npj | climate and atmospheric science

www.nature.com/npjclimatsci

ARTICLE OPEN

Check for updates

Lightning nowcasting with aerosol-informed machine learning and satellite-enriched dataset

Ge Song[1], Siwei Li[1,2,3]✉ and Jia Xing[4]

Accurate and timely prediction of lightning occurrences plays a crucial role in safeguarding human well-being and the global environment. Machine-learning-based models have been previously employed for nowcasting lightning occurrence, offering advantages in computation efficiency. However, these models have been hindered by limited accuracy due to inadequate representation of the intricate mechanisms driving lightning and a restricted training dataset. To address these limitations, we present a machine learning approach that integrates aerosol features to more effectively capture lightning mechanisms, complemented by enriched satellite observations from the Geostationary Lightning Mapper (GLM). Through training a well-optimized LightGBM model, we successfully generate spatially continuous (0.25° by 0.25°) and hourly lightning nowcasts over the Contiguous United States (CONUS) during the summer season, surpassing the performance of competitive baselines. Model performance is evaluated using various metrics, including accuracy (94.3%), probability of detection (POD, 75.0%), false alarm ratio (FAR, 38.1%), area under curve of precision–recall curve (PRC-AUC, 0.727). In addition to the enriched dataset, the improved performance can be attributed to the inclusion of aerosol features, which has significantly enhanced the model. This crucial aspect has been overlooked in previous studies. Moreover, our model unravels the influence of aerosol composition and loading on lightning formation, indicating that high aerosol loading consisting of sulfates and organic compounds tends to enhance lightning activity, while black carbon inhibits it. These findings align with current scientific knowledge and demonstrate the immense potential for elucidating the complex mechanisms underlying aerosol-associated lightning phenomena.

npj Climate and Atmospheric Science (2023)6:126 ; https://doi.org/10.1038/s41612-023-00451-x

INTRODUCTION

Lightning, a prominent cause of natural human fatalities, poses a significant threat to modern society, resulting in over 4000 deaths globally each year[1,2]. Additionally, it leads to significant economic losses, with the United States alone experiencing around 1 billion US dollars in damages annually. Timely and accurate prediction of lightning occurrences plays a vital role in facilitating emergency preparedness and protective measures. Moreover, lightning serves as a primary natural source of nitrogen oxides, thereby exerting considerable influence on atmospheric chemistry[3], underscoring the criticality of lightning prediction in safeguarding human well being and the global environment.

Lightning commonly occurs during the formation of thunderstorms, which are typically characterized by high moisture levels and an unstable atmosphere[4–8]. Numerical models can explicitly simulate lightning formation by incorporating parameterized microphysics processes[9,10]. However, current numerical models struggle to strike a balance between high lightning detectability and low false alarm rates (FAR), thereby limiting their applicability in lightning forecasting[11–13]. Additionally, the computational demands of lightning simulation within numerical models impede the efficiency of lightning nowcasting, where timeliness is crucial in domains such as aviation and manufacturing. In contrast, observation-based data-driven lightning models have emerged as efficient methods for achieving accurate lightning nowcasts, leveraging ground-truth samples at a lower computational cost. For example, Mostajabi et al.[14] were pioneers in exploring data-driven models for lightning nowcasting in the future hour with remarkable accuracy by solely utilizing weather variables. Furthermore, the inherent capacity of machine learning models to capture nonlinear characteristics enables high performance even with simple and practical feature inputs. So far, a series of machine learning models have been explored to predict the occurrence of lightning with meteorological variables either from weather station, or assimilated meteorological model and weather radar, including artificial neural network and decision tree[15], light gradient-boosting machine (LightGBM)[16], support vector machines and random forest[17] and long short-term memory recurrent neural network[18]. Current machine learning models demonstrate high efficiency; however, they still encounter challenges with high FAR at high probability of detection (POD) levels[17]. This limitation may be attributed to insufficient training datasets and incomplete feature data utilized in previous models, which will be thoroughly elucidated in subsequent sections.

