remote sensing

MDPI

Article

Precipitation Estimation Using FY-4B/AGRI Satellite Data Based on Random Forest

Yang Huang [1,2,†], Yansong Bao [1,2,†], George P. Petropoulos [3], Qifeng Lu [4,*], Yanfeng Huo [5] and Fu Wang [4]

[1] Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, CMA Key Laboratory for Aerosol-Cloud-Precipitation, Nanjing University of Information Science & Technology, Nanjing 210044, China; 202212050019@nuist.edu.cn (Y.H.); ysbao@nuist.edu.cn (Y.B.)
[2] School of Atmospheric Physics, Nanjing University of Information Science & Technology, Nanjing 210044, China
[3] Department of Geography, Harokopio University of Athens, EI. Venizelou 70, Kallithea, 17671 Athens, Greece; gpetropoulos@hua.gr
[4] Earth System Modeling and Prediction Center, China Meteorological Administration, Beijing 100081, China; wangfu@cma.gov.cn
[5] Anhui Institute of Meteorological Sciences, Hefei 230031, China; huoyanfeng@ahmi.org.cn
* Correspondence: luqf@cma.gov.cn
† These authors contributed equally to this work.

Abstract: Precipitation is the basic component of the Earth's water cycle. Obtaining high-resolution and high-precision precipitation data is of great significance. This paper establishes a precipitation retrieval model based on a random forest classification and regression model during the day and at night with FY-4B/AGRI Level1 data on China from July to August 2022. To evaluate the retrieval effect of the model, the GPM IMERG product is used as a reference, and the retrieval results are compared against those of the FY-4B/AGRI operational precipitation product. In addition, the retrieval results are analyzed according to different underlying surfaces. The results showed that compared with the FY-4B/AGRI operational precipitation product, the retrieval model can better identify precipitation and capture precipitation areas of light rain, moderate rain, heavy rain and torrential rain. Among them, the probability of detection (POD) of the day model increased from 0.328 to 0.680, and the equitable threat score (ETS) increased from 0.252 to 0.432. The POD of the night model increased from 0.337 to 0.639, and the ETS score increased from 0.239 to 0.369. Meanwhile, the precipitation estimation accuracy of the day model increased by 38.98% and that of the night model increased by 40.85%. Our results also showed that due to the surface uniformity of the ocean, the model can identify precipitation better on the ocean than on the land. Our findings also indicated that for the different underlying surfaces of the land, there is no significant difference in each evaluation index of the model. This is a strong argument for the universal applicability of the model. Notably, the results showed that, especially for more vegetated areas and areas covered by water, the model is capable of estimating precipitation. In conclusion, the precipitation retrieval model that is proposed herein can better determine precipitation regions and estimate precipitation intensities compared with the FY-4B/AGRI operational precipitation product. It can provide some reference value for future precipitation retrieval research on FY-4B/AGRI.

Keywords: FY-4B/AGRI; random forest; precipitation retrieval; underlying surfaces

Citation: Huang, Y.; Bao, Y.; Petropoulos, G.P.; Lu, Q.; Huo, Y.; Wang, F. Precipitation Estimation Using FY-4B/AGRI Satellite Data Based on Random Forest. *Remote Sens.* **2024**, *16*, 1267. https://doi.org/10.3390/rs16071267

Academic Editor: Kenji Nakamura

Received: 16 February 2024
Revised: 29 March 2024
Accepted: 1 April 2024
Published: 3 April 2024

1. Introduction

As the most active element of the atmosphere, precipitation is the basic component of the Earth's water cycle and plays an indispensable role in the atmospheric process for various space–time scales [1]. China is located in the East Asian monsoon area. The uneven distribution of annual variations in summer wind and the time and space distribution of precipitation can easily lead to the occurrence of droughts and floods, causing serious

losses to China's industrial and agricultural production and economy [2]. Therefore, obtaining high-resolution and high-precision precipitation data is of great significance to study surface water cycle processes, monitor natural disasters and simulate the watershed hydrology process.

At present, there are three main methods for the acquisition of precipitation data: the first is ground rainfall stations' observations, the second is meteorological radar observations and the third is satellite remote sensing observations [3]. The observations of ground rainfall stations are the most direct way to obtain precipitation data, which can accurately reflect the precipitation at a certain point. However, this requires the deployment of very dense rain gauges to accurately observe the characteristics of rainfall in a region [4]. Meteorological radar observations can obtain precipitation data over large areas and with a high level of spatiotemporal accuracy. However, the observation range of precipitation by a single radar is limited, as the accuracy of rainfall measurement decreases with the increase in distance from the radar, and factors such as clutter, beam blockage and abnormal propagation also need to be considered [5]. Satellite remote sensing utilizes satellites acquiring visible light, infrared and passive microwave radiation information to make indirect observations of precipitation. It has the advantage of a high spatiotemporal resolution and wide coverage area and can capture rapidly changing weather phenomena in a short time [6].

Early retrieval based on satellite optical data relies on the relationship between infrared cloud top temperature and rainfall probability and intensity, that is, a high cloud top indicates strong convection and heavy precipitation [7]. This retrieval method is suitable for deep convective precipitation processes. These deep convective clouds can be easily identified in infrared or water vapor channels [8]. Yet, stratus precipitation processes caused by frontal cyclones at mid-latitudes show considerable deficiencies [9] and are characterized by relatively uniform and warm cloud top temperatures. There is little difference between precipitation and non-precipitation regions. Thus, retrieval techniques based only on the infrared cloud top temperature will lead to an underestimate of the detected precipitation area [10]. Therefore, for the retrieval of warm cloud precipitation with a relatively low cloud top height, especially stratospheric cloud precipitation, it is necessary to consider the physical properties of water droplets or ice particles in the cloud [11]. Thus, the multi-channel threshold method was later developed to extract cloud radiation characteristics and physical parameters from satellite multi-spectral data to further improve the precipitation retrieval algorithm. However, these threshold algorithms rely heavily on the parametric relationship between cloud physical properties and the precipitation process. Based on this theoretical background, only a few variables are used, and there is actually a nonlinear relationship between remote sensing information and precipitation, so it is difficult to improve the accuracy of the algorithm [12–15].

With the improvement of computing power, artificial intelligence has ushered in a wave of development. Machine learning (ML) is a method of data prediction after training a model with a large amount of data as its input. Quantitative precipitation retrieval using ML methods can supplement traditional physical driving methods and is also a reliable and effective way to further improve the accuracy of satellite precipitation retrieval [16]. A number of scholars have carried out research in this area. For instance, Kuhnlein et al. [17] classified convective and stratiform precipitation regions during the day, at night and at twilight using the random forest (RF) method, further improving the accuracy of their precipitation estimations. Lazri et al. [18] built a combined model of support vector machine (SVM), artificial neural network (ANN) and random forest (RF) classifiers to improve the classification of convective precipitation and stratiform precipitation. The developed scheme was superior to the different classifiers used alone, reaching a total classification accuracy of 97.40%. Ma et al. [19] extracted some features related to precipitation based on Himawari-8 satellite data and topographic height data in East Asia by using a gradient decision tree. Their results showed the algorithm had a higher hit rate and a lower false alarm rate as a whole. In another study, Min et al. [20] established an RF classification

and regression model and adopted sample balance technology to optimize it. The results showed the precipitation fall area and intensity were basically consistent with those of the GPM product. Hirose et al. [21] used an RF to estimate precipitation based on Himawari-8 satellite data, which have a higher estimation accuracy for heavy rain from warm precipitation clouds. Kong et al. [22] used an SVM and the quadratic curve-fitting statistical analysis method to estimate precipitation based on Himawari-8 satellite data. The results of this study showed the SVM retrieval results were superior to those of the quadratic curve-fitting analysis. Wang et al. [23] used the dictionary learning and regularization constraint methods to estimate precipitation. Their findings showed the retrieval precipitation and GPM precipitation to have a good degree of similarity. Zhang et al. [24] used an RF to build a retrieval model to estimate precipitation levels based on Himawari-8 satellite data and the GFS numerical forecast product. Their results showed that the RF could describe the precipitation contour well. Guan et al. [25] used an RF to retrieve precipitation data based on FY-4A/AGRI Level1 data. Their results showed the RF had a higher accuracy than that of the AGRI's precipitation product during the day and at night. In general, ML methods have a higher accuracy in estimating precipitation than traditional physical methods. But there are still some problems that we need to solve. There are still numerous errors for heavy rain and torrential rain, and there is no research on precipitation retrieval for the whole region of China based on FY-4B/AGRI satellite data currently. As we all know, ML is a black-box-like algorithm, and it is difficult to determine causal relationships. What can we do to improve ML to make it better? Firstly, we should select more appropriate model feature variables as the input to models. Secondly, an uneven sample distribution will lead an ML model to often overestimate the majority of samples and underestimate the minority of samples, which is more obvious in precipitation samples. Thus, we should select the optimal sample proportion when training the model. And thirdly, because of the complex and diverse terrain of China, it is necessary to evaluate the precipitation retrieval effect of different underlying surfaces.

In order to optimize the FY-4B/AGRI precipitation estimation algorithm and improve the accuracy of satellite precipitation retrieval, this paper establishes the FY-4B/AGRI precipitation estimation algorithm based on an RF. The algorithm is based on the analysis of the relationship between FY-4B/AGRI observation data and precipitation, as well as the infrared signal characteristics of precipitation clouds. Thus, more feature variables related to clouds and precipitation are considered in the input of the model. To evaluate the model predictions, the GPM IMERG product is used as a reference and the retrievals are compared with those of the FY-4B/AGRI operational precipitation product to evaluate the retrieval effect of the model. At the same time, the retrieval results are compared and analyzed according to different underlying surfaces.

2. Materials and Methods

2.1. Materials

2.1.1. FY-4B/AGRI Level1 Data

FY-4B is the first operational satellite of FY-4. The Advanced Geostationary Radiation Imager (AGRI) is one of the main payloads of the FY-4B. FY-4B/AGRI has four full-disk observations per hour and ninety-five full-disk observations per day (at 14:15 UTC, the satellite undergoes maintenance, so no observations are made). FY-4B/AGRI has a total of fifteen channels covering a wavelength range of 0.45~13.6 μm, including six visible/near-infrared channels (1–6 channels) and nine infrared channels (7–15 channels), as shown in Table 1. The range of FY-4B/AGRI Level1 data used in this study is 73°E~135°E, 3°N~60°N (including Chinese mainland and adjacent to the Bohai Sea, the Yellow Sea, the East China Sea and the South China Sea on the edge of Chinese mainland).

Table 1. FY-4B/AGRI specifications.

Band	Central Wavelength (μm)	Spectral Bandwidth (μm)	Spatial Resolution (km)	Main Applications
1	0.47	0.45~0.49	1	Visibility, Aerosol
2	0.65	0.55~0.75	0.5	Visibility, Vegetation
3	0.825	0.75~0.90	1	Vegetation, Aerosol
4	1.379	1.371~1.386	2	Cirrus cloud
5	1.61	1.58~1.64	2	Cloud/Snow, Water cloud/Ice cloud
6	2.225	2.10~2.35	2	Cirrus cloud, Aerosol
7	3.75	3.50~4.00 (high)	2	Cloud, Fire point
8	3.75	3.50~4.00 (low)	4	Earth's surface
9	6.25	5.80~6.70	4	Upper-level water vapour
10	6.95	6.75~7.15	4	Mid-level water vapour
11	7.42	7.24~7.60	4	Lower-level water vapour
12	8.55	8.3~8.8	4	Cloud
13	10.80	10.30~11.30	4	Cloud, Surface temperature
14	12.00	11.50~12.50	4	Cloud, Total water vapor
15	13.3	13.00~13.60	4	Cloud

2.1.2. GPM IMERG Product

The Integrated Multi-satellitE Retrieval for GPM (IMERG) is the latest generation of multi-satellite fusion retrieval precipitation data designed specifically for the Global Precipitation Measurement Mission (GPM) [26]. IMERG is based on the mutual calibration and inversion of microwave (PMW), infrared (IR) and other precipitation observations with a high spatial and temporal resolution for GPM satellite constellation. Thus, it is the best global algorithm currently available. IMERG offers three products: Early Run (ER), Late Run (LR), and Final Run (FR), covering the latitude range of 90°N to 90°S. Among them, FR is released from GPM IMERG after a delay of approximately 3.5 months, using monthly precipitation analysis from ground observation stations for its calibration, which is mainly applied in scientific research [27]. Research shows GPM IMERG FR product can accurately capture the precipitation distribution characteristics of Chinese mainland on the whole [28]. In this study, FR product is used as reference data for precipitation retrieval, with a spatial resolution of 0.1° and a temporal resolution of 30 min.

2.1.3. FY-4B/AGRI Operational Precipitation Product

The operational precipitation product is the pure satellite estimated precipitation results, generated using FY-4B/AGRI infrared channel data, without ground rain gauge correction [29]. The FY-4B/AGRI Level1 data at each time step are inverted to obtain the corresponding precipitation product; thus, the temporal resolution of the precipitation product is also 15 min and the spatial resolution is 4 km.

2.1.4. Topographic Data

ETOPO2v2 is a global topographic model developed by the National Geophysical Data Center (NGDC), part of the National Oceanic and Atmospheric Administration (NOAA). It includes the topography of the world's land and ocean, with a spatial resolution of 0.03°.

2.1.5. Land Cover Type Data

These data were released by Liu Liangyun's team of the Institute of Aerospace Information Innovation, Chinese Academy of Sciences. All Landsat satellite data from 1984 to 2020 are used to produce a global 30 m fine land cover dynamic monitoring product from 1985 to 2020. The product follows the classification system of baseline data in 2020, including 29 land cover types, with an update cycle of 5 years and a spatial resolution of 30 m [30].

