remote sensing

MDPI

Article

Quantitative Precipitation Estimates Using Machine Learning Approaches with Operational Dual-Polarization Radar Data

Kyuhee Shin [1], Joon Jin Song [2], Wonbae Bang [1] and GyuWon Lee [1,*]

[1] Department of Astronomy and Atmospheric Sciences, Center for Atmospheric REmote sensing (CARE), Kyungpook National University, Daegu 41566, Korea; kyuhee@knu.ac.kr (K.S.); wonbaebang@knu.ac.kr (W.B.)
[2] Department of Statistical Science, Baylor University, Waco, TX 76798-7140, USA; Joon_Song@baylor.edu
* Correspondence: gyuwon@knu.ac.kr; Tel.: +82-53-950-6361

Abstract: Traditional radar-based rainfall estimation is typically done by known functional relationships between the rainfall intensity (R) and radar measurables, such as R–Z_h, R–(Z_h, Z_{DR}), etc. One of the biggest advantages of machine learning algorithms is the applicability to a non-linear relationship between a dependent variable and independent variables without any predefined relationships. We explored the potential use of two supervised machine learning methods (regression tree and random forest) in rainfall estimation using dual-polarization radar variables. The regression tree does not require normalization and scaling of data; however, this method is quite unstable since each split depends on the parent split. Since the random forest is an ensemble method of regression trees, it has less variability in prediction compared with regression trees, but consumes more computer resources. We considered several different configurations for machine learning algorithms with different sets of dependent and independent variables. The random forest model was appropriately tuned. In the test of variable importance, the specific differential phase (differential reflectivity) was the most important variable to predict the rainfall rate (residual that is the difference between the true rainfall rate and the one estimated from the R–Z relationship). The models were evaluated by 10-fold cross-validation. The best model was the random forest model using a residual with the non-classified training set. The results indicated that the machine learning algorithms outperformed the traditional R–Z relationship. Then, we applied the best machine learning model to an S-band dual-polarization radar (Mt. Myeonbong) and validated the result with ground rain gauges. The results of the application to radar data showed that the estimates of the residuals had spatial variability. The stratiform and weak rain areas had positive residuals while convective areas had negative residuals, indicating that the spatial error structure driven by the R–Z relationship was well captured by the model. The rainfall rates of all pixels over the study area were adjusted with the estimated residuals. The rainfall rates adjusted by residual showed excellent agreement with the rain gauge, especially at high rainfall rates.

Keywords: machine learning; rainfall estimation; polarimetric radar; R–Z relationship

Citation: Shin, K.; Song, J.J.; Bang, W.; Lee, G. Quantitative Precipitation Estimates Using Machine Learning Approaches with Operational Dual-Polarization Radar Data. *Remote Sens.* **2021**, *13*, 694. https://doi.org/10.3390/rs13040694

Academic Editor: Francisco J. Tapiador
Received: 26 December 2020
Accepted: 10 February 2021
Published: 14 February 2021

1. Introduction

Quantitative precipitation estimates (QPE) are a major area of interest within the field of dual-polarization radar. With the advent of polarimetric radar, QPE algorithms using dual-polarization radar variables have been developed in recent decades [1–5]. The dual-polarization radar observes the differential reflectivity (Z_{DR}, dB), specific differential phase (K_{DP}, km^{-1}), and cross coefficient (ρ_{HV}), as well as the reflectivity (Z, mm^6 m^{-3} or dBZ). Polarimetric variables help to overcome several issues in QPE, such as miscalibration of the radar transmitter or receiver, attenuation in precipitation, and partial beam blockage. The use of these polarimetric variables can provide improved QPE [6]. In addi-

tion, since various microphysical information, such as the shape, size, and number concentration of raindrops, is provided using the horizontal and vertical polarization information, QPE using dual-polarization variables provides higher accuracy than using the reflectivity factor in horizontal polarization [7–9].

A simple form of radar-based QPE can be performed by an empirical relation between Z and the rainfall rate (R). Marshall and Parmer [10] introduced the R–Z relationship ($Z = 200\ R^{1.6}$), which explains the empirical relationship between Z and R; however, the R–Z relationship is sensitive to the variability of the drop size distribution (DSD, N(D) in $m^{-3}\ mm^{-1}$), which causes the uncertainty of the QPE using the R–Z relationship [11,12]. To deal with the uncertainty in the R–Z relationship, rainfall estimation based on dual-polarization radar variables that provide various information on raindrops was proposed [13–15].

Z_{DR} is a good measurement of the median volume diameter, and K_{DP} is linearly related to the rainfall rate as it has a lower moment than Z. As a result, rainfall estimation using these variables is more robust with respect to the variability of N(D) [13,16]. It is possible to significantly improve the over/underestimation in rainfall estimation using the R–Z relationship in a strong rainfall rate of 10 mm h^{-1} or more [17]. These empirical relationships have limitations in explaining the complex nonlinearity between R and radar variables, which leads to errors in rainfall estimation.

Recently, researchers have shown an increased interest in QPE based on machine learning using remote sensing data [18–20]. Machine learning (ML) methods, such as decision tree (DT), random forest (RF), and artificial neural networks (ANNs), are techniques for discovering the relationship between independent variables and a dependent variable based on sample data without any preliminary assumptions, including linearity. DT is divided into classification tree (CT) for a discrete dependent variable and regression tree (RT) for a continuous dependent variable [21]. DT can take into account interactions and nonlinearity between variables.

RF is an ensemble method that consists of a number of DTs and shows more accurate prediction performance than a single DT [22]. RF can reduce the variance by lowering the correlation between DTs with randomly selected independent variables. RF can be fitted in parallel, as several DTs are independently generated. Ouallouche et al. [20] performed rainfall estimation based on RF using data retrieved from the satellite, such as the cloud top height, cloud top temperature, cloud phase, and textural features.

As a result, the rain rates estimated by RF greatly agreed with those measured by rain gauges. Kusiak et al. [23] applied five data-mining approaches, including RF and DT, to estimate the rainfall using rain gauge data and Z measured from the Doppler WSR-88D radar of the National Weather Service's Next-Generation Radar (NEXRAD) system. They compared the statistics over the methods, but did not compare them with the empirical relationships. The neural network model showed good performance with the lowest mean absolute error (MAE), and the results were lower in order of support vector machine (0.19), k-nearest neighbor (0.22), CT and RT (0.26), and RF (0.27).

ANNs are ML algorithms that have been inspired by the human neuron-synaptic neural network structure. ANNs are actively applied to atmospheric remote sensing data, as this is effective in extracting characteristics and trends of complex data structures and is suitable for modeling non-linear relationships [24]. Chiang et al. [25] estimated rainfall with a recurrent neural network (RNN) using Z measured from the C-band radar for typhoon periods. The RNN produced better hourly rainfall estimates than those from R–Z relationships in terms of the root mean square error (RMSE).

As a result of the comparison of 48-hours rainfall accumulations, the rainfall estimates obtained from the R–Z relationship were underestimated with a relative bias larger than –45%, while those from the RNN had a relative bias within ±5%. Chen et al. [26] proposed a deep neural network (DNN) approach in rainfall estimation using simulated dual-polarization radar variables based on the N(D) measured from disdrometers. The

rainfall rates based on the DNNs model were almost consistent with the rainfall rates computed directly from the N(D).

They compared the hourly rainfall estimates based on the proposed algorithm using Colorado State University-Chicago Illinois (CSU-CHILL) radar data with 11 rain gauges and showed excellent agreement between the estimates and the measurements from the gauges. Although ANNs have been popularly used in a variety of applications due to the advantage of describing nonlinearity between variables, the main shortcoming we encounter with these methods is the black box problem, which makes interpretation of the process difficult. For instance, these methods provide little insight into how the independent variables influence the learning and prediction processes. On the other hand, tree-based ML methods (DT and RF) provide ease of interpretation with determination of the variable importance.

Little research has been done on ML-based rainfall estimation using dual-polarization radar. Using the dual-polarization variables allows us to incorporate microphysics information, such as the shape and the number concentration of raindrops into the rainfall estimation. The objective of this study is to improve the accuracy of rainfall estimation based on polarimetric radar parameters using machine learning methods—specifically, tree-based methods (DT and RF).

We used observed drop size distributions, N(D), measured using a two-dimensional video disdrometer (2DVD) to simulate rainfall intensity and radar variables. The ML models were trained with these simulated R and radar variables and cross-validated to check the degree of fitting. The best ML model was independently applied into Mt. Myeonbong (MYN) S-band dual-polarization radar data for rainfall estimation. The estimated rainfall rate was verified using the rain gauge data of automatic weather stations (AWSs) within the radar observation range.

2. Data

2.1. Training Dataset: 2DVD data

In this study, 2DVD data were used to train ML models. The 2DVD is an optical instrument that detects precipitation particles and uses two orthogonal cameras to detect the shadow of particles falling into the observation area. Microphysics information, such as the diameter of particles (D, mm), fall velocity (V_f, m s^{-1}), and the *axis* ratios can be obtained by measuring the shadow of precipitation particles [27]. This information can be contaminated by observing particles that fall into the observation area after being hit by a disdrometer and broken, or by the mismatching of particles in the image processing.

These outliers are removed by comparison with the empirical relationship between D and V_f [28]. In addition, an N(D) that has one or more channels with a zero number concentration is considered as discontinuous N(D) and eliminated [29]. A total of 41 diameter channels of N(D) from 0 mm to 10.25 mm at 0.25 intervals were used. A 1-min rain rate (R, mm h^{-1}) was calculated from quality-controlled 1-min N(D) using Equation (1):

$$R_{2DVD} = \frac{\pi}{6} \int_0^{D_{max}} N(D)\, V_f(D)\, D^3\, dD \quad (1)$$

where dD is the diameter interval at each diameter bin.

Table 1 shows the observation locations, observation periods, the number of 1-min N(D), and the maximum 1-min rain rate from the 2DVD data. The total number of training data was 51,302, measured in Oklahoma (OKL, USA), Daegu (DAE, ROK), Boseong (BOS, ROK), and Jincheon (JIN, ROK) to secure the diversity of microphysical processes. We used data observed in spring (April 2019) and autumn (October 2018) as well as summer data (May to September) for seasonal variety. The 2DVD data in Oklahoma were obtained by the National Severe Storms Laboratory (NSSL), National Oceanic and Atmospheric Administration (NOAA), and the data in Jincheon were provided by the Weather Radar

Center (WRC), Korea Meteorological Administration (KMA). The other data were collected by the Center for Atmospheric REmote sensing (CARE), Kyungpook National University (KNU).

We discarded the time if the radar reflectivity was greater than 55 dBZ in order to exclude hail particles in the analysis [6]. We also removed the time if the rainfall rate was less than 0.1 mm h^{-1} because the disdrometer typically underestimates the rainfall rate when small drops (diameter < 0.7 mm) are dominant [30]. The maximum rainfall rate was larger in OKL than the sites in ROK. The median rainfall rate (radar reflectivity) varied from 0.99 mm h^{-1} (22.92 dBZ) to 1.78 mm h^{-1} (27.72 dBZ). Table 1 shows the clearly different statistical characteristics of rainfall in different climates (OKL vs. ROK, different regions in ROK) [29]. Unlike the maximum rainfall rate, the maximum reflectivity showed less discrepancy. These characteristics certainly have impact on the ML models and will be discussed later.

Table 1. The two-dimensional video disdrometer (2DVD) data used in this study.