First, previous studies primarily relied on ground-based lightning detection networks and sensors onboard polar orbit satellites, which exhibit significant limitations in terms of detection efficiency and spatial coverage per overpass, thereby constraining the accuracy of observation-based models for lightning prediction[19–22]. Along with the development of the geostationary satellites, real-time monitoring lightning occurrence across the space becomes available. Particularly, the sensor Geostationary Lightning Mapper (GLM) onboard the Geostationary Operational Environmental Satellites (GOES) can capture the detailed characteristics of lightning occurrence at full spatiotemporal coverage to support analysis including diagnosis of the current numerical models[23,24], investigation on the association of natural events in the climate system[25–27] and risk prevention[28,29]. Such high

[1]Hubei Key Laboratory of Quantitative Remote Sensing of Land and Atmosphere, School of Remote Sensing and Information Engineering, Wuhan University, 430000 Hubei, China. [2]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. [3]Hubei Luojia Laboratory, Wuhan University, 430000 Hubei, China. [4]Department of Civil and Environmental Engineering, the University of Tennessee, Knoxville TN 37996, USA. ✉email: siwei.li@whu.edu.cn

Table 1. Parameters and evaluation metrics of model prediction skill.

Parameters acronym	Full name	Definition	Expression
TP	True Positive	Number of correctly predicted lightning-active samples	-
FP	False Positive	Number of lightning-inactive samples wrongly predicted as lightning-active	-
FN	False Negative	Number of lightning-active samples wrongly predicted as lightning-inactive	-
TN	True Negative	Number of correctly predicted lightning-inactive samples	-
POD	Probability of Detection	Proportion of correctly detected lightning-active samples	$\frac{TP}{TP+FN}$
FAR	False Alarm Ratio	Proportion of samples predicted as lightning-active but no lightning is observed	$\frac{FP}{TP+FP}$
CSI	Critical Success Index	Ratio of correct prediction of lightning-active samples to all predictions that need or are made as lightning-active	$\frac{TP}{TP+FP+FN}$
HSS	Heidke Score Skill	A measure to evaluate the fractional improvement of the forecast over the forecast merely due to chance	$\frac{2\times(TP\times TN-FP\times FN)}{(TP+FN)\times(FN+TN)+(TP+FP)\times(FP+TN)}$
PRC-AUC	Area Under Curve of Precision–Recall Curve	Summary of model binary responses regarding to different thresholds	-
Accuracy	-	Proportion of correctly prediction samples	$\frac{TP+TN}{TP+FP+FN+TN}$

detection efficiency of cloud-to-ground (CG) and intracloud (IC) lightning derived from GLM has great potentials to provide reliable lightning occurrence record[30] for the observation-based model for lightning prediction.

Second, a crucial limitation of current machine learning models for lightning prediction is their exclusive reliance on meteorological information, neglecting the significant influence of aerosols on lightning patterns. However, observational studies have demonstrated the substantial impact of aerosols on lightning formation[31,32]. On the one hand, aerosols can stimulate convection, promoting particle collision and enhancing charge dissipation. On the other hand, aerosols possess notable radiative properties that suppress particle activation[33–36]. Furthermore, distinct aerosol components exhibit diverse pathways of influence on lightning discharges[37]. Therefore, incorporating detailed aerosol information becomes essential to enhance the performance of lightning prediction models. Satellite-based measurements provide real-time monitoring of aerosol chemical components at a comprehensive spatiotemporal scale. Additionally, studies have indicated a close relationship between near-surface aerosols and PM2.5, enabling timely monitoring of near-surface aerosol distribution[38,39]. By leveraging well-designed machine learning models that incorporate aerosol information and utilizing satellite-enriched datasets, significant improvements in lightning prediction performance are anticipated.

In this study, we aimed to enhance the existing lightning nowcasting model by incorporating aerosol information, specifically the aerosol optical depth and composition, in addition to conventional meteorological variables and ground-based networks. Furthermore, we utilized observations from geostationary satellite as the primary data source and the label for the lightning nowcasting model, considering their stability and data availability. The evaluation results were assessed using common metrics of nowcasting and forecasting research. Our findings demonstrate the effectiveness of aerosol-informed machine learning in predicting lightning occurrences within the next hour, while also providing valuable insights into the role of aerosols in lightning formation.