2.2. Data Pre-Processing

This study classifies the land cover types of the study area into 6 categories according to the Classification of Land Use Status Quo [31]: farmland (paddy field and dry land), woodland (natural woodland, artificial gardens and shrubland, etc.), grassland (natural or semi-natural herbaceous vegetation), bare land (natural bare land and uncultivated land after harvest), artificial surfaces (urban and rural residential sites, industrial mines and roads, etc.) and water bodies (rivers and lakes, etc.), as shown in Table 2.

Table 2. Land cover types and corresponding labels.

Land Cover Types	Labels
Farmland	10, 20
Woodland	12, 51, 52, 61, 62, 71, 72, 81, 82, 91, 92, 120, 121, 122
Grassland	11, 130, 140, 150, 152, 153
Bare land	200, 201, 202
Artificial surfaces	190
Water bodies	210

At the same time, FY-4B/AGRI Level1 data, FY-4B/AGRI operational precipitation product, topographic data, land cover type data and GPM IMERG product are matched in time and space to establish spatio-temporal matching dataset.

Based on the space–time of the GPM IMERG product, the average value of FY-4B/AGRI Level1 data is matched to FY-4B/AGRI operational precipitation product in the corresponding period ((1) time matching). For example, the GPM IMERG product in 0900UTC-0930UTC corresponds to the average value of two satellite images in this period. Next, the nearest satellite pixel within 4 km is searched ((2) spatial matching of satellite data). Thirdly, data within 4 km are searched and the mean value is taken to represent the central point information ((3) spatial matching of topographic data). Lastly, the data within 4 km are searched and the mode is taken to represent the central point information ((4) spatial matching of land cover type data). Since the channels of visible and near-infrared wavelengths are not available at night, the dataset is divided according to the solar zenith angle (SZ) into the daytime dataset ($SZ < 85^\circ$) and the night dataset ($SZ \geq 85^\circ$) [32].

2.3. Methods

2.3.1. Overall Technical Route

Based on the FY-4B/AGRI Level1 data of China from July to August 2022, the precipitation retrieval model is established to estimate precipitation using RF. The algorithm's technical process is shown in Figure 1. Firstly, the FY-4B/AGRI Level1 data, topographic data and GPM IMERG product are pre-processed and the spatio-temporal matching dataset is established. Then, the dataset is divided into day and night datasets according to the SZ. The dataset of July 2022 is used for modelling (day dataset: 15,921,868; night dataset: 109,100,161), and 1/4 of the dataset is randomly taken as the training dataset and 3/4 as the validation dataset. In order to test the time extension of the model, the dataset of August 2022 is used as an independent testing dataset (day dataset: 13,962,058; night dataset: 10,658,703). In this study, sklearn in Python is used to realize RF model. In the first step, the precipitation identification model is established to determine whether each satellite pixel has rain using the RF classification model. The precipitation condition of GPM IMERG product, namely non-precipitation and precipitation, is as the target variable. And the trained model is applied to the independent testing dataset to determine the precipitation region. In the second step, non-precipitation regions in the dataset are removed, and the precipitation estimation model is established using the RF regression model. The $\geq$0.1 mm/h precipitation data in the GPM IMERG product are the target variable. And the trained model is also applied to the testing dataset to retrieve precipitation intensity. Finally, the GPM IMERG product is used as the reference data, and the retrieval results are

compared with those of the FY-4B/AGRI operational precipitation product to evaluate the retrieval effect of the model.

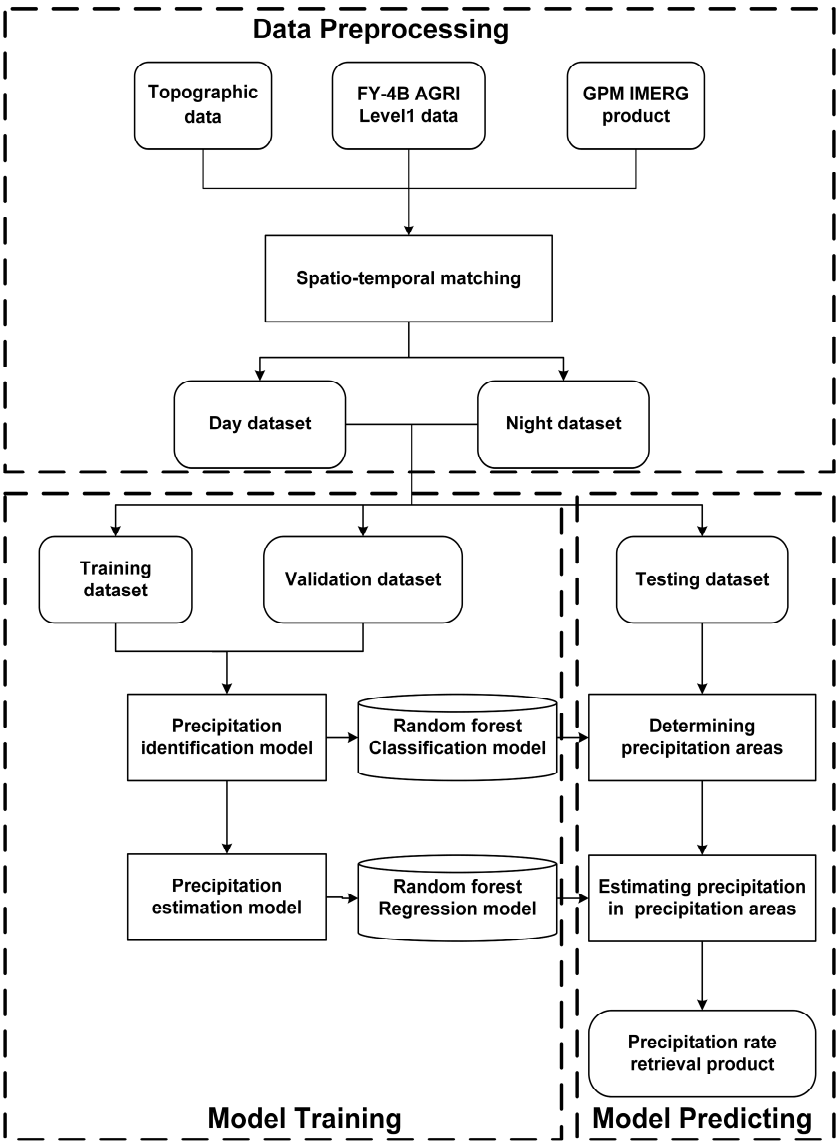

Figure 1. Technical flow of precipitation retrieval algorithm.

2.3.2. Random Forest

RF is a classic bagging model whose weak learner is the classification and regression tree (CART) decision tree model, which can be used for classification and regression [33]. RF training can be calculated in parallel, can estimate thousands of explanatory variables and can capture nonlinear modes between target variables and features. Compared with the traditional multiple linear regression model or parametric regression model and the neural network, RF does not need to set the function form in advance and can reduce the computational load [34]. Most importantly, the importance between variables can also be shown.

The main flow of RF is as follows: In the original dataset, random sampling is put back to form different sample datasets, and then different decision tree models are built according to these datasets. For each sample, if there are M input variables, and then k feature subsets are selected from M features. Then, the optimal splitting feature among k features is used to split the node. For classification tree, CART selects the optimal node to segment the data according to the Gini index minimum principle; for regression tree, CART selects the optimal node to segment the data according to the square error minimization

criterion and stops the growth until the branch stop criterion is met. The final result is obtained based on the mode (for the classification model) or the average (for the regression model) of these decision tree models.

2.3.3. Selection of Feature Variables

For the identification and estimation of precipitation, it is necessary to select feature variables related to precipitation as the input of the model. Considering the tropical deep convective precipitation process and mid-latitude stratospheric cloud precipitation process, physical variables related to precipitation process are selected, mainly including cloud top height (CTH), cloud top temperature (CTT), cloud water path (CWP), cloud phase (CP) and water vapor (WV). CTH reflects the level of cloud development [35], and clouds with significant vertical development contain a large amount of precipitation. CTT is widely used to estimate precipitation in convective systems [36], and a colder cloud top temperature corresponds to a higher precipitation rate. CWP reflects the size of water or ice particles and the optical thickness of the cloud [37]. A cloud region with higher cloud water path, that is, the cloud region with high optical thickness and large effective particle radius, has a large amount of cloud water and a higher probability and intensity of precipitation. CP reflects whether the upper part of the cloud is ice cloud or water cloud [37]. The rainfall process is related to the ice particles in the upper part of the cloud, and the cloud phase information can be used to distinguish the clear sky. WV can be used to represent different sensitivity characteristics to cloud water vapor.

The optical and microphysical properties of clouds can be determined by satellite observations. Visible channels can be used to observe clouds during day, with large differences in albedo values observed between clouds and clear sky. Near-infrared channels can be used to distinguish between snow and clouds; the albedo of snow is significantly higher than the albedo of lower clouds composed of water droplets. The radiation characteristics of the infrared channels are more sensitive to the size and distribution of water vapor condensation. An increase in particle radius leads to an increase in transmittance and cloud emissivity and a decrease in reflectivity. Therefore, the optical and microphysical properties of clouds can be deduced by appropriate satellite spectral channels and their combination, so as to reflect the radiation signal characteristics of the precipitation clouds.

The 10.8 μm infrared brightness temperature can be used to obtain information about CTT and CTH, especially for convective clouds [35,36]. Since the 10.8 μm band is located in the atmospheric window region and is relatively transparent, the influence of water vapor above the cloud top on the radiation in the window region is negligible, so the brightness temperature of the 10.8 μm channel in the window region is usually regarded as the CTT [38].

The generation of precipitation is the result of the continuous development of clouds. In general, the development of clouds at a single station shows that the brightness temperature of the cloud top is constantly decreasing. The attenuation of clouds at a single station shows that the brightness temperature of the cloud top is rising. Therefore, the change in the brightness temperature of the cloud top can be used as an indicator of the growth factor of the clouds and can also reflect the change in its precipitation intensity [22]. Research shows that the mean and variance of adjacent samples are usually used to represent the environment around pixel points [39]; thus, the brightness temperature variance is calculated as follows: a satellite image is divided into 5×5 panes, and the average brightness temperature of all pixels at 10.8 μm is calculated in each pane. Then, the variance between each pixel and the average brightness temperature are calculated. Then, the cloud top brightness temperature gradient of each pixel is calculated to represent the cloud top brightness temperature change. Taking the satellite pixel (i,j) as the centre, the cloud top brightness temperature gradient modulus ($Gm(IR_{10.8})$) of this point is shown in Formula (1).

$$Gm(IR_{10.8}) = \sqrt{(IR_{10.8}(i-1,j-1) - IR_{10.8}(i+1,j+1))^2 + (IR_{10.8}(i+1,j-1) - IR_{10.8}(i-1,j+1))^2} \quad (1)$$

The brightness temperature differences ($\Delta T_{6.25-10.8}$ and $\Delta T_{7.42-12}$) between the water vapor channel and the long-wave infrared channel are sensitive to changes in CTH [40] and can be used to distinguish convective clouds from non-precipitation cirrus clouds.

CP information such as ice or water clouds above clouds can be obtained by the difference between ($\Delta T_{8.55-10.8}$ and $\Delta T_{10.8-12}$). At these two wavelengths, the radiation absorption characteristics of ice particles and water particles in the upper clouds are different [41]. Water particles absorb more between 10.8 μm and 12 μm than between 8.55 μm and 10.8 μm, while ice particles absorb more between 8.55 μm and 10.8 μm than between 10.8 μm and 12 μm, so the cloud phase can be distinguished by the difference in absorption.

When the cloud top brightness temperature is low, cirrus clouds may exist, which are generally non-precipitation clouds [38]. The brightness temperature at 1.379 μm and 13.3 μm can be used to identify cirrus clouds [42].

Visible and near-infrared channels can provide information about CWP, and the larger the reflectivity of the channel, the larger CWP. During day, the visible reflection of sunlight is closely related to the optical thickness, while the near-infrared reflection is closely related to the effective particle radius of the cloud. At night, because the channels of visible and near-infrared wavelengths are not available, the difference ($\Delta T_{3.75-7.42}$ and $\Delta T_{3.75-10.8}$) is chosen to indirectly reflect the information of the cloud water path [36]. With the increase in the particle radius, the scattering effect of particles in the 3.75 μm channel is stronger than that in the 7.42 μm and 10.8 μm channels, and thinner clouds have stronger penetration. At the same time, the infrared radiation of a cloud with a larger optical thickness at 3.75 μm is greater than that at 7.42 μm and 10.8 μm channels.

WV channels can be used to represent different sensitivity characteristics of cloud water vapor. Although most of these channels have similar components, due to differences in sensitivity, they also have independent components for the characteristics of cloud top surface to a certain extent. Therefore, the similar characteristics of water vapor channels are represented by the sum of WV channels, while the difference between 6.25 μm and 7.42 μm represents the wavelength-dependent effects on emissivity difference in the cloud tops at WV channels [19].

Since satellites detect cloud top information in both visible and infrared channels, ground information related to precipitation should also be considered. Precipitation has local characteristics. Studies have shown that for annual precipitation, topographic elevation and topographic relief are the main topographic factors that affect the spatial heterogeneity of precipitation [43]. Increases and changes in topographic height will slow down the horizontal wind speed and affect the vertical wind speed at the bottom. Mountain fluctuations can force the air flow to rise or divert from both sides. Thus, digital elevation model (DEM) and orographic variation (OV) are added to the precipitation retrieval.

At the same time, considering that the energy received by the satellite detection instrument is related to the observation zenith angle of the satellite, the satellite zenith angle (SAZ) is added to the retrieval as additional information.

Table 3 shows the physical-related features and their specific expressions that need to be used in the model during day and at night. There are 18 feature variables in the day model and 16 feature variables in the night model.

Table 3. Physical-related features used by the model during the day and at night and their specific expressions.

