

Development of Cross-Domain Artificial Neural Network to Predict High-Temporal Resolution Pressure Data

Young Hwan Choi ¹ and Donghwi Jung ^{2,*}

¹ Department of Civil and Architectural Engineering and Mechanics, University of Arizona, Tucson, AZ 85721, USA; younghwanchoi@email.arizona.edu

² School of Civil, Environmental and Architectural Engineering, Korea University, Seoul 02841, Korea

* Correspondence: donghwiku@gmail.com

Received: 3 April 2020; Accepted: 6 May 2020; Published: 8 May 2020

Abstract: Forecasting hydraulic data such as pressure and demand in water distribution system (WDS) is an important task that helps ensure efficient and accurate operations. Despite high-performance data prediction, missing data can still occur, making it difficult to effectively operate WDS. Though the pressure data are directly related to the rules of operation for pumps or valves, few studies have been conducted on pressure data forecasting. This study proposes a new missing and incomplete data control approach based on real pressure data for reliable and efficient WDS operation and maintenance. The proposed approach is: (1) application of source data from high-resolution, real-world pressure data; (2) development of a cross-domain artificial neural network (CDANN), combining the standard artificial neural networks (ANNs) and the cross-domain training approach for missing data control; and (3) analysis of standard data mining according to external factors to improve prediction accuracy. To verify the proposed approach, a real-world network located in South Korea was used, and the forecasting results were evaluated through performance indicators (i.e., overall, special points, and percentage errors). The performance of the CDANN is compared with that of standard ANNs, and CDANN was found to provide better predictions than traditional ANNs.

Keywords: water distribution system; missing data control; pressure data prediction; cross-domain artificial neural network; data categorization standard

1. Introduction

Since the advent of the Fourth Industrial Revolution, techniques for estimating demand- and pressure-based data are improving for facets of water distribution system (WDS) including planning, design, operation, and strategic decisions [1–4]. For example, WDS operators need to know the magnitude and pattern of future user demand. This information is important for proactively and efficiently satisfying user demands for reservoirs, water treatment plants, and pump stations [5,6]. WDS also need to predict the water demand 20–30 years into the future in order to develop new water sources and/or expand existing water treatment plants.

Past data-driven technology was focused on demand forecasting techniques for the expansion of WDS, which were necessary to determine the size and layout of the systems for reliable and realistic planning and design [7]. Previous studies have attempted to develop stochastic models for data forecasting. Most of the WDS data are time series that can exhibit more complex profiles through comparison with other infrastructure data. Stochastic process models, which can be formulated in discrete- or continuous-time, are more advanced alternatives that can be used to model these complex

profiles. Traditionally, auto-regressive integrated moving average (ARIMA)-based models have been used for understanding and modeling the WDS demand. ARIMA-based models typically solve the problem as a linear correlation among variables [8–10]. Billings and Jones [2] used these models and applied the mathematical formulations of processes that obeyed specific probabilistic and statistical laws; thus, their simulated forecasts resulted in a series of outcomes for each period over a given time span. The value of stochastic models in forecasting demand data lies in their ability to quantify estimates of the level of uncertainty associated with forecast values. However, these techniques did not always produce predictions with sufficient accuracy. To mitigate this problem, several advanced data forecasting models have been applied more recently, such as artificial intelligence (AI) approaches. Artificial neural networks (ANNs) and fuzzy logic techniques of forecasting water demand are advanced methods that are classified as nonparametric approaches [2,11–13], which are applied to both long- and short-term demand forecasting. Herrera et al. [14] performed a comprehensive comparison of various predictive methods for hourly water demand forecasting, suggesting the use of support vector regression (SVR) as one of the models through which it is possible to achieve better results. Furthermore, the application of ANN models for water demand forecasting has typically involved comparing the performances of ANN models with those of conventional regression models [14–17] and time series analysis models [17].

For the aforementioned studies, forecasting water demand were essential at the infrastructure development stage. However, major infrastructure systems, including water utilities, have recently been installed and operated in most urban areas; hence, research on the optimal operation and maintenance of WDS is essential. Furthermore, to enable safe operation and management of systems, and to perform effective valve and pump maneuvers, water utilities need to be acquainted with local real-time end-user behavior regarding water consumption [18]. Therefore, hereafter, the focus of data forecasting studies should switch from planning and design to operation and maintenance of WDS.

Therefore, data forecasting studies have recently been conducted for efficient system operation and maintenance of water utilities [19–23]. These studies have performed water demand or pressure forecasting that generated theoretical synthetic data for real-time pump operation. It is well-known that though water demand is an unknown variable, it can be estimated via theoretical forecasting or past demand-trend analysis, and the nodal pressure can be calculated using the estimated water demand by one of the available WDS hydraulic solvers (i.e., EPANET). However, as this pressure value is also a simulation result and not a measured value, the application of pressure data, when sufficient field measurements are available, is necessary for the accurate and efficient operation and management of WDS. Therefore, for optimal WDS component (e.g., pump, valve) operation, utilizing water pressure data, especially the measured water pressure or the forecasting data estimated by realistic pressure [24], is better than using water demand.

Recently, advanced sensor technologies have been expanding the development of demand and pressure estimation techniques with measurements from advanced sensors (e.g., advanced metering infrastructure (AMI)). Therefore, applying data forecasting techniques that consider uncertainty provides a basis for accurately quantifying infrastructure such as the WDS. Furthermore, the risk of water shortages and revenue losses can be reduced, thereby enabling the optimization of operational and investment decisions. However, the data gathering process often produces incomplete or missing values for various reasons such as interference in the network connection, malfunction of the data collector, or sensor failure, leading to data scarcity [25].

Especially while managing water resources in small- to medium-sized utilities, incomplete data present serious challenges for the development and operation of water infrastructure failure-prediction models [26]. Moreover, most of the regression-based water system models assume that input is provided as a complete data matrix. In the case of the dataset having missing values (one or more), the models would perform deletions listwise or pairwise, or substitute missing values with mean values [3,22]. In addition, to ensure preventive maintenance, repair, or replacement of the water systems, the researchers proposed different risk assessment frameworks using different methods such as the analytic hierarchy process, fuzzy expert system, artificial neural network, multicriteria decision analysis, and proportional hazard model [3,27–32].

In other engineering fields, approaches to dealing with incomplete and missing data have been actively researched. Acuna and Rodriguez [33] compared three imputation methods (i.e., mean imputation, median imputation, and k-NN imputation) using twelve datasets and two classifiers. Kornelsen and Coulibaly [34] demonstrated data-driven approaches against conventional infilling techniques (i.e., the statistical and interpolation infilling approaches) for the imputation of missing values in a distributed soil moisture dataset. Inman et al. [35] compared two imputation approaches (the cubic spline and multiple imputations) and two clustering techniques (autocorrelation-based fuzzy clustering and wavelet-based clustering) on the electrical demand data of a commercial building. Nelwamondo et al. [36] developed the expectation maximization (EM) algorithm, which was combined with the auto-associative neural network and genetic algorithm (GA), to solve the problem of missing data imputation. Sim et al. [37] applied and compared several imputation models (i.e., the performance of listwise deletion, mean imputation, group mean imputation, predictive mean imputation, hot-deck, k-NN, and k-means clustering) in the hypothetical computing application dataset to identify the best approach. However, these approaches estimate that if any value is missing, it is assumed to be zero or the representative values (e.g., mean value) that are considered the neighborhood values in the training process. The drawback of these techniques is that they lead to a severe loss of information, hinder the model accuracy, and introduce decision making biases [38,39].

