remote sensing

MDPI

Article

Estimation of Precipitation Area Using S-Band Dual-Polarization Radar Measurements

Joon Jin Song [1], Melissa Innerst [1,2], Kyuhee Shin [3], Bo-Young Ye [4], Minho Kim [1], Daejin Yeom [3] and GyuWon Lee [3,*]

[1] Department of Statistical Science, Baylor University, Waco, TX 76798, USA; Joon_Song@baylor.edu (J.J.S.); innerst@juniata.edu (M.I.); Minho_Kim1@baylor.edu (M.K.)
[2] Department of Mathematics, Juniata College, Huntingdon, PA 16652, USA
[3] Center for Atmospheric Remote Sensing (CARE), Department of Astronomy and Atmospheric Sciences, Kyungpook National University, Daegu 41566, Korea; kyuhee@knu.ac.kr (K.S.); djyeom@knu.ac.kr (D.Y.)
[4] Institute of Environmental Studies, Pusan National University, Busan 46241, Korea; boyoungye@pusan.ac.kr
* Correspondence: gyuwon@knu.ac.kr; Tel.: +82-53-950-6361

Abstract: Estimating precipitation area is important for weather forecasting as well as real-time application. This paper aims to develop an analytical framework for efficient precipitation area estimation using S-band dual-polarization radar measurements. Several types of factors, such as types of sensors, thresholds, and models, are considered and compared to form a data set. After building the appropriate data set, this paper yields a rigorous comparison of classification methods in statistical (logistic regression and linear discriminant analysis) and machine learning (decision tree, support vector machine, and random forest). To achieve better performance, spatial classification is considered by incorporating latitude and longitude of observation location into classification, compared with non-spatial classification. The data used in this study were collected by rain detector and present weather sensor in a network of automated weather systems (AWS), and an S-band dual-polarimetric weather radar during ten different rainfall events of varying lengths. The mean squared prediction error (MSPE) from leave-one-out cross validation (LOOCV) is computed to assess the performance of the methods. Of the methods, the decision tree and random forest methods result in the lowest MSPE, and spatial classification outperforms non-spatial classification. Particularly, machine-learning-based spatial classification methods accurately estimate the precipitation area in the northern areas of the study region.

Keywords: classification; machine learning; statistical learning; precipitation area estimation; spatial classification

check for updates

Citation: Song, J.J.; Innerst, M.; Shin, K.; Ye, B.-Y.; Kim, M.; Yeom, D.; Lee, G. Estimation of Precipitation Area Using S-Band Dual-Polarization Radar Measurements. *Remote Sens.* **2021**, *13*, 2039. https://doi.org/10.3390/rs13112039

Academic Editor: Simone Lolli

Received: 25 March 2021
Accepted: 18 May 2021
Published: 21 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Precipitation plays a major role in the water cycle, which has direct impacts on human life. Accurate rainfall estimation is therefore essential to reduce damage from natural disasters such as drought, flood, and typhoon. Specially, identification of precipitating regions is of interest to many fields, such as agriculture, hydrology, and atmospheric science. A variety of equipment and techniques have been used to accurately estimate precipitation variability over time and space.

Ground-based equipment gauge has been widely used as a reference due to its accurate measurement of precipitation [1]. The gauge directly measures the amount of precipitation in limited monitoring sites, and spatial interpolation is often employed to predict precipitation at unobserved locations [2–7]. It is also used to compare with other remote sensing equipment as the ground truth value [8–12]. However, the accuracy of precipitation amount and detection is limited by the quantitative resolution of gages such as the size of bucket. Precipitation detector and visibility weather sensor from automatic weather stations (AWS) are additionally used to overcome this limitation in detecting the

existence of precipitation [13,14]. These instruments can yield accurate point measurements of precipitation occurrence but are limited into spatial coverage such as point measurements at the AWS site, thus resulting no observation over oceanic and mountainous regions, and suffer from observational errors [15,16].

To overcome these limitations, a weather radar is commonly used as a remote sensing instrument. The radar is superior in space-time resolution and spatial coverage is used for understanding spatial distribution of precipitation. To accurately estimate precipitation, it is essential to remove the contamination due to ground clutter and beam blockage over complex terrain. Many techniques to classify non-meteorological echoes have been developed using statistical methods such as fuzzy logic classifier, Bayesian approach, and neural network [17–22]. The contamination from beam blockage has also been corrected using a vertical profile of radar parameter and a beam blockage fraction estimated by radar beam propagation of standard atmosphere and digital elevation model (DEM) simulation of the beam blocked fraction [23,24]. A hybrid scan method consisting of the lowest radar bins was developed using multiple elevation angles to avoid the non-metrological echoes and beam blockage [25–27]. Furthermore, the spatial variabilities of dual-polarization parameters are useful features to identify non-meteorological echoes [22,28]. Ryzhkov and Zrnic [29] also showed that dual-polarimetric variables such as the correlation coefficient and the different phase are good indicators of non-meteorological echoes. Kwon et al. [27] developed Hybrid Surface Rainfall (HSR), which can improve rainfall estimation by using dual-polarization parameters and the hybrid scan method by reflecting current atmospheric conditions. However, HSR still shows high uncertainty at a long distance due to beam broadening and subsequent contamination by bright band. In addition, the uncertainty of precipitation estimation is increased due to the effects of beam height and wind drift [30,31]. According to Yan and Bárdossy [32], the correction of displacement between beam height and surface precipitation by using combined data of radar and ground measurement improves the precipitation estimation. The surface precipitation area identified in this study can thus further improve the QPE by correctly eliminating no precipitation area while retaining precipitation area. The complete description of precipitation area can be widely applied in various fields, such as weather forecast [33], hydrological and land surface modeling [32], development of precipitation sensor, and plant water requirement [34].

The objective of this paper is to develop an optimal framework for mapping precipitation areas using dual-polarization radars. The main contribution of this paper is twofold. First, this paper presents a novel data framework for improving the precipitation area estimation. Two types of sensors, rain detector (AWS_RE) and present weather sensor (VIS_WW), are considered. To generate the binary variable by accumulating one-minute data at ten-minute average data, two thresholds are employed. Lastly, two models, non-spatial and spatial models, are considered in the classification. The spatial model includes the longitude and latitude of station location in the set of covariates. Second, this paper provides a rigorous comparison of the classification methods for precipitation area estimation. There are numerous statistical and machine learning methods for classification problems. Among them, we consider five classification methods to estimate precipitation areas based on the accumulated data and spatial characteristics. Two statistical methods (logistic regression and linear discriminant analysis) and three machine learning methods (support vector machine, decision tree, and random forest) are used in this paper. For each method, as addressed earlier, two classification models are employed, non-spatial classification and spatial classification. The former considers only radar measurements as predictors, while the latter incorporates spatial information into classification, called spatial classification. It is expected that we can improve the accuracy of precipitation area estimation using spatial information. Additionally, we study and compare the events classified into two types of storms (stratiform and convective).

2. Methodology

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$ be a set of p covariates for each binary response y_i, $i = 1, \cdots, n$. In this paper, we consider two statistical and three machine learning classification methods for precipitation area estimation. To assess the performance of the classifiers, the mean-squared prediction error (MSPE) for each method is calculated using leave-one-out cross validation (LOOCV).

2.1. Logistic Regression

In linear regression, the response Y can be modeled directly with covariates due to its continuity. However, when the response is binary, it is more appropriate to model the probability of a response being a particular outcome, $p = P(Y = 1|\boldsymbol{x})$. A common approach to modeling p is the logistic regression (LR) model,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \tag{1}$$

where β_j is the regression coefficient of x_j. The left-hand side of the equation above is known as the log-odds or the logit of probability (p). Note that p is not linear in terms of $\boldsymbol{x}$, while the logit is. The model can be rewritten as

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{2}$$

which is an S-shaped curve with predictions constrained between 0 and 1. To estimate the model parameters, $\beta_1, \cdots, \beta_p$, the maximum likelihood method is commonly used. Once we obtain the estimates, making predictions $\hat{p}$ is simply achieved by substituting the values of the estimates for the regression parameters.

2.2. Linear Discriminant Analysis

As its name implies, linear discriminant analysis (LDA) involves modeling a response variable as a linear combination of the covariates. Welch [35] and Fisher [36] were both credited for developing LDA, and the latter is used in this paper.

Fisher's approach to the problem of finding an optimal classification rule is to find the linear combination of the covariates, which maximizes the between-group variance while also minimizing the within-group variance. Let $\mathbf{B}$ be the between-group covariance matrix and $\mathbf{W}$ be the within-group covariance matrix. Then, we seek to find b such that

$$\frac{b'\mathbf{B}b}{b'\mathbf{W}b} \tag{3}$$

is maximized. The solution, or linear discriminant, is the eigenvector corresponding to the largest eigenvalue of $\mathbf{W}^{-1}\,\mathbf{B}$. This ratio of the between-group variance to the within-group variance can be thought of as a signal-to-noise ratio. In other words, Fisher's approach involves maximizing the signal-to-noise ratio. When the number of covariates is large, the calculations can become complicated. Unlike Welch's approach, the method does not make any assumptions about the underlying distributions of the data. However, it is necessary to assume that the covariance matrices are identical for all groups.

2.3. Support Vector Machine

Support vector machine (SVM) is a powerful tool for solving classification and regression problems in machine learning. If data are completely separable into two classes, there are infinitely many linear boundaries accurately separating into the two classes, and it is challenging to choose the optimal boundary based solely on accuracy. Vapnik [37] defined an alternate metric, the margin, which is the smallest distance from the observations to a

hyperplane separating them. The maximal margin hyperplane is farthest from the training observations. The decision function can be written as

$$D(\mathrm{u}) \;=\; \beta_0 + \sum_{j=1}^{P} \beta_j u_j \;=\; \beta_0 + \sum_{i=1}^{n} y_i \alpha_i \mathrm{x}_i' \mathrm{u} \tag{4}$$

where y_i is the sign of the class ($y_i = 1$ for Group 1 and $y_i = -1$ for Group 2), α_i is the model parameter, and $\mathrm{x}_i'\mathrm{u}$ is the dot product between the observed data and the data points that fall on the margin boundary. In the completely separable case mentioned previously, the set of nonzero α parameter corresponds to the data points that fall on the boundary of the margin, known as support vectors. The classifier is supported solely by these data points. If $D(\mathrm{u}) > 0$, the new sample is classified into Group 1, whereas, if $D(\mathrm{u}) < 0$, the new sample is classified into Group 2.