Area	Period [Year]	Number of 1-Min Data	Median of 1-min Rain Rate [mm h^{-1}]	Median of 1-min Reflectivity [dBZ]	Maximum of 1-min Rain Rate [mm h^{-1}]	Maximum of 1-min Reflectivity [dBZ]
Oklahoma, USA (OKL)	1996–2006 (May to September)	7944	1.78	27.72	133.39	54.88
Daegu (DAE)	2011–2012 (May to September)	7516	1.36	25.09	99.24	52.50
Boseong (BOS)	2013–2015, 2018 (May to September)	12,083	0.99	22.92	93.39	53.95
	2018 (October)	713				
Jincheon (JIN)	2013–2015, 2018 (May to September)	22,731	1.06	23.55	76.46	54.11
	2018 (October)	315				
	2019 (April)	545				
Total		51,302			-	

Dual-polarization radar variables are obtained by T-matrix scattering simulation [31], and 5-min time average values of Z_h, Z_{DR}, and K_{DP} were additionally used to consider the movement of the precipitation system ($Z_{h\,5min}$, $Z_{DR\,5min}$, and $K_{DP\,5min}$). The T-matrix method used to calculate the polarimetric radar variables is one of the most widely used tools for computing light scattering by non-spherical particles based on directly solving Maxwell's equations. This approach can simulate theoretical radar measurements for homogeneous and rotationally symmetric non-spherical particles. Backward scattered fields yield Z_H, Z_{DR}, and ρ_{HV}, while forward scattered fields produce K_{DP} [17]. The control conditions and the values used in this study are shown in Table 2. The radar wavelength was 11.01 cm

(S-band), and the elevation angle of the radar was set at 0°. The raindrop shape formula suggested by Thurai et al. [32] was used.

Table 2. The control conditions and their corresponding values used in the T-matrix scattering simulation.

Characteristics	Values
Radar wavelength	11.01 cm (S-band)
Radar elevation angle	0°
Environment temperature	23 °C
Drop shape formula	Taken from Thurai et al. (2007)

2.2. Operational MYN S-band Dual-polarization Radar Data

The weather radar used in this study was the MYN S-band dual-polarization radar operated by the Korea Meteorological Administration (KMA). Table 3 shows the detailed specifications of the MYN radar. A volume scan of nine elevation angles (0°, 0.39°, 0.83°, 2°, 2.88°, 4.06°, 5.67°, 7.88°, and 10.94°) was performed every 10 min with a beam width of 0.92°. The measured parameters were Z_H, Z_{DR}, K_{DP}, ρ_{HV}, etc. Z_H and Z_{DR} were calibrated through post-processing. The averaged Z_H calibration bias was calculated by the self-consistency principle of Z_H and K_{DP}, and the averaged Z_{DR} calibration bias was conducted by comparing the Z_{DR} and Z_H radar measurements with the theoretical relationship between the same parameters simulated by the 2DVD [33,34]. The averaged calibration bias of Z_H was −5.0 dBZ, and the averaged calibration bias of Z_{DR} was 0.03 dB.

Table 3. Specifications of the Mt. Myeongbong (MYN) S-band dual-polarization radar.

Parameter	Value
Frequency (wavelength)	2272 MHZ (10 cm, S-band)
Location	36° 10′ 45″ N, 128° 59′ 50″ E
Height	1136 m
Beam width	0.92°
Elevation angles	0°, 0.39°, 0.83°, 2°, 2.88°, 4.06°, 5.67°, 7.88°, and 10.94°
Maximum range	285 km

Rain gauge data from a total of 192 rain gauges in KMA AWSs within the MYN radar observation range (150 km) were used to verify the radar QPE (Figure 1). Each AWS was equipped with two sizes of tipping-bucket rain gauge, 0.1 and 0.5 mm, which measure the 1-min R. In this study, the 10-minute average rain rate was used to match the time resolution with the radar data observed at 10-minute intervals, and missing values were excluded when calculating the 10-minute average rain rate. We analyzed six rainfall cases for QPE and verification. The rainfall events included stratiform rain (Cases 1, 2, 3, and 6) and convective rain (Cases 4 and 5) from 2017 to 2018 (Table 4).

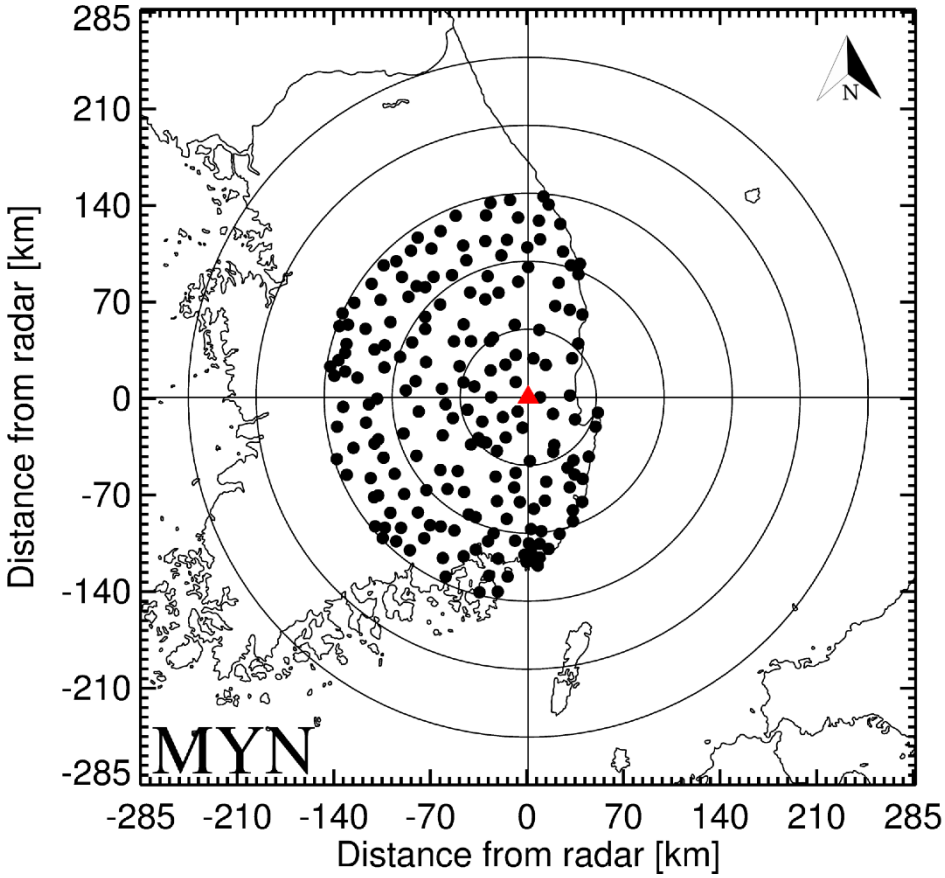

Figure 1. The locations of 192 automatic weather stations (dots) within the radar observation range of the MYN radar (red filled triangle). Black rings denote radar range rings with a 50 km interval.

Table 4. List of precipitation period and rain types of rainfall events.

Case No.	Period (LST)	Rain Type
1	0000–1200 August 14 2017	Stratiform
2	0200–1100 September 11 2017	Stratiform
3	0200–0700 July 01 2018	Stratiform
4	1700 Aug. 27–0600 August 28 2018	Convective
5	1000–1600 September 03 2018	Convective
6	0500–1000 September 07 2018	Stratiform

3. Methods

3.1. Machine Learning

In this study, RT and RF were used for rainfall estimation. Figure 2 shows a schematic diagram of the RT. We defined N to be the number of RTs, and M to be the number of independent variables. A node was divided into sub-nodes with the lowest variance [21]. A recursive binary partition was conducted for each node until a stop condition was met. The most important independent variable was placed at the top of the tree as a root node. The node divided from the root node is called the intermediate node, and the node that reaches the last is called the terminal node.

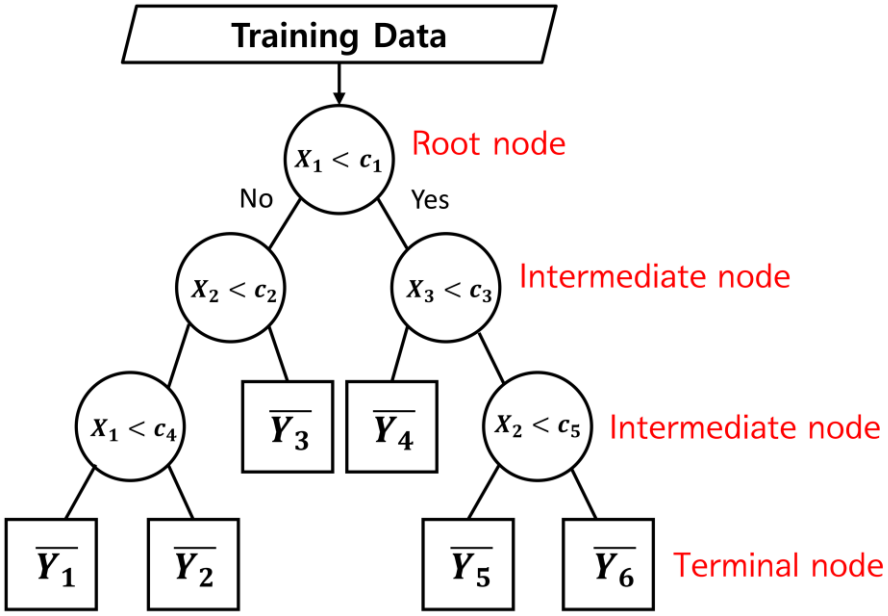

Figure 2. Schematic diagram of the regression tree. X_i is an independent variable, and c_i denotes an optimal splitting criterion value for the corresponding independent variables. The $\overline{Y_j}$ denotes the average value of the data belonging to the *j*th terminal node.

The RF was based on N RTs with N bootstrap samples (see Figure 3). The bootstrap samples were generated by sampling with replacement. Each RT was grown with the splitting rules using the different independent variables selected randomly. The final prediction was given by the average of the predictions from all RTs. In the RF, the importance of independent variables was measured through the increase of the node purity. The independent variable with the highest increase of node purity played a major role in the prediction.

RF can be optimized by tuning two parameters when generating RTs. One is the number of RTs (n_{tree} = N), and the other is the number of independent variables that are randomly sampled (m_{try} < M). Liaw and Wiener [35] suggested that n_{tree} was 500, and m_{try} was $\sqrt{M}$ for classification, as opposed to a third of M for regression. Kühnlein et al. [19] compared out-of-bag (OOB) errors by changing the n_{tree} and m_{try} to improve the predictability of RF, which was used to select important independent variables and to compute the error of the unbiased estimate [22]. To determine the optimal values with the lowest OOB error in this study, we considered the range of values for each tuning parameter, from 400 to 700 for n_{tree} and from 1 to the number of independent variables for m_{try}. The optimal n_{tree} and m_{try} were applied to each model.

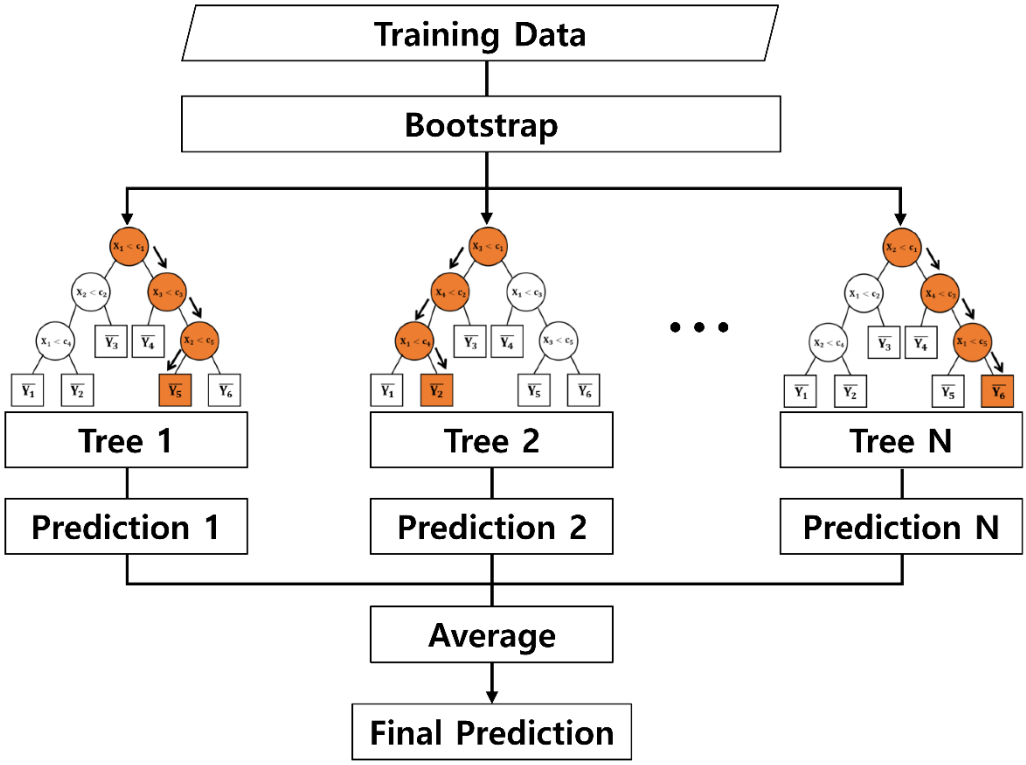

Figure 3. Schematic diagram of the random forest (RF) model.