Feature Variables	Day	Night
CTH	$\Delta T_{6.25-10.8}$ $\Delta T_{7.42-12}$	$\Delta T_{6.25-10.8}$ $\Delta T_{7.42-12}$
CTT	$IR_{10.8}$ $Var(IR_{10.8})$ $Gm(IR_{10.8})$	$IR_{10.8}$ $Var(IR_{10.8})$ $Gm(IR_{10.8})$

Table 3. *Cont.*

Feature Variables	Day	Night
CWP	$VIS_{0.65}$ $VIS_{0.825}$ $NIR_{1.61}$ $NIR_{2.225}$	$\Delta T_{3.75-7.42}$ $\Delta T_{3.75-10.8}$
CP	$\Delta T_{8.55-10.8}$ $\Delta T_{10.8-12}$ $NIR_{1.379}$ $IR_{13.3}$	$\Delta T_{8.55-10.8}$ $\Delta T_{10.8-12}$ $NIR_{1.379}$ $IR_{13.3}$
WV	$WV_{6.25+6.95+7.42}$ $WV_{6.25-7.42}$	$WV_{6.25+6.95+7.42}$ $WV_{6.25-7.42}$
Topography	DEM OV	DEM OV
Satellite zenith angle	SAZ	SAZ

2.3.4. Model Tuning and Testing

In the model, there are two important parameters that need to be adjusted: one is the total number of trees 'n' in the forest, which plays a crucial role in the sensitivity of the model's performance. The other is the number of features 'k' available for each node, which more directly affects the performance of the model. It determines the difference between each tree and indirectly affects the stability of the model against a high amount of noise data. Therefore, in the present study, the number of decision trees 'n' ranges from 10 to 1000, and the number of features 'k' ranges from 1 to the maximum number of feature variables of the model (18 during day and 16 at night), respectively, to find the optimal parameters of the model.

In RF modelling, there will be some data (about 36.8%) that are never randomly selected, which are called "out-of-bag data". These out-of-bag data are not used by the model for training, and sklearn can help us test the model with them.

In the classification model, ROC-AUC score of out-of-bag data is used as the evaluation index for parameter adjustments. The ROC-AUC curve is shown in Figure 2. ROC-AUC is one of the most important evaluation indicators to check the performance of a classification model. Receiver operating characteristic (ROC) is a probability curve with false positive rate (FPR) and true positive rate (TPR) as its axes [44]. Area under the curve (AUC) is the area under the ROC curve used to measure the performance of the classifier. The closer the value of AUC is to 1, the better the classifier performance is.

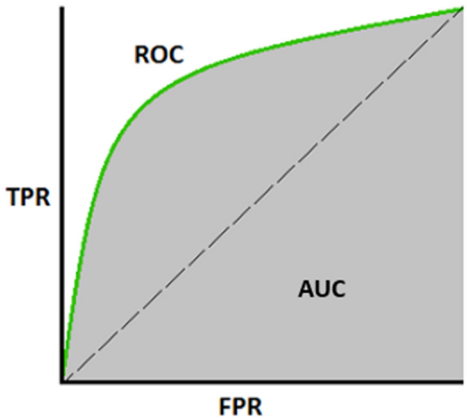

Figure 2. ROC-AUC curve (The dash line represents the classification ability equal to 0, also known as the pure opportunity line, and the green line represents the ROC curve).

In the regression model, mean square error (MSE) of the out-of-bag data is used as the evaluation index for parameter adjustments. MSE is generally used to detect the

deviation between the predicted value and the true value of the model. The smaller MSE is, the smaller the error of the model's predictions, which is estimated according to the formula below:

$$MSE = \frac{\sum_{i=1}^{n}(S_i - I_i)^2}{n} \quad (2)$$

where S represents model's prediction data, I represents observation data and n represents the number of matching points.

In addition, in the process of model training, the ratio of non-precipitation and precipitation in the dataset is 4:1, which contains a low number of precipitation pixels. Direct training of randomly selected dataset will lead to serious underestimation of precipitation samples. Therefore, the method of random subsampling is used to reduce the samples of non-precipitation to solve the problem of sample imbalance. The proportion of non-precipitation and precipitation samples is set as 4:1, 3:1, 2:1 and 1:1 for training, to evaluate the effect of the model and select the optimal sample proportion.

The precipitation test classification table is used to describe the relationship between model's predictions and observations, as shown in Table 4.

Table 4. Classification table of precipitation.

	GPM IMERG: Precipitation	**GPM IMERG: Non-Precipitation**
RF Prediction: Precipitation	NA	NB
RF Prediction: Non-precipitation	NC	ND

At the same time, false-alarm ratio (FAR), probability of detection (POD), critical success index (CSI) and equitable threat score (ETS) are introduced to evaluate the accuracy of precipitation identification model [45].

$$FAR = \frac{NB}{NA + NB} \quad (3)$$

$$POD = \frac{NA}{NA + NC} \quad (4)$$

$$CSI = \frac{NA}{NA + NB + NC} \quad (5)$$

$$ETS = \frac{NA - dr}{NA + NB + NC - dr}, \quad dr = \frac{(NA + NB) \cdot (NA + NC)}{NA + NB + NC + ND} \quad (6)$$

where FAR indicates the proportion of the area with no actual precipitation in the total predicted precipitation area predicted by the model. POD represents the possibility that the model can correctly identify precipitation pixels when precipitation is actually observed. CSI represents the proportion of precipitation pixels in correctly classified pixels when the correctly classified non-precipitation points are moved out. ETS indicates the proportion of pixels that are correctly classified after considering contingencies compared with that of random prediction [46]. The smaller the FAR is, the higher the POD, CSI and ETS are, the lower the false alarm rate of precipitation events and the better its precipitation identification ability.

According to the classification standard of rainfall, hourly rainfall can be divided into four classes: light rain (0.1~1.5 mm/h), moderate rain (1.5~7.0 mm/h), heavy rain (7.0~15.0 mm/h) and torrential rain ($\geq$15.0 mm/h) [47]. In the process of model training, the ratio of light rain, moderate rain, heavy rain and torrential rain in the dataset is 75:20:3:1, including a low number of pixels of heavy rain and torrential rain. Direct training of the randomly selected dataset will lead to a serious overestimation of light rain samples and a serious underestimation of heavy rain and torrential rain samples. The model may not learn the characteristics of heavy rain and torrential rain samples, resulting in a large error. Therefore, the upsampling method is used to greatly increase the number of samples of moderate rain, heavy rain and torrential

rain to solve the problem of sample imbalance. The ratio of light rain, moderate rain, heavy rain and torrential rain is set as 75:20:3:1 and 1:1:1:1, respectively, for training to evaluate the effect of the model and select the optimal sample ratio.

Correlation coefficient (R), root mean square error (RMSE) and average error (BIAS) are used to evaluate the precipitation estimation model's effect [48].

$$R = \frac{\sum_{i=1}^{n}(S_i - \overline{S}) \cdot (I_i - \overline{I})}{\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2 \cdot \sum_{i=1}^{n}(I_i - \overline{I})^2}} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(S_i - I_i)^2}{n}} \quad (8)$$

$$BIAS = \frac{\sum_{i=1}^{n} S_i - I_i}{n} \quad (9)$$

where S represents model prediction data, I represents observation data and n represents the number of matching points. The larger R is, the closer the model prediction data and the observed data are and the higher the data authenticity is. The smaller the RMSE, the smaller the degree of dispersion between the model prediction data and the observed data. The smaller the BIAS, the smaller the difference between the predicted data of the model and the observed data.

3. Results

3.1. Precipitation Identification Model

The model is realized through the following three steps: model tuning; model training and validation; and model testing. This section analyses the optimal sample proportion and parameters of the model tuning and summarizes the precipitation identification results in the validation dataset and testing dataset, respectively.

Figure 3 shows the sample distribution of different ratios during the day and at night after random subsampling. Tables 5 and 6 show the validation results for different sample ratios during the day and at night, respectively. The results show that as the proportion of non-precipitation and precipitation samples in the training set decreases, the FAR and POD values increase continuously, while the CSI and ETS values increase at first and then significantly decrease when the ratio is one to one. Although the validation results with a one-to-one sample ratio have the highest POD value, the FAR value is the highest and the ETS value is the lowest, which means that the false alarm ratio is very large and the model has a high misjudgment rate in predicting precipitation. Therefore, after a comprehensive comparison of all the validation results, a two-to-one ratio of non-precipitation to precipitation is selected as the optimal sample proportion.

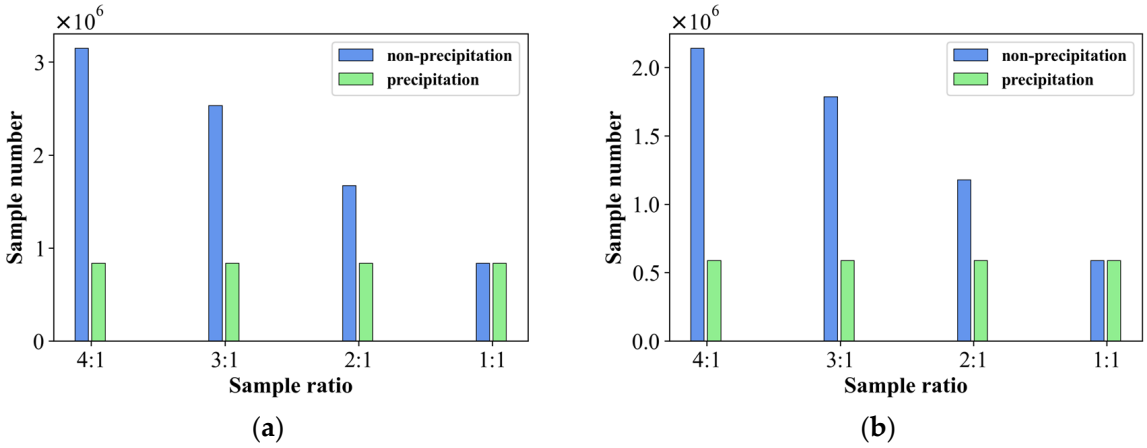

Figure 3. Sample number of different ratios (**a**) during the day and (**b**) at night.

Table 5. Validation results for different ratios during day in the precipitation identification model.

Non-Precipitation: Precipitation	FAR	POD	CSI	ETS
4:1	0.226	0.610	0.518	0.439
3:1	0.300	0.658	0.535	0.452
2:1	0.326	0.744	0.547	0.454
1:1	0.436	0.856	0.515	0.400

Table 6. Validation results for different ratios at night in the precipitation identification model.

Non-Precipitation: Precipitation	FAR	POD	CSI	ETS
4:1	0.248	0.593	0.496	0.413
3:1	0.277	0.639	0.513	0.425
2:1	0.346	0.740	0.530	0.430
1:1	0.453	0.859	0.502	0.379

Figure 4 shows the change in ROC-AUC score for different numbers of parameters in the model during the day and at night. Before the number of trees is 200, the score follows a significant increasing trend. After the number of trees is 200, the score no longer significantly changes, and the more trees there are, the more computer memory and time will be required. Therefore, 500 trees are selected as the optimal number of trees in the precipitation identification model. At the same time, when the number of features is 11 during the day and 12 at night, the score is the largest. Therefore, 11 and 12 features are selected as the number of optimal features in the precipitation identification model during the day and at night, respectively.

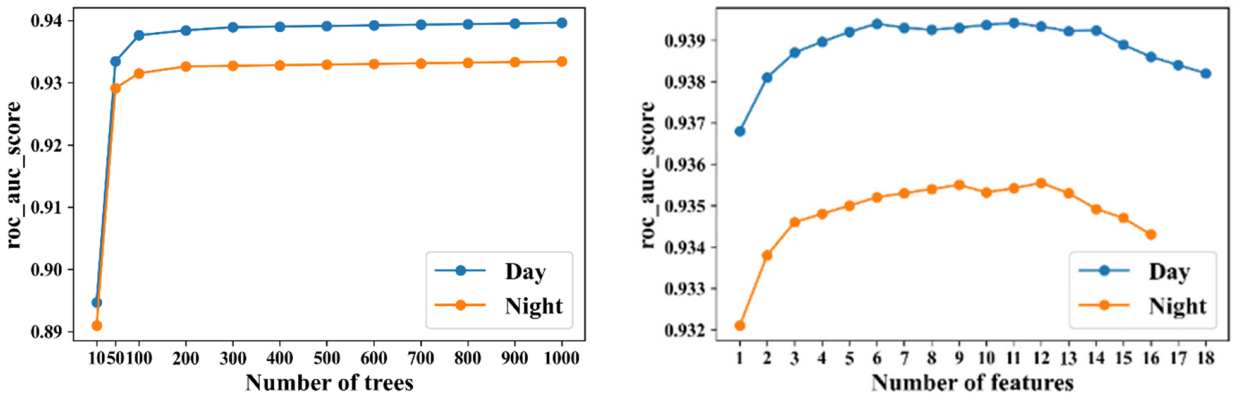

Figure 4. ROC-AUC score for different numbers of (**a**) trees and (**b**) features.

The RF classifier will give each input feature a specific weight so that the importance score of each feature can be calculated from the sum of the reduced Gini coefficients of all nodes split on that feature in all of the decision trees. Thus, we can use the 'feature_importances_' to output the importance of each feature. As shown in Figure 5, the infrared bright temperature and infrared bright temperature difference, which represent cloud microphysical parameters such as CTH, CP, CWP and CTT, have made great contributions. Among them, CTH is the most important. At the same time, DEM, the topography variable, and SAZ, the satellite observation angle variable, are also of high importance.