2.4. Decision Tree

Decision tree (DT) methods involve segmenting the covariate space into a number of simple and distinct regions. To split the covariate space into the regions, recursive binary splitting is commonly used, starting at the top of the tree and successively splitting the covariate space. Each split is indicated via two new branches further down the tree.

In order to perform this splitting, all covariates are considered and one is chosen with the lowest impurity measure. There are several impurity measures available for DT. For instance, the Gini index for the two-class case is defined by

$$G \;=\; \hat{p}_{m1}(1 - \hat{p}_{m1}) + \hat{p}_{m2}(1 - \hat{p}_{m2}) \tag{5}$$

where $\hat{p}_{mk}$ is the probability of the k^{th} class in the m^{th} region. This index measures how often a randomly selected element would be incorrectly identified. To predict a new observation, the region to which it belongs is identified, and it is assigned to the class with the largest class probability, $\hat{p}_{mk}$, in that region.

2.5. Random Forest

A single decision tree can produce highly variable estimates. To decrease this variance, an ensemble approach is a reasonable choice, combining the outcomes of several decision trees to a single classification. Random forests are widely used ensemble learning methods for classification and regression, combining predictions from multiple models to improve predictive performance [38]. These classifications can be thought of as votes, and the new observation is classified into the class with the most votes. An alternative approach is to use predicted probabilities. For a given class, the number of trees which classified the new observation into that class is divided by the total number of trees, which is considered the predicted probability for that class. The new observation is classified into the class with the highest probability. The variance reduction achieved by this method can be seen visually in the form of a smoother Receiver Operating Characteristic (ROC) curve, and numerically in the form of a lower Area Under the Curve (AUC).

In this method, a number of decision trees are built on bootstrapped training samples. The trees must be built in such a way that they are uncorrelated. Thus, at each split in a tree, a random sample of m covariates out of the full set of p covariates is chosen as split candidates. Then, one of the m covariates is chosen to split on. This prevents a very strong covariate from always being used for the first split in a tree, which leads to a unique set of trees. This leads to a lower variance than would be achieved using a single tree or bagged trees.

2.6. Spatial Classification

This paper focuses on the prediction of a binary response at unsampled locations. This prediction problem can be thought of a spatial classification problem. Utilizing spatial information in classification problems can enhance classification accuracy. To

this end, we include spatial information, longitude and latitude in this study, in the set of covariates, which is referred to as Model 2. Different spatial information could be incorporated if available. To evaluate the effectiveness of the spatial information, the non-spatial classification model (Model 1) with only non-spatial predictors is compared with the spatial classification model (Model 2). In this study, both models include the radar variables as non-spatial predictors, but longitude and latitude are used as spatial predictors in only Model 2 (spatial classification model).

2.7. Leave-One-Out Cross Validation

Leave-one-out cross validation (LOOCV) is conducted to assess the performance of the classification methods mentioned above. Cross validation methods involve splitting the set of observations into two parts, a training set and a test set. The model is trained on the observations in the training set, and predictions are made for the observations in the test set. Then, the mean squared prediction error (MSPE) is calculated by comparing the actual and predicted values for the observations in the test set. In LOOCV, the test set consists of only a single observation and the model is trained on the remaining $n - 1$ observations. This is repeated n times, where n is the number of observations, and each observation is left out one by one. The LOOCV estimate of the MSPE is given by

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (6)$$

where y_i is the actual value of the i^{th} observation in the test set, and $\hat{y}_i$ is the predicted value of the i^{th} observation in the test set. Note that the observations in the test set are classified into one of two groups. Thus, y_i and $\hat{y}_i$ can both only take on one of two values—either 0 or 1. This constrains the MSPE to values between 0 and 1. Moreover, the MSPE can be thought of as the percent of incorrectly classified observations in the test set. Therefore, lower MSPE values are preferable.

3. Data Description

Data from the Mountain Myeonbong S-band dual polarization radar and the Automatic Weather Station (AWS) are used in this study. This radar has a beam width of 0.92° and scans plan position indicators (PPI) at nine elevation angles (0°, 0.39°, 0.83°, 2°, 2.88°, 4.06°, 5.67°, 7.88°, 10.94°) for every 10 min (Table 1). The observed radar variables are reflectivity (Z_H), differential reflectivity (Z_{DR}), specific differential phase shift (K_{DP}), and cross-correlation coefficient (ρ_{HV}). The biases in Z_H and Z_{DR} were calibrated by post-processing. The averaged calibration bias of Z_H is −3.9 dBZ calculated by using the self-consistency constraint between Z_H and K_{DP} [11,39]. The averaged calibration bias of Z_{DR} is 0.03 dB obtained by comparison of observed data with the average Z_H–Z_{DR} relationship derived from data from disdrometer [11]. The HSR radar data are generated by using the lowest radar elevation angles which are not affected by ground echoes, beam blockage, and non-meteorological echoes [27]. The radar parameters in polar coordinates are converted to Cartesian coordinates with a horizontal resolution of 1 km × 1 km over a 240 km range. The pixel values with no echoes were replaced as Z_H = −32 dBZ, Z_{DR} = −8 dB, K_{DP} = 0° km^{-1}, ρ_{HV} = 0 [40,41].

Table 1. Characteristics of the S-band dual-polarization radar at Mt. Myeonbong (MYN).

Parameter	Value
Frequency (wavelength)	2,727 MHz (11 cm, S-band)
Location	36°10′45″N, 128°59′50″E
Height	1136 m
Beam Width	0.92°
Elevation angle	0°, 0.39°, 0.83°, 2°, 2.88°, 4.06°, 5.67°, 7.88°, 10.94°
Max range	285 km

The AWS data are collected from two devices, rain (precipitation) detector (AWS_RE) and present weather sensor, using optical sensors (VIS_WW). These data are quality controlled by the physical limit check, the climate range check, the step check, the persistence check, the internal consistency check, and the median filter [42]. The total of 596 AWSs are used over the area covered by radar, and time resolution is 1 min. These data are recorded as 1 if rain is detected, and 0 otherwise. The present weather sensor computed the visibility and present weather using measured optical extinction/scattering, temperature, humidity, and precipitation rate with observed precipitation particles. The outputs of present weather sensor are WMO (World Meteorological Organization) 4680 code (drizzle, rain, showery precipitation, etc.) from 0 to 99, which is converted to binary data. The binary variable is set to 1 if the present weather code is greater than 40, indicating precipitation, and 0 otherwise. These binary data (AWS_RE and VIS_WW) are used as response variables in this study. The rain detector is installed in every AWS, while the present weather sensor is available in some stations. The map of the stations is shown in Figure 1.

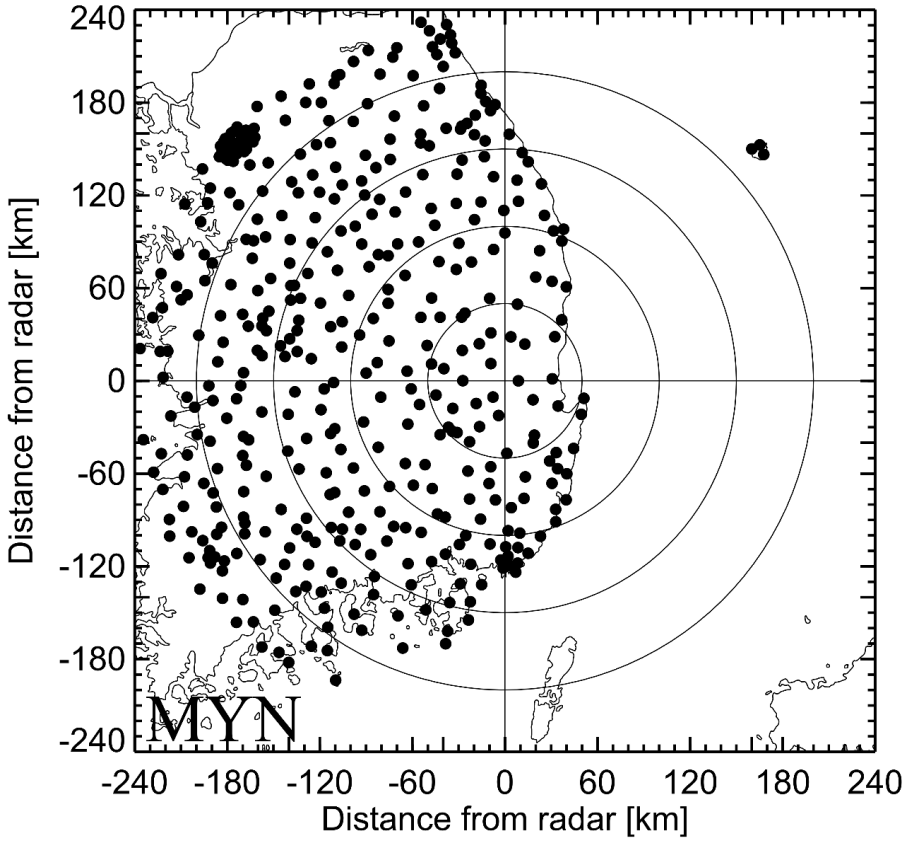


Figure 1. The locations of Automatic Weather Stations (dots) within the radar observation range of the MYN radar. Black circles denote radar range rings with a 50 km interval.

To match the time resolution between the radar data (every 10 min) and the AWS data (every minute), we consider two different ways to accumulate the AWS data using two thresholds. For the first threshold (Type 1 threshold), the station is classified as "wet" (that is, avg (AWS_RE) $\geq$ 0.5 or avg (VIS_WW) $\geq$ 0.5) if rain was observed for at least five out of ten minutes. The second threshold (Type 2 threshold) defines a station as "wet" if rain was observed for at least one out of ten minutes (that is, avg (AWS_RE) $\geq$ 0.1 or avg (VIS_WW) $\geq$ 0.1). The resulting pair of the radar data and the AWS data were produced by extracting the values of the nearest radar pixel to each AWS. The observations with missing values are eliminated. Table 2 presents the eight scenarios by considering two sensors (AWS_RE and VIS_WW), two thresholds (Type 1 and Type 2), and two models (Model 1: non-spatial and Model 2: spatial).

Table 2. The eight scenarios made by considering two sensors, two thresholds, and two models.

