Improving Radar-Based Precipitation Nowcasts with Machine Learning Using an Approach Based on Random Forest

YIWEN MAO[a] AND ASGEIR SORTEBERG[a,b]

[a] *Geophysical Institute, University of Bergen, Bergen, Norway*

(Manuscript received 20 May 2020, in final form 23 September 2020)

ABSTRACT: A binary classification model is trained by random forest using data from 41 stations in Norway to predict the precipitation in a given hour. The predictors consist of results from radar nowcasts and numerical weather predictions. The results demonstrate that the random forest model can improve the precipitation predictions by the radar nowcasts and the numerical weather predictions. This study clarifies whether certain potential factors related to model training can influence the predictive skill of the random forest method. The results indicate that enforcing a balanced prediction by resampling the training datasets or lowering the threshold probability for classification cannot improve the predictive skill of the random forest model. The study reveals that the predictive skill of the random forest model shows seasonality, but is only weakly influenced by the geographic diversity of the training dataset. Finally, the study shows that the most important predictor is the precipitation predictions by the radar nowcasts followed by the precipitation predictions by the numerical weather predictions. Although meteorological variables other than precipitation are weaker predictors, the results suggest that they can help to reduce the false alarm ratio and to increase the success ratio of the precipitation prediction.

SIGNIFICANCE STATEMENT: Machine learning can be useful in improving weather forecasts with relatively inexpensive computational efforts. Specifically, this study has demonstrated that radar nowcasts can be improved by integrating the information from radar and numerical weather prediction using the random forest method. The random forest method's performance shows seasonality but is only weakly influenced by the geographic diversity of the training dataset. Also, there is no need to use specific strategies to address the imbalance of the precipitation and no precipitation frequency from the observations during model training. However, future study is needed to identify better predictor choices to further improve the random forest method.

KEYWORDS: Classification; Hindcasts; Nowcasting; Statistical forecasting

1. Introduction

The methods that are primarily used for nowcasting precipitation over a short period of time (e.g., <6 h) can be classified into two categories: 1) methods based on numerical weather prediction (NWP) models, and 2) methods based on extrapolating radar echoes (Dixon and Wiener 1993; Li et al. 1995; Germann and Zawadzki 2002; Mandapaka et al. 2012; Hwang et al. 2015; Shi et al. 2015; Zou et al. 2019). To accurately nowcast precipitation at a local station using NWP-based methods, models with fine spatial-temporal resolutions need to be run, which can be computationally expensive (Reyniers 2008; Shi et al. 2015). Radar based precipitation nowcasts, which are based on extrapolating radar echoes, can provide a high spatial (1 km) and temporal resolution (10 min; Reyniers 2008), but good predictive skill is not always guaranteed. Nowcasting precipitation using both methods (radar echoes and NWP) is not perfect, and the complexity of current nowcasting systems varies greatly. Some systems are based on simple tracking algorithms, while others require a variety of observations that are processed using sophisticated algorithms. Furthermore, the biggest limitation with radar-based precipitation nowcasting is the difficulty in predicting the development of new precipitation areas, and it is not clear whether methods based on sophisticated algorithms are more accurate than simple methods (Reyniers 2008).

Machine learning (ML) methods, also called data-driven methods, have gained popularity in recent years. In general, ML requires a transfer function that links predictors (input) to predictands (output) based on historical data, and new predictions can be made by feeding new predictor data into a transfer function. Their advantages include that they can formulate complex data relationships (e.g., nonlinearity) solely based on historical data. They are relatively easy to implement with a low computational cost, but their performance can be compared to physical models (Mosavi et al. 2018). ML methods are often applied in predictions of natural hazards (e.g., floods, landslides, and avalanches). For example, Liu et al. (2020) apply three different ML algorithms to spatial modeling of shallow landslides in Norway.

In this study, we explore whether predictions made by radar nowcasts and NWP can be improved by ML models. If they can, ML has the potential to be a practical method to improve current nowcasting models. The performance of ML models

[b] Current affiliation: Geophysical Institute, University of Bergen, Bergen, Norway.

Corresponding author: Yiwen Mao, yiwen.mao@uib.no

DOI: 10.1175/WAF-D-20-0080.1

can be improved by model training. Therefore, empirical tests and trials are needed. Many algorithms are available to build an ML model, but this study will not focus on choosing an optimal ML algorithm. Instead, a commonly known and robust algorithm, random forest (RF; Breiman 2001), is used to build a binary classification model for predicting whether there is precipitation or no precipitation.

Generally speaking, precipitation data are often characterized by unbalanced frequencies of precipitation and no precipitation. The number of precipitation hours should be much less than the number of hours without precipitation at most locations. Just by always guessing the majority class (i.e., no precipitation in this study), one can achieve the accuracy of more than 50%, but such prediction does not consider the minority class. Therefore, the accuracy can be misleading, as the minority class is overlooked (Japkowicz and Stephen 2002; Guo et al. 2008). There are two common strategies to address the imbalance of the majority and minority class: 1) assigning different costs for the predictions of the two classes and 2) resampling the original datasets to obtain balanced datasets before training (Chawla et al. 2002). This study investigates whether it is beneficial to use these strategies in the problem of predicting precipitation and no precipitation.

The geographic locations of the data included for training may also influence the performance of the classification model built by the RF. If the same physics governs the predictand–predictor relationship for all regions, the predictand–predictor relationship at different areas can be modeled by one universal function. In this case, only one ML model is needed to make predictions at different locations. Adding training data from far away locations may improve the prediction because the number of training samples increases. On the other hand, if the predictand–predictor relationships are strongly impacted by local factors, training a local model using data from other regions will not be useful. Also, predictors' strength may vary with seasons. In addition, the variables that are most relevant to the model results often cannot be determined in advance. Thus, additional variables (other than nowcasts results from the radar and NWP) may be needed to counterbalance the negative influences of incorrect predictions made by the radar nowcasts and NWP.

There are two main objectives of this study. The first is to clarify how the aforementioned factors related to model training (i.e., methods used to deal with the unbalanced binary classes, geographic locations and seasonality of training data, and the choice of predictors) can influence the performance of the RF binary classification used for precipitation prediction. The second is to assess whether predictive skills of NWP and radar based nowcasts can be further improved by using an RF model. The rest of the paper is organized as follows. Section 2 presents the data, and the methodology is described in section 3. Section 4 presents the main results. Discussions and conclusions are provided in section 5.

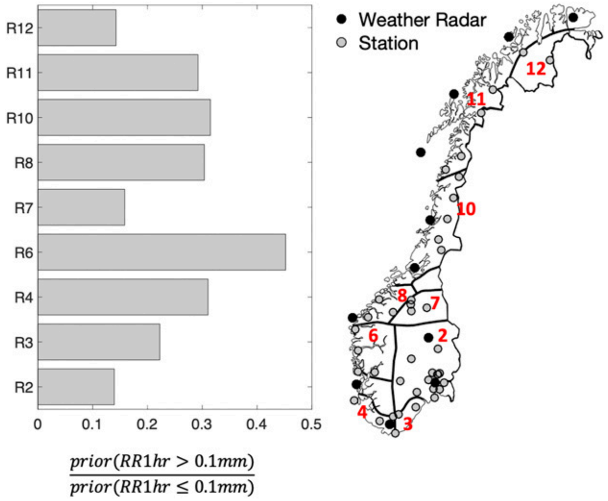

FIG. 1. (right) The locations of the 41 stations used in this study are depicted. They are grouped into 9 of the 13 precipitation regions. The locations of weather radars in Norway are also shown. (left) The ratio of the prior probability of precipitation to that of no precipitation in each region.

2. Data

All data used in this study (radar nowcasts, NWP, and historical observations) are obtained from the Norwegian Meteorological Institute (2011). The algorithm for the radar nowcasts used in this study is described in (Chambolle and Pock 2011), The NWP model used in this study is the convective-permitting operational weather prediction model Applications of Research to Operations at Mesoscale (AROME). The model covers Scandinavia and the Nordic Seas with a horizontal resolution of 2.5 km (Müller et al. 2017).

The Norwegian meteorological institute has divided Norway into 13 precipitation regions based on EOF and cluster analysis of seasonal and annual precipitation variability (Hanssen-Bauer and Nordli 1998). Specifically, these regions are characterized by highly correlated time series of monthly precipitation data. The 41 stations considered in this study are grouped into 9 of the 13 regions (Fig. 1). The data available for each station considered in the study are divided into two subsets. The first subset consists of data collected from February 2017 to December 2017, and the second subset consists of data collected from February 2018 to December 2018.

The two subsets are used as training and testing datasets, respectively. The details of the datasets are presented in section 2b. Only the results tested on the dataset of 2018 (and trained by the dataset of 2017) are presented in the paper. The same conclusions as presented in this paper can also be found when the training and testing datasets are interchanged. The datasets of the predictand and predictors (section 2b) at each station are at least 73% complete for the period considered in this study, and most datasets (35 out of 41 stations) are more than 80% complete.

a. *An example of precipitation predictions by radar or numerical weather predictions (NWP) and observations*

An example at a central eastern station in Norway is used to compare the accuracy of radar nowcasts with that of NWP in predicting precipitation and no precipitation (Fig. 2). It shows

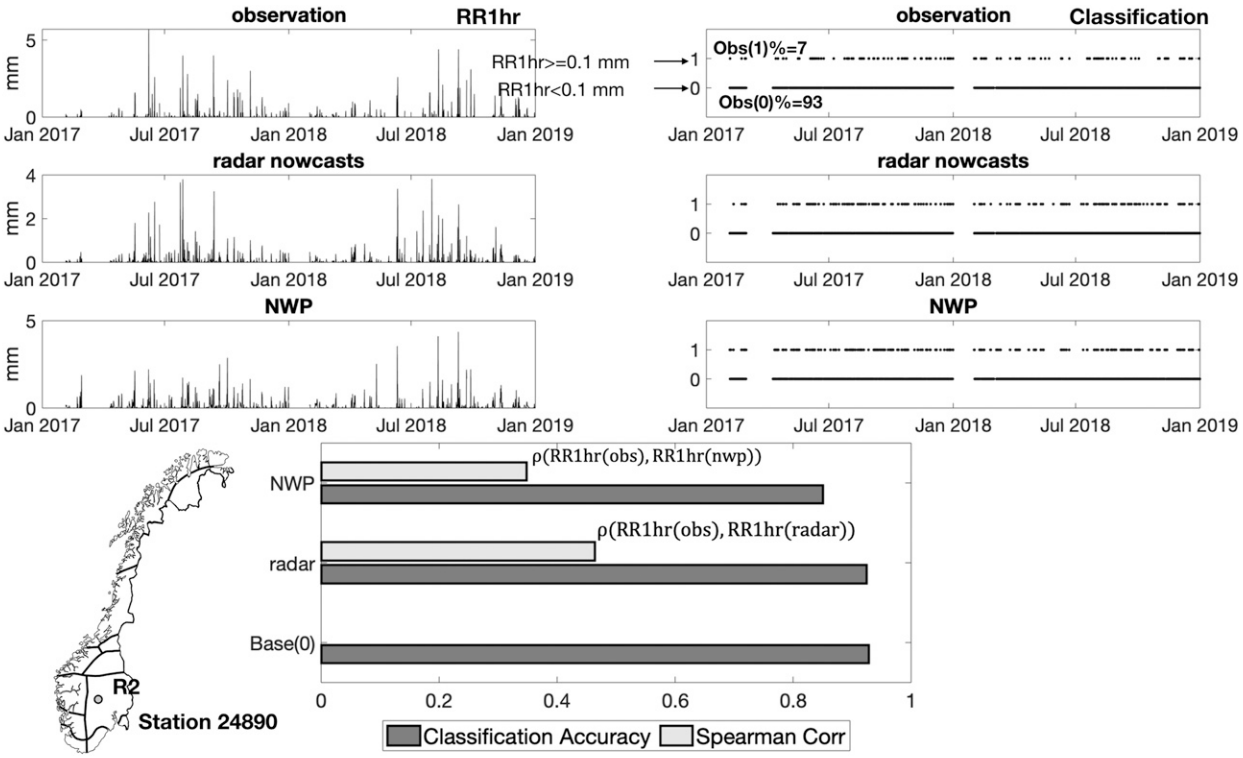

FIG. 2. (top) Accumulated precipitation in 1 h (RR1hr) and the corresponding classification of precipitation and no precipitation are shown for the observation, NWP, and radar nowcasts for the station 24890 (Bromma, Nes) in region 2. (bottom) The darker bar plot shows the accuracy of the precipitation classification (the percentage of correct predictions out of all predictions) by the radar nowcast (radar), NWP, and a base prediction of always guessing no precipitation [Base(0)]. The lighter bar plot indicates the Spearman correlation coefficients between the RR1hr from observation (obs) and those from the NWP/radar nowcasts.