This paper is organized as follows. In the following section, we present the results and analysis of the lightning nowcasting model we proposed, utilizing aerosol information and geostationary satellite observations. Subsequently, we proceed with a discussion of the findings and draw our conclusions.

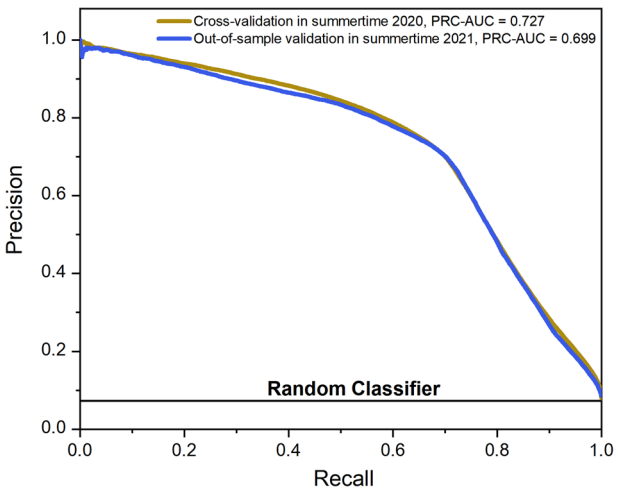

Fig. 1 Evaluation of the lightning prediction model presented by precision–recall curve. The model is evaluated under two schemes: 10-fold cross-validation and out-of-sample validation with testing set from summertime 2021.

RESULTS

Performance and transferability of prediction model

To train and validate the lightning prediction model, we collected a lightning database comprising observations from Geostationary Lightning Mapper (GLM) onboard geostationary satellite (GOES-16). The data was labeled based on the presence or absence of lightning activity. In addition, meteorological and aerosol data were obtained from forecast products provided by the Copernicus Atmosphere Monitoring Service (CAMS) and used as the input features for the model. Various validation schemes and evaluation metrics were employed to assess the performance of the model. Detailed information regarding to the validation methods can be found in the Methods section, while Table 1 presents the evaluation parameters and metrics used. Figure 1 illustrates the performance of the proposed lightning prediction model, which was trained and cross-validated during the summer of 2020. The precision–recall curve in the figure depicts the tradeoff between precision and recall at difference thresholds (labeled as "Cross-validation in summertime 2020" in Fig. 1). The model exhibited promising lightning nowcasting ability, as evidenced by a PRC-

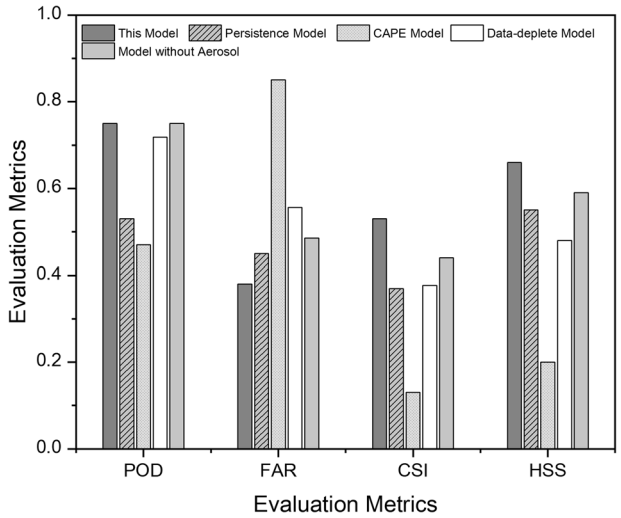

Fig. 2 Evaluation of the applicability of lightning occurrence in CONUS by comparing this model with other baseline models. The baseline models include the Persistence model, CAPE model, data-deplete model which is trained with lightning mapping array (LMA) and model without aerosol as input. The evaluation metrics include POD, FAR, CSI and HSS.

AUC of 0.727 for the LightGBM model. Notably, the shape of the precision–recall curve indicates that the model can maintain a low proportion of false alarm predictions when higher precision is desired. Specifically, at a threshold where model achieves a POD of 75%, a common level in existing models, the FAR was determined to be 38%. Comparing the precision–recall curve of this LightGBM model with that of a random classifier highlights the model's ability to effectively distinguish between lightning and non-lightning cases.