Scenarios	Type of Sensors	Threshold	Regression Model
S1	AWS	Type 1: Avg (AWS_RE) $\geq$ 0.5	Model 1: Non-spatial
S2	AWS	Type 2: Avg (AWS_RE) $\geq$ 0.1	Model 1: Non-spatial
S3	AWS	Type 1: Avg (AWS_RE) $\geq$ 0.5	Model 2: Spatial
S4	AWS	Type 2: Avg (AWS_RE) $\geq$ 0.1	Model 2: Spatial
S5	VIS	Type 1: Avg (VIS_WW) $\geq$ 0.5	Model 1: Non-spatial
S6	VIS	Type 2: Avg (VIS_WW) $\geq$ 0.1	Model 1: Non-spatial
S7	VIS	Type 1: Avg (VIS_WW) $\geq$ 0.5	Model 2: Spatial
S8	VIS	Type 2: Avg (VIV_WW) $\geq$ 0.1	Model 2: Spatial

Eleven rain events are selected from June to September in 2017 (Table 3). A total of 318 10-min data are used to estimate the precipitation area. The PPI images of Z_H and ρ_{HV} for two precipitation types (stratiform and convective) are shown in Figure 2. Z_H at the low elevation angle (0.0°) in the stratiform rain shows widespread precipitation area, and the rainfall intensity is quite weak (Figure 2a). Another typical characteristic of stratiform rainfall is the distinct melting layer. Low ρ_{HV} at the high elevation angle (4.06°) appears in a ring shape, indicating a melting layer (Figure 2b). In the convective case, Z_H at the low elevation angle is stronger (>35 dBZ) than that of the stratiform case. The echoes are aligned northeast to southwest (Figure 2c) and the melting layer is not shown in ρ_{HV} at the high elevation angle (Figure 2d).

Table 3. Characteristics of the rainfall event used in analysis.

No.	Date (MM/DD/YYYY)	Rainfall Duration [Hours]	The Number of 10-min Data	
			Stratiform	Convective
1	06/06/2017	5	30	0
2	07/03/2017	3	14	14
3	07/04/2017	3	18	0
4	07/14/2017	2	12	0
5	07/28/2017	2	12	0
6	07/29/2017	3	18	0
7	07/31/2017	7	42	0
8	08/09/2017	3	0	18
9	08/14/2017	12	72	0
10	08/15/2017	4	0	24
11	09/11/2017	9	54	0
Total number of data			262	56

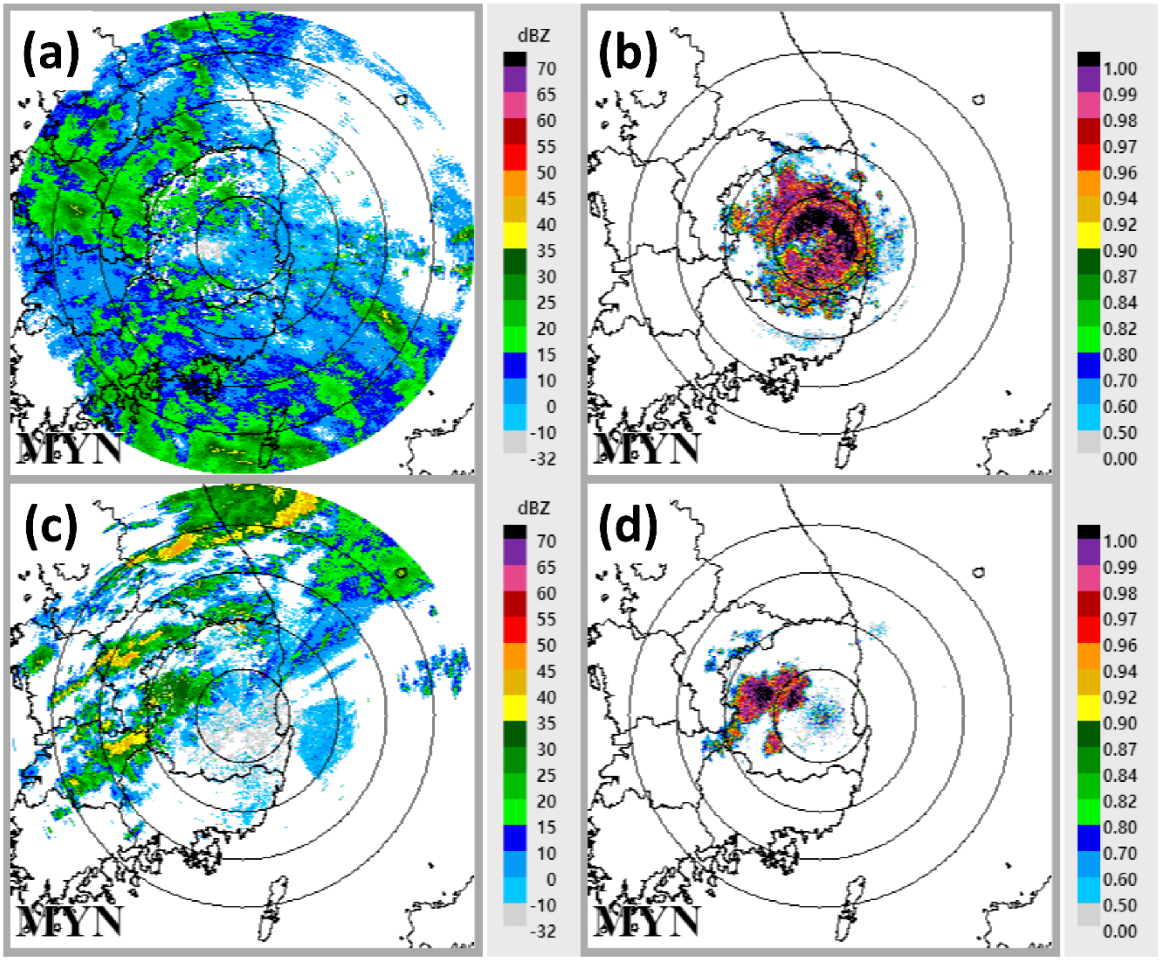


Figure 2. PPIs of (**a**) Z_H for the elevation angle of 0.0° and (**b**) ρ_{HV} for the elevation angle of 4.06° in the stratiform case. (**c**,**d**) are the same as (**a**,**b**) except for the convective case, respectively.

The precipitation type of each event is classified as stratiform and convective cases using the spatial autocorrelation of precipitation field. The spatial autocorrelation of precipitation is known to depend on precipitation types [43–46]. Thus, the spatial autocorrelation of rainfall rate can be used to quantitatively determine between convective and stratiform precipitation systems. Rainfall rate (R) estimated by Z_H ($Z_H = 200R^{1.6}$) from HSR data is analyzed to calculate spatial autocorrelation with the maximum horizontal length of 100 km. An ellipse fit of decorrelation length is given by e-folding distances which measure the spatial extent for the variations of precipitation field. In this study, the time steps in which with the length of the major axis of the ellipse is more than 130 km are considered the stratiform type, while others are classified as the convective type. An example of spatial autocorrelation for stratiform and convective type is shown in Figure 3. The length of the major axis of the ellipse is 143.50 km (107.96 km) in the stratiform (convective) case. The length of the minor axis (43.52 km) as well as the axis ratio (0.40) is much smaller in this convective case. The length of the major axis is almost 100 km. These are the characteristics of banded convective precipitation.

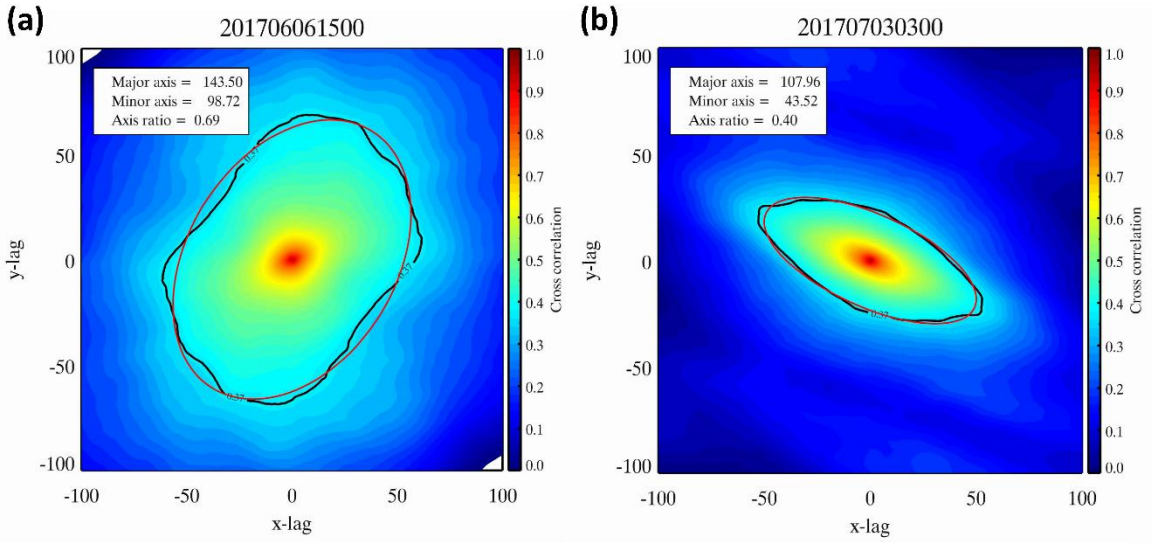


Figure 3. Spatial autocorrelation of two-dimensional rainfall rate of (**a**) the stratiform case (06/06/2017 1500 KST) and (**b**) the convective case (07/03/2017 0300 KST).

4. Results

First of all, we compare the eight scenarios described in Table 2 with statistical and machine learning classification methods in order to find an optimal data framework for precipitation area estimation, followed by an extensive comparison of classifiers. In this comparison, three factors are considered, sensor types (AWS and VIS), threshold types (0.1 and 0.5), and model choice (spatial and non-spatial). The medians of the MSPE values obtained from each of the eight scenarios are reported in Table 4. For the sensor types, it is found that the MSPE values of AWS are mostly smaller than those of VIS, which indicates that the rain detectors are preferable. It is interesting to note that the Type 1 threshold (0.5) offers the smaller median for AWS and VIS. To investigate the effectiveness of spatial information in precipitation area estimation, non-spatial (Model 1) and spatial (Model 2) models are compared under the same sensor and threshold. Notice that spatial model significantly outperforms non-spatial model for AWS and VIS. In addition, tree-based methods (DT and RF) in S3 provide the most precise estimation of precipitation area in terms of the MSPE. Figure 4 plots the distribution of the MSPE values for each scenario over all methods and events, showing the smallest variability in S3. Thus, in the remainder of this section, S3 is used to compare classification methods over different types of rain events.

