atmosphere

MDPI

Article

Typhoon Quantitative Rainfall Prediction from Big Data Analytics by Using the Apache Hadoop Spark Parallel Computing Framework

Chih-Chiang Wei * and Tzu-Hao Chou

Department of Marine Environmental Informatics & Center of Excellence for Ocean Engineering, National Taiwan Ocean University, Keelung 20224, Taiwan; 10781001@mail.ntou.edu.tw

* Correspondence: ccwei@ntou.edu.tw

Received: 4 July 2020; Accepted: 15 August 2020; Published: 17 August 2020

Abstract: Situated in the main tracks of typhoons in the Northwestern Pacific Ocean, Taiwan frequently encounters disasters from heavy rainfall during typhoons. Accurate and timely typhoon rainfall prediction is an imperative topic that must be addressed. The purpose of this study was to develop a Hadoop Spark distribute framework based on big-data technology, to accelerate the computation of typhoon rainfall prediction models. This study used deep neural networks (DNNs) and multiple linear regressions (MLRs) in machine learning, to establish rainfall prediction models and evaluate rainfall prediction accuracy. The Hadoop Spark distributed cluster-computing framework was the big-data technology used. The Hadoop Spark framework consisted of the Hadoop Distributed File System, MapReduce framework, and Spark, which was used as a new-generation technology to improve the efficiency of the distributed computing. The research area was Northern Taiwan, which contains four surface observation stations as the experimental sites. This study collected 271 typhoon events (from 1961 to 2017). The following results were obtained: (1) in machine-learning computation, prediction errors increased with prediction duration in the DNN and MLR models; and (2) the system of Hadoop Spark framework was faster than the standalone systems (single I7 central processing unit (CPU) and single E3 CPU). When complex computation is required in a model (e.g., DNN model parameter calibration), the big-data-based Hadoop Spark framework can be used to establish highly efficient computation environments. In summary, this study successfully used the big-data Hadoop Spark framework with machine learning, to develop rainfall prediction models with effectively improved computing efficiency. Therefore, the proposed system can solve problems regarding real-time typhoon rainfall prediction with high timeliness and accuracy.

Keywords: typhoon; precipitation; machine learning; big data; Hadoop; Spark

1. Introduction

Taiwan is an island in East Asia. The latitude and longitude of Taiwan are 21° N–25° N and 120° E–122° E, respectively. As shown in Figure 1, situated in the main tracks of typhoons in the Northwestern Pacific Ocean, Taiwan is frequently stricken by typhoons and heavy rainfall [1]. Tropical cyclones form in tropical oceans, and nearly 90% of tropical cyclones form on sea surface of 27 °C in regions that are approximately 20° in latitude. Approximately 80 typhoons are generated annually in the world, and typhoons from the Northwestern Pacific Ocean are the strongest [2]. The typhoon brings abundant rainwater that fills the reservoir, and it also causes losses of life, including in flooding in some areas and landslides.

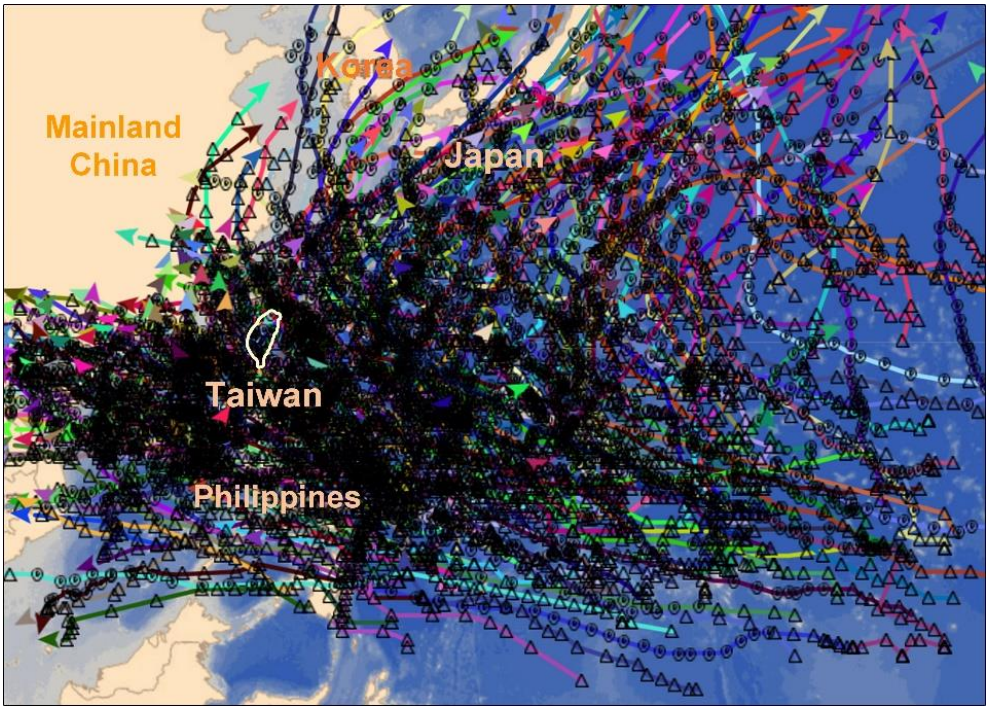

Figure 1. Historical typhoon events (2008–2017) that affected East Asia (The map was produced by the Central Weather Bureau, Taiwan. http://www.cwb.gov.tw).

This study established a typhoon rainfall prediction model to predict rainfall quantity when typhoons strike. The prediction model can help people of all industries prepare for heavy rainfall and prevent disasters in advance, to reduce economic loss and casualties. This study used Northern Taiwan as the research area. Records of past typhoon events indicated that when the typhoon center passes through the land of Northern Taiwan, it often brings heavy rainfall and causes major disasters [3]. For example, Typhoon Herb in 1996 caused 51 deaths, and 22 people went missing. In 2008, Typhoon Jangmi caused two deaths, and two people went missing. Thus, establishing a real-time and accurate typhoon rainfall prediction model is imperative.

Following the emergence of artificial intelligence in recent years, machine learning enables machines learn to the rules from the input data in various algorithms that are similar to the rules of thumb that are generated from computer autolearning. Machine learning can be applied for big-data analysis. Therefore, machine learning has been extensively applied in various fields in recent years [4–6]. In terms of the development of rainfall prediction models, the conventional linear regression model has been constantly applied because of its comprehensive statistical theory foundation [7–9]. Machine-learning models include artificial neural networks, decision trees, support vector regressions, random forests, and Bayesian networks [10–20]. In addition, hybrid machine-learning models have been developed, such as the adaptive network-based fuzzy inference system [21]. Moreover, Maier and Dandy [22], Antolik [23], Maier et al. [24], Madsen et al. [25], Maçaira et al. [26], and Paulo Vitor de Campos Souza [27] have conducted in-depth reviews of rainfall prediction models and water-resources-related models, using machine-learning algorithms. Newly developed models in recent years, such as deep neural networks (DNNs), convolutional neural networks, and deep belief networks, can be applied to image processing for atmospheric science-related research (e.g., rainfall retrieval, typhoon track prediction, and wind speed prediction [28–31]).

To facilitate the efficient computation of machine-learning algorithms, this study used a popular big-data technology—the Hadoop Spark distributed computing framework, which is a cost-efficient and feasible parallel computing system—as a feasible option [32,33]. Hadoop is a computing platform for storing and managing large sets of data and originates from the open-source code of the Apache

Software Foundation (ASF) [34]. The ASF is an American nonprofit corporation that supports Apache software projects [35]. Hadoop was originally designed to handle massive amounts of data efficiently and inexpensively. It has the intelligence necessary to run the distributed file system and parallel processing work [36]. Hadoop system consists of the Hadoop Distributed File System (HDFS) [37] and MapReduce framework [38,39] (see Section 3.2 for the concepts behind HDFS and MapReduce). Furthermore, Apache Spark is a flexible cluster-computing framework that was originally developed by AMPLab at the University of California, Berkeley [40]. Spark is a new-generation big-data computing framework that can improve the computational efficiency of MapReduce. The difference between Spark and MapReduce is that MapReduce accesses data in hard disks, whereas the computation using Spark involves memories. According to Xin et al. [41], the computational speed of Spark is approximately 10 times faster than that of MapReduce.

The literature on efforts to reduce computational time in water resources and engineering applications can be found; for example, Hu et al. [42] presented a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis, using Hadoop. The total running time of the coupled models is reduced by 80%. Hu et al. [43] presented a framework for the global sensitivity analysis to socio-hydrological models. This framework can find a balance between the heavy computational burden associated with the model execution and the number of model evaluations. The balance was achieved through the combination of Hadoop-based cloud computing and polynomial chaos expansion. Qureshi and Koubaa [33] investigated the single-board computer-based clusters in energy efficient data centers in the context of big-data applications. Hadoop was deployed on two low-cost single-board computer-based clusters, using Raspberry Pi and Odroid Xu-4 platforms. In terms of power efficiency, for smaller workloads, the Xu-4 cluster outperforms the Raspberry Pi and HDM clusters; and for low-intensity workloads, the Xu-4 cluster fares 37% better than the HDM Cluster.

To the best of our knowledge, applications in rainfall estimation and prediction using Hadoop Spark big-data technology are scarce. This suggested that, given the thriving development of big-data technology, the novel Hadoop Spark system could provide a potential to increase the performance of data storage and reduce the computing time required for data analysis and models' building. The Hadoop Spark system was proposed to improve computing efficiency, as required by rainfall prediction models, thereby enabling real-time typhoon rainfall prediction that requires high timeliness. Accordingly, this study had two objectives: (1) use machine-learning models to establish typhoon prediction models and improve rainfall prediction accuracy, and (2) use the newly Hadoop Spark big-data computing framework to accelerate machine-learning computation and improve the timeliness of real-time rainfall prediction.

2. Data Sources and Preprocessing

The research area was the Taipei Metropolitan Area in Northern Taiwan (red circle in Figure 2). The Taipei Metropolitan Area is the most populated area in Taiwan (approximately 6.95 million people). The four surface observation stations (i.e., Tamsui, Anbu, Taipei, and Keelung) in this area were used as the experimental sites. The Anbu and Taipei stations were located in the Taipei Basin, and the Tamsui and Keelung stations were in coastal areas.

The data source was the Central Weather Bureau (CWB) of Taiwan. This study collected ground meteorological data from seven surface observation stations (the four target stations in the research area and three adjacent stations, i.e., Pengjiayu, Su-ao, and Yilan, near research area) and typhoon climatological data from typhoon warning sheets.

Table 1 presents the latitude, longitude, and altitude of all surface observation stations. Data from 1961 to 2017 were collected (a total of 271 typhoon incidents in 57 years are listed in Appendix A Table A1). According to the CWB typhoon classification, 79 severe typhoons (defined as typhoons with a maximum wind speed at the center reaching 51.0 m/s or higher), 119 moderate typhoons (32.7–51.0 m/s), and 73 mild typhoons (32.7 m/s or lower) were observed.

