TYPE Original Research
PUBLISHED 11 November 2025
DOI 10.3389/feart.2025.1669417

Improving prediction of short-duration heavy rainfall in Guangxi, China during the pre-summer rainy season based on Fengyun-4A lightning frequency and a machine learning algorithm

Weixiang Huang[1], Xiaoli Luo[2]*, Wei Zhang[1], Bo Feng[1], Xiaofei Xia[1], Chenying Yi[1] and Weijun Hu[3]

[1]Guangxi Power Grid Equipment Monitoring and Diagnosis Engineering Technology Research Center, Electric Power Research Institute of Guangxi Power Grid Co., Ltd., Nanning, China, [2]Guangxi Meteorological Observatory, Nanning, China, [3]Fangchenggang Power Bureau of Guangxi Power Grid Co., Ltd., Fangchenggang, China

Check for updates

OPEN ACCESS

EDITED BY
Ismail Gultepe,
Ontario Tech University, Canada

REVIEWED BY
Yali Zhu,
Chinese Academy of Sciences (CAS), China
Amar Jyothi Kasimahanthi,
National Centre for Medium Range Weather Forecasting, India

*CORRESPONDENCE
Xiaoli Luo,
✉ xlpeace2003@126.com

RECEIVED 19 July 2025
REVISED 11 October 2025
ACCEPTED 20 October 2025
PUBLISHED 11 November 2025

CITATION
Huang W, Luo X, Zhang W, Feng B, Xia X, Yi C and Hu W (2025) Improving prediction of short-duration heavy rainfall in Guangxi, China during the pre-summer rainy season based on Fengyun-4A lightning frequency and a machine learning algorithm. *Front. Earth Sci.* 13:1669417. doi: 10.3389/feart.2025.1669417

Introduction: This study investigated the relationships between short-duration heavy rainfall (SDHR) events and lightning activity over Guangxi, China, during the pre-summer rainy season from 2019 to 2023.

Methods: The analysis was conducted using the satellite-retrieved IMERG precipitation dataset and the Fengyun-4A lightning mapping imager (FY-4A/LMI). We employed a Random Forest machine learning model to assess the value of one-hour antecedent lightning frequency as a predictor for SDHR. The model's interpretability was further examined using SHapley Additive exPlanation (SHAP) value analysis.

Results: The results revealed distinct spatiotemporal variations. SDHR events first occurred in eastern Guangxi in April, expanded westwards in May, and covered the entire region by June. Lightning activity peaked in April, decreased in May, and increased again in June. Both exhibited a unimodal diurnal cycle, peaking at nocturnal-to-morning hours; however, SDHR intensity consistently reached its maximum at 21:00 UTC, approximately one hour later than the lightning peak. The mean number of lightning flashes per SDHR event decreased from 8.58 in April to 6.14 in May and 6.10 in June. Incorporating lightning frequency into the Random Forest model substantially enhanced SDHR prediction accuracy, reducing the mean absolute error by 4.42% (April), 6.02% (May), and 4.29% (June). The coefficient of determination (R^2) for SDHR amount increased from 0.29 to 0.35 in April, 0.38 to 0.45 in May, and 0.22 to 0.29 in June.

Discussion: SHAP analysis confirmed the positive contribution of lightning frequency to rainfall intensity prediction throughout the entire study period. This positive contribution exhibited a monotonically increasing trend when lightning frequency was below the threshold of 15, and lightning frequency was found to amplify its influence on the model output through interactions with other predictors. Collectively, these results underscore the value of lightning observations as robust predictors for improving short-term heavy rainfall forecasts.

KEYWORDS

short-duration heavy rainfall, FY-4A/LMI, random forest, SHAP value, feature interaction

1 Introduction

Short-duration heavy rainfall (SDHR) events are major meteorological hazards that frequently trigger flash floods, urban inundation, and landslides (Brooks and Stensrud, 2000; Yuan et al., 2014), thus posing significant threats to human life and socioeconomic development. Globally, observational evidence indicates a marked increase in heavy rainfall frequency over recent decades under climate warming (Chiappa et al., 2024; Han et al., 2025), with approximately 23% of the world's population now exposed to one-in-100-year flood risks (Rentschler et al., 2022). This trend is particularly pronounced in southeastern China, where extreme hourly precipitation events have shown significant intensification and have occurred predominantly during the afternoon and midnight hours (Wang et al., 2023).

Thunderstorm systems often produce both lightning and heavy rainfall simultaneously, and their combined effects can significantly amplify societal impacts—primarily through power infrastructure disruptions (Liu et al., 2017) and intensified flash flooding (Bahari et al., 2023). Due to these compounded hazards, the relationship between lightning activity and heavy rainfall has been widely investigated. Early studies using Tropical Rainfall Measuring Mission (TRMM) data identified a strong correlation between total lightning flash rates and rainfall rates in winter thunderstorms over the eastern Mediterranean, particularly at monthly and seasonal scales (Price and Federmesser, 2006). Further spatial analyses have demonstrated that all lightning types exhibit consistent relationships with rainfall, though positive cloud-to-ground flashes are more strongly linked to precipitation amounts than negative flashes (Soula and Chauzy, 2000). Microphysical differences between convective systems with and without lightning are particularly notable: lightning-producing storms typically exhibit stronger updrafts, deeper cloud depths, and more active ice-phase processes compared to non-lightning storms (Matthee et al., 2014). Recent observations reveal that approximately 43% of SDHR events coincide with lightning activity, with the strongest correlation occurring when precipitation follows lightning by 5–10 min (Zhou K. H. et al., 2022). Given these findings, significant effects have been made to incorporate by lightning data into heavy rainfall forecasting. Lightning data assimilation has been widely used to enhance convective weather prediction, demonstrating a positive impact on short-term precipitation forecasts (Torcasio et al., 2021). For instance, integrating Fengyun-4A (FY-4A) lightning data into numerical model improves cloud initialization, leading to better 12-h precipitation forecasts, with particularly notable improvements in nowcasting (1–2 h) (Liu et al., 2019; Xu et al., 2020). However, numerical weather prediction (NWP) methods still exhibit considerable uncertainty in quantitative precipitation forecasting due to the complex physical mechanisms involved. Recent studies suggest that machine learning (ML) approaches can outperform traditional NWP models in heavy rainfall prediction (Zhou K. H. et al., 2022). Lightning data has proven particularly valuable in ML-based thunderstorm nowcasting (Leinonen et al., 2022) and enhances the robustness of large hail prediction (Czernecki et al., 2019). These findings highlight the potential of lightning observations—especially in data-sparse regions—to improve convective rainfall estimation (Kochtubajda et al., 2013).

The pre-summer rainy season (PSRS) in South China, spanning April to June, marks the onset of the East Asian summer monsoon (Huang et al., 2012; Xiao et al., 2025). This period is characterized by frequent intense heavy rainfall (Lu et al., 2022), often triggering devastating flash floods and significant economic losses. Concurrently, lightning activity surges from April to June, making South China as the country's most lightning-prone region and resulting in severe lightning-related casualties and infrastructure damage (Zhang et al., 2011; Guan et al., 2024). Guangxi, a major province in South China, exhibits pronounced spatiotemporal variations in both SDHR events and lightning activity (Jiang et al., 2017; Lu et al., 2022; Guan et al., 2024; Guo et al., 2024). This study investigates the climatological characteristics of lightning activity and SDHR events in Guangxi, along with their interrelationships. Furthermore, we evaluate the utility of lightning data in improving SDHR prediction using a machine learning approach.

This paper is structured as follows: Section 2 describes the data and methodology. Section 3 presents the spatial distributions of SDHR events and lightning activity during the PSRS. Section 4 analyzes their diurnal variations, while Section 5 explores SDHR prediction using machine learning. Finally, Section 6 summarizes the key findings.

2 Study area, datasets, and methodology

2.1 Study area

This study focuses on Guangxi Province (*104°26′–112°04′E, 20°54′–26°20′N*), which is located in southern China (Figure 1). The region exhibits distinct topographic characteristics: it is bounded by the Yunnan-Guizhou Plateau to the northwest and the Beibu Gulf to the south, resulting in a general topographic gradient that descends from northwest to southeast. The northeastern area features Mao'er Mountain, which has the highest peak in Guangxi, with an elevation of 2,142 m, whereas the central and southern regions comprise predominantly hilly and mountainous terrain (<800 m) with complex meso- and microscale topographic features. These pronounced elevation differences significantly enhance topographic lift effects, which are known to play a crucial role in heavy rainfall formation in Guangxi (Liao et al., 2022). Guangxi

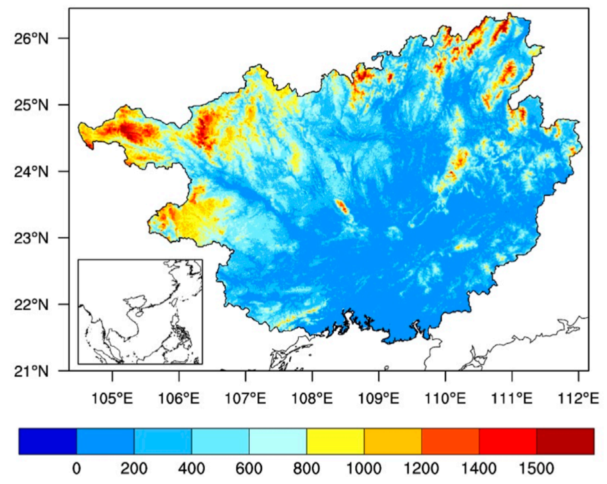

FIGURE 1
Location and topographic height of Guangxi Province, China (shaded, units: m).

experiences intense precipitation during the PSRS, with northern Guangxi representing one of the primary heavy rainfall centres in this season (Luo et al., 2020).