Table 4. Medians of MSPE for the eight scenarios and different classifiers and average values of all different classifiers.

Scenarios	LR	LDA	SVM	DT	RF	Average
S1	0.2290	0.2312	0.2325	0.2386	0.2386	0.2330
S2	0.2713	0.2706	0.2704	0.2715	0.2728	0.2713
S3	0.2069	0.2100	0.1800	0.1681	0.1702	0.1870
S4	0.2249	0.2286	0.1908	0.1772	0.1870	0.2017
S5	0.2599	0.2588	0.2564	0.2614	0.2635	0.2600
S6	0.2901	0.2844	0.2874	0.2922	0.2891	0.2886
S7	0.2204	0.2196	0.1953	0.2014	0.1929	0.2059
S8	0.2227	0.2220	0.2000	0.2083	0.1953	0.2097

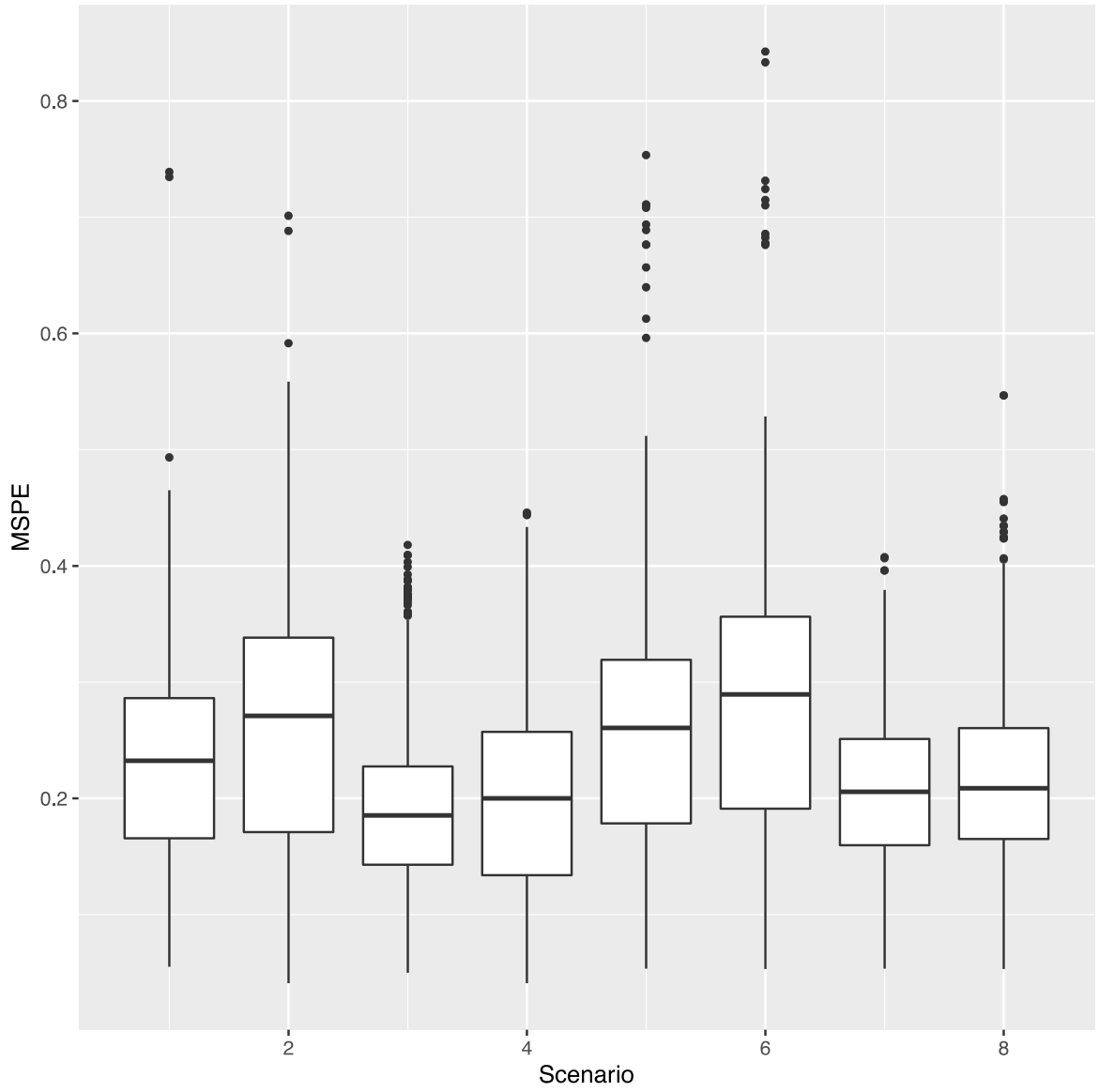


Figure 4. Boxplot of the MSPE for the eight scenarios. Each boxplot represents the distribution of the MSPE values over all methods and events. The black dots are outliers.

To illustrate precipitation area estimation for the entire area of interest, the case at 09:40 KST on 31 July 2017 is chosen, and the observed radar measurements are shown in Figure 5. The statistical and machine learning methods addressed in Section 2 are used. The estimated precipitation areas (no precipitation: blue, precipitation: red) are given in Figure 6. The true values from AWS are shown by triangles (no precipitation) and circles (precipitation). The maps yielded by the statistical methods present similar spatial patterns (two upper figures for LR and LDA). Substantial misclassification, estimating wet areas when the stations are observed as dry, has been found in the northern areas. However, the machine learning methods correctly estimate the areas as dry. RF produces a smoother classification boundary than DT.

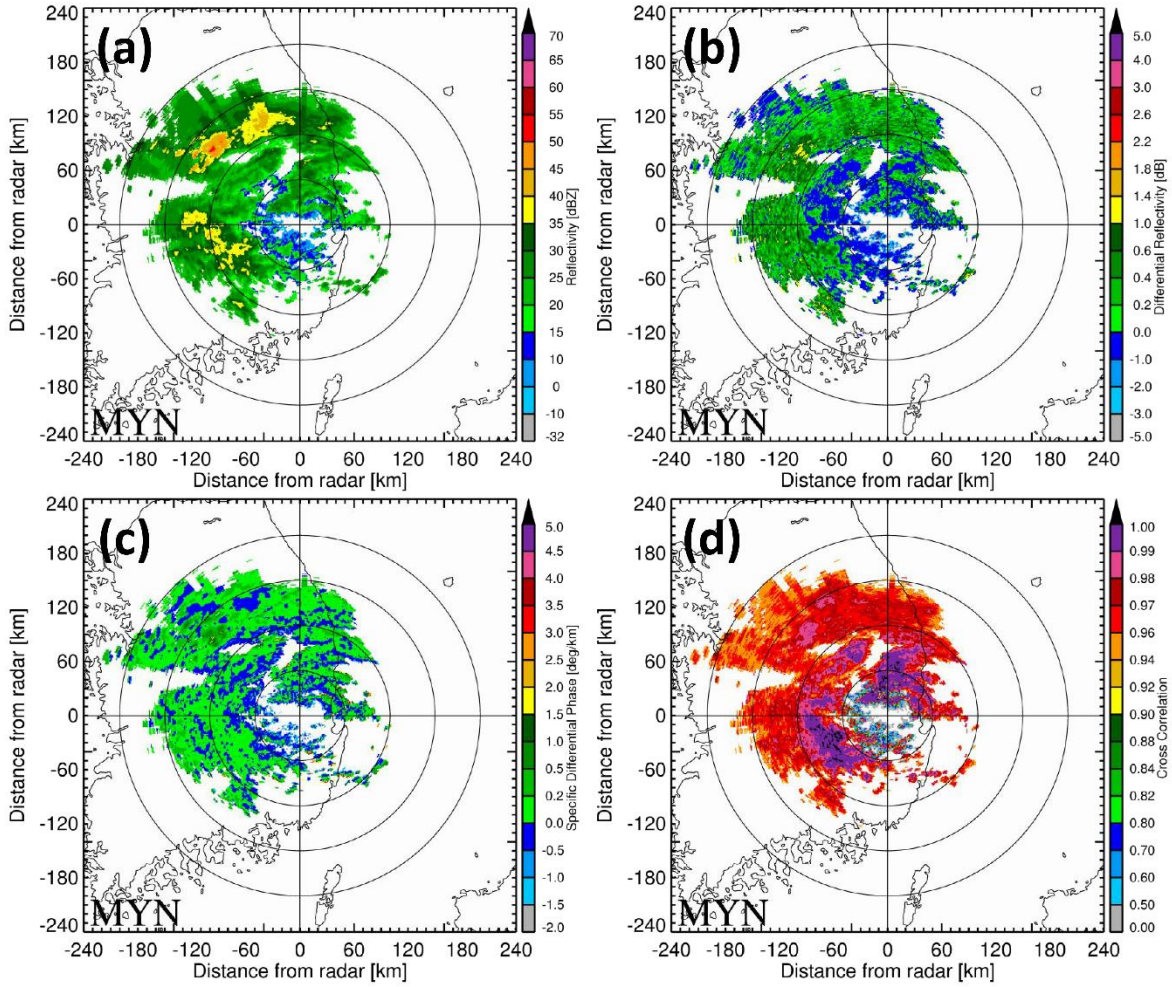


Figure 5. Radar measurements, (**a**) reflectivity (Z_H), (**b**) differential reflectivity (Z_{DR}), (**c**) specific differential phase (K_{DP}), and (**d**) cross-correlation coefficient (ρ_{HV}), at 09:40 KST on 31 July 2017.