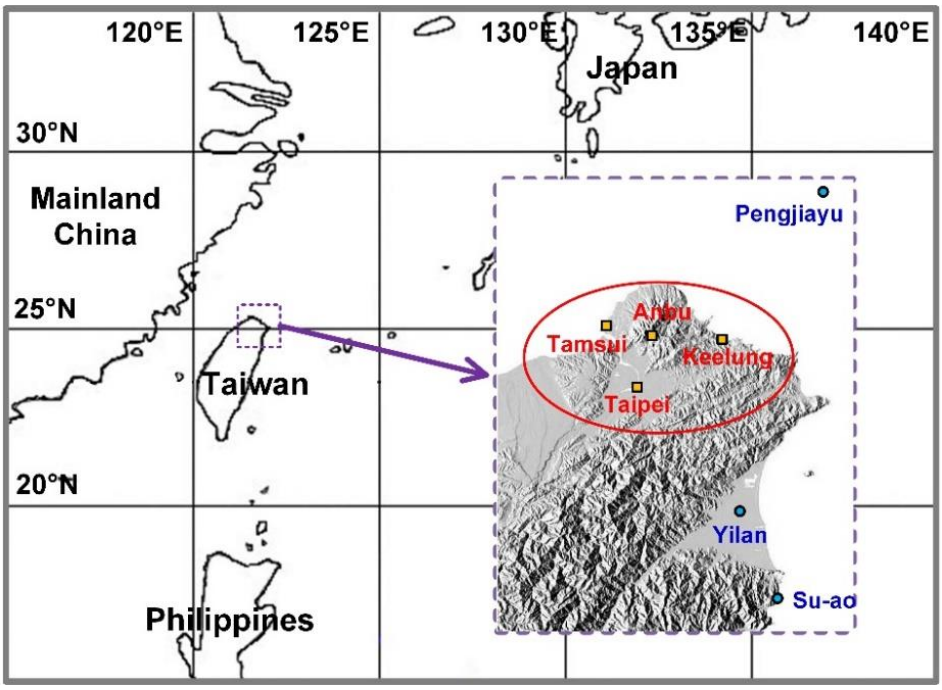

Figure 2. Research area location and observation station distribution.

Table 1. Geographic location and altitude of surface observation stations.

Station	Latitude (°N)	Longitude (°E)	Altitude (m)
Tamsui	25.1649	121.4489	19
Anbu	25.1826	121.5297	826
Taipei	25.0377	121.5149	7
Keelung	25.1333	121.7405	27
Pengjiayu	25.6280	122.0797	102
Su-ao	24.5967	121.8574	25
Yilan	24.7640	121.7565	8

Typhoon attributes on typhoon warning sheets include pressure in the typhoon center (code: W_1), distance between the typhoon center and the target station (W_2), radius of winds over 15.5 m/s (W_3), moving speed of the typhoon (W_4), and maximum wind speed of the typhoon center (W_5). The W_2 can be obtained by calculating the latitudes and longitudes of the typhoon center and target stations. The attribute data statistics of typhoon warning sheets are provided in Appendix A Table A2. The meteorological attributes of surface observation stations include the air pressure on the ground (Y_1), temperature on the ground (Y_2), dew point on the ground (Y_3), relative humidity (Y_4), vapor pressure on the ground (Y_5), surface wind velocity (Y_6), surface wind direction (Y_7), distance from the typhoon center (Y_8), and precipitation (Y_9). The attribute data statistics of typhoon warning sheets are provided in Appendix A Tables A3 and A4.

This study used correlation analysis to select input variables that were suitable for the target stations. According to Reference [44], a correlation coefficient of $|r| \geq 0.3$ represents moderate to high correlation, and $|r| < 0.3$ represents low correlation. Therefore, this study adopted whether the correlation coefficient between the attribute data and rainfall of each target station was greater or equal to 0.3 as the selection criteria. As presented in Table 2, the target stations of Tamsui, Anbu, Taipei, and Keelung, respectively, selected 15, 20, 15, and 12 attributes. The results revealed that Y_4 and Y_8 in each station were crucial attributes that manifested in the high correlation between the distance from the typhoon center and relative humidity to rainfall.

Table 2. Selected attributes of each target station.

Target Station	Selected Attributes
Tamsui	W_2, Y_4, and Y_8 of Tamsui; Y_8 of Anbu; Y_4 and Y_8 of Taipei; Y_4, Y_6, and Y_8 of Keelung; Y_4, Y_6, and Y_8 of Pengjiayu; Y_1 of Su-ao; Y_1 and Y_8 of Yilan
Anbu	Y_2, Y_4, Y_6, and Y_8 of Tamsui; W_2 and Y_8 of Anbu; Y_2, Y_4, and Y_8 of Taipei; Y_2, Y_4, Y_6, and Y_8 of Keelung; Y_6 and Y_8 of Pengjiayu; Y_1 and Y_8 of Su-ao; Y_1, Y_6, and Y_8 of Yilan
Taipei	Y_4 and Y_8 of Tamsui; Y_1 and Y_8 of Anbu; W_2, Y_4, and Y_8 of Taipei; Y_4, Y_6, and Y_8 of Keelung; Y_1 and Y_6 of Pengjiayu; Y_1 of Su-ao; Y_1 and Y_8 of Yilan
Keelung	Y_8 of Tamsui; Y_8 of Anbu; Y_4 and Y_8 of Taipei; W_2, Y_4, Y_6, and Y_8 of Keelung; Y_6 and Y_8 of Pengjiayu; Y_8 of Su-ao; Y_8 of Yilan

3. Methodology

This study designed a big-data computing framework analysis system to estimate rainfall during typhoons. Figure 3 displays the flowchart of the design. The flowchart consists of three sections of main tasks: data preprocessing, the computing environment, and modeling and evaluation.

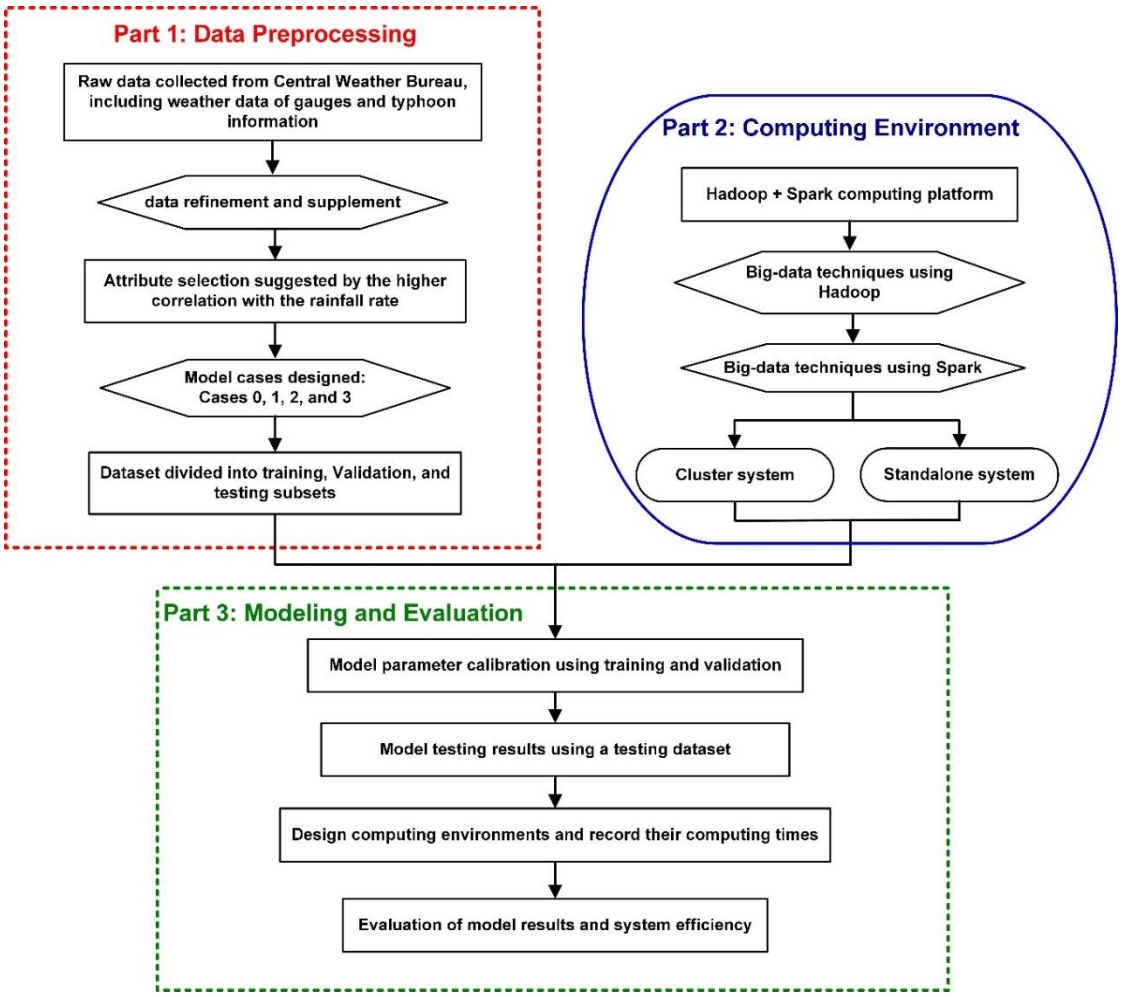

Figure 3. Method design flowchart.

The first section involved the collection of data associated with historical typhoon events in the study area. Data collection, data processing, and attribute selection were conducted according to aforementioned methods (Section 2). This study setup several cases to establish rainfall prediction models. Case 1 used all data records regarding typhoon events that affected Taiwan. The number of 16,957 hourly records was collected from 1961 to 2017 (totally 271 typhoons), in this case. Case 2 involved typhoons that passed through the research area. We found that 101 typhoons (6549 hourly records) can be refined by the collected 271 typhoons of Case 1. In addition, rainfall intensity that reaches heavy rain (rainfall quantity reaching 40 mm/h) can easily cause disasters, according to Reference [45]. Therefore, Case 3 was constituted of selected typhoon events that reached the level of heavy rain. We found that 80 typhoons (4771 hourly records) were refined by the collected 271 typhoons of Case 1. In terms of data segmentation, this study divided the data into training, validation, and testing sets (as illustrated in Section 3.1).

The second section involved developing a computing platform of the Apache Hadoop Spark 2.0 distributed parallel computing framework. The development process of the framework is illustrated in Section 3.2. This study assigned a computer as the master (or NameNode) to establish multiple clusters of the Hadoop Spark computing framework to facilitate the management of multiple slave computers (or DataNode) and allocate all computational resources, thereby optimizing computing efficiency.

The third section adopted the machine-learning model to establish rainfall prediction models. This study used DNNs and multiple linear regressions (MLRs) to establish prediction models under the Hadoop Spark distributed framework. The model-training process is explained in Section 4. This study used the R programming language-based integrated development environment software RStudio to establish machine-learning models. RStudio and the Hadoop Spark distributed framework were connected to facilitate modeling and computation tasks. This study recorded execution times under a standalone computer and the Hadoop Spark parallel computing framework to evaluate the efficiency of the Hadoop system. In addition, the quality of the prediction results was evaluated by using the following indicators, including the root mean squared error (RMSE), relative RMSE (rRMSE), mean absolute error (MAE), relative MAE (rMAE), and coefficient of determination (R-Squared). The equations are as follows.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\left(p_i^{\mathrm{Pre}} - p_i^{\mathrm{Obs}}\right)^2 / n} \tag{1}$$

$$\mathrm{rRMSE} = \frac{\mathrm{RMSE}}{\overline{p^{\mathrm{Obs}}}} \tag{2}$$

$$\mathrm{MAE} = \frac{1}{n} \times \sum_{i=1}^{n}\left|p_i^{\mathrm{Pre}} - p_i^{\mathrm{Obs}}\right| \tag{3}$$

$$\mathrm{rMAE} = \frac{\mathrm{MAE}}{\overline{p^{\mathrm{Obs}}}} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(p_i^{\mathrm{Obs}} - p_i^{\mathrm{Pre}}\right)^2}{\sum_{i=1}^{n}\left(p_i^{\mathrm{Obs}} - \overline{p^{\mathrm{Obs}}}\right)^2} \tag{5}$$

where n is the total sets of data, p_i^{Obs} is the ith observation value, p_i^{Pre} is the ith prediction value, and p^{Obs} is the mean of the observation values.

