water

MDPI

Article

Urban Flood Prediction Using Deep Neural Network with Data Augmentation

Hyun Il Kim and Kun Yeun Han *

Department of Civil Engineering, Kyungpook National University, 80 Daehak-ro, Buk-gu Daegu 41566, Korea; hyunn228@gmail.com

* Correspondence: kshanj@knu.ac.kr

Received: 11 February 2020; Accepted: 19 March 2020; Published: 22 March 2020

Abstract: Data-driven models using an artificial neural network (ANN), deep learning (DL) and numerical models are applied in flood analysis of the urban watershed, which has a complex drainage system. In particular, data-driven models using neural networks can quickly present the results and be used for flood forecasting. However, not a lot of data with actual flood history and heavy rainfalls are available, it is difficult to conduct a preliminary analysis of flood in urban areas. In this study, a deep neural network (DNN) was used to predict the total accumulative overflow, and because of the insufficiency of observed rainfall data, 6 h of rainfall were surveyed nationwide in Korea. Statistical characteristics of each rainfall event were used as input data for the DNN. The target value of the DNN was the total accumulative overflow calculated from Storm Water Management Model (SWMM) simulations, and the methodology of data augmentation was applied to increase the input data. The SWMM is one-dimensional model for rainfall-runoff analysis. The data augmentation allowed enrichment of the training data for DNN. The data augmentation was applied ten times for each input combination, and the practicality of the data augmentation was determined by predicting the total accumulative overflow over the testing data and the observed rainfall. The prediction result of DNN was compared with the simulated result obtained using the SWMM model, and it was confirmed that the predictive performance was improved on applying data augmentation.

Keywords: urban flood; deep neural network; flood prediction; data augmentation

1. Introduction

It is very important to predict the possibility of flooding in urban basins and ascertain the degree of flooding in advance. Hazardous flooding events that cause residential and commercial property damage can occur in many forms and with frequent frequencies. Numerical analysis, trend analysis, and flood prediction for various rainfall scenarios can be utilized as important basic data for urban planning or flood response. One- and two-dimensional flood analysis programs can be used for this purpose; however, in the case of a numerical-based model, it may take some time to adjust parameters, collect data, and perform post-processing of output data.

For this reason, data-driven models using machine learning have recently been applied to the prediction of urban runoff [1]. Granata et al. [2] predicted runoff in urban regions via support vector regression (SVR) and compared it with the results of EPA-SWMM simulations. Although there was a lack of accuracy in estimating the peak runoff, the possibility of improvement was confirmed through verification. If more data are applied to the learning of SVR, the prediction power of peak flow will be further enhanced. The study of input data for a machine learning method is also important to find a reliable prediction model. Talei et al. [3] suggested the evaluation result of rainfall and discharge inputs with adaptive network fuzzy inference systems (ANFIS). In this study, the criteria for determining the optimal rainfall-runoff analysis model among sequential, pruned sequential, and non-sequential

time-series input data were presented. In addition to a general artificial neural network (ANN), the deep learning (DL) technique has begun to be applied to the water resources field to improve the predictive power of hydrologic data and to include more ideas in the model. According to Shen [4], the application of the DL technique to the water resources and hydrology fields is becoming more common, and it is likely to yield good predictions and analysis results in the water resources field as well as in most scientific fields. Hu et al. [5] used a long short-term memory (LSTM)-based DL model for rainfall-runoff simulation and used data on 86 rainfall-runoff patterns. This research was possible because the database for LSTM could be collected by hydrologic observations in the target river basin.

Although many studies have been conducted to predict rainfall, hydrologic data, and flood events using ANN and DL techniques, research related to the supplementation of input data for DL is deficient. Indeed, some studies have raised the problem of insufficiency data for application to a data-driven model. Li et al. [6] raised the problem of the lack of input data and attempted to predict the possibility of flooding in urban regions. For flood forecasting, a conceptual model considering the drainage system was combined with the logistic regression technique. The hydraulic model was compared with the prediction results, and the verification results that could be developed were shown. Nikhil et al. [7] analyzed a dam basin that had been recently affected by a disaster, which was caused by dam outflow. The dam outflow could not be predicted due to the lack of data. Rainfall, flow discharge, and water level data were used to predict the flood downstream, and as observation data were added to the ANN, the predictive power was shown to decrease. It appears that the performance of a neural network becomes degraded when using uncertain observational data when data are insufficient. However, this study does not provide a sure solution for a lack of data.

In this study, urban flood prediction was performed using a deep neural network (DNN), and the problem of a lack of data was resolved by data augmentation and observed rainfall data that caused the urban flooding. The total amount of overflow was predicted via the trained DNN for the drainage basin of Samseong-dong, Seoul, which was damaged by heavy flooding during the years of 2010–2011. The input data for DNN learning consisted of 10 statistical characteristics, including total rainfall, rainfall intensity, hourly maximum rainfall, standard deviation, skewness, and kurtosis. The total accumulative overflow calculated by SWMM simulations was selected as the target value. To apply abundant data to the predictive model, data augmentation was applied. The predictive power according to the amount of augmented data was evaluated to estimate the applicability of the suggested methodology. In this paper, the effectiveness of data augmentation for the urban flood prediction is discussed. In addition, this paper attempts to confirm the optimal input combination by using statistical characteristics of rainfall distribution. The proposed methodology can supplement the lack of data on actual rainfall that has caused flooding and establish an optimal database that can be applied to DL techniques that require various data patterns. Finally, a prediction model is selected to predict the exact total accumulative overflow for the study area.

2. Methodology

2.1. Observed Rainfall

Heavy rainfall that caused severe flooding damage to cities (Seoul, Incheon, Cheonan, Chungju, Busan, Ulsan, Changwon city area) was investigated. For collecting rainfall data, the Automated Synoptic Observing System (ASOS) and observed data from the Automatic Weather System (AWS) were obtained from the Korea Meteorological Administration (KMA) website [8]. The actual rainfall from 2009 to 2017 was investigated and a total of 70 heavy rainfall events were collected. The duration of rainfall was considered to be 6 hours, since most urban floods occur within 6 hours [9]. The total rainfall ranged from a minimum of 7.5 mm to a maximum of 284.4 mm, and the rainfall that did not cause flooding was also considered for input data of the DNN. By collecting data on observed cases of rainfall that induced urban flood, it is possible to compensate for the lack of rainfall data in the study

area. Table 1 shows the names and codes of observatories, the dates of observation, and the amounts of rainfall employed in this study.

Table 1. Survey of observed rainfall nationwide.

	Rainfall Station	Days of Observation	Total Rainfall (mm)
Seoul Area	Seoul (108) ASOS	21 September 2010, 27 July 2011	267.5, 116.5
	Gangnam (400) AWS	21 September 2010, 27 July 2011, 15 August 2012, 22 July 2013	253.5, 184.5, 131.5, 140.5
	Kwanak Mountain (116) Weather Radar	21 September 2010, 27 July 2011	110.0, 184.5
	Seocho (401) AWS	21 September 2010, 27 July 2011, 15 August 2012, 22 July 2013	258.0, 201.0, 129.5, 128.0
	Gangdong (402) AWS	21 September 2010, 27 July 2011	275.5, 111.5
	Songpa (403) AWS	21 September 2010, 27 July 2011	252.0, 194.5
	Gangseo (404) AWS	21 September 2010, 27 July 2011	70.0, 178.5
	Yangchoen (405) AWS	21 September 2010, 27 July 2011	52.0, 180.5
	Dobong (406) AWS	21 September 2010	166.5
	Nowon (407) AWS	21 September 2010, 27 July 2011	57.5, 88.5
	Dongdaemun (408) AWS	21 September 2010, 27 July 2011	232.5, 127.5
	Jungrang (409) AWS	21 September 2010, 27 July 2011	267.5, 112.0
	KMA (410) AWS	21 September 2010, 27 July 2011	251.5, 164.5
	Mapo (411) AWS	21 September 2010, 27 July 2011	241.5, 177.5
	Seodaemun (412) AWS	21 September 2010, 27 July 2011	154.5, 121.5
	Gwangjin (413) AWS	21 September 2010, 27 July 2011	249.0, 117.0
	Seongbuk (414) AWS	21 September 2010, 27 July 2011	160.0, 106.5
	Yongsan (415) AWS	21 September 2010, 27 July 2011	180.0, 136.5
	Eunpyeong (416) AWS	21 September 2010, 27 July 2011	238.5, 25.5
	Geumcheon (417) AWS	21 September 2010, 27 July 2011	160.0, 47.0
	Hangang (418) AWS	21 September 2010, 27 July 2011	240.5, 150.0
	Seongdong (421) AWS	21 September 2010, 27 July 2011	216.5, 118.0
	Bukak Mountain (422) AWS	27 July 2011	125.0
	Gangbuk (424) AWS	21 September 2010, 27 July 2011	119.5, 81.0
	Namhyeon (425) AWS	27 July 2011	248.0
	Kwanak (509) AWS	21 September 2010, 27 July 2011	91.0, 245.0
	Yeongdeungpo (510) AWS	21 September 2010, 27 July 2011	260.0, 181.0
	Gwacheon (590) AWS	21 September 2010, 27 July 2011	63.5, 140.0