that radar nowcasts are often more accurate than NWP. In this example, the accuracy (i.e., the percentage of correct predictions out of all predictions) of the dichotomous predictions of Class 1 (precipitation) and Class 0 (no precipitation) by the NWP and radar nowcasts is 0.85 and 0.92, respectively. The better accuracy of the radar nowcasts also corresponds to a higher degree of covariation (i.e., higher Spearman correlation coefficient) between the precipitation observations and precipitation predictions.

Second, the observations are dominated by Class 0 (no precipitation), which is typical of most regions in Norway. In this example, approximately 7% of all observations are precipitation events. Therefore, guessing no precipitation (i.e., predicting 0) for all cases can give an impressively high accuracy of 0.93, which is close to the accuracy of the radar nowcasts and exceeds that of NWP by approximately 9%. To surpass the simple prediction of always guessing 0 (no precipitation), it is desirable to improve the predictive skills of NWP and radar nowcasts by some other methods, such as an RF model.

b. *The predictand and predictors of the random forest (RF) model*

In this study, the RF model is used to predict whether there is precipitation within a given hour of observation at a station. The predictand of the RF model is the observed accumulated precipitation in 1 h, denoted as RR1hr, with an observational interval of 6 h (i.e., 0000–0100, 0600–0700, 1200–1300, and 1800–1900 UTC). The predictand is labeled as Class 1 if RR1hr exceeded or equaled to 0.1 mm. Otherwise, it is labeled as Class 0.

This study considers 17 variables as potential predictors for the RF model. The predictors consist of variables that are directly related to the precipitation predictions by the radar nowcasts (1 and 2 in Table 1) and NWP (3 and 4 in Table 1). In addition, some other NWP-derived variables (5–17 in Table 1) are also used as predictors. These variables are chosen because of their potential relevance to precipitation and the availability of data on the variables during the period of study. In particular, the K index is a measure of thunderstorm potential (George 2014), and it is defined as $K = T_{850\mathrm{mb}} - T_{500\mathrm{mb}} + \mathrm{Td}_{850\mathrm{mb}} - T_{700\mathrm{mb}} - \mathrm{Td}_{700\mathrm{mb}}$, where T and Td denote temperature and dewpoint temperature at the specified pressure level (1 mb = 1 hPa). This study also includes the wind speed normal to the topographic aspect in the predictors, as this variable can address the potential influence of topography on precipitation. A topographic aspect is the compass direction that a slope faces.

The variables from the radar nowcasts and NWP are forecasting values that correspond to the observational hour. The forecast lead time of the radar nowcasts is 2 h, and the

TABLE 1. List of all variables considered as predictors for the RF model.

	Full name	Abbreviation
1	Accumulated precipitation in 1 h from the radar nowcasts	RR1hr(Radar)
2	Duration of precipitation in 1 h from the radar nowcasts	tL(Radar)
3	Accumulated precipitation in 6 h from AROME	RR6hr(NWP)
4	Accumulated precipitation in 1 h from AROME	RR1hr(NWP)
5	Air pressure at sea level	SLP
6	Air temperature at 2 m	T2m
7	Fog area fraction	Fog AF
8	Low-type cloud area fraction	LowC AF
9	Medium-type cloud area fraction	MediumC AF
10	High-type cloud area fraction	HighC AF
11	Relative humidity at 2 m	RH2m
12	Zonal wind U at 10 m	U10m
13	Meridional wind V at 10 m	V10m
14	Atmospheric boundary layer thickness	ABLT
15	Wind speed normal to the topographic aspect	WSNT
16	Average dewpoint depression from 1000 to 500 mb	DD(1000–500 mb)
17	K index	K

forecast reference time is 6 h before the observational hour for AROME. Both the main cycles of AROME and the observational hours for the predictands are 0000, 0600, 1200, and 1800 UTC, so the interval of 6 h is the shortest between a forecast reference time and the next available hour of observation. In other words, the AROME model is initialized 6 h before the observational hour. The grid spacing of AROME is 2.5 km × 2.5 km, while that of the radar nowcasts is 1 km × 1 km.

Each predictor variable listed in Table 1 is the average of two expressions as shown in Eq. (1). The first expression is the value at the grid point closest to each station. Given that the atmosphere is a continuum, the weather at one location is influenced by its surroundings. Therefore, the second expression is the average grid values weighted by the inverse of the distance from the center of a square of 100 km × 100 km, centered at the grid point closest to the station. The 100 km × 100 km area is large enough to include all grid points the values of which can contribute meaningfully to the result of the inverse distance weighted average:

$$\begin{cases} p_1 = p_c, & \text{if} \quad d(x_c, x_i) = 0, \\ p_2 = \dfrac{\sum\limits_{i=1}^{N} w_i p_i}{\sum\limits_{i=1}^{N} w_i}, & \text{if} \quad d(x_c, x_i) > 0, \end{cases} \tag{1}$$

where

- p_1 and p_2 are the two expressions of a predictor;
- x_c and p_c denote the grid point closest to each station and the corresponding predictor, respectively;
- x_i is the ith grid point in the area of 100 km × 100 km centered at x_c;
- N is the number of grid points in the area; and
- $w_i = 1/d(x_c, x_i)$ denotes the weight of the ith grid point (i.e., the inverse distance from x_c).

3. Methodology

a. Random forest

This study uses RF to implement the ML model for predicting precipitation. This section gives a brief overview of RF. Random forest belongs to the ensemble learning, which is a process based on generating many simple models and aggregating their results (Hastie et al. 2009). The individual model for RF classification is a tree-based classification model. A node in the tree represents a spilt point on a predictor variable. Each terminal node of a tree (i.e., a leaf) is one particular classification (i.e., the prediction). The tree structure can be created by repeatedly dividing the training data (i.e., binary splitting). The splitting point of each predictor variable is determined based on some cost function. For classification, a commonly used cost function is the Gini index which measures the degree of homogeneity of the groups created by each split. The splitting process stops when terminal nodes contain a minimum number of training data samples, or the tree's growth reaches a maximum depth. The test data can be passed down the tree structure (i.e., various nodes) created by training until terminal nodes are reached. In this way, new predictions are made.

Tree models can output posterior probability for each class. Following the definition of the Statistics and Machine Learning Toolbox of MATLAB (MathWorks 2019), the posterior probability of a class associated with the tree model can be defined as the number of splitting sequences that lead to the classification of the class divided by the number of all possible splitting sequences. Posterior probability can be converted to classification through the choice of a threshold probability p, that is,

- Posterior probability ≥ p → Predicting Class 1
- Posterior probability < p → Predicting Class 0

The posterior probability for the RF classification model is defined as the mean posterior probability for each class of all tree models used to build the RF model. The prior probability in this study is defined as the fraction of training samples of each class out of all training samples.

Tree models are prone to have a high level of noise, which makes the prediction results unstable as a small change in training data can lead to very different sequences of splitting. One remedy is bagging; that is, aggregating the noisy results obtained from many tree models. For classification models, bagging refers to building a committee of tree models using bootstrapped samples from the training dataset, and the individual classification by each tree model is analogous to casting a vote. The result of the classification is the majority vote of the committee. Random forest modifies the procedures of bagging by ensuring all tree models within the committee are decorrelated. Decorrelation is achieved by randomly selecting a subset of predictor variables to form the splitting sequence during the tree-growing process (Breiman 2001).

An important feature of RF is the use of out-of-bag (OOB) samples, which are samples not included in the tree growing process. Specifically, the prediction of $z_i = (x_i, y_i)$ in the input space is derived from the majority vote of tree models constructed using bootstrap samples in which z_i is not included. Therefore, the training process is identical to that obtained by N-fold cross validation (Hastie et al. 2009). The OOB error can be used to find the number of trees needed for the RF model, and the training is terminated once the OOB error stabilizes. For each training case in this study, at least 500 trees are used, which is not smaller than the number of trees that stabilized the OOB error.

b. *Evaluation of classification models*

The model results in this study are a dichotomous prediction: precipitation exceeding 0.1 mm within the hour of observation (RR1hr ≥ 0.1 mm) is labeled as Class 1, and no precipitation (RR1hr < 0.1 mm) is labeled as Class 0. The 2 × 2 contingency table summarizes all possible prediction outcomes (Table 2). Various metrics for evaluating the model performance can be derived from the contingency table.

Since the prior probability of Class 1 is much smaller than that of Class 0 in most regions (Fig. 2), the contingency table is dominated by correct rejection D. The most common metric accuracy is defined as

$$\mathrm{ACC} = \frac{A + D}{A + B + C + D}, \tag{2}$$

where D is predominantly larger than A, B, and C, the accuracy will always be good even though Class 1 (i.e., the minority) is greatly misclassified.

Therefore, this study focuses on metrics that can reflect the skills of the classification of the minority Class 1 (RR1hr ≥ 0.1 mm). They are the probability of detection (POD), probability of false detection (POFD), false alarm ratio (FAR), success ratio (SR), critical success index (CSI), skill score relative to the base prediction of always predicting 0 (SS0), and frequency bias. The meanings of these metrics in the context of this study are summarized in Table 3, and their formulas are listed below (Wilks 2011; Inness and Dorling 2012):

$$\mathrm{POD} = \frac{A}{A + C}, \tag{3}$$

$$\mathrm{POFD} = \frac{B}{B + D}, \tag{4}$$

$$\mathrm{FAR} = \frac{B}{A + B}, \tag{5}$$

$$\mathrm{SR} = \frac{A}{A + B}, \tag{6}$$

$$\mathrm{CSI} = \frac{A}{A + B + C}, \tag{7}$$

$$\mathrm{SS0} = \frac{\mathrm{ACC} - \mathrm{ACC}(0)}{1 - \mathrm{ACC}(0)}, \tag{8}$$

TABLE 2. A 2 × 2 contingency table for a binary classification; A, B, C, and D represent the number of cases belonging to each category.