The trained model is applied to the validation dataset. As shown in Table 7, the precipitation prediction effect during the day is better. The FAR score is 0.326, that is, for 32.6% of the cases in which precipitation is predicted, the IMERG does not observe precipitation. The POD score is 0.744, that is, the probability of the model correctly identifying precipitation pixels is 74.4% when precipitation is observed. The CSI score is 0.547 and the ETS score is

0.454. At night, the FAR score is 0.346, the POD score is 0.740, the CSI score is 0.530 and the ETS score is 0.430.

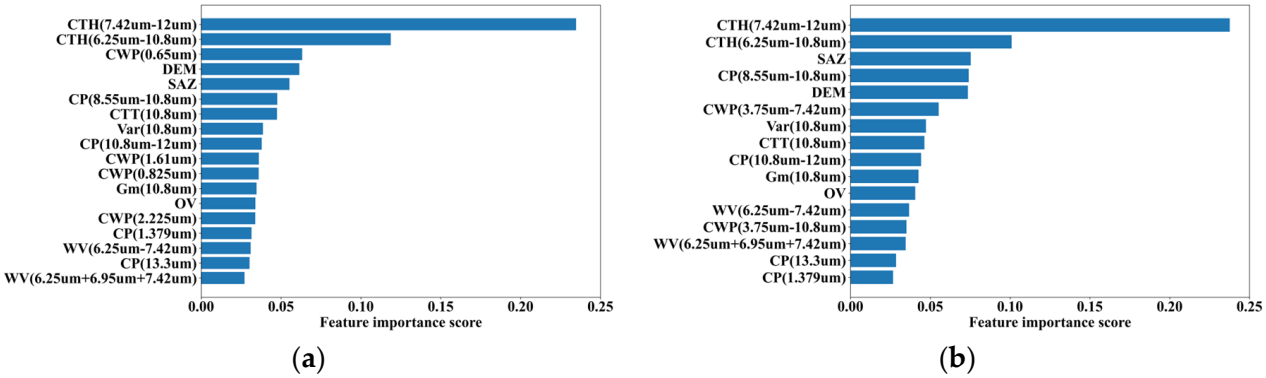

Figure 5. Importance ranking of feature variables in precipitation identification model (**a**) during the day and (**b**) at night.

Table 7. Evaluation indicators of the identification model.

	FAR	POD	CSI	ETS
Day	0.326	0.744	0.547	0.454
Night	0.346	0.740	0.530	0.430

The independent testing dataset is used to determine the precipitation region and is compared with the FY-4B/AGRI operational precipitation product of the same period, as shown in Table 8. The results show that the FAR of the model is higher than that of the operational product. The POD, CSI and ETS scores are all higher than those of the operational product. Overall, the model is better at identifying precipitation. From the comparison of the day and night models, the FAR score of the night model is higher, and the POD, CSI and ETS scores are not as high as those of the day model, which may be due to the relatively more input feature variables of the day model and the fact that it learns more precipitation features.

Table 8. Evaluation indicators of precipitation identification model and FY-4B/AGRI operational precipitation product.

Evaluation Indicators	Retrieval Model		Operational Product	
	Day	Night	Day	Night
FAR	0.385	0.448	0.319	0.393
POD	0.680	0.639	0.328	0.337
CSI	0.477	0.421	0.284	0.277
ETS	0.432	0.369	0.252	0.239

A boxplot of the evaluation indicators (FAR, POD, CSI and ETS) is used to represent the hour-by-hour results of the precipitation identification model on the testing dataset, as shown in Figure 6. The red line represents the mean value, and the red dot represents the outlier. The lower end and upper end of the blue box correspond to 25% and 75% after all the values are arranged from small to large. The short horizontal line on the upper boundary of the box represents the maximum value except the outlier, and the horizontal line on the lower boundary represents the minimum value except the outlier. During the daytime period (2200 UTC~1000 UTC, corresponding to 06:00~18:00 Beijing time), the model shows relatively high POD, CSI and ETS scores and a relatively low FAR score. The change in each index is relatively stable. During the night period (1100 UTC~2100 UTC, corresponding to 19:00~05:00 Beijing time), the performance of the model is relatively poor, and the change in indicators is large. Meanwhile, the transition of the indicators between day and night is relatively continuous.

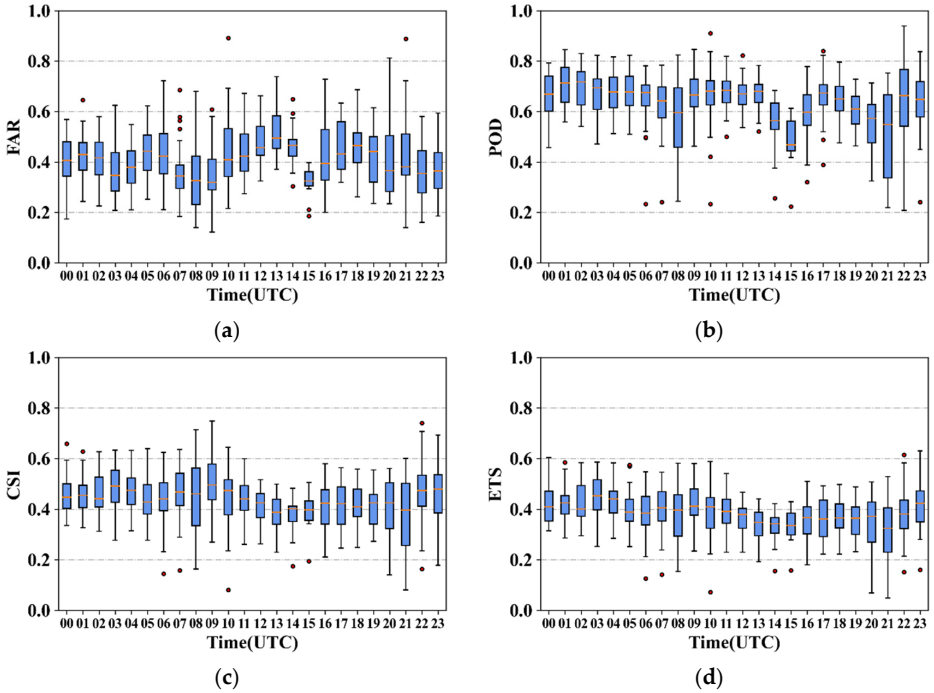

Figure 6. Hour-by-hour results of precipitation identification model on testing dataset (**a**) FAR, (**b**) POD, (**c**) CSI and (**d**) ETS.

We take 0200 UTC~0230 UTC on 18 August 2022 and 1500 UTC~1530 UTC on 9 August 2022 as an example to evaluate the effect of the day and night models, respectively, as shown in Figures 7 and 8, where the blue region represents precipitation and the colourless region represents non-precipitation. Compared with the GPM IMERG product, the model is basically accurate in retrieving the location of the precipitation region, but the region is relatively large. Particularly in the western region, the night model does not identify precipitation well. In the area where the GPM IMERG product observed precipitation, the FY-4B/AGRI operational precipitation product did not observe precipitation, especially in the coastal area.

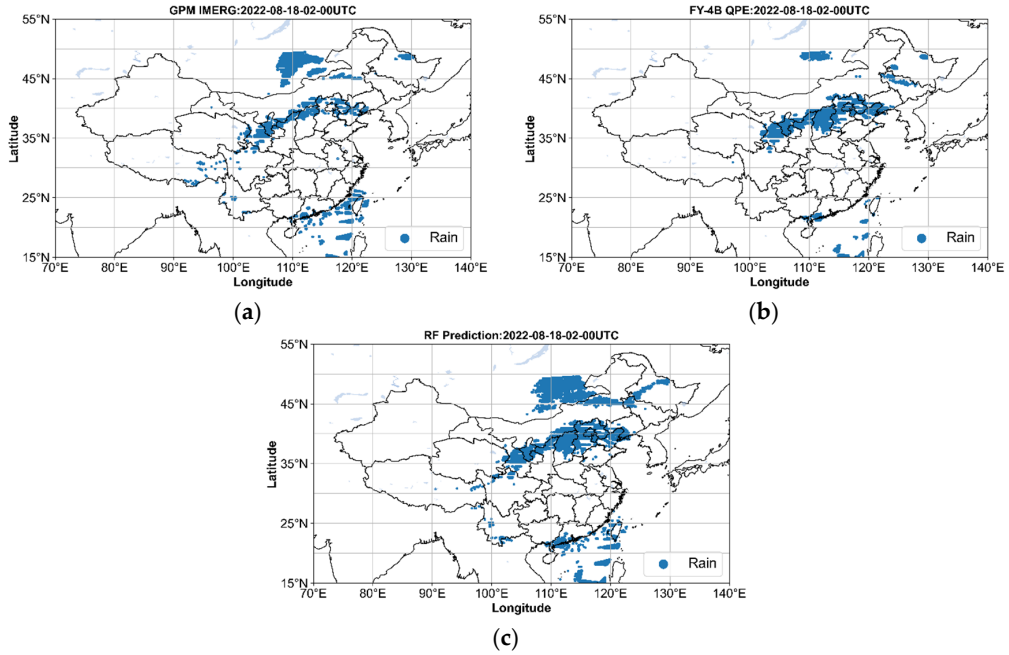

Figure 7. Precipitation identification results from 0200 UTC to 0230 UTC on 18 August 2022: (**a**) GPM IMERG product, (**b**) FY-4B/AGRI operational precipitation product and (**c**) RF model.

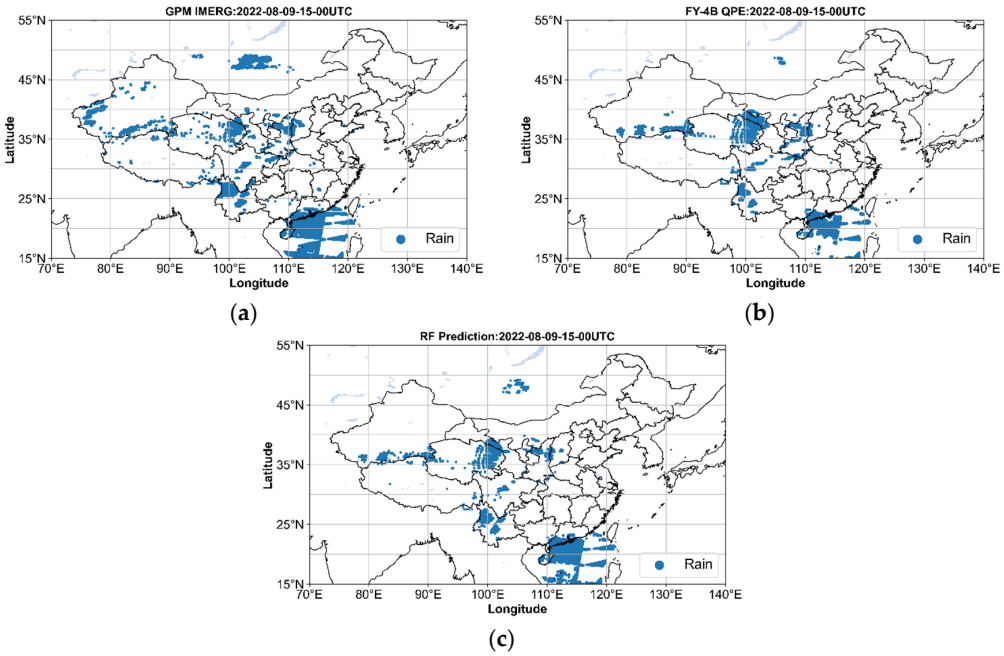

Figure 8. Precipitation identification results of 1500 UTC~1530 UTC on 9 August 2022: (**a**) GPM IMERG precipitation product, (**b**) FY-4B/AGRI operational precipitation product and (**c**) RF model inversion results.

The retrieval results for the ocean and land are compared and analysed, respectively, as shown in Figure 9. The results show that the ocean has a better ability to identify precipitation than the land, and the ocean has a lower FAR score and higher POD, CSI and ETS scores.

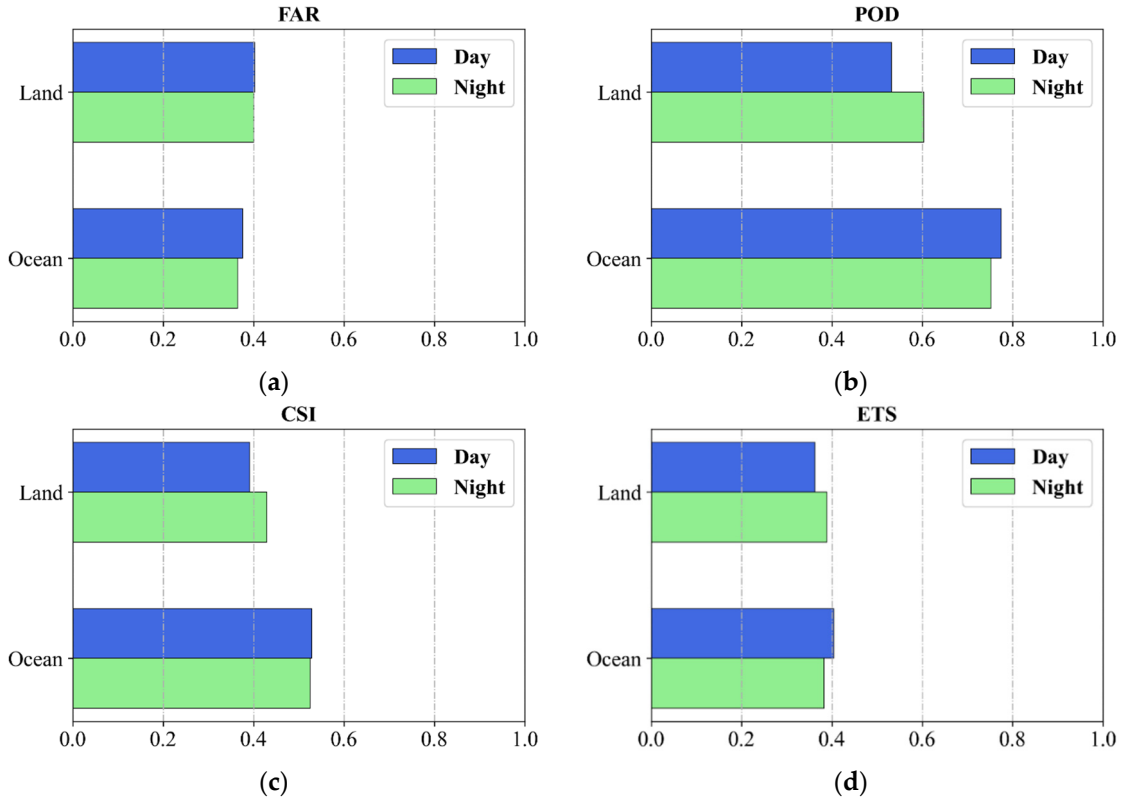

Figure 9. Evaluation indicators of ocean and land, (**a**) FAR, (**b**) POD, (**c**) CSI and (**d**) ETS.