3.2. Rainfall Estimation

3.2.1. R–Z Relationship

For validation of the ML models used in this study, we derived the empirical rainfall estimation relationships based on dual-polarization radar variables. Several R–Z relationships were considered with different thresholds of K_{DP} and Z_{DR}. The R–Z relationship calculated using all training data is shown as Equation (2), and Equation (3) was retrieved with the data below the threshold of K_{DP} and Z_{DR} (K_{DP} < 0.04° km^{-1} and Z_{DR} < 0.3 dB). Equation (4) was constructed with the data above the thresholds of K_{DP} and Z_{DR} (K_{DP} ≥ 0.04° km^{-1} or Z_{DR} ≥ 0.3 dB). The equations are assumed to have a power-law ($Y = aX^b$), and the parameters a and b were estimated using the weighted total least squares method [36].

$$R(Z_h) = 0.030\ Z_h^{0.667}\ (Z_h = 197\ R^{1.50}) \quad (2)$$

$$R(Z_h) = 0.017\ Z_h^{0.806}\ (Z_h = 151\ R^{1.24}) \text{ when } K_{DP} < 0.04° \text{ km}^{-1} \text{ and } Z_{DR} < 0.3 \text{ dB} \quad (3)$$

$$R(Z_h) = 0.012\ Z_h^{0.769}\ (Z_h = 318\ R^{1.30}) \text{ when } K_{DP} \geq 0.04° \text{ km}^{-1} \text{ or } Z_{DR} \geq 0.3 \text{ dB} \quad (4)$$

$$R(K_{DP}) = 42.6\ K_{DP}^{0.720} \quad (5)$$

$$R(Z_h, Z_{DR}) = 0.003\ Z_h^{0.913}\ Z_{DR}^{-0.647} \quad (6)$$

3.2.2. ML-Based Estimation

The 2DVD data above the thresholds of K_{DP} and Z_{DR} (K_{DP} ≥ 0.04° km^{-1} or Z_{DR} ≥ 0.3 dB) were used as training data. The dual-polarization radar parameters simulated by the T-matrix and the 5-min time average values of parameters were used as the independent variables. In this study, we investigated the impacts of three factors on the estimation accuracy. First, three types of dependent variable are considered: R_{2DVD} calculated by N(D)

(M1), the residual, $\varepsilon = R(Z_h) - R_{2DVD}$, between R_{2DVD} and R—which was computed from Equation (4) (M2)—and the normalized residual, $\bar{\varepsilon} = \varepsilon/\overline{R(Z_h)}$, (M3).

Second, two different groups of the independent variables were used, and the difference between the two groups included K_{DP}. We denote the two groups by KY for the group with K_{DP} and KN for the group without K_{DP}. Lastly, data binning was implemented to group individual observations into specific bins determined by reflectivity (Table 5), allowing us to train the models locally. The local training with the binned training data and global training with the entire training data are denoted by CY and CN, respectively. In the local training (CY) with RF, n_{tree} and m_{try} with the lowest OBB error for each model were utilized.

Table 5. The reflectivity (Z_H) interval and the number of training data.

Class No.	Interval [dBZ]	Number of Observations
1	$0 \leq Z_H < 20$	599
2	$20 \leq Z_H < 30$	11,023
3	$30 \leq Z_H < 40$	9369
4	$40 \leq Z_H$	1639

K_{DP} and Z_{DR} provide more accurate rainfall estimation than $R(Z_h)$ for heavier rainfall, whereas the improvement is not often significant for lighter rainfall due to the noises of K_{DP} and Z_{DR} [13,37–39]. Silvestro et al. [40] proposed a rainfall estimation algorithm that makes use of the best empirical relationships depending on the thresholds of K_{DP} and Z_{DR}, which outperformed $R(Z_h)$ for real-time applications. In this study, the rainfall estimation based on this algorithm with different thresholds of K_{DP} and Z_{DR} was performed, which is shown in Figure 4. When K_{DP} was greater than 0.04° km^{-1}, KY was utilized for rainfall estimation, KN was used if K_{DP} was less than 0.04° km^{-1}, and Z_{DR} was 0.3 dB or more. The rainfall was estimated by the R–Z relationship (Equation (3)) when both K_{DP} and Z_{DR} were less than the thresholds. The models used in this study are summarized in Table 6.

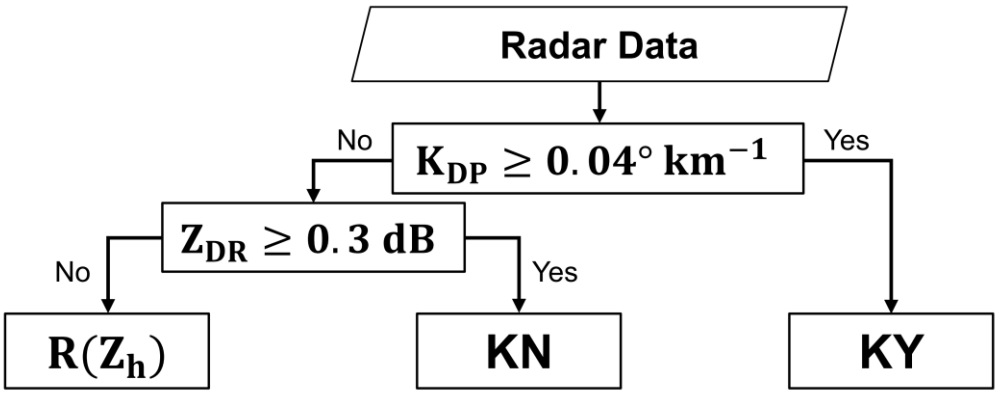

Figure 4. Flowchart of the rainfall estimation in this study.

Table 6. Summary of the models used in this study.

Independent Variables	Training Set	Dependent Variables		
		Rain rate (R_{2DVD}) (M1)	Residual ($\varepsilon = R(Z_h) - R_{2DVD}$) (M2)	Normalized residual ($\bar{\varepsilon} = \varepsilon/\overline{R(Z_h)}$) (M3)

Z_h, Z_{DR}, K_{DP}, ρ_{HV}, $Z_{h\,5min}$, $Z_{DR\,5min}$, $K_{DP\,5min}$ (KY)	Not classified training set (CN)	M1KYCN	M2KYCN	M3KYCN
	Classified by reflectivity interval (CY)	M1KYCY	M2KYCY	M3KYCY
Z_h, Z_{DR}, ρ_{HV}, $Z_{h\,5min}$, $Z_{DR\,5min}$ (KN)	Not classified training set (CN)	M1KNCN	M2KNCN	M3KNCN
	Classified by reflectivity interval (CY)	M1KNCY	M2KNCY	M3KNCY

3.2.3. Validation

The trained models were verified by 10-fold cross-validation. Six statistics were used to assess the performance of the ML models: The root mean square error (RMSE), mean absolute error (MAE), bias, correlation coefficient (CORR), coefficient of efficiency (COE) [41], and normalized error (1-NE),

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(R_{est_i} - R_{obs_i}\right)^2} \quad (7)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|R_{est_i} - R_{obs_i}\right| \quad (8)$$

$$Bias = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_{est_i}}{R_{obs_i}}\right) \quad (9)$$

$$CORR = \frac{\sum_{i=1}^{N}(R_{est_i} - \overline{R_{est}})(R_{obs_i} - \overline{R_{obs}})}{\sqrt{\sum_{i=1}^{N}(R_{est_i} - \overline{R_{est}})^2 \sum_{i=1}^{N}(R_{obs_i} - \overline{R_{obs}})^2}} \quad (10)$$

$$COE = 1 - \frac{\sum_{i=1}^{N}\left(R_{obs_i} - R_{est_i}\right)^2}{\sum_{i=1}^{N}\left(R_{obs_i} - \overline{R_{obs}}\right)^2} \quad (11)$$

$$1 - NE = 1 - \frac{\left(\frac{1}{N}\sum_{i=1}^{N}\left|R_{est_i} - R_{obs_i}\right|\right)}{\overline{R_{obs}}} \quad (12)$$

where N is the number of observations in test data, R_{est} represents the estimated rainfall rate, and R_{obs} is the observed rainfall rate.

3.3. Application to Operational Radar Data

The ML model with the highest accuracy in the 10-fold cross-validation using 2DVD was applied to the rainfall estimation using the operational dual-polarization radar data. The operational radar data used were conducted using the Hybrid Surface Rainfall method (HSR) of the MYN S-band dual-polarization radar data. The HSR is a technique of generating a rainfall field using the data of the lowest elevation angle that is not affected

by ground clutter, beam blockage, and non-meteorological echoes. It is applied to the radar data in polar coordinates, and we selected the rain field using the threshold and calibrated for the radar bias (Z_H and Z_{DR}).

When determining the rain field, that is, eliminating non-meteorological echoes or removing artifacts, the threshold values of the following texture were used as follows [42,43]: 0.95 for ρ_{HV}, 0.1 for $\delta(\rho_{HV})$, 4.0 dB for $\delta(Z_H)$, 4.0 dB for $\delta(Z_{DR})$, and 15.0° for $\delta(\Phi_{DP})$. Here, $\delta(x)$ is the radial texture of variable x with a window size of 10. The HSR data at polar coordinates are converted to Cartesian coordinates [42].

The ML-based rainfall estimation was performed for each grid point in the Cartesian coordinates. Similar to the training data, the 5 × 5 km spatial average of Z_h, Z_{DR}, and K_{DP} were considered as the independent variables by taking into account the movement of the precipitation system. The estimated rainfall rate was verified by the rain gauges in the AWSs within the radar observation range (150 km). The rainfall rate estimated from the radar grid point closest to the AWS was compared with the average rainfall rate measured by the rain gauges.

4. Results

4.1. Rainfall Estimation from Simulated Dual-Polarization Variables

Figure 5 shows the fitted tree models for M1KYCN and M1KNCN (with and without K_{DP}). See Table 6 for the symbols such as M1, KY, CN, etc. We found that K_{DP} plays an important role in rainfall estimation with M1KYCN because it is used as a splitting criterion at the root node. On the other hand, K_{DP} is not shown in the tree based on M1KNCN, while Z_h is the most crucial for the model.

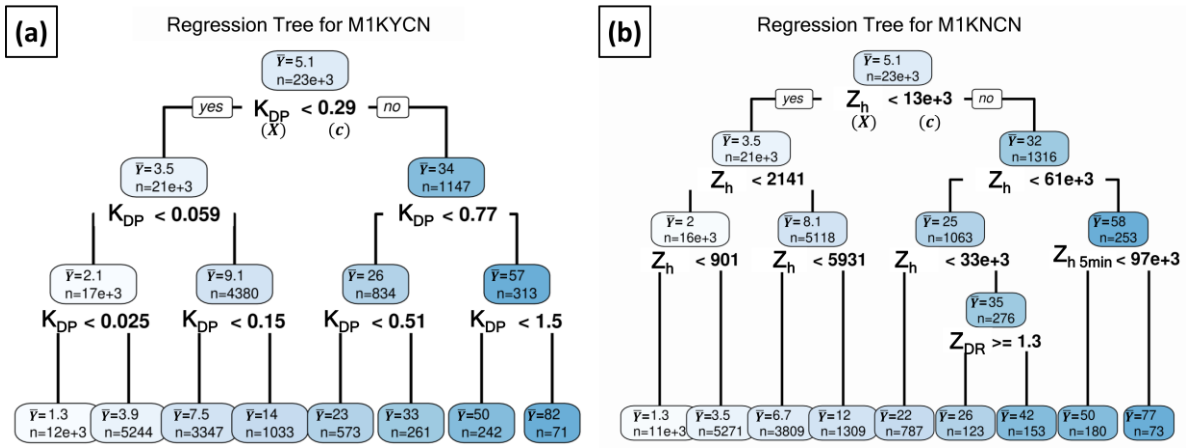

Figure 5. Results of the classification by the (**a**) M1KYCN and (**b**) M1KNCN regression trees. n is the number of observations. *X*, *c*, and $\bar{Y}$ are the same as in Figure 2.