We further conducted an evaluation of the proposed model regarding its transferability, which refers to the ability of the trained model to be applied to a different temporal range of interest. In this validation scheme, the model was trained using datasets from summertime 2020 and subsequently tested on data from summertime 2021. Figure 1 illustrates the performance of the model when applied to summertime 2021 (labeled as "Out-of-sample validation in summertime 2021" in Fig. 1). Notably, little contrast is observed compared to the performance in summertime 2020. The transferred model exhibits a slightly reduced PRC-AUC of 0.699 compared to its application in 2020. This indicates excellent model transferability, implying the model's potential to be incorporated into parameterization models and used for interpreting lightning occurrence during numerical simulation in the future.

To evaluate the effectiveness of the proposed lightning prediction model compared to commonly used baseline models, namely the Persistence model and the CAPE model (as described in detail in Supplementary Method 1), we conducted a comprehensive model intercomparison. Evaluation of these models is based on established metrics used in previous studies to compare lightning occurrence models, namely POD, FAR, Critical Success Index (CSI), and Heidke Skill Score (HSS). The results are presented in Fig. 2, where the proposed model achieves the highest POD (0.75 for this model, 0.53 for the Persistence model, and 0.47 for the CAPE model), CSI (0.53 for this model, 0.37 for the Persistence model, and 0.13 for the CAPE model), and HSS (0.66 for this model, 0.55 for the Persistence model, and 0.20 for the CAPE model). Additionally, the proposed model exhibits the lowest FAR (0.38 for this model, 0.45 for the Persistence model, and 0.85 for the CAPE model), indicating its

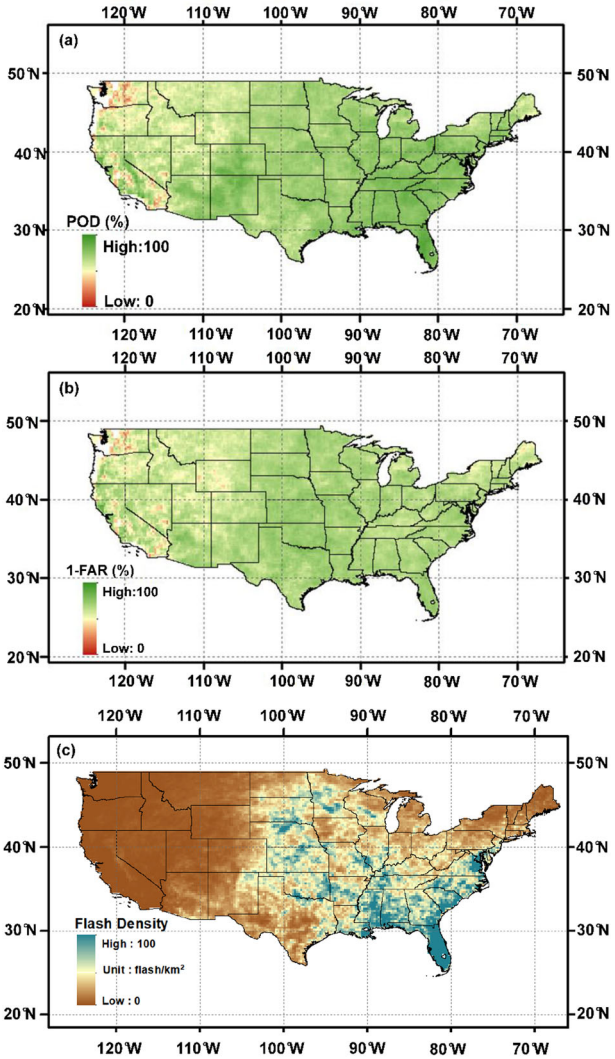

Fig. 3 Distribution of model performance in CONUS, and distribution of lightning densities during 2020 summertime. **a** Spatial distribution of recall (POD) in CONUS; **b** spatial distribution of precision (1-FAR) in CONUS; and **c** the spatial distribution of lightning density during 2020 summertime.

superior ability to accurately capture lightning occurrence and outperform the baseline prediction approaches.