The land is subdivided according to different land cover types, and the precipitation identification results are evaluated, respectively, as shown in Figure 10. The results show that there is no significant difference in the evaluation indicators of the underlying surfaces of different land cover types, indicating the universal applicability of the model. In the day model, water bodies perform better. In the night model, woodland performs better, with a lower FAR score and higher POD, CSI and ETS scores.

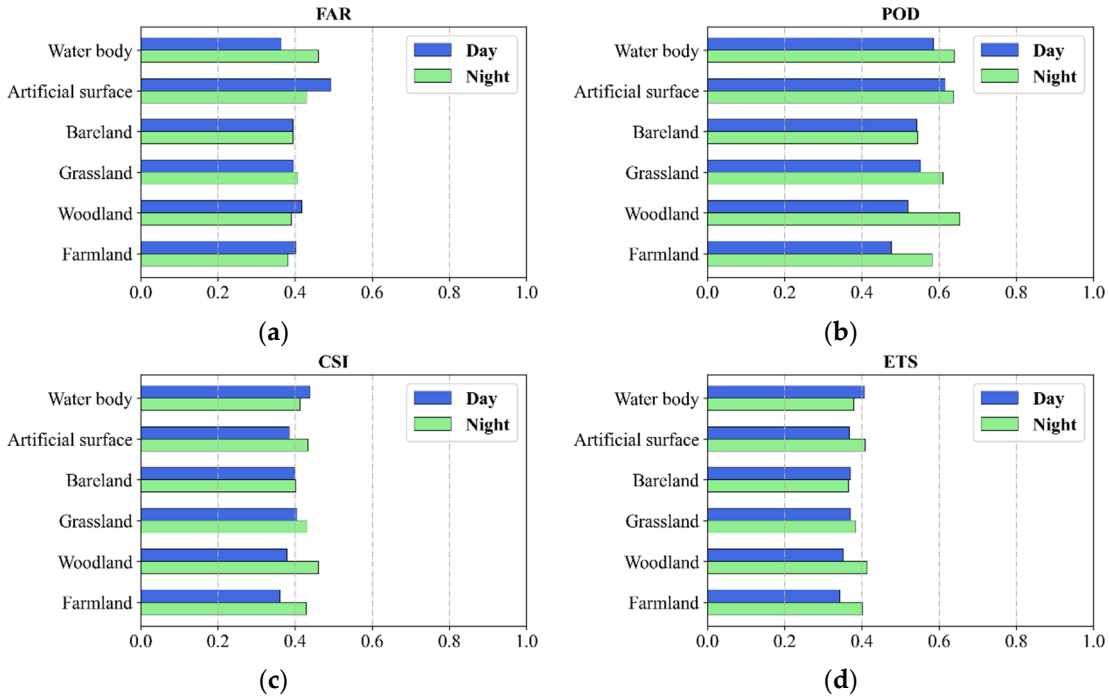

Figure 10. Evaluation indicators of different underlying surfaces of land, (**a**) FAR, (**b**) POD, (**c**) CSI and (**d**) ETS.

3.2. Precipitation Estimation Model

The model is also realized through three steps. This section analyses the optimal sample proportion and parameters for model tuning as well as summarizes the precipitation estimation results in the validation dataset and testing dataset, respectively.

Figure 11 shows the sample's distribution for different ratios during the day and at night after random upsampling. Tables 9 and 10 show the validation results for different ratios of day and night, respectively. The results show that when the sample size for moderate rain, heavy rain and heavy rain is increased, and the sample proportion of light rain, moderate rain, heavy rain and torrential rain is set to 1:1:1:1, the BIAS and RMSE values of moderate rain, heavy rain and heavy rain are reduced, and the model effect is better. Therefore, 1:1:1:1 is selected as the optimal sample proportion.

Table 9. Validation results for different ratios during day in the precipitation estimation model.

Sample Ratio	Light Rain		Moderate Rain		Heavy Rain		Torrential Rain	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
75:20:3:1	0.698	1.122	−0.488	2.101	−4.639	5.909	−14.852	17.987
1:1:1:1	0.878	1.292	−0.234	2.025	−4.457	5.713	−14.306	17.367

Table 10. Validation results for different ratios at night in the precipitation estimation model.

Sample Ratio	Light Rain		Moderate Rain		Heavy Rain		Torrential Rain	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
75:20:3:1	0.722	1.179	−0.541	2.175	−4.985	6.212	−15.003	17.871
1:1:1:1	0.910	1.374	−0.237	2.131	−4.767	5.986	−14.481	17.334

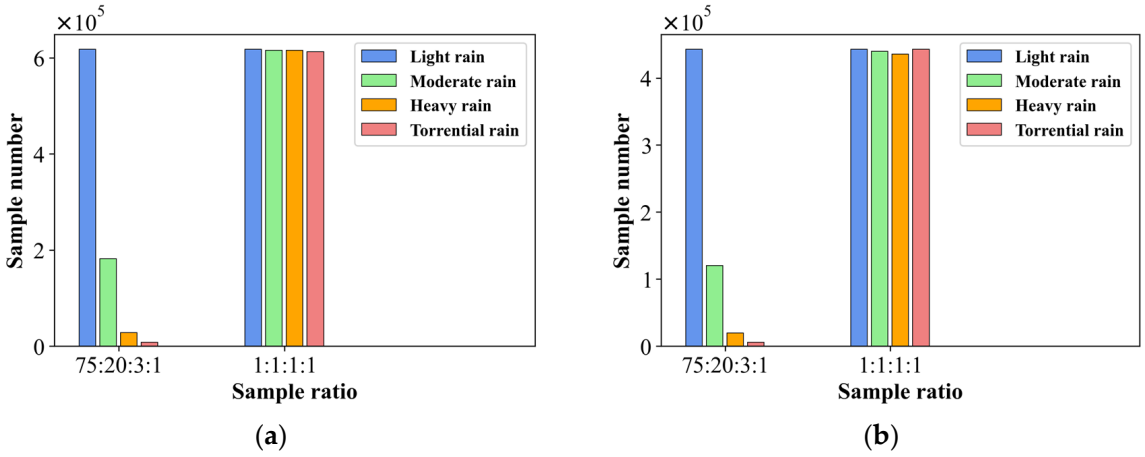

Figure 11. Sample numbers for different ratios during (**a**) day and (**b**) night.

Figure 12 shows the change in the MSE score under different numbers of parameters in the model during the day and at night. Before the number of trees is 200, the MSE value has a significant decreasing trend, and after 200, the MSE value no longer changes significantly. Therefore, 500 trees are selected as the optimal number of trees in the precipitation estimation model. When the number of features is 11 during both the day and at night, the MSE value is the smallest. Therefore, 11 features are selected as the number of optimal features in the precipitation estimation model.

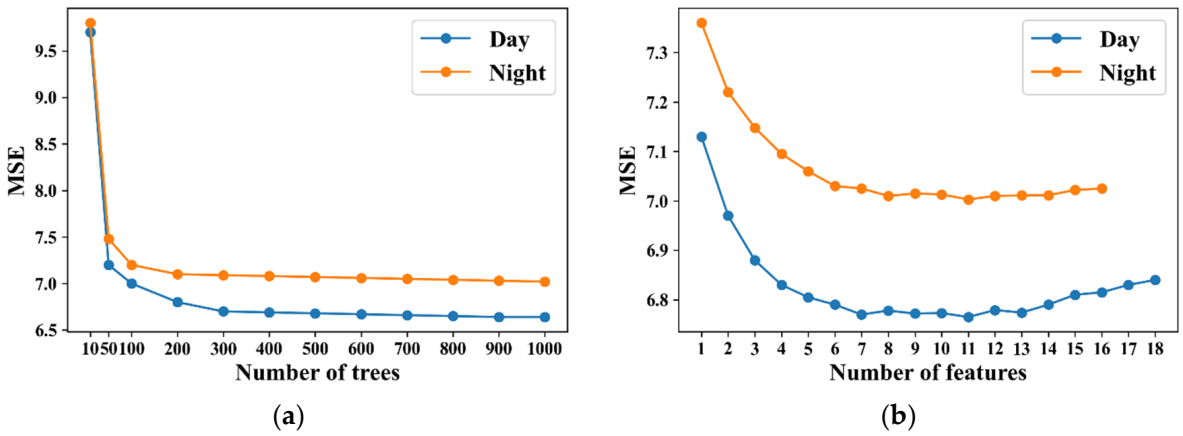

Figure 12. MSE for different numbers of (**a**) trees and (**b**) features.

The importance of each feature of the model can also be output, as shown in Figure 13. The infrared bright temperature and infrared bright temperature difference, which represent cloud microphysical parameters such as CTH, CWP, CTT and CP, still make a large contribution. Meanwhile, in the precipitation estimation model, the contribution of WV to model training is also large, especially in the night model. DEM and SAZ are also important for precipitation estimation.

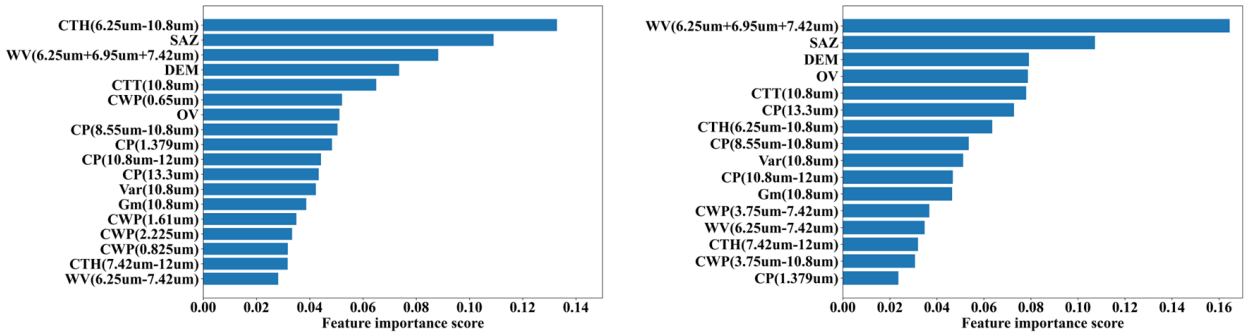

Figure 13. Importance ranking of feature variables in precipitation estimation model (**a**) during day and (**b**) at night.

The trained model is applied to the validation dataset, as shown in Table 11. The precipitation retrieval by the day model is consistent with the GPM IMERG product, and the precipitation intensity error is small. At night, the R is 0.604, BIAS is 0.332 mm/h and RMSE is 2.558 mm/h. A scatterplot of the precipitation intensity is shown in Figure 14, where red indicates a high density of data. In the light rain region, the model shows a significant overestimation, while in the heavy rain and torrential rain regions, the model shows an underestimation.

Table 11. Evaluation indicators of the precipitation estimation model.

	R	BIAS	RMSE
Day	0.631	0.308	2.495
Night	0.604	0.332	2.558

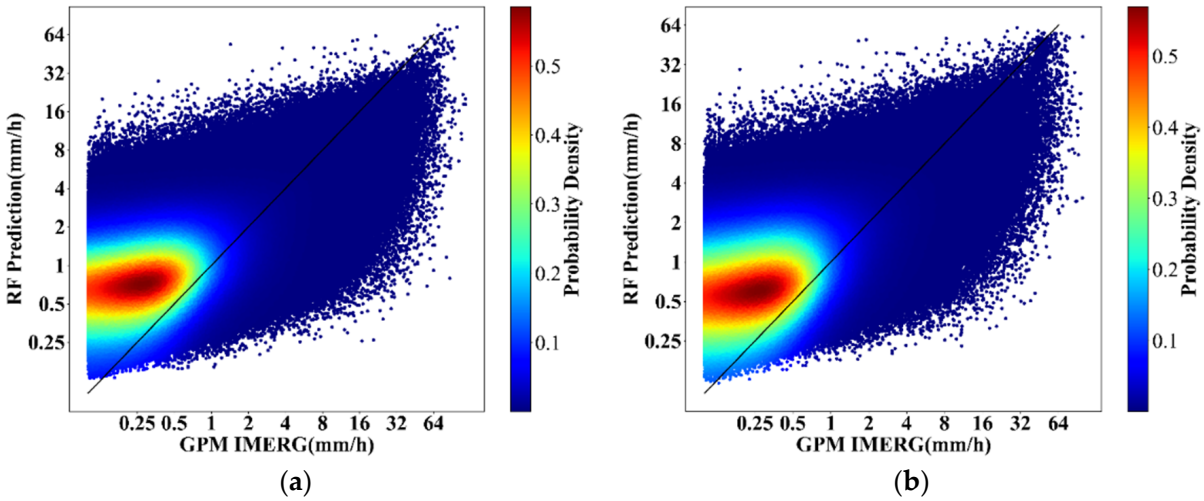

Figure 14. Density scatterplot of precipitation intensity with GPM IMERG and RF predictions (**a**) during day and (**b**) at night.