	Rainfall Station	Days of Observation	Total Rainfall (mm)
Other Areas	Incheon (112) ASOS	23 July 2017	69.3
	Cheonan (232) ASOS	16 July 2017	223.6
	Chungju (127) ASOS	16 July 2017	284.4
	Busan (159) ASOS	16 July 2009, 27 July 2011, 15 July 2012, 25 August 2014, 11 September 2017	243.0, 229.5, 114.0, 91.0, 254.1,
	Ulsan (152) ASOS	5 October 2016, 11 September 2017	91.2
	Changwon (155) ASOS	25 August 2014, 5 October 2016	209.5, 116.7

2.2. Urban Runoff Simulation

EPA-SWMM, which is used to calculate the total accumulative overflow in the study area, is an urban runoff model that can be used for both rainfall–runoff and river routing, such as the analysis of surface and underground flows, and flows in the drainage pipe network in urban basins with drainage systems [10]. Prior to analysis of the scope of flooding for urban basins, the runoff for the rainfall was calculated using SWMM (version 5.1, United States Environmental Protection Agency, Washington, DC, USA), and operation of the pump station was applied with reference to the report of the Seoul Comprehensive Planning System for Flood Damage Reduction (2015). RUNOFF and EXTRAN were used among the execution blocks of the SWMM model, and the RUNOFF block performed initial calculations for the outflow, water quality, and inflow hydrographs for drainage basins based on the rainfall scenario [11]. The EXTRAN block was used to calculate the flow rate and depth of the drainage pipe system using the output data of the runoff block, and it was possible to analyze the backflow and overflow amount in the pipe based on the flow rate and the water level in the drainage network for each calculation [12]. For analysis of the overflow according to various rainfall events in the study area, the Saint–Venant equations (Equations (1) and (2)) were used.

$$Q = W \times \frac{1}{n}\left(d - d_p\right)^{\frac{5}{3}} \times S^{\frac{1}{2}} \tag{1}$$

$$\frac{\partial Q}{\partial t} + gAS_f - 2V\frac{\partial A}{\partial t} - V^2\frac{\partial A}{\partial x} + gA\frac{\partial H}{\partial x} = 0 \tag{2}$$

Here, Q is runoff (m^3/s), W is the subwatershed width (m), n is the Manning's roughness coefficient, d is the depth (m), d_p is the ground reservoir lost depth (m), S is the subwatershed slope, A is the surface flow cross-sectional area of sub-watershed (m^2), and V is the surface flow velocity (m/s). One-dimensional runoff analysis results of SWMM were used as the target value of the DNN model, although they had some limitations [13]. Because there is no observed flow data in the study area, simulated SWMM results were considered as a reasonable runoff from rainfall events.

2.3. Deep Neural Network

The DNN is the most basic DL technique that can be used for DL by using two or more hidden layers in a general ANN. The basic structure of a DNN is shown as Figure 1a, where $x_1, x_2, \ldots, x_n$ represent the input data with n attributes, and $\hat{y}$ represents the value predicted through the DNN. When the neural network layer is deepened, it can have a high level of expressive power and is strong in terms of elaborate learning and expression efficiency. The DNN can require large amounts of data and must be supported by high-performance computing technology and data storage capabilities. DNNs are also composed of an input layer, a hidden layer, and an output layer, and various activation functions can be applied to ensure that input data is completely reflected the hidden layer [14].

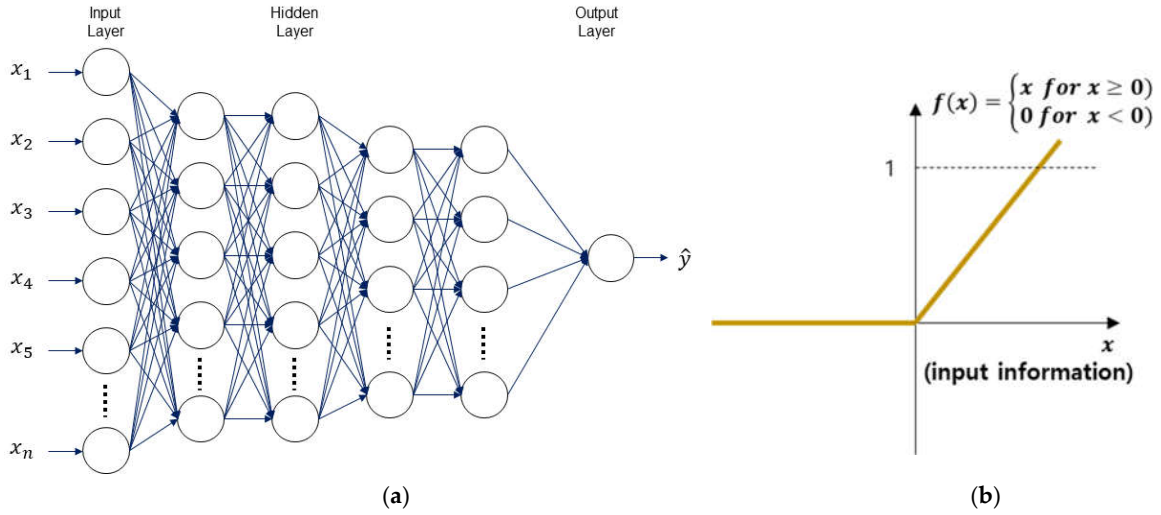


Figure 1. (**a**) Deep neural network (DNN) and (**b**) rectified linear units (ReLU) activation function.

The rectified linear units (ReLU) function as shown in Figure 1b was used as the activation function of hidden layers. As a result, if x is greater than 0, the derivative value is 1, and even if input data go through several hidden layers, the characteristics of the data remain without disappearing to the output layer [15]. Adam used for optimization for the learning process of DNN, which is effective for dealing with nonlinear problems, including outliers [16]. In this study, these parameters could improve the accuracy and reliability of the learning process.