In order to emphasize the spatial effectiveness of the lightning prediction model across the CONUS, the spatial distribution of two key metrics, POD and FAR, are presented using validation datasets in Fig. 3a, b. The results demonstrate that both metrics exhibit higher values in the southeastern CONUS, which aligns with regions characterized by elevated lightning density. Furthermore, it is observed that the spatial distribution of model performance correlates with the distribution of lightning density, as depicted in Fig. 3c. Specifically, regions with sparse lightning occurrence exhibit lower POD values, reaching approximately 30% in areas where flash densities fall below 0.05 flashes per square kilometer. This phenomenon can be attributed to the imbalanced dataset used in the machine learning process, where samples with infrequent lightning occurrences may contribute to the lower performance in these regions.

Improvement from enriched dataset from GLM

Previous studies have indicated that machine-learning-based lightning prediction models, which utilize data from radar and

ground-based lightning networks, demonstrate moderate now-casting skills. In this study, we present an enhancement to the model's accuracy by incorporating data-enriched observations from GLM onboard the geostationary satellite GOES-16, which provides full spatial coverage. We compare the proposed model and the lightning prediction model simulated on the basis of the publicly accessible lightning mapping array (LMA) observations (labeled as "data-deplete" model in the following), which has been widely investigated to study the lightning patterns[40,41]. In comparison to the proposed model (data-enriched model), the data-deplete model utilizes observations from LMA to predict the lightning. Details of the LMA and their corresponding center locations are explained in Supplementary Method 2 and presented in Supplementary Table 1. Figure 2 illustrates the superior performance of the data-enriched model compared to the data-deplete model across all evaluation metrics. Although the POD of data-enriched model (75%) is only slightly higher than data-deplete model (72%), the FAR of data-deplete model (56%) is significantly higher than that of the data-enriched model (36%). Additionally, the CSI and HSS for the data-deplete model (38% and 48% respectively) indicate inherent deficiencies in the model without the enrichment provided by geostationary satellite observations.

Such improvement in model performance is also attributed to the detection stability offered by the geostationary satellite observations. In comparison to space-borne observations, the ground-based detection network exhibits a decreasing detection efficiency as the distance from the network center increases. We demonstrate that the data-deplete model experiences a decline in accuracy with increasing distance from the center of the network, as depicted in Supplementary Figure 1. Specifically, the model's CSI decreases by 0.78% and the FAR increases by 1% per 50 km away from the network center. The slopes of model metrics regarding to the distance have been determined to be statistically significant on a one-tailed t-test with p-value of less than 0.05. The performance of the model based on LMA becomes increasingly limited in regions where no local LMA equipment is available due to its dependency on distance. In contrast, the enhanced stability provided by observations from geostationary satellites protects the prediction model from reduced robustness and expands its applicability to a broader range of regions.

Aerosol enhances the predictability of lightning occurrence

Numerous studies have documented the impact of aerosols on long-range lightning occurrence, with different aerosol components exhibiting distinct effects, either suppressing or enhancing lightning activity[37,42,43]. This study analyzes the diurnal variability of lightning and aerosols, aiming to uncover the temporal patterns of lightning occurrence and aerosol behavior. According to Supplementary Fig. 2, the distribution of lightning occurrence exhibits a pronounced preference for the afternoon and evening hours, with limited observations during the morning hours. This distinct temporal pattern indicates that predictors with temporal characteristics possess the potential to forecast lightning occurrences. The diurnal variation of aerosol optical depth (AOD), as depicted in Supplementary Figure 2b, aligns with the pattern observed for lightning occurrence. This consistency suggests that aerosol information can serve as a reliable temporal predictor for lightning events. We further fitted the anomalies of diurnal variability with mean lightning density in the hours when the mean lightning density exceeds 0.001 flash/km^2, a high correlation (Pearson's $r = 0.897$) is observed as in Supplementary Fig. 2d, while a lower correlation for temperature (Pearson's $r = 0.772$) as in Supplementary Figure 2e. To further explore the indication effect of multiple factors, a Time lagged cross correlation (TLCC) analysis is used to reveal the time-series indication effect of these factors[44]. As shown in Supplementary Fig. 3, AOD exhibits