Moreover, traditional forecasting approaches using AI techniques have assumed that the training and test data are drawn from the same data distribution; thus, the data are not suitable for addressing situations where new unlabeled data are obtained or training data are insufficient [40]. If the distribution of training data has similar trends, the above problems regarding the lack of training data can be addressed by increasing the dataset using synthetic data generation methods (e.g., linear regression, parallel data generation, random but deterministic, obfuscated data) [41–44]. Hence, for optimal WDS operation and maintenance, effective and appropriate approaches for dealing with the missing values of the database are essential.

Therefore, this study proposes a methodology to handle missing or incomplete pressure data for efficient WDS operation. Toward this objective, this study applies three schemes. First, to increase the accuracy of the pressure data prediction under missing data or limited number of data conditions, this study applies data forecasting based on real data as source data. The applied source data were measured from three pressure meters at one-minute intervals in real-world WDS located in South Korea. Second, to improve the forecasting performance, this study proposes a new approach to control missing data, called the cross-domain artificial neural network (CDANN), combining the standard artificial neural networks and the cross-domain training approach [45], covering missing source data by replacing the target data with those generated from one or more source data. Third, because the performance of data prediction differs depending on the type and category of training data, data mining was performed before the data forecasting process, incorporating factors affecting water pressure. The data from a group of pressure meters are compared with the forecasting performance according to training data, considering various characteristics such as the day of the week, time, and temperature. The proposed pressure data forecasting approach can be applied for effective operation in real-world WDS that do not have sufficient number of installed pressure meters.

2. Pressure Data Forecasting Model

The proposed pressure data forecasting model is based on methodology in order to generate reliable forecast data through various training combinations using limited observed data, which are difficult to obtain owing to space and budget constraints in WDS. Toward this objective, this chapter first introduces the region in which the present study was conducted and the characteristics of the real data; then, it describes various combinations of training data according to the data characteristics. Finally, the combined ANN and cross-domain training approach, as well as the forecasting methods in this study, are discussed along with performance measurements.

2.1. Pressure Variation in the Study Area through EPANET

In this study, to verify the proposed pressure forecasting model, the Galsan network (Figure 1) in Seosan-si, South Korea, has been applied. This network consists of 1 pressure zone, 88 pipes, and 88 nodes, and it supplies a flow rate of approximately 0.00288 (m³/s) in an area of 1.57 km². The altitude of this area is approximately 30–60 m, as most of this area is mountainous terrain; the highlands and lowlands are mixed. For this reason, if the water does not sufficiently pressurize in the highlands area, the minimum required pressure cannot be satisfied; however, with sufficient pressurization in the highlands, a high-pressure section is generated in the lowlands.

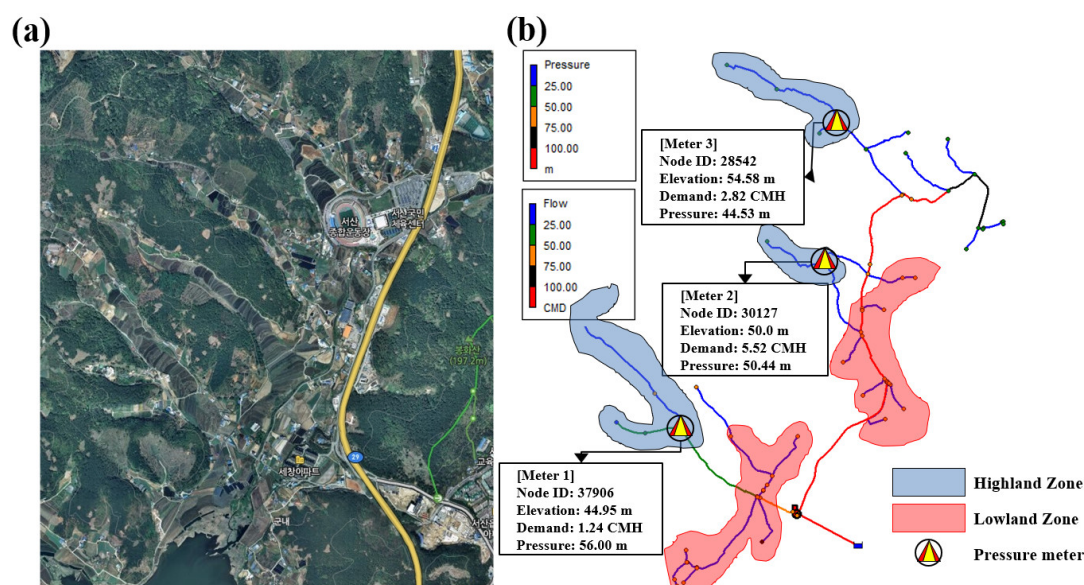


Figure 1. Description of the Galsan network: (a) a topographical map (from Google Maps) and (b) the network configuration.

Therefore, as illustrated in Figure 1b, three pressure meters were installed in this area for effective pressure control to satisfy the pressure constraints in the high- and low-pressure zones. To choose the location of the pressure meters, hydraulic modeling was performed using EPANET considering the variance of pressure. The pressure meters were installed where variation of the end node pressure can be effectively measured. Four data collection devices, one LTE router, and three WCDMA were installed at the same locations. The observed pressure data had high resolution. Data were collected at 1-min intervals, totaling approximately 1,296,000 observations (10 months) from May 2019.

Table 1 is an example of the pressure data collected. However, to effectively operate WDS, the system requires more data from various measured points than was obtained from the three pressure meters because this area has a high elevation. Therefore, this study proposes an unknown pressure data forecasting model using data from three real pressure meters.

Table 1. Example of observed pressure data in the Galsan network.

Date, Time	Pressure (m)			
	Meter 1	Meter 2	Meter 3	Average
2019-10-23, 00:01	5.080	3.152	4.107	4.113
2019-10-23, 00:02	4.952	3.091	4.040	4.028
2019-10-23, 00:03	5.024	2.978	4.002	4.001
...				
2019-10-23, 14:23	4.945	3.101	4.059	4.036
2019-10-23, 14:24	5.019	3.046	4.028	4.031

2019-10-23, 14:25	4.955	2.965	3.956	3.959
...			...	
2019-10-23, 23:58	5.131	3.238	4.037	4.135
2019-10-23, 23:59	5.106	3.193	4.158	4.152

2.2. Various Combinations of Training Data

The operation of WDS involves control of reservoirs, pumps, and valves according to consumer demand patterns, which, in the planning stages, depend on the average water consumption per customer and water consumption trends (e.g., in industrial, commercial, and residential areas). However, these demands are estimated by similar water consumption locations or past consumption patterns, and actual water consumption is different from estimated values. Therefore, the pressure data forecasting as well as demand forecasting is essential for effective real-time operation of WDS. Furthermore, in previous studies, the pattern of demand has been known to fluctuate on monthly, weekly, and daily scales. Additionally, the hydrological factors such as the trend of greater consumption in summer than in winter cause changes in water demand. As the water pressure in WDS hydraulically relates to water demand, this study evaluates the temporal effects (e.g., month, day of the week, time of day, and season) and determines the most appropriate influencing factors for the optimal pressure forecast. Therefore, this chapter introduces the standard of data mining according to the various combinations of training data in the forecasting approach used in this study.

Training dataset 1: The first training dataset is for the same day of the week. Among the data obtained through the three pressure meters, the results of the analysis of the pressure trend for two weeks in October are shown in Figure 2. The average pressure values on the same days of the week are slightly different; however, the trends are similar. This analysis reveals that, in accordance with previous studies related to water demand forecasting, the pressure variations also show similar changes over the days of the week. Therefore, this dataset includes training data from the same day of the week for effective data forecasting.