For the error function, the mean absolute error (MAE) shown in Equation (3) was used to estimate the loss rate for learning. Here, n is the number of target data and, in this study, it represents the total number of total accumulative overflows to be predicted. y is the amount of total accumulative overflow calculated using EPA-SWMM, and $\hat{y}$ is the total amount of accumulative overflow predicted by the DNN. In order to make up for the shortcomings of MAE's size-dependent error analysis, the mean absolute percentage error (MAPE) was also used. The MAE and MAPE were applied in the error analysis to verify the performance of the DNN.

$$\text{MAE} = \frac{1}{n}\sum\left|\hat{y} - y\right| \tag{3}$$

$$\text{MAPE} = \frac{1}{n}\sum\left|\frac{\hat{y} - y}{y}\right| \times 100(\%) \tag{4}$$

2.4. Data Augmentation

Data augmentation is a method of increasing the quantitative amount of data by finely tuning the initial input data. It is effective for neural network training to make invariant predictions [17]. In other words, some changes are made to specific data x to obtain $x\prime$ and this is treated as new input data. In general, it is also an effective method to uniformly apply random noise along a Gaussian distribution [18]. In this study, it is necessary to build a DL model with only observed rainfall scenario and so, it was determined that the application of data augmentation is effective. Therefore, data augmentation was attempted on the input data. The standard deviation for the properties (total rainfall, maximum rainfall by duration, rainfall intensity, etc.) of various samples (rainfall scenarios) was multiplied by C(constant number) and $N(0,\ 1)$ (the noise of a Gaussian distribution). Then, the new set of input data was constructed by adding multiplication results to the initial input data. In Equation (5), x represents the original data, $\sigma_i(i = 1 \sim 9)$ is the standard deviation of each 9 data feature, noise is a random number between 0 and 1 generated in the Gaussian distribution, $x\prime$ represents newly generated data, and this was used as new input data. The constant number of C is user-defined hyper-parameter and it was empirically specified as a 0.05 value. The reason for using 0.05 in C is that when the data

augmentation technique is applied to the observed rainfall, it changed less than 10 mm. There are not many examples of the application of data augmentation in the water resources field, and the focus should be on confirming the validity and reliability of this technique. The final goal was to increase the accuracy of the prediction of the total overflow rate by linking data augmentation and DL.

$$x' = x + C \times \sigma_i \times N(0,\ 1) \tag{5}$$

The flowchart for applying data augmentation is shown in Figure 2. The database of rainfall and the total accumulative overflow was constructed by simulation results of SWMM which validated by the usage of FLO-2D (two-dimensional flood analysis program) and flood marks. The DNN was trained using the initial database, and the prediction of total accumulative overflow was performed on the new input data, which came out of the data augmentation. An outlier check for the predicted results was performed by comparing with SWMM simulation results. Excel was used to check whether the negative prediction result was shown, and the unacceptable value was predicted compared to the simulated results. Once the data were confirmed to be sound, the processed input data and the predicted total accumulative overflow data were added to the database for DL. The proposed series of processes was repeated, and this study attempted to identify how many rounds of data augmentation would be best according to the input data combination.

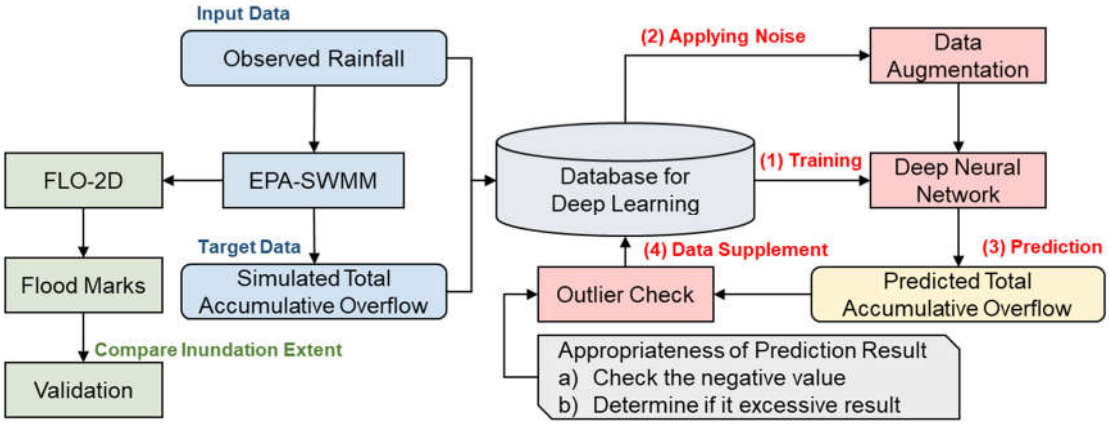
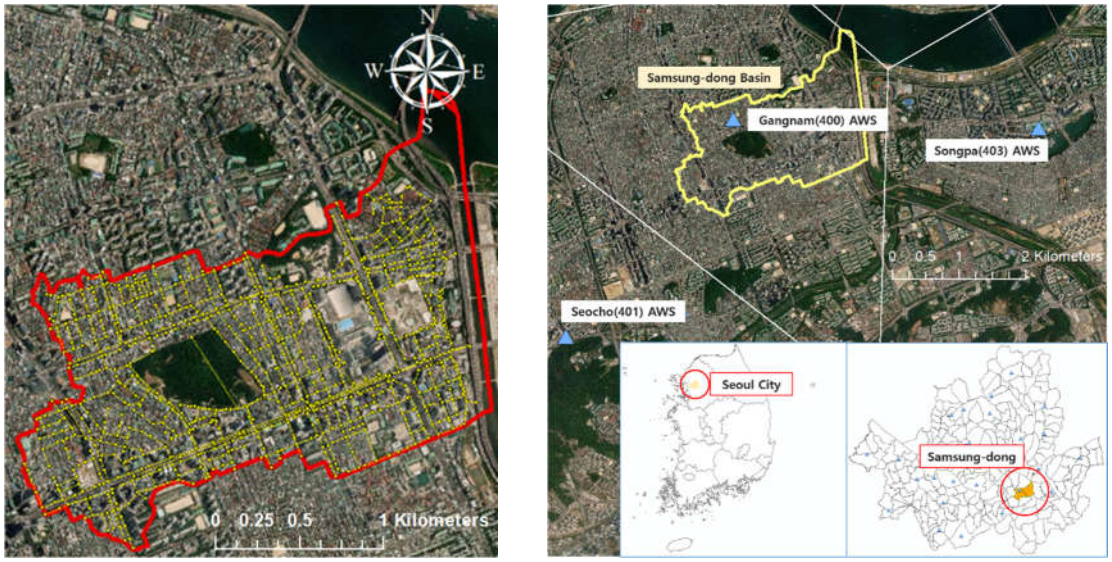


Figure 2. Flowchart for data augmentation and training the DNN.

3. Application

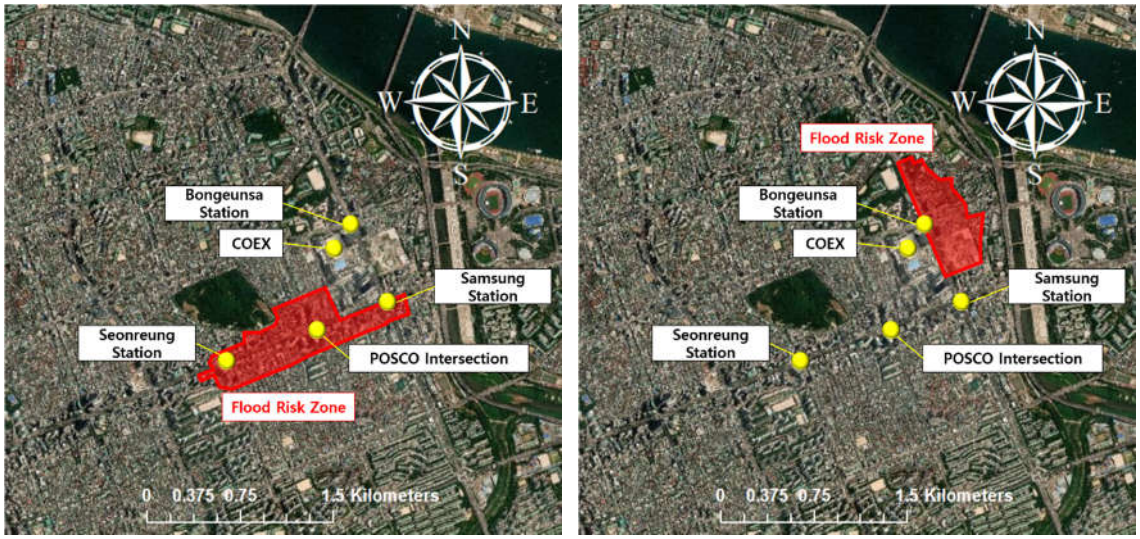
3.1. Study Area

The subject area was set to the Samseong-dong district including Samseong 1 and Samseong 2 drainage basins in Seoul, Korea. Figure 3a reveals the drainage basin borders, drainpipe network, and manhole locations. Figure 3b shows the nearby AWS stations around the study area. The Gangnam AWS station in the Samsung-dong basin. Seocho and Songpa AWS stations are 4.1 km, 4.6 km from Gangnam Observatory, respectively. The drainpipe network and manhole information were used in performing SWMM-based one-dimensional urban runoff analysis. The rainfall on the urban surface flows into the Tancheon stream through a conduit network. The subject area experienced extreme flood damage because of torrential rain on 21 September 2010 and 27 July 2011. It appears that an urban flood occurred because of the relatively low land, complicated drainage system, and inflow of rain exceeding the conduit capacity. According to Seoul City's Storm and Flood Damage Reduction Master Plan [19], urban flood-prone areas (red section in Figure 4) in Samseong-dong were Samseong-Seonreung Stations (around Teheran street, Figure 4a) and Bongeunsa-Samseong Stations (Yeongdong street, Figure 4b). It is likely that in the event of torrential rain exceeding sewage conduit capacity, a flood might have occurred in the two regions described above. According to the inundation trace map in 2010 and 2011, a flood took place along the street between Samseong Station and Seonreung Station.