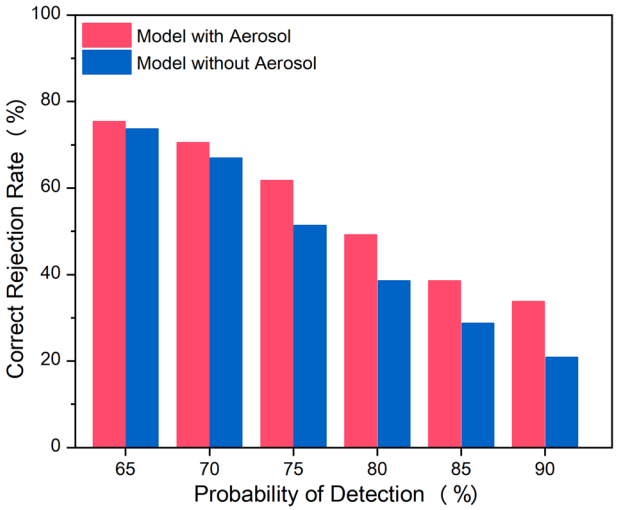

Fig. 4 Differences of model performances at difference POD levels. The two models (with and without aerosol information as model input) are compared in terms of the correct rejection rate (1-FAR) given different conditions of POD.

outstanding synchronicity with the lightning occurrence with no offset and maintains high correlation at offset of −1 h, indicating the trend of AOD can both well mark the trend of lightning occurrence at the current moment and predict the lightning occurrence in the following hour, while meteorological variables including relative humidity (offset of +4 h) and temperature (offset of −2 h) show inferior indication of lightning occurrence. Thus, aerosol observations show great potential to indicate the occurrence of lightning in terms of temporal variation.

Consequently, enhanced lightning prediction performance can be anticipated through the utilization of better-designed machine-learning models incorporating aerosol information. The performance of models with and without aerosol information is compared in application scenarios where a high POD is required, as shown in Fig. 4. When the POD threshold is below 70%, the contribution of aerosol information to the predictability of lightning is minimal. This phenomenon can be attributed to the information contributed by meteorological data and historical lightning records. These additional sources of information aid in capturing the spatial and temporal patterns of lightning occurrences, thereby reducing the influence of aerosol. However, at higher levels of POD, the significant contribution of aerosol information becomes more prominent, dominating the prediction performance. When the POD threshold is set above 75%, the difference in correct rejection rates between models with and without aerosol information exceeds 10%. This indicates that aerosol data provides valuable information for lightning prediction, as evidenced by the improved correct rejection rates as the demand for POD increases. However, achieving precise predictions of lightning occurrence with a high POD above 80% remains challenging without a thorough understanding of the intricate mechanisms involving aerosols and other factors that quantify the complete process of lightning formation. While the results suggest that aerosol information can enhance the predictability of lightning, there is still room for further improvement in the model's performance. This can be achieved by incorporating additional quantifiable features related to the formation of lightning.

The statistical distribution of AOD is depicted in Fig. 5a, revealing that in the majority of cases, AOD values are below 0.2, accounting for approximately 80% of the total cases. The analysis then focused on the relationship between model performance and aerosol loading, as illustrated in Fig. 5b. As AOD increases up