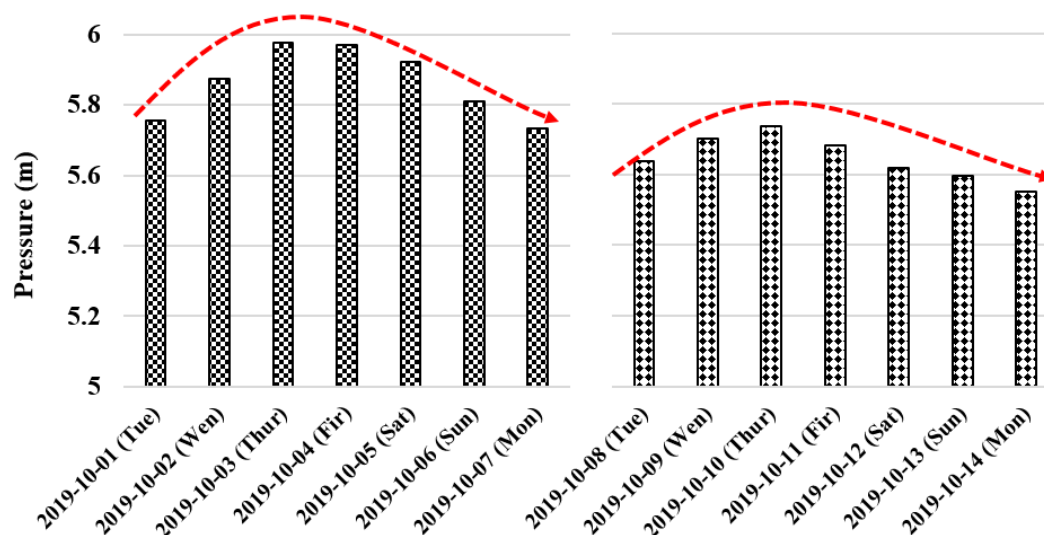


Figure 2. Example of Galsan network average water pressure data in the 1st and 2nd weeks of October

Training dataset 2: The second training dataset considers the same time period over a day. Dataset 1 considers similar characteristics of water consumption trends on the same day of the week. Dataset 2 considers similar characteristics of water consumption, even if the day of the week is different when water is used in the same time period (Figure 3). In addition, as the hydraulic variations of WDS possess the time series characteristic wherein the hydraulic results of the previous analysis are influenced at a later time, training using data of the same time period is effective for forecasting unknown data. Therefore, the data combination of the same time period divides the day

into four periods (Period 1: 00:00–06:00, Period 2: 06:00–12:00, Period 3: 12:00–18:00, and Period 4: 18:00–24:00) for use as training data.

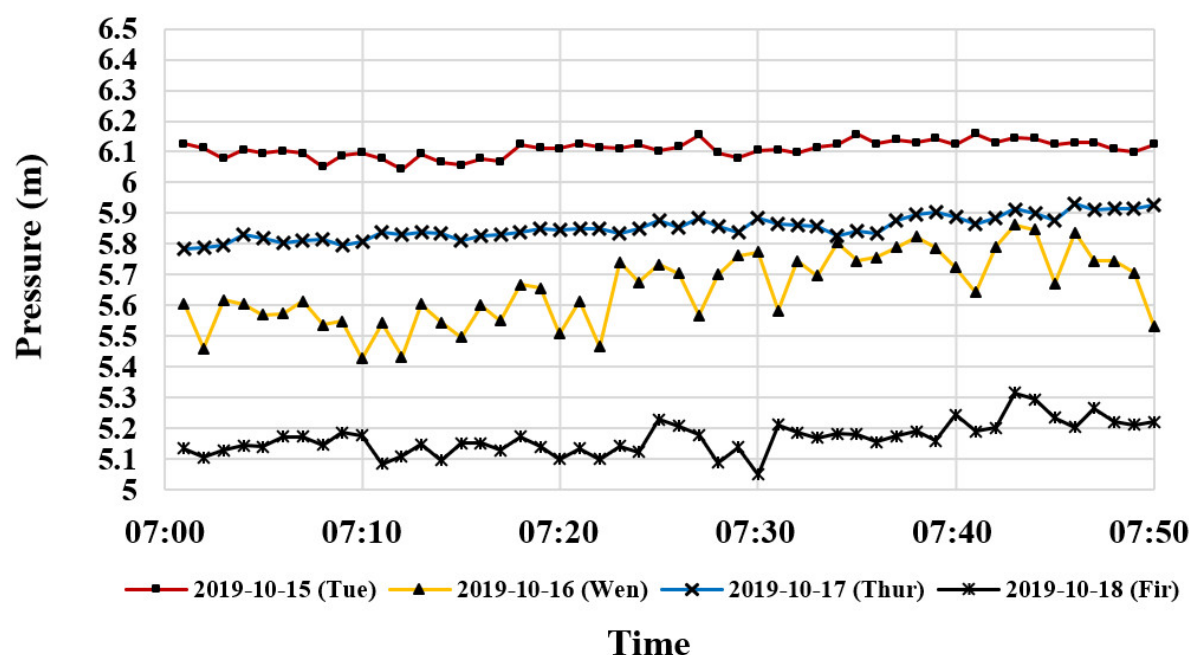


Figure 3. Example of Galsan network water pressure data for Period 2 in October.

Training dataset 3: This set includes data training according to season. The season focuses on temperature, and Gibbs [46] showed that daily average temperature affects water consumption. Particularly in water consumption, there is no significant change at 0–20°C, although water consumption increases significantly above 20°C [11,47]. Therefore, the data training is performed using data combinations according to temperature. These combinations group days with average temperature <5°. Each combination is divided into four steps from 0–25°C and is used as training data.

2.3. Cross-Domain Artificial Neural Network

The data forecasting approach implemented in this study is a cross-domain artificial neural network (CDANN). The standard artificial neural network (ANN) is a powerful computational model under the explicit data condition, wherein the variables involved in the unknown data have complex non-linear relationships with each other [48,49]. Typically, the ANN consists of the input layer, where the data are inputted into the model as training data; the hidden layer(s), which perform weighting and data processing; and finally, the output layer, where the results of the ANN are produced. Each layer consists of basic elements called neurons. The neuron is a non-linear algebraic function, parameterized with boundary values [50]. The signals passing through the neurons are modified by weights and transfer functions. This process is repeated until the output layer is reached. The input, hidden, and output layers are the parameters of the ANN, and the parameters are applied depending on the problem. If the number of hidden neurons is less than ideal, the network cannot learn the process correctly, and if there are excessive neurons, then training requires a longer time duration, and the network may over-fit [51,52]. However, in the training step, if the amount of input data are very little or the correlation between source data and target data are very low, the model may not be able to be trained correctly, even with a large number of hidden neurons. In the data used in this study, the correlation between source data and target data is high, and the form is similar. Therefore, to address the problems that may occur because of the characteristics of the input variables, the CDANN, which combines the cross-domain training approach with an artificial neural network, is used as the forecasting model in our study.

