., E.A.M.-R. and S.A.N.-T.; visualization, D.B.-T.; supervision, R.B.-B.; project administration, D.B.-T. and R.B.-B.; funding acquisition, R.B.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from Tecnológico de Monterrey and CONACYT (CVU: 1080546).

Data Availability Statement: The data were provided by a telemetry company based in Mexico City named Virtual Wave Control (VWC). The original dataset is private and the company's property and was allowed to be used only for this work.

Acknowledgments: The authors would like to thank: CONACYT (CVU: 1080546), Tecnológico de Monterrey, VWC, and all the Reviewers for all the support and counsel.

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

Abbreviations

1D CNN	One-dimensional convolutional neural network
ACF	Auto-correlation function
AE	Autoencoder
AIC	Akaike's information criterion
ANN	Artificial neural networks
ARIMA	Autoregressive Integrated Moving Average
CNN	Convolutional neural network
DL	Deep learning
DMAs	District-metered areas
DT	Decision tree
KNN	K-nearest neighbor
lps	Liters per second
MA	Moving average
MAPE	Mean absolute percentage error
MFCC	Mel frequency cepstrum coefficient
ML	Machine learning
NA	Not available
PACF	Partial auto-correlation function
RMSE	Root-mean-squared error
SNR	Signal-to-noise ratio
SVM	Support vector machine
TF	Transfer Function
TFCNN	Time–frequency convolutional neural network
WDSs	Water distribution systems
XGBoost	Extreme gradient boosting

References

1. Mekonnen, M.M.; Hoekstra, A.Y. Four billion people facing severe water scarcity. *Sci. Adv.* **2016**, *2*, e1500323. [CrossRef] [PubMed]
2. United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420)*; United Nations: New York, NY, USA, 2019.
3. Aguirre, D.R.; Espinoza, V. *El Gran reto del Agua en la Ciudad de México*; Sistema de Aguas de La Ciudad de México: Mexico City, Mexico, 2012.
4. Pineda-Pablos, N.; Salazar-Adams, A. Cities and drought in Mexico. Water management as a mitigation critical strategy. *Tecnol. Cienc. Agua* **2016**, *7*, 95–113. Available online: https://www.cabdirect.org/cabdirect/abstract/20183068452 (accessed on 29 January 2021).
5. Ortega-Ballesteros, A.; Iturriaga-Bustos, F.; Perea-Moreno, A.J.; Muñoz-Rodríguez, D. Advanced Pressure Management for Sustainable Leakage Reduction and Service Optimization: A Case Study in Central Chile. *Sustainability* **2022**, *14*, 12463. [CrossRef]
6. Mashhadi, N.; Shahrour, I.; Attoue, N.; El Khattabi, J.; Aljer, A. Use of machine learning for leak detection and localization in water distribution systems. *Smart Cities* **2021**, *4*, 1293–1315. [CrossRef]
7. Sun, C.; Parellada, B.; Puig, V.; Cembrano, G. Leak localization in water distribution networks using pressure and data-driven classifier approach. *Water* **2019**, *12*, 54. [CrossRef]
8. Ares-Milián, M.J.; Quiñones-Grueiro, M.; Verde, C.; Llanes-Santiago, O. A leak zone location approach in water distribution networks combining data-driven and model-based methods. *Water* **2021**, *13*, 2924. [CrossRef]
9. Fereidooni, Z.; Tahayori, H.; Bahadori-Jahromi, A. A hybrid model-based method for leak detection in large scale water distribution networks. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 1613–1629. [CrossRef]
10. Yu, T.; Chen, X.; Yan, W.; Xu, Z.; Ye, M. Leak detection in water distribution systems by classifying vibration signals. *Mech. Syst. Signal Process.* **2023**, *185*, 109810. [CrossRef]
11. Chen, J.; Tang, P.; Rakstad, T.; Patrick, M.; Zhou, X. Augmenting a deep-learning algorithm with canal inspection knowledge for reliable water leak detection from multispectral satellite images. *Adv. Eng. Inform.* **2020**, *46*, 101161. [CrossRef]
12. Sousa, D.P.; Du, R.; Mairton Barros da Silva, J., Jr.; Cavalcante, C.C.; Fischione, C. Leakage detection in water distribution networks using machine-learning strategies. *Water Supply* **2023**, *23*, 1115–1126. [CrossRef]
13. Fares, A.; Tijani, I.A.; Rui, Z.; Zayed, T. Leak detection in real water distribution networks based on acoustic emission and machine learning. *Environ. Technol.* **2022**, 1–17. [CrossRef]
14. Bykerk, L.; Valls Miro, J. Detection of Water Leaks in Suburban Distribution Mains with Lift and Shift Vibro-Acoustic Sensors. *Vibration* **2022**, *5*, 21. [CrossRef]