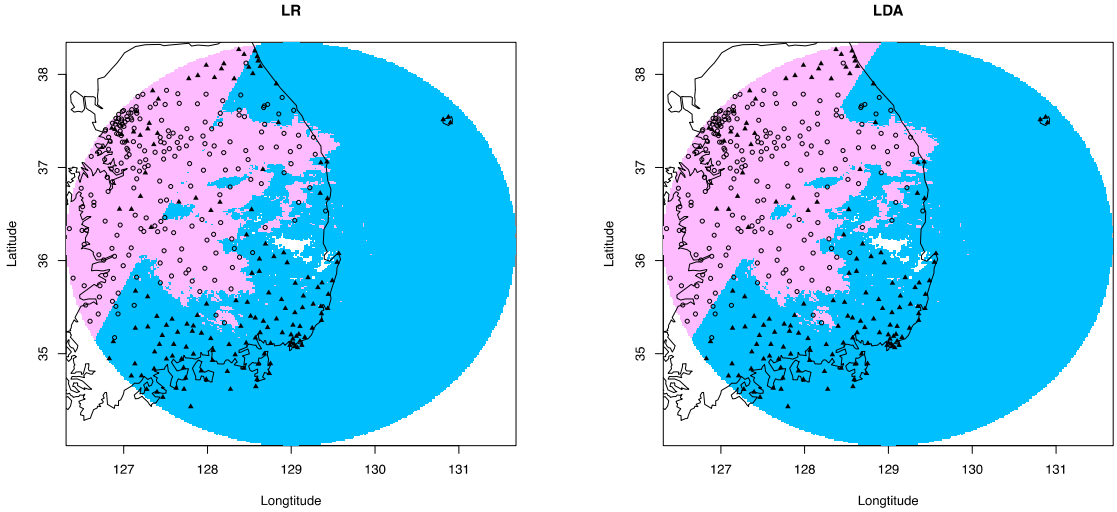


Figure 6. *Cont.*

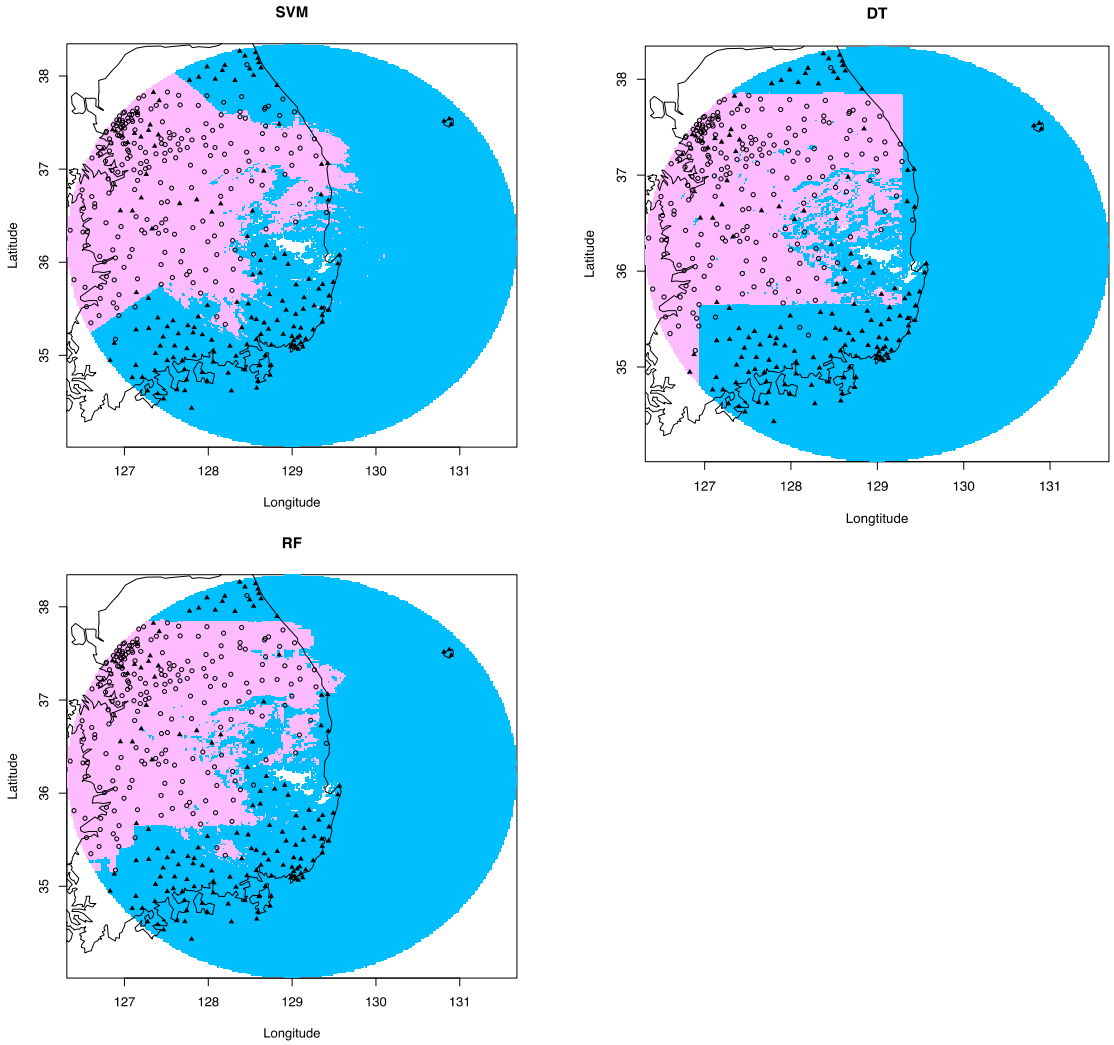


Figure 6. Estimated precipitation areas, dry (blue) and wet (red), on 31 July 2017. Circle and triangle indicate wet and dry observed in the stations, respectively.

Box plots of the MSPE values for each of the five classification methods are shown in Figure 7. The overall variability of the MPSE values is similar for all five methods. The DT method has a slightly lower variability. The statistical methods provide nearly the same distribution of the MSPE, whereas the machine learning methods yield slightly different boxplots. Note that the machine learning methods apparently achieve lower MSPE, and DT and RF methods have the lowest median of the MSPE values. The spatial patterns of the MSPE values are visualized in Figure 8. The spatial patterns of the statistical methods, LR and LDA, are similar, indicating large values in central and western regions. The errors show systematic structure in space. The machine learning methods also show similar spatial patterns, and RF has lower MSPE values on some stations in southern and northern regions. The errors are nearly random in space.

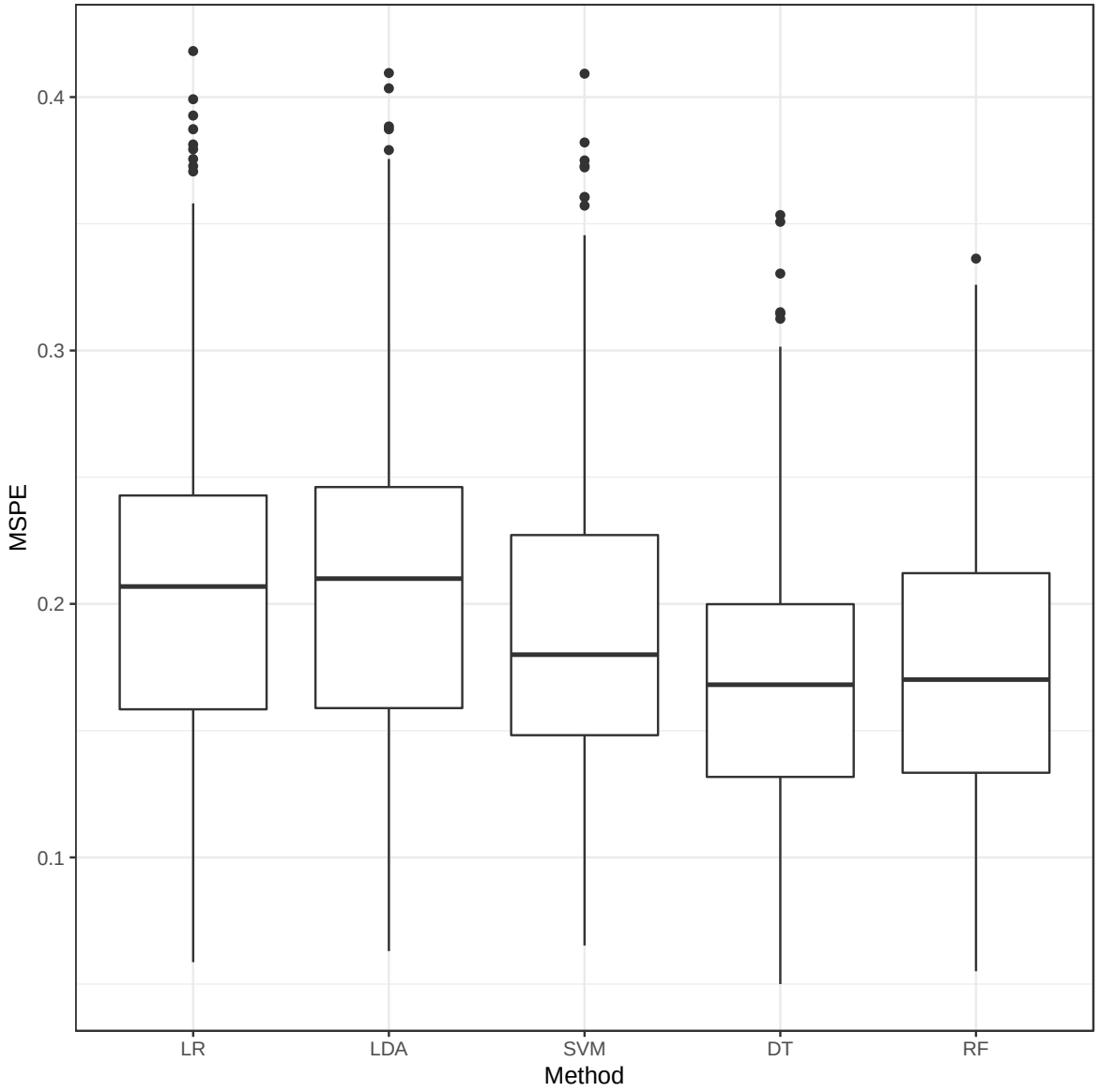


Figure 7. Boxplots of the MSPE for the five classification methods.

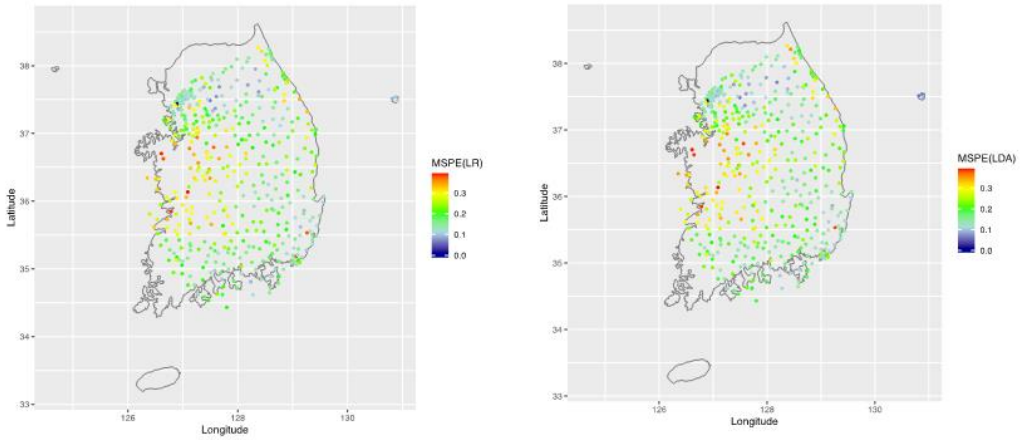


Figure 8. *Cont*.