	Observe 1	Observe 0
Predict 1	A (hit)	B (false alarm)
Predict 0	C (miss)	D (correct rejection)

$$\mathrm{bias} = \frac{A + B}{A + C}, \tag{9}$$

where ACC [Eq. (2)] is the accuracy of the classification model, and ACC(0) is the accuracy of the base prediction of always predicting 0.

In addition, SR can be expressed as SR = 1 − FAR. CSI considers both false alarms B and missed predictions of precipitation C; therefore, CSI is a more balanced metric than POD and SR (Nurmi 2003). CSI and bias can be expressed by POD and SR as

$$\mathrm{CSI} = \frac{1}{\dfrac{1}{\mathrm{SR}} + \dfrac{1}{\mathrm{POD}} - 1}, \tag{10}$$

$$\mathrm{bias} = \frac{\mathrm{POD}}{\mathrm{SR}}. \tag{11}$$

Therefore, the quantities: POD, SR (FAR), CSI, and the bias can be visualized in one diagram (Roebber 2009), with SR as the x axis and POD as the y axis.

SS0 ≤ 0 indicates that the classification model is not more accurate than simply guessing Class 0 (the majority class: precipitation) for all occasions, and bias < 1 and bias > 1 indicate underprediction and overprediction of precipitation events, respectively (Inness and Dorling 2012).

Moreover, since the RF classification model can output posterior probability, different classification results can be achieved by changing the threshold probability. The receiver operating characteristics (ROC) curve provides a visual representation of the general goodness of the classification model by plotting POFD versus POD as the threshold probability varies. The area under the ROC curve (AUC) associates the characteristics of the ROC with a number which varies between 0 (the worst model) and 1 (the best model).

Finally, the empirical 90% confidence intervals based on bootstrap are calculated for all metrics evaluations. Specifically, each test dataset is sampled with replacement 10 000 times. All sampled test datasets are the same size as the original one, and the metrics are evaluated on each sampled dataset. A sequence of differences between the metrics of the original dataset and those of the sampled datasets is calculated, and the 95th and 5th percentiles of the sequence are extracted to construct the empirical 90% confidence interval for the metrics of the original dataset.

c. *Factors influencing RF predictability*

This study compares the test results of different RF models trained using a controlled variation of each factor of interest to demonstrate how the factors can influence the predictive skills of the RF model.

TABLE 3. Descriptions of various metrics used in this study.

Abbreviation	Full name	Meanings	Best	Worst
POD	Probability of detection	The fractional success of predicting precipitation out of all occasions when precipitation is observed	1	0
POFD	Probability of false detection	The number of times precipitation is falsely predicted out of all occasions when precipitation is not observed	0	1
FAR	False alarm ration	The number of times precipitation is falsely predicted out of all precipitation predictions	0	1
SR	Success ratio	The fractional success of precipitation prediction out of all precipitation predictions	1	0
CSI	Critical success index	The fractional success of precipitation prediction out of all precipitation predictions and missed predictions of precipitation	1	0
SS0	Skill score relative to a base prediction	The relative improvement of the classification model from the base prediction of always predicting the majority Class 0 (no precipitation)	1	<0
bias	Frequency bias	Quantifies whether the RF model tends to predict the precipitation more (bias > 1) or less (bias < 1) often than it is actually observed	1	

1) FACTOR 1: THRESHOLD PROBABILITY AND RESAMPLING METHODS

The time series of observed precipitation is characterized by the unbalanced prior probability of precipitation and no precipitation. Classification models that are trained using unbalanced datasets tend to predict the majority class more frequently while ignoring the minority class (Longadge and Dongre 2013). First, the study assesses whether the problem caused by the unbalanced prior probability of the training data can be solved simply by choosing a different threshold probability for classification. In particular, the focus has been on lowering the threshold probability to allow more predictions of the minority class.

Next, the prediction results of a model trained without resampling (No RS) and of models trained using 1) oversampling (OS) of the minority class and 2) undersampling (US) of the majority class have been compared. OS and US are two common methods for addressing the problem of the unbalanced training dataset.

Oversampling is achieved by applying the synthetic minority oversampling technique (SMOTE; Chawla et al. 2002), which constructs synthetic minority class samples in the feature space. Undersampling is achieved by randomly removing some samples that belong to the majority class. Resampling using OS and US results in balanced training datasets.

2) FACTOR 2: GEOGRAPHIC DIVERSITY AND SEASONALITY OF TRAINING DATASETS

This study assesses whether it is possible to use one general model for the whole country or whether separate models are necessary for different locations. A related question is how the geographic diversity of the training data can influence the predictive skills of the RF classification model. To address this question, the trained models have been classified into three types based on the geographic diversity of the training data.

The first type is the RF model trained for each station by only using the data from the station. The second type is the RF model trained for each precipitation region (Fig. 1), which is obtained by pooling the training data from all stations in the region. The third type is the RF model trained by pooling training data from all stations. The three types of models are labeled as 1) Local, 2) Region, and 3) All, respectively.

In terms of the diversity of locations included in the training data, the first type (Local) is the least diverse because all training data are derived from a specific location. There is no geographic difference between the training and test data when applying a local model to predict the test data at the station. The third model (All) is the most diverse, since the training data are derived from stations all over the country. The training data in this case include locations that are very remote from the stations where the predictions are needed (i.e., the location of test data). The geographic diversity of the second model (Region) is between that of Local and All.

Moreover, since the NWP and radar nowcast's predictive skills vary with seasons, it is expected that the RF model may also be influenced by seasonality. Four seasonal models have been trained using only data from winter, summer, spring, and fall from the training dataset and tested on the respective seasonal data of the test dataset.

3) FACTOR 3: CHOICE OF PREDICTORS

Two different approaches have been used to analyze the influence of each predictor on the RF classification model.

The first approach obtains the ranking of predictor importance by permutating the training data of each predictor. Specifically, a baseline metric is evaluated on the training dataset, and a feature column (i.e., each predictor) of the training dataset is reshuffled. Then, the metric is evaluated again. The difference between the metrics before and after the permutation is calculated, and a larger difference suggests that the predictor is more important.

However, when two predictors are collinear, shuffling the values of one of them will not prevent the information of this predictor being fed into the RF model, as the RF model can

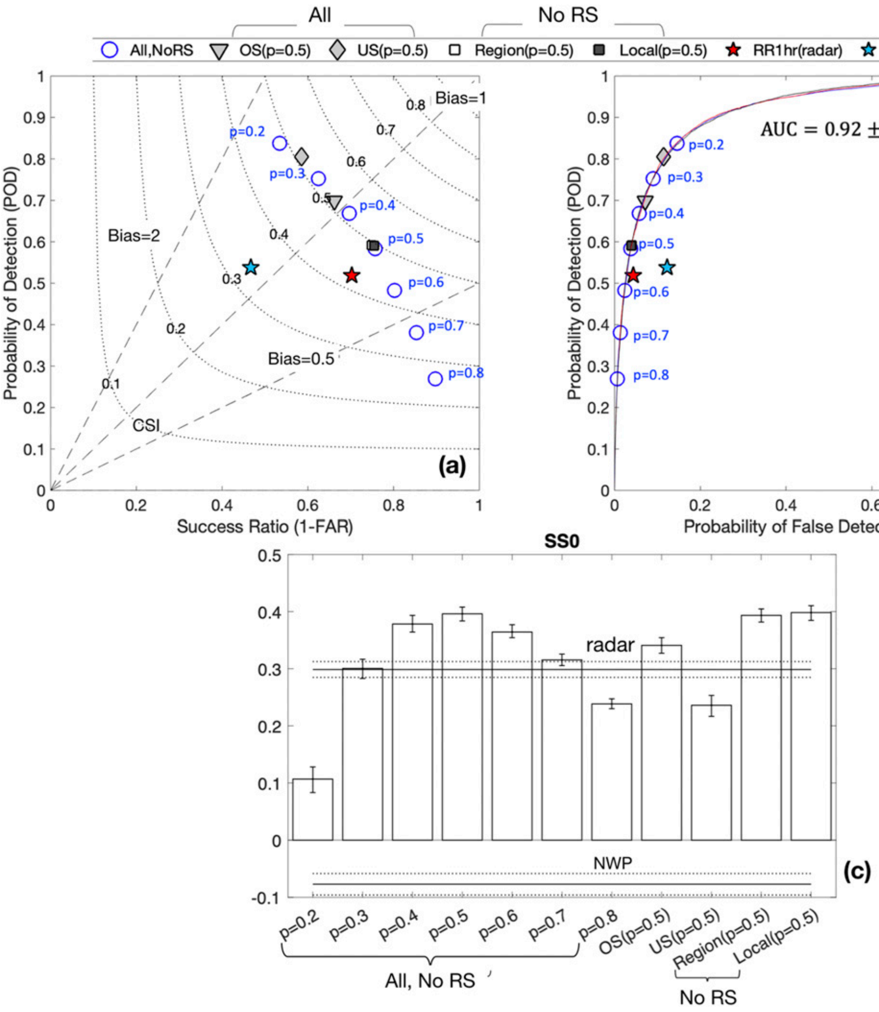

FIG. 3. Results of the RF models with different training datasets and tested on test datasets from all stations. All, Region, and Local refer to training data from all stations, stations in a region, and a local station, respectively. No RS refers to the use of the original dataset without resampling. OS (US) refers to resampling by oversampling (undersampling), and p denotes the threshold probability for classification. (a) The metrics (POD, SR, FAR, CSI, and bias). (b) POFD and ROC curves are shown, and the AUC values for all training cases considered are approximately 0.92 as labeled. (c) Values of SS0 for various training cases. The same metrics used to evaluate the two benchmarks: the precipitation predictions by the radar nowcasts and NWP are also shown for comparison. The error bars and dashed lines in the plot of SS0 indicate the 90% bootstrap confidence intervals.

obtain the same information from the correlated predictor. Therefore, the method based on permutation may not reflect the actual ranking of importance (Scikit-Learn Developers 2019). To solve this problem, the predictors are grouped into clusters by performing hierarchical clustering of the Spearman correlation matrix of predictors and keeping a single predictor from each cluster. Furthermore, a variable of random values is added as a reference predictor. The random variable will not add any information to the RF model; therefore, in an ideal situation it should have the lowest measure of importance among all predictors.

The second approach examines the relationship between the metrics of the RF classification model and the Spearman correlation coefficient between the precipitation observations (the predictand) and each predictor, denoted as ρ[RR1hr(obs), predictor]. Intuitively, the predictive information of a predictor can be assessed by the degree of covariation between the time series of the predictor and the precipitation observation. If the better predictive skills of the RF model correspond to higher covariations between the precipitation observation and a predictor, assessed by ρ[RR1hr(obs), predictor], then the predictor has a positive influence on the RF model.

A numerical experiment based on bootstrapping has been used to examine the relationships between the metrics of the RF classification models and ρ[RR1hr(obs), predictor]. Specifically, the test data of all 41 stations used in this study

TABLE 4. The empirical 90% bootstrap confidence intervals for various metrics shown in Fig. 3. All, Region, and Local refer to RF models trained using data from all stations, stations in each region, and each local station, respectively. No RS, OS, and US denote no resampling, oversampling, and undersampling, respectively. Precipitation predictions by the radar nowcasts and NWP are the two benchmarks.