The increase of node purity in RF for CN is shown in Table 7. To account for potential overfitting in the random forest, we optimized the tuning parameters, the number of RT (n_{tree}), and the number of independent variables randomly sampled (m_{try}). The values were expressed according to the inclusion of K_{DP} for each model. For M1KY, the node purity rose the most (713,059) when K_{DP} was used as the criterion for split. Then, in the order of Z_h (426,349) $K_{DP\,5min}$ (311,490), and $Z_{h\,5min}$ (147,971), the increase of node purity was higher. In the case of M1KN, the node purity increased the most when the node was divided by Z_h. This indicates that Z_h played the most major role in the estimation of the rainfall rate.

Unlike M1KY, the increase in the node purity of Z_{DR} was the highest in the M2KY. This indicates that the errors, which cannot be explained by the R–Z relationship, were most closely related to Z_{DR}. The second most important variable in M2KY is $Z_{DR\,5min}$. Similar to KY, Z_{DR} was the most important variable in M2KN, and the importance was the highest in the order of Z_h and $Z_{DR\,5min}$. The tendency of increasing node purity of M3 was

similar to that of M2 regardless of whether K_{DP} was included or not. As expected, ρ_{HV} was mostly shown to be less important, as it has a very small fluctuation in rainfall cases.

Table 7. Increase of the node purity in RFs for CN. The highest increases of node purity are highlighted in bold.

		M1	M2	M3
KY	Z_h	426,349	75,422	2734
	Z_{DR}	36,434	**207,260**	**7153**
	K_{DP}	**713,053**	29,997	1019
	ρ_{HV}	3883	22,859	641
	$Z_{h\ 5min}$	147,971	19,286	635
	$Z_{DR\ 5min}$	36,618	87,982	2993
	$K_{DP\ 5min}$	311,490	11,326	385
KN	Z_h	**832,321**	101,335	3347
	Z_{DR}	106,485	**200,627**	**6948**
	ρ_{HV}	10,998	22,430	712
	$Z_{h\ 5min}$	565,867	32,965	1192
	$Z_{DR\ 5min}$	147,220	97,432	3308

Table 8 presents the increase in the node purity of CY. In M1KY, K_{DP} is shown as the most important variable in all reflectivity intervals as in CN (Table 6), and Z_h is indicated as the second most important variable. Similar to CN, M2KY and M2KN always show Z_{DR} as the most important variable. The result of M3 is omitted since it appears with the same tendency as that of M2.

Table 8. Increase of the node purity in RFs for CY. The highest increases of node purity are highlighted in bold.

		M1				M2			
		$0 \le Z_H < 20$	$20 \le Z_H < 30$	$30 \le Z_H < 40$	$40 \le Z_H$	$0 \le Z_H < 20$	$20 \le Z_H < 30$	$30 \le Z_H < 40$	$40 \le Z_H$
KY	Z_h	0.803	1351	21,850	97,115	0.080	103	3064	51,941
	Z_{DR}	0.240	357	10,560	24,184	**0.355**	**484**	**16,522**	**166,202**
	K_{DP}	**1.197**	**1861**	**35,962**	**255,737**	0.112	155	4303	17,446
	ρ_{HV}	0.041	30	2625	2413	0.035	62	4856	26,171
	$Z_{h\ 5min}$	0.202	579	7727	29,802	0.081	77	1225	8406
	$Z_{DR\ 5min}$	0.192	265	8624	15,266	0.194	257	11,103	58,468
	$K_{DP\ 5min}$	0.273	875	15,248	88,529	0.089	130	2167	6925
KN	Z_h	**1.771**	**2987**	**51,455**	**261,552**	0.107	224	62,096	62,096
	Z_{DR}	0.363	490	15,227	46,510	**0.418**	**545**	**153,789**	**153,789**
	ρ_{HV}	0.0616	34	3188	9538	0.051	55	33,457	33,457
	$Z_{h\ 5min}$	0.443	1443	20,805	148,684	0.130	181	14,208	14,208
	$Z_{DR\ 5min}$	0.296	350	11,760	43,250	0.230	261	67,743	67,743

The scatterplots of the 10-fold cross-validation for the RT are displayed in Figure 6. The discontinuity of the estimated value is a prominent problem of RT. For the weak R_{2DVD} ($R_{2DVD} < 20$ mm h^{-1}), there is a continuity of data because rainfall is often estimated using the R–Z relationship when the R_{2DVD} is weak and the Z_{DR} and K_{DP} are small. In M1CY

(Figure 6b), since the training data are divided by the reflectivity interval, the discontinuity problem is rectified, and the 1-NE value increased 7.08% compared to M1CN. M2 (Figure 6c,d) and M3 (Figure 6e,f) demonstrate similar results and showed a more continuous value than M1, with a higher positive CORR value (CORR > 0.98). Additionally, CY appears lower in the RMSE, MAE, and bias values compared with those of CN, and the CORR, COE, and 1-NE values are higher.

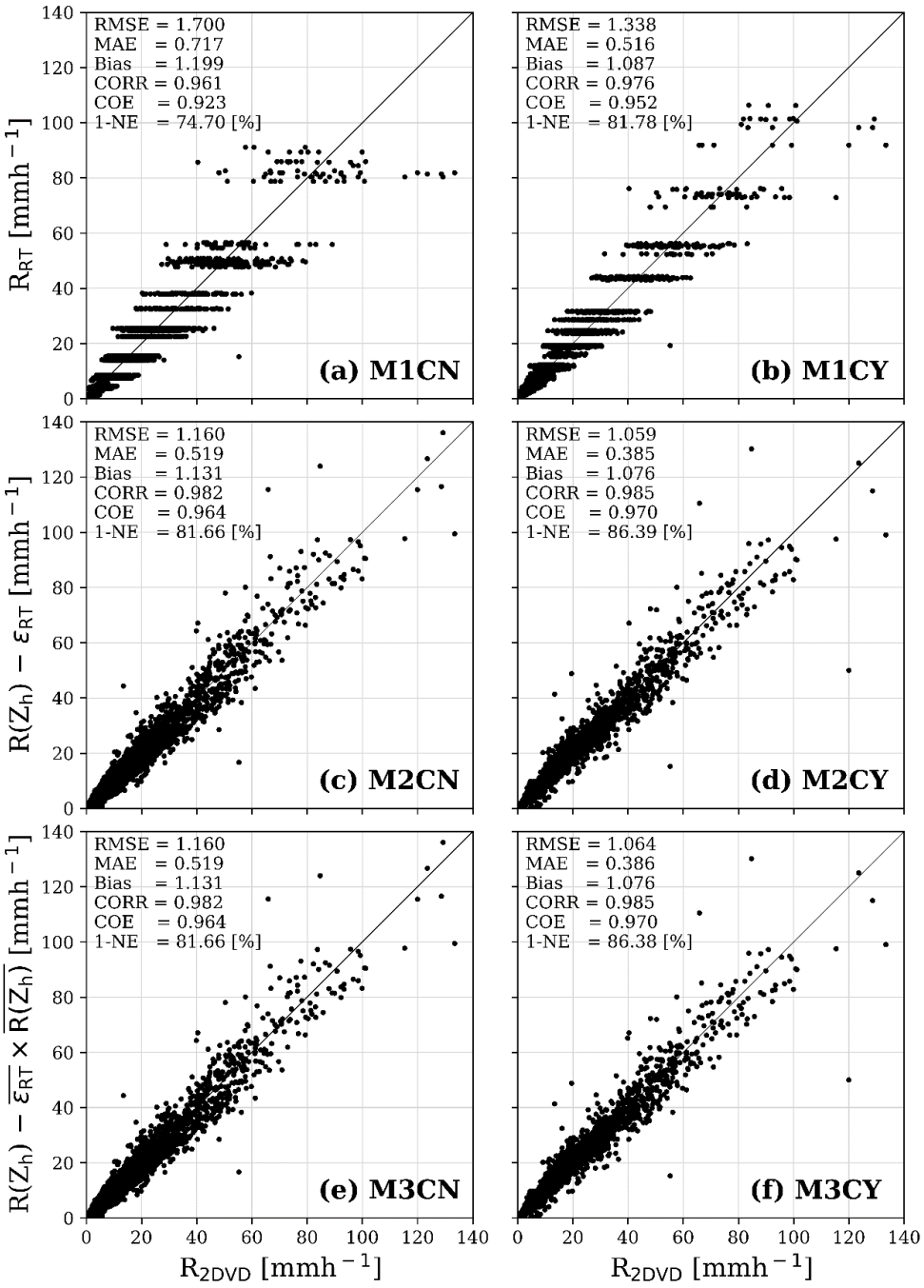

Figure 6. Scatter plots of R_{2DVD} and the rainfall rate estimated using a regression tree (RT) with 10-fold cross-validation. (**a**) M1CN, (**b**) M1CY, (**c**) M2CN, (**d**) M2CY, (**e**) M3CN, and (**f**) M3CY.

Figure 7 presents the results of the 10-fold cross-validation of RF. Overall, RF-based models show improved accuracy compared to the RT-based models in Figure 6. In M1CN (Figure 7a), the CORR value and the COE value are higher than 0.98, even though it pre-

sented the worst statistics among the RF models. Compared to M1CN, the underestimation in the strong R_{2DVD} (R_{2DVD} > 80 mm h^{-1}) was corrected in M2CN and M3CN (Figure 7c,e), and the RMSE reduced by 4.93%. The RMSE, MAE, and bias decreased in M1CY (Figure 7b) compared with in M1CN, whereas the RMSE, MAE, and bias increased in M2CY and M3CY (Figure 7d,f).

M2CN outperformed the other models with an RMSE value of 0.598, MAE value of 0.255, CORR value of 0.995, and 1-NE value of 90.99%. The M3CN shows a similar performance with M2CN. There was no significant improvement between CN (left panels) and CY (right panels). Therefore, the ensemble effect by CY was less affected in the RF-based models compared with in the RT-based model. This is because RF itself is based on the ensemble. The interesting thing is that CN in RF even showed a slightly better score than CY. This could be explained by the fact that CN uses more training data and is free from possible local features in CY.