15. Guo, G.; Yu, X.; Liu, S.; Ma, Z.; Wu, Y.; Xu, X.; Wang, X.; Smith, K.; Wu, X. Leakage detection in water distribution systems based on time–frequency convolutional neural network. *J. Water Resour. Plan. Manag.* **2021**, *147*, 04020101. [CrossRef]
16. Pérez-Pérez, E.D.J.; López-Estrada, F.R.; Valencia-Palomo, G.; Torres, L.; Puig, V.; Mina-Antonio, J.D. Leak diagnosis in pipelines using a combined artificial neural network approach. *Control. Eng. Pract.* **2021**, *107*, 104677. [CrossRef]
17. Moulik, S.; Majumdar, S.; Pal, V.; Thakran, Y. Water Leakage Detection in Hilly Region PVC Pipes using Wireless Sensors and Machine Learning. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics—Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020; pp. 1–2. [CrossRef]
18. Xue, P.; Jiang, Y.; Zhou, Z.; Chen, X.; Fang, X.; Liu, J. Machine learning-based leakage fault detection for district heating networks. *Energy Build.* **2020**, *223*, 110161. [CrossRef]
19. Taghlabi, F.; Sour, L.; Agoumi, A. Prelocalization and leak detection in drinking water distribution networks using modeling-based algorithms: A case study for the city of Casablanca (Morocco). *Drink. Water Eng. Sci.* **2020**, *13*, 29–41. [CrossRef]
20. Rai, A. Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141. [CrossRef]
21. Li, H.; Li, J.; Guan, X.; Liang, B.; Lai, Y.; Luo, X. Research on overfitting of deep learning. In Proceedings of the 2019 15th International Conference on Computational Intelligence and Security (CIS), Macao, China, 13–16 December 2019; pp. 78–81.
22. Elsaraiti, M.; Merabet, A. A Comparative Analysis of the ARIMA and LSTM Predictive Models and Their Effectiveness for Predicting Wind Speed. *Energies* **2021**, *14*, 6782. [CrossRef]
23. Viccione, G.; Guarnaccia, C.; Mancini, S.; Quartieri, J. On the use of ARIMA models for short-term water tank levels forecasting. *Water Supply* **2020**, *20*, 787–799. [CrossRef]
24. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
25. Van der Walt, J.C.; Heyns, P.S.; Wilke, D.N. Pipe network leak detection: Comparison between statistical and machine learning techniques. *Urban Water J.* **2018**, *15*, 953–960. [CrossRef]
26. Liu, Y.; Ma, X.; Li, Y.; Tie, Y.; Zhang, Y.; Gao, J. Water Pipeline Leakage Detection Based on Machine Learning and Wireless Sensor Networks. *Sensors* **2019**, *19*, 5086. [CrossRef]
27. Quy, T.B.; Kim, J.M. Leakage Detection of Water-Induced Pipelines Using Hybrid Features and Support Vector Machines. In *Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing*; Bhatia, S., Tiwari, S., Mishra, K., Trivedi, M., Eds.; Springer: Singapore, 2019; Volume 924. [CrossRef]
28. Vrachimis, S.G.; Eliades, D.G.; Taormina, R.; Kapelan, Z.; Ostfeld, A.; Liu, S.; Kyriakou, M.; Pavlou, P.; Qiu, M.; Polycarpou, M.M. Battle of the leakage detection and isolation methods. *J. Water Resour. Plan. Manag.* **2022**, *148*, 04022068. [CrossRef]
29. Islam, M.R.; Azam, S.; Shanmugam, B.; Mathur, D. A Review on Current Technologies and Future Direction of Water Leakage Detection in Water Distribution Network. *IEEE Access* **2022**, *10*, 107177–107201. [CrossRef]