The cross-domain training approach can be applied when the source dataset (D^s) and target dataset (D^t) are strongly correlated. This approach has generally been used for image processing [40,53,54]; however, recently, the trends of input and output have been applied to similar engineering problems such as pulse data [55]. Feature replication combines all samples from both D^s and D^t and attempts to learn the generalities between the two datasets by replicating parts of the original feature vector, x_i , for different domains, following equation 1:

$$X_i^s = [x_1, x_2, \dots, x_i], x_i \in D^s \quad X_i^t = [x_1^*, x_2^*, \dots, x_i^*], x_i^* \in D^t \quad (1)$$

Equation 2 is an example of the data combination for the CDANN. If the given data are X_1^s , X_2^s , X_1^t , and X_2^t with the same data distribution, these data are replicated using the cross-domain approach to X_1^{s*} , X_2^{s*} , X_1^{t*} , and X_2^{t*} . Subsequently, the replicated data and original data are composed of the data combination. The comparison of the data combination method between the ANN and CDANN is as follows. While the standard ANN combines the input and target data as given by $(X_1^s - X_2^s)$ and $(X_1^t - X_2^t)$, the CDANN combines the original data and replica until all combinations are generated for the input or target data such as $(X_1^s - X_1^t)$, $(X_1^s - X_2^t)$, $(X_2^s - X_1^t)$, $(X_2^s - X_2^t)$, $(X_1^t - X_1^s)$, $(X_1^t - X_2^s)$, $(X_2^t - X_1^s)$, and $(X_2^t - X_2^s)$.

$$\begin{aligned} \text{Given data} &= \begin{cases} X_1^s = [a, b, c], X_2^s = [d, e, f] \\ X_1^t = [g, h, i], X_2^t = [j, k, l] \end{cases} \\ \text{Generating Replica} &= \begin{cases} X_1^{s*} = [a^*, b^*, c^*], X_2^{s*} = [d^*, e^*, f^*] \\ X_1^{t*} = [g^*, h^*, i^*], X_2^{t*} = [j^*, k^*, l^*] \end{cases} \end{aligned}$$

Data combination with ANN and CDANN

$$\begin{aligned} \text{Standard ANN} &= \begin{pmatrix} X_1^s = [a, b, c] \\ X_1^t = [g, h, i] \end{pmatrix}, \begin{pmatrix} X_2^s = [d, e, f] \\ X_2^t = [j, k, l] \end{pmatrix} \quad (2) \\ \text{CDANN} &= \begin{pmatrix} X_1^s = [a, b, c] \\ X_1^{t*} = [g^*, h^*, i^*] \end{pmatrix}, \begin{pmatrix} X_1^s = [a, b, c] \\ X_2^{s*} = [d^*, e^*, f^*] \end{pmatrix}, \begin{pmatrix} X_2^s = [d, e, f] \\ X_1^{t*} = [g^*, h^*, i^*] \end{pmatrix}, \begin{pmatrix} X_2^s = [d, e, f] \\ X_2^{s*} = [d^*, e^*, f^*] \end{pmatrix}, \begin{pmatrix} X_1^t = [g, h, i] \\ X_1^{t*} = [g^*, h^*, i^*] \end{pmatrix}, \begin{pmatrix} X_1^t = [g, h, i] \\ X_2^{t*} = [j^*, k^*, l^*] \end{pmatrix}, \begin{pmatrix} X_2^t = [j, k, l] \\ X_1^{t*} = [g^*, h^*, i^*] \end{pmatrix}, \begin{pmatrix} X_2^t = [j, k, l] \\ X_2^{t*} = [j^*, k^*, l^*] \end{pmatrix} \end{aligned}$$

The cross-domain training process can be useful in the case of fewer input variables or when the input and output show a similar trend. The structure of the CDANN is composed of the standard ANN and a cross-domain pre-training process, as shown in Figure 4.

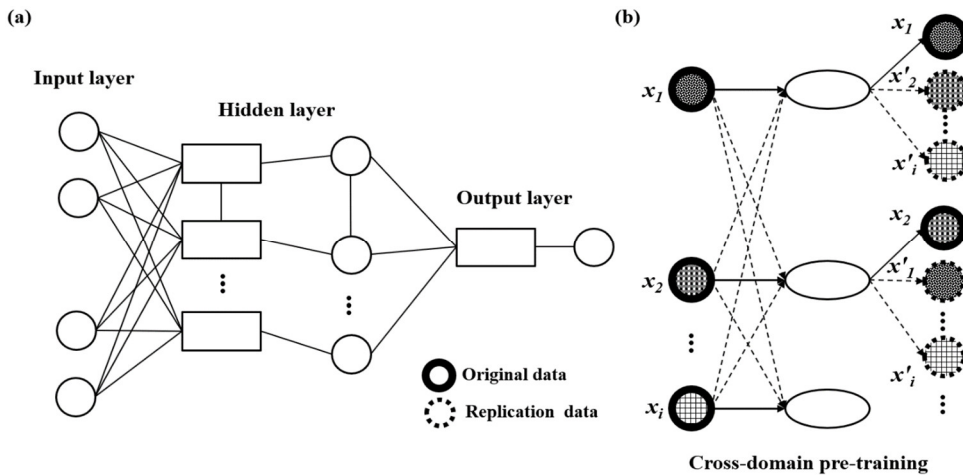


Figure 4. Concept of the cross-domain artificial neural network (CDANN): (a) artificial neural network and (b) cross-domain pre-training.

The training process of the standard ANN distributes the error to arrive at a best fit or minimum error. However, if the amount of inputs and target data are insufficient, the model cannot be sufficiently trained. Therefore, for the cross-domain pre-training technique, the number of input variables can be increased by creating replicas of the user input variables and combining them. This process improves the accuracy of training, compared to the normal training process, by increasing the amount of training data when there is a high correlation between input data and target data. After generating the new input variable through cross-domain pre-training, the process of the standard ANN is followed. Information passes through the network in a forward direction, the network predicts an output, and minimization of error is achieved through several iterations.

2.4. Performance Measures and Error Evaluation

Performance evaluation parameters measure the forecast accuracy and help develop a more robust model by modifying the existing parameters or model formulation to reduce errors in forecasts. The predictive performance is evaluated by comparing the observed and forecasted pressure data. When comparing the forecasting performances for various parameter options or models, the model with the least performance value is considered the most accurate. However, as various performance measures have different characteristic forecasting errors (e.g., overall error, special points error, and percentage error), they are required in order to evaluate the model performance objectively. Model performance can vary depending on the performance measure that is applied.

Hence, this study has applied four performance measures to compare the model performance quantitatively: mean absolute error (MAE) (Equation 3), mean absolute percentage error (MAPE) (Equation 4), mean squared error (MSE) (Equation 5), and root mean squared error (RMSE) (Equation 6). MAE is appropriate for evaluating the least deviation from the average. MAPE is a dimensionless parameter that is similar to MAE but is expressed as a percentage. MSE and RMSE penalize the models that have large deviations, and owing to this characteristic, these performance measures have previously been used in studies on water pressure and demand data forecasting [20,56].

$$\text{Mean Absolute Error (MAE)} = \frac{1}{N} \sum_{t=1}^N |x_{obs.}^t - x_{fore.}^t| \quad (3)$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \frac{100}{N} \sum_{t=1}^N \left| \frac{x_{obs.}^t - x_{fore.}^t}{x_{obs.}^t} \right| \quad (4)$$

$$\text{Mean Squared Error (MSE)} = \frac{1}{N} \sum_{t=1}^N (x_{obs.}^t - x_{fore.}^t)^2 \quad (5)$$

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{MSE} \quad (6)$$

where N is the number of variables, $x_{obs.}$ is an observed data set, and $x_{fore.}$ is a forecasted data set.

3. Application and Results

This section shows the results of pressure forecasting obtained by the three cases of various combinations of training data using the proposed CDANN technique for effective pressure data forecasting in real world WDS when the given data are insufficient. The Galsan block pressure data applied in this study was obtained from June 2019 to November 2019 at 1-min intervals, and the daily average temperature was obtained from Korea Water Management Information System (WAMIS). All computations were performed using MS Excel 2019, and the neural network/data Manager was developed using MATLAB 2019 (MathWorks, Inc., Natwick, MA).