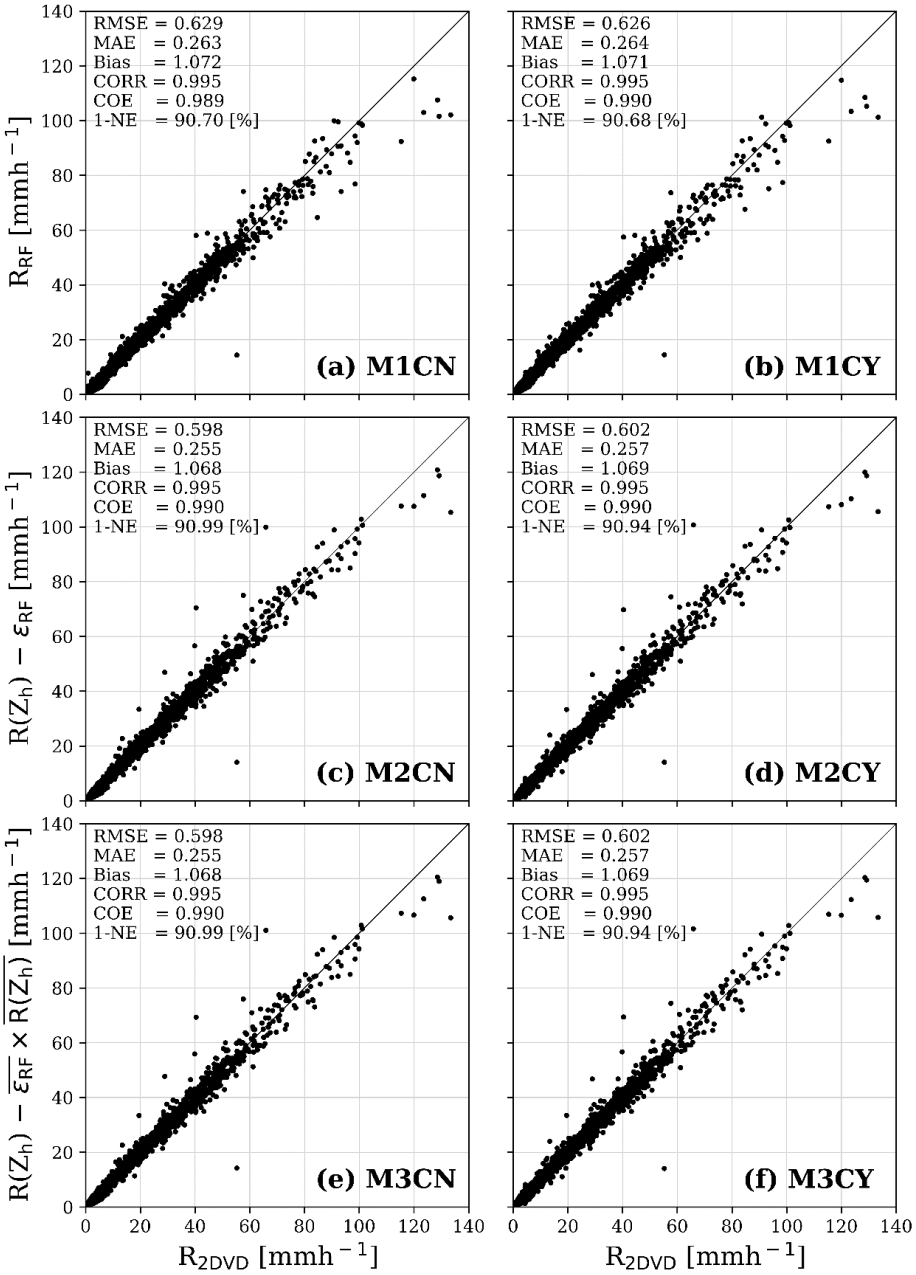

Figure 7. Scatter plots of R_{2DVD} and the rainfall rate estimated by RF using 10-fold cross-validation. (**a**) M1CN, (**b**) M1CY, (**c**) M2CN, (**d**) M2CY, (**e**) M3CN, and (**f**) M3CY.

The scatterplots of R estimated by the empirical R–Z relationship are shown in Figure 8. The rainfall rate estimated by $R(Z_h)$ calculated from the entire training data (Figure 8a) presents a large dispersion overall and a tendency to underestimate at a rainfall rate of 60 mm h^{-1} or higher. When the rainfall rate was estimated based on Equations (3) and (4) according to the threshold of K_{DP} and Z_{DR} (Figure 8b), the RMSE and MAE increased compared to the case of using one $R(Z_h)$ due to overestimating at 40–100 mm h^{-1}; however, the underestimation of the high rainfall rate ($R > 100$ mm h^{-1}) was resolved.

Estimation using K_{DP} (Figure 8c) had the lowest RMSE value (1.509) and the highest CORR value (0.974) and COE value (0.939) among the estimations using empirical relationships; however, there was still underestimation overall. On the other hand, in the case of the relationship using Z_h and Z_{DR} (Figure 8d), the highest 1-NE value (85.24%) and the lowest MAE value (0.418) are presented, but most of the R_{2DVD} values were overestimated. Among the ML models, the model that showed the lowest performance was M1CN-RT, with an RMSE value of 1.700 and CORR value of 0.961 (see Figure 6). This indicates that all ML-based models outperformed these empirical relationship-based approaches.

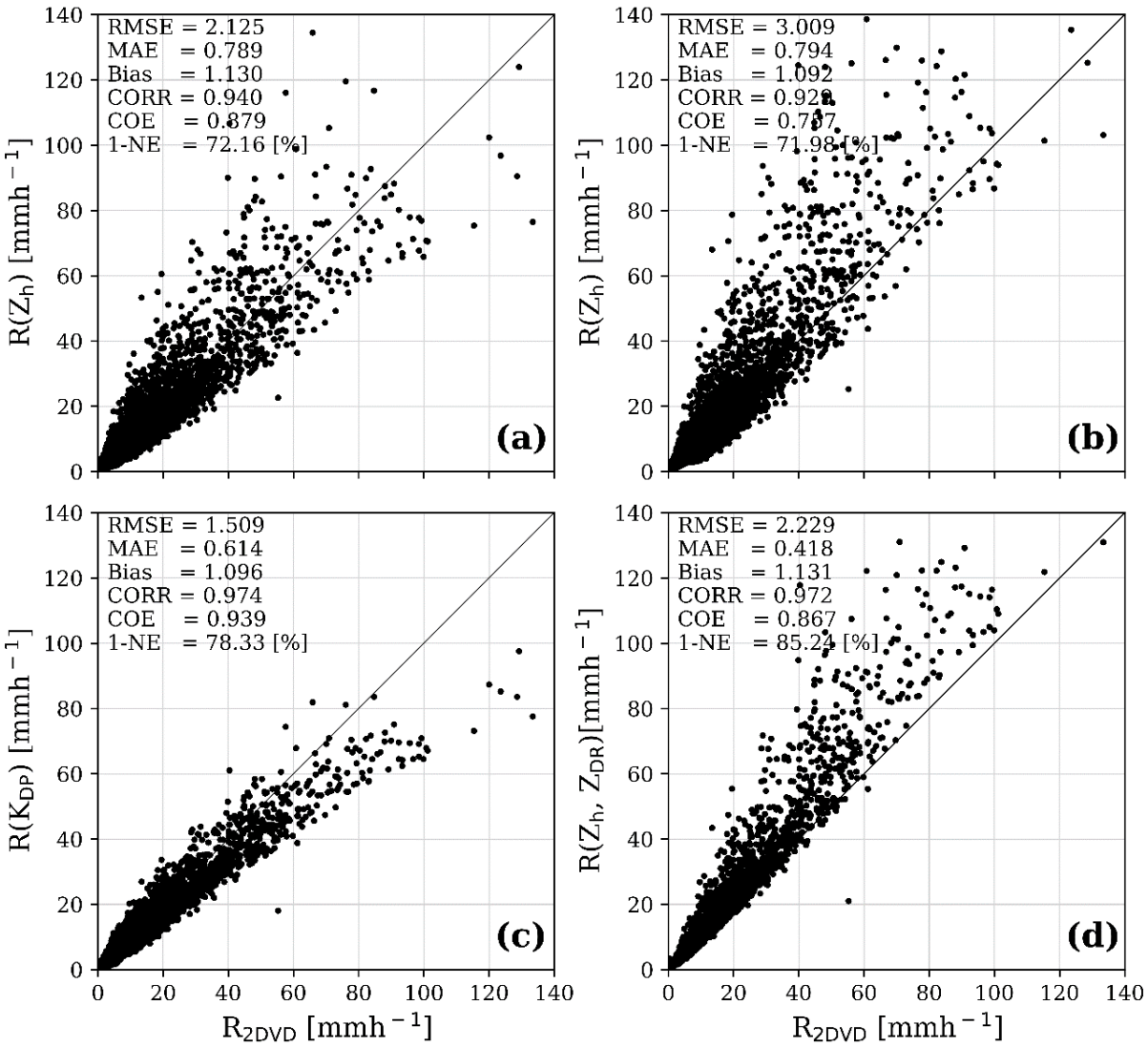

Figure 8. Scatter plots of the rainfall rate from empirical relationships. (**a**) $R(Z_h)$, (**b**) adjusted $R(Z_h)$, (**c**) $R(K_{DP})$, and (**d**) $R(Z_h, Z_{DR})$.

4.2. Rainfall Estimation from Operational Radar

In the results of the 10-fold cross-validation of the ML models, the M2CN-RF and M3CN-RF showed the best and similar performance in terms of the RMSE and CORR. In this section, we chose M2CN-RF for rainfall estimation with MYN S-band dual-polarization radar data. The HSR images of the rainfall rate of Case 1 and Case 2 are displayed in Figures 9 and 10. $R(Z_h)$ was estimated by Equation (3) when K_{DP} and Z_{DR} were below the threshold value and by Equation (4) when they were above the threshold value (Figures 9a,10a).

The M2CN-RF (Figures 9c,10c) adjusted the rainfall rate from $R(Z_h)$ by applying the residuals. In the HSR rainfall image, the gray region ($\varepsilon = 0$) in Figures 9c and 10c correspond to the area in which $R(Z_h)$ was replaced due to the threshold values of K_{DP} and Z_{DR}. A positive ε value indicates that $R(Z_h)$ was overestimated, while a negative ε value indicates that $R(Z_h)$ was underestimated. For reference, the HSR of the radar observed variables is shown in Figures 9(d)–(f) and 10(d)–(f).

In Case 1 (1000 LST on August 14, 2017), which is a stratiform case, the area of $\varepsilon = 0$ is wider due to lower values of K_{DP} and Z_{DR}. Although the $R(Z_h)$ field is highly correlated with the Z_h field, the ε field is less correlated with the Z_h field. The large (smaller) value of Z_{DR} is related to the larger positive (negative) ε. The ε is less correlated with K_{DP} in this stratiform case since most of the K_{DP} value is smaller. In Case 2 (0730 LST on September 11, 2017), which is a stratiform rain event with embedded convection, the area of $\varepsilon = 0$ is smaller than that in Case 1. Similar to Case 1, the positive ε area is highly correlated with the higher value of Z_{DR}. Interestingly, the region of higher K_{DP} with $Z_h > 50$ dBZ in the south direction had the largest negative values of ε.

In both cases, it can be seen that ε appeared in space with significant structure, indicating that the error generated from the R–Z relationship had a spatial structure. This result is not surprising because inhomogeneous microphysical processes cause the natural spatial DSD variation and result to the spatial structure of residual from R(Z) [11,12]. The ML model (M2CN-RF) uses the simulated dual-polarization variables, which do not have instrumental noises of the radar, such as sampling error, beam broadening, beam blockage, and miscalibration of the radar [12].

As a result, we can expect the spatial structure of residual will be masked if the order of the magnitude of instrumental noises is comparable to that of natural variation of DSDs. The random variability of instrumental noise can be reduced by averaging the samples. Since the M2CN-RF is an ensemble-based model, the spatial structure of residual can be clearer as the instrumental noises are diminished by averaging the prediction of each RT.

Figures 11 and 12 show the verification of the period average rainfall rate (12 h for Case 1, and 11 h for Case 2) for estimation with the empirical relationships and ML. In Case 1, the RMSE value of ML ($R(Z_h) - \varepsilon_{RF}$) is 1.207 and the CORR is 0.773 (Figure 11a), which shows improved performance compared to those estimated by the empirical relationships. When calculated by Equation (2) (Figure 11b), there is a tendency to underestimate when $R > 5$ mm h^{-1}. This underestimation is slightly improved by applying Equations (3) and (4) based on the threshold values of K_{DP} and Z_{DR}, and the CORR value increases (Figure 11c).

Underestimation still appears with the heavier rainfall rate. This underestimation of the rainfall rate estimated by $R(Z_h)$ is improved by ML in Figure 11a, adjusting the rainfall rate with ε_{RF}. The estimation based on Equation (5) results in substantial underestimation with a low CORR value and the worst performance (Figure 11d). The results estimated by Equation (6) show a severe overestimation of strong R_{AWS} ($R_{AWS} > 5$ mm h^{-1}) (Figure 11e).