30. Choi, J.; Im, S. Application of CNN Models to Detect and Classify Leakages in Water Pipelines Using Magnitude Spectra of Vibration Sound. *Appl. Sci.* **2023**, *13*, 2845. [CrossRef]
31. Shen, Y.; Cheng, W. A Tree-Based Machine Learning Method for Pipeline Leakage Detection. *Water* **2022**, *14*, 2833. [CrossRef]
32. Cody, R.A.; Narasimhan, S. A field implementation of linear prediction for leak-monitoring in water distribution networks. *Adv. Eng. Inform.* **2020**, *45*, 101103. [CrossRef]
33. Fabbiano, L.; Vacca, G.; Dinardo, G. Smart water grid: A smart methodology to detect leaks in water distribution networks. *Measurement* **2020**, *151*, 107260. [CrossRef]
34. Tornyeviadzi, H.M.; Seidu, R. Leakage detection in water distribution networks via 1D CNN deep autoencoder for multivariate SCADA data. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106062. [CrossRef]
35. Kammoun, M.; Kammoun, A.; Abid, M. Experiments based comparative evaluations of machine learning techniques for leak detection in water distribution systems. *Water Supply* **2021**, *22*, 628–642. [CrossRef]
36. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 112, p. 18.
37. Schaffer, A.L.; Dobbins, T.A.; Pearson, S.A. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: A guide for evaluating large-scale health interventions. *BMC Med. Res. Methodol.* **2021**, *21*, 58. [CrossRef]
38. ArunKumar, K.E.; Kalaga, D.V.; Kumar, C.M.S.; Chilkoor, G.; Kawaji, M.; Brenza, T.M. Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). *Appl. Soft Comput.* **2021**, *103*, 107161. [PubMed]
39. Mestre, G.; Portela, J.; Rice, G.; San Roque, A.M.; Alonso, E. Functional time series model identification and diagnosis by means of auto-and partial autocorrelation analysis. *Comput. Stat. Data Anal.* **2021**, *155*, 107108. [CrossRef]
40. Nury, A.H.; Hasan, K.; Alam, M.J.B. Comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh. *J. King Saud Univ.-Sci.* **2017**, *29*, 47–61. [CrossRef]
41. Katoch, R.; Sidhu, A. An application of ARIMA model to forecast the dynamics of COVID-19 epidemic in India. *Glob. Bus. Rev.* **2021**. [CrossRef]
42. Helmer, R.M.; Johansson, J.K. An exposition of the Box-Jenkins transfer function analysis with an application to the advertising-sales relationship. *J. Mark. Res.* **1977**, *14*, 227–239. [CrossRef]

43. von Asmuth, J.R.; Bierkens, M.F.; Maas, K. Transfer function-noise modeling in continuous time using predefined impulse response functions. *Water Resour. Res.* **2002**, *38*, 23-1–23-12. [CrossRef]
44. DelSole, T.; Tippett, M.K. Correcting the corrected AIC. *Stat. Probab. Lett.* **2021**, *173*, 109064. [CrossRef]
45. Barua, A.; Ahmed, M.U.; Begum, S. A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions. *IEEE Access* **2023**, *11*, 14804–14831. [CrossRef]
46. Li, X.; Li, Z. The application of linear and nonlinear water tanks case study in teaching of process control. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *113*, 012165. [CrossRef]