Geographic diversity	Resampling	Threshold probability	ΔSR	ΔPOD	ΔCSI	Δbias	ΔPOFD
All	No RS	$p = 0.2$	0.007	0.007	0.007	0.023	0.003
All	No RS	$p = 0.3$	0.008	0.009	0.008	0.018	0.002
All	No RS	$p = 0.4$	0.009	0.009	0.009	0.014	0.002
All	No RS	$p = 0.5$	0.008	0.009	0.008	0.012	0.002
All	No RS	$p = 0.6$	0.010	0.010	0.010	0.012	0.001
All	No RS	$p = 0.7$	0.009	0.009	0.008	0.010	0.001
All	No RS	$p = 0.8$	0.010	0.008	0.008	0.008	0.001
All	OS	$p = 0.5$	0.008	0.008	0.006	0.015	0.002
All	US	$p = 0.5$	0.007	0.007	0.007	0.019	0.003
Region	No RS	$p = 0.5$	0.008	0.010	0.009	0.013	0.002
Local	No RS	$p = 0.5$	0.009	0.009	0.008	0.013	0.001
Benchmarks							
Radar			0.010	0.010	0.009	0.014	0.002
NWP			0.008	0.008	0.006	0.018	0.002

have been combined, and the test data at each station consist of all predictors and the precipitation observation [RR1hr(Obs)]. Overall, 10 000 test units have been created by repeatedly drawing a random sample of 500 data points from the pooled test data. Various metrics for evaluating the RF model, as well as ρ[RR1hr(obs), predictor], have been calculated for each test unit, and the pattern between the metrics and ρ[RR1hr(obs), predictor] is displayed by the probability distribution based on the 10 000 test units.

Since the bootstrapping experiment is based on the data of the 41 stations, no additional information can be added by bootstrapping. However, the results of the bootstrapping experiment help to delineate the underlying relationship between quantities of interest, since 41 data points may not be

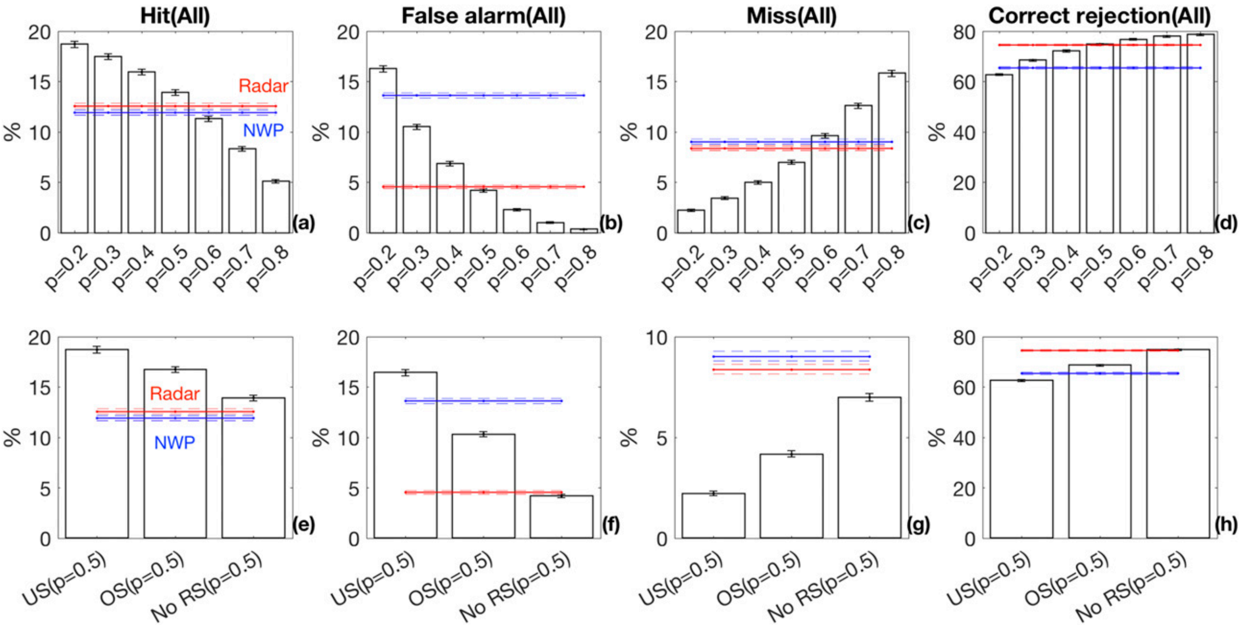

FIG. 4. (top) The percentage of test cases belonging to hit, false alarm, miss, and correct rejection for the RF model trained by using the original data (without resampling) from all stations but with different threshold probability p for classification. (bottom) The percentage of test cases belonging to hit, false alarm, miss, and correct rejection for the RF models trained by using data from all stations with and without resampling (No RS). OS and US refer to oversampling and undersampling, respectively. The red and blue solid lines indicate the values of the benchmarks: the radar nowcasts and NWP, respectively. The error bars and dashed lines indicate the corresponding 90% bootstrap confidence intervals for the RF models and benchmarks.

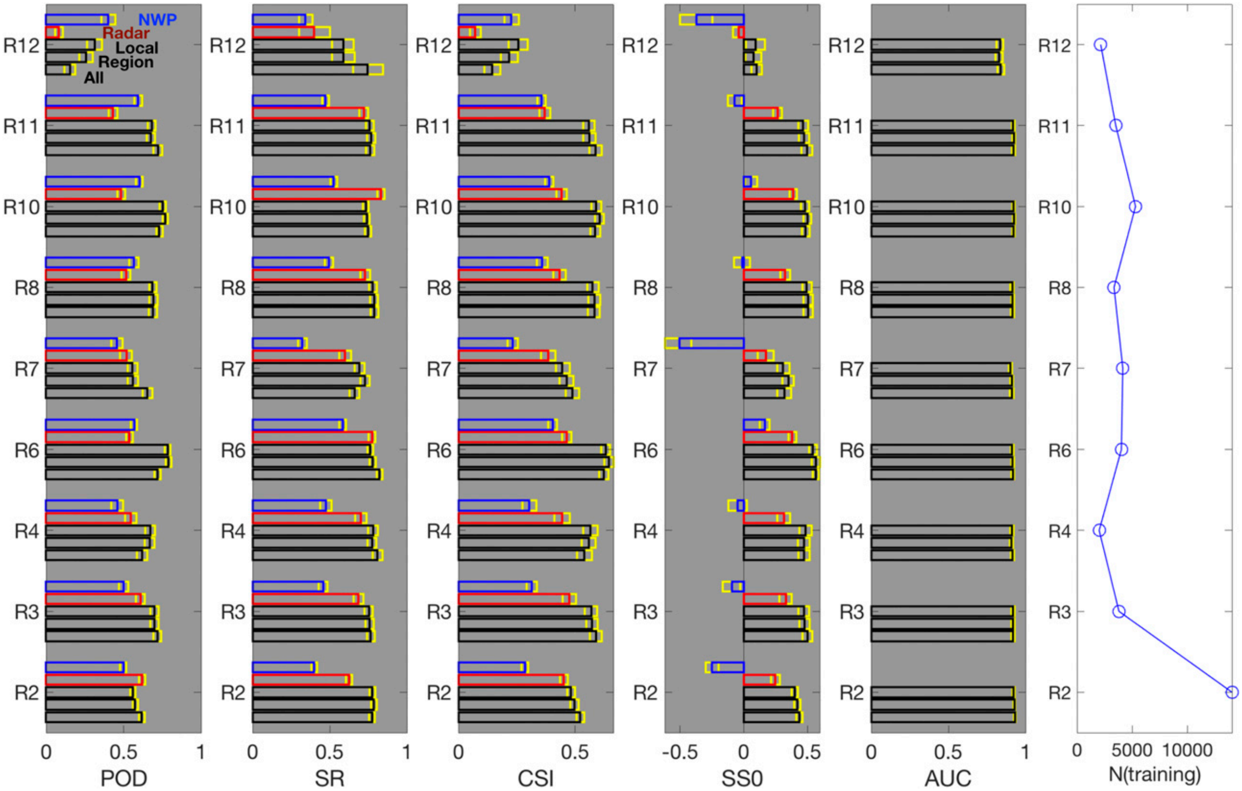

FIG. 5. The metrics of POD, SR, CSI, SS0, and AUC for evaluating RF models with training datasets consisting of data from a local station (Local), stations from a region (Region), and all stations (All), respectively, and tested on the test datasets in each region. The number of training samples in each region is shown in the last column. The same metrics (except AUC) for evaluating the two benchmarks: the precipitation predictions by the radar nowcasts and NWP are also shown for comparison. The yellow rectangles indicate the empirical 90% bootstrap confidence intervals.

enough to display the complete pattern. The number of data points in each test unit, 500, is chosen subjectively, but the choice ensures a broad range of values for all quantities of interest (i.e., metrics of the RF model and the Spearman correlation coefficients between the predictand and predictors).

Moreover, a third approach is used to assess whether additional meteorological variables, other than those directly related to the precipitation predictions from the radar nowcasts and NWP, can contribute to the predictive skills of RF. Specifically, the predictors in Table 1 can be divided into three subsets:

(i) Variables directly related to the precipitation prediction by the radar nowcasts (1 and 2 in Table 1);
(ii) Variables directly related to the precipitation prediction by AROME (3 and 4 in Table 1); and
(iii) Meteorological variables output from AROME other than precipitation (5–17 in Table 1).

Random forest models built by different subsets of predictors are compared with each other. The subsets of predictors are i and ii (precipitation only), iii (no precipitation), i (radar nowcasts only), and ii and iii (NWP only).Two complementary subsets of all test data (Test 1 and Test 2) have been used to evaluate the RF models. Specifically, the precipitation observations in Test 1 are correctly predicted by either the radar nowcasts or NWP and are misclassified by both the radar nowcast and NWP in Test 2.

d. Comparing RF with NWP and radar nowcasts

The same bootstrapping experiment described in section 3c(3) has also been used to examine the relationship between the predictive skill of the RF model and that of precipitation predictions by the radar nowcasts and NWP, to determine whether the RF model can further improve the predictions made by the radar nowcasts and NWP. For each of the 10 000 test units of the bootstrapping experiment, various metrics described in section 3b have been calculated for the RF model, as well as for the corresponding precipitation predictions by the radar nowcasts and NWP. The same metrics have also been calculated for the test data of all 41 stations. The metrics for the RF and for the radar/NWP have been compared using the probability distribution based on the 10 000 test units as well as the scatterplot of the 41 stations.