The independent testing dataset is used to estimate the precipitation intensity in precipitation areas and is compared with the FY-4B/AGRI operational precipitation product for the same period, as shown in Table 12. The BIAS between the model and the product is positive, indicating that the amount of precipitation is overestimated by the model and the operational product as a whole. The RMSE of the model is lower than that of the operational product both during the day and at night, indicating that the retrieval accuracy of the model is higher than that of the operational product. The error of the day model is

smaller than that of the night model, indicating that the day model has a better retrieval effect on the precipitation intensity. In general, it is noted that the accuracy of the day model is improved by 38.98% compared with that of the operational product. Similarly, the night model's accuracy is improved by 40.85%. The formula for the improvement in accuracy is as follows:

$$\left(RMSE_{product} - RMSE_{\mathrm{model}}\right) / RMSE_{product} \quad (10)$$

Table 12. Evaluation indicators of precipitation estimation model and FY-4B/AGRI operational precipitation product.

Evaluation Indicators	Retrieval Model		Operational Product	
	Day	Night	Day	Night
R	0.441	0.421	0.254	0.311
BIAS	0.744	1.029	0.598	1.517
RMSE	2.832	3.127	4.641	5.291

The BIAS and RMSE values of different precipitation levels are also evaluated, as shown in Figure 15. The results show that the BIAS value of both the model and operational product is positive for light rain, indicating that both the model and operational product overestimate the amount of precipitation. At the moderate rain level, the BIAS of the day model is negative, indicating that the day model underestimates the amount of rainfall, while the BIAS of the day product as well as the night model and product are positive, indicating that the amount of precipitation is overestimated. And it is worth mentioning that the BIAS of moderate rain is quite small. For heavy rain and torrential rain, both the model and operational product underestimate the amount of precipitation on the whole, and the degree of underestimation increases with the increase in precipitation level. At the same time, the RMSE increases with the increase in precipitation level. The RMSE of the day model is smaller than that of the night model, and the RMSE of the model is lower than that of the operational product. For torrential rain, the RMSE of the day model is smaller than that of the day operational product, but the RMSE of the night model is larger, indicating that the accuracy of the model is relatively low when retrieving precipitation at night.

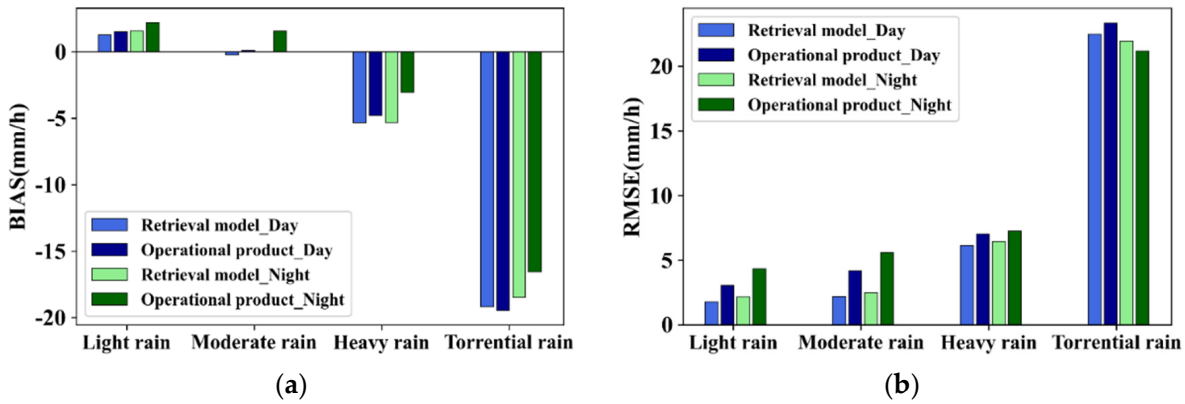

Figure 15. (**a**) BIAS and (**b**) RMSE between retrieval model and operational product at different precipitation levels.

Figure 16 shows the hour-by-hour results (R, BIAS snd RMSE) of the precipitation estimation model for the testing dataset. During the daytime (2200 UTC~1000 UTC), the change in each index is relatively stable. During the night period (1100 UTC~2100 UTC), the performance of the model is relatively poor, and the variation in indicators is large. In

particular, in the period of 1100 UTC~1500 UTC, the R value is relatively decreased, and the BIAS and RMSE values are relatively increased. Additionally, the transition of indicators between day and night is relatively continuous, and there are no large fluctuations.

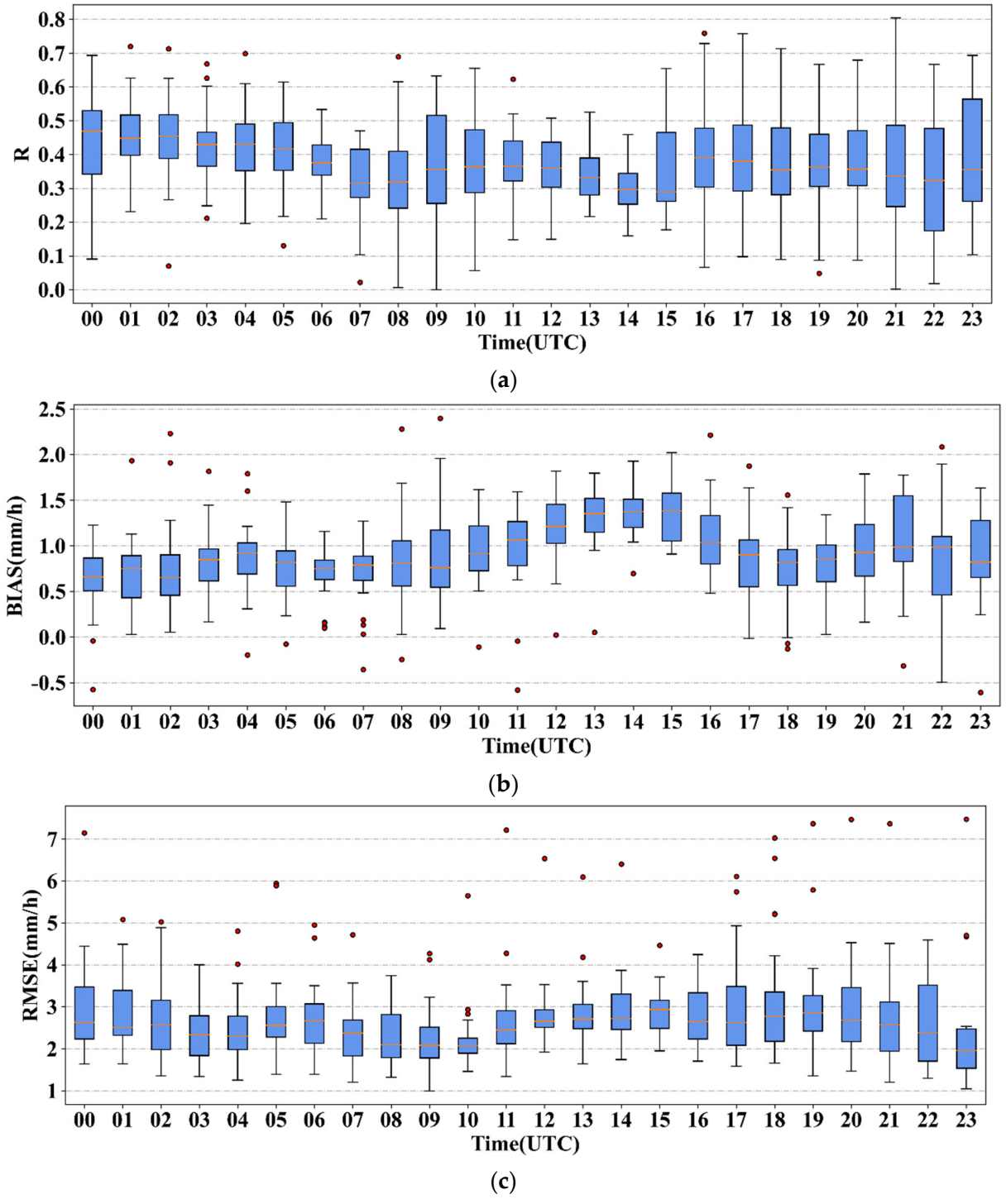

Figure 16. Hour-by-hour results of the precipitation estimation model for the testing dataset (**a**) R, (**b**) BIAS and (**c**) RMSE.

We take 0200 UTC~0230 UTC on 18 August 2022 and 1500 UTC~1530 UTC on 9 August 2022 as an example to evaluate the effect of the day and night models, respectively, as shown in Figures 17 and 18, where the colour code represents the precipitation intensity. Compared with the GPM IMERG product, the model is able to better capture the precipitation regions

of light rain, moderate rain, heavy rain and torrential rain, but the precipitation intensity is still underestimated for heavy rain and heavy rain. In the light-rain regions, some overestimations appear. The FY-4B/AGRI operational precipitation product cannot capture some light-rain and moderate-rain regions, and there are obvious overestimations in moderate-rain and heavy-rain regions.

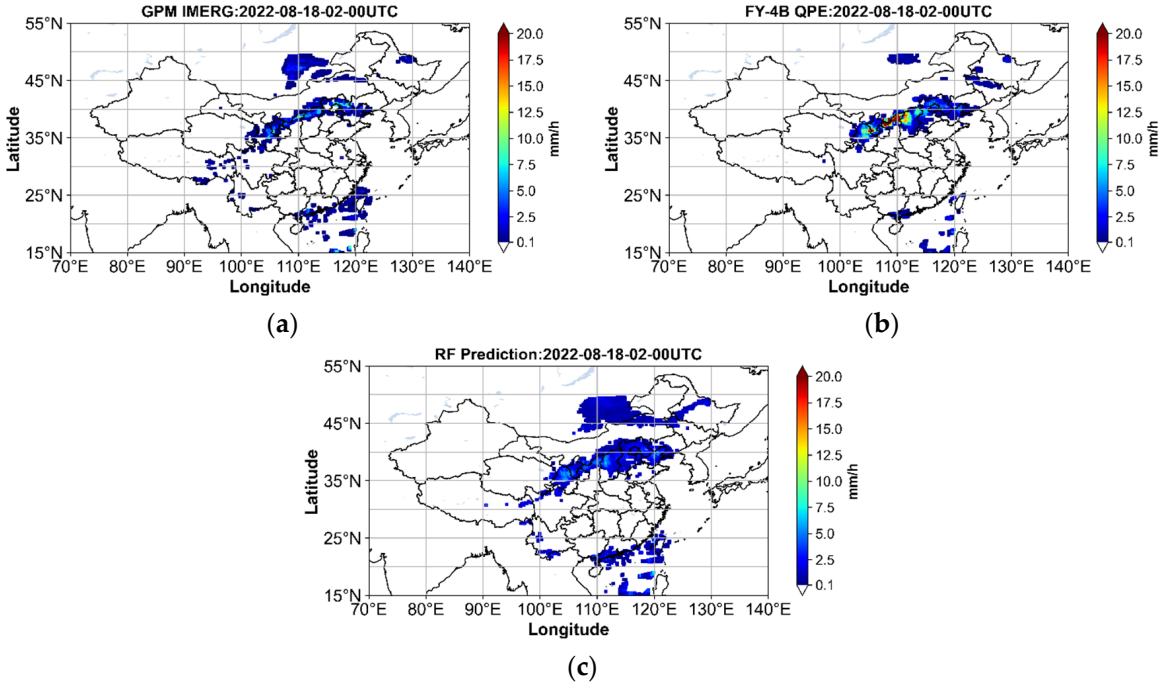

Figure 17. Precipitation estimation results from 0200 UTC to 0230 UTC on 18 August 2022: (**a**) GPM IMERG product, (**b**) FY-4B/AGRI operational precipitation product and (**c**) RF model.

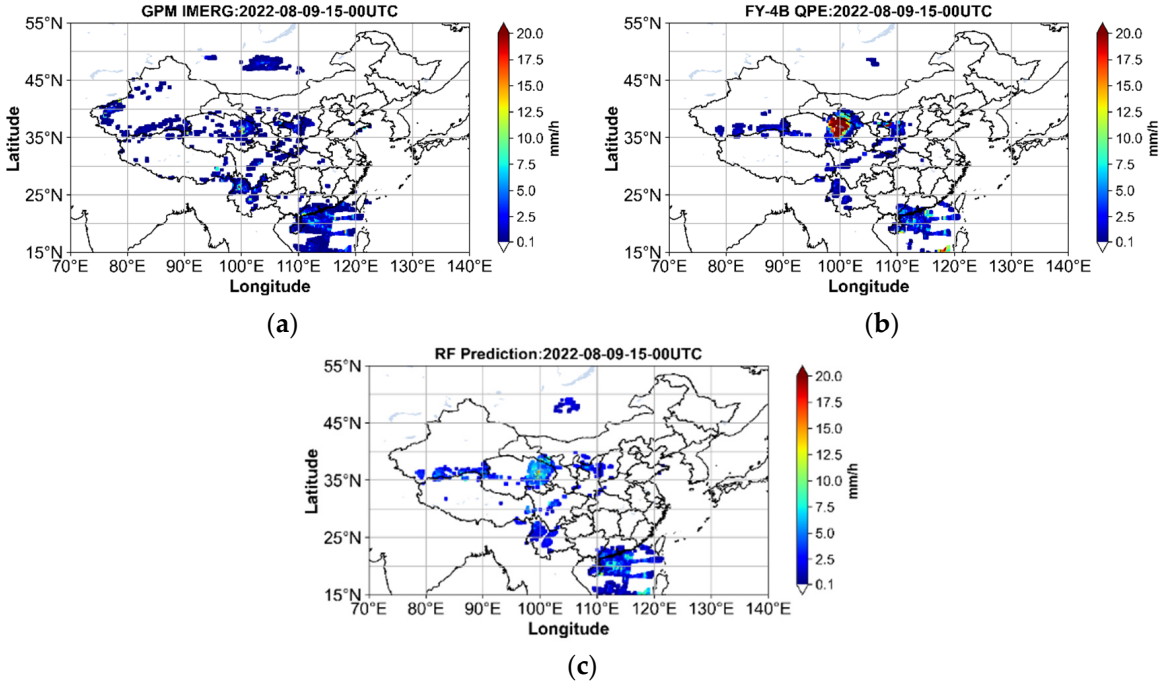

Figure 18. Precipitation estimation results of 1500 UTC~1530 UTC on 9 August 2022: (**a**) GPM IMERG product, (**b**) FY-4B/AGRI operational precipitation product and (**c**) RF model.