3.1. Model Formulation

First, a sensitivity analysis of the parameters of the CDANN (e.g., training function, adaption learning function, performance function, number of layers, number of neurons, and transfer function) was performed by comparing the model performance depending on parameter variation, applying this model for the effect of each parameter, and comparing the performance indices for 252 cases (14 training functions $\times$ 2 learning functions $\times$ 3 learning functions $\times$ 3 transfer functions). As a result of the sensitivity analysis, we used the Bayesian regularization backpropagation (TRAINBR) algorithm for the training function, the gradient descent weight, and bias learning function (LEARNGD) for the adaptation learning function, and log-sigmoid (LOGSIG) for the transfer function. The applied data for this study have been categorized as 70% training, 15% validation, and 15% test datasets. Among the used data, since the training pressure datasets (day of the week, time period, daily average temperature) have different pressure ranges, the normalization process has been performed during pre-processing using the min-max normalization approach, which determines a minimum and maximum value in the training data, and each value is normalized within 0 to 1. The normalization constants that were used in the training dataset were also used in the validation and test sets.

Subsequently, to evaluate the impact of various training data combinations, the cross-domain pre-training approach was applied. Figure 5 details various training data combinations proposed in this study. The first training dataset forecasts the unknown data meter 3 (M3) using data from the same day of the week. The input and target data are used in the training step, and this training focuses on the relationship between input and target. When combining the training data using the cross-domain approach, the replica data and original data are combined as a pair. Subsequently, the training data are composed until all combinations are generated for the input or target data.

In Figure 5a, the meter 3 (M3) data from the fourth week, from amongst the given one-month data, is assumed to be the unknown data. The CDANN learns the relationship between the input data and target data using the training data and makes a replica of 11 data points except for the unknown data in the fourth week and meter 3 data (W4, M3). Using 22 data points including the replica, each data point to be located at the input or target is combined. First, in a box (dash-outlined), when (W1, M1) is an input, the target data to be predicted are (W1, M2). In this process, the input and target data are used in the training stage. As the correlation between data of the same day is higher than that of a different day, the same day (W1) data and data from a different meter (M1 and M2) are combined. Likewise, all combinations such as (W2, M1)-(W2, M2) and (W3, M1)-(W3, M2) are constructed. CDANN cross-uses data from each meter as the input and generates target data (by turns) for full nonlinear relationships among different meter data to be embedded in its structure. The hidden layer eventually improves the forecasting performance. To predict the unknown data at test periods, the output is generated by inputting either (W4, M1) or (W4, M2), composed of the above combination rolls.

Figure 5b presents a configuration of the training data combination using the same time period data and applies it to three cases in order to combine the data, considering the time series characteristics. Case 1 performs training, considering the continuous time among data of the same pressure meter as (M1, TS1), (M1, TS2). As the pressure data of WDS are continuous, the training data combination is performed using continuous time data such as TS1 and TS2. Case 2 compiles a training dataset by combining different meter data from the same time series. The combinations of input and target data cross each other according to the same process as the first training set combination approach. Case 3 sets training data, considering the same time period and meter data for different days.

Figure 5c describes the combination of training data considering the same season. In this study, the season indicates that if the difference in the daily average temperature is 5°C, the same season is assumed, such as 10–15°C, 15–20°C, 20–25°C, and 25–30°C. The configuration of the given data is similar to the first training data combination. After data configuration, based on these training data combinations (i.e., (a) day of the week, (b) time period, and (c) daily average temperature), the replica data are generated and the process of setting the input and target data is performed according to the cross-domain training approach, which is seen in Figure 5a.

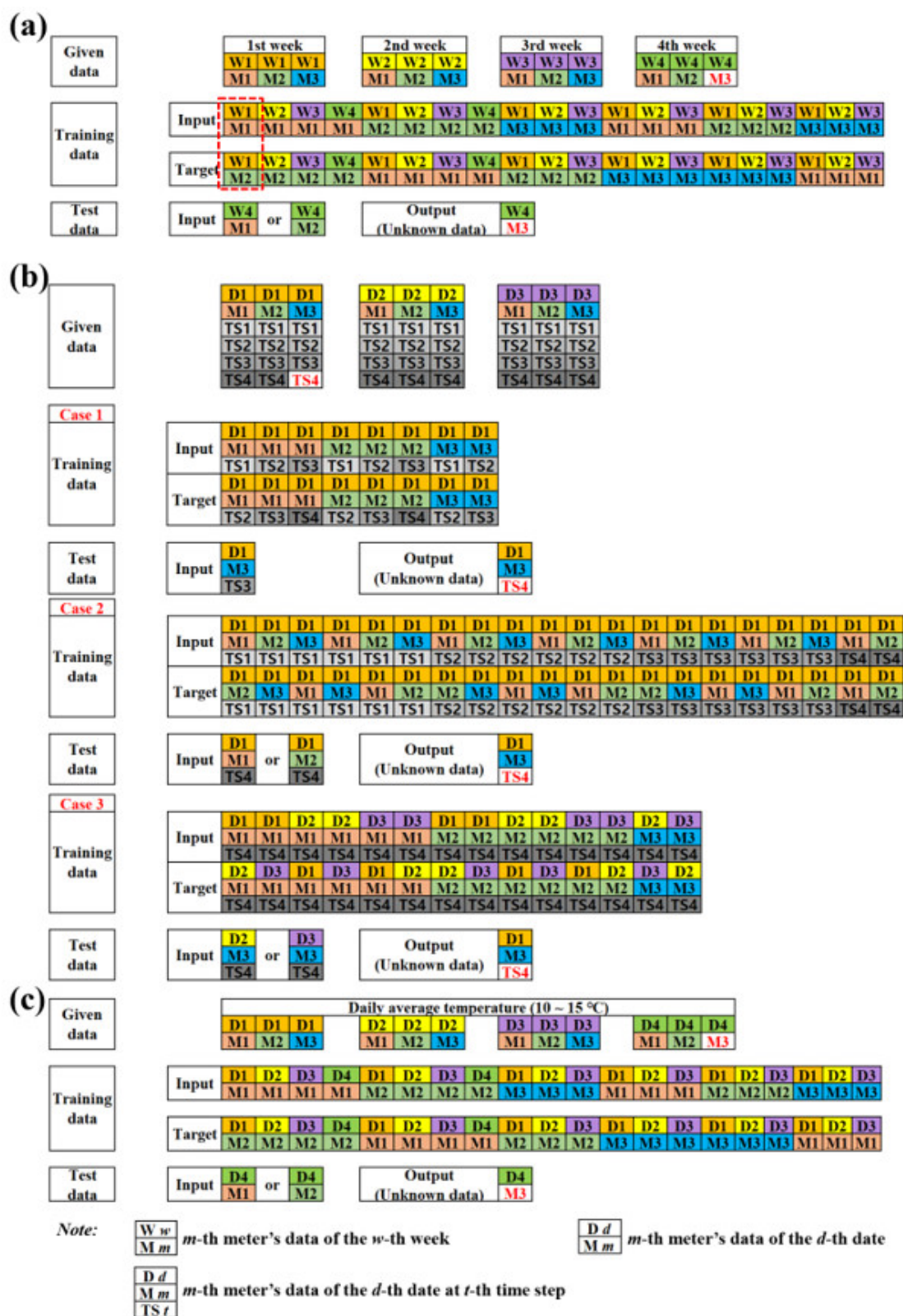


Figure 5. Depiction of training data combinations: (a) day of the week, (b) time period, (c) season (temperature).

3.2. Forecasting Results and Discussions

The Galsan network was used for validation to verify the application of the proposed reliable pressure data forecasting approach. First, unknown data forecasting is performed using the proposed three training data combinations and their performance is evaluated against the observed data in

Figure 6 and Table 2. The graphical results show the observed data and two trial results among ten predictions. Furthermore, to evaluate stable forecasting results of this study, the data forecasting is performed 10 times individually. Figure 6 shows the forecasting results for 10-time trials and two cases. Further, to increase the training efficiency, the normalization process is applied in the pre-data-processing for each training dataset (day of the week, time period, daily average temperature) because each dataset has a different pressure range.