3.1. Data Division

The duration of the prediction time in this study was 1–6 h. The predicted rainfall quantity of the ith hour is $(\hat{p}_{t+i})_{i=1,6}$. As analyzed previously, the predictor variables at each target station of lead time = 1–6 h were the same as in Table 2. This study divided the data proportionally into training, validation,

and testing sets. In Cases 1–3, approximately 90% of typhoon events were randomly selected as training and validation sets in a ratio of 7:3 during the data segmentation process. The remaining 10% were used as the testing sets.

During data segmentation, 27 typhoon events were selected as the testing sets from the events ranked as having "heavy rain" regarding rainfall intensity to maintain consistency among the testing sets of all model cases. The typhoons in the testing sets were Pamela (1961), Elsie (1969), Fran (1970), Bess (1971), Amy (1977), Maury (1981), Jeff (1985), Susan (1988), Yancy (1990), Herb (1996), Otto (1998), Xangsane (2000), Nakri (2002), Nangka (2003), Haitang (2005), Longwang (2005), Pabuk (2007), Jangmi (2008), Parma (2009), Megi (2010), Nanmadol (2011), Tembin (2012), Kong-Rey (2013), Fung-Wong (2014), Goni (2015), Aere (2016), and Nesat (2017). The remaining 244 (approximately 90%) typhoons were divided randomly by a ratio of 7:3 into training and validation sets.

3.2. Computing Environment

This study established a Hadoop Spark distributed framework system. The HDFS can be extended from a single server to multiple servers. The NameNode is responsible for managing and maintaining the HDFS directory system and controlling the reading and writing of data. Multiple DataNodes are responsible for data storage. Figure 4a presents the concept map of the system in which DataNodes can be multiple clusters. The HDFS was designed to treat hardware malfunctioning as normal instead of abnormal, access streaming data, process large-scale datasets, simplify consistency models, prioritize mobile computation to mobile data, and develop cross-hardware and software platforms [46]. MapReduce is a parallel computing framework (Figure 4b). The computing process consists of two steps: Map and Reduce. The Map step divides works into subworks to be implemented by separate multiple DataNodes. The Reduce step integrates all DataNode computation results and transmits the final computational result back to the NameNode. The MapReduce method enables the parallel processing of a massive quantity of data on multiple devices, thereby considerably improving data-processing efficiency [47].

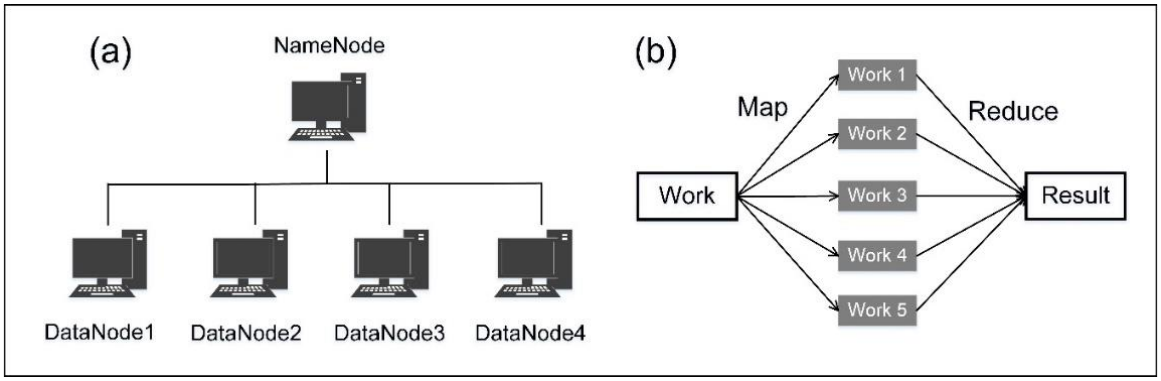

Figure 4. Schematic diagram of (**a**) HDFS system and (**b**) MapReduce framework.

This study tested the operating performance of a standalone computer and cluster systems, to understand the computing efficiency of the Hadoop distributed framework. Table 3 presents the hardware equipment (including brands, chip sets, and motherboards) of this study. The server computers of the cluster system adopted the E-series central processing unit (CPU), and that of the standalone computers adopted the I-series CPU. The clock rates of the server and standalone computers could reach 3.5 GHz, and the memory capacity was consistently DDR4-2400 16G.

This study connected four server computers in a series, to construct the Hadoop distributed framework system. During the system establishment process, we used Yet Another Resource Negotiator (YARN), which replaced the MapReduce engine in the first version of Hadoop and provided resource management and job scheduling in the Hadoop distributed processing platform [48]. Although YARN

is the data refinery layer and is a processing level for scheduling parallel computing jobs, this structure makes the complexities of distributed computing abstract [47].

Table 3. Hardware equipment of the experimental systems.

Equipment	Cluster System	Standalone PC
Brand and model	ASUS-TS300E9	GIGABYTE-P55
CPU	E3-1240v6 (3.5GHz)	I7-6700HQ (3.5GHz)
Chipset	Intel C236 Chipset	Intel C236 Chipset
Memory	DDR4-240016G	DDR4-240016G
Number of computers	4	1

For the system hardware equipment, a computer was designated as the master and functioned as the NameNode in the HDFS and ResourceManager in YARN. The remaining three computers (coded as data1, data2, and data3) functioned as DataNodes in the HDFS and NodeManager in YARN. Table 4 specifies the IP address and functions of master, data1, data2, and data3.

Table 4. Hadoop multi-node cluster.

Name	IP Address	HDFS	YARN
master	192.168.0.100	NameNode	ResourceManager
data1	192.168.0.101	DataNode	NodeManager
data2	192.168.0.102	DataNode	NodeManager
data3	192.168.0.103	DataNode	NodeManager

YARN = Yet Another Resource Negotiator.

4. Modeling and Evaluation

This section explains the establishment of the rainfall prediction models of the four target stations (i.e., Tamsui, Anbu, Taipei, and Keelung). During the DNN modeling process, three hidden layers were adopted, and the parameters to be calibrated included the learning rate and neuron number in the hidden layers. This study used the trial-and-error method to calibrate the parameters. First, the learning rate was fixed at 0.1, to test the neuron number in each hidden layer. In the first hidden layer, 1–20 neurons were tested. After the optimal neuron number of the first hidden layer was determined, the optimal neuron numbers of the second and third hidden layers were sequentially tested. After all neuron numbers of all the hidden layers were determined, the learning rate was calibrated. The learning rate was calibrated with an interval of 0.1, within a range from 0 to 1, to determine the optimal solution.

The various model parameters for forecasting horizons ranging from 1 to 6 h were separately calibrated by using the trial-and-error method (similar to the process in the one-hour-ahead forecast). Table 5 summarizes calibrated parameters of the lead time from 1 to 6 h in the three model cases.

For building MLR models, the attributes selected were the same as those used in the DNN models. The regression equation represents a straight line or plane that minimizes the squared differences between predicted and obtained output values [49]. This study used a stepwise regression method and specified selection criteria based on the statistical probability associated with each field. The criteria and stepwise estimation were used to add and remove fields [50,51].

4.1. Results and Comparisons

This section describes how testing sets were used to predict rainfall based on the DNN and MLR models. Figure 5 displays the performance of the evaluation indicators of the testing sets, including absolute error indicators (i.e., MAE and RMSE) and the squared correlation coefficient (R^2).

The evaluation results of the models in each station revealed that, 1–6 h in the future, the DNN model had superior performance to that of the MLR model in the Tamsui, Anbu, Taipei, and Keelung stations.

Table 5. Deep neural network (DNN) parameter calibrations of lead time = 1–6 h at target stations.

Station	Case	Parameter	Lead Time (h)					
			1	2	3	4	5	6
Tamsui	1	Layers 1–3	(6,4,6)	(2,1,5)	(1,2,6)	(1,2,1)	(1,1,1)	(1,1,1)
		Learning rate	0.3	0.1	0.2	0.1	0.4	0.2
	2	Layers 1–3	(3,2,4)	(2,3,2)	(2,1,3)	(3,2,4)	(3,3,1)	(3,3,5)
		Learning rate	0.6	0.4	0.3	0.1	0.1	0.1
	3	Layers 1–3	(1,6,4)	(2,5,4)	(3,7,3)	(1,3,1)	(1,3,6)	(1,3,4)
		Learning rate	0.2	0.2	0.1	0.1	0.1	0.1
Anbu	1	Layers 1–3	(7,4,8)	(6,3,4)	(6,4,4)	(6,3,2)	(4,5,6)	(5,5,7)
		Learning rate	0.4	0.4	0.3	0.4	0.3	0.3
	2	Layers 1–3	(3,6,4)	(3,5,4)	(3,4,5)	(3,3,4)	(6,4,5)	(5,3,6)
		Learning rate	0.5	0.3	0.2	0.2	0.5	0.4
	3	Layers 1–3	(4,7,7)	(3,5,7)	(3,6,7)	(2,5,1)	(3,5,3)	(5,4,6)
		Learning rate	0.4	0.3	0.4	0.4	0.3	0.2
Taipei	1	Layers 1–3	(4,3,5)	(1,3,3)	(1,2,2)	(1,2,1)	(1,3,5)	(1,1,1)
		Learning rate	0.3	0.3	0.3	0.2	0.1	0.1
	2	Layers 1–3	(2,6,2)	(2,5,7)	(1,5,4)	(2,3,5)	(2,5,2)	(4,4,1)
		Learning rate	0.7	0.1	0.1	0.2	0.1	0.1
	3	Layers 1–3	(1,8,6)	(1,5,3)	(2,6,7)	(2,5,4)	(5,1,1)	(2,3,3)
		Learning rate	0.6	0.1	0.1	0.1	0.3	0.1
Keelung	1	Layers 1–3	(3,1,1)	(3,2,2)	(5,1,1)	(4,2,2)	(4,4,2)	(5,4,4)
		Learning rate	0.1	0.2	0.1	0.3	0.1	0.1
	2	Layers 1–3	(6,4,2)	(4,3,5)	(3,3,5)	(3,2,3)	(2,3,7)	(1,2,5)
		Learning rate	0.1	0.3	0.1	0.1	0.1	0.1
	3	Layers 1–3	(5,5,6)	(5,3,5)	(3,1,2)	(4,2,1)	(2,3,3)	(1,3,2)
		Learning rate	0.1	0.1	0.1	0.1	0.1	0.1

To facilitate comparison between the model performance of different stations, we plotted rRMSE line charts of the future 1–6 h. The results are shown in Figure 6.

- In Case 1, the DNN model prediction result (Figure 6a) exhibited the most favorable performance in the Anbu station, followed by that in the Taipei, Tamsui, and Keelung stations. The MLR prediction results (Figure 6d) were most favorable in the Anbu station, followed by those in the Taipei, Keelung, and Tamsui stations (no significant difference was attained).
- In Case 2, the DNN prediction results (Figure 6b) were the most favorable in the Anbu station, followed by the Tamsui, Taipei, and Keelung stations. The MLR prediction results of the Anbu station were the most favorable (Figure 6e), followed by those in the Taipei, Keelung, and Tamsui stations (similar results).
- In Case 3, the DNN prediction results (Figure 6c) of the Anbu station were the most favorable, followed by those of the Taipei, Tamsui, and Keelung stations. The MLR prediction results of the Anbu station were the most favorable (Figure 6f), followed by those of the Taipei, Keelung, and Tamsui stations (with similar results).

These results revealed that the DNN model had a more favorable performance in the Anbu station than it did in the other stations. By contrast, the Anbu station also had the most favorable MLR results, and the other stations had similar results.