(**a**) (**b**)

Figure 3. Study area and nearby Automatic Weather System (AWS) stations. (**a**) shows the drainage system, and (**b**) shows rainfall observatory.



(**a**) (**b**)

Figure 4. Flood hazard zone in Samsung-dong. (**a**) shows Samsung–Seonreung Station, and (**b**) shows Bongeunsa–Samsung Station.

3.2. SWMM Validation with Flood Trace Mark

To decide the adequacy of the one-dimensional urban runoff analysis results, this study performed two-dimensional inundation analysis, using overflow at each manhole point and attempted to compare the results with those for flood-prone areas and flood marks. It is because there are only records of flood marks and data on the water level and discharge in the conduit are absent. To validate the adequacy of the SWMM model results, the two-dimensional model adopted the use of the finite difference method (FDM)-based hydraulic analysis program 'FLO-2D'. FLO-2D is a grid-based hydraulic model designed to analyze flood wave propagation approved by the Federal Emergency Management Agency (FEMA) [20]. If the overflows of each manhole were calculated from SWMM, these were entered into the input file of FLO-2D. The exact location of a manhole in the two-dimensional space was found by using the GIS program (spatial join tool in ArcGIS). The scope of calculation on the target basin was examined through an expected flood map and flood history records. After completing the construction of two-dimensional grids, flood waves were analyzed, using the continuity and momentum equations.

Sufficient time was assigned for simulation, delivering flood waves between buildings and to the road as much as possible. To properly reflect the influence of buildings and roads on the flood waves, two-dimensional inundation analysis was performed on 5 m^2 grids. Using Equations (6) and (7), the synthetic roughness coefficient was calculated, and the result (0.025) was applied. Here, n is the synthetic roughness coefficient and n_0 is the bed roughness coefficient. In addition, θ refers to building coverage (%) while n_1, n_2, and n_3 are 0.06 (farmland), 0.047 (road), and 0.05 (others), respectively. Furthermore, A_1, A_2, and A_3 denote the agricultural area, road area, and other land-use areas, respectively, while h (m) refers to the depth of water [21,22]. Considering that the study area is a metropolitan watershed, only the roughness coefficient of the roads (n_2) was used in Equation (7). In addition, this study applied 0.014 as the roughness coefficient of the roads by referring to the road design manual [23].

$$n^2 = n_o^2 + 0.02 \times \frac{\theta}{100 - \theta} \times h^{4/3} \tag{6}$$

$$n_o^2 = \frac{n_1^2 A_1 + n_2^2 A_2 + n_3^2 A_3}{A_1 + A_2 + A_3} \tag{7}$$

As shown in Figure 5, the flood mark of 2010 and 2011 was compared to the two-dimensional flood analysis results applying observed rainfall for 21 September 2010 and 27 July 2011. Flood marks were prepared by Seoul city. The hydrograph of the SWMM result (10 min interval) for each observed rainfall was indicated in Figure 5b,d. The ratio of inundation was calculated in 5 × 5 grid units. The percentage of simulated inundation grid within the flood mark was 75.8% and 71.2% for 2010 and 2011 rainfall events, respectively. In addition, 67% and 63.4% of the two-dimensional simulation results for 2010 and 2011 events occurred within the flood mark from Samseong Station to Seonreung Station. Because a flood trace map is prepared at 0.3 m or deeper, and Figure 5b,d represents the flood depth through surface flow analysis results reflecting roads and buildings, it is concluded that two-dimensional flood analysis and SWMM simulation results in the study area are acceptable.

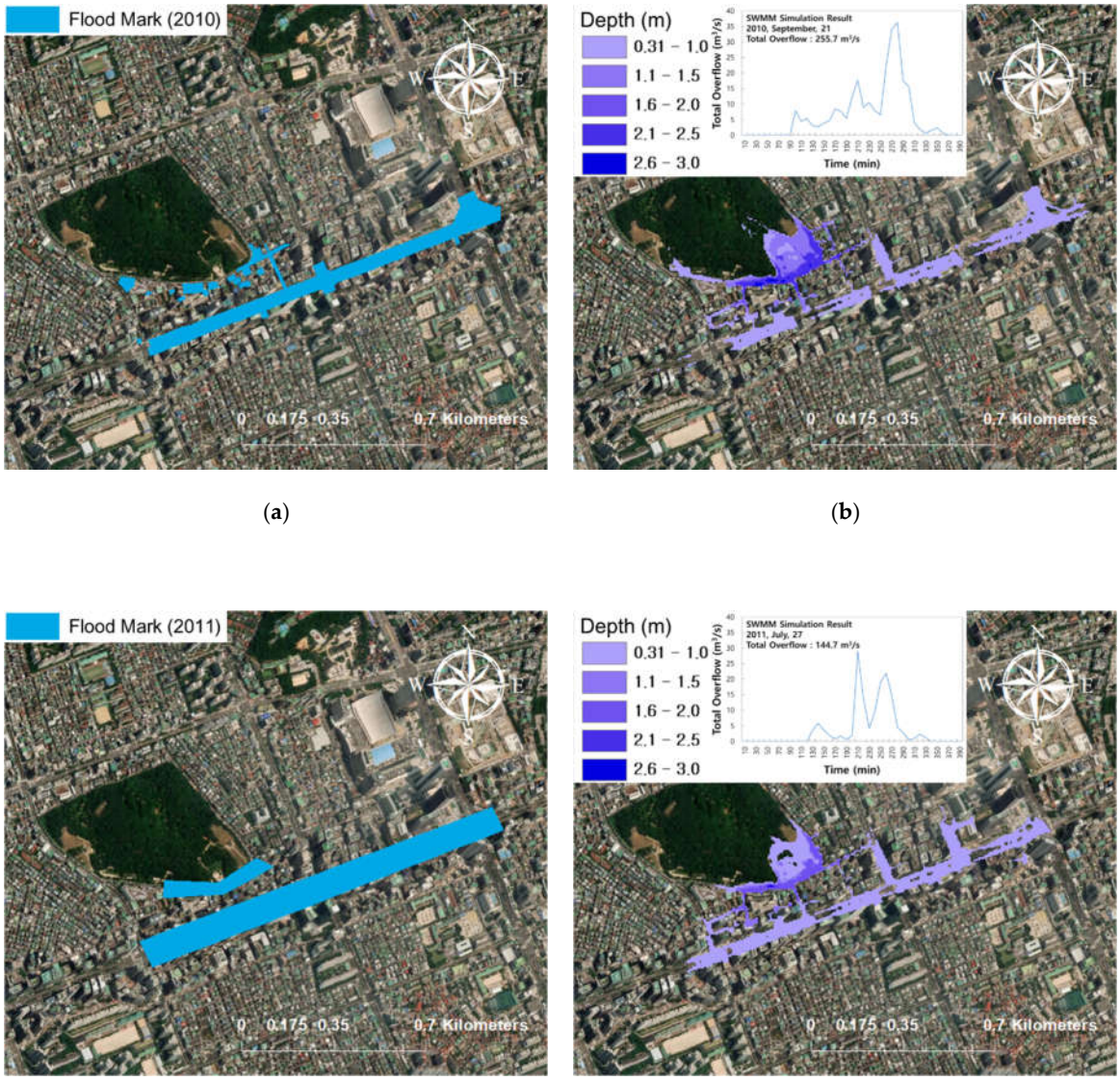


Figure 5. Verification of 2D flood analysis results. (**a**) shows flood mark (2010), (**b**) shows simulation result (21 September 2010), (**c**) shows flood mark (2011), and (**d**) shows simulation result (27 July 2011).