The pressure data applied in this simulation use the data for the same day of the week in October 2019 only. For example, the data for Tuesday in October totaled 17,280 ($=3 \text{ m} \times 1,440/\text{day} \times 4 \text{ days}$), excluding holidays with different water consumption patterns.

Table 2. Pressure data forecasting results of the Galsan network using performance measures.

Scenarios	Simulation	MAE (10^{-4})	MAPE (%)	MSE (10^{-3})	RMSE (10^{-2})
Training dataset 1 (Day of the week)	Mon.	2.56	2.74	2.95	5.43
	Tue.	2.85	2.72	2.82	5.31
	Fri.	2.93	2.85	3.17	5.63
	Sun.	2.11	2.47	2.03	4.51
Training dataset 2 (Time period)	Case 1	6.83	11.41	6.58	8.11
	Case 2	6.50	12.94	6.33	7.95
	Case 3	5.90	9.48	5.12	7.15
Training dataset 3 (Daily average temperature)	10–15°C	3.43	3.59	3.76	6.13
	15–20°C	3.23	3.13	3.70	6.08
	20–25°C	4.09	3.95	4.55	6.74
	25–30°C	3.76	3.24	3.46	5.89

As section 2.3 mentioned, CDANN is useful in the case of fewer input variables or when the input and output show a similar trend. In case of the applied Galsan network, since there are only three pressure meter datasets, it has a limited amount of training data. CDANN can surpass the traditional forecasting approach when there is a small amount of input data by increasing the number of input variables via creating replicas of the user input variables and combining them. The performance of CDANN was expressed by comparing the original pressure data and forecasting data using various performance indices (i.e., MAE, MAPE, MSE, RMSE). The overall results show that the performance is, on average, below 10% MAPE, 0.1 RMSE, and 0.001 MAE regardless of the kind of training dataset. In the case of the pressure trend, a slight difference is observed depending on the day of the week. It is clearly observed that the trend in water pressure varies with water demand. In the case of Monday, Tuesday, and Sunday, the prediction results agree well, but the results of Friday show a slight variation between the observed and predicted data. This implies that there is a more variation in consumption on Friday than on other days. This shows that water users have a more irregular life pattern on Friday in comparison to other days. With regard to performance evaluation, the forecasting on Sunday shows the minimum error of 2.47% (MAPE). According to MSE (2.03) and RMSE (4.51), the overall forecasting data have small deviations. Likewise, in the graphical results, it is observed that the prediction error for Friday is the largest, and the MSE indicates that overall matching is worse than the forecasting of peak points. Figure 6b presents the prediction results, considering time period training. The applied pressure data was acquired in September 2019, and each numerical representation of the time period data is 10,800 ($=60/\text{h} \times 6 \text{ h} \times 30 \text{ days}$). Furthermore, because each case of training dataset 2 is considered to have time series characteristics, the amount of applied data are slightly different. However, the results of the prediction do not show a significant influence. The forecasting trend in training dataset 2 is in accordance with the overall observed data with time, although it cannot predict peak values. This implies that according to the training based on the time period, water consumption has considerable variations.

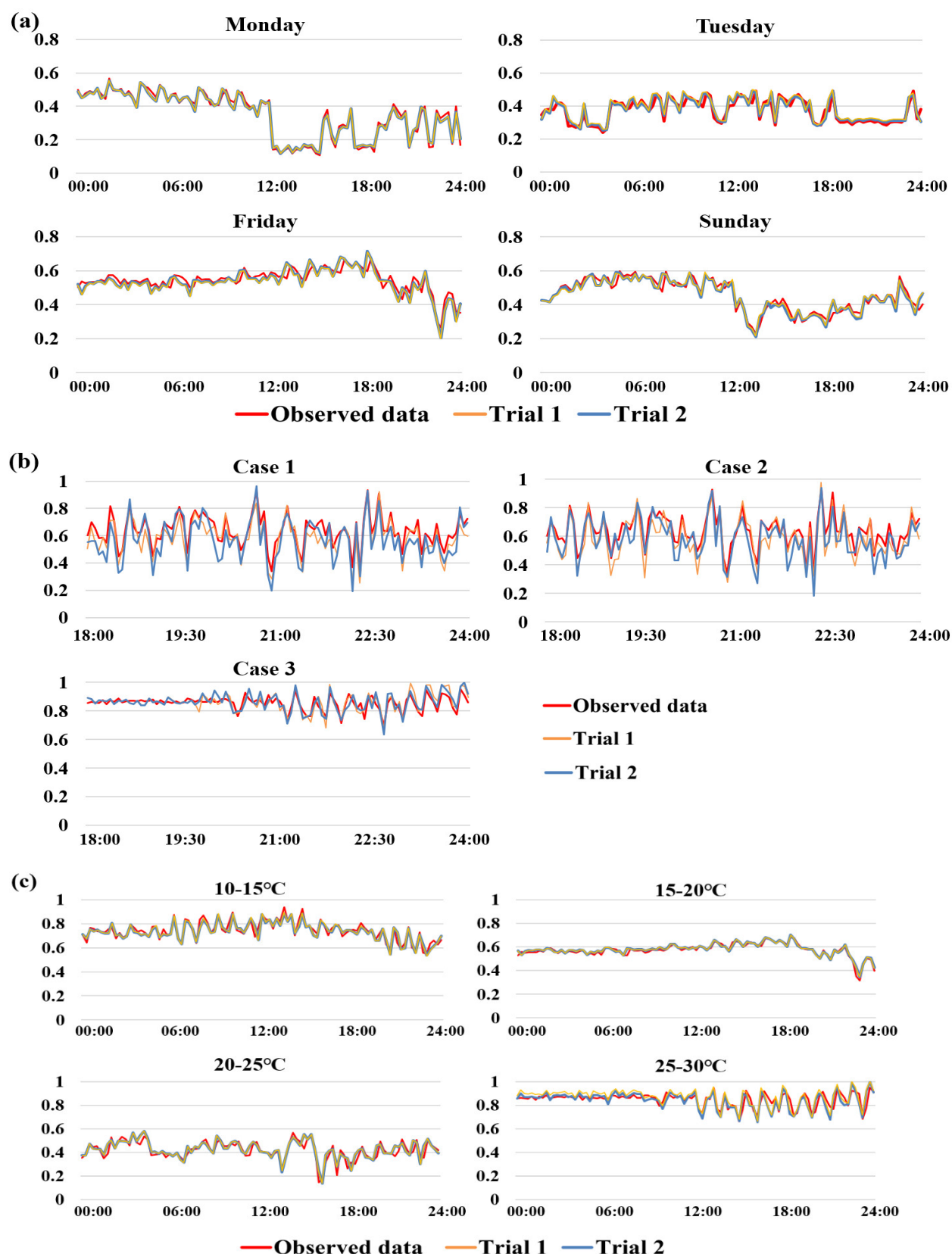


Figure 6. Pressure data forecasting trend: (a) day of the week, (b) time period, and (c) season (temperature).

In addition, a comparison of the three cases indicates that the best prediction is that of Case 3, which trains the same time period and meter data for different dates. These results indicate that in forecasting pressure data with strong time series characteristics, training should be performed at the

same time step, and the sensitivity in forecasting the same pressure meter is higher than that of the date of measure. The third training dataset is a prediction considering the daily average temperature. This set divides the four cases, depending on temperature, at 5°C intervals. For this simulation, the data that is at a similar temperature between June 2019 and November 2019 is used. In Figure 6c, the trend and values of pressure are different depending on the daily average temperature. The pressure values for 15–25°C are less than those for 25–30°C and 15–25°C. As the network is located in a rural area, it uses a large quantity of water during the farming season (15–25°C). Furthermore, the predictive results considering the daily average temperature were superior to the training dataset 2. Most of the cases for MAPE demonstrate approximately 3% error, and the overall forecasting ability is also outstanding based on the MSE and RMSE indices.