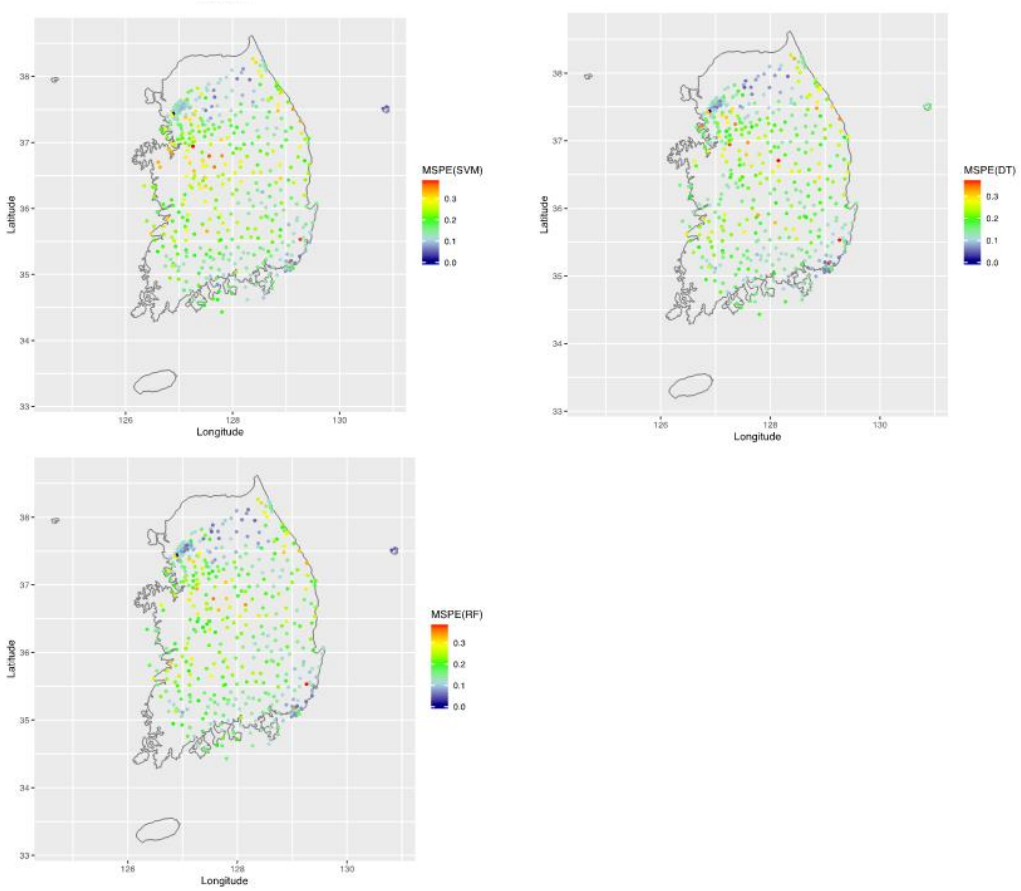


Figure 8. Spatial distributions of the MSPE for S3.

Figure 9 presents the boxplots of the MSPE values for each of these two types of rain events and of all events. As expected, the stratiform category has smaller MSPE than the convective category. However, there are some outliers in stratiform. The convective category showed the largest error.

Figure 10 displays the boxplots of the MSPE for the five classification methods by precipitation types. It indicates that the machine learning methods, SVM, DT, and RF, mostly outperform the statistical methods, LR and LDA, in terms of the MSPE. Note that there are outliers on the high end for all classification methods. For the convective category, the median of the MSPE for DT is considerably lower than that of the other methods. Similar to the stratiform category, the machine learning methods have smaller medians than the statistical methods. Interestingly, there are no outliers in the boxplots for the convective category. The combined category shows slightly larger medians than the stratiform categories. Likewise, the machine learning methods perform better than the statistical methods except SVM. A few outliers of MSPE are detected in DT and RF.

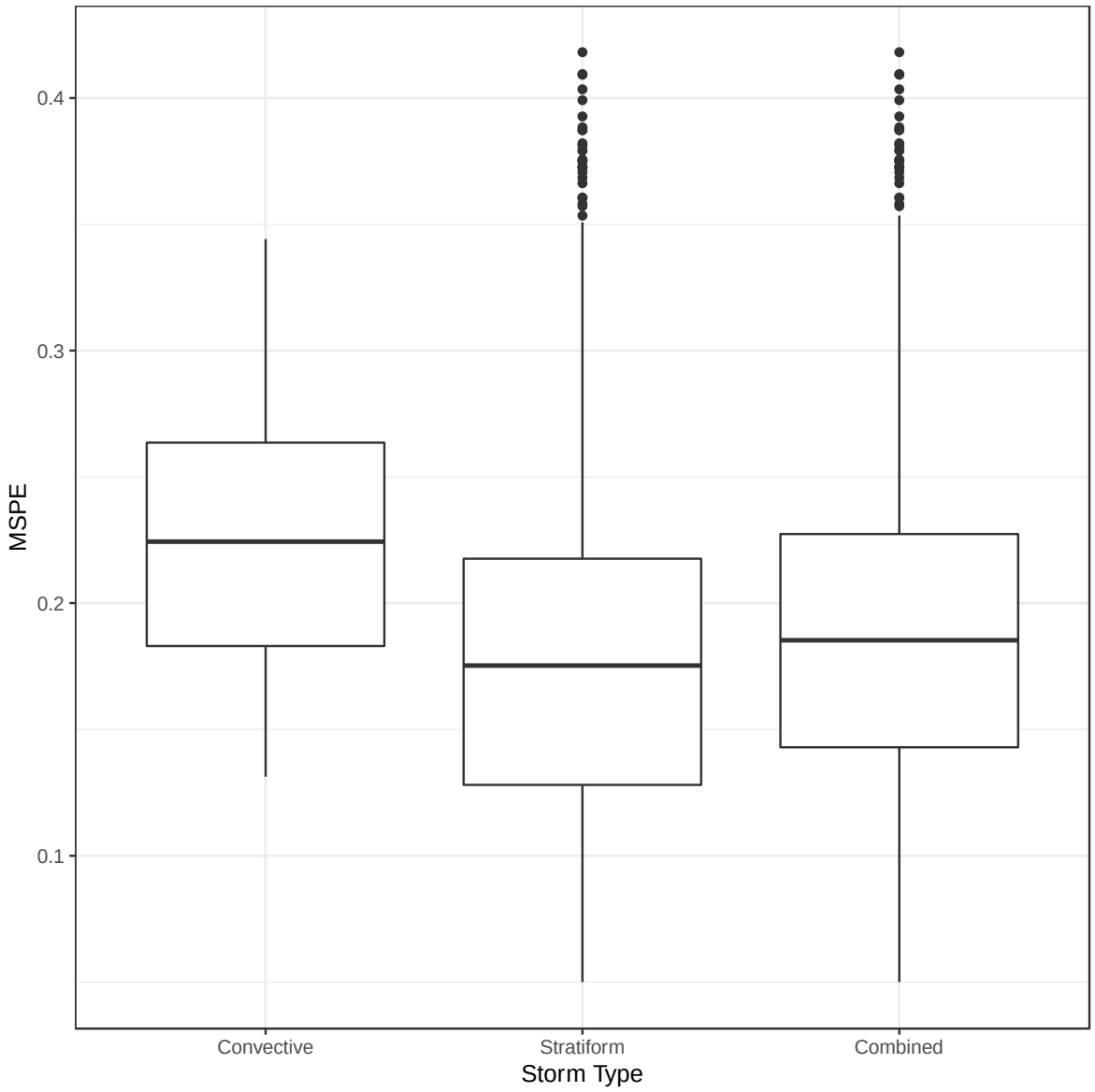


Figure 9. Boxplots of the MSPE for the three types of rain events.

In Figure 11, the estimated precipitation areas for the two types, convective and stratiform, are compared with the two selected classification methods, LDA and RF, that yield overall the worst and best results, respectively. For the convective case, RF correctly estimates wet areas in the western region, while LDA misclassifies them as dry. Similarly, RF provides more precise estimates in the northern and south-western regions for the stratiform case. It is worth pointing out that the two dry stations in the south-western region are accurately estimated by RF even though they are surrounded by wet areas.