3.3. Input Data

Because the observed heavy rainfall is diverse in terms of the shape of distribution and location of the peak, it is hard to predict urban flood in real-time. It appears that runoff patterns in the urban basins are influenced by rainfall distribution characteristics. Based on 70 observed rainfall events, this study calculated the total rainfall, maximum 1-hour rainfall, maximum 2-hour rainfall, maximum 3-hour rainfall, rainfall intensity, peak rainfall position(%), standard variation, skewness, inter-event time, and kurtosis. Using the same statistical features of rainfall events as those of the DNN input data, this study attempted to build a total accumulative overflow prediction model in which spatiotemporal characteristics of rainfall were learned. The maximum rainfall for a duration of 1–3 h represented the maximum value of ordinate with 1-, 2-, and 3-hour durations while rainfall intensity was estimated by dividing the 6-hour duration of the total rainfall. The peak rainfall location represents the location of the peak value of the rainfall distribution in percentage. The inter-event time is obtained by estimating the time, which represents 0 rainfall events in minutes. The results of statistical analysis on a part of the collected rainfall event data are stated in Tables 2 and 3.

Table 2. Observed rainfall and statistical features (four observation stations in the Seoul area).

Statistical Characteristic	Gangnam Station	Gangnam Station	Dobong Station	Gangbuk Station
Observation Date	21 September 2010	27 July 2011	21 September 2010	27 July 2011
Total Rainfall (mm)	253.50	184.50	166.50	81.00
Max. Rainfall in 1 h (mm)	78.00	71.00	57.00	25.00
Max. Rainfall in 2 h (mm)	136.50	105.50	106.00	38.00
Max. Rainfall in 3 h (mm)	181.50	143.00	136.00	41.00
Rainfall Intensity (mm/h)	42.25	30.75	27.75	13.35
Peak Rainfall Location (%)	25.00	55.56	66.67	77.78
Standard Deviation (mm)	5.17	4.38	4.68	2.49
Skewness	0.39	1.28	0.40	1.05
Inter-event Time (min)	60	10	100	130
Kurtosis	−0.4973	1.2289	−1.4715	0.3522

Table 3. Observed rainfall and statistical features (four observation stations in other areas).

Statistical Characteristic	Busan Station	Ulsan Station	Changwon Station	Cheonan Station
Observation Date	11 September 2017	5 October 2016	5 October 2016	16 July 2017
Total Rainfall (mm)	254.10	233.80	116.70	223.60
Max. Rainfall in 1 h (mm)	84.90	103.10	74.90	69.30
Max. Rainfall in 2 h (mm)	144.90	166.40	97.10	118.50
Max. Rainfall in 3 h (mm)	176.20	203.30	107.20	142.50
Rainfall Intensity (mm/h)	42.35	38.97	19.45	37.27
Peak Rainfall Location (%)	36.11	77.78	75.00	75.00
Standard Deviation (mm)	5.15	6.06	4.96	4.05
Skewness	1.05	0.89	2.25	0.33
Inter-event Time (min)	0	0	20	0
Kurtosis	0.6922	−0.4508	4.7215	−0.8722

The statistical characteristics on rainfall event data were entered into the DNN. In this study, four different cases of input data sets were applied (Table 4). Then, a correlation analysis of the total accumulative overflow of statistical characteristics was performed. It is estimated that the correlation between total rainfall and total accumulative overflow is relatively higher than other characteristics. In terms of a peak rainfall position (%), the correlation converged to 0, not being used in combination with input data. In addition, a relatively high positive correlation with total accumulative overflow was found at maximum 3-hour rainfall and rainfall intensity as well. In terms of skewness, kurtosis, and inter-event time on observed rainfall, a negative correlation with the target value was observed. Case 1 reveals the use of all statistical analysis values excluding a peak rainfall position while Case 2 refers to a combination of input data without two sub-factors among characteristics in negative correlation. Case 3 uses five data sets in a high positive correlation. Lastly, Case 4 uses the highest positive correlation data and three of the negative correlation data. Regardless of the combination of input data, the target values were the same as the total accumulative overflow.

Table 4. Combination of input data.

CASE 1 (9 Inputs)	CASE 2 (6 Inputs)	CASE 3 (5 Inputs)	CASE 4 (4 Inputs)	R^2 with Total Accumulative Overflow
Total Rainfall	Total Rainfall	Total Rainfall	Total Rainfall	0.8144
Max. Rainfall in 1 h	Max. Rainfall in 1 h	Max. Rainfall in 1 h	-	0.5698
Max. Rainfall in 2 h	Max. Rainfall in 2 h	Max. Rainfall in 2 h	-	0.5114
Max. Rainfall in 3 h	Max. Rainfall in 3 h	Max. Rainfall in 3 h	-	0.3957
Rainfall Intensity	Rainfall Intensity	Rainfall Intensity	-	0.566
Standard Deviation	Standard Deviation	-	-	0.3623
Skewness	-	-	Skewness	(−) 0.3778
Kurtosis	-	-	Kurtosis	(−) 0.2499
Inter-event Time	-	-	Inter-event Time	(−) 0.1087

3.4. Prediction Model and Data Augmentation

A total of 69 rainfall events were used for DNN learning except for the prediction target rainfall among a total of 70 observed rainfall events. The DNN was composed of one input layer, eight hidden layers, and one output layer, and three hidden layers were set to have 18, 18, 12, 12, 12, 10, 10, and 8 nodes, respectively. The number of hidden layers and the number of nodes of each hidden layer were determined through trial and error. The final structure of DNN was determined by the empirical process and the MAE of the testing process. The summarized process of finding the optimal structure of the DNN is shown in Table 5. The activation function of the hidden layer uses the ReLU function. In the case of the activation function for output nodes, the linear function was used to derive the results in units of floating decimal points. The epoch for maximum learning was set to 5000. In DNN learning, 70% of input data was used for training, 20% was used for validation, and 10% was used for testing. The data for training, validation, and testing were randomly selected in 69 rainfall events. This is to avoid over-fitting a particular data set. As another way to avoid over-fitting, the early stopping function was used. If the MAE value for validation does not decrease even after repeating the DNN training, the learning process was stopped early.

Table 5. Empirical process for finding the DNN structure.

No.	MAE (Testing)	Number of Batches	Structure
1	19.63	4	4/9/18/12/10
2	21.02	4	4/9/18/18/12/12/10
3	23.2	4	4/9/18/18/12/10
4	20.84	4	4/18/18/12/12/12/10/10
5	19.81	8	8/18/18/12/12/12/10/10
6	19.64	8	8/18/18/12/12/12/10/10
7(Selected)	17.51	8	18/18/12/12/12/10/10/8
8	18.28	10	18/18/12/12/12/10/10/8
9	19.05	8	18/18/18/12/12/12/10/10/8

The MAE with training and validation process is shown in Table 6. Table 7 shows the predictive results for the testing data through MAE and MAPE analysis. In order to show the overall error variation according to the application of the data augmentation, the number of data augmentation was applied up to 10 times. The reason for doing data augmentation up to 10 times is that too much data augmentation can make the original data meaningless. The dataset for DNN learning is increased by applying noise to the original data. In addition, gaps in the original data can be filled with noise-added data through the data augmentation technique. For these reasons, the DNN used in this study learns a more abundant rainfall-runoff pattern by data augmentation more than using only simulated results. Considering that abundant data is a fundamental condition of deep learning models, it is expected that

the predictive power of the suggested model would be improved. As data augmentation was applied, the training, validation, and testing error were decreased (Tables 6 and 7). In the case of the MAE, the overall decrease was observed without significant variation. However, in the case of the MAPE, which shows the error ratio of the observed value and the variability appeared to be large depending on the number of times data augmentation was applied.