2.2 Data

The Integrated Multi-satellite Retrievals for Global Precipitation Measurement (GPM-IMERG) dataset, which was developed by the U.S. GPM team (Huffman et al., 2020), is a level-3 gridded precipitation product and it is freely available through the website (https://disc.gsfc.nasa.gov/). This study used the V07A final run product, which provides high-resolution precipitation estimates at 0.1° spatial resolution and 30-min temporal intervals. Extensive validation studies have demonstrated that IMERG is one of the most accurate high-resolution satellite precipitation products currently available (Tang et al., 2020; Pradhan et al., 2022), with particularly good performance in South China's complex climatological conditions (Zhang et al., 2023).

The Fengyun-4A Lightning Mapping Imager (FY-4A/LMI) dataset provides lightning observations with a nadir spatial resolution of 7.8 km. This study used level 2 products that are publicly available through the National Satellite Meteorological Center of China (http://satellite.nsmc.org.cn/DataPortal/cn/home/index.html). The LMI instrument characteristics and data processing algorithms are detailed in the manuscripts of Hui and Guo (2021) and Hui et al. (2023). The dataset comprises three hierarchical lightning products: single "event" (LMIE), "group" (LMIG), and "flash" data. Each LMIE represents an individual lightning signal with an associated occurrence time, location, and spectral radiance, whereas an LMIG corresponds to either a complete lightning strike or a K-change process during intracloud flashes. Statistical analyses reveal consistent ratios among these products of approximately 9:3:1 (LMIE:LMIG:Flash) (Chen et al., 2021). Recent validation studies have demonstrated the reliability of LMI observations. Wu et al. (2024) reported that most LMIGs contain fewer than 10 LMIEs, with a mean value of 3.6 LMIEs per LMIG. The LMI data show excellent consistency with ground-based lightning detection data from the China Lightning Location Network (LLNC). The number of LMI groups is also close to the number of ground-based LLNC cloud-to-ground events (Cao et al., 2021). While LMI detects more intense lightning activity in northern Guangxi and river valleys of southwestern Guangxi than the ground-based China Meteorological Administration (CMA) Lightning Detection Network Advanced TOA and Direction (ADTD) system data, Pan (2023) confirmed that this discrepancy does not preclude a strong agreement in their monthly and diurnal variations. Given this demonstrated reliability and the close correspondence between LMIG counts and LLNC-detected cloud-to-ground events, we employed the LMIG dataset in this study to investigate the relationship between the LMIG and heavy precipitation events.

This study also utilizes the ERA5-Land global reanalysis dataset (Joaquín et al., 2021), a state-of-the-art reanalysis product specifically designed for land surface applications, including 10 variables such as 2-m air temperature, 2-m dewpoint temperature, surface latent heat flux, surface sensible heat flux, 10-m zonal wind, 10-m meridional wind, 10-m wind speed, surface pressure, surface albedo, and skin temperature. The dataset, with a high horizontal resolution of 0.1 ° × 0.1 ° and hourly intervals, was employed to drive the ML model. All analyses focus on the April–June periods during 2019–2023, with temporal references consistently reported in Coordinated Universal Time (UTC) throughout the study.

2.3 Methodology

In this study, short-duration heavy rainfall (SDHR) events were defined as events in which the precipitation exceeds 20 mm within 1 hour in Guangxi (Lu et al., 2022), established by the China Meteorological Administration. To identify these events, we derived hourly IMERG precipitation data by accumulating native half-hourly products. And the hourly precipitation from satellite-based IMERG in a specific grid cell greater than 20 mm/h will be recorded as a SDHR event. Similarly, we processed the LMIE lightning observations by regridding them to 0.1° resolution and accumulating them to hourly totals.

This study employs the Random Forest (RF) algorithm to predict the SDHR in Guangxi. As an ensemble learning method, RF combines bootstrap aggregation (bagging) with random subspace selection, constructing multiple decision trees through the Classification and Regression Trees (CART) algorithm. This approach demonstrates particular advantages in handing high-dimensional feature spaces and complex nonlinear relationships, making it particularly suitable for precipitation prediction (Wolfensberger et al., 2021; Anco-Valdivia et al., 2025). The analysis incorporates 28,537, 37,184, and 60,591 SDHR events for April, May, and June, respectively. For each month, the dataset was randomly partitioned into training (70%) and testing (30%) subsets.

Given the coarser resolution (0.25 ° × 0.25 °) of atmospheric variables (e.g., vertical velocity, geopotential height) from ERA5, which is inadequate for extracting discriminative predictors over a small region experiencing multiple simultaneous heavy rainfall events, we opted not to incorporate high-level atmospheric variables

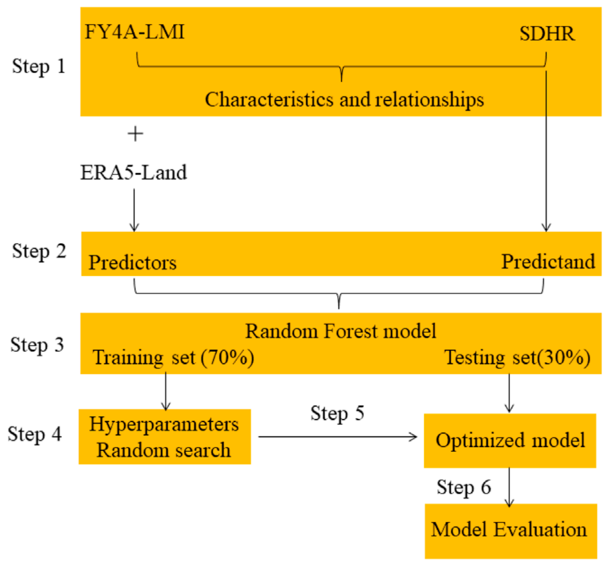

FIGURE 2
Flowchart for model construction.

as predictors. Instead, lightning frequency was selected due to its utility as a sensitive indicator of convective activity. And thirteen predictive variables were used to train the RF model, comprising (1) ten ERA5-Land derived variables mentioned in Section 2.2. (2) Lightning frequency from FY-4A satellite observations. (3) Geographic coordinates (longitude and latitude) to account for spatial variability.

As shown in the flowchart in Figure 2, the construction of the RF model involves four key steps: (1) extraction of predictors from ERA5-Land and lightning frequency data corresponding to the SDHR events; (2) random partitioning of the dataset for each month into training (70%) and testing (30%) subsets; (3) utilization of the training set to train the RF model, with hyperparameters optimized via a random search procedure; and (4) final evaluation of the optimized model's performance using the testing set. Wu et al. (2017) reported that short-duration rainfall events correlate most strongly with lightning flashes within a narrow temporal window of less than 25 min before and after rainfall initiation. Similarly, Zhou Q. L. et al. (2022) observed that the lightning-rainfall relationship is statistically significant only within a 1-h timeframe. Therefore, in order to achieve the maximum forecast lead time, the lightning frequency leading the SDHR events by 1 hour was selected as the primary predictor for the RF model. Similarly, all other meteorological variables were incorporated as hourly data, temporally aligned to precede the events by 1 hour.

Model performance was quantitatively assessed using four metrics: mean square error (MSE) (Equation 1), root mean square error (RMSE) (Equation 2), mean absolute error (MAE) (Equation 3), and coefficient of determination (R^2) (Equation 4), calculated as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2 \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \tag{2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{m}|O_i - P_i| \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}\left(\overline{O_i} - O_i\right)^2} \tag{4}$$

where O and P denote the observed and predicted values, n is the number of samples. Overbars denote the mean values.

3 Spatial distribution of SDHR events and lightning density

Figures 3a–c present the five-year mean frequency of SDHR events from April to June. The spatial distribution reveals distinct monthly variations: in April (Figure 3a), SDHR events occurred primarily in eastern Guangxi, with a maximum frequency centre (<9 events) located in northeastern Guangxi. By May (Figure 3b), SDHR activity extended westwards, although it remained infrequent along coastal areas. In June (Figure 3c), SDHR events became widespread across all of Guangxi, with frequencies exceeding 12 events in northern Guangxi. The spatial patterns and event frequencies derived from the IMERG dataset show good agreement with ground-based observations from automatic meteorological stations (Lu et al., 2022; Liao et al., 2022), thus validating the reliability of the IMERG dataset for this analysis.