Figure 19 shows the retrieval effect of the model on different underlying surfaces. The results show that the RMSE of the ocean is higher than that of the land. The day model is better for the underlying surface of the land, and there is no significant difference between the model under different underlying surfaces.

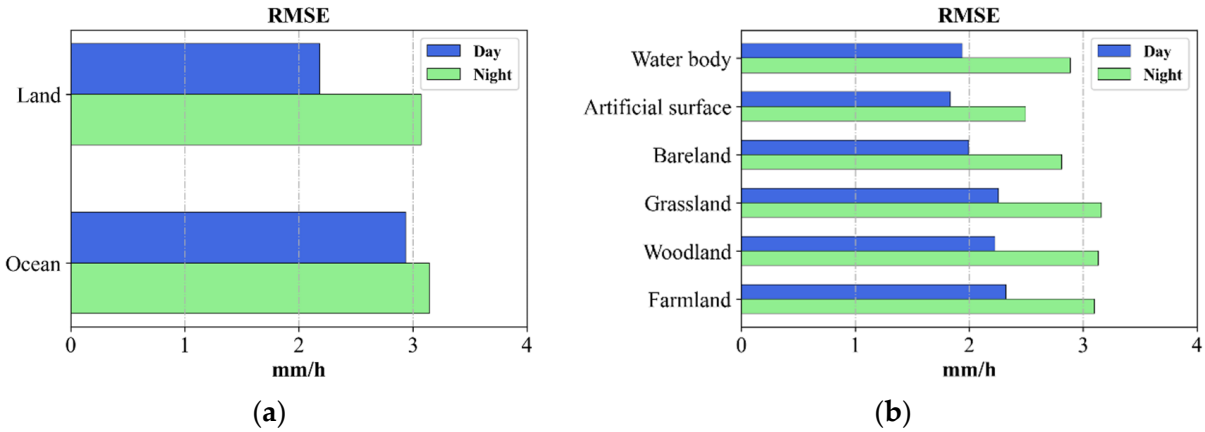

Figure 19. RMSE comparison of different underlying surfaces: (**a**) land and sea classification, (**b**) land classification.

3.3. Comparision with Ground Rain Gauge Data

To demonstrate the advantages and value of the proposed model in applications, the ground rain gauge data with a higher accuracy are selected for independent validation for the testing dataset. Before the evaluation, we match the rain gauge data with the retrieval data. Firstly, we convert Beijing time to UTC time. Because the temporal resolution of the rain gauge data is one hour, the average of the retrieval data in one hour is matched with the rain gauge data at this hour. Secondly, we search for the retrieval data closest to the ground rain gauge station [49]. The evaluation indicators are shown in Figures 20 and 21. The results show that when using ground rain gauge data as our reference, the accuracy of the precipitation retrieval is lower than that of the GPM IMERG product, but the overall trend is consistent. In the precipitation identification model, the POD, CSI and ETS scores are all higher than those of the FY-4B/AGRI operational product, and in the precipitation estimation model, the retrieval error is also smaller. Overall, the model is better at identifying and estimating precipitation.

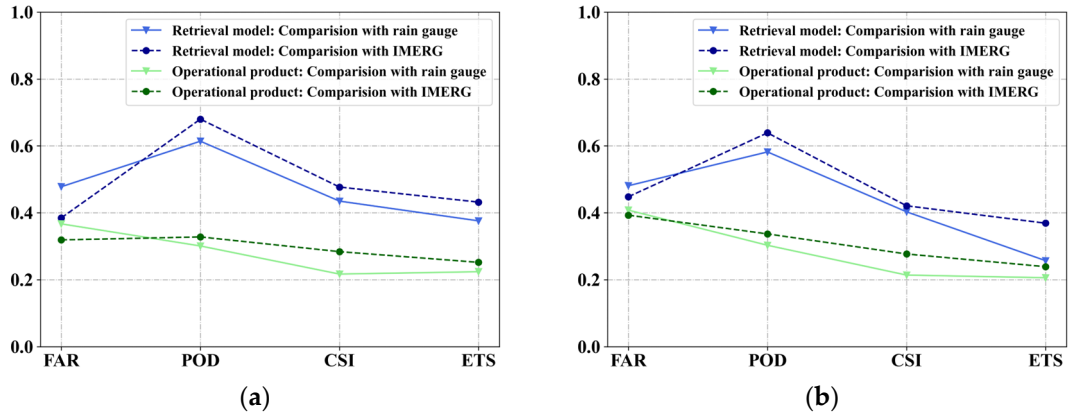

Figure 20. Evaluation indicators of precipitation identification model and FY-4B/AGRI operational precipitation product based on the ground rain gauge data and GPM IMERG product (**a**) during the day and (**b**) at night.

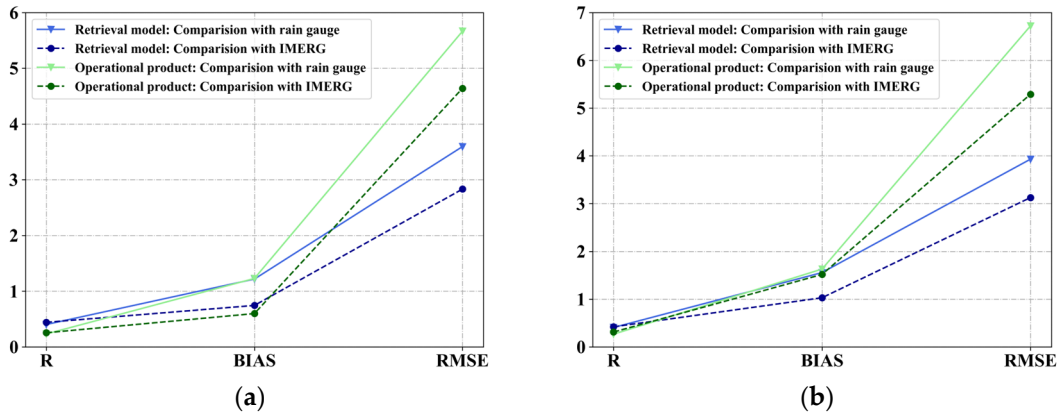

Figure 21. Evaluation indicators of precipitation estimation model and FY-4B/AGRI operational precipitation product based on the ground rain gauge data and GPM IMERG product (**a**) during the day and (**b**) at night.

4. Discussion

This study focuses on establishing an FY-4B/AGRI precipitation identification and estimation model during the day and at night based on an RF. To further improve the accuracy of satellite precipitation retrieval, especially for heavy rain and torrential rain, we made the following improvements during the training of the model.

Firstly, we selected 18 feature variables in the day model and 16 feature variables in the night model related to precipitation as the input of the model. These variables can better reflect the characteristics of clouds and precipitation. Among them, in the precipitation identification model, CTH is the most important, indicating that the cloud region with a large cloud top height and a low cloud top temperature have a large amount of cloud water and a higher probability and intensity of rainfall. In the precipitation estimation model, the contribution of WV to model training is important, especially in the night model, indicating that abundant water vapor conditions have an important effect on precipitation. Meanwhile, in the summer, due to the strong influence of the East Asian summer monsoon, the prevailing southeast wind brings abundant water vapor. Under a high temperature, the water vapor condenses with the uplift of terrain, which is the main mechanism for generating precipitation. Therefore, the impact of DEM on precipitation is also significant. At the same time, SAZ, the satellite observation angle variable, is also of high importance. For some features that contribute relatively little, their relatively small contribution may come from overlapping with other information. In the decision tree algorithm, a feature's contribution to an overlapping part will be assigned to the most distinguishable feature variable first, while the contribution of other features will not be recorded, although they may have partially overlapping information. Thus, this results in the underestimation of some features' contributions.

Secondly, we tested the effect of different proportions of sample size on the accuracy of the model training and selected the optimal sample proportion. In the precipitation identification model, the method of random subsampling is used to reduce the samples of non-precipitation. The proportion of non-precipitation to precipitation samples is set as 4:1, 3:1, 2:1 and 1:1 for training. In the precipitation estimation model, the upsampling method is used to greatly increase the number of samples of moderate rain, heavy rain and torrential rain. The proportion of light rain, moderate rain, heavy rain and torrential rain is set as 75:20:3:1 and 1:1:1:1, respectively, for training. The results show differences in sample size do have an impact on the accuracy of the model. After a comprehensive comparison of all the validation results, a 2:1 ratio of non-precipitation to precipitation and a 1:1:1:1 ratio of light rain, moderate rain, heavy rain and torrential rain are selected as the optimal sample proportion, respectively. Such settings enhance the ability of the model to learn minority samples and reduce the retrieval error.

Thirdly, we evaluated the retrieval effect of different underlying surfaces. The ocean has a better ability to identify precipitation than the land. Its better performance may be due to the fact that compared with the land, the ocean's surface is more uniform. And there are more heavy rainfall events over the ocean, so these unpredictable heavy rainfall events also inevitably increase the frequency of large errors over the ocean [20]. In different underlying surfaces of the land, water bodies perform better in the day model. The evaporation of river water increases the amount of water vapor in the atmosphere and increases cloud cover and precipitation, so the model is better able to identify precipitation. In the night model, woodland has a better performance. Its better performance may be attributed to the characteristics of woodland. Its lower reflectivity and higher absorption rate provide heat for the generation of showers. A large amount of water vapor from forest transpiration is quickly transported to the sky, which promotes precipitation, especially convective precipitation and topographic precipitation.

Overall, the proposed model is better at identifying precipitation. Its POD, CSI and ETS scores are all higher than those of the operational product, while the FAR score of the model is higher than that of the operational product. This may be due to the fact that the model subsamples the non-precipitation samples and removes a large number of them, which leads to an overestimation of retrieval precipitation pixels. To demonstrate the value of the proposed model in an application, we also used the ground rain gauge data as our reference to conduct a comparative experiment. When using the ground rain gauge data as our reference, its accuracy of precipitation retrieval is lower than that of the GPM IMERG product, but the overall trend is consistent. On the one hand, this is because the random forest model uses the GPM IMERG product as the target data, and the retrieval model is designed and tuned to replicate the precipitation intensity as provided by the GPM IMERG product. The retrieval results are closer to those of the GPM IMERG product, so the error between the GPM IMERG product and ground rain gauge data is also substituted into the retrieval results. On the other hand, there are still temporal and spatial differences in terms of the spatio-temporal matching between the retrieval data and ground rain gauge data.

5. Conclusions

In the present study, a new precipitation retrieval model is proposed during the day and at night using FY-4B/AGRI Level1 satellite data, based on a random forest classification and regression model. The use of the proposed method is evaluated for China for the time period from July to August 2022. The GPM IMERG product is used as a reference to evaluate the retrieval effect of the model. Our key study findings are summarized as follows:

(1) Compared with the FY-4B/AGRI operational precipitation product, the retrieval model is better able to identify precipitation and better able to capture precipitation areas of light rain, moderate rain, heavy rain and torrential rain. During the day, the POD score increased from 0.328 to 0.680, the CSI score increased from 0.284 to 0.477 and the ETS score increased from 0.252 to 0.432. At night, the POD score increased from 0.337 to 0.639, the CSI score increased from 0.277 to 0.421 and the ETS score increased from 0.239 to 0.369.

(2) The precipitation estimation accuracy of the retrieval model is higher than that of the FY-4B/AGRI operational precipitation product, in which the accuracy of the day model increased by 38.98% and that of the night model by 40.85%. Moreover, the retrieval error of the model increases with the increase in precipitation level. For light rain, both the model and operational product overestimate the amount of precipitation. For moderate rain, the day model underestimates the amount of precipitation, while the day product, the night model and the night product overestimate the amount of precipitation. And it is worth mentioning that the BIAS score of moderate rain is quite small. For heavy rain and torrential rain, both the model and product underestimate the amount of precipitation on the whole, and the degree of underestimation increases with the increase in precipitation level.

(3) In our comparative analysis of different underlying surfaces, due to the surface uniformity of the ocean, the model can identify precipitation better on the ocean than on the land. For different underlying surfaces of the land, there is no significant difference in each evaluation index of the model, indicating the universal applicability of the model. Particularly for more vegetated areas and areas covered by water, the model is able to accurately estimate precipitation.

In conclusion, the precipitation retrieval model established in this study can better determine precipitation regions and estimate precipitation intensity compared with the FY-4B/AGRI operational precipitation product. It can provide some reference value for users carrying out future precipitation retrieval research on FY-4B/AGRI and those who need information on dynamic changes in precipitation with a high spatial and temporal resolution.

It should be pointed out this study only trained and verified the precipitation data in the summer, and the model may not perform well in other seasons of the year. In the future, we will consider constructing a precipitation retrieval model with a longer applicable period. Our results also showed that different underlying surfaces do have an impact on the retrieval of precipitation data. We can consider to include these factors as variables in the model. In addition, the dataset is split in two subsets, day and night, according to the solar zenith angle. Thus, the day model and the night model were established separately. However, it should be noted that since China spans over four to five timezones, it is very likely that the night samples are affected by sunlight, and vice versa. We will divide the time period into day, dusk and night or consider a unified model for the entire day in later works. In this study, a resampling method is adopted to change the distribution of samples so as to improve the ability of the precipitation retrieval model to identify a few types of samples. This method sacrifices the accuracy of the majority of class samples. In the future, more suitable sampling methods or retrieval algorithms will be considered to further improve the model's retrieval accuracy for heavy rain and torrential rain.