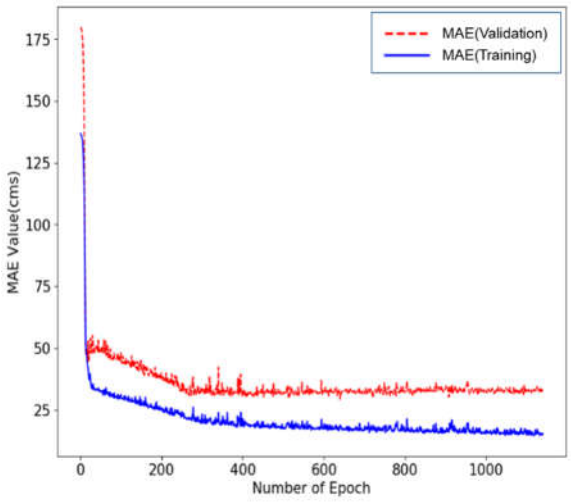
Table 6. Training and validation error with data augmentation.

Training and Validation Result		Input Criterion							
		CASE 1 (MAE)		CASE 2 (MAE)		CASE 3 (MAE)		CASE 4 (MAE)	
		Train.	Val.	Train.	Val	Train.	Val	Train.	Val
Original DNN		17.13	25.24	10.79	14.80	8.42	11.28	25.18	18.32
Number of Data Augmentations Applied	1	10.40	15.04	4.01	10.27	6.68	9.83	9.80	21.71
	2	7.45	11.17	5.09	7.69	3.81	5.65	10.37	13.52
	3	5.88	6.81	5.43	5.29	3.84	4.70	8.63	11.79
	4	6.69	7.25	4.30	5.33	3.98	4.48	7.98	11.25
	5	5.33	5.01	4.91	4.23	3.82	3.25	7.24	6.08
	6	5.06	4.71	3.16	3.28	3.11	3.08	7.40	7.77
	7	5.1	7.74	3.06	4.62	3.02	4.88	6.43	8.83
	8	4.92	3.94	3.24	2.42	2.95	2.70	7.15	6.04
	9	5.12	6.15	2.82	3.09	3.6	2.94	5.60	5.80
	10	4.17	4.71	3.43	3.74	3.31	3.97	6.04	5.78

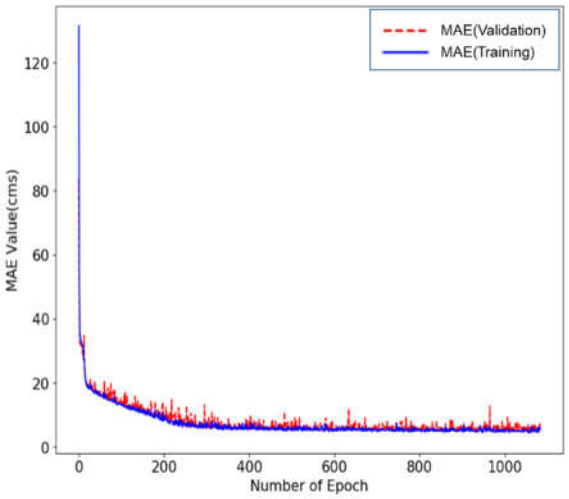
Table 7. Testing error with data augmentation.

Testing Result		Input Criterion							
		CASE 1		CASE 2		CASE 3		CASE 4	
		MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Original DNN		23.87	61.75	13.02	37.83	11.11	36.89	23.77	49.30
Number of Data Augmentations Applied	1	13.91	46.94	7.92	48.75	5.90	21.31	27.32	71.77
	2	10.06	11.04	5.41	6.18	4.75	6.46	12.75	14.28
	3	10.6	9.48	7.12	7.13	5.99	5.46	14.74	14.72
	4	8.33	12.62	4.20	6.73	3.92	6.76	11.80	22.24
	5	9.17	16.29	4.44	18.99	5.46	35.13	8.40	27.83
	6	6.23	57.27	3.39	6.96	6.21	14.34	7.96	12.97
	7	5.13	15.91	4.12	11.19	2.77	8.07	6.54	8.93
	8	1.96	11.56	3.77	27.99	3.77	7.24	7.27	42.32
	9	6.14	13.56	3.07	21.74	3.15	22.83	6.44	32.0
	10	4.46	34.42	3.38	11.49	4.22	7.83	4.15	7.02

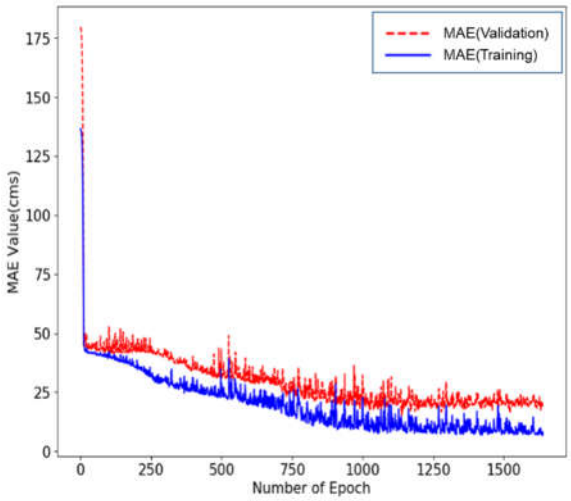
Figure 6 shows the results of DL training for each combination of the input data after applying the initial input data and data augmentation. The horizontal axis of each graph represents an epoch, and the vertical axis represents the mean absolute error (MAE). The red dotted line shows the graph for the validation data set, and the blue solid line shows the error graph for the training data set. Each graph shows that after applying data augmentation, predictive power is improved not only for the training data set for the DL model, but also for the validation data set as well. In other words, the data augmentation technique improves the analysis result of validation data, rather than simply over-fitting with the training data.


MAE(Validation)
MAE(Training)
MAE Value(cms)
Number of Epoch

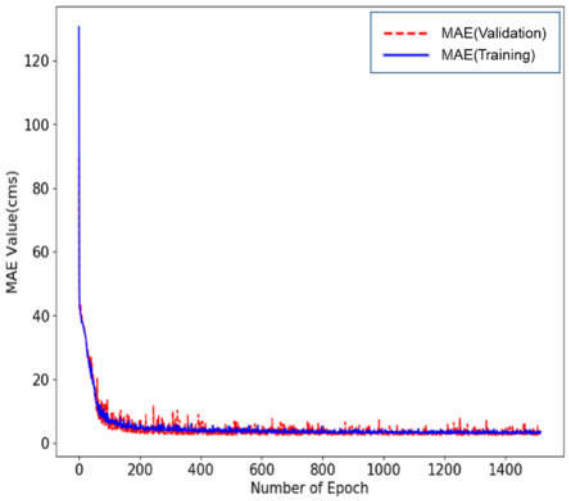
(**a**)


MAE(Validation)
MAE(Training)
MAE Value(cms)
Number of Epoch

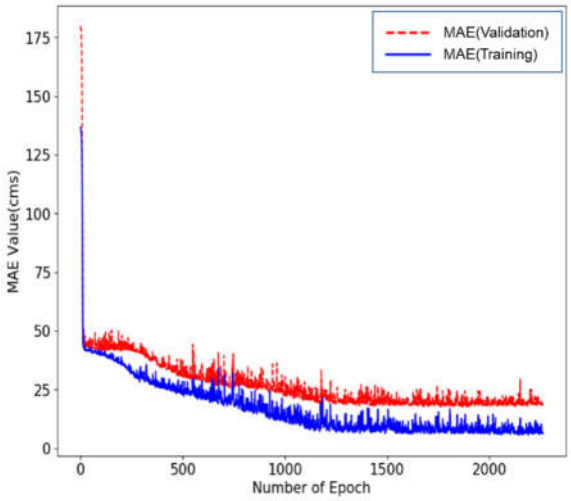
(**b**)