The distinct spatial distribution patterns of SDHR events from April to June can be attributed to topographic influences and seasonal transitions during the PSRS. The South China Sea summer monsoon (SCSSM) typically begins in mid-May (Peng et al., 2022) and provides a natural division of the PSRS into pre- and post-monsoon periods. These two phases exhibit fundamentally different precipitation mechanisms (Jiang et al., 2017; Peng et al., 2022; Xiao et al., 2025). During the pre-monsoon period, convective rainfall is triggered primarily by front systems that result from the convergence of midlatitude cold air masses and tropical warm air masses. The cold-dry air penetrates northeastern Guangxi through the low-lying Xiang-Gui Corridor (Figure 1), where it converges with warm-wet southerly flows to form quasi-stationary fronts (Liao et al., 2022). This frontal activity explains the concentration of SDHR events in northeastern Guangxi during April and May (Figures 3a,b). Following monsoon onset, the dominant precipitation mechanism shifts to orographic lifting of moisture-laden southerly winds. The prevailing monsoon flow transports abundant tropical moisture that is subsequently uplifted along the windward slopes of northern Guangxi (Figure 1), which generates widespread heavy rainfall events (Figure 3c). This transition accounts for the observed expansion of SDHR activity across the entire region in June.

The contribution of SDHR events to the total precipitation is substantial, with SDHR accounting for more than 40% of the monthly rainfall in certain regions from April to June (Figures 3d–f). The spatial patterns of these proportional contributions closely mirror the frequency distribution of SDHR events (Figures 3a–c).

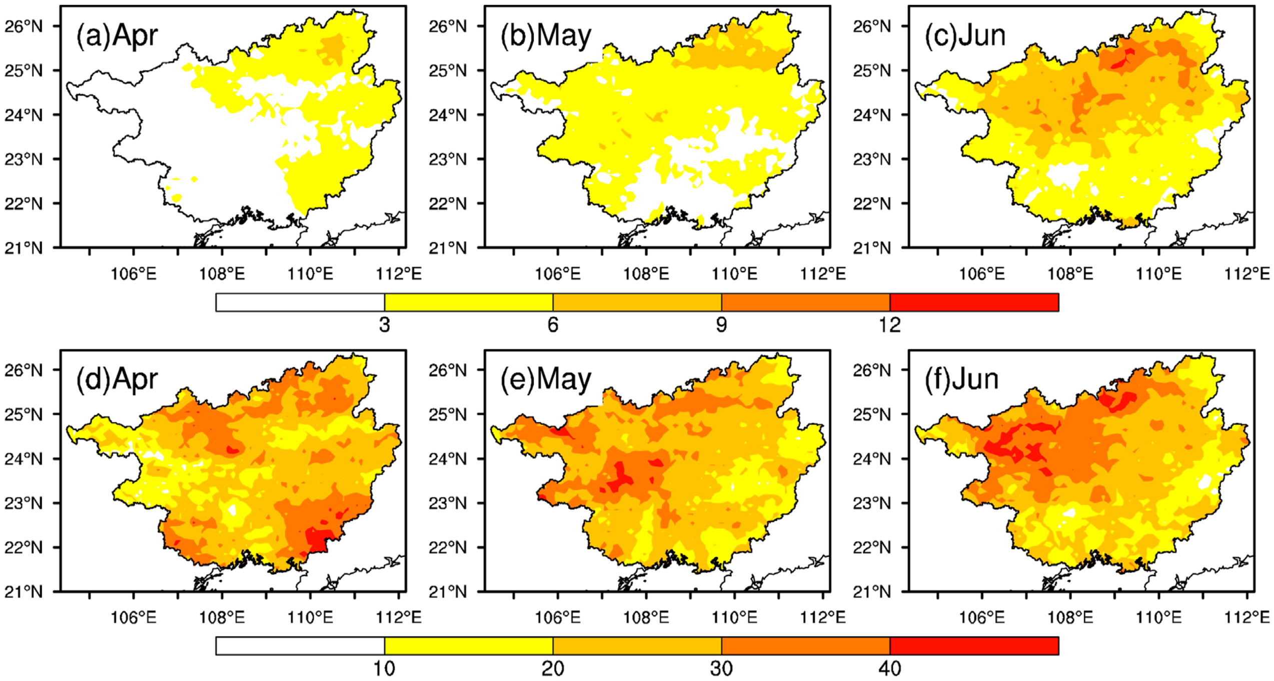

FIGURE 3
Number of SDHR events **(a–c)** and percentage of SDHR amount relative to the total monthly rainfall **(d–f**, units: %) in April **(a,d)**, May **(b,e)**, and June **(c,f)**.

Notably, in April (Figure 3d), while southeastern and southwestern Guangxi experienced fewer SDHR events, these intense rainfall episodes constituted a disproportionately large fraction (~40%) of the total precipitation. This heightened contribution underscores the critical role of SDHR in regional water budgets and its potential to trigger devastating flash floods (Wolfensberger et al., 2021). Similar patterns emerge in other subregions and months: SDHR represents substantial proportions of the total rainfall in midwestern Guangxi during May (Figure 3e) and northwestern Guangxi in June (Figure 3f), despite relatively lower event frequencies in these areas. These findings highlight how even infrequent SDHR events can dominate local precipitation regimes, particularly in orographically complex regions.

The spatial distributions of the lightning flash density associated with the SDHR events are presented in Figures 4a–c. Our analysis reveals strong spatial correlations between lightning flashes and SDHR events (Figures 3a–c), with spatial correlation coefficients of 0.67 (April), 0.56 (May), and 0.73 (June), which are statistically significant at the 95% confidence level. However, unlike the progressive increase in the SDHR frequency from April to June (Figures 3a–c) and the total lightning activity (Guan et al., 2024), the SDHR-related lightning density peaked in April, decreased in May, and resurged in June. This pattern aligns with findings of Ranalkar and Chaudhari (2009), who reported higher pre-monsoon lightning activity in South Asia.

Spatially, the lightning activity was concentrated in eastern Guangxi (~0.4 fL km^{-2}, Figure 4a) in April, shifted to the southwest in May (Figure 4b), and finally migrated to northern Guangxi in June (Figure 4c), which closely matched the SDHR distributions (Figure 3c). The peak lightning flash density exceeded 0.5 fL km^{-2} throughout the study period. Notably, the SDHR-related lightning accounted for more than 50% of total lightning activity (Figures 4d–f), even in southern Guangxi, where SDHR events were less frequent (Figures 3a,b). This predominance suggests that (1) lightning activity is predominantly associated with SDHR events and (2) weaker rainfall systems rarely produce substantial lightning.

The strong coupling between SDHR events and lightning activity is evident not only in their spatial distributions (Figures 3, 4) but also in their daily variations (Figure 5). Statistical analysis revealed significant positive correlations (r > 0.6, p-value <0.05) between daily SDHR frequency and lightning counts in regions with frequent SDHR activity (Figure 3). However, this relationship weakens considerably in southern Guangxi during April–June, which is likely due to the predominance of weaker convective systems in this region during these months. This spatial exception aligns with previous findings that southern Guangxi's primary SDHR season typically commences in July (Lu et al., 2022).

4 Diurnal variations

Consistent with previous findings (Lu et al., 2022; Liao et al., 2022), SDHR events in Guangxi demonstrated pronounced diurnal variability (Figure 6). Our analysis reveals similar nocturnal-dominated patterns during April–June, with peak occurrences consistently observed at 21:00 UTC. The diurnal cycle exhibited a unimodal distribution across all 3 months, with minimal afternoon activity. While April and May presented comparable daytime frequencies (~200 events), June showed a marked increase to approximately 350 events. Notably, the daily increase in SDHR