Figure 7 plots the overall prediction errors of 1–6 h. The overall average performance of the DNN model (Figure 7a) presented the lowest mean of the four stations in Case 3 (rRMSE = 1.982), the second lowest mean in Case 2 (rRMSE = 2.025), and the highest mean in Case 1 (rRMSE = 2.037). Regarding the overall average performance of the MLR model (Figure 7b), the means of the four stations from the

lowest to highest were Case 1 (rRMSE = 2.143), Case 2 (rRMSE = 2.166), and Case 3 (rRMSE = 2.160). These results revealed that the prediction model using Case 3 could yield more accurate prediction values than could other model cases.

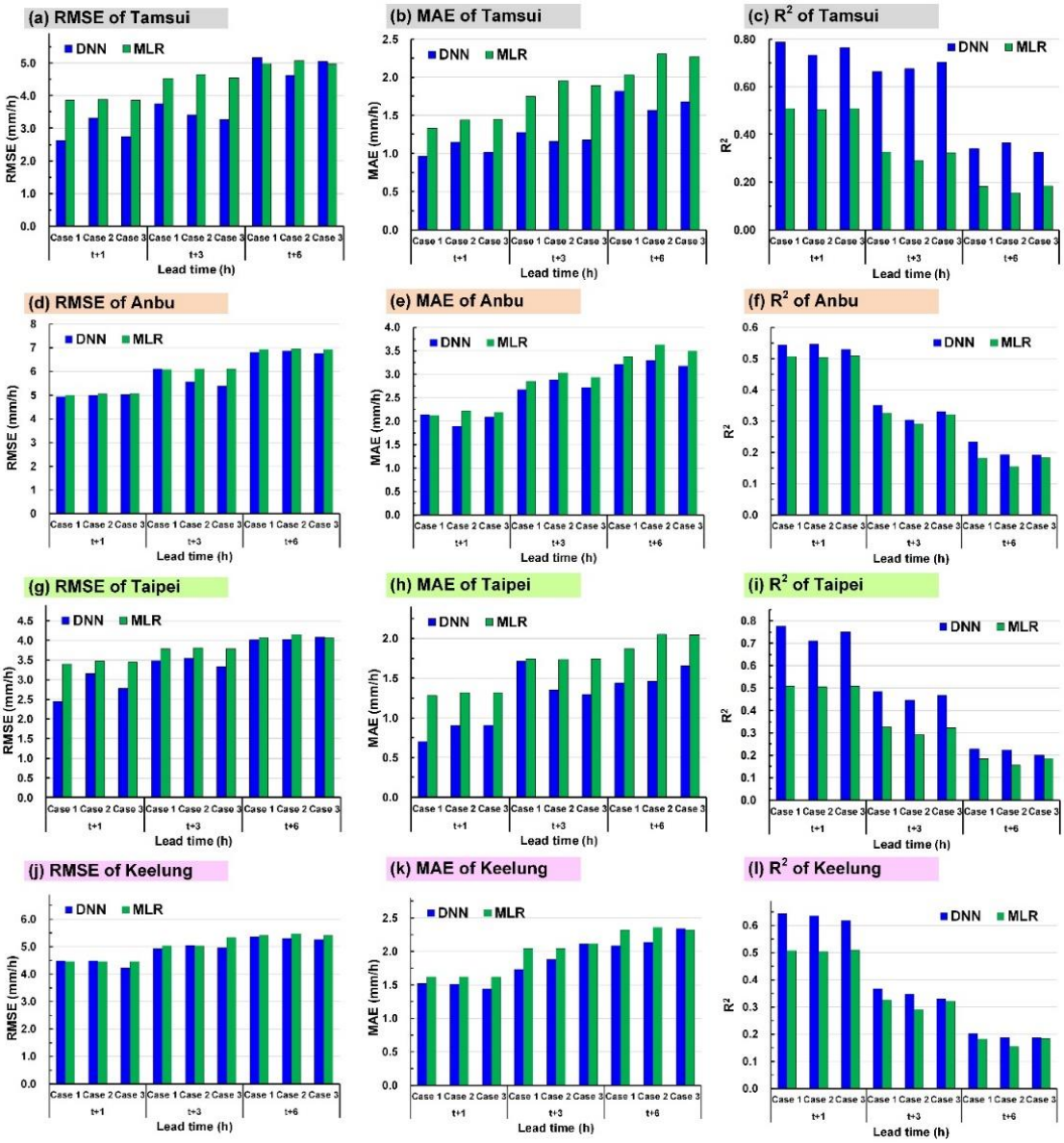

Figure 5. Performance levels of the root mean squared error (RMSE), mean absolute error (MAE), and R^2 by using a testing set at the (**a**–**c**) Tamsui station, (**d**–**f**) Anbu station, (**g**–**i**) Taipei station, and (**j**–**l**) Keelung station.

4.2. Simulation of Typhoons

This section details the simulation of two typhoon events: Typhoon Herb (1996) and Typhoon Jangmi (2008). After Typhoon Herb landed in Taiwan, its track (Figure 8a) went from east to west, thereby penetrating the research area and bringing a massive rainfall quantity to the area. The rainfall duration of Typhoon Herb in the research area was approximately 40 h. After Typhoon Jangmi landed in Taiwan, its track (Figure 8b) went from southeast to northwest, passed through the southwestern part of the research area (Figure 8a), and brought considerable rainfall to the area. The rainfall duration of Typhoon Jangmi in the research area was approximately 72 h.

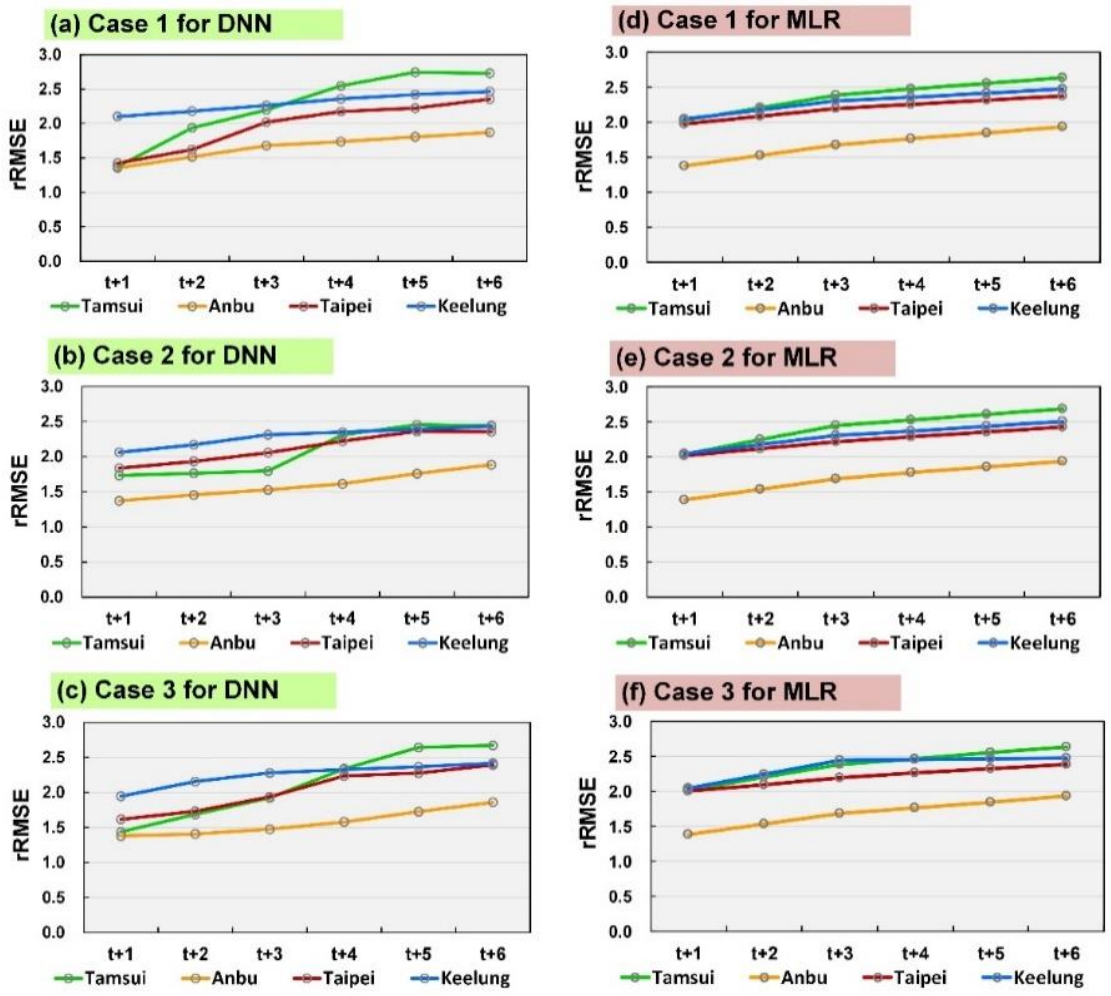

Figure 6. Comparison of performance levels among lead time 1–6 h: (**a**–**c**) DNN model for Cases 1–3, respectively, and (**d**–**f**) MLR model for Cases 1–3, respectively.

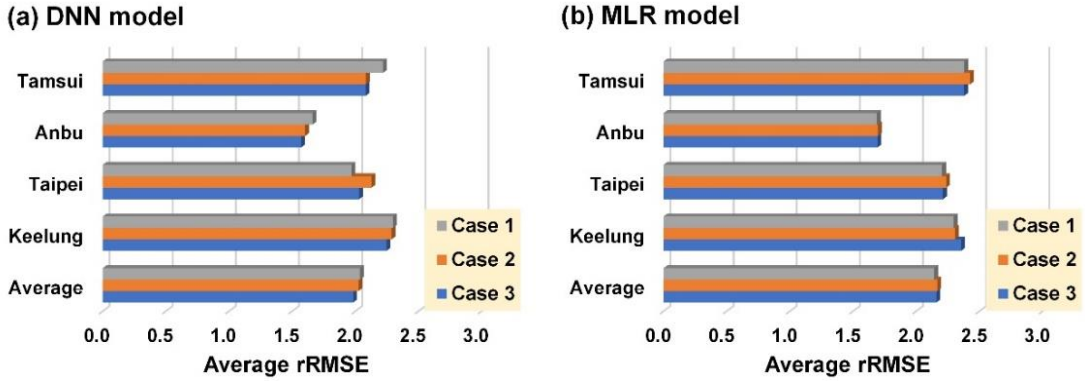

Figure 7. Comparison of average performance for (**a**) DNN model and (**b**) MLR model.

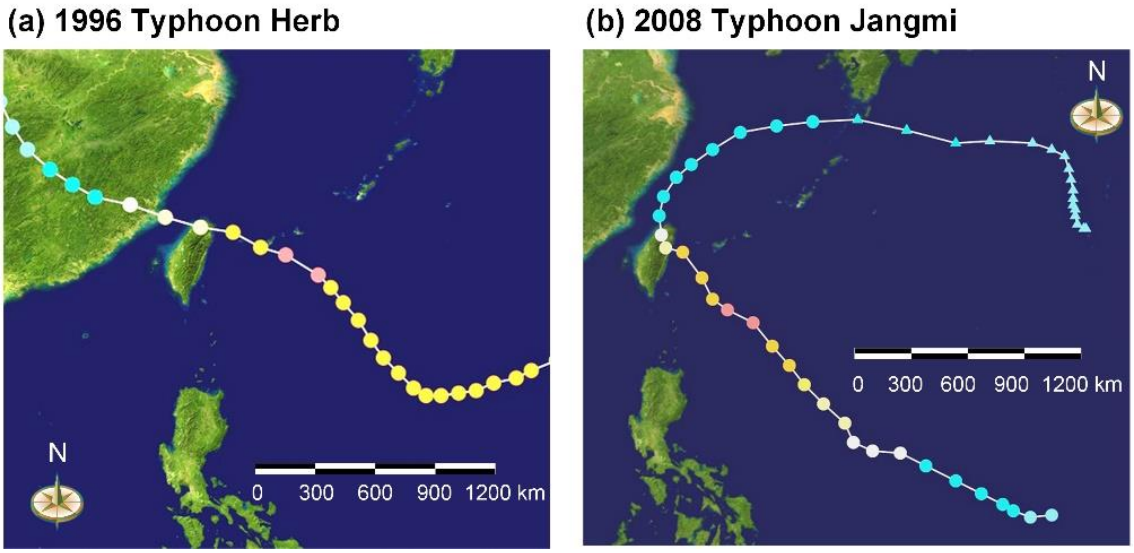

Figure 8. Track of Typhoons: (**a**) Herb and (**b**) Jangmi.