4. Results

The RF models can be tested on three test cases: 1) the test dataset of each station, 2) combining the test datasets of

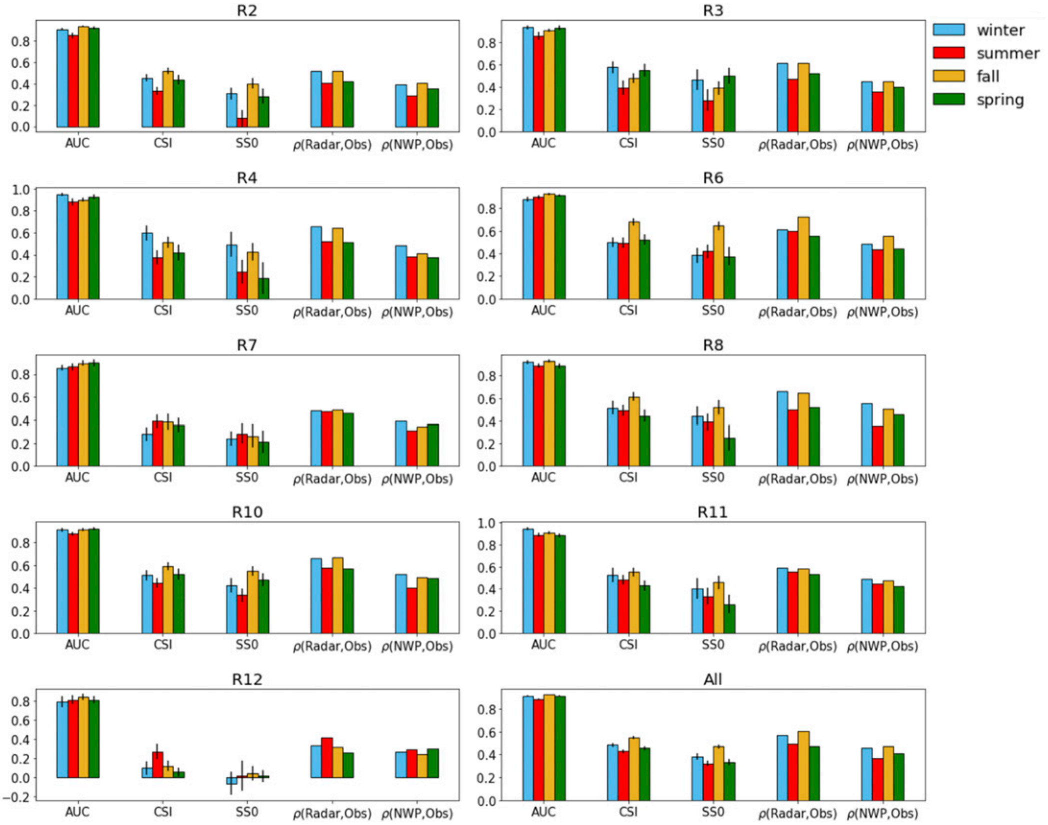

FIG. 6. The metrics of AUC, CSI, and SS0, for evaluating the seasonal RF models trained and tested using data from winter, summer, fall, and spring, respectively, for each region as well as for all regions combined (labeled as All). The Spearman correlations between precipitation from observations and radar nowcasts/NWP for each season are also displayed. The error bars indicate the corresponding 90% bootstrap confidence intervals.

stations in each region (Fig. 1), and 3) combining the test datasets of all 41 stations. The results of the RF models were also compared with the predictions (precipitation or no precipitation) by the radar nowcasts and NWP, which were used as the two benchmarks for comparison with the RF classification models.

Most results presented in sections 4a and 4b were tested on test case 3. The results are displayed in Fig. 3, and the corresponding 90% empirical bootstrap confidence intervals for various metrics are listed in Table 4, but they are too small to be clearly marked in Fig. 3, except for the SS0.

a. *Comparing various threshold probability and resampling methods*

Figures 3a and 3b demonstrates that, as the threshold probability decreased from $p = 0.8$ to $p = 0.2$, the POD increased, but the probability of false detection (POFD) and the FAR also increased, thereby lowering the SR. Moreover, the SS0 appeared to be associated with the threshold probability $p = 0.5$ (Fig. 3c). The values of the critical success index (CSI) did not vary noticeably from $p = 0.5$ to $p = 0.2$ but began to decrease for $p > 0.5$ (Fig. 3a). Overprediction of precipitation was associated with $p < 0.4$, and underprediction of precipitation occurred when $p > 0.4$ (Fig. 3a). The metrics associated with the RF model were better than the corresponding metrics for both benchmarks only at $p = 0.5$.

The upper panel of Fig. 4 summarizes the effects of changing the threshold probability p on the four elements of the contingency table. As the threshold probability decreased, the percentage of hits (Fig. 4a) increased and the percentage of misses (Fig. 4c) decreased, which contributed positively to the predictive skill of the RF model. However, the percentage of false alarms increased notably (Fig. 4b) and the correct rejection was also reduced (Fig. 4d), which lowered the predictive skill of the RF model. Since the percentages of hits, false alarms, and misses associated with $p = 0.5$ were all better than those of the radar nowcasts and NWP, $p = 0.5$ was chosen as the default threshold probability for classification in this study.

Figures 3 and 4e–4h demonstrate that resampling the training data to increase the proportion of the minority class had the same effects as lowering the threshold probability from $p = 0.5$. In particular, the results of oversampling (OS) and undersampling (US), as well as decreasing the threshold probability p, were like moving along the ROC curve toward the point of (1, 1) in the ROC space (Fig. 3b); in other words,

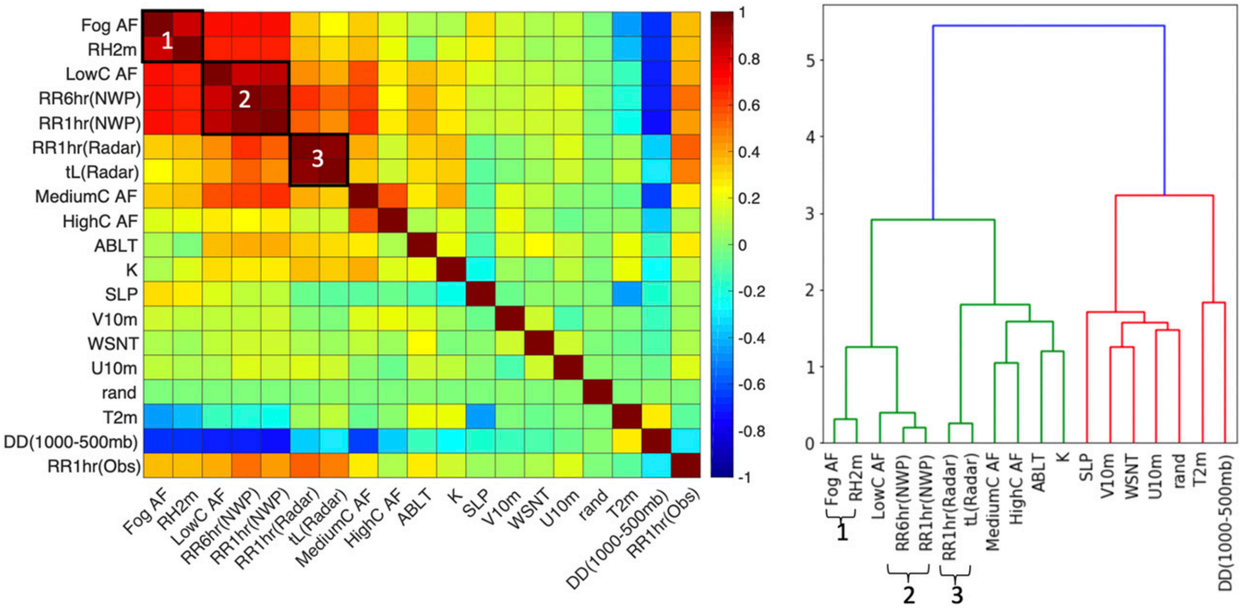

FIG. 7. (left) The heat map of all predictors listed in Table 1 and the precipitation observations for the dataset consisting of data from all stations. (right) The dendrogram used to visualize the hierarchical clustering of the Spearman correlations of the predictors. The three clusters of variables with strong collinearity are marked.

both POD and POFD increased. In the diagram of SR versus POD (Fig. 3a), the results of the OS, US, and decreasing p from 0.5 were similar to those obtained from moving along the same CSI contour toward higher POD and lower SR [Eq. (10)]. Also, the values of SS0 for US and OS decreased by approximately 40% and 13%, respectively, from that of no resampling (No RS) with $p = 0.5$ (Fig. 3c). Overall, the results displayed on Figs. 3 and 4 show that resampling the training datasets and lowering the threshold probability for classification are not effective in improving the predictive skill of the RF model. Therefore, the results discussed in the rest of this paper are based on the case of No RS and $p = 0.5$.

b. *Influences of geographic diversity and seasonality of training datasets*

Figure 3 also demonstrates that there were negligible differences among the RF models trained by the datasets consisting of data from a local station, stations in a region and all stations. The differences among the Local, Region, and All models were further verified on the test data of each region, as shown in Fig. 5. The difference between the Region model and Local model was negligible for all metrics considered in any region. The CSI, SR, and POD of the All model differed by around 10% or more from those of the Local model and the Region model in regions 4, 6, 7, and 12, but the SS0 and AUC were approximately the same for all three models in all regions.

Moreover, Fig. 6 shows that the variations of AUC of the four seasons were not substantial. However, the values of CSI and SS0 show that predictive skills of summer and spring were lower than those of winter and fall when considering the test results of pooling test data of all regions. Although the seasonal variations of predictive skills in terms of CSI and SS0 differed in regions, it was common that either winter or fall values stood out as the best, and the lowest values among the four seasons were often found in summer and spring in most regions. However, region 12 was exceptional as the summer CSI was more than 50% higher than the other three seasons. Moreover, the Spearman correlations between the precipitation predictions by radar nowcasts/NWP and observation were more likely to be higher in fall and winter than in summer and spring, but region 12 was a notable exception.

c. *Influences of chosen predictors*

This section summarizes the results of using three different approaches [outlined in section 3c(3)] to analyze the influence of the chosen predictors on the predictive skills of the RF model.

1) FIRST APPROACH: RANKING IMPORTANCE BY PERMUTATION OF PREDICTOR DATA

The heatmap of all predictors listed in Table 1 is shown in Fig. 7. The heatmap indicates that some predictors were highly correlated. The dendrogram in Fig. 7 shows the hierarchical clustering of predictors based on the Spearman correlations of the predictors. Three clusters were identified: 1) Fog AF and RH2m, 2) LowC AF, RR6hr(NWP), and RR1hr(NWP), and 3) RR1hr(Radar) and tL(Radar). Furthermore, the heatmap indicates that variables belonging to these three clusters had higher Spearman correlation coefficients with the precipitation observations (i.e., the predictand) than the other variables.