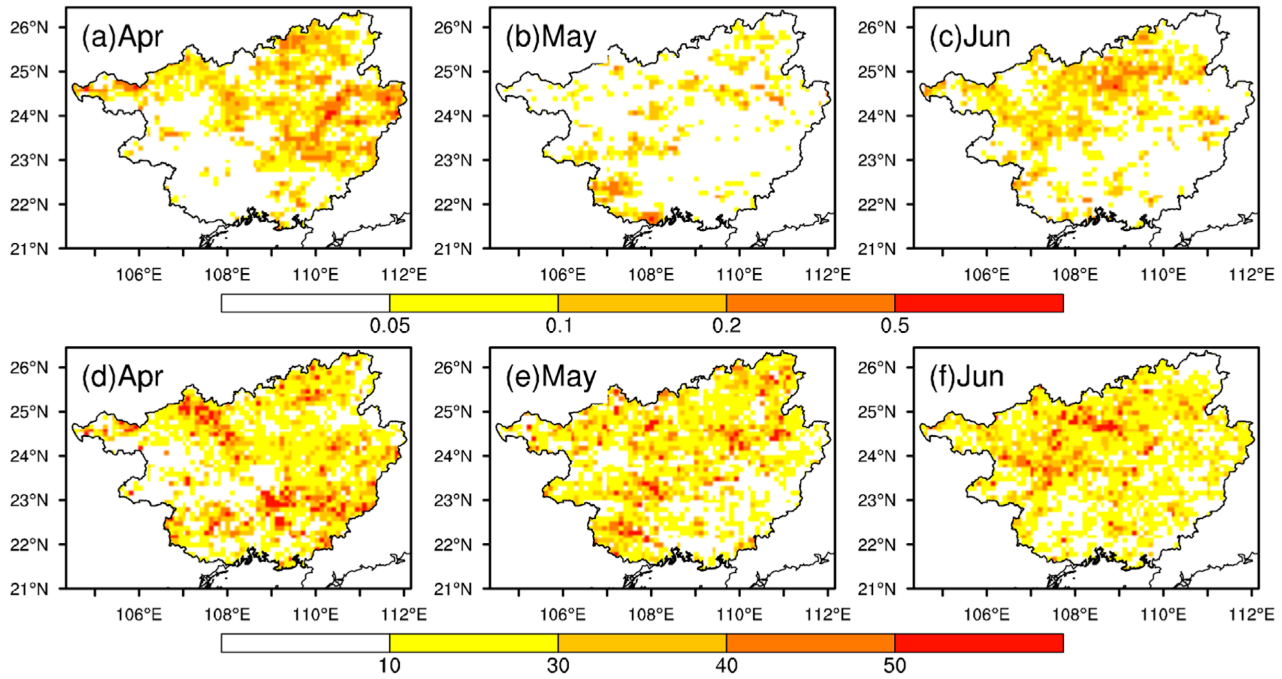

FIGURE 4
Density of the lightning frequency (**(a–c)**, units: fl km^{-2}) and percentage of the lightning frequency during SDHR events relative to the total lightning frequency **(d–f)**, units: %) in April **(a,d)**, May **(b,e)**, and June **(c,f)**.

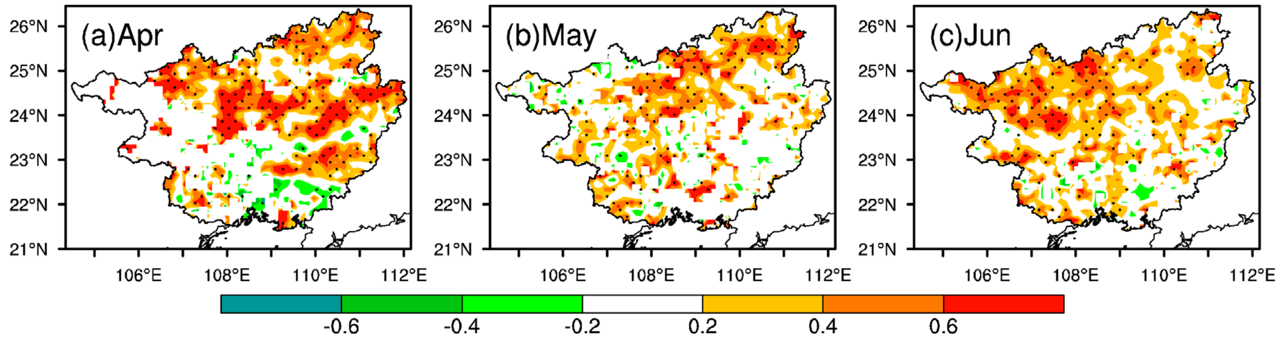

FIGURE 5
Correlation coefficients between the daily mean number of SDHR events and the frequency of lightning events for **(a)** April, **(b)** May, and **(c)** June. The dotted areas indicate that the coefficients are significant at the 95% confidence level.

activity began at approximately 14:00 UTC in all months, with event counts reaching their respective maxima of 502.8 (April), 758.0 (May), and 872.6 (June). This nocturnal enhancement may be attributable to the intensification of low-level southwesterly winds during nighttime hours (Jiang et al., 2017). Importantly, the diurnal patterns derived from IMERG data strongly agree with ground-based rain gauge observations (Lu et al., 2022), thus further validating the reliability of the IMERG dataset for SDHR detection.

The lightning activity associated with SDHR events exhibited diurnal variations remarkably similar to those of the SDHR events themselves (Figure 6), with both displaying distinct unimodal patterns peaking during nocturnal-to-morning hours. However, lightning activity consistently preceded SDHR events by approximately 1 hour, with maximum frequencies reached at 20:00 UTC compared with 21:00 UTC for SDHR. The peak lightning frequency showed significant monthly variation, with April having the highest value (2635.2 flashes), followed by June (1590.2 flashes) and May (1246.4 flashes). Daytime lightning activity (00:00–08:00 UTC) remained minimal throughout the study period (<100 flashes), which likely resulted from both reduced SDHR occurrence and the documented diurnal variation in FY-4A/LMI detection efficiency, which was notably greater during nighttime hours (Wu et al., 2024). Despite these variations, the temporal coupling between hourly time series of lightning frequency and SDHR events in Figure 6 remained

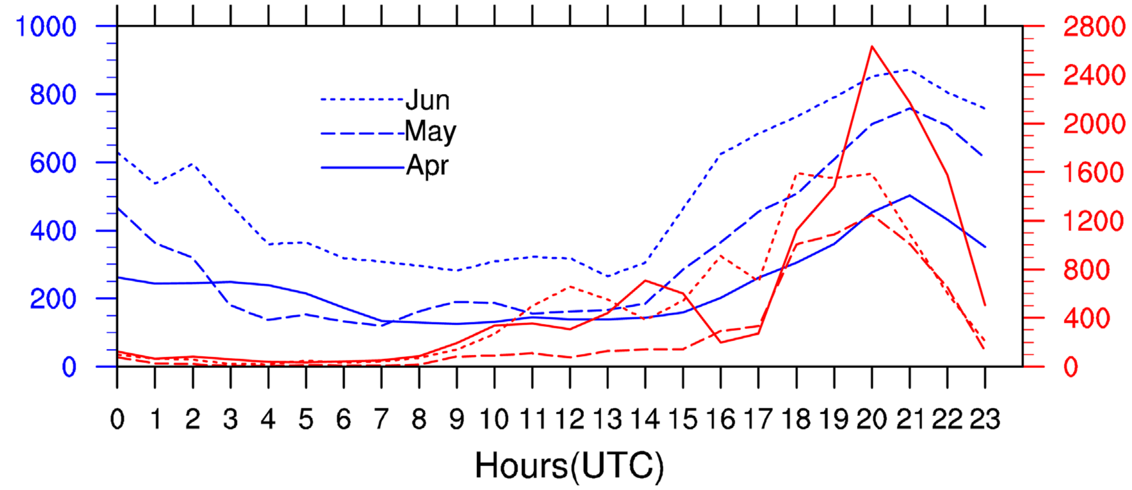

FIGURE 6
Diurnal variations in total SDHR events (blue lines) and lightning frequency (red lines).

statistically robust, with correlation coefficients of 0.79 (April), 0.81 (May), and 0.64 (June), all of which are significant at the 99% confidence level.

Figure 7 presents boxplots of the diurnal variations in the lightning frequency during SDHR events from April to June. As shown in Figure 7, the diurnal variations in the lightning frequency during a single SDHR event were more complicated than the total lightning frequency shown in Figure 6. Nocturnal SDHR events generally resulted in greater lightning activity than daytime events, with the latter typically exhibiting only about 2 flashes per event. In April (Figure 7a), the lightning frequency showed multiple peaks (5~7 flashes) at 09:00–10:00, 14:00–15:00, and 19:00–22:00 UTC, with various extreme events exceeding 25 flashes. May (Figure 7b) showed a single prominent peak (3~4 flashes) during 18:00–20:00 UTC, whereas June (Figure 7c) showed bimodal peaks (3~4 flashes) from 11:00–13:00 and 18:00–20:00 UTC. On average, each SDHR event produced 8.58, 6.14, and 6.10 lightning flashes in April, May, and June, respectively, indicating significantly increased electrical activity during April's pre-monsoon period. These variations likely reflect the effects of differences in convective intensity and environmental conditions on charge separation processes.

5 Prediction of the SDHR

Previous analyses have demonstrated a strong correlation between the SDHR and lightning frequency, both in terms of spatial distribution and diurnal variability. To quantitatively assess the contribution of lightning frequency to predict the SDHR, we conducted two distinct experiments using the RF model: one incorporating lightning frequency as a predictor and another excluding it, while maintaining identical configurations for all other model parameters.

Figure 8 presents a comparative evaluation of the two approaches, displaying scatter plots of predicted versus observed SDHR amounts during April–June. The results clearly indicate that predictions generated by the lightning-frequency-inclusive model exhibit significantly better agreement with observations, as evidenced by their closer alignment with the 1:1 line. This improvement is further quantified by consistently lower error metrics (MSE, RMSE, MAE) and higher R^2 in the lightning-frequency model compared to its counterpart. Specifically, integrating lightning frequency reduced prediction errors by 7.92% (MSE), 4.11% (RMSE), and 4.42% (MAE) in April; 12.06% (MSE), 6.24% (RMSE), and 6.02% (MAE) in May; and 8.33% (MSE), 4.25% (RMSE), and 4.29% (MAE) in June. Concurrently, the R^2 values improved from 0.29 to 0.35 (April), 0.38 to 0.45 (May), and 0.22 to 0.29 (June), underscoring the added predictive skill conferred by lightning frequency data.