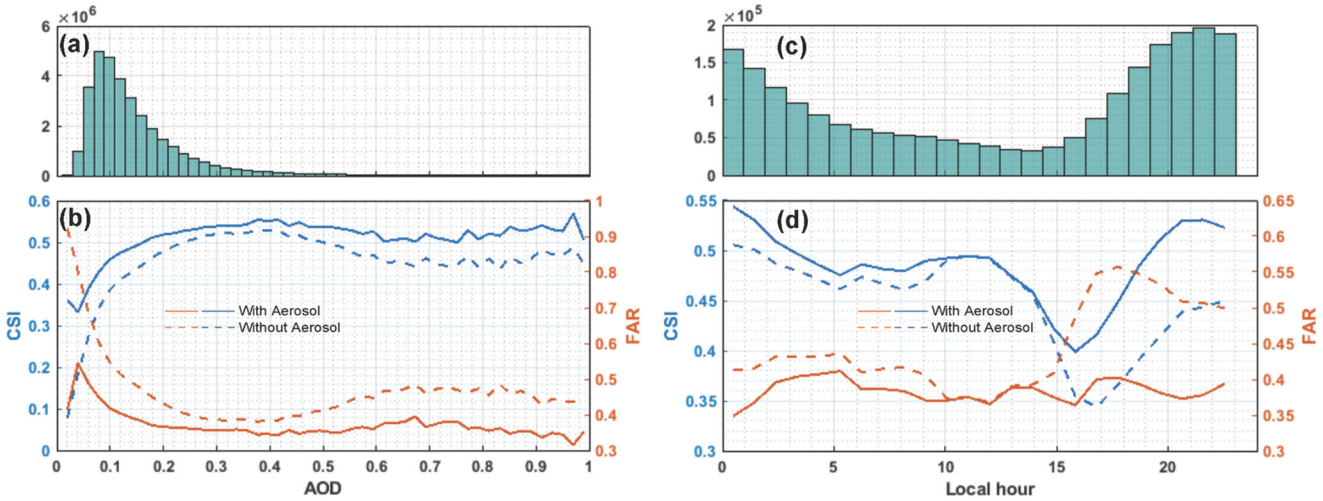

Fig. 5 The model performance regarding AOD and local hours for model with and without aerosol information. a Histogram of data samples in terms of AOD; **b** model performance in terms of CSI at different AOD levels for the models with and without aerosol information as input; **c** histogram of data samples in terms of the local hour; **d** diurnal distribution of the model performance in terms of CSI for the models with and without aerosol information as input.

to 0.2, both models demonstrate an improvement in terms of CSI and FAR. In comparison to the model without AOD, the proposed model exhibits better robustness, with the FAR ranging from 30% to 50%, while the other model struggles to predict lightning occurrences with a FAR exceeding 50% for AOD values below 0.1, which represents around 40% of the total cases. The difference in CSI between the models remains consistently significant in low AOD situations. However, as AOD increases to 0.4, the contrast between the two models gradually diminished. As AOD continues to rise, potentially indicating an air pollution event, the discrepancy in model performances further expands. At this stage, the inclusion of aerosol information can reduce approximately 40% of false warning reports.

Aerosol observations have a significant impact the temporal variability of model performance. As depicted in Fig. 5d, the lightning prediction model with AOD demonstrates stable performance throughout the day. During local hours before 2 pm, both models exhibit only minor differences. However, after 2 pm, aerosol features play a crucial role in improving the model performance. The inclusion of aerosol information results in a reduction of FAR by 0.10–0.15 after 2 pm, indicating that 25% of false early warnings are avoided by considering aerosol features. Similarly, for CSI, the aerosol information contributes to an elevation of CSI with by 0.05–0.10. It is important to note that the time period during which the model's performance is enhanced by aerosol information (3–11 pm) does not entirely overlap with the time range of relatively high lightning occurrence (6 pm–2 am). This suggests that the aerosol information does not directly indicate the immediate occurrence of lightning but rather provides insight into the trend of future lightning occurrence. This observation aligns well with the finding presented in Supplementary Fig. 3.

The contribution of aerosol observations exhibits varying patterns across different regions. Supplementary Figure 4 illustrates the model's enhancement through aerosol observations in CONUS. In most regions, aerosols demonstrate a positive impact on the POD of the lightning prediction model, particularly in the southeastern and Midwestern regions of CONUS. The significance of aerosols becomes evident in the southeastern CONUS, which experiences the highest flash densities. The results indicate that the occurrence of lightning becomes more detectable with the aid of aerosol observations, resulting in a remarkable enhancement of approximately 10%. Regarding the reduction in FAR due to aerosols, the distribution of model enhancement follows similar patterns to those observed in terms of the POD and CSI, with the southeastern and Midwestern CONUS regions benefiting the most from aerosol observations. However, there are still certain areas where the incorporation of aerosol features could potentially impair the model's performance, especially in the west coast regions. This could be attributed to different aerosol effect regimes, particularly the spatial distribution of aerosols (e.g., black carbon from wildfires[45]) on the west coast.