In addition, this study compares the predictive results to show the effect of the CDANN. Table 3 illustrates good performance for MAPE and RMSE of the CDANN. For all scenarios from the aspect of overall deviation error, MAPE values are lower than the forecasting results achieved by performing the standard ANN training technique. In the case of training dataset 1, the average MAPE value of standard ANNs is approximately 3.3, which is greater than that of the CDANN error by approximately 0.6 (23%) and that of other prediction errors by 17.1%. The CDANN has a greater amount of training data than the standard ANN owing to the creation of a replica of the training data. Furthermore, as the input and target data have similar trends, the relationship of these data is more efficient for data training. For these reasons, the results show that the training approach using cross-domain outperforms the other approaches.

Table 3. Comparative results between normal and cross-domain training approaches.

Scenarios	Simulation	MAPE (%)		RMSE (10^{-2})	
		Standard ANN	CDANN	Standard ANN	CDANN
Training dataset 1 (Day of the week)	Mon.	3.29	2.74	6.52	5.43
	Tue.	3.54	2.72	6.90	5.31
	Fri.	3.48	2.85	6.87	5.63
	Sun.	3.05	2.47	5.57	4.51
Training dataset 2 (Time period)	Case 1	12.67	11.41	9.00	8.11
	Case 2	14.10	12.94	8.67	7.95
	Case 3	10.71	9.48	8.08	7.15
Training dataset 3 (Daily average temperature)	10–15°C	4.34	3.59	7.42	6.13
	15–20°C	3.66	3.13	7.11	6.08
	20–25°C	4.38	3.95	7.48	6.74
	25–30°C	3.86	3.24	7.01	5.89

4. Conclusions

Considerable effort has been undertaken to improve data forecasting for the operation and maintenance of WDS. These major efforts have included water demand forecasting by applying statistical approaches and, more recently, artificial intelligence. However, even though the pressure data are directly affected and more useful for reliable and effective WDS operation (i.e., pump stations or pressure reduction valves), there have been few efforts made to address water pressure forecasting problems. Moreover, despite applying high technology such as artificial intelligence for the data gathering process, there are limitations to the effective operation of WDS if the applied data are scarce and incomplete or missing for various reasons (e.g., interference of network connection, malfunction of the data collector, sensor failure).

Therefore, this study introduced a new missing and incomplete data control approach based on real pressure data for reliable and efficient WDS operation and maintenance. In addition, to improve the performance of data prediction, the standard for data mining, considering the characteristics of factors related to the pressure variants, was analyzed. The proposed approach has three sub-techniques: (1) application of data based on high-resolution, real-pressure data to increase the

accuracy of data prediction; (2) development of a cross-domain artificial neural network (CDANN) combining the standard artificial neural networks and the cross-domain training approach for the missing data control; and (3) analysis of the standard of data mining considering factors affected by water pressure (i.e., day of the week, time period, and daily average temperature) to improve predictive performance.

The proposed missing data handling techniques applied a CDANN incorporating three combinations of training data (i.e., day of the week, time period, and daily average temperature). Among the three training data combinations, the same day of the week and daily average temperature are identified as the best approaches except for the same time periods. In addition, this study has been compared with the proposed three training approaches and the normal ANN training technique to evaluate performance. For this case, it can improve the forecasting accuracy by approximately 17% compared to the normal training approach for observed data.

This study has several limitations that future studies should address. More hydrological factors (precipitation, reservoir level, inflow, and humidity) should be considered for improving the performance of forecasting. A sensitivity analysis of various forecasting methods (e.g., various machine learning algorithms) should be carried out, and performance should be compared to achieve the best forecasting results. Finally, using the forecasted data, a pump and valve operation rule should be formulated to investigate its impact on the quality of the ultimate solution, and thereby confirm the findings of this study.

In the future, this study will be expected to benefit data control approaches using pressure data where the amount of data is otherwise insufficient. In particular, the CDANN combines standard ANNs and the cross-domain training approach in order to overcome the drawbacks of traditional machine learning algorithms. The proposed technique can greatly contribute to the improvement of various high-performance machine learning algorithms in the future.

Author Contributions: Y.H.C. surveyed the previous studies, wrote the original manuscript, and conducted simulations. D.J. conceived the original idea of the proposed method and supervised. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported (1) by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1C1C1006481) and (2) by a Korea University Grant.

Acknowledgment: The authors are grateful for the funding agencies, the National Research Foundation of Korea (NRF) and Korea University.

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

References

1. Gardiner, V.; Herrington, P. *Water Demand Forecasting*, 1st ed.; Spon Press: Norwich, UK, 1990.
2. Billings, B.; Jones, C. *Forecasting Urban Water Demand*, 2nd Ed.; American Waterworks Association: Denver, CO, USA, 2008.
3. Kanakoudis, V.K.; Tolikas, D.K. The role of leaks and breaks in water networks: Technical and economical solutions. *J. Water Supply Res. Technol. — AQUA* **2001**, *50*, 301–311.
4. Tsitsifli, S.; Kanakoudis, V.; Bakouros, I. Pipe networks risk assessment based on survival analysis. *Water Resour. Manag.* **2011**, *25*, 3729.
5. Kanakoudis, V.; Gonelas, K. Forecasting the residential water demand, balancing full water cost pricing and non-revenue water reduction policies. *Procedia Eng.* **2014**, *89*, 958–966.
6. Kanakoudis, V.; Gonelas, K. Analysis and calculation of the short and long run economic leakage level in a water distribution system. *Water Util. J.* **2016**, *12*, 57–66.
7. Behboudian, S.; Tabesh, M.; Falahnezhad, M.; Ghavanini, F.A. A long-term prediction of domestic water demand using preprocessing in artificial neural network. *J. Water Supply Res. Technol.* **2014**, *63*, 31–42.
8. 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.
9. Voicu, O.; Wong, Y.S. On the construction of a nonlinear recursive predictor. *J. Comput. Appl. Math.* **2006**, *190*, 393–407.