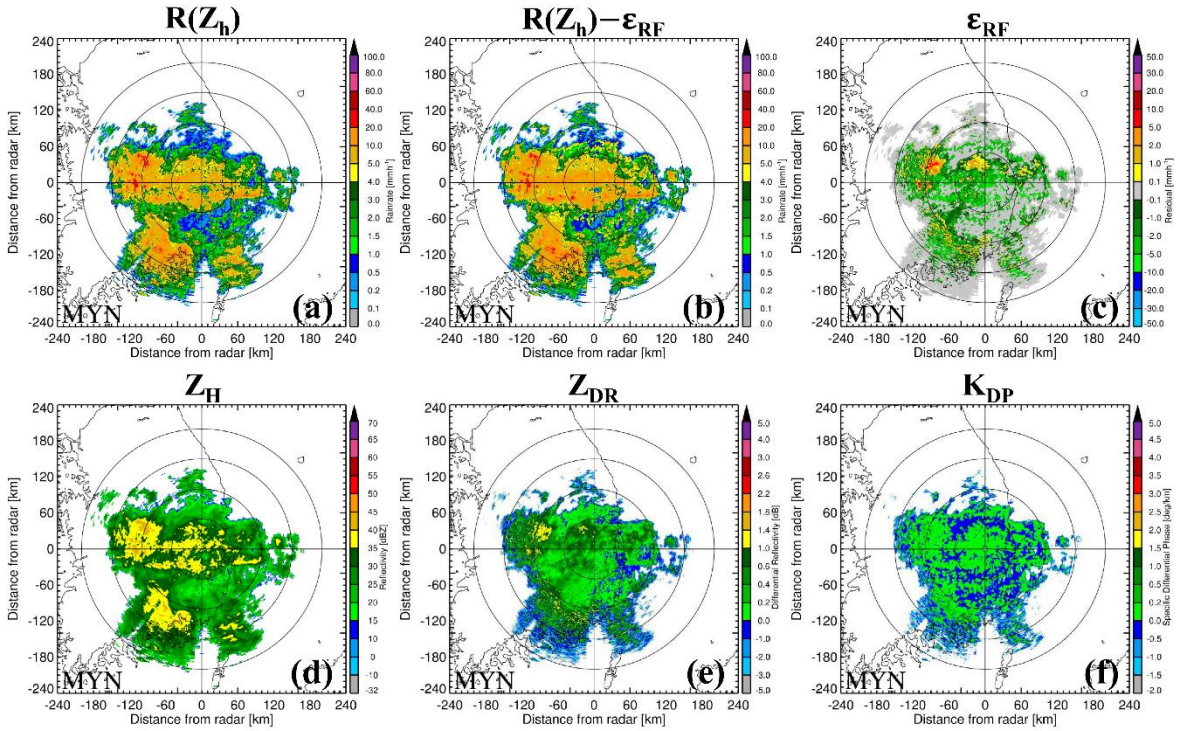

Figure 9. Hybrid Surface Rainfall method (HSR) images of the (**a**) estimated rain rate by Z–R relationship ($R(Z_h)$), (**b**) adjusted rain rate ($R(Z_h) - \boldsymbol{\varepsilon_{RF}}$), (**c**) estimated residual by RF ($\boldsymbol{\varepsilon_{RF}}$), (**d**) Z_H, (**e**) Z_{DR}, and (**f**) K_{DP} at Case 1 (1000 LST on August 14, 2017).

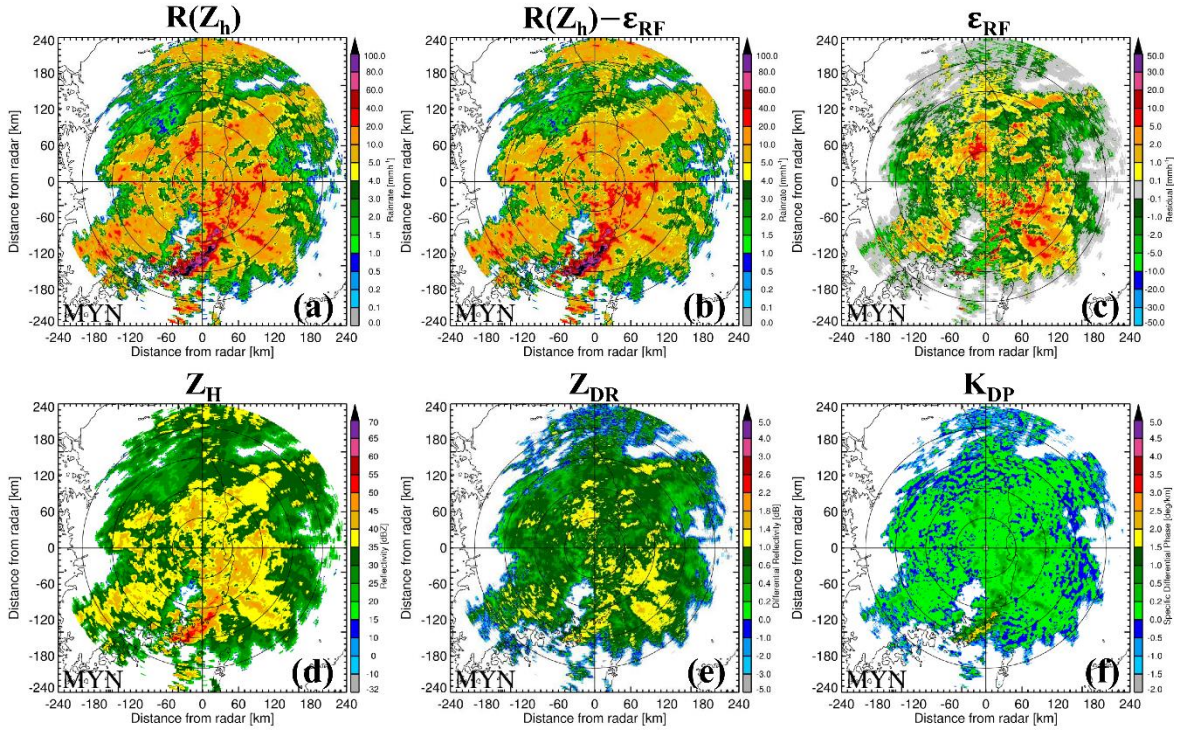

Figure 10. HSR images of the (**a**) estimated rain rate by Z–R relationship ($R(Z_h)$), (**b**) adjusted rain rate ($R(Z_h) - \boldsymbol{\varepsilon_{RF}}$), (**c**) estimated residual by RF ($\boldsymbol{\varepsilon_{RF}}$), (**d**) Z_H, (**e**) Z_{DR}, and (**f**) K_{DP} at Case 2 (0730 LST on September 11, 2017).

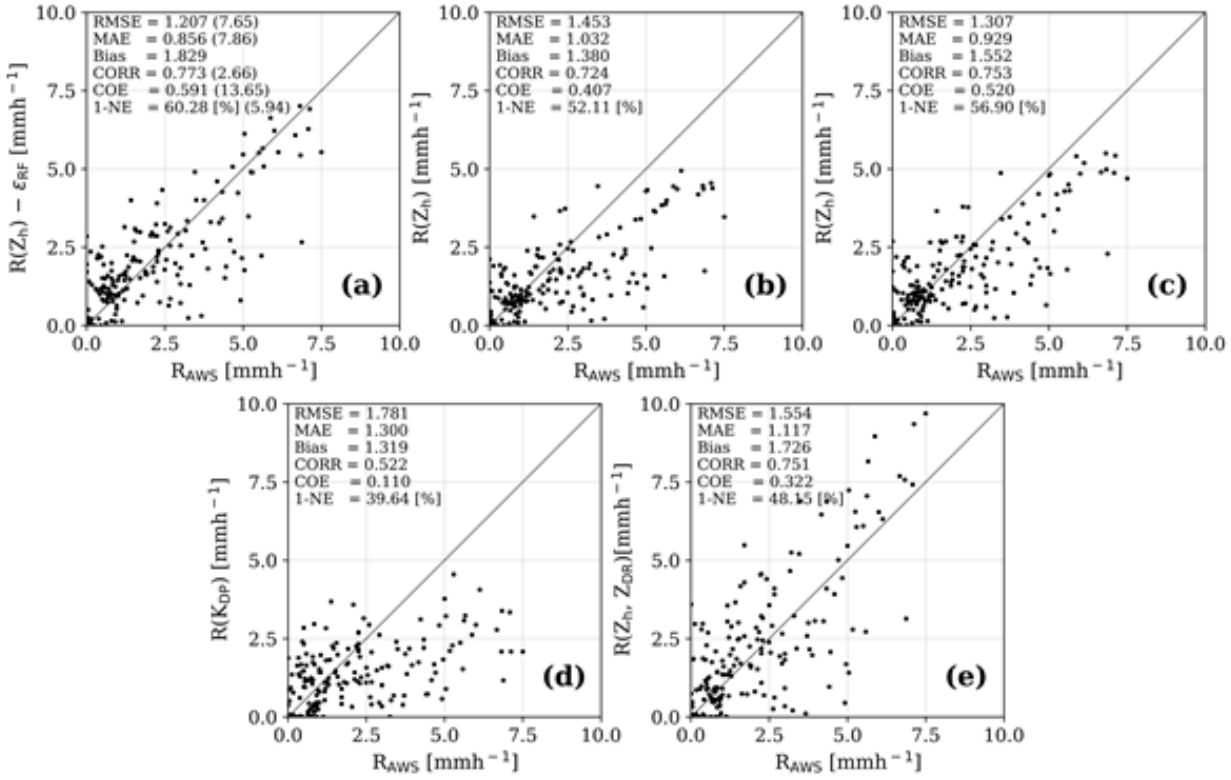

Figure 11. Scatter plots of (**a**) R(Z_h)− $\boldsymbol{\varepsilon_{RF}}$, (**b**) R($Z_h$), (**c**) adjusted R($Z_h$), (**d**) R($K_{DP}$), and (**e**) R($Z_h$, Z_{DR}) for Case 1 (August 14, 2017). R_{AWS} is the rainfall rate from the ground rain gauge. Values in the parentheses (in (a)) represent the improvement percentages from the best performance of the empirical relationship.

In Case 2, a similar trend as seen in Case 1 appears. R(Z_h)− ε_{RF} also outperforms the empirical relationships in all statistics. The underestimations of R(Z_h) that appear in R_{AWS} stronger than 15 mm h^{-1} (Figure 12b,c) are corrected by ε_{RF} (Figure 12a), leading to the decreases of RMSE from 2.59 and 1.89 to 1.74, respectively. Analogous to Case 1, Figure 12d presents an overall underestimation, and the RMSE value also shows the largest value. R(Z_h, Z_{DR}) overestimated the strong rainfall rate (R_{AWS} > 25 mm h^{-1}) (Figure 12e).

A total of six rainfall events from 2017 to 2018 were used to verify the five different models (Figure 13 and Table 9). The scatter plots of the event averaged rainfall rate are displayed, and the same color indicates the same event (Figure 6). The statistics are shown for rainfall types and models (Table 1). In general, the ML model (R(Z_h)− $\boldsymbol{\varepsilon_{RF}}$) outperformed all the empirical relationships, with an RMSE value of 1.039, MAE value of 0.593, CORR value of 0.959, COE value of 0.912, and 1-NE value of 75.24%. The estimation of the ML model showed the most consistency with the one-to-one line.

On the other hand, R(K_{DP}) tended to underestimate (Figure 13d), and R(Z_h, Z_{DR}) overestimated in weak rain with R_{AWS} < 17.5 mm h^{-1}. According to the rainfall types, R(Z_h)− $\boldsymbol{\varepsilon_{RF}}$ had a higher CORR (0.956), COE (0.905), and 1-NE (77.18%) in the stratiform rain, whereas it had a lower RMSE (0.612), MAE (0.330), and bias (1.041) in the convective rain. The 1-NE of the stratiform (convective) rain varied from 60% (10%) to 77% (56%). The poor performance of the convective rain was due to the smaller size of rain cells. Most of the convective rain showed smaller areas of precipitation and short-lived cells.