MAE(Validation)
MAE(Training)
MAE Value(cms)
Number of Epoch

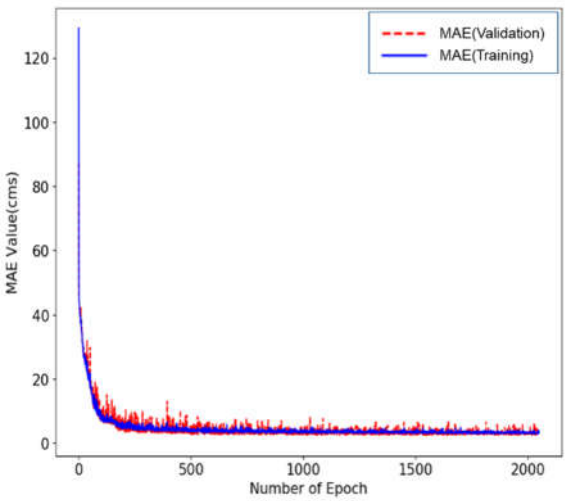
(**c**)


MAE(Validation)
MAE(Training)
MAE Value(cms)
Number of Epoch

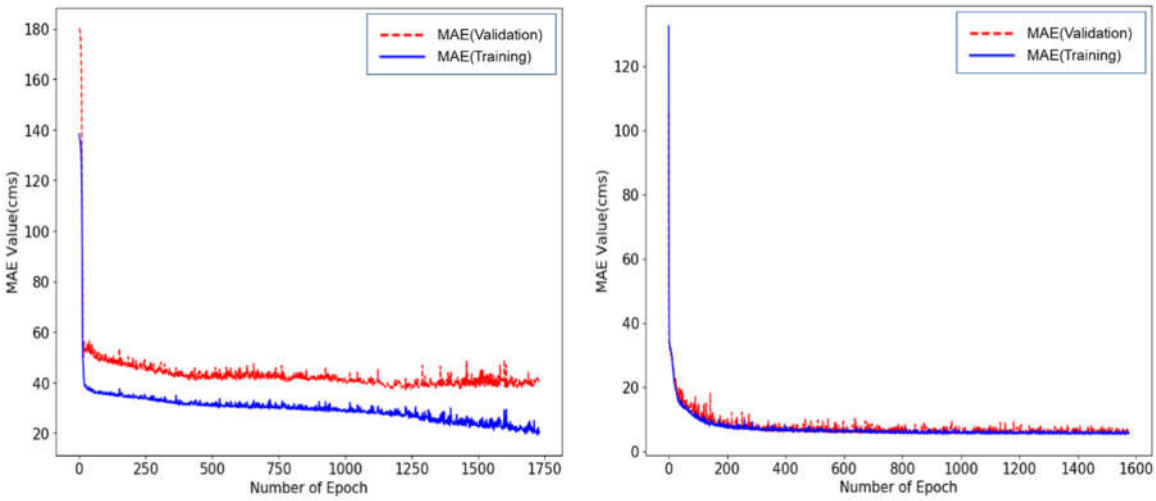
(**d**)


MAE(Validation)
MAE(Training)
MAE Value(cms)
Number of Epoch

(**e**)


MAE(Validation)
MAE(Training)
MAE Value(cms)
Number of Epoch

(**f**)



(**g**) (**h**)

Figure 6. The trend of mean absolute error (MAE) with validation and testing data. (**a**) Shows Case 1: original data, (**b**) shows Case 1: after the 10th data augmentation, (**c**) shows Case 2: original data, (**d**) shows Case 2: after the 10th data augmentation, (**e**) shows Case 3: original data (**f**) shows Case 3: after the 10th data augmentation, (**g**) shows Case 4: original data and (**h**) shows Case 4: after the 10th data augmentation.

3.5. Prediction Results

The result of SWMM simulations with consideration of the observed rainfall on 27 July 2011, was used as the target data for total accumulative overflow prediction (Figure 7). The DNN trained with 69 rainfall scenarios was used to perform the prediction. The total rainfall for the target rainfall was 184.5 mm, the rainfall intensity was 30.75 mm/h, the 1-, 2-, and 3-hour maximum rainfall was 71.0, 105.5, and 143.0 mm, respectively, the standard deviation was 4.38 mm, the skewness was 1.27, the kurtosis was 1.23, and the inter-event time was 10 min. As a result of EPA-SWMM simulations, the total accumulated overflow was calculated as 144.7 m^3/s. Table 6 shows the predictive results before applying data augmentation and the predictive results of the DNN according to the number of data augmentations applied. When the data augmentation was applied four times for the input data of CASE 1, seven times for CASE 2, ten times for CASE 3, and once for CASE 4, predictions were the closest to the results of SWMM. The absolute average was calculated for the correlation values between the target value and the statistical characteristics in each input data (refer to Table 4). They were calculated as 0.4395, 0.5366, 0.5715, and 0.3877 in CASE 1, CASE 2, CASE 3, and CAS 4, respectively, and the higher the mean of the correlation coefficients was, the lower was the standard deviation of the predictive results according to data augmentation. The absolute average correlation is shown in Table 8.

In the case of CASE 4, which had the lowest absolute mean value of the correlation coefficient, the difference with the results of SWMM was confirmed to increase as data augmentation was applied. In contrast, in CASE 3, which had the highest mean value of the correlation coefficient, it was shown that predictive power could be increased by attempting data augmentation. In Table 8, * (asterisk) indicates that the predictive results of the DNN for each input data combination was the closest to the results of SWMM.

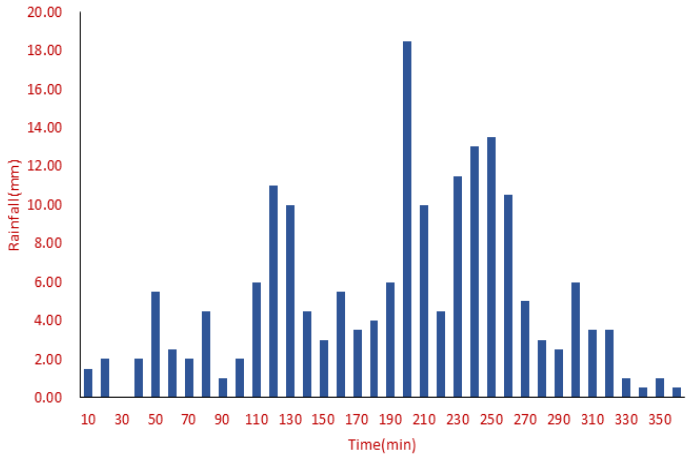


Figure 7. Observed rainfall on 27 July 2011.

Table 8. Results of the simulation and prediction.

Total Accumulative Overflow (m^3/s)		Input Criterion			
		CASE 1	**CASE 2**	**CASE 3**	**CASE 4**
Average of Absolute R-square (refer to R^2 in Table 4)		0.4395	0.5366	0.5715	0.3877
Simulated with SWMM				144.7	
Prediction with Original DNN		168.66	159.10	142.15	149.90
Predict with Data Augmentation	1	174.92	147.67	143.36	145.61 *
	2	149.27	160.80	152.75	140.37
	3	130.88	166.44	149.01	134.75
	4	146.63 *	148.31	150.18	132.25
	5	152.48	147.14	146.77	119.99
	6	154.49	148.82	160.14	142.34
	7	162.85	143.57 *	147.71	136.49
	8	129.85	160.21	160.21	136.32
	9	151.82	149.65	162.69	137.04
	10	147.39	142.69	144.51 *	138.47
Standard Deviation		13.21	7.58	6.32	6.47

4. Conclusions

In general, there is a lack of historical data on flooding for a single urban drainage basin and basic data for constructing a data-driven model. In this study, a DNN was used to predict the total accumulative overflow in an urban drainage basin. Cases in which significant damage was caused by flooding in the target area were limited to the events of 21 September 2010, and 27 July 2011, and therefore an additional 6-hour rainfall observation on the national scale was applied. SWMM was used to calculate the total accumulative overflow in 10-minute units. Data augmentation was used to increase the quantitative amount of input data, and the variation in the predictive performance was examined for each of the input data conditions. The main findings are as follows.