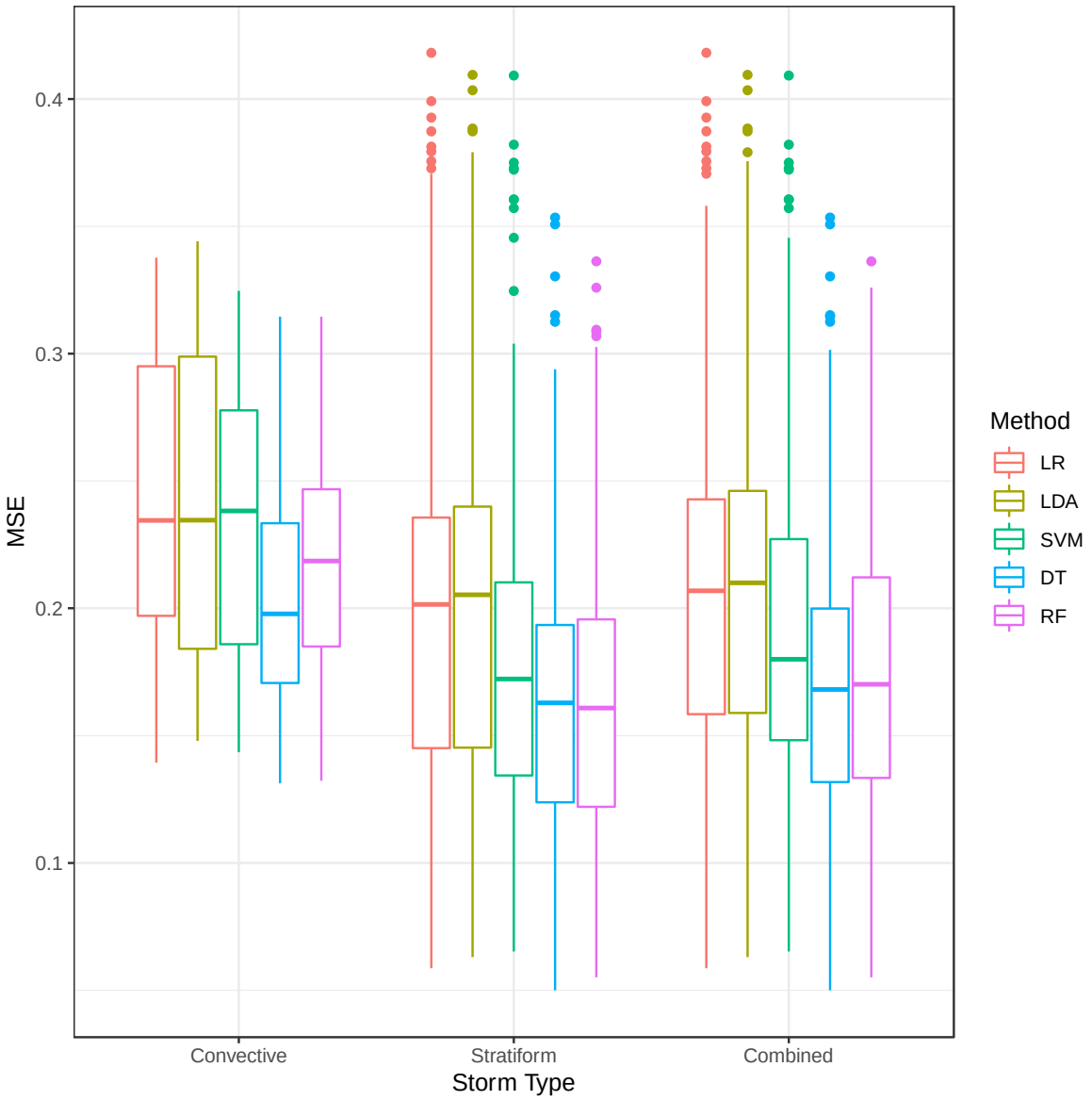


Figure 10. Boxplots of the MSPE for the three types of events by classification method.

For validation purposes, we perform the precipitation area estimations with data from different years, 2018 and 2019, and a different validation method. In the previous analysis, the models are trained by 10 min data and evaluated by LOOCV, which is applicable in real time. In this analysis, we consider temporally aggregated data sets acquired by combining all 10 min data sets in each event and hold-out validation in order to examine the sensitivity of model performance to temporal resolution and validation method. Table 5 presents the MSPE performance of the five models for the six events and shows similar results as compared to the previous analysis. The machine learning methods mostly outperform the statistical methods, and the tree-based methods are more appealing than the other methods.

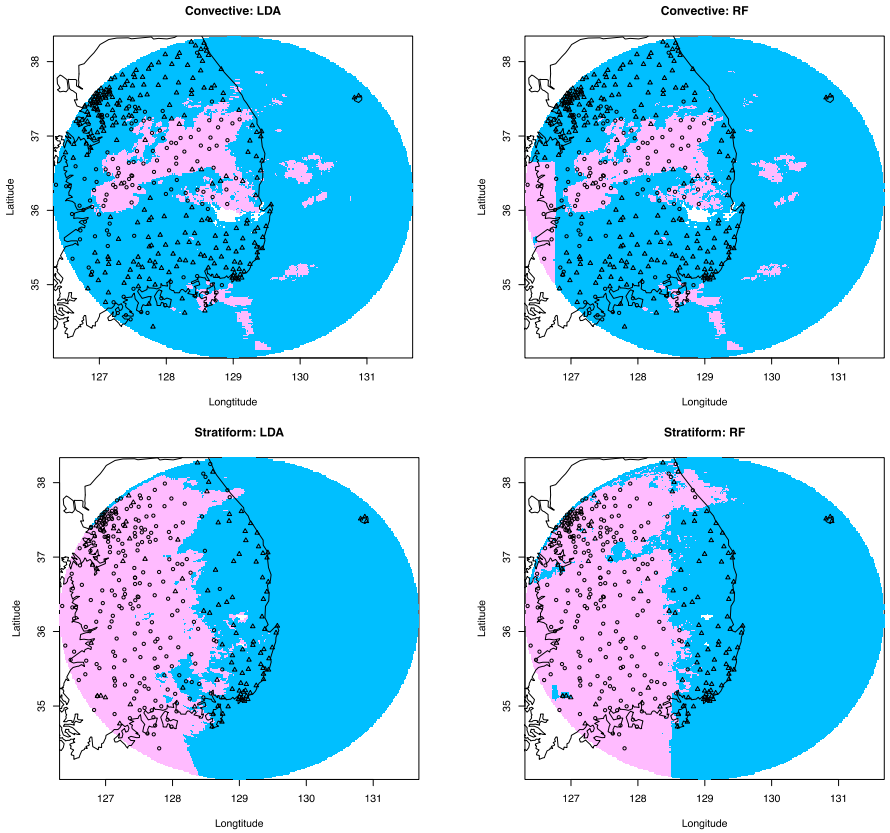


Figure 11. Estimated precipitation areas of LDA and RF for two types (stratiform and a convective) of cases. Estimated dry (blue), estimated wet (red), observed dry (triangle), and observed wet (circle).

Table 5. MSPE of the five models for the validation data sets. Each data is divided into the training set (80%) and the test set (20%).

Date (MM/DD/YYYY)	LR	LDA	SVM	DT	RF	Rainfall Duration [Hours]	The Number of 10-min Data
2018/04/23	0.2813	0.2900	0.2726	0.2701	0.2548	28	168
2018/06/27	0.0790	0.0819	0.0772	0.0790	0.0698	6	42
2018/08/26	0.1267	0.1289	0.1276	0.1278	0.1229	128	768
2018/09/03	0.1019	0.1043	0.1008	0.1024	0.0945	22	132
2019/04/09	0.2015	0.2001	0.1947	0.1836	0.1631	10	60
2019/05/18	0.1143	0.1203	0.1006	0.1048	0.0781	9	54

5. Discussion

Our results indicate that the best model is the tree-based methods (DT and RF), which does not require retraining once it is trained. The tree-based model is computationally efficient and easy to deploy at any machine. Wolfensberger et al. [47] also mentioned that the RF-based polarimetric radar QPE algorithm can easily be used operationally. Shin et al. [48] also showed significant improvement in QPE with the RF. Given that statement and the fact that any additional scaling of input variables is not needed in our study, it is obvious that our model is suitable for operational application.

We have conducted the precipitation area estimation using MYN S-band radar data and AWS data. Microphysics is dependent on different climate and regions because of the variability of the drop size distribution [49]. Therefore, in terms of QPE, when the location has changed, the values of dual-polarization variables related with microphysics information would be changed. However, estimation of precipitation area is less affected

by locations due to the range of polarimetric variables, which indicates that precipitation signal is less dependent on the regions.

6. Conclusions

We have studied precipitation area estimation using ground-based measurements and remote sensing products. Several rain events classified as either stratiform, or convective are used for analysis. We considered three factors, rain sensors, thresholds, and models, to present a novel data framework for improving the estimation. As a result, Scenario 3 with AWS, Type 1 threshold, spatial model, yields better performance than other scenarios. With the data framework, a rigorous comparison of the classification methods for precipitation area estimation was conducted. It is found that the machine learning methods yield better predictive performance than the statistical methods. Specifically, the tree-based methods, such as decision trees and random forests, are attractive methods for both stratiform and convective storms.

This study classified rain events as one of two categories, stratiform or convective, according to their precipitation patterns. Typically, this classification is originated from spatial extension and temporal duration of precipitation. Due to various physical conditions and characteristics, it seems to be more effective to estimate precipitation areas separately for different categories. To this end, it is an imperative task to accurately classify an event in data preprocessing.

The results presented in this paper indicate that no single method exhibits clear superiority over others in different events. In those situations, multi-model inference is an alternative by combining multiple statistical and machine learning methods in order to improve overall predictive performance. In statistics, model averaging is often used to improve predictive ability and account for uncertainty due to model choice. The most common method is Bayesian model averaging (BMA) that places prior probability on each model and determines the posterior distribution as the weighted average of the individual models.

The complex physical processes governing the precipitation process have substantial variability over space and time. In this paper, we consider incorporating spatial information into the precipitation area estimation. The methods addressed in the paper can be extended to incorporate spatial-temporal information by adding a dynamic component to the estimation.

In this study, we have found that spatial classification outperforms non-spatial classification. This finding can be related to censoring radar data. Radar measurements are often subject to censoring due to detection limits. In this case, there are several methods to deal with this problem. Firstly, analyzing a complete data set by deleting censored data is a common approach, which is inefficient because the information by discarding the censored data is lost. The other common approach is to replace censored values with a fixed value such as detection limit, as we have done in this paper. It is well known that these naïve methods result in bias estimation. The censored radar measurements in the data sets have the same values replaced by the practical detection limits, which is quite noninformative in the classification. However, spatial classification can provide additional information for classifying precipitation area by borrowing information from near observations.

The rain-detecting devices installed on the ground monitoring stations would be subject to measurement error. Since the variable of interest in this study is a binary variable, the measurement error problem is known as misclassification in statistics. We expect that future studies accounting for misclassification in the estimation will produce better predictive abilities than standard estimation.

Author Contributions: This work was made possible by significant contributions from all authors. Conceptualization, J.J.S. and G.L.; methodology, J.J.S., M.I., and G.L.; data preparation: K.S. and J.J.S.; software, J.J.S., M.I., K.S., and M.K.; validation, K.S., B.-Y.Y. and D.Y.; formal analysis, J.J.S., B.-Y.Y., M.K. and D.Y.; investigation, J.J.S. and G.L.; writing—original draft preparation, J.J.S., B.-Y.Y., and K.S.; writing—review and editing, J.J.S. and G.L.; visualization, J.J.S. and K.S.; supervision, J.J.S. and

G.L.; funding acquisition, J.J.S. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by "Development of integrated technology for utilization of Korea weather radar (KMA2021-00220)" project of the Weather Radar Center, Korea Meteorological Administration. This work was funded by the Korea Meteorological Administration Research and Development Program under Grant KMI2020-00910.