To quantify the relative importance of individual predictors in the RF model, we employed the SHapley Additive exPlanation (SHAP) values (Lundberg and Lee, 2017). Figure 9 presents the mean absolute SHAP value–indicating predictor importance–and the SHAP values themselves, which reveal how low or high values of each predictor influence the model's output. The analysis focuses on the top 11 most influential predictors across the 3-month period.

As shown in Figure 9a, the RF model identifies 2-m surface air temperature (t2m) as the most critical predictor for April, followed by 10-m meridional wind speed (v10) and 10-m wind speed (wspd10). In May (Figure 9b), v10 emerges as the dominant factor, with longitude and 2-m dewpoint temperature (d2m) ranking second and third, respectively. For June (Figure 9c), the top three predictors are t2m, d2m, and surface latent heat flux (slhf). Notably, the most influential predictors in all 3 months exhibit an inverse relationship with SDHR, as higher values of these predictors correspond to negative SHAP values, suggesting a suppressive effect on heavy rainfall. Lightning frequency (lf), while less prominent in April and May (ranking 10th), rises to fourth place in June. Importantly, its SHAP values demonstrate a consistent positive correlation with SDHR amounts across all

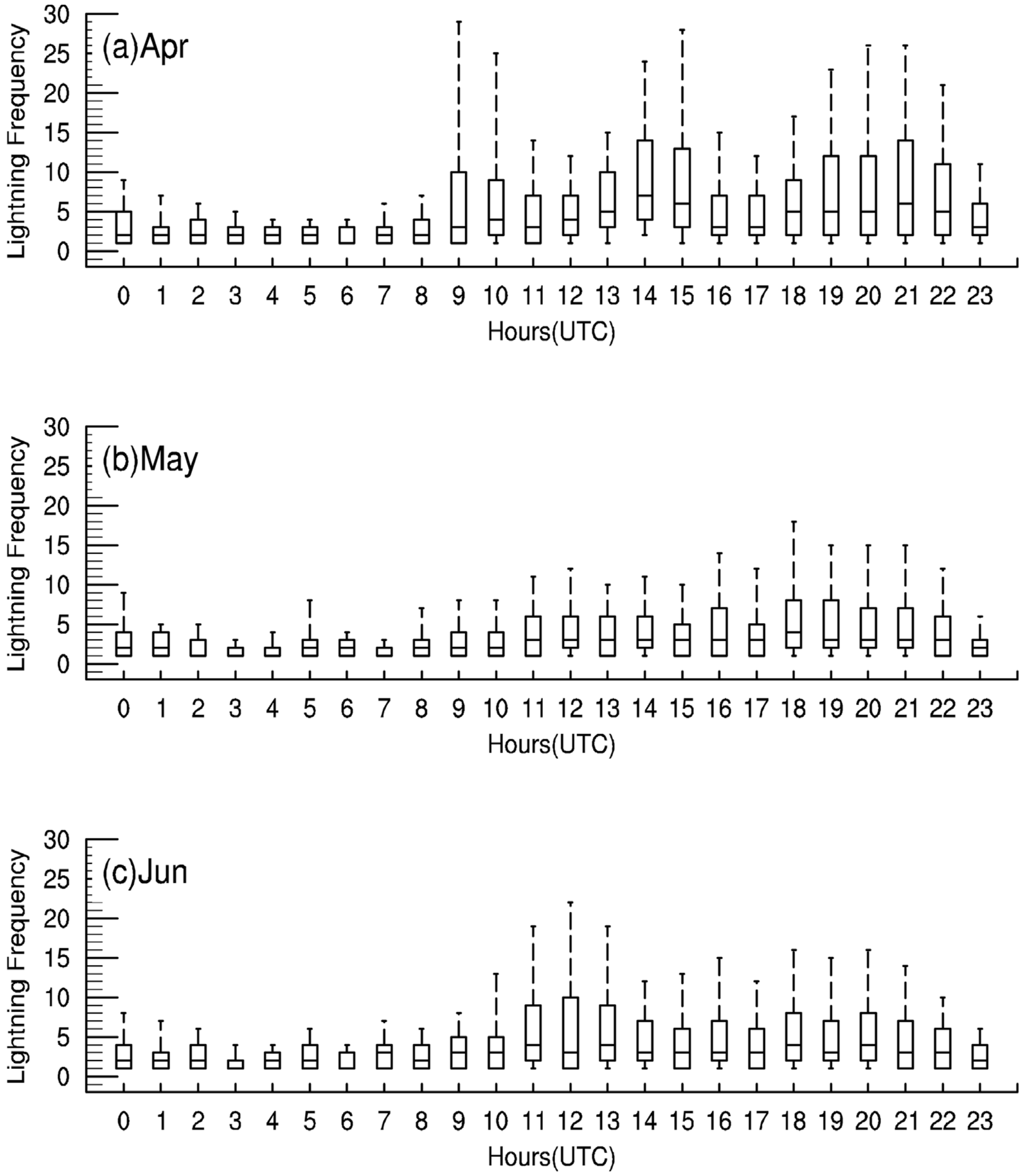

FIGURE 7
Box and whisker plots of the diurnal variations in the lightning frequency for single SDHR events in April **(a)**, May **(b)**, and June **(c)**. The centerline of each box indicates the median, and the bottom and top lines of the box indicate the 25th and 75th percentiles, respectively. The bottom and top of the dashed vertical lines indicate the 10th and 90th percentiles, respectively.

months: higher lightning frequency enhances the model's predictive capability by providing a physically meaningful signal for intense precipitation events.

Figure 9 demonstrates that higher lighting frequency is associated with heavier rainfall, despite its relative weak feature importance in April and May. A key question arises: Does the SDHR increase monotonically with lighting frequency? As depicted in Figure 10, lightning frequency exhibits a predominantly positive contribution to most SDHR events. However, SHAP values increase with lightning frequency only up to a threshold of ~15 during the PSRS, beyond which they stabilize despite further increases in lightning frequency. This suggests that lightning frequency has a monotonically positive influence on SDHR below this threshold but becomes negligible above it.

Another intriguing observation is how lighting frequency–a less important variable (ranked 10th in feature importance for April and May in Figure 9)–still contributes significantly to SDHR prediction. This can be partially explained by its interactions with other prominent variables (Figure 10). In April (Figure 10a), a notable interaction exists between lightning frequency and t2m (the most important variable). Surface warming enhances convective development and then produces more lightning frequency, which in turn favors heavier rainfall. Thus, the effect of lightning frequency on the model is amplified through its synergy with t2m. In May (Figure 10b), lightning frequency