(1) Flood analysis was performed on the drainage basin of Samseong-dong, Seoul. The total accumulative overflow results for each heavy rainfall event were calculated using 70 observed rainfall events and EPA-SWMM. The characteristics of each rainfall event were analyzed and the correlation with the total accumulative overflow was calculated. As a result, rainfall characteristics that had a high correlation with urban outflow phenomena were identified. Although the highest correlation was found for the total rainfall and rainfall intensity, it was also found that maximum rainfall in 1- to 3-hour units during 6 hours of heavy rainfall was also highly correlated. It was found that the peak rainfall location did not have significant influence during rainfall events

and that kurtosis, skewness, and rainfall time for rainfall events had a negative correlation with urban runoff.

(2) As data augmentation was applied, it was found that the mean absolute error (MAE) and mean absolute percentage error (MAPE) values of the predictive results for testing data decreased for all input data combinations, which verified that prediction performance for data that were not applied to the training can be sufficiently improved using data augmentation. The largest difference of error analysis between the initial input condition (predicted with original DNN, Table 6) and after applying the 10th data augmentation was indicated at the CASE 4-based DNN model. Although input data with low correlation was used in CASE 4, it was judged that data augmentation could be helpful to partially overcome the poor predictive power.

(3) The total accumulative overflow for the rainfall event on 27 July 2011, was predicted using the DNN constructed according to input data combinations. Predictions were made according to the data augmentation method, and the predictions of CASE 3 using a highly correlated input data combination were the closest to the results of SWMM. In the case of CASE 4, which used the least correlated input data, as more data augmentations were applied, the poorer the predictive results were. Although data augmentation can be used to make up for the lack of input data and reduce errors in learning, it is necessary to conduct proper correlation analysis between the input data and the target value data beforehand. It took 14 minutes for the one-dimensional urban runoff analysis of the SWMM model to be completed; however, predictions using the DNN took 2–3 seconds. In an event of heavy rainfall causing actual flooding, it is likely to save a lot of time in estimating the degree of urban flooding.

(4) By successfully performing predictions using observed rainfall data and applying data augmentation, basic research on data supplementation techniques in the data-based analysis could be performed. The result of the proposed DNN model is expected to be used as basic data for the real-time flood response in urban areas. If a predictive model is constructed not only for the drainage basin in Samseong-dong but also for all drainage basins in Seoul, it seems that it would be possible to use practically for the entire flood forecasting and warning system in Seoul. Furthermore, if the predicted total accumulative overflow is linked with the expected inundation map, the rapid-simulation of a two-dimensional flood map could be possible. This methodology would be helpful to identify the flood risk area in an urban watershed.

Author Contributions: Conceptualization, H.I.K.; Methodology, H.I.K.; Writing original draft, H.I.K.; Supervision, K.Y.H.; Writing-review & editing, K.Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Environment Industry & Technology Institute(KEITI) though Water Management Research Program, funded by Korea Ministry of Environment(MOE)(79609).

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

References

1. Mosavi, A.; Ozturk, P.; Chau, K.W. Flood Prediction Using Machine Learning Models: Literature Review. *Water* **2018**, *10*, 1536. [CrossRef]
2. Granata, F.; Gargano, R.; Marinis, G. Support Vector Regression for Rainfall-Runoff Modeling in Urban Drainage: A Comparison with the EPA's Storm Water Management Model. *Water* **2016**, *8*, 69. [CrossRef]
3. Talei, A.; Chua, L.H.C.; Wong, T.S.W. Evaluation of Rainfall and Discharge Inputs used by Adaptive Network-based Fuzzy Inference Systems (ANFIS) in rainfall-runoff modeling. *J. Hydrol.* **2010**, *391*, 248–262. [CrossRef]
4. Shen, C. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resour. Res.* **2018**, *54*, 8558–8593. [CrossRef]
5. Hu, C.; Wu, Q.; Li, H.; Jian, S.; Li, N.; Lou, Z. Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water* **2018**, *10*, 1543. [CrossRef]

6. Li, X.; Willems, P. A Data-Driven Hybrid Urban Flood Modeling Approach. In *EPiC Series in Engineering, HIC 2018, Proceedings of the 13th International Conference on Hydroinformatics, Palermo, Italy, 1–6 July 2018*; Loggia, G.L., Freni, G., Puleo, V., Marchis, M.D., Eds.; EasyChair: Manchester, UK, 2018; Volume 3, pp. 1193–1200.
7. Nikhil, B.C.; Arjun, N.; Keerthi, C.; Sreerag, S.; Ashwin, H.N. Flood Prediction using Flow and Depth Measurement with Artificial Neural Network in Canals. In Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), Erode, India, 27–29 March 2019; pp. 798–801.
8. Korea Meteorological Agency Meteorological Database. 2020. Available online: https://data.kma.go.kr (accessed on 2 September 2019).
9. Yoon, S.S.; Bae, D.H.; Choi, Y.J. Urban Inundation Forecasting Using Predicted Radar Rainfall: Case Study. *J. Korean Soc. Hazard Mitig.* **2014**, *14*, 117–126. [CrossRef]
10. Huber, W.C.; Dickson, R.E. *Storm Water Management Model. User's Manual Version 4*; Environmental Protection Agency: Washington, DC, USA, 1988.
11. United States Environmental Protection Agency (EPA). *Storm Water Management Model User's Manual Version 5.0*; Environmental Protection Agency: Washington, DC, USA, 2010.
12. Park, J.H.; Kim, S.H.; Bae, D.H. Evaluating Appropriateness of the Design Methodology for Urban Sewer System. *J. Korea Water Resour. Assoc.* **2019**, *52*, 411–420.
13. Mark, O.; Weesakul, S.; Apirumanekul, C.; Arronnet, S.B.; Djordjevic, S. Potential and Limitations of 1D Modeling of Urban Flooding. *J. Hydrol.* **2004**, *299*, 284–299. [CrossRef]
14. Izumi, T.; Miyoshi, M.; Kobayashi, N. Runoff Analysis Using a Deep Neural Network. In Proceedings of the 12th International Conference on Hydroscience & Engineering, Hydro-Science & Engineering for Environmental Resilience, Taiwan, China, 6–10 November 2016.
15. Agarap, A.F. Deep Learning using Rectified Linear Units (ReLU). *arXiv* **2019**, arXiv:1803.08375v2.
16. Kingma, D.P.; Ba, J.L. ADAM: A Method for Stochastic Optimization. *arXiv* **2015**, arXiv:1412.6980.
17. Cui, X.; Goel, V.; Kingsbury, B. Data Augmentation for Deep Neural Network Acoustic Modeling. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1469–1477.
18. Xu, Y.; Jia, R.; Mou, L.; Li, G.; Chen, Y.; Lu, Y.; Jin, Z. Improved Relation Classification by Deep Recurrent Neural Networks with Data Augmentation. *arXiv* **2016**, arXiv:1601.03651v2.
19. Seoul Metropolitan City. *Comprehensive Plan for Storm and Flood Damage Reduction*; Seoul Metropolitan City: Seoul, Korea, 2015; Volume 1, pp. 374–375.
20. Risi, R.D.; Jalayer, F.; Paola, F.D. Meso-scale hazard zoning of potentially flood prone areas. *J. Hydrol.* **2015**, *527*, 316–325. [CrossRef]
21. Moore, M.R. Development of a High-Resolution 1D/2D Coupled Flood Simulation of Charles City, Iowa. Master's Thesis, University of Iowa, Iowa City, IA, USA, 2011.
22. Son, A.L.; Kim, B.H.; Han, K.Y. A study on prediction of inundation area considering road network in urban area. *J. Korean Soc. Civ. Eng.* **2015**, *35*, 307–318. [CrossRef]
23. Korea Institute of Civil Engineering and Building Technology. *Road Design Manual*; KICT: Korea, Seoul, 2001.