Author Contributions: Conceptualization, Y.H. (Yang Huang) and Y.B.; methodology, Y.H. (Yang Huang) and Y.B.; software, Y.H. (Yang Huang); validation, Y.H. (Yang Huang); formal analysis, Y.H. (Yang Huang) and G.P.P.; data curation, Y.H. (Yang Huang); writing—original draft preparation, Y.H. (Yang Huang); writing—review and editing, Y.B., Q.L., G.P.P., Y.H. (Yanfeng Huo) and F.W.; visualization, Y.H. (Yang Huang); supervision, Y.B.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (2022YFC3004102), the Natural Science Foundation of China (U2242212), the Fengyun Application Pioneering Project (FY-APP-2022.0603), the Fengyun Application Pioneering Project (2022) Xu Jianmin Meteorological Satellite Innovation Center Project (FY-APP-ZX-2022.0208), the Water Science and Technology Project of Jiangsu Province (2023022) and the Shanghai Aerospace Science and Technology Innovation Foundation (SAST2021-032).

Data Availability Statement: FY-4B data are available free of charge upon registration at the FENGYUN Satellite Data Center (https://satellite.nsmc.org.cn/ (accessed on 3 August 2023)). GPM IMERG data are available at the Precipitation Data Directory (https://gpm.nasa.gov/data/directory (accessed on 9 October 2023)) upon registration. All other data will be made available upon request.

Acknowledgments: The authors would like to thank the China National Satellite Meteorological Center and Global Precipitation Measurement website, which are freely accessible to the public. Also, the authors would like to sincerely thank Python for offering powerful analysis computational tools such as scikit-learn. The authors would like to thank the editor and anonymous reviewers for their constructive comments and suggestions for improving the manuscript.

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

References

1. Zhang, H.; Xu, Y.; Dou, S. TRMM Downscaling Data of Yangtze Based on GWR Model. *Res. Soil Water Conserv.* **2021**, *28*, 149–162. [CrossRef]
2. Zhang, G.; Li, Z.; Song, Y.; Wu, Y.; Wang, X. Spatial Patterns of Chang Trend in Rainfall of China and Role of East Asia Summer Monsoon. *Arid Land Geogr.* **2011**, *1*, 34–42. [CrossRef]
3. Wu, J.; Dong, W.; Zhang, Y.; Chen, Y.; Xu, H.; Chen, X. Application of Multi-source Rainfall Data in the Flash Flood Forecast of Guanshan River Basin. *Eng. J. Wuhan Univ.* **2021**, *54*, 72–81. [CrossRef]
4. Richard, F. *Quantitative Precipitation Estimation in the National Weather Service, Hydrology Laboratory, Office of Hydrologic Development, National Weather Service, 3 April 2023*; National Weather Service: Silver Spring, Maryland, USA, 2023. Available online: https://hdsc.nws.noaa.gov/pub/hdsc/data/papers/articles/hrl/papers/wsr88d/MPE_workshop_NWSTC_lecture1_121305.pdf (accessed on 20 July 2023).
5. Sokol, Z.; Szturc, J.; Orellana-Alvear, J.; Popová, J.; Jurczyk, A.; Célleri, R. The Role of Weather Radar in Rainfall Estim-ation and Its Application in Meteorological and Hydrological Modelling-A Review. *Remote Sens.* **2021**, *13*, 351. [CrossRef]
6. De Coning, E.; Poolman, E. South African Weather Service operational satellite based precipitation estimation technique: Applications and improvements. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1131–1145. [CrossRef]
7. Adler, R.F.; Negri, A.J. A Satellite Technique to Estimate Tropical Convective and Stratiform Rainfall. *J. Appl. Meteorol.* **1988**, *27*, 30–51. [CrossRef]
8. Levizzani, V. Satellite Rainfall Estimations: New Perspectives for Meteorology and Climate from the EURAINSA T Project. *Ann. Geophys.* **2003**, *46*, 363–372. [CrossRef]
9. Ebert, E.E.; Janowiak, J.E.; Kidd, C. Comparison of Near-real-time Precipitation Estimates from Satellite Observations and Numerical Models. *Bull. Am. Meteorol. Soc.* **2007**, *88*, 47–64. [CrossRef]
10. Arkin, P.A.; Meisner, B.N. The Relationship between Large-scale Convective Rainfall and Cold Cloud over the Western Hemisphere during 1982–1984. *Mon. Weather Rev.* **1987**, *115*, 51–74. [CrossRef]
11. Kidd, C.; Kniveton, D.R.; Todd, M.C.; Bellerby, T.J. Satellite Rainfall Estimation Using Combined Passive Micro-wave and Infrared Algorithms. *J. Hydrometeorol.* **2003**, *4*, 1088–1104. [CrossRef]
12. Nauss, T.; Kokhanovsky, A.A. Discriminating Raining from Non-raining Clouds at Mid-latitudes Using Multispectral Satellite Data. *Atmos. Chem. Phys.* **2006**, *6*, 5031–5036. [CrossRef]
13. Roebeling, R.A.; Holleman, I. SEVIRI Rainfall Retrieval and Validation Using Weather Radar Observations. *J. Geophys. Res.* **2009**, *114*, D21202. [CrossRef]
14. Kühnlein, M.; Thies, B.; Nauss, T.; Bendix, J. Rainfallrate Assignment Using MSG SEVIRI Data—A Promising Approach to Spaceborne Rainfall Rate Retrieval for Midlatitudes. *J. Appl. Meteorol. Clim.* **2010**, *49*, 1477–1495. [CrossRef]
15. Feidas, H.; Giannakos, A. Identifying Precipitating Clouds in Greece Using Multispectral Infrared Meteosat Second Generation Satellite Data. *Theor. Appl. Clim.* **2011**, *104*, 25–42. [CrossRef]
16. Rivolta, G.; Marzano, F.S.; Coppola, E.; Verdecchia, M. Artificial Neural-network Technique for Precipitation Now-casting from Satellite Imagery. *Adv. Geosci.* **2006**, *7*, 97–103. [CrossRef]
17. Kühnlein, M.; Appelhans, T.; Thies, B. Improving the Accuracy of Rainfall Rates from Optical Satellite Sensors with Machine Learning-A Random Forest-Based Approach Applied to MSG SEVIRI. *Remote Sens. Environ.* **2014**, *141*, 129–143. [CrossRef]
18. Lazri, M.; Ameur, S. Combination of Support Vector Machine, Artificial Neural Network and Random Forest for Improving the Classification of Convective and Stratiform Rain Using Spectral Features of SEVIRI Data. *Atmos. Res.* **2017**, *203*, 118–129. [CrossRef]
19. Ma, L.; Zhang, G.P.; Lu, E. Using the Gradient Boosting Decision Tree to Improve the Delineation of Hourly Rain Areas during the Summer from Advanced Himawari Imager Data. *J. Hydrometeorol.* **2018**, *19*, 761–776. [CrossRef]
20. Min, M.; Bai, C.; Guo, J. Estimating Summertime Precipitation from Himawari-8 and Global Forecast System Based on Machine learning. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2557–2570. [CrossRef]
21. Hirose, H.; Shige, S.; Yamamoto, M.K.; Higuchi, A. High Temporal Rainfall Estimations from Himawari-8 Multiband Observations Using the Random-forest Machine-learning Method. *J. Meteorol. Soc. Jpn.* **2019**, *97*, 689–710. [CrossRef]
22. Kong, X.; Li, C.; Chen, X. Precipitation Retrieval Based on Multi-channel Data of Himawari-8 Satellite in Hedong Area of Gansu Province. *J. Meteorol. Res. Appl.* **2020**, *41*, 54–60. [CrossRef]
23. Wang, G.; Wang, D.; Wu, R. Application Study of Himawari-8/AHI Infrared Spectral Data on Precipitation Signal Recognition and Retrieval. *J. Infrared Millim. Waves* **2020**, *39*, 251–262. [CrossRef]
24. Zhang, Y.; Wu, K.; Zhang, J. Estimating Rainfall with Multi-Resource Data over East Asia Based on Machine Learning. *Remote Sens.* **2021**, *13*, 3332. [CrossRef]
25. Guan, L.; Zhong, Y. Retrieval of Surface Rainfall Using Random Forest Algorithm Based on FY-4A AGRI Observations. *Prog. Geophys.* **2023**, *38*, 1931–1938. (In Chinese) [CrossRef]
26. Ren, Y.; Yong, B.; Lu, D.; Chen, H. Evaluation of the Integrated Multi-satellite Retrievals (IMERG) for Global Precipitation Measurement (GPM) Mission over the Mainland China at Multiple Scales. *J. Lake Sci.* **2019**, *31*, 560–572. [CrossRef]
27. Xiao, L.; Zhang, A.; Min, C. Evaluation of GPM Satellite-based Precipitation Estimates during Three Tropical-related Extreme Rainfall Events. *Plateau Meteorol.* **2019**, *38*, 993–1003. [CrossRef]

28. Shi, L.; Feng, W.; Lei, Y.; Wang, Z.; Zheng, Q. Accuracy evaluation of daily GPM precipitation product over Mainland China. *Meteorol. Mon.* **2022**, *48*, 1428–1438. [CrossRef]
29. You, R. Satellite Quantitative Precipitation Estimation Method. In Proceedings of the 35th Annual Meeting of the Chinese Meteorological Society S21 Satellite Meteorology and Ecological Remote Sensing, Hefei, China, 24 October 2018.
30. Liu, L.; Zhang, X.; Gao, Y.; Chen, X.; Shuai, X.; Mi, J. Finer-Resolution Mapping of Global Land Cover: Recent Developments, Consistency Analysis, and Prospects. *J. Remote Sens.* **2021**, *2021*, 5289697. [CrossRef]
31. Shuai, C.; Sha, J.; Lin, J. Spatial Difference of the Relationship between Remote Sensing Index and Land Surface Temperature under Different Underlying Surfaces. *J. Geo-Inf. Sci.* **2018**, *20*, 1657–1666. [CrossRef]
32. Wang, S.; Cui, P.; Zhang, P. FY-3C/VIRR SST Algorithm and Cal/Val Activities at NSMC/CMA. In *Ocean Remote Sensing and Monitoring from Space*; SPIE: Bellingham, WA, USA, 2014; pp. 94–101. [CrossRef]
33. Deng, N.; Cui, Y.; Guo, Y. Frequency Ratio-random Forest-model-based Landslide Susceptibility Assessment. *Sci. Technol. Eng.* **2020**, *20*, 13990–13996. [CrossRef]
34. Wang, W.; Yao, Z.; Jia, S. Application Research on Random Forest Algorithm in the Statistical Test of Rainfall Enhancement Effect. *Meteorol. Environ. Sci.* **2018**, *41*, 111–117. [CrossRef]
35. Lazri, M.; Ameur, S.; Mohia, Y. Instantaneous Rainfall Estimation Using Neural Network from Multispectral Observations of SEVIRI Radiometer and its Application in Estimation of Daily and Monthly Rainfall. *Adv. Space Res.* **2014**, *53*, 138–155. [CrossRef]
36. Thies, B.; Nauss, T.; Bendix, J. Precipitation Process and Rainfall Intensity Differentiation Using Meteosat Second Generation Spinning Enhanced Visible and Infrared Imager Data. *J. Geophys. Res.* **2008**, *113*, D23206. [CrossRef]
37. Ackerman, S.A.; Strabala, K.I.; Menzel, W.P.; Frey, R.A.; Moeller, C.C.; Gumley, L.E. Discriminating Clear Sky from Clouds with MODIS. *J. Geophys. Res.* **1998**, *103*, 32141–32157. [CrossRef]
38. Sun, S.; LI, W.; Huang, Y. Retrieval of Precipitation by Using Himawari-8 Infrared Images. *Acta Sci. Nat. Univ. Pekinenis* **2019**, *55*, 215–226. [CrossRef]
39. Behrangi, A.; Hsu, K.L.; Imam, B. Evaluating the Utility of Multispectral Information in Delineating the Areal Extent of Precipitation. *J. Hydrometeorol.* **2009**, *10*, 684–700. [CrossRef]
40. Fritz, S.; Laszlo, I. Detection of Water Vapor in the Stratosphere over Very High Clouds in the Tropics. *J. Geophys. Res. Atmos.* **2012**, *98*, 22959–22967. [CrossRef]
41. Baum, B.A.; Platnick, S. Introduction to MODIS Cloud Products. In *Earth Science Satellite Remote Sensing*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 74–91. [CrossRef]
42. Mecikalski, J.R.; Bedka, K.M. Forecasting Convective Initiation by Monitoring the Evolution of Moving Cumulus in Daytime GOES Imagery. *Mon. Weather Rev.* **2006**, *134*, 49–78. [CrossRef]
43. Zeng, L.; Gao, Y.; Jiang, Y. Scale Effects of Terrain Factors on Precipitation in East China. *Adv. Earth Sci.* **2022**, *37*, 535–548. [CrossRef]
44. Lei, X.; Zhang, G.; Yao, Q. Research on Automatic Recognition of Agricultural Machine Image Based on Convolutional Neural Network. *J. Chin. Agric. Mech.* **2022**, *43*, 140–147. [CrossRef]
45. WWRP/WGNE Joint Working Group on Forecast Verification Research Forecast Verification: Issues, Methods and FAQ. Available online: http://www.cawcr.gov.au/projects/verification/ (accessed on 7 January 2015).
46. Ma, S.; Chen, C.; He, H. Experiment and Verification of the Convective-scale Ensemble Forecast Based on BGM. *Plateau Meteorol.* **2018**, *37*, 495–504. [CrossRef]
47. *GB/T 28592–2012*; Grade of Precipitation. National Meteorological Center: Beijing, China, 2012.
48. Toté, C.; Patricio, D.; Boogaard, H.; Vander Wijngaart, R.; Tarnavsky, E.; Funk, C. Evaluation of Satellite Rainfall Estimates for Drought and Flood Monitoring in Mozambique. *Remote Sens.* **2015**, *7*, 1758–1776. [CrossRef]
49. Zhong, Y. Testing and evaluation of quantitative precipitation estimation product from Fengyun 4 satellite. *J. Agric. Catastrophol.* **2021**, *11*, 96–98. [CrossRef]