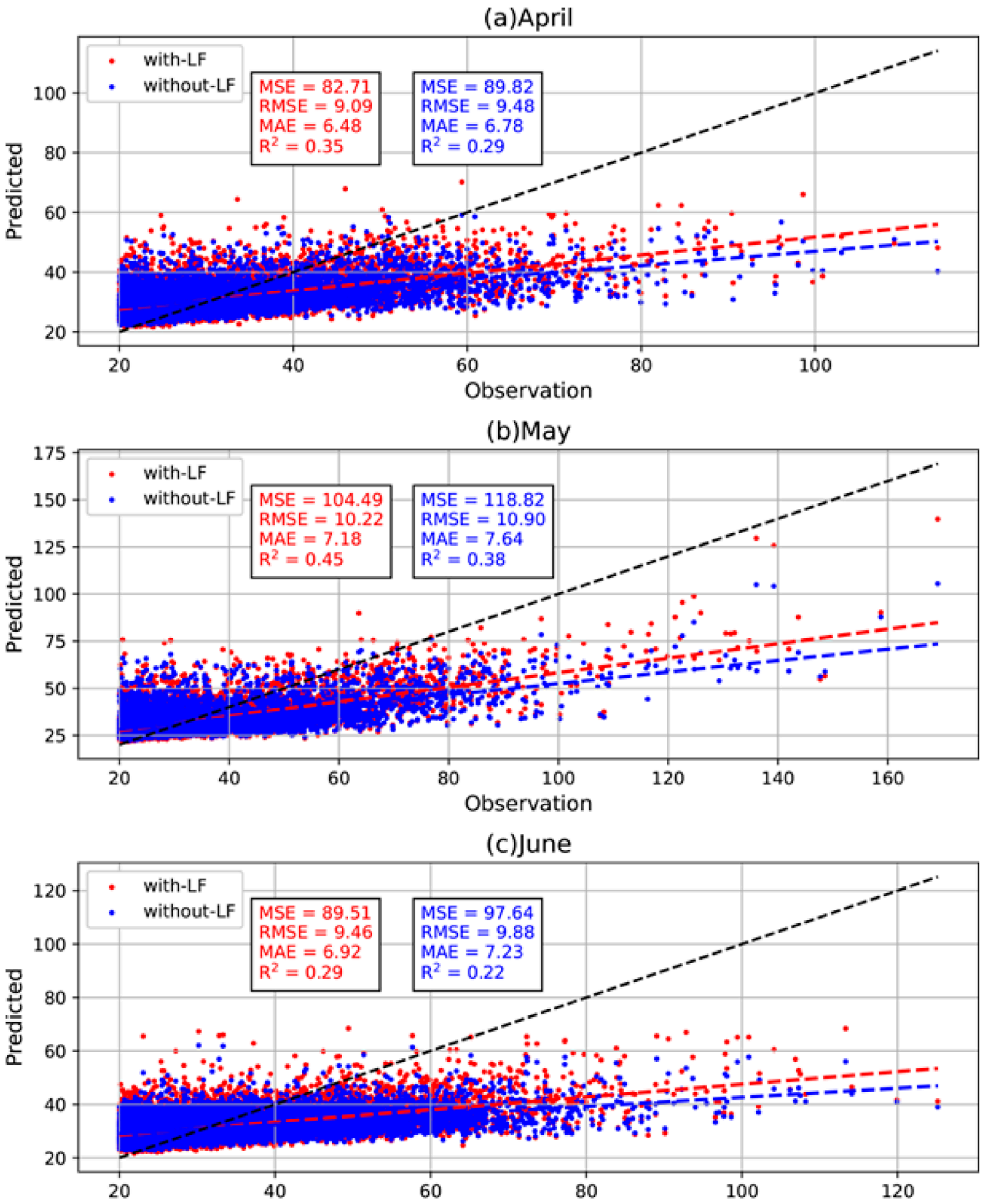

FIGURE 8
Scatter plot of predicted SDHR versus observed SDHR (units: mm h^{-1}) on the testing set in **(a)** April, **(b)** May, and **(c)** June. Red (blue) dots indicate the ML model trained with (without) lightning frequency. Dash red and blue lines represent the corresponding regression lines.

interacts strongly with longitude (the second most important variable in Figure 9b). Lower longitude regions promote higher lightning frequency, while lower longitude itself is associated with heavier rainfall (Figure 9b). Consequently, the combined interaction effect appears multiplicative in May, corresponding to the greatest error reduction observed in Figure 7. By June, lightning frequency becomes more relatively important (ranked fourth, Figure 9c), yet it no longer exhibits linear interactions with other variables (Figure 10c). In summary, predictors influence the model output not only through their individual effects but also via their interactions, which can significantly modulate their contributions.

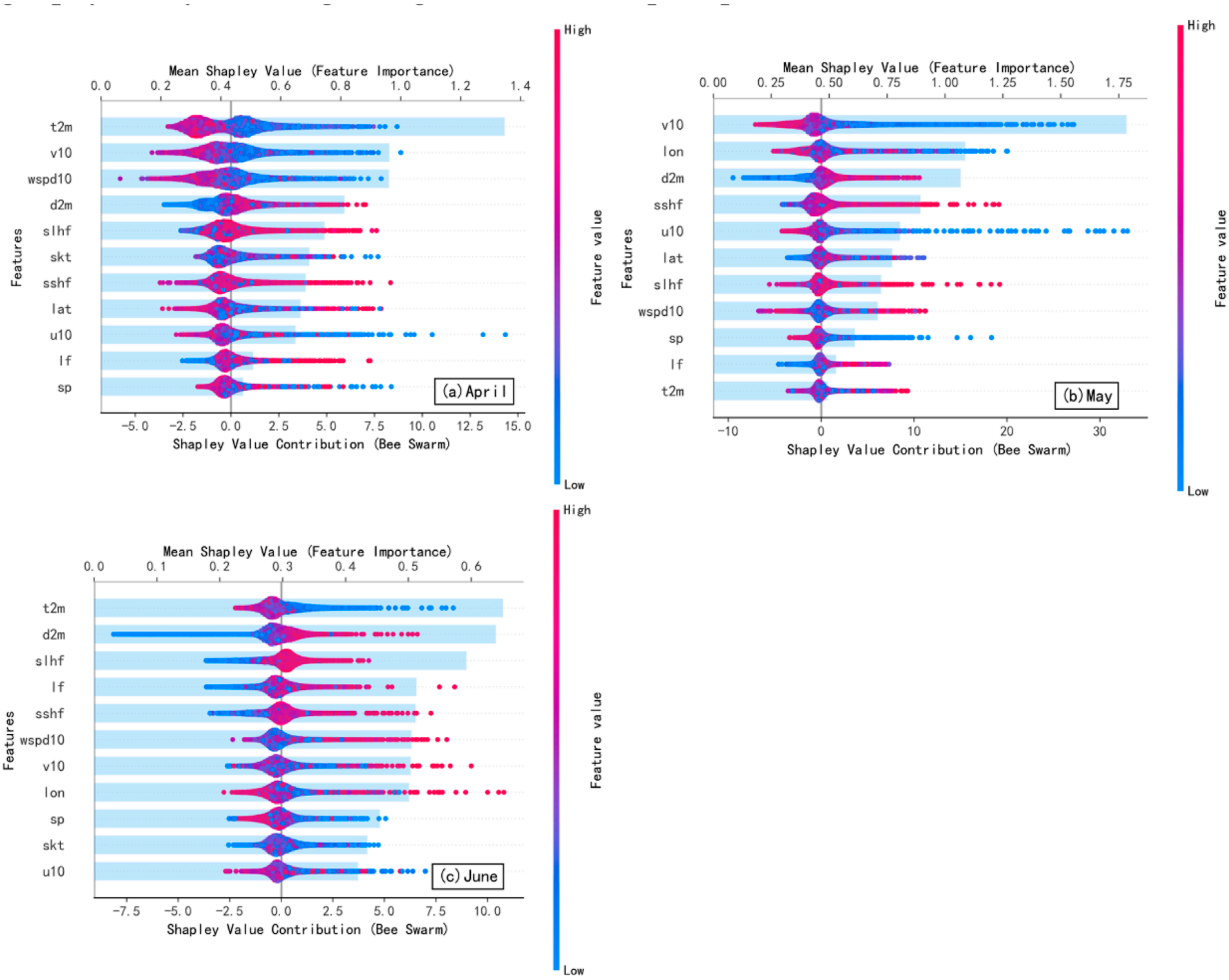

FIGURE 9
Feature importance and bee swarm plot of the top 11 features for the SDHR in **(a)** April, **(b)** May, and **(c)** June.

6 Conclusion and discussion

This study investigates the relationship between SDHR events and lightning activity over Guangxi, China, utilizing the satellite merged precipitation dataset IMERG and the FY-4A/LMI dataset during the PSRS from 2019 to 2023. Quantitative analysis further reveals that incorporating lightning frequency as a predictor in the RF model significantly reduces heavy rainfall prediction errors. The key findings are summarized below:

During the PSRS, SDHR events initially developed in eastern Guangxi in April, with a pronounced maximum center located in the northeastern part of Guangxi. The SDHR activity exhibited a westwards expansion into western Guangxi in May. In June, the number of SDHR events increased substantially and all of Guangxi was covered with widespread occurrence. The lightning activity associated with these SDHR events displayed distinct monthly variations: it peaked in April, decreased notably in May, and resurged in June. Spatially, the lightning density was strongly correlated with the SDHR distribution from April to June. From a diurnal variation perspective, both SDHR events and lightning activity exhibited pronounced unimodal distributions, with peak occurrences predominantly during the nocturnal-to-morning period. SDHR events consistently peaked at 21:00 (UTC) across all 3 months (April-June). In contrast, lightning activity reached its maximum intensity at least 1 hour earlier than SDHR events. An analysis of lightning flashes associated with individual SDHR events revealed multiple nighttime peaks in lightning frequency, which contrasted with the unimodal distribution observed in the total lightning frequency. Statistically, the mean number of lightning flashes per SDHR event was 8.58 in April, which decreased to 6.14 in May and 6.10 in June.