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

Acknowledgments: We would like to thank to the Weather Radar Center for operating the weather radar and for providing radar and AWS 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. Qin, Y.; Chen, Z.; Shen, Y.; Zhang, S.; Shi, R. Evaluation of satellite rainfall estimates over the Chinese Mainland. *Remote Sens.* **2014**, *6*, 11649–11672. [CrossRef]
2. Franke, R. Scattered Data Interpolation: Tests of Some Method. *Math. Comput.* **1982**, *38*, 181. [CrossRef]
3. Yilmaz, K.K. Towards Improved Modeling for Hydrologic Predictions in Poorly Gauged Basins. Ph.D. Thesis, University of Arizona, Tucson, AZ, USA, 2007. Available online: https://repository.arizona.edu/handle/10150/195252 (accessed on 17 May 2021).
4. Dai, Q.; Bray, M.; Zhuo, L.; Islam, T.; Han, D. A Scheme for Rain Gauge Network Design Based on Remotely Sensed Rainfall Measurements. *J. Hydrometeorol.* **2017**, *18*, 363–379. [CrossRef]
5. Song, J.J.; Kwon, S.; Lee, G.W. Incorporation of Parameter Uncertainty into Spatial Interpolation Using Bayesian Trans-Gaussian Kriging. *Adv. Atmos. Sci.* **2015**, *32*, 413–423. [CrossRef]
6. Waken, R.J.; Song, J.J.; Kwon, S.; Min, K.H.; Lee, G.W. A flexible and efficient spatial interpolator for radar rainfall estimation. *J. Appl. Stat.* **2018**, *45*, 829–844. [CrossRef]
7. Ryu, S.; Song, J.J.; Kim, Y.; Jung, S.H.; Do, Y.; Lee, G.W. Spatial Interpolation of Gauge Measured Rainfall Using Compressed Sensing. *Asia Pac. J. Atmos. Sci.* **2021**, *57*, 331–345. [CrossRef]
8. Gorgucci, E.; Scarchilli, G.; Chandrasekar, V. Operational Monitoring of Rainfall over the Arno River Basin Using Dual-Polarized Radar and Rain Gauges. *J. Appl. Meteorol.* **1996**, *35*, 1221–1230. [CrossRef]
9. Adeyewa, Z.D.; Nakamura, K. Validation of TRMM radar rainfall data over major climatic regions in Africa. *J. Appl. Meteorol.* **2003**, *42*, 331–347. [CrossRef]
10. Wang, J.; Wolff, D.B. Evaluation of TRMM ground-validation radar-rain errors using rain gauge measurements. *J. Appl. Meteorol. Climatol.* **2010**, *49*, 310–324. [CrossRef]
11. Kwon, S.-H.; Lee, G.; Kim, G. Rainfall Estimation from an Operational S-Band Dual-Polarization Radar: Effect of Radar Calibration. *J. Meteorol. Soc. Jpn.* **2015**, *93*, 65–79. [CrossRef]
12. Lee, J.W.; Lee, E.-H. Evaluation of daily precipitation estimate from integrated MultisatellitE Retrievals for GPM (IMERG) data over South Korea and East Asia. *Atmosphere* **2018**, *28*, 273–289. [CrossRef]
13. Gultepe, I.; Milbrandt, J.A. Probabilistic parameterizations of visibility using observations of rain precipitation rate, relative humidity, and visibility. *J. Appl. Meteorol. Climatol.* **2010**, *49*, 36–46. [CrossRef]
14. Kim, M.-S.; Kwon, B. Rainfall Detection and Rainfall Rate Estimation Using Microwave Attenuation. *Atmosphere* **2018**, *9*, 287. [CrossRef]
15. Xie, P.; Arkin, P.A. Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Clim.* **1996**, *9*, 840–858. [CrossRef]
16. Rahimi, A.R.; Holt, A.R.; Upton, G.J.G.; Cummings, R.J. Use of dual-frequency microwave links for measuring path-averaged rainfall. *J. Geophys. Res. Atmos.* **2003**, *108*. [CrossRef]
17. Grecu, M.; Krajewski, W.F. An efficient methodology for detection of anomalous propagation echoes in radar reflectivity data using neural networks. *J. Atmos. Ocean. Technol.* **2000**, *17*, 121–129. [CrossRef]
18. Berenguer, M.; Sempere-Torres, D.; Corral, C.; Sanchez-Diezma, R. A fuzzy logic technique for identifying nonprecipitating echoes in radar scans. *J. Atmos. Ocean. Technol.* **2006**, *23*, 1157–1180. [CrossRef]
19. Cho, Y.-H.; Lee, G.; Kim, K.-E.; Zawadzki, I. Identification and removal of ground echoes and anomalous propagation using the characteristics of radar echoes. *J. Atmos. Ocean. Technol.* **2006**, *23*, 1206–1222. [CrossRef]
20. Rico-Ramirez, M.A.; Cluckie, I.D. Classification of ground clutter and anomalous propagation using dual-polarization weather radar. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1892–1904. [CrossRef]
21. Hubbert, J.C.; Dixon, M.; Ellis, S.M.; Meymaris, G. Weather radar ground clutter. Part I: Identification, modeling, and simulation. *J. Atmos. Ocean. Technol.* **2009**, *26*, 1165–1180. [CrossRef]
22. Ye, B.-Y.; Lee, G.; 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. [CrossRef]

23. Andrieu, H.; Creutin, J.D.; Delrieu, G.; Faure, D. Use of a weather radar for the hydrology of a mountainous area. Part I: Radar measurement interpretation. *J. Hydrol.* **1997**, *193*, 1–25. [CrossRef]
24. Kabeche, F.; Figueras i Ventura, J.; Fradon, B.; Boumahmoud, A.A.; Dupuy, S.; Westrelin, S.; Tabary, P. Quantitative precipitation estimation(QPE) in the French Alps with a densenetwork of polarimetric X-band radars. In Proceedings of the 35th Conference on Radar Meteorology, Pittsburgh, PA, USA, 26–30 September 2011; American Meteorological Society: Boston, MA, USA, 2011; pp. 11–150. Available online: https://ams.confex.com/ams/35Radar/webprogram/Paper191894.html (accessed on 17 May 2021).
25. Fulton, R.A.; Breidenbach, J.P.; Seo, D.-J.; Miller, D.A.; O'Bannon, T. The WSR-88D rainfall algorithm. *Weather Forecast.* **1998**, *13*, 377–395. [CrossRef]
26. Chang, P.-L.; Lin, P.-F.; Dao Jou, B.J.; Zhang, J. An application of reflectivity climatology in constructing radar hybrid scans over complex terrain. *J. Atmos. Ocean. Technol.* **2009**, *26*, 1315–1327. [CrossRef]
27. 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. [CrossRef]
28. Giuli, D.; Gherardelli, M.; Freni, A.; Seliga, T.A.; Aydin, K. Rainfall and clutter discrimination by means of dual-linear polarization radar measurements. *J. Atmos. Ocean. Technol.* **1991**, *8*, 777–789. [CrossRef]
29. Ryzhkov, A.V.; Zrnic, D.S. Polarimetric rainfall estimation in the presence of anomalous propagation. *J. Atmos. Ocean. Technol.* **1998**, *15*, 1320–1330. [CrossRef]
30. Liu, H.; Chandrasekar, V.; Gorgucci, E. Detection of rain/no rain condition on the ground based on radar observations. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 696–699. [CrossRef]
31. Lauri, T.; Koistinen, J.; Moisseev, D. Advection-based adjustment of radar measurements. *Mon. Weather Rev.* **2012**, *140*, 1014–1022. [CrossRef]
32. Yan, J.; Bárdossy, A. Short time precipitation estimation using weather radar and surface observations: With rainfall displacement information integrated in a stochastic manner. *J. Hydrol.* **2019**, *574*, 672–682. [CrossRef]
33. Dong, J.; Crow, W.T.; Reichle, R. Improving Rain/No-Rain Detection Skill by Merging Precipitation Estimates from Different Sources. *J. Hydrometeorol.* **2020**, *21*, 2419–2429. [CrossRef]
34. Kuśmierek-Tomaszewska, R.; Zarski, J.; Dudek, S. Meteorological automated weather station data application for plant water requirements estimation. *Comput. Electron. Agric.* **2012**, *88*, 44–51. [CrossRef]
35. Welch, B.L. Note on discriminant functions. *Biometrika* **1939**, *31*, 218–220. [CrossRef]
36. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
37. Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; ISBN 9781475732641.
38. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
39. 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. [CrossRef]
40. Rinehart, R.E. *Radar for Meteorologists*, 5th ed.; Rinehart Publications: Saint Joseph, MO, USA, 2010; ISBN 9780965800235.
41. Kumjian, M.R. Principles and applications of dual-polarization weather radar. Part I: Description of the Polarimetric Radar Variables. *J. Oper. Meteorol.* **2013**, *1*, 226–242. [CrossRef]
42. KMA. *Real-Time Quality Control System for Meteorological Observation Data (I) Application, 11-1360000-000206-01*; Technical Note 2006-2; KMA: Seoul, Korea, 2006; p. 157.
43. Zawadzki, I.I. Statistical Properties of Precipitation Patterns. *J. Appl. Meteorol.* **1973**, *12*, 459–472. [CrossRef]
44. Zawadzki, I.I. On Radar-Raingage Comparison. *J. Appl. Meteorol.* **1975**, *14*, 1430–1436. [CrossRef]
45. Datta, S.; Jones, W.L.; Roy, B.; Tokay, A. Spatial Variability of Surface Rainfall as Observed from TRMM Field Campaign Data. *J. Appl. Meteorol.* **2003**, *42*, 598–610. [CrossRef]
46. Lee, C.K.; Lee, G.; Zawadzki, I.; Kim, K.-E. A Preliminary Analysis of Spatial Variability of Raindrop Size Distributions during Stratiform Rain Events. *J. Appl. Meteorol. Climatol.* **2009**, *48*, 270–283. [CrossRef]
47. Wolfensberger, D.; Gabella, M.; Boscacci, M.; Germann, U.; Berne, A. RainForest: A random forest algorithm for quantitative precipitation estimation over Switzerland. *Atmos. Meas. Tech.* **2021**, *14*, 3169–3193. [CrossRef]
48. 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. [CrossRef]
49. 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**, *21*, 2675–2690. [CrossRef]