To remove strong collinearity to rank the importance of predictors, one variable was kept in each cluster. The predictors in

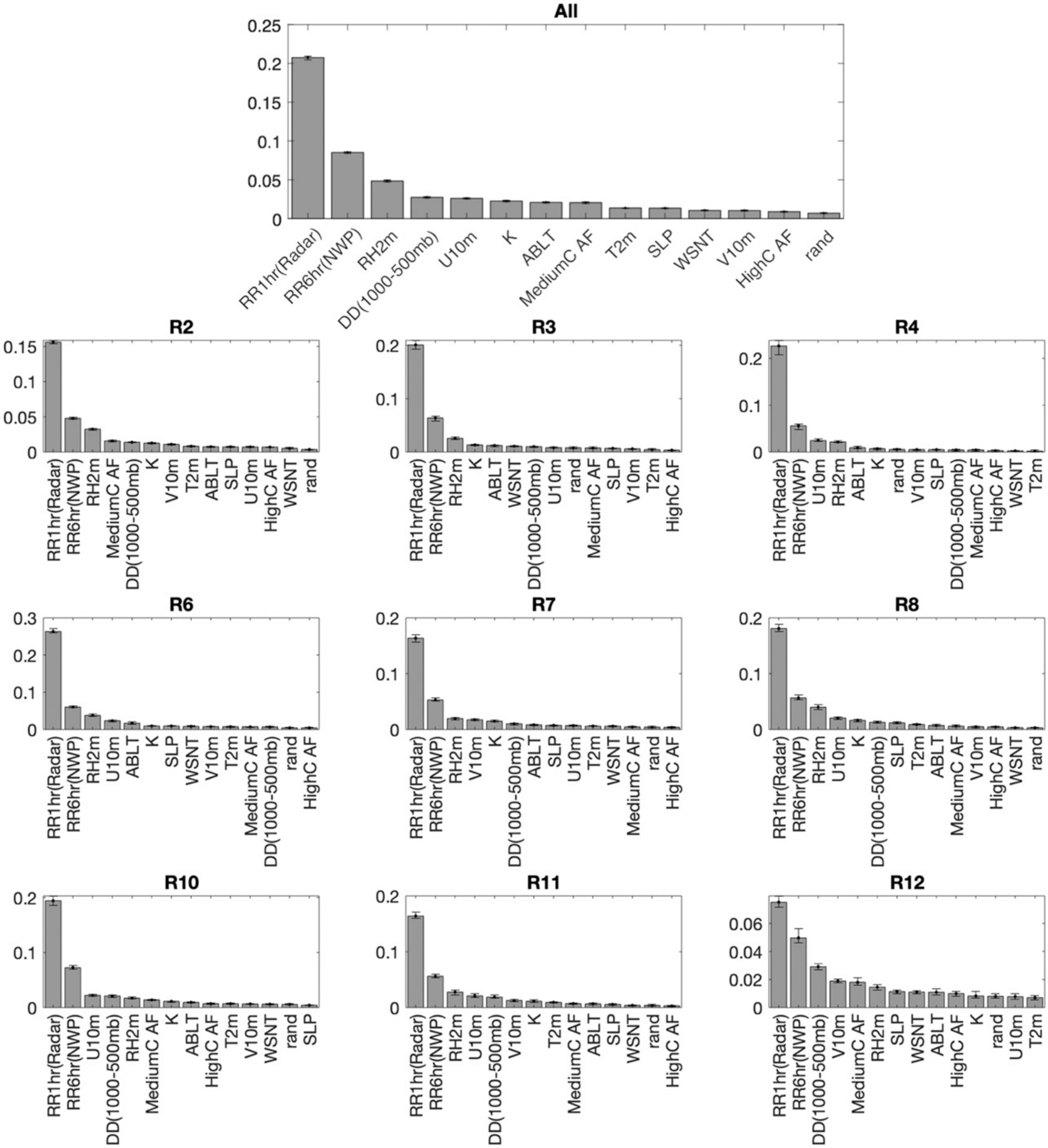

FIG. 8. Ranking importance by permutations of uncorrelated predictors for the RF models trained by using the training datasets from (top) all stations and (others) from stations in each region. The error bars indicate the range between the minimum and maximum values of importance ranking obtained by repeating the process 10 times.

Table 1 without strong collinearity were chosen as RR1hr(Radar), RR6hr(NWP), SLP, T2m, MediumC AF, HighC AF, RH2m, U10m, V10m, ABLT, WSNT, DD(1000–500mb), and *K*. Moreover, a sequence of random numbers denoted as rand was added as a reference predictor [as explained in section 3c(3)]. These predictors were used to train RF models, and Fig. 8 shows the ranking of predictor importance by permutation of these predictors.

The results indicate that the most and second most important predictors were precipitation predictions by the radar nowcasts and NWP. Specifically, RR1hr(Radar) was at least 2.7 times more important than that of RR6hr(NWP),

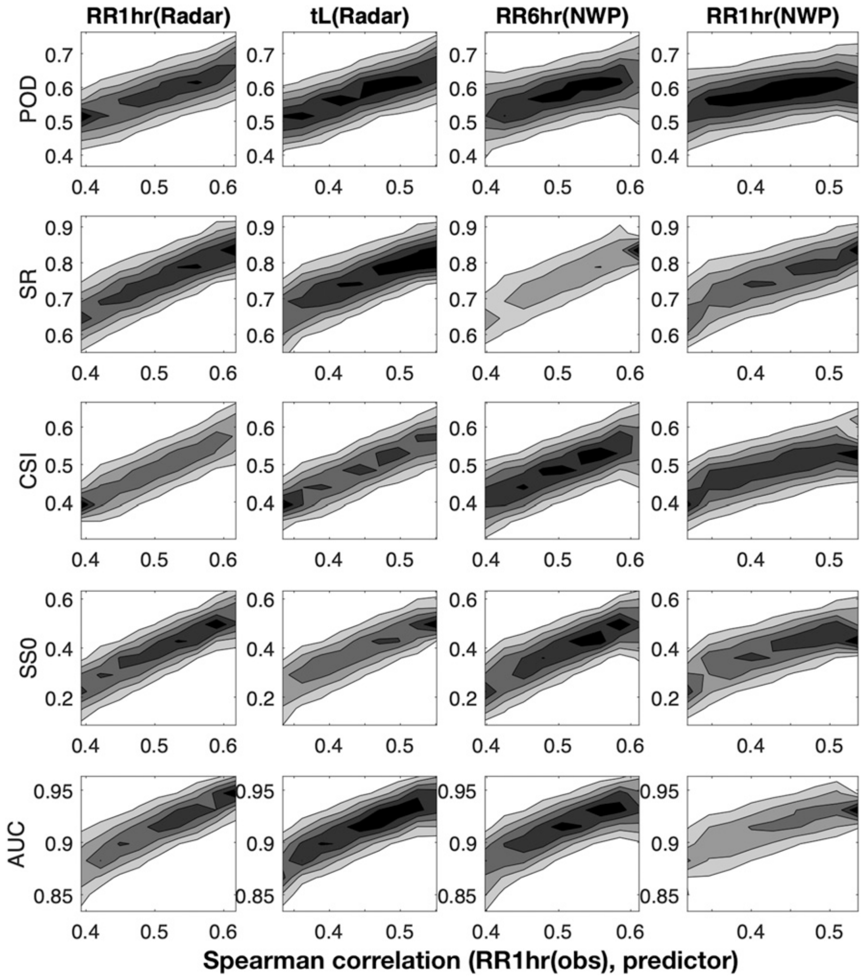

FIG. 9. Probability density functions (PDF) of various metrics (POD, SR, CSI, SS0, and AUC) conditioned on the Spearman correlations between the precipitation observations and the predictors: (first column) RR1hr(Radar), (second column) tL(Radar), (third column) RR6hr(NWP), and (fourth column) RR1hr(NWP). The full names of the predictors are listed in Table 1. The PDF is derived from a numerical experiment based on bootstrapping as described in section 3c(3).

and RR1hr(Radar) was at least 4.5 times more important than that of the third variable for most training cases except the model trained by data in region 12. Notably, the differences in importance among the first three variables were less pronounced for the model trained by data in region 12.

2) SECOND APPROACH: SPEARMAN CORRELATION BETWEEN THE PRECIPITATION OBSERVATIONS AND PREDICTORS

This study has also asked whether meteorological variables other than precipitation output from the NWP (AROME) have positive influences on the predictive skills of the RF model, despite their low importance. To answer this question, bootstrapping was used [section 3c(3)] to qualitatively examine how the covariation between each predictor and the precipitation observation can be related to the metrics of the RF model.

The metrics POD, SR, CSI, SS0 and AUC were strongly correlated to the Spearman correlation coefficients between the precipitation observations and predictions by the radar and AROME [i.e., RR1hr(Radar), tL(Radar), RR6hr(NWP), and RR1hr(NWP)] as shown in Fig. 9. Some of these metrics were also influenced by the Fog AF, LowC AF, MediumC AF, RH2m, U10m, ABLT, DD(1000–500 mb), and K (Fig. 10). However, the influences of these variables were less pronounced than RR1hr(Radar), tL(Radar), RR6hr(NWP), and RR1hr(NWP). Some of these variables belonged to the same cluster of strong collinearity shown in section 4c(1).

Some variables showed no apparent relationship between the metrics of the RF model and the Spearman correlation

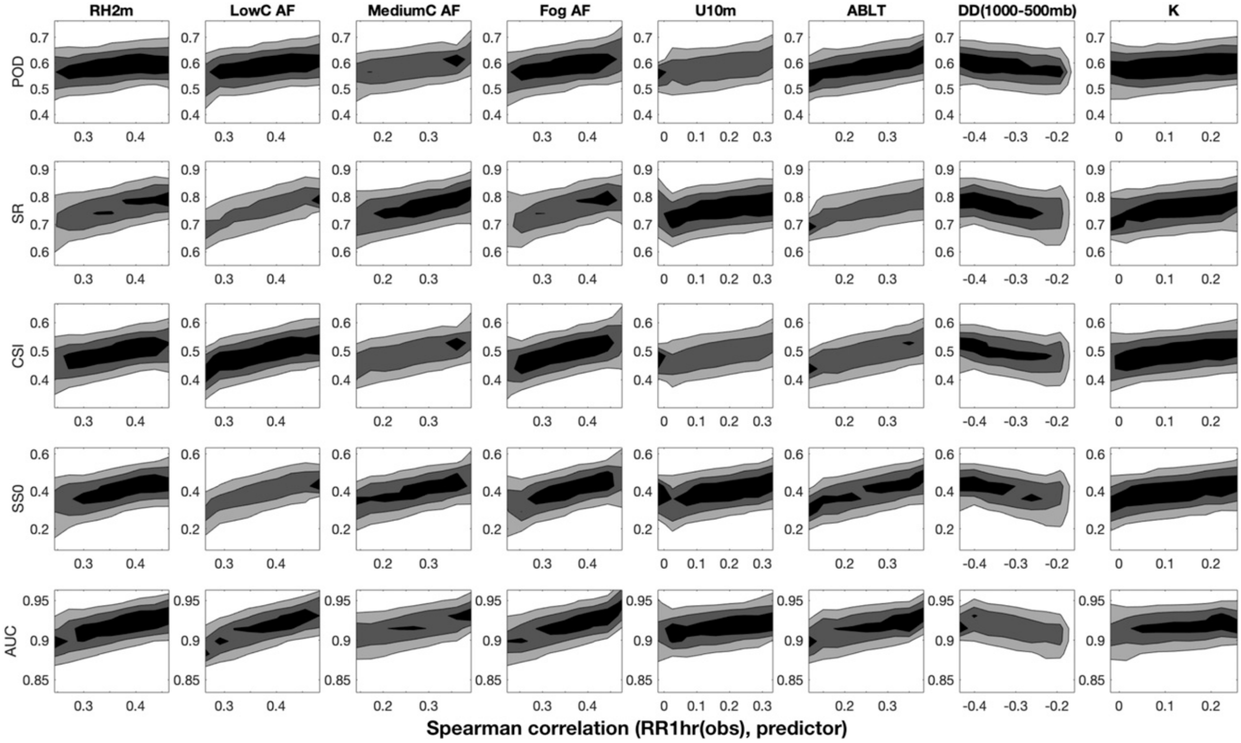

FIG. 10. As in Fig. 9, but for the predictors labeled at the top of each column.

coefficients between the variable and precipitation observation. These variables were the T2m, SLP, WSNT, V10m, and HighC AF (Fig. 11). These variables also had the lowest rankings in importance besides the random variable for the RF All model, as shown in the top panel of Fig. 8.