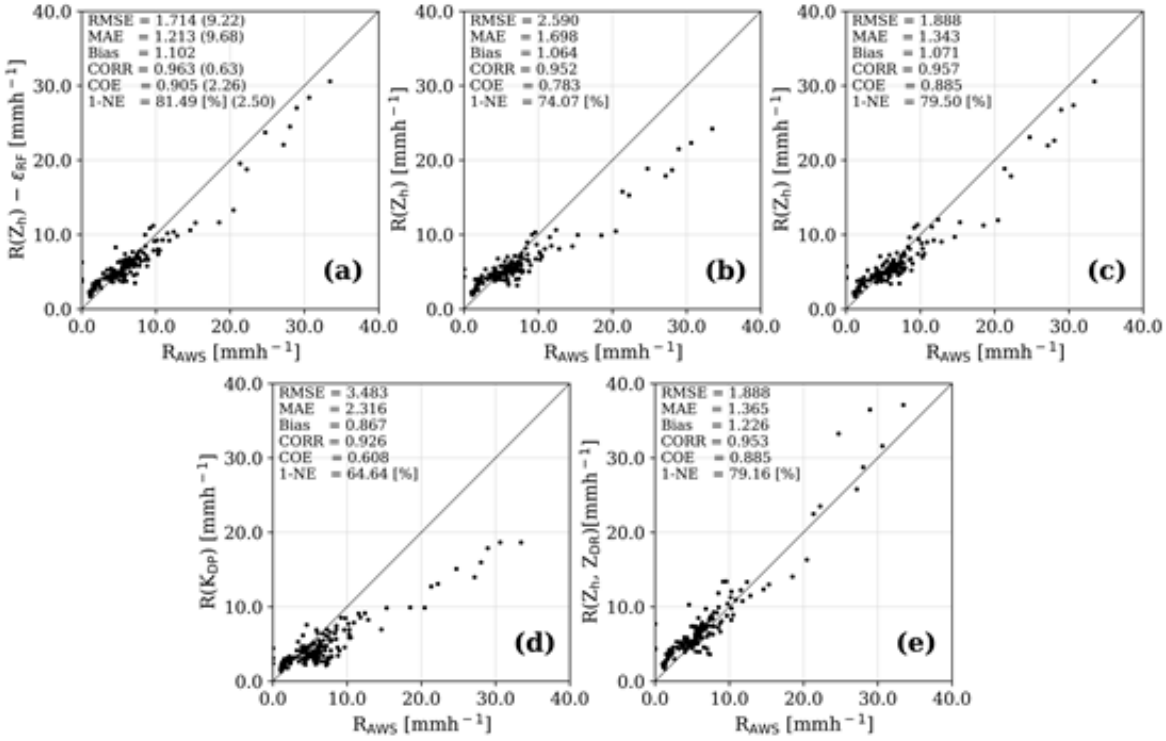

Figure 12. Scatter plots of (**a**) $R(Z_h) - \boldsymbol{\varepsilon_{RF}}$, (**b**) $R(Z_h)$, (c) adjusted $R(Z_h)$, (**d**) $R(K_{DP})$, and (**e**) $R(Z_h, Z_{DR})$ for Case 2 (September 11, 2017). Values in parentheses (in (a)) represent the improvement percentage from the best performance of the empirical relationship.

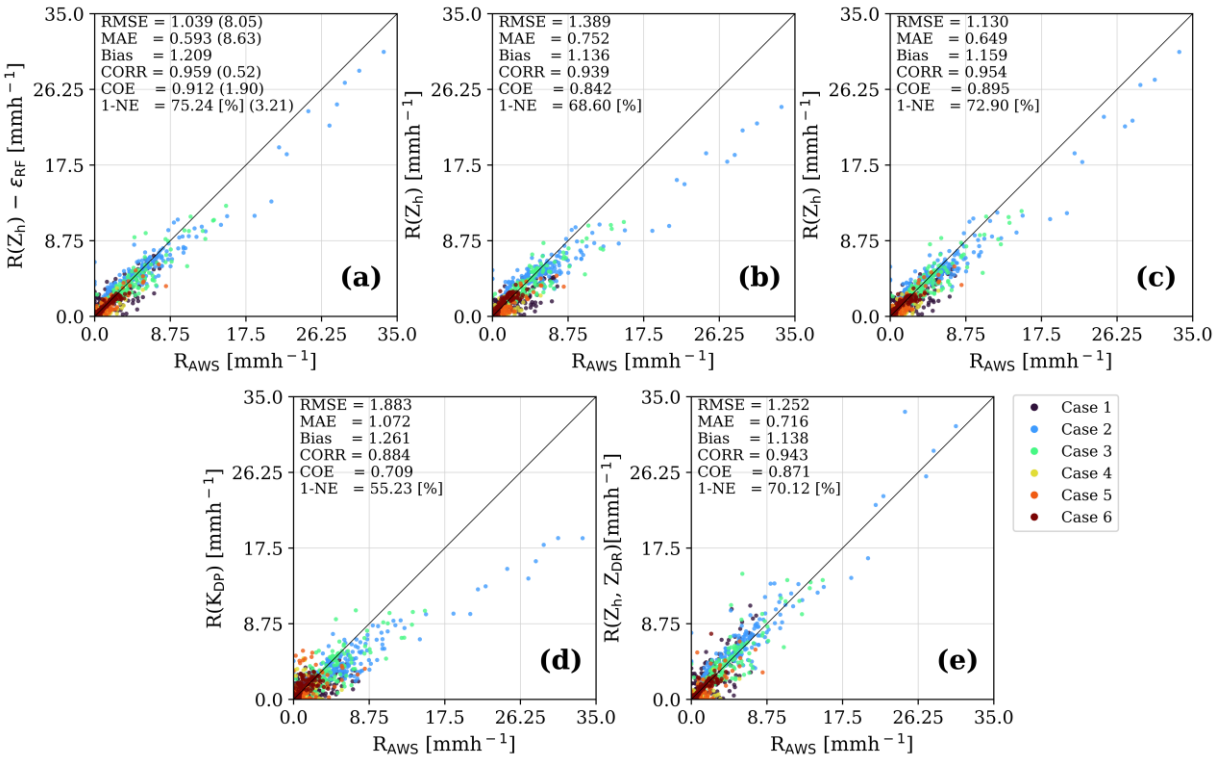

Figure 13. Scatter plots of (**a**) $R(Z_h) - \boldsymbol{\varepsilon_{RF}}$, (**b**) $R(Z_h)$, (**c**) adjusted $R(Z_h)$, (**d**) $R(K_{DP})$, and (**e**) $R(Z_h, Z_{DR})$, with different cases represented by different colors. Values in parentheses (in (a)) represent the improvement percentage from the best performance of the empirical relationship.

Table 9. Accuracy of the rainfall estimation with the different models for stratiform, convective, and all cases. The highest values of statistics are highlighted in bold, and the values in parentheses represent the improvement percentage from the best performance of the empirical relationship. Root mean square error (RMSE), mean absolute error (MAE), bias, correlation coefficient (CORR), coefficient of efficiency (COE) [41], and normalized error (1-NE).

Type	Method	RMSE	MAE	Bias	CORR	COE	1-NE [%]
Stratiform	$R(Z_h) - \boldsymbol{\varepsilon_{RF}}$	**1.237 (8.37)**	**0.770 (8.66)**	1.294	**0.956 (0.52)**	**0.905 (2.03)**	**77.18 (2.88)**
	$R(Z_h)$	1.676	0.843	1.190	0.936	0.825	70.79
	Adjusted $R(Z_h)$	1.350	0.985	1.210	0.951	0.887	75.02
	$R(K_{DP})$	2.217	1.346	**1.129**	0.896	0.695	60.09
	$R(Z_h, Z_{DR})$	1.469	0.916	1.307	0.941	0.866	72.83
Convective	$R(Z_h) - \boldsymbol{\varepsilon_{RF}}$	**0.612 (3.92)**	**0.330 (5.98)**	1.041	**0.873 (2.22)**	**0.731 (3.10)**	**55.79 (5.36)**
	$R(Z_h)$	0.716	0.382	0.971	0.833	0.632	48.82
	Adjusted $R(Z_h)$	0.637	0.351	**1.040**	0.854	0.709	52.95
	$R(K_{DP})$	1.204	0.671	1.539	0.463	-0.040	10.07
	$R(Z_h, Z_{DR})$	0.803	0.418	0.762	0.763	0.537	40.98
Total	$R(Z_h) - \boldsymbol{\varepsilon_{RF}}$	**1.039 (8.05)**	**0.593 (8.63)**	1.209	**0.959 (0.52)**	**0.912 (1.90)**	**75.24 (3.21)**
	$R(Z_h)$	1.389	0.752	**1.136**	0.939	0.842	68.60
	Adjusted $R(Z_h)$	1.130	0.649	1.159	0.954	0.895	72.90
	$R(K_{DP})$	1.883	1.072	1.261	0.884	0.709	55.23
	$R(Z_h, Z_{DR})$	1.252	0.716	1.138	0.943	0.871	70.12

5. Conclusions

The ML-based rainfall estimation using dual-polarization radar variables was explored with simulated and observed variables. The ML methods, RT and RF, used in this study allowed us to model the nonlinear relationship between the dependent and the independent variables and to identify important independent variables. The ML methods were first trained with the DSDs observed from 2DVD. In this study, we also considered three types of dependent variables (R: M1, the residual $\varepsilon = R(Z_h) - R_{2DVD}$: M2, and the normalized residual $\bar{\varepsilon} = \varepsilon/\overline{R(Z_h)}$: M3), two groups of independent variables (with K_{DP}: KY and without K_{DP}: KN), and two types of training data (categorized with intervals of Z_h: CY, and overall data without categorization: CN).

In the CY models of RF, the number of RTs and independent variables used was optimized. As a result of ML using DSDs from 2DVD, the K_{DP} was identified as the most important variable for rainfall estimation in both the M1KY-RT and M1KY-RF, while Z_h served as the most significant variable in the M1KN. This is an outcome of the fact that the K_{DP} can be approximated with the closest moment of DSDs to R, that is 4.2–5.6th moment and the Z_h with the sixth moment of DSDs [12,14]. In M2 and M3, Z_{DR} was the most crucial to explain the error (or residual) occurring in the R–Z relationships.

The ML methods were compared with the empirical relationships through 10-fold cross-validation. In the cross-validation, $R(Z_h)$, KN, and KY were first determined with the threshold values of K_{DP} and Z_{DR} due to the noises of K_{DP} and Z_{DR} for light rain (Figure

4), and the other models were then subsequently applied. Since the estimation in RT took the average of the terminal nodes, discontinuity in the estimation and underestimation (overestimation) at the strong (weak) rainfall rate within the node was often shown; however, residual ε corrects these problems. Similarly, CY is also improved with a lower value of RMSE. Since RF takes an ensemble of RT, the discontinuity problem in RT disappears; however, underestimation was still found above 100 mm h^{-1} in the M1CN-RF. Adjusting the rainfall rate with the estimated ε resolved the underestimation issue, and the estimation was close to the true value.

Compared to the empirical relationships, all the ML models showed improved evaluation statistics compared with the R–Z relationships. Even the worst ML model (M1CN-RT) showed meaningful improvement of the RMSE value (1.700) compared with the R–Z relationships (RMSE of 2.125 and 3.009). The M2CN-RF outperformed all the empirical relationships and presented a higher CORR value and lower RMSE value. The M2CN-RF, the model with the best performance among the ML models trained with 2DVD data, was applied to the MYN S-band dual-polarization radar data.

In the stratiform case (Case 1), most of the ε values were zero in the weak rainfall rate region, while the ε values were positive in the region of larger Z_{DR}. The negative ε values were large in the convection region, in particular, in the region of larger K_{DP} values in CASE 2. In addition, when estimated by the R–Z relationship, a significant underestimation was shown in heavier rainfall regions and was corrected by ε that was estimated by the ML models. In addition, the significant spatial structure of ε appeared and was highly correlated with Z_{DR} positively, and K_{DP} negatively. The evaluation with six rainfall events indicated that the ML model outperformed the empirical relationships regardless of the rainfall type (i.e., stratiform or convective). The statistics according to the rainfall type show that the ML-based QPE for stratiform cases had a higher CORR, COE, and 1-NE compared with the convective cases.

There was a dependency of the estimation accuracy on the trained data set. When we trained the RF model with DSD data from the DAE that was nearest to the MYN radar, the higher rainfall rate was systematically underestimated, likely due to the limited range of rainfall intensity. A recent study showed a significant discrepancy of DSD characteristics due to different climate and main forcings [29]. The addition of the OKL data significantly extended the rainfall prediction range, particularly in higher ranges, and resolved the underestimation of higher rainfall. As a result, we added additional DSDs data from OKL, BOS, and JIN in the training data set to broaden the variability of DSDs. This improved the accuracy in the overall statistics.

Through this study, we investigated the potential application of rainfall estimation based on ML using polarimetric radar data. We envision that more accurate rainfall estimation can be achieved by applying the RF model considering ε trained with 2DVD data to a large number of operational radars. In particular, the ML model improved the estimates in the heavy rain region, which were underestimated in the empirical relationship. This approach would be useful in the analysis and forecasting of severe weather. Future studies could include extending ML models to various radars and rainfall cases in different weather conditions.