We trained the RF models using two distinct sets of predictors—one incorporating lightning frequency data and the other excluding it—to evaluate its impact on heavy rainfall prediction. The results demonstrate that including 1-h antecedent lightning frequency significantly reduces forecasting errors for heavy precipitation, with the MAE decreasing by 4.42%, 6.02%, and 4.29% in April, May, and June, respectively. SHAP value reveals that lightning frequency exhibits a consistent positive

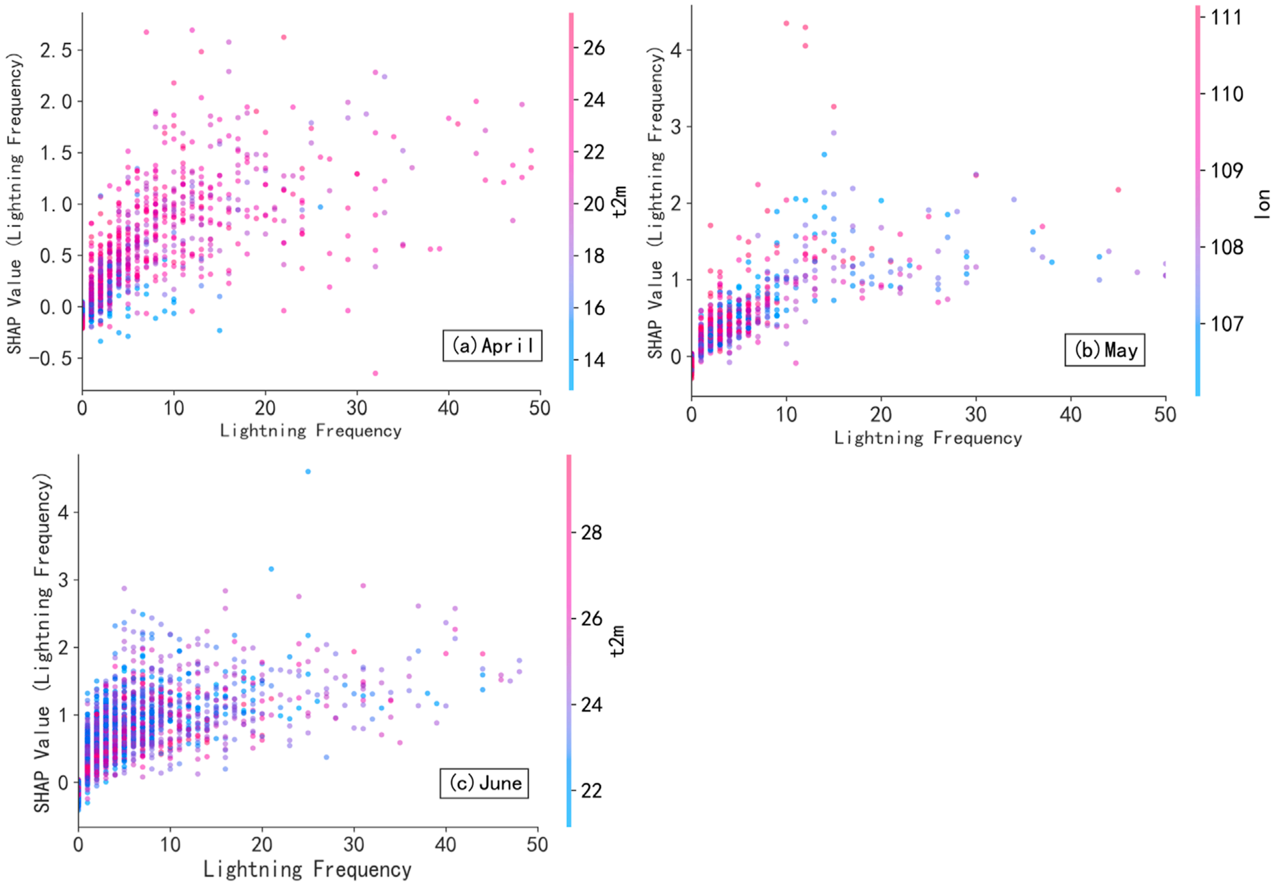

FIGURE 10
Dependence plot of lightning frequency and 2-m air temperature in April **(a)**, longitude in May **(b)**, and 2-m air temperature in June **(c)** based on SHAP value.

contribution to heavy rainfall prediction—higher lightning frequencies correspond to heavier rainfall. The SHAP analysis further demonstrates a nonlinear threshold effect of lightning frequency on SDHR prediction, where its contribution increases monotonically below ~15 flashes before saturating. Although lightning frequency ranks relatively low (10th) in individual feature importance during April-May, it plays a significant role through synergistic interactions with key atmospheric and geographic factors. Specifically, its coupling with t2m enhances convective activity in April, while its interaction with longitude reveals geographic modulation of precipitation mechanisms in May, where lower longitudes simultaneously favor both lightning occurrence and heavy rainfall. These findings highlight how variable interactions can substantially modify individual contributions, providing crucial insights into the complex mechanisms governing extreme precipitation events. The study underscores the importance of considering both threshold effects and interaction mechanisms in improving the predictability of severe convective weather systems.

Despite its advantages in many aspects, IMERG exhibits a systematic underestimation of heavy rainfall intensity. Moreover, it struggles to accurately capture small-scale heavy precipitation events, despite its relatively high spatial resolution (10 km), particularly at the urban flood scale (Xu et al., 2024). Given the increasing frequency of extreme precipitation under global warming, improving the prediction of such small-scale events demands greater attention.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://satellite.nsmc.org.cn/DataPortal/cn/home/index.html; https://cds.climate.copernicus.eu/#!/search?text=ERA5&type=dataset.

Author contributions

WH: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. XL: Conceptualization, Supervision, Writing – review and editing. WZ: Software, Writing – review and editing. BF: Visualization, Writing – review and editing. XX: Methodology, Writing – review and editing. CY: Data curation, Funding acquisition, Project administration, Resources, Supervision,

Writing – review and editing. WH: Validation, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors declare that this study received funding from Guangxi Power Grid Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication. This work was sponsored by the Science and Technology Project of Guangxi Power Grid Co., Ltd. "Research and Application of Power Grid Lightning Strike Status Perception and Hierarchical Defense Strategies: Topic 1: Perception of Lightning Strike Status and Risk Management of Transmission and Distribution Lines under Severe Thunderstorm Conditions, along with Research on Hierarchical Defense Strategies (Grant No. GXKJXM20230272)" and the Natural Science Foundation of Guangxi (Grant No. 2023GXNSFAA026511).

Conflict of interest

Authors WH, WZ, BF, XX, and CY were employed by Guangxi Power Grid Equipment Monitoring and Diagnosis Engineering Technology Research Center, Electric Power Research Institute of Guangxi Power Grid Co., Ltd.

Author WH was employed by Fangchenggang Power Bureau of Guangxi Power Grid Co., Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anco-Valdivia, J., Valencia-Félix, S., Espinoza Vigil, A. J., Anco, G., Booker, J., Juarez-Quispe, J., et al. (2025). A methodology based on random forest to estimate precipitation return periods: a comparative analysis with probability density functions in Arequipa, Peru. *Water* 17, 128. doi:10.3390/w17010128

Bahari, N., Mohammad, S. A., Esa, M. R. M., Ahmad, M. R., Ahmad, N. A., and Malek, Z. A. (2023). Analysis of lightning flash rate with the occurrence of flash floods and hailstorms in peninsular Malaysia. *J. Atmos. Solar-Terr. Phy.* 250, 106121. doi:10.1016/j.jastp.2023.106121

Brooks, H. E., and Stensrud, D. J. (2000). Climatology of heavy rain events in the United States from hourly precipitation observations. *Mon. Weather Rev.* 128, 1194–1201. doi:10.1175/1520-0493(2000)128<1194:cohrei>2.0.co;2

Cao, D. J., Lu, F., Zhang, X. H., and Yang, J. (2021). Lightning activity observed by the FengYun-4A lightning mapping imager. *Remote Sens.* 13, 3013. doi:10.3390/rs13153013

Chen, Z. X., Qie, X. S., Sun, J. Z., Sun, J. Z., Xiao, X., Zhang, Y. X., et al. (2021). Evaluation of Fengyun-4A lightning mapping imager (LMI) performance during multiple convective episodes over beijing. *Remote Sens.* 13, 1746. doi:10.3390/rs13091746

Chiappa, J., Parsons, D. B., Furtado, J. C., and Shapiro, A. (2024). Short-duration extreme rainfall events in the central and eastern United States during the summer: 2003–2023: trends and variability. *Geophy. Res. Lett.* 17, e2024GL110424. doi:10.1029/2024gl110424

Czernecki, B., Taszarek, M., Marosz, M., Półrolniczak, M., Kolendowicz, L., Wyszogrodzki, A., et al. (2019). Application of machine learning to large hail prediction-the importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5. *Atmos. Res.* 227, 249–262. doi:10.1016/j.atmosres.2019.05.010

Guan, L., Tian, F. Y., Zheng, Y. G., Cao, Y. C., and Liu, Z. M. (2024). A study on lightning activities and environmental physical characteristics in China from 2010 to 2019 (in Chinese). *Meteorol. Sci. Tech.* 52, 858–868.

Guo, Y. Y., Shen, Y., and Xu, Y. F. (2024). Co-variability between the summer cloud-to-ground lightning and precipitation over south China and their possible connections with meteorological parameters (in Chinese). *J. Trop. Meteor.* 40, 373–388.

Han, X. L., Chen, Q. X., and Fu, D. S. (2025). Trends in extreme precipitation and associated natural disasters in China, 1961–2021. *Climate* 13, 74. doi:10.3390/cli13040074

Huang, R. H., Chen, J., Wang, L., and Lin, Z. (2012). Characteristics, processes, and causes of the spatio-temporal variabilities of the East Asian monsoon system. *Adv. Atmos. Sci.* 29, 910–942. doi:10.1007/s00376-012-2015-x

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K. L., Joyce, R. J., Kidd, C., et al. (2020). "Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG)," in *Satellite precipitation measurement. Advances in global change research*. Editor V. Levizzani (Cham: Springer International Publishing), 343–353.