Figures 9 and 10, respectively, depict the prediction and observation values of the hyetographs of Typhoons Herb and Jangmi. As shown in Figure 9, the maximum hourly rainfall of Typhoon Herb at the Tamsui, Anbu, Taipei, and Keelung stations were 37.0, 51.8, 34.6, and 41.6 mm/h, respectively. The highest hourly rainfall was observed in the Anbu station. The cumulative total rainfall quantities from high to low were at the Anbu (560 mm), Tamsui (246 mm), Keelung (217 mm), and Taipei stations (35 mm). The maximum hourly rainfall of Typhoon Jangmi at the Tamsui, Anbu, Taipei, and Keelung stations was 36.0, 42.5, 19.5, and 25.5 mm/h, respectively. The highest hourly rainfall was observed in the Anbu station. The total rainfall quantities from high to low occurred at the Anbu (563 mm), Tamsui (392 mm), Keelung (280 mm), and Taipei stations (228 mm). The results revealed that terrain-affected typhoon circulations terrain could easily cause heavy rainfall and the accumulation of tremendous rainfall quantity at the Anbu station. Because the Taipei station is situated in the Taipei Basin, the rainfall quantity from the typhoons was less there than in the other stations.

Figures 9 and 10 present the prediction results when the lead time was 1, 3, and 6 h. The results revealed that, as prediction duration increased, the prediction accuracy of each model decreased gradually. The simulation results of Typhoons Herb and Jangmi revealed that, when the lead time was 1 h (Figure 9a,d,g,j and Figure 10a,d,g,j), the prediction values of the DNN and MLR models roughly matched the observation values. When the lead time reached 3 or 6 h, the prediction values of the MLR model tended to underestimate high rainfall and overestimate low rainfall. A possible cause is that the MLR used statistical linear regression and adopted the means as the prediction results. Compared with the MLR, the DNN model better reflected fluctuations in future rainfall possibly because neuron weighting was adjusted to facilitate the learning of rainfall estimation during modeling.

Figures 11 and 12 depict the absolute and relative error indicator results of the two typhoons. The figures indicate that a greater absolute error of RMSE was generated at the Anbu station than that of the other three stations during both typhoons, whereas a lower RMSE was observed at the Taipei station. By contrast, the Anbu station exhibited a lower rRMSE than did the other three stations, thereby generating results that were consistent with those described in a previous section.

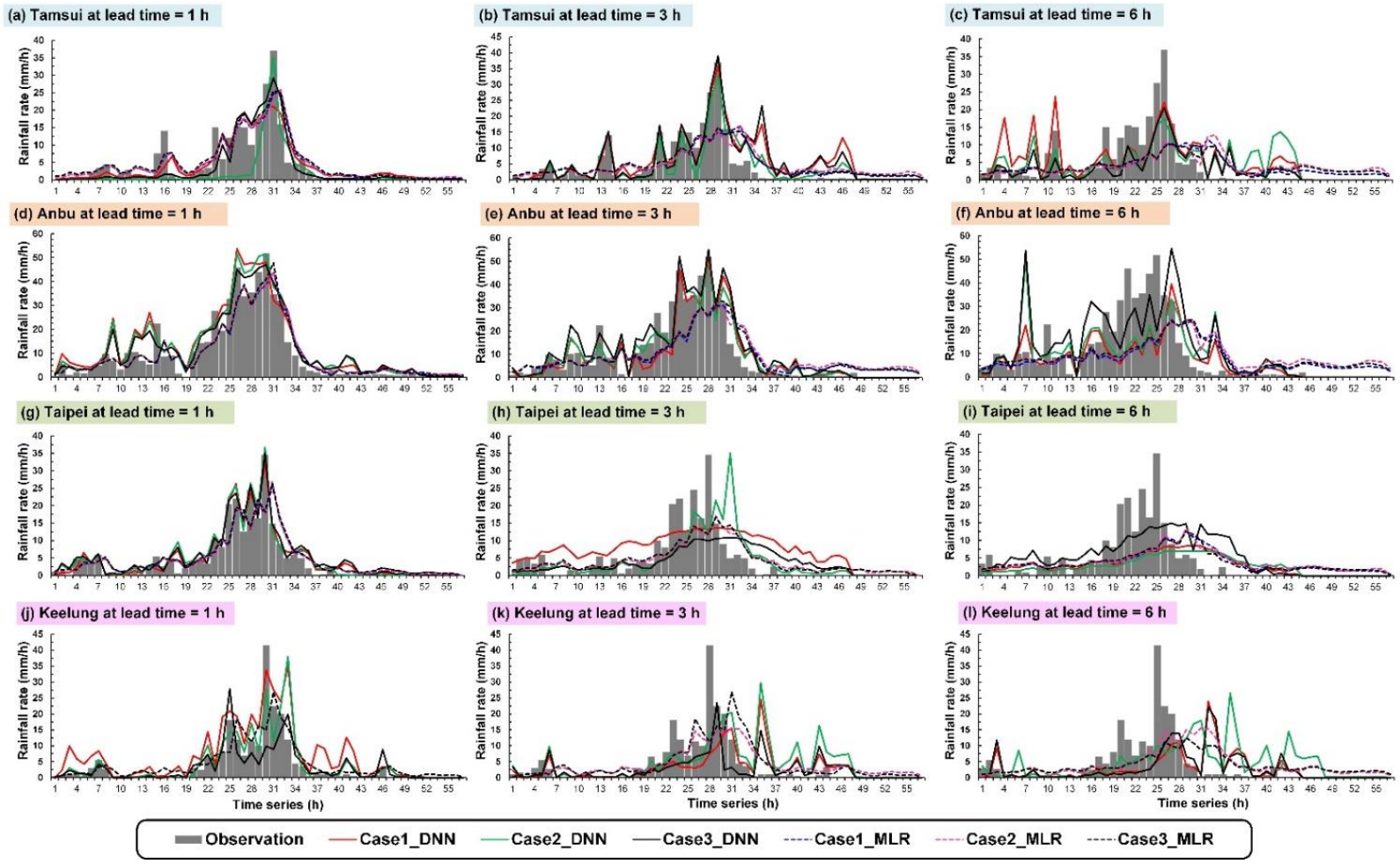

Figure 9. Simulation results of Typhoon Herb: (**a**–**c**) Tamsui station at lead time = 1, 3, and 6 h, respectively, (**d**–**f**) Anbu station at lead time = 1, 3, and 6 h, respectively, (**g**–**i**) Taipei station at lead time = 1, 3, and 6 h, respectively, and (**j**–**l**) Keelung station at lead time = 1, 3, and 6 h, respectively.

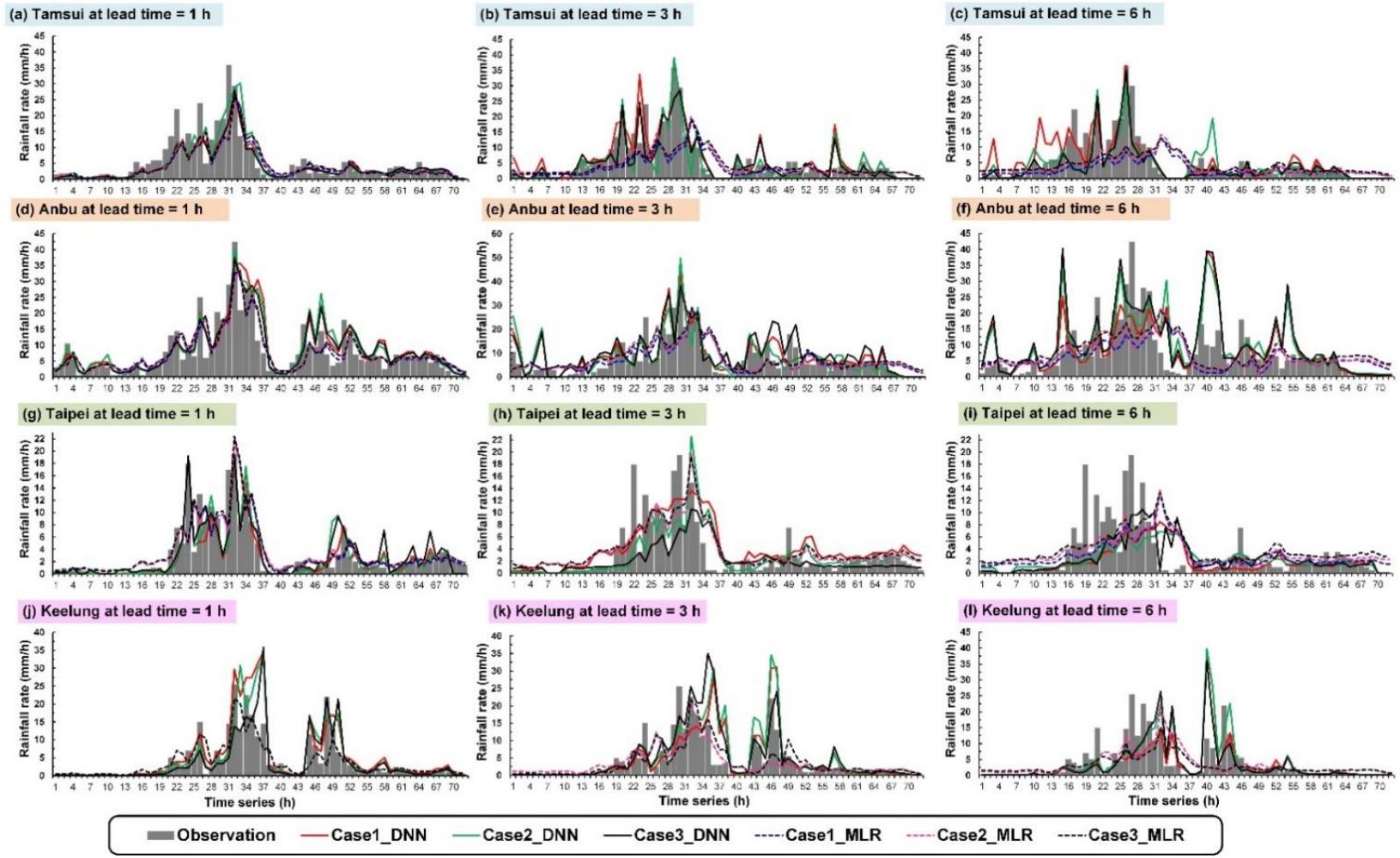

Figure 10. Simulation results of Typhoon Jangmi: (**a**–**c**) Tamsui station at lead time = 1, 3, and 6 h, respectively, (**d**–**f**) Anbu station at lead time = 1, 3, and 6 h, respectively, (**g**–**i**) Taipei station at lead time = 1, 3, and 6 h, respectively, and (**j**–**l**) Keelung station at lead time = 1, 3, and 6 h, respectively.