10. Alvisi, S.; Franchini, M.; Marinelli, A. A short-term, pattern-based model for water-demand forecasting. *J. Hydroinform.* **2007**, *9*, 39–50.
11. Zhou, S.L.; McMahon, T.A.; Walton, A.; Lewis, J. Forecasting daily urban water demand: A case study of Melbourne. *J. Hydrol.* **2000**, *236*, 153–164.
12. Bennett, C.; Stewart, R.A.; Beal, C.D. ANN-based residential water end-use demand forecasting model. *Expert Syst. Appl.* **2013**, *40*, 1014–1023.
13. Tiwari, M.K.; Adamowski, J.F. Medium-term urban water demand forecasting with limited data using an ensemble wavelet-bootstrap machine-learning approach. *J. Water Resour. Plan. Manag.* **2014**, *141*, 04014053.
14. Herrera, M.; Torgo, L.; Izquierdo, J.; Perez-Garcia, R. Predictive models for forecasting hourly urban water demand. *J. Hydrol.* **2010**, *387*, 141–150.
15. Jentgen, L.; Kidder, H.; Hill, R.; Conrad, S. Energy management strategies use short-term water consumption forecasting to minimize cost of pumping operations. *J. Am. Water Works Ass.* **2007**, *99*, 86–94.
16. Cutore, P.; Campisano, A.; Kapelan, Z.; Modica, C.; Savic, D. Probabilistic prediction of urban water consumption using the SCEMUA algorithm. *Urban Water J.* **2008**, *5*, 125–132.
17. Firat, M.; Yurdusev, M.A.; Turan, M.E. Evaluation of artificial neural network techniques for municipal water consumption modeling. *Water Resour. Manag.* **2009**, *23*, 617–632.
18. Ghiassi, M.; Zimbra, D.K.; Saidane, H. Urban water demand forecasting with a dynamic artificial neural network model. *J. Water Resour. Plan. Manag.* **2008**, *134*, 138–146.
19. Brentan, B.M.; Luvizotto, E., Jr.; Herrera, M.; Izquierdo, J.; Pérez-García, R. Hybrid regression model for near real-time urban water demand forecasting. *J. Comput. Appl. Math.* **2017**, *309*, 532–541.
20. Wang, Y.; Puig, V.; Cembrano, G. Non-linear economic model predictive control of water distribution networks. *J. Process Control.* **2017**, *56*, 23–34.
21. Candelieri, A.; Perego, R.; Archetti, F. Intelligent pump scheduling optimization in water distribution networks. In Proceedings of the International Conference on Learning and Intelligent Optimization, Kalamata, Greece, 10–15 June 2018, pp. 352–369.
22. Doghri, M.; Duchesne, S.; Poulin, A. Impacts of the integration of water demand prediction in real time control of water distribution systems. In Proceedings of the WDSA/CCWI Joint Conference Proceedings, Kingston, ON, Canada, 23–25 July 2018; Volume 1.
23. Abu-Mahfouz, A.M.; Hamam, Y.; Page, P.R.; Adedeji, K.B.; Anele, A.O.; Todini, E. Real-time dynamic hydraulic model of water distribution networks. *Water*. **2019**, *11*, 470.
24. Kanakoudis, V.; Gonas, K. Applying pressure management to reduce water losses in two Greek cities' WDSs: Expectations, problems, results and revisions. *Procedia Eng.* **2014**, *89*, 318–325.
25. Wood, A.; Lence, B.J. Using water main break data to improve asset management for small and medium utilities: District of Maple Ridge, BC. *J. Infrastruct. Syst.* **2009**, *15*, 111–119.
26. Haider, H.; Sadiq, R.; Tesfamariam, S. Performance indicators for small-and medium-sized water supply systems: A review. *Environ. Rev.* **2014**, *22*, 1–40.
27. Schafer, J.L.; Olsen, M.K. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivar. Behav. Res.* **1998**, *33*, 545–571.
28. Fares, H.; Zayed, T. Hierarchical fuzzy expert system for risk of failure of water mains. *J. Pipeline Syst. Eng. Pract.* **2010**, *1*, 53–62.
29. Francisque, A.; Shahriar, A.; Islam, N.; Betrie, G.; Binte Siddiqui, R.; Tesfamariam, S.; Sadiq, R. A decision support tool for water mains renewal for small to medium sized utilities: A risk index approach. *J. Water Supply Res. Technol.* **2014**, *63*, 281–302.
30. Jafar, R.; Shahrour, I.; Juran, I. Application of Artificial Neural Networks (ANN) to model the failure of urban water mains. *Math. Comput. Model.* **2010**, *51*, 1170–1180.
31. Rogers, P.D.; Grigg, N.S. Failure assessment model to prioritize pipe replacement in water utility asset management. In Proceedings of the Water Distribution Systems Analysis Symposium, Cincinnati, OH, USA, 27–30 August 2006; pp. 1–17.
32. Kanakoudis, V.K. Vulnerability based management of water resources systems. *J. Hydroinformatics* **2004**, *6*, 133–156.
33. Acuna, E.; Rodriguez, C. The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications*; Springer: Berlin, Heidelberg, 2004; pp. 639–647.
34. Kornelsen, K.; Coulbaly, P. Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset. *J. Hydrol. Eng.* **2014**, *19*, 26–43.

35. Inman, D.; Elmore, R.; Bush, B. A case study to examine the imputation of missing data to improve clustering analysis of building electrical demand. *Build. Serv. Eng. Res. Technol.* **2015**, *36*, 628–637.
36. Nelwamondo, F.V.; Mohamed, S.; Marwala, T. Missing data: A comparison of neural network and expectation maximization techniques. *Curr. Sci.* **2007**, *93*, 1514–1521.
37. Sim, J.; Lee, J.S.; Kwon, O. Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. *Math. Probl. Eng.* **2015**, *2015*, 538613.
38. Honaker, J.; King, G. What to do about missing values in time-series cross-section data. *Am. J. Pol. Sci.* **2010**, *54*, 561–581.
39. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019; Volume 793.
40. Zhuang, F.; Luo, P.; Xiong, H.; Xiong, Y.; He, Q.; Shi, Z. Cross-domain learning from multiple sources: A consensus regularization perspective. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1664–1678.
41. Barker, E.B.; Kelsey, J.M. *Recommendation for Random Number Generation Using Deterministic Random Bit Generators (Revised)*; US Department of Commerce, Technology Administration, National Institute of Standards and Technology, Computer Security Division, Information Technology Laboratory: Washington, DC, USA, 2007, pp. 800–890.
42. Hoag, J.E. *Synthetic Data Generation: Theory, Techniques and Applications*; University of Arkansas: Fayetteville, AR, USA, 2008.
43. Hoag, J.E.; Thompson, C.W. A Parallel General-Purpose Synthetic Data Generator. In *Data Engineering*; Springer: Boston, MA, USA, 2009; pp. 103–117.
44. Berkovsky, S.; Kuflik, T.; Ricci, F. The impact of data obfuscation on the accuracy of collaborative filtering. *Expert Syst. Appl.* **2012**, *39*, 5033–5042.
45. Luo, P.; Zhuang, F.; Xiong, H.; Xiong, Y.; He, Q. Transfer learning from multiple source domains via consensus regularization. In Proceedings of the 17th ACM conference on Information and knowledge management, Napa Valley, CA, USA, 26–30 October 2008, pp. 103–112.
46. Gibbs, K.C. Price variable in residential water demand models. *Water Resour. Res.* **1978**, *14*, 15–18.
47. Kozłowski, E.; Kowalska, B.; Kowalski, D.; Mazurkiewicz, D. Water demand forecasting by trend and harmonic analysis. *Arch. Civ. Mech. Eng.* **2018**, *18*, 140–148.
48. Gallant, S.I. *Neural Network Learning and Expert Systems*; The MIT Press: Cambridge, MA, USA, 1993.
49. Smith, M. *Neural Networks for Statistical Modelling*; Van Nostrand Reinhold: New York, NY, USA, 1994; p. 235.
50. Dreyfus, G.; Martinez, J.-M.; Samuelides, M.; Gordon, M.B.; Badran, F.; Thiria, S.; Herault, L. *Reseaux de Neurones: Methodologie et Applications*; Editions Eyrolles: Paris, France, 2002.
51. Karunanithi, N.; Grenney, W.J.; Whitley, D.; Bovee, K. Neural networks for river flow prediction. *ASCE J. Comput. Civil Eng.* **1994**, *8*, 210–220.
52. Govindaraju, R.S. Artificial neural network in hydrology. II: Hydrologic application, ASCE task committee application of artificial neural networks in hydrology. *J. Hydrol. Eng.* **2017**, *5*, 124–137.
53. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, Australia, 6–11 August 2017, pp. 1857–1865.
54. Yamaguchi, M.; Koizumi, Y.; Harada, N. AdaFlow: Domain-adaptive Density Estimator with Application to Anomaly Detection and Unpaired Cross-domain Translation. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3647–3651.
55. Geosling, E.; Pollak, J.; Hooper, R. Advancing water science through community collaboration. *Environ. Earth Sci.* **2015**, *73*, 1919–1924.
56. Altunkaynak, A.; Özger, M.; Çakmak, M. Water consumption prediction of Istanbul city by using fuzzy logic approach. *Water Resour. Manag.* **2005**, *19*, 641–654.