Hui, W., and Guo, Q. (2021). Preliminary characteristics of measurements from Fengyun-4A lightning mapping imager. *Int. J. Remote Sens.* 42, 4922–4941. doi:10.1080/01431161.2021.1906983

Hui, W., Zhang, W. J., Lyu, W. T., Zhang, Y. J., and Li, P. F. (2023). On-orbit response characteristics of Fengyun-4A lightning mapping imager (LMI) and their impacts on LMI detection. *J. Atmos. Oce. Tech.* 40, 1657–1674. doi:10.1175/jtech-d-22-0126.1

Jiang, Z. N., Zhang, D. L., Xia, R. D., and Qian, T. T. (2017). Diurnal variations of presummer rainfall over southern *China. J. Clim.* 30, 755–773. doi:10.1175/jcli-d-15-0666.1

Joaquín, M. S., Emanuel, D., Anna, A. P., Clément, A., Gabriele, A., Gianpaolo, B., et al. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* 13, 4349–4383. doi:10.5194/essd-13-4349-2021

Kochtubajda, B., Burrows, W. R., Liu, A., and Patten, J. K. (2013). Surface rainfall and cloud-to-ground lightning relationships in Canada. *Atmos. Ocean.* 51, 226–238. doi:10.1080/07055900.2013.780154

Leinonen, J., Hamann, U., Germann, U., and Mecikalski, J. R. (2022). Nowcasting thunderstorm hazards using machine learning: the impact of data sources on performance. *Nat. Hazards Earth Sci.* 22, 577–597. doi:10.5194/nhess-22-577-2022

Liao, S. S., Zhuo, J., Luo, J. Y., Ling, S. L., and Lu, B. F. (2022). Analysis of characteristics of the extreme short-time heavy rainfall in Guangxi in rain season (in Chinese). *Torr. Rain Disas.* 41, 308–314. doi:10.3969/j.issn.1004-9045.2022.03.077

Liu, S., Yang, X., and Yang, J. (2017). Research on solutions to transmission capacity limitations of transmission corridors. *Northeast Electr. Power Technol.* 38, 51–54. doi:10.3969/j.issn.1004-7913.2017.05.014

Liu, R. X., Liu, J., Pessi, A., Hui, W., Cheng, W., and Huang, F. X. (2019). Preliminary study on the influence of FY-4 lightning data assimilation on precipitation predictions. *J. Trop. Meteor.* 25, 528–541. doi:10.16555/j.1006-8775.2019.04.009

Lu, W. P., Wang, J. D., and Sun, J. W. (2022). Spatio-temporal distribution characteristics of short-time heavy precipitation in Guangxi based on automatic meteorological observation stations (in Chinese). *J. Meteor. Res. Appl.* 43, 91–97. doi:10.19849/j.cnki.CN45-1356/P.2022.4.15

Lundberg, S. M., and Lee, S. I. (2017). A unifed approach to interpreting model predictions. *Adv. Neural Inf. Proces. Syst.* 30. doi:10.48550/arXiv.1705.07874

Luo, Y. L., Xia, R. D., and Chan, J. C. L. (2020). Characteristics, physical mechanisms, and prediction of pre-summer rainfall over south China: research progress during 2008–2019. *J. Metero. Soc.* 98, 19–42. doi:10.2151/jmsj.2020-002

Matthee, R., Mecikalski, J. R., Carey, L. D., and Bitzer, P. M. (2014). Quantitative differences between lightning and nonlightning convective rainfall events as observed with polarimetric radar and MSG satellite data. *Mon. Weather Rev.* 142, 3651–3665. doi:10.1175/mwr-d-14-00047.1

Pan, Y. C. (2023). *Spatial and temporal distribution of lightning over Guangxi based on observations from lightning mapping imager onboard fengyun satellite*. Nanning, China: Nanning Normal University.

Peng, D. D., Zhou, T. J., Sun, Y., and Lin, A. L. (2022). Interannual variation in moisture sources for the first rainy season in south China estimated by the FLEXPART model. *J. Clim.* 35, 745–761. doi:10.1175/jcli-d-21-0289.1

Pradhan, R. K., Markonis, Y., Vargas Godoy, M. R., Villalba-Pradas, A., Andreadis, K. M., Nikolopoulos, E. I., et al. (2022). Review of GPM IMERG performance: a global perspective. *Remote Sens. Environ.* 268, 112754. doi:10.1016/j.rse.2021.112754

Price, C., and Federmesser, B. (2006). Lightning-rainfall relationships in mediterranean winter thunderstorms. *Geophy. Res. Lett.* 33, 2005GL024794–16. doi:10.1029/2005gl024794

Ranalkar, M., and Chaudhari, H. (2009). Seasonal variation of lightning activity over the Indian subcontinent. *Meteor. Atmos. Phys.* 104, 125–134. doi:10.1007/s00703-009-0026-7

Rentschler, J., Salhab, M., and Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. *Nat. Commun.* 13, 3527. doi:10.1038/s41467-022-30727-4

Soula, S., and Chauzy, S. (2000). Some aspects of the correlation between lightning and rain activities in thunderstorms. *Atmos. Res.* 56, 355–373. doi:10.1016/s0169-8095(00)00086-7

Tang, G. Q., Clark, M. P., Papalexiou, S. M., Ma, Z. Q., and Hong, Y. (2020). Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets. *Remote Sens. Environ.* 240, 111697. doi:10.1016/j.rse.2020.111697

Torcasio, R. C., Federico, S., Prat, A. C., Panegrossi, G., Adderio, L. P., and Dietrich, S. (2021). Impact of lightning data assimilation on the short-term precipitation forecast over the central Mediterranean sea. *Remote Sens.* 13, 682. doi:10.3390/rs13040682

Wang, X. Y., Jiang, W. G., Wu, J. J., Hou, P., Dai, Z. J., Rao, P. Z., et al. (2023). Extreme hourly precipitation characteristics of mainland China from 1980 to 2019. *Int. J. Clim.* 43, 2989–3004. doi:10.1002/joc.8012

Wolfensberger, D., Gabella, M., Boscacci, M., Germann, U., and Berne, A. (2021). RainForest: a random forest algorithm for quantitative precipitation estimation over Switzerland. *Atmos. Meas. Tech.* 14, 3169–3193. doi:10.5194/amt-14-3169-2021

Wu, F., Cui, X. P., Zhang, D. L., and Qiao, L. (2017). The relationship of lightning activity and short-duration rainfall events during warm seasons over the beijing metropolitan region. *Atmos. Res.* 195, 31–43. doi:10.1016/j.atmosres.2017.04.032

Wu, A. K., Cai, L. J., Guo, J. C., and Ding, M. (2024). Comparative study on lightning data between the Fengyun-4A satellite and the national lightning monitoring network in Guizhou, China. *J. Atmos. Solar-Terres. Phy.* 256, 106194. doi:10.1016/j.jastp.2024.106194

Xiao, Z. X., Duan, A. M., Chen, Y. H., Wei, C., and Luo, X. L. (2025). Different mechanisms of Eurasian snow cover on precipitation during the early and late pre-rainy season in south China. *Sci. China Earth Sci.* 68, 882–897. doi:10.1007/s11430-024-1490-2

Xu, G. Q., Huang, S. Y., and Zhao, C. Y. (2020). Influence of FY-4A lightning data on numerical forecast of convective weather (in Chinese). *Meteor. Mon.* 46, 1165–1177. doi:10.7519/j.issn.1000-0526.2020.09.004

Xu, J. Y., Qi, Y. C., Li, D. H., and Zhao, Z. F. (2024). Can IMERG QPE product capture the heavy rain on urban flood scale? *Sci. Total Environ.* 933, 173022. doi:10.1016/j.scitotenv.2024.173022

Yuan, W., Sun, W., Chen, H., and Yu, R. (2014). Topographic effects on spatiotemporal variations of short-duration rainfall events in warm season of central north China. *J. Geophys. Res. Atmos.* 119 (11), 234. doi:10.1002/2014jd022073

Zhang, W. J., Meng, Q., Ma, M., and Zhang, Y. J. (2011). Lightning casualties and damages in China from 1997 to 2009. *Nat. Hazards* 57, 465–476. doi:10.1007/s11069-010-9628-0

Zhang, Y., Zheng, X., Li, X. F., Lyu, J. X., and Zhao, L. L. (2023). Evaluation of the GPM-IMERG V06 final run products for monthly/annual precipitation under the complex climatic and topographic conditions of China. *J. App. Meteor. Climatol.* 62, 929–946. doi:10.1175/jamc-d-22-0110.1

Zhou, K. H., Sun, J. S., Zheng, Y. G., and Zhang, Y. T. (2022). Quantitative precipitation forecast experiment based on basic NWP variables using deep learning. *Adv. Atmos. Sci.* 39, 1472–1486. doi:10.1007/s00376-021-1207-7

Zhou, Q. L., Cui, X. P., and Hao, S. F. (2022). The statistical relationship of lightning activity and short-duration rainfall events over guangzhou, China, in 2017. *Weather Forecast* 37, 601–615. doi:10.1175/waf-d-21-0161.1