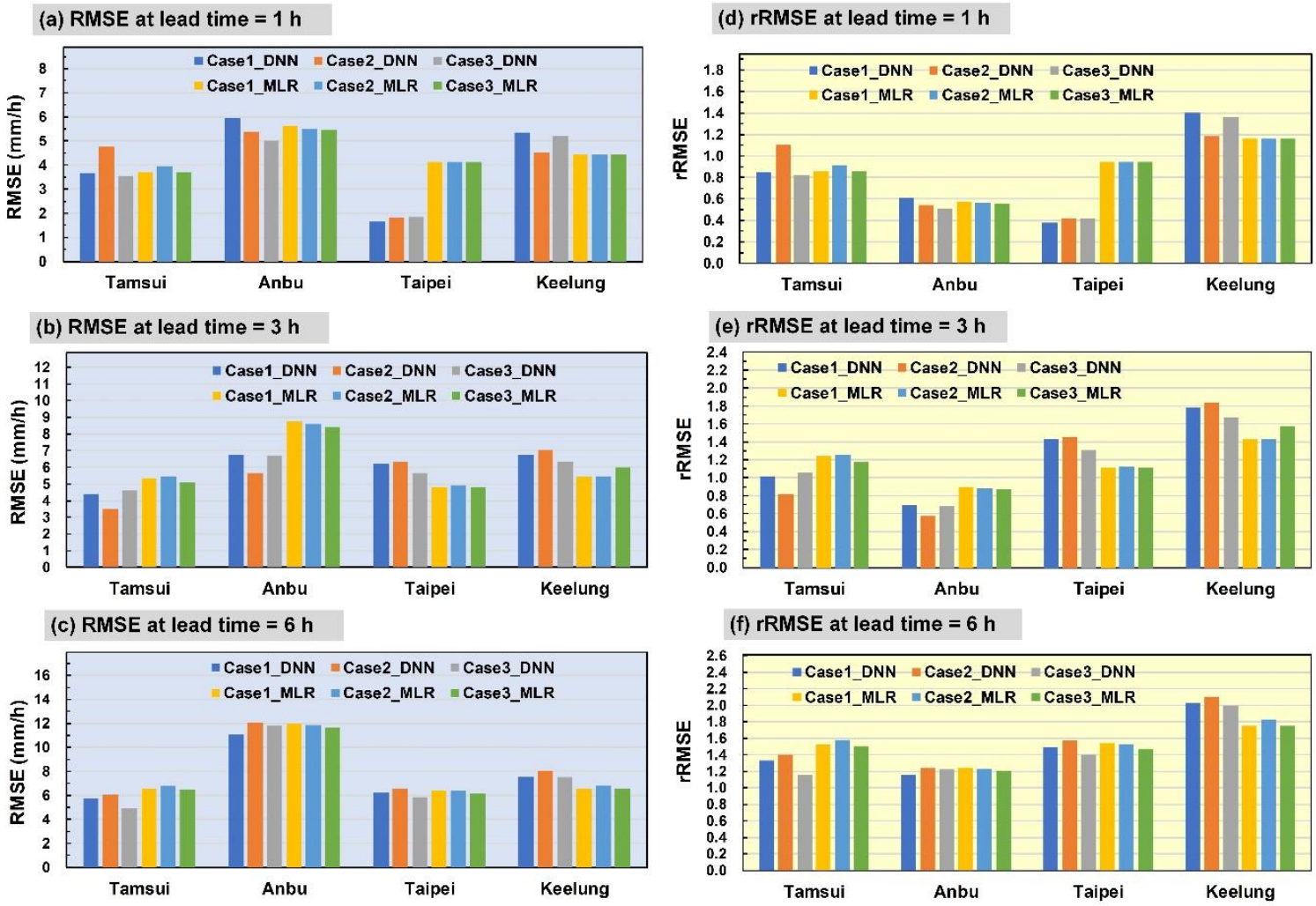

Figure 11. Performance levels of Typhoon Herb: (**a**–**c**) RMSE measures at lead time = 1, 3, and 6 h, respectively, and (**d**–**f**) rRMSE measures at lead time = 1, 3, and 6 h, respectively.

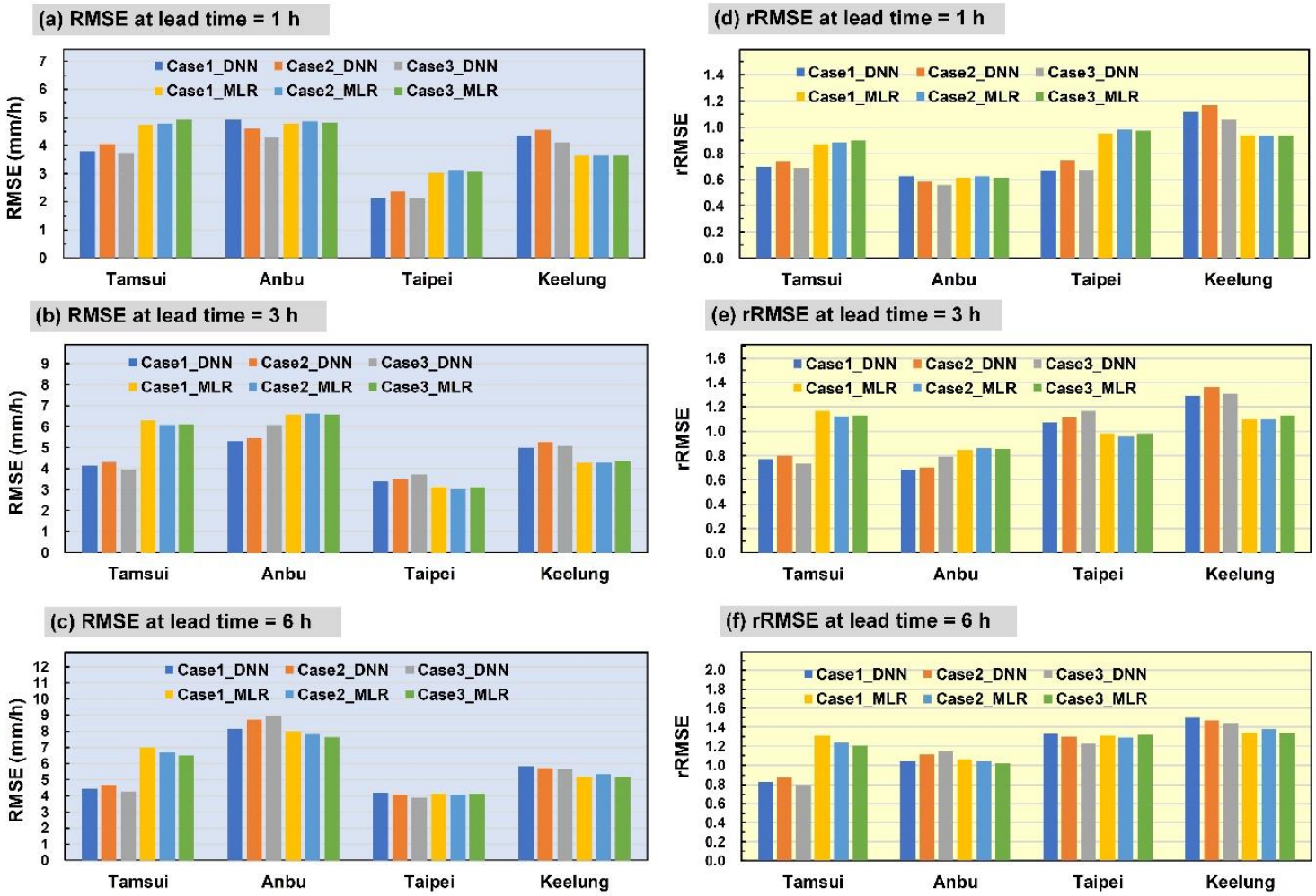

Figure 12. Performance levels of Typhoon Jangmi: (**a**–**c**) RMSE measures at lead time = 1, 3, and 6 h, respectively, and (**d**–**f**) rRMSE measures at lead time = 1, 3, and 6 h, respectively.

5. Efficiency of Computation Environments

This section presents the evaluation of computer computation performance. First, a single-server computer (E3 CPU) was used to test the computing time of the DNN, in Cases 1–3, which respectively contained 16,957, 6549, and 4771 records. Figure 13 shows that, in a single-server environment, in Cases 1–3, the computing time decreased following the decreasing number of records. However, the data quantity run in the three cases was not enough to present differences in the computing efficiency of computation environments that require a massive quantity of data.

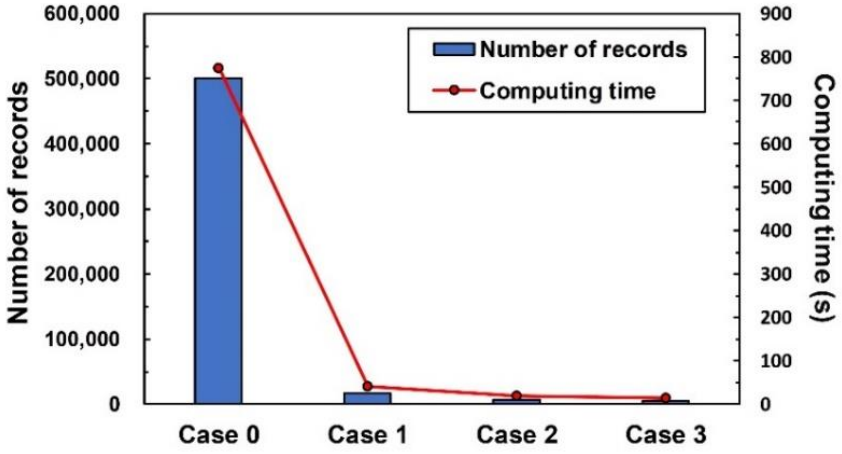

Figure 13. Computing efficiency among model cases by using a single server (E3).

This study designed a dataset (coded as Case 0) to understand computing performance when data increased. The attribute data were obtained from 57 years of hourly station data (the maximum duration with available records from the stations in the research area). The total number of records reached approximately 500,000. Figure 13 indicates that, when the data quantity increased substantially, the computing time of Case 0 increased rapidly and considerably. Because Case 0 can be easily used to evaluate computing efficiency under various computation environments, it was adopted in the following evaluation process.

According to the computer equipment described in Section 3.2, we simulated three computation environments: (1) a standalone system with only one I7 CPU computer (hereinafter referred to as I7), (2) a single server with only one E3 CPU computer (hereinafter referred to as E3), and (3) a Hadoop cluster system that serially connected four E3 CPU computers (hereinafter referred to as Hadoop). Table 6 presents the computing times of the Tamsui, Anbu, Taipei, and Keelung stations under the three computation environments. In the table, "CPU" represents the computing time of the CPU, and "USER" represents the total computing time after the CPU computing time was deducted (including time for data transmission in the system and the transmission of network packets). In addition, Table 6 provides a comparison of the DNN and MLR modeling times. The results revealed that the USER time was longer than the CPU time in all three models. In DNN modeling, a large quantity of parameters is required to test the analysis results; therefore, the USER and CPU times were longer than those in MLR modeling. Moreover, when evaluating the computing performance of the three computation environments (I7, E3, and Hadoop), we observed the mean computing time of the four stations and found that the Hadoop USER and CPU times were considerably shorter than those in the I7 and E3 computation environments.

Figure 14 displays the comparison of the total computing times of the DNN and MLR models. Figure 14a indicates that the MLR model had extremely high computing efficiency, and the total computing time of I7, E3, and Hadoop was within 2 s. By contrast, Figure 14b revealed the considerable difference between the total computing times among the computation environments in the DNN

model, in which Hadoop was 27 and 9 times faster than I7 and E3, respectively. The experimental results revealed that when complex machine-learning model computation is required, the Hadoop Spark framework, based on big-data technology, can be applied to develop an efficient computation environment and system.