3) THIRD APPROACH: PARTITION OF PREDICTORS AND TEST DATA

A third approach has been used to study overall, whether additional meteorological variables not directly related to precipitation contribute to the predictive skills of the RF model. The results of the third approach are shown in Fig. 12. All predictors listed in Table 1 have been divided into three subsets i, ii, and iii as described in section 3c(3). When all test data were used, the RF model with only predictors of subset iii (no precipitation) resulted in the lowest predictive skill. The RF model with predictors of subsets i and ii (only precipitation) had the highest predictive skill among all RF models with subsets of predictors. However, its predictive skill was still lower than the RF model with all predictors (Fig. 12a). The RF model with NWP only predictors ii and iii was slightly better than the one with radar nowcast only predictors of subset i. Overall, Fig. 12a shows that the metrics of the RF model were improved by including subset iii (weather variables other than precipitation) in addition to predictors from subsets i and ii. However, the improvement was limited, generally less than 15%.

The test data subsets, Test 1 and Test 2, represented the best and worst scenarios of the precipitation predictions by the radar nowcasts and NWP (Fig. 12b). In particular, Test 2 represented the situation when both the radar nowcast and NWP failed to predict precipitation. In this situation, the RF model that included subsets iii as predictors [i.e., RF(iii), RF(ii), (iii)] as predictors could not improve POD. However, the SR was improved by approximately 55% from that of the RF model with only subsets i and ii as predictors.

d. Comparing precipitation predictions by RF with NWP and radar nowcast

The bottom panel of Fig. 13 shows the comparison of precipitation predictions by the two benchmarks: the radar nowcasts and NWP with the RF All model tested on the dataset consisting of test data from all stations. Overall, the metrics (POD, SR, CSI, and SS0) for evaluating the RF model exceeded those for the radar nowcasts and NWP. The increases of the metrics by using the RF model from the two benchmarks were generally moderate except for the SS0. In particular, the RF model increased the value of SS0 from −0.08 (NWP) to 0.39. The comparison of bias also indicates that NWP overpredicted precipitation whereas the RF model and radar nowcasts underpredicted precipitation.

The top and middle panels of Fig. 13 show the comparisons of metrics tested on the bootstrapped samples [section 3c(3)] and test datasets of individual stations. The results further demonstrate that 1) better predictive skills of the radar nowcasts and NWP corresponded to better performance of

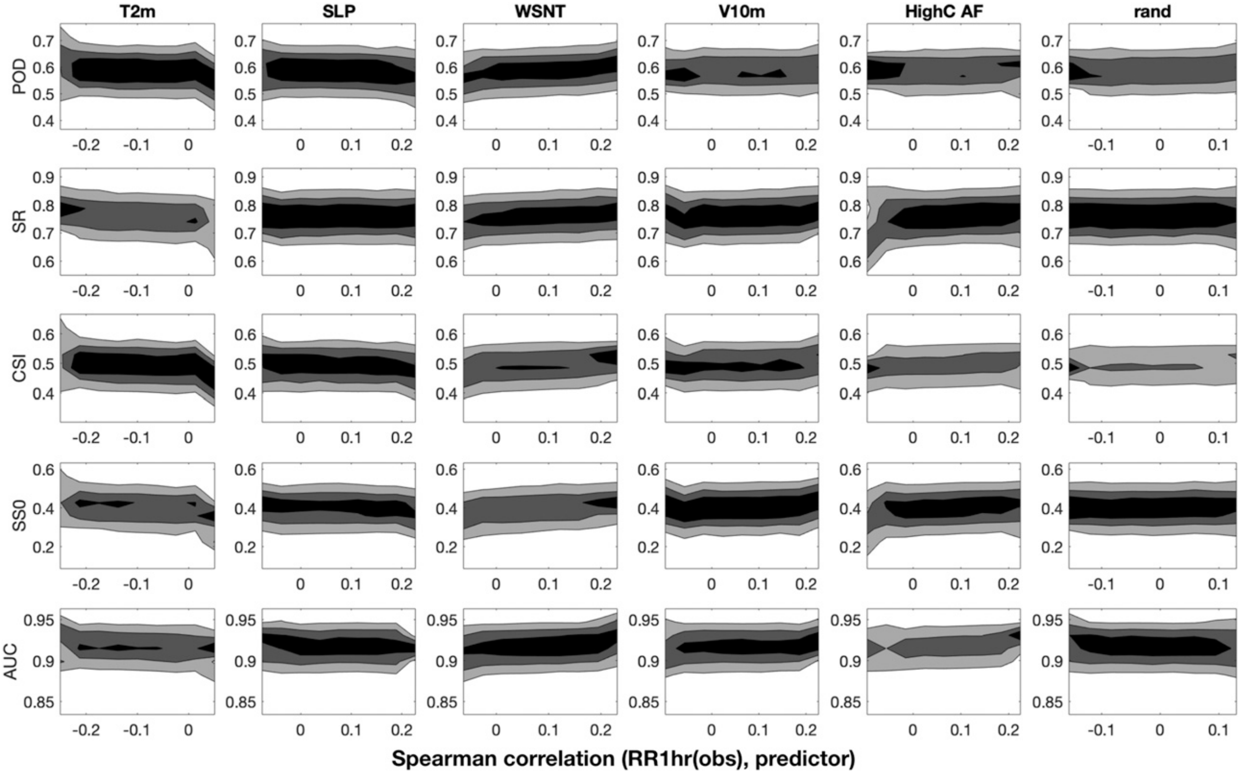

FIG. 11. As in Fig. 9, but for the predictors labeled at the top of each column.

the RF model in general, and 2) the metrics of the RF model (POD, SR, CSI, and SS0) exceeded those of the radar nowcasts and NWP for most cases, but the improvement in POD was less evident than other metrics. Moreover, the comparison of bias indicates that the RF model underpredicted precipitation (i.e., bias < 1) for approximately 87% of all stations. The radar nowcasts underpredicted precipitation for about 75% of all stations. However, the NWP overpredicted precipitation (bias > 1) for 95% of all stations.

5. Discussion and conclusions

This study examines some potential factors related to training an RF model for processing the precipitation predictions from radar nowcasts and NWP. These factors are 1) typical strategies for addressing imbalanced datasets for binary classification (i.e., lowering the threshold probability for classification, resampling the training datasets), 2) the geographic diversity and seasonality of the training datasets, and 3) the choice of predictors.

The results in section 4a suggest that neither resampling the training dataset nor decreasing the threshold probability from $p = 0.5$ can improve the predictive skill of the RF model. Both resampling and lowering the threshold probability redistribute the test cases belonging to hit, false alarm, miss, and correct rejection from that of No RS ($p = 0.5$) by increasing the frequency of the minority predictions. The positive contributions (more hit and less miss) resulting from the increased minority predictions are always counterbalanced by the negative contributions (mostly the increased false alarms and decreased correct rejections).

Furthermore, OS achieves the redistribution of the elements of the contingency table by adding repetitive predictive information from the training data of the minority class, whereas US deletes predictive information from the training data of the majority class. The added and deleted training cases contain either correct or incorrect information for prediction. Therefore, the predictive skills of OS and US do not necessarily exceed those of the original training dataset (No RS) as shown by the results of the study. Figure 3 shows that the SS0 of OS was approximately three times that of US, and Fig. 4 shows that US altered the distribution of the four elements of the contingency table more than OS did. The results suggest that deleting correct predictions can have more negative influences on predictive skills than adding incorrect predictions.

The results of section 4b suggest that the differences among the All model, the Region model, and the Local model are not large enough to make any one of them noticeably more or less accurate. Therefore, the geographic diversity of the training dataset is not a decisive factor in influencing the predictive skills of the RF model. This suggests that training one model for a large region may not reduce the quality of the forecast comparing to training separate models for separate regions or individual stations.

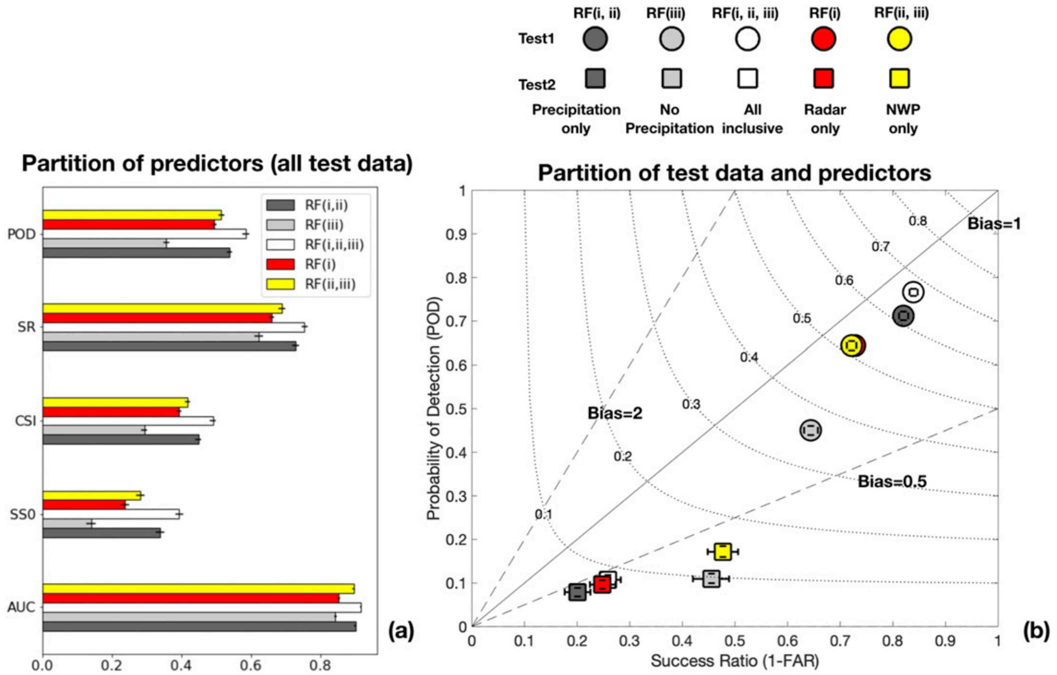

FIG. 12. The predictors are partitioned into three subsets: (i) the precipitation predictions by the radar nowcasts (1 and 2 in Table 1), (ii) the precipitation predictions by NWP (3 and 4 in Table 1), (iii) the variables output from NWP other than precipitation (5–17 in Table 1). (left) The metrics (POD, SR, CSI, SS0, and AUC) for evaluating the RF models with different combinations of predictor subsets i, ii, iii tested on the dataset consisting of test data from all stations. (right) Various metrics for evaluating the same set of RF models as in the left panel, but tested on two complementary subsets of all test data (Test 1 and Test 2). The precipitation observations in Test 1 are correctly predicted by either the radar nowcasts or NWP and are misclassified by both the radar nowcasts and NWP in Test 2. Error bars: the empirical 90% bootstrap confidence intervals.

The results of section 4c suggest that the predictive skill of the RF model is driven by the skills of precipitation predictions from the radar nowcasts and NWP. Although the RF model is not as influenced by meteorological variables other than precipitation, the results of section 4c(3) suggest that additional meteorological variables not directly related to precipitation can help to improve the RF model by reducing the false alarm ratio thereby increasing the success ratio of predicting precipitation.

The seasonal variations of RF models' predictive skills generally correspond to the strength of the Spearman correlations between the radar/NWP prediction and observations in different seasons (Fig. 6). For NWP, it is well known that summer weather events are more likely to be attributed to local-scale convections, which may not be well resolved by numerical weather models. Also, because the summer precipitation events caused by local-scale convections are often short-lived, extrapolation of radar echoes may not respond quickly enough to track and predict the precipitation system. Therefore, predictive skills are better in cold seasons (winter and fall) and worse in warm seasons (summer and spring) in most regions, but region 12 is a notable exception. The results suggest that both radar and NWP in this region perform poorly during winter, even worse than in summer. One potential reason could be that the region is relatively dry during winter comparing to other regions. The uncertainty of precipitation detection by radar and NWP can be greater with less frequent precipitation events.