Author Contributions: This work was made possible by significant contribution from all authors. Conceptualization, G.L. and K.S.; methodology, K.S., J.J.S., W.B., and G.L.; software, K.S. and J.J.S.; validation, K.S. and G.L.; formal analysis, K.S., W.B., and G.L.; investigation, K.S. and G.L.; resources, K.S. and W.B.; writing—original draft preparation, K.S. and W.B.; writing—review and editing, J.J.S. and G.L.; visualization, K.S.; supervision, G.L.; funding acquisition, G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Korea Meteorological Administration Research and Development Program under Grant KMI2020-00910 and by the Korea Environmental Industry & Technology Institute (KEITI) of the Korea Ministry of Environment (MOE) as "Advanced Water Management Research Program" (79615).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: This paper is based on Kyuhee Shin's thesis. We thank the Korea Meteorological Administration for acquiring the radar and AWS data and to Dr. Alexander V. Ryzhkov and Dr. Terry Schuur from National Severe Storm Laboratory for providing their valuable 2DVD data. We also greatly appreciate students and researchers in CARE, KNU for constructive discussions.

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

References

1. Ryzhkov, A.V.; Zrnić, D.S. Comparison of Dual-Polarization Radar Estimators of Rain. *J. Atmos. Ocean. Technol.* **1995**, *12*, 249–256.
2. Brandes, E.A.; Zhang, G.; Vivekanandan, J. Experiments in Rainfall Estimation with a Polarimetric Radar in a Subtropical Environment. *J. Appl. Meteorol.* **2002**, *41*, 674–685, doi:10.1175/1520-0450(2002)041<0674:EIREWA>2.0.CO;2.
3. Matrosov, S.Y.; Cifelli, R.; Kennedy, P.C.; Nesbitt, S.W.; Rutledge, S.A.; Bringi, V.N.; Martner, B.E. A comparative study of rainfall retrievals based on specific differential phase shifts at X- and S-band radar frequencies. *J. Atmos. Ocean. Technol.* **2006**, *23*, 952–963, doi:10.1175/JTECH1887.1.
4. Bringi, V.N.; Rico-Ramirez, M.A.; Thurai, M. Rainfall Estimation with an Operational Polarimetric C-Band Radar in the United Kingdom: Comparison with a Gauge Network and Error Analysis. *J. Hydrometeorol.* **2011**, *12*, 935–954, doi:10.1175/JHM-D-10-05013.1.
5. Thompson, E.J.; Rutledge, S.A.; Dolan, B.; Thurai, M.; Chandrasekar, V. Dual-polarization radar rainfall estimation over tropical oceans. *J. Appl. Meteorol. Climatol.* **2018**, *57*, 755–775, doi:10.1175/JAMC-D-17-0160.1.
6. Bringi, V.N.; Chandrasekar, V. *Polarimetric Doppler Weather Radar: Principles and Applications*; Cambridge University Press: Cambridge, UK, 2001; ISBN 9780521623841.
7. Zrnić, D.S.; Ryzhkov, A. Advantages of Rain Measurements Using Specific Differential Phase. *J. Atmos. Ocean. Technol.* **1996**, *13*, 454–464, doi:10.1175/1520-0426(1996)013<0454:AORMUS>2.0.CO;2.
8. Friedrich, K.; Germann, U.; Gourley, J.J.; Tabary, P. Effects or radar beam shielding on rainfall estimation for the polarimetric C-band radar. *J. Atmos. Ocean. Technol.* **2007**, *24*, 1839–1859, doi:10.1175/JTECH2085.1.
9. Kumjian, M.R. Principles and applications of dual-polarization weather radar. Part II: Warm and cold season applications. *J. Oper. Meteorol.* **2013**, *1*, 243–264, doi:10.15191/nwajom.2013.0120.
10. Marshall, J.S.; Palmer, W.M.K. The distribution of raindrops with size. *J. Meteorol.* **1948**, *5*, 165–166, doi:10.1175/1520-0469(1948)005<0165:TDORWS>2.0.CO;2.
11. Maki, M.; Park, S.G.; Bringi, V.N. Effect of natural variations in rain drop size distributions on rain rate estimators of 3 cm wavelength polarimetric radar. *J. Meteorol. Soc. Jpn.* **2005**, *83*, 871–893, doi:10.2151/jmsj.83.871.
12. Lee, G. Sources of errors in rainfall measurements by polarimetric radar: Variability of drop size distributions, observational noise, and variation of relationships between R and polarimetric parameters. *J. Atmos. Ocean. Technol.* **2006**, *23*, 1005–1028, doi:10.1175/JTECH1899.1.
13. Sachidananda, M.; Zrnić, D.S. Rain rate estimates from differential polarization measurements. *J. Atmos. Ocean. Technol.* **1987**, *4*, 588–598, doi:10.1175/1520-0426(1987)004<0588:RREFDP>2.0.CO;2.
14. Ryzhkov, A.V.; Giangrande, S.E.; Schuur, T.J. Rainfall Estimation with a Polarimetric Prototype of WSR-88D. *J. Appl. Meteorol.* **2005**, *44*, 502–515, doi:10.1175/jam2213.1.
15. Cifelli, R.; Chandrasekar, V.; Lim, S.; Kennedy, P.C.; Wang, Y.; Rutledge, S.A. A new dual-polarization radar rainfall algorithm: Application in Colorado precipitation events. *J. Atmos. Ocean. Technol.* **2011**, *28*, 352–364, doi:10.1175/2010JTECHA1488.1.
16. Seliga, T.A.; Bringi, V.N.; Al-Khatib, H.H. A Preliminary Study of Comparative Measurements of Rainfall Rate Using the Differential Reflectivity Radar Technique and a Raingage Network. *J. Appl. Meteorol.* **1981**, *20*, 1362–1368.
17. Ryzhkov, A.V.; Zrnic, D.S. *Radar Polarimetry for Weather Observations*; Springer Atmospheric Sciences; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; ISBN 9783030050931.
18. Teschl, R.; Randeu, W.L.; Teschl, F. Improving weather radar estimates of rainfall using feed-forward neural networks. *Neural Netw.* **2007**, *20*, 519–527, doi:10.1016/j.neunet.2007.04.005.
19. Kühnlein, M.; Appelhans, T.; Thies, B.; Nauss, T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI. *Remote Sens. Environ.* **2014**, *141*, 129–143, doi:10.1016/j.rse.2013.10.026.
20. Ouallouche, F.; Lazri, M.; Ameur, S. Improvement of rainfall estimation from MSG data using Random Forests classification and regression. *Atmos. Res.* **2018**, *211*, 62–72, doi:10.1016/j.atmosres.2018.05.001.
21. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; Routledge: New York, NY, USA, 1984; ISBN 9780412048418.
22. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
23. Kusiak, A.; Wei, X.; Verma, A.P.; Roz, E. Modeling and prediction of rainfall using radar reflectivity data: A data-mining approach. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2337–2342, doi:10.1109/TGRS.2012.2210429.

24. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716, doi:10.1016/j.rse.2020.111716.
25. Chiang, Y.M.; Chang, F.J.; Jou, B.J.D.; Lin, P.F. Dynamic ANN for precipitation estimation and forecasting from radar observations. *J. Hydrol.* **2007**, *334*, 250–261, doi:10.1016/j.jhydrol.2006.10.021.
26. Chen, H.; Chandrasekar, V.; Cifelli, R. A Deep Learning Approach to Dual-Polarization Radar Rainfall Estimation. In Proceedings of the 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC), New Delhi, India, 9–15 March 2019; pp. 1–2.
27. Kruger, A.; Krajewski, W.F. Two-Dimensional Video Disdrometer: A Description. *J. Atmos. Ocean. Technol.* **2002**, *19*, 602–617, doi:10.1175/1520-0426(2002)019<0602:TDVDAD>2.0.CO;2.
28. Atlas, D.; Srivastava, R.C.; Sekhon, R.S. Doppler radar characteristics of precipitation at vertical incidence. *Rev. Geophys.* **1973**, *11*, 1–35.
29. Bang, W.; Lee, G.; Ryzhkov, A.; Schuur, T.; Lim, K.-S.S. Comparison of microphysical characteristics between southern Korea and Oklahoma using two-dimensional video disdrometer data. *J. Hydrometeorol.* **2020**, 1–61, doi:10.1175/JHM-D-20-0087.1.
30. Thurai, M.; Gatlin, P.; Bringi, V.N.; Petersen, W.; Kennedy, P.; Notaroš, B.; Carey, L. Toward completing the raindrop size spectrum: Case studies involving 2D-video disdrometer, droplet spectrometer, and polarimetric radar measurements. *J. Appl. Meteorol. Climatol.* **2017**, *56*, 877–896, doi:10.1175/JAMC-D-16-0304.1.
31. Mishchenko, M.I.; Travis, L.D.; Mackowski, D.W. T-matrix computations of light scattering by nonspherical particles: A review. *J. Quant. Spectrosc. Radiat. Transf.* **1996**, *55*, 535–575, doi:10.1016/0022-4073(96)00002-7.
32. Thurai, M.; Huang, G.J.; Bringi, V.N.; Randeu, W.L.; Schönhuber, M. Drop shapes, model comparisons, and calculations of polarimetric radar parameters in rain. *J. Atmos. Ocean. Technol.* **2007**, *24*, 1019–1032, doi:10.1175/JTECH2051.1.
33. Lee, G.; Zawadzki, I. Radar calibration by gage, disdrometer, and polarimetry: Theoretical limit caused by the variability of drop size distribution and application to fast scanning operational radar data. *J. Hydrol.* **2006**, *328*, 83–97, doi:10.1016/j.jhydrol.2005.11.046.
34. Kwon, S.; Lee, G.; Kim, G. Rainfall Estimation from an Operational S-Band Dual-Polarization Radar: Effect of Radar Calibration. *J. Meteorol. Soc. Jpn. Ser. II* **2015**, *93*, 65–79, doi:10.2151/jmsj.2015-005.
35. Liaw, A.; Wiener, M. Classification and regression by random forest. *R News* **2002**, *2*, 18–22.
36. Amemiya, Y. Generalization of the TLS approach in the errors-in-variables problem. In *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*; Van Huffel, S., Ed.; SIAM: Philadelphia, PA, USA, 1997; pp. 77–86.
37. Chandrasekar, V.; Bringi, V.N. Error Structure of Multiparameter Radar and Surface Measurements of Rainfall Part I: Differential Reflectivity. *J. Atmos. Ocean. Technol.* **1988**, *5*, 783–795.
38. Chandrasekar, V.; Bringi, V.N.; Balakrishnan, N.; Zrnić, D.S. Error Structure of Multiparameter Radar and Surface Measurements of Rainfall. Part III: Specific Differential Phase. *J. Atmos. Ocean. Technol.* **1990**, *7*, 621–629, doi:10.1175/1520-0426(1990)007<0621:ESOMRA>2.0.CO;2.
39. Chandrasekar, V.; Gorgucci, E.; Scarchilli, G. Optimization of Multiparameter Radar Estimates of Rainfall. *J. Appl. Meteorol.* **1993**, *32*, 1288–1293, doi:10.1175/1520-0450(1993)032<1288:OOMREO>2.0.CO;2.
40. Silvestro, F.; Rebora, N.; Ferraris, L. An algorithm for real-time rainfall rate estimation by using polarimetric radar: RIME. *J. Hydrometeorol.* **2009**, *10*, 227–240, doi:10.1175/2008JHM1015.1.
41. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290, doi:10.1016/0022-1694(70)90255-6.
42. Kwon, S.; Jung, S.-H.; Lee, G. Inter-comparison of radar rainfall rate using Constant Altitude Plan Position Indicator and hybrid surface rainfall maps. *J. Hydrol.* **2015**, *531*, 234–247, doi:10.1016/j.jhydrol.2015.08.063.
43. Ye, B.Y.; Lee, G.W.; Park, H.M. Identification and removal of non-meteorological echoes in dual-polarization radar data based on a fuzzy logic algorithm. *Adv. Atmos. Sci.* **2015**, *32*, 1217–1230, doi:10.1007/s00376-015-4092-0.