Table 6. Computing time of both models with their computation environments.

Station	Model	Standalone (I7)		Single Server (E3)		Cluster System (Hadoop)	
		USER	CPU	USER	CPU	USER	CPU
Tamsui	MLR	0.58	0.13	0.56	0.05	0.58	0.05
	DNN	2134.0	175.6	697.5	59.1	76.4	8.6
Anbu	MLR	0.95	0.21	0.70	0.18	0.58	0.06
	DNN	2299.6	200.1	751.6	67.3	82.3	9.8
Taipei	MLR	0.58	0.08	0.57	0.07	0.56	0.06
	DNN	2252.7	216.4	736.3	72.8	80.6	10.6
Keelung	MLR	0.47	0.14	0.50	0.05	0.47	0.05
	DNN	2043.7	157.7	667.9	53.0	73.2	7.7
Average	MLR	0.65	0.14	0.58	0.09	0.55	0.05
	DNN	2182.5	187.5	713.3	63.0	78.1	9.2

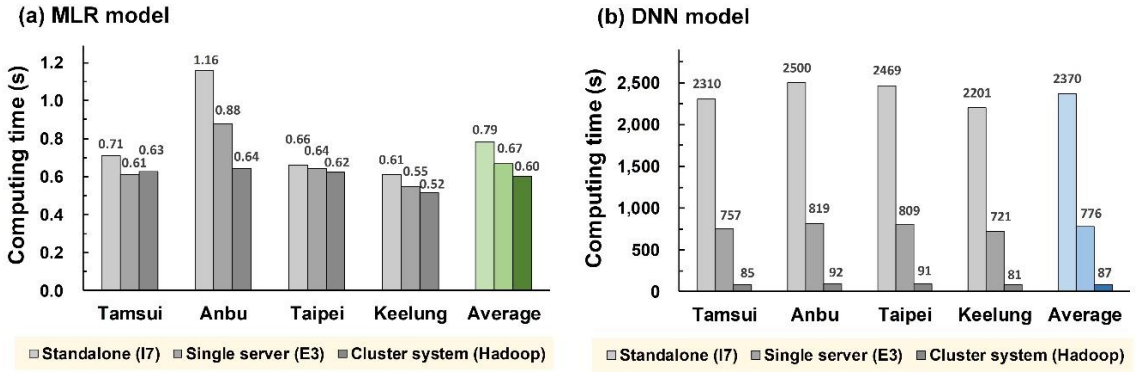

Figure 14. Comparison of total computational time for (**a**) MLP model, and (**b**) DNN model.

6. Conclusions

The study developed a computing environment by using big-data technology to accelerate machine-learning algorithms when building rainfall prediction models during typhoons. This study used machine-learning models that comprised DNNs and MLRs to establish rainfall prediction models. The big-data technology used was the Hadoop Spark distributed cluster-computing framework. The Hadoop system consisted of the HDFS and MapReduce framework, and next-generation Spark technology was used to improve the efficiency of the distributed computing.

The research area was Northern Taiwan, where four surface observation stations (i.e., Tamsui, Anbu, Taipei, and Keelung) were selected as the experimental sites. The sources include meteorological data from CWB of Taiwan from 1961 to 2017 from typhoon warning and ground stations. This study screened relevant attribute data based on individual model cases of the stations. The prediction duration was 1–6 h. To understand the computing performance of the Hadoop distributed framework, we tested the computing performance of standalone computers and cluster systems, using three computation environments: a standalone system with one I7 CPU computer, a single server with one E3 CPU computer, and a Hadoop system with four E3 CPU computers.

Through the experiments, we obtained the following findings: (1) in machine-learning computation, prediction errors increased with prediction duration in the DNN and MLR models. Regarding station prediction performance, the DNN model performed more favorably in the Anbu station, followed by

that in the Taipei, Tamsui, and Keelung stations, whereas the MLR model performed more favorably in the Anbu station and generated similar results in the other three stations. (2) Of the three computation environments that were used for testing the Hadoop Spark distributed cluster-computing framework, the Hadoop system had a faster execution speed than did the standalone systems (single I7 CPU and single E3 CPU). In models that required complex computation (e.g., DNN model parameter test), the big-data Hadoop Spark framework could be used to develop a highly efficient computation environment and system. In practical applications, the framework can be implemented through the Internet to connect more than the four computers that were used in this study. Therefore, following the development of big-data technology, highly nonlinear problems (e.g., rainfall simulation) can be solved more rapidly to enable the efficient development of prediction systems.

In summary, this study successfully used the big-data technology Hadoop Spark and combined machine learning to develop rainfall prediction models with effectively improved computing efficiency. Therefore, the proposed system can solve problems concerning real-time typhoon rainfall prediction with high timeliness and accuracy.

Future studies are likely to focus on enhancing the feature-selection method. Among the numerous feature-selection methods available, according to Maier et al. [24] and Wu et al. [52], two primary approaches are typically adopted: the model-free approach (e.g., correlation-based criterion method and mutual information method) and the model-based approach (e.g., stepwise selection method and ad hoc method). The most commonly used measure of statistical dependence for input selection is a correlation measure, which is classified as a model-free approach. This selection has the disadvantage of measuring only the linear dependence between variables [53,54]. For this research, we used the selection of variables based on the correlation measure, without considering the multicollinearity among inputs. Therefore, we suggest applying a model-based approach, such as the stepwise selection method proposed by Efroymson [55], which is a one-step-at-a-time approach and is based on t-tests of the individual parameters, known to be severely affected by multicollinearity (i.e., highly correlated predictors).

Author Contributions: C.-C.W. conceived of and designed the experiments and wrote the manuscript; T.-H.C. and C.-C.W. carried out this experiment and analysis of the data and discussed the results. All authors have read and agreed to the published version of the manuscript.

Funding: The support for this study under Grant No. MOST109-2622-M-019-001-CC3 that was provided by the Ministry of Science and Technology, Taiwan, is greatly appreciated.

Acknowledgments: The authors acknowledge the data provided by Taiwan's Central Weather Bureau. This manuscript was edited by Wallace Academic Editing.

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

Appendix A. Typhoons and Data Attributes Over 1961–2017

Table A1. Collected typhoons.

Year	Typhoon	Year	Typhoon	Year	Typhoon
1961	Betty, Elsie, June, Lorna, Pamela, Sally	1980	Ida, Norris, Percy, Betty	1999	Sam
1962	Kate, Opal, Wanda, Amy, Dinah	1981	Ike, June, Maury, Agnes, Clara, Irma	2000	Kai-tak, Bilis, Prapiroon, Bopha, Yagi, Xangsane, Bebinca
1963	Wendy, Gloria	1982	Andy, Cecil, Dot, Ken	2001	Cimaron, Chebi, Utor, Trami,
1964	Betty, Doris, Ida, Sally, Tilda	1983	Wayne, Ellen, Forrest	2002	Rammasun, Nakri, Sinlaku
1965	Dinah, Harriet	1984	Wynne, Alex, Freda, Holly, June	2003	Kujira, Nangka, Soudelor, Imbudo, Morakot, Vamco, Krovanh, Dujuan, Melor

Table A1. *Cont.*

Year	Typhoon	Year	Typhoon	Year	Typhoon
1966	Judy, Alice, Cora, Elsie	1985	Hal, Jeff, Nelson, Val, Brenda	2004	Conson, Mindulle,
1967	Anita, Clara, Nora, Carla, Glida	1986	Nancy, Peggy, Wayne, Wayne, Wayne, Abby	2005	Haitang, Matsa, Sanvu, Talim, Khanun, Damrey, Longwang
1968	Wendy, Elaine	1987	Thelma, Vernon, Alex, Cary, Dinah, Gerald, Lynn	2006	Chanchu, Ewiniar, Bilis, Kaemi, Saomai, Bopha, Shanshan
1969	Viola, Betty, Elsie, Flossie	1988	Susan, Warren, Nelson	2007	Pabuk, Sepat, Wipha, Krosa, Mitag
1970	Olga, Wilda, Fran	1989	Sarah	2008	Kalmaegi, Fung-wong, Nuri, Sinlaku, Hagupit, Jangmi
1971	Lucy, Nadine, Agnes, Bess	1990	Marian, Ofelia, Percy, Robyn, Yancy, Abe, Dot	2009	Linfa, Molave, Morakot, Parma
1972	Susan, Winnie, Betty	1991	Amy, Brenda, Ellie, Mireille, Nat, Ruth, Seth	2010	Lionrock, Namtheun, Meranti, Fanapi, Megi
1973	Joan, Nora	1992	Bobbie, Mark, Omar, Polly, Ted	2011	Aere, Songda, Meari, Muifa, Nanmadol
1974	Jean, Lucy, Wendy, Bess	1993	Tasha, Yancy, Abe	2012	Talim, Doksuri, Saola, Haikui, Kai-tak, Tembin, Tembin, Jelawat
1975	Nina, Betty, Elsie	1994	Tim, Caitlin, Doug, Fred, Gladys, Seth	2013	Soulik, Cimaron, Trami, Kong-rey, Usagi, Fitow
1976	Ruby, Billie	1995	Deanna, Gary, Janis, Kent, Ryan	2014	Hagibis, Matmo, Fung-wong
1977	Ruth, Thelma, Vera, Amy	1996	Cam, Gloria, Herb, Sally, Zane	2015	Noul, Chan-hom, Linfa, Soudelor, Goni, Dujuan
1978	Olive, Rose, Della, Ora	1997	Winnie, Amber, Cass, Ivan	2016	Nepartak, Meranti, Malakas, Megi, Aere
1979	Gordon, Hope, Irving, Judy	1998	Nichole, Otto, Yanni, Zeb, Babs	2017	Nesat, Haitang, Hato, Guchol, Talim

Table A2. Statistical data on typhoons.

Attribute	Range	Mean
Pressure at typhoon center (hPa)	15–1000	957.6
Latitude (°N) of typhoon center	15–29.5	22.3
Longitude (°E) of typhoon center	113.2–133.7	122.4
Radius of winds over 15.5 m/s (km)	0–400	206.7
Moving speed of typhoon (km/h)	0–65	17.1
Maximum wind speed of typhoon center (m/s)	12–216	74.5

Table A3. Statistical data on ground weather at the target stations of Tamsui and Anbu.

Attribute	Tamsui Station		Anbu Station	
	Range	Mean	Range	Mean
Air pressure on the ground (hPa)	957–1022	1000.5	871–929	912.2
Temperature on the ground (°C)	15.1–38.2	27.2	9.5–30.2	21.4
Dew point on the ground (°C)	10.8–30	23.1	8.5–25.4	20.2
Relative humidity (%)	2.4–100	79.5	42–100	93.0
Vapor pressure on the ground (hPa)	11.3–42.4	28.5	11.1–34.1	23.8
Surface wind velocity (m/s)	0–29.3	3.7	0–41.8	6.5
Surface wind direction (°)	0–360	140.4	0–360	228.3
Precipitation (mm)	0–86.8	1.3	0–119.5	2.9

Table A4. Statistical data on ground weather at the target stations of Taipei and Keelung.