It is also noticeable that the CSI values of radar nowcasts in nearly all regions exceed that of NWP, except in region 12 where the CSI of radar nowcasts is only around one-third of that of NWP (Fig. 5). Moreover, the difference in importance between the top two predictors (precipitation predictions by radar nowcasts and NWP) is less pronounced in region 12 than other regions (Fig. 8), which also suggests that the predictive skills of radar nowcasts in region 12 is lower than in other regions.

The above results suggest that radar nowcasts in region 12 are abnormal and could be outliers. In general, radar has exceptional advantages over other observing systems in nowcasting precipitation because it collects the information of precipitation particles in three dimensions with a high spatial and temporal resolution (Wang et al. 2017). However, rainfall estimation by radar is also subject to errors of various sources, such as beam shielding, ground clutter, anomalous propagation (Testik and Gebremichael 2010). The unusually low CSI values and less prominent importance score of radar nowcasts in region 12 suggest that radar measurements used for

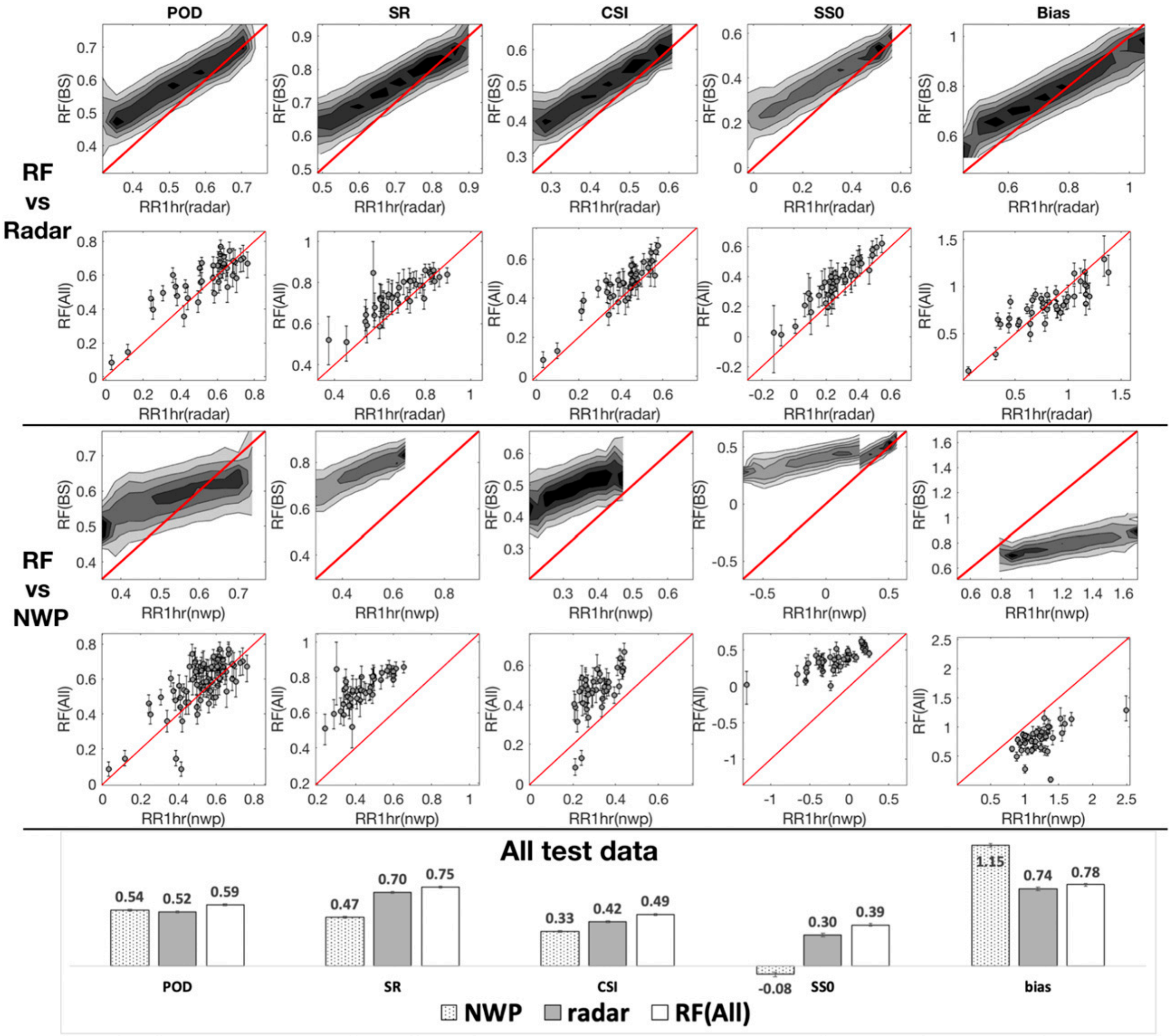

FIG. 13. (bottom) Metrics (POD, SR, CSI, SS0, and bias) evaluating the precipitation predictions by the radar nowcasts, NWP, and the RF(All) model trained by using training data from all stations. The metrics are for test data from all stations. The top and middle panels show the comparisons of the metrics for evaluating the precipitation predictions by the radar nowcasts/NWP and the RF(All) tested on the bootstrapped samples (the filled contour) and test data from each station (the scatterplot), respectively. The error bars indicate the empirical 90% bootstrap confidence intervals.

nowcasting precipitation in region 12 are subject to some systematic errors. Further investigations are needed to address the issue.

In conclusion, there are robust improvements in most of the verification measures tested in the study as shown in section 4d. However, it is beneficial to further improve the RF model by identifying additional useful predictors other than those examined in this study in case of low predictive skills of radar nowcasts and NWP, such as in region 12.

Acknowledgments. This research was supported by the Geophysical Institute of the University of Bergen, Norway and the Norwegian Meteorological Institute.

Data availability statement. The data of the predictand and predictors (section 2b) for building the RF model and the geographic information of the 41 stations (Fig. 1) used in the study are available at http://dx.doi.org/10.17632/smxkyhtdvj.3 (Mao 2020).

REFERENCES

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Chambolle, A., and T. Pock, 2011: A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, **40**, 120–145, https://doi.org/10.1007/s10851-010-0251-1.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002: SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357, https://doi.org/10.1613/jair.953.

Dixon, M., and G. Wiener, 1993: TITAN: Thunderstorm identification, tracking, analysis, and nowcasting—A radar-based methodology. *J. Atmos. Oceanic Technol.*, **10**, 785–797, https://doi.org/10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2.

George, J. J., 2014: *Weather Forecasting for Aeronautics*. Academic Press, 684 pp.

Germann, U., and I. Zawadzki, 2002: Scale dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Wea. Rev.*, **130**, 2859–2873, https://doi.org/10.1175/1520-0493(2002)130<2859:SDOTPO>2.0.CO;2.

Guo, X., Y. Yin, C. Dong, G. Yang, and G. Zhou, 2008: On the class imbalance problem. *2008 Fourth Int. Conf. on Natural Computation*, Jinan, China, IEEE, 192–201, https://doi.org/10.1109/ICNC.2008.871.

Hanssen-Bauer, I., and P. Nordli, 1998: Annual and seasonal temperature variations in Norway 1876-1997. DNMI Rep. 25, 98 pp.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Science & Business Media, 763 pp.

Hwang, Y., A. J. Clark, V. Lakshmanan, and S. E. Koch, 2015: Improved nowcasts by blending extrapolation and model forecasts. *Wea. Forecasting*, **30**, 1201–1217, https://doi.org/10.1175/WAF-D-15-0057.1.

Inness, P. M., and S. Dorling, 2012: *Operational Weather Forecasting*. John Wiley & Sons, 248 pp.

Japkowicz, N., and S. Stephen, 2002: The class imbalance problem: A systematic study. *Intell. Data Anal.*, **6**, 429–449, https://doi.org/10.3233/IDA-2002-6504.

Li, L., W. Schmid, and J. Joss, 1995: Nowcasting of motion and growth of precipitation with radar over a complex orography. *J. Appl. Meteor.*, **34**, 1286–1300, https://doi.org/10.1175/1520-0450(1995)034<1286:NOMAGO>2.0.CO;2.

Liu, Z., G. Gilbert, J. M. Cepeda, A. O. K. Lysdahl, L. Piciullo, H. Hefre, and S. Lacasse, 2020: Modelling of shallow landslides with machine learning algorithms. *Geosci. Front.*, https://doi.org/10.1016/j.gsf.2020.04.014, in press.

Longadge, R., and S. Dongre, 2013: Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707.

Mandapaka, P. V., U. Germann, L. Panziera, and A. Hering, 2012: Can Lagrangian extrapolation of radar fields be used for precipitation nowcasting over complex alpine orography? *Wea. Forecasting*, **27**, 28–49, https://doi.org/10.1175/WAF-D-11-00050.1.

Mao, Y., 2020: Random forest nowcasts. Accessed 19 May 2020, https://doi.org/10.17632/smxkyhtdvj.3.

MathWorks, 2019: Predict responses using ensemble of bagged decision trees-matlab-mathworks nordic. Accessed 1 December 2019, https://se.mathworks.com/help/stats/treebagger.predict.html.

Mosavi, A., P. Ozturk, and K. Chau, 2018: Flood prediction using machine learning models: Literature review. *Water*, **10**, 1536, https://doi.org/10.3390/w10111536.

Müller, M., and Coauthors, 2017: AROME-MetCoOp: A nordic convective-scale operational weather prediction model. *Wea. Forecasting*, **32**, 609–627, https://doi.org/10.1175/WAF-D-16-0099.1.

Norwegian Meteorological Institute, 2011: Free access to weather- and climate data from Norwegian meteorological institute from historical data to real time observations. eKlima, accessed 1 May 2018, https://www.met.no/en/free-meteorological-data/Download-services.

Nurmi, P., 2003: Recommendations on the verification of local weather forecasts. ECMWF Tech. Memo. 430, 19 pp., https://www.ecmwf.int/en/elibrary/11401-recommendations-verification-local-weather-forecasts.

Reyniers, M., 2008: Quantitative precipitation forecasts based on radar observations: Principles, algorithms and operational systems. Royal Meteorological Institute of Belgium Publ. Scientifique et Technique 52, 62 pp.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Scikit-Learn Developers, 2019: Permutation importance with multicollinear or correlated features. Accessed 1 December 2019, https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *NIPS'15: Proc. 28th Int. Conf. on Neural Information Processing Systems*, Cambridge, MA, NIPS, 802–810, https://dl.acm.org/doi/10.5555/2969239.2969329.

Testik, F. Y., and M. Gebremichael, 2010: *Rainfall: State of the Science*. Wiley Online Library, 287 pp.

Wang, Y., and Coauthors, 2017: Guidelines for nowcasting techniques. WMO Publ. 1198, 82 pp., https://library.wmo.int/doc_num.php?explnum_id=3795.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Zou, H., S. Wu, J. Shan, and X. Yi, 2019: A method of radar echo extrapolation based on TREC and Barnes filter. *J. Atmos. Oceanic Technol.*, **36**, 1713–1727, https://doi.org/10.1175/JTECH-D-18-0194.1.