Attribute	Taipei Station		Keelung Station	
	Range	**Mean**	**Range**	**Mean**
Air pressure on the ground (hPa)	954–1023	1001.6	954–1021	1000.5
Temperature on the ground (°C)	16.1–37.3	27.5	15.6–36.7	27.3
Dew point on the ground (°C)	11.2–28.5	23.3	9.4–28.6	23.4
Relative humidity (%)	37–100	78.9	46–100	80.1
Vapor pressure on the ground (hPa)	13.3–38.9	28.8	11.8–37.1	29.0
Surface wind velocity (m/s)	0–28.9	3.9	0–28.5	5.0
Surface wind direction (°)	0–360	134.9	0–360	131.7
Precipitation (mm)	0–76	1.3	0–95.3	1.4

References

1. Wei, C.C. Study on wind simulations using deep learning techniques during typhoons: A case study of Northern Taiwan. *Atmosphere* **2019**, *10*, 684. [CrossRef]
2. Kang, S.C.; Shiu, R.S.; Wu, T.H. Development of typhoon search program with human manipulation consideration. In Proceedings of the Conference for Disaster Management in Taiwan, Disaster Management Society of Taiwan, Taipei, Taiwan, 16 November 2012.
3. Wei, C.C.; Cheng, J.Y. Nearshore two-step typhoon wind-wave prediction using deep recurrent neural networks. *J. Hydroinform.* **2020**, *22*, 356–367. [CrossRef]
4. Burgin, M.; Klinger, A. Experience, generations, and limits in machine learning. *Theor. Comput. Sci.* **2004**, *317*, 71–91. [CrossRef]
5. Wei, C.C. Radial basis function networks combined with principal component analysis to typhoon precipitation forecast in a reservoir watershed. *J. Hydrometeorol.* **2012**, *13*, 722–734. [CrossRef]
6. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
7. Yeh, T.C. Typhoon rainfall over Taiwan area: The empirical orthogonal function modes and their applications on the rainfall forecasting. *Terr. Atmos. Ocean. Sci.* **2002**, *13*, 449–468. [CrossRef]
8. Lee, C.S.; Huang, L.R.; Shen, H.S.; Wang, S.T. A climatology model for forecasting typhoon rainfall in Taiwan. *Nat. Hazards* **2006**, *37*, 87–105. [CrossRef]
9. Hsu, N.S.; Wei, C.C. A multipurpose reservoir real-time operation model for flood control during typhoon invasion. *J. Hydrol.* **2007**, *336*, 282–293. [CrossRef]
10. Hall, T.; Brooks, H.E.; Doswell, C.A. Precipitation forecasting using a neural network. *Weather Forecast.* **1999**, *14*, 338–345. [CrossRef]
11. Fox, N.I.; Wikle, C.K. A Bayesian quantitative precipitation nowcast scheme. *Weather Forecast.* **2005**, *20*, 264–275. [CrossRef]
12. Nasseri, M.; Asghari, K.; Abedini, M.J. Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. *Expert Syst. Appl.* **2008**, *35*, 1415–1421. [CrossRef]
13. Biondi, D.; De Luca, D.L. A Bayesian approach for real-time flood forecasting. *Phys. Chem. Earth* **2012**, *42–44*, 91–97. [CrossRef]
14. Wei, C.C. Wavelet support vector machines for forecasting precipitations in tropical cyclones: Comparisons with GSVM, regressions, and numerical MM5 model. *Weather Forecast.* **2012**, *27*, 438–450. [CrossRef]
15. Kühnlein, M.; Appelhans, T.; Thies, B.; Nauß, T. Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests. *J. Appl. Meteorol. Climatol.* **2014**, *53*, 2457–2480. [CrossRef]
16. Wei, C.C. Simulation of operational typhoon rainfall nowcasting using radar reflectivity combined with meteorological data. *J. Geophys. Res. Atmos.* **2014**, *119*, 6578–6595. [CrossRef]
17. Wei, C.C.; You, G.J.Y.; Chen, L.; Chou, C.C.; Roan, J. Diagnosing rain occurrences using passive microwave imagery: A comparative study on probabilistic graphical models and "black box" models. *J. Atmos. Ocean. Technol.* **2015**, *32*, 1729–1744. [CrossRef]
18. Diez-Sierra, J.; del Jesus, M. Subdaily rainfall estimation through daily rainfall downscaling using random forests in Spain. *Water* **2019**, *11*, 125. [CrossRef]

19. Ko, C.M.; Jeong, Y.Y.; Lee, Y.M.; Kim, B.S. The development of a quantitative precipitation forecast correction technique based on machine learning for hydrological applications. *Atmosphere* **2020**, *11*, 111. [CrossRef]
20. Xiang, B.; Zeng, C.; Dong, X.; Wang, J. The application of a decision tree and stochastic forest model in summer precipitation prediction in Chongqing. *Atmosphere* **2020**, *11*, 508. [CrossRef]
21. Asklany, S.A.; Elhelow, K.; Youssef, I.K.; El-wahab, M.A. Rainfall events prediction using rule-based fuzzy inference system. *Atmos. Res.* **2011**, *101*, 228–236. [CrossRef]
22. Maier, H.R.; Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: A review of modeling issues and applications. *Environ. Model. Softw.* **2000**, *15*, 101–124. [CrossRef]
23. Antolik, M.S. An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.* **2000**, *239*, 306–337. [CrossRef]
24. Maier, H.R.; Jain, A.; Dandy, G.C.; Sudheer, K. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environ. Model. Softw.* **2010**, *25*, 891–909. [CrossRef]
25. Madsen, H.; Lawrence, D.; Lang, M.; Martinkova, M.; Kjeldsen, T.R. Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. *J. Hydrol.* **2014**, *519*, 3634–3650. [CrossRef]
26. Maçaira, P.; Thomé, A.M.; Oliveira, F.L.; Ferrer, A.L. Time series analysis with explanatory variables: A systematic literature review. *Environ. Model. Softw.* **2018**, *107*, 199–209. [CrossRef]
27. Paulo Vitor de Campos Souza, P.V. Fuzzy neural networks and neuro-fuzzy networks: A review the main techniques and applications used in the literature. *Appl. Soft Comput.* **2020**, *92*, 106275. [CrossRef]
28. Tao, Y.; Gao, X.; Ihler, A.; Sorooshian, S.; Hsu, K. Precipitation identification with bispectral satellite information using deep learning approaches. *J. Hydrometeorol.* **2017**, *18*, 1271–1283. [CrossRef]
29. Wang, H.Z.; Wang, G.B.; Li, G.Q.; Peng, J.C.; Liu, Y.T. Deep belief network based deterministic and probabilistic wind speed forecasting approach. *Appl. Energy* **2016**, *182*, 80–93. [CrossRef]
30. Wang, J.H.; Lin, G.F.; Chang, M.J.; Huang, I.H.; Chen, Y.R. Real-time water-level forecasting using dilated causal convolutional neural networks. *Water Resour. Manag.* **2019**, *33*, 3759–3780. [CrossRef]
31. Wei, C.C.; Hsieh, P.Y. Estimation of hourly rainfall during typhoons using radar mosaic-based convolutional neural networks. *Remote Sens.* **2020**, *12*, 896. [CrossRef]
32. Emani, C.K.; Cullot, N.; Nicolle, C. Understandable big data: A survey. *Comput. Sci. Rev.* **2015**, *17*, 70–81. [CrossRef]
33. Qureshi, B.; Koubaa, A. On energy efficiency and performance evaluation of single board computer based clusters: A Hadoop case study. *Electronics* **2019**, *8*, 182. [CrossRef]
34. Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; Franklin, M.J. Apache Spark: A unified engine for big data processing. *Commun. Acm* **2016**, *59*, 56–65. [CrossRef]
35. International Data Corporation (IDC). Big Data Big Opportunities. 2018. Available online: http://www.emc.com/microsites/cio/articles/big-data-bigopportunities/LCIA-BigDataOpportunities-Value.pdf (accessed on 25 July 2018).
36. Dailey, W. The Big Data Technology Wave. 2019. Available online: https://www.skillsoft.com/courses/5372828-thebig-data-technology-wave/ (accessed on 18 March 2019).
37. Borthakur, D. The Hadoop Distributed File System: Architecture and Design. 2007. Hadoop Projection Website. Available online: http://svn.apache.org/repos/asf/hadoop/common/tags/release-0.16.3/docs/hdfs_design.pdf (accessed on 1 July 2020).
38. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. Acm* **2008**, *51*, 107–113. [CrossRef]
39. Bechini, A.; Marcelloni, F.; Segatori, A. A MapReduce solution for associative classification of big data. *Inf. Sci.* **2016**, *332*, 33–55. [CrossRef]
40. Armbrust, M.; Xin, R.S.; Lian, C.; Huai, Y.; Liu, D.; Bradley, J.K.; Meng, X.; Kaftan, T.; Franklin, M.J.; Ghodsi, A. Spark SQL: Relational Data Processing in Spark. In Proceedings of the ACM SIGMOD/PODS Conference, Melbourne, Australia, 31 May– 4 June 2015; ACM Press: New York, NY, USA, 2015.
41. Xin, R.S.; Rosen, J.; Zaharia, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Shark: SQL and Rich Analytics at Scale. In Proceedings of the ACM SIGMOD/PODS Conference, New York, NY, USA, 22–27 June 2013; ACM Press: New York, NY, USA, 2013.

42. Hu, Y.; Cai, X.; DuPont, B. Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using Hadoop. *Environ. Model. Softw.* **2015**, *70*, 149–162. [CrossRef]
43. Hu, Y.; Garcia-Cabrejo, O.; Cai, X.; Valocchi, A.J.; DuPont, B. Global sensitivity analysis for large-scale socio-hydrological models using Hadoop. *Environ. Model. Softw.* **2015**, *73*, 231–243. [CrossRef]
44. Taylor, R. Interpretation of the correlation coefficient: A basic review. *J. Diagn. Med Sonogr.* **1990**, *1*, 35–39. [CrossRef]
45. Central Weather Bureau (CWB). 2019. Available online: http://www.cwb.gov.tw/V8/C/K/announce.html (accessed on 1 December 2019).
46. Villegas-Ch, W.; Palacios-Pacheco, X.; Luján-Mora, S. Application of a smart city model to a traditional university campus with a big data architecture: A sustainable smart campus. *Sustainability* **2019**, *11*, 2857. [CrossRef]
47. Ajah, I.A.; Nweke, H.F. Big data and business analytics: Trends, platforms, success factors and applications. *Big Data Cogn. Comput.* **2019**, *3*, 32. [CrossRef]
48. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Khan, S.U. The rise of "big data" on cloud computing: Review and open research issues. *Inf. Syst.* **2015**, *47*, 98–115. [CrossRef]
49. Lin, C.; Wei, C.C.; Tsai, C.C. Prediction of influential operational compost parameters for monitoring composting process. *Environ. Eng. Sci.* **2016**, *33*, 494–506. [CrossRef]
50. Genell, A.; Nemes, S.; Steineck, G.; Dickman, P.W. Model selection in medical research: A simulation study comparing Bayesian model averaging and stepwise regression. *BMC Med Res. Methodol.* **2010**, *10*, 108. [CrossRef] [PubMed]
51. Wei, C.C. Comparing lazy and eager learning models for water level forecasting in river-reservoir basins of inundation regions. *Environ. Model. Softw.* **2015**, *63*, 137–155. [CrossRef]
52. Wu, W.; May, R.J.; Maier, H.R.; Dandy, G.C. A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. *Water Resour. Res.* **2013**, *49*, 7598–7614. [CrossRef]
53. Bennett, N.D.; Croke, B.; Guariso, G.; Guillaume, J.H.A.; Hamilton, S.H.; Jakeman, A.; Marsili-Libelli, S.; Newham, L.T.; Norton, J.P.; Perrin, C.; et al. Characterising performance of environmental models. *Environ. Model. Softw.* **2013**, *40*, 1–20. [CrossRef]
54. Wei, C.C. Comparing single- and two-segment statistical models with a conceptual rainfall-runoff model for river streamflow prediction during typhoons. *Environ. Model. Softw.* **2016**, *85*, 112–128. [CrossRef]
55. Efroymson, M.A. *Multiple Regression Analysis*; Ralston, A., Wilf, H.S., Eds.; Mathematical Methods for Digital Computers, John Wiley: New York, NY, USA, 1960; pp. 191–203.