Contribution of aerosol components by model interpretation

The Shapley Additive ExPlanation Approach (SHAP) method serves as a valuable tool for interpreting machine learning models and analyzing their features. Figure 6 presents a feature importance analysis, revealing the top 10 features that contribute the most to lightning prediction. The complete names corresponding to the feature acronyms can be found in Supplementary Table 2. Among these variables, flash density emerges as the strongest predicator of lightning, while the importance of aerosol components varies. Sulfate stands out as the most influential predictor for lightning occurrence, followed by sea salt, black carbon and organic compounds, which display moderate importance in the prediction. The contribution of aerosol composition and optical depths underscores their high relevance in lightning prediction. Weather variables, traditionally used as predictors of lightning, exhibit moderate ability to nowcast lightning. For instance, relative humidity demonstrates the highest predictability among weather variables, aligning with previous knowledge of lightning formation mechanisms and further supporting the notion that lightning occurrence favors high moisture levels[4–8]. The SHAP analysis proves to be a valuable tool in identifying whether aerosol composition positively or negatively affects lightning occurrence. Increasing levels of aerosols components such as sulfate, organic compound and sea salt correspond to higher SHAP values, indicating their interpretation as enhancing factors for lightning according to the machine learning model. Conversely, black carbon exhibits a negative effect on lightning occurrence. Such results are consistent with previous research on the impact of aerosol components on lightning[46,47]. A more detailed comparison and analysis with the existing knowledge base is provided in the Discussion section. Consequently, optimizing the prediction model necessitates a combination of aerosol characterization and weather variables.

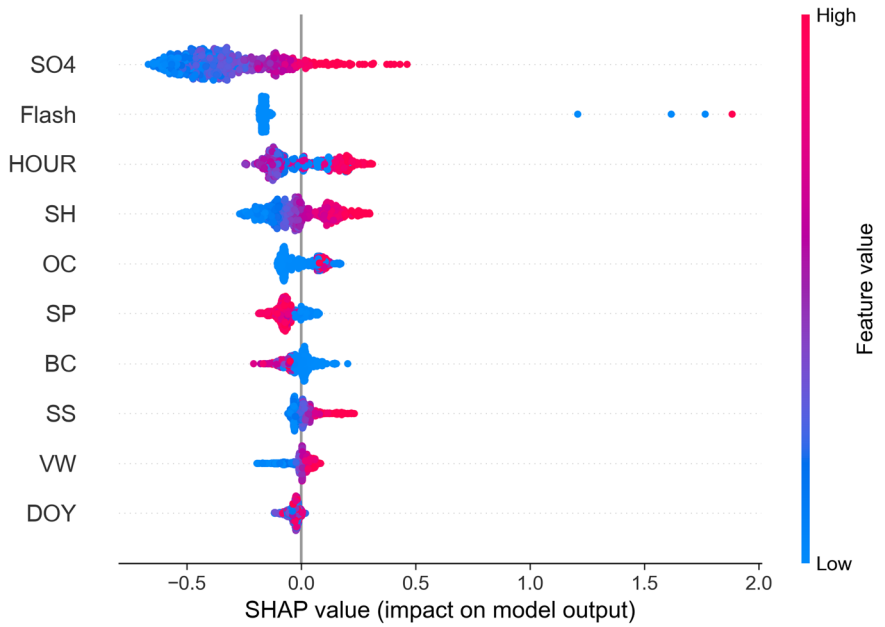

Fig. 6 Feature importance demonstrated by SHAP value of the machine learning model. The feature importance is ranked by the mean absolute SHAP value of the features.

DISCUSSION

In this study, we propose a highly accurate observation-based data-driven model for lightning occurrence prediction, utilizing the LightGBM gradient boosting framework. Our model integrates aerosol observations and meteorological variables, making it one of the most precise lightning prediction models currently available (accuracy = 94.3%, POD = 75%, FAR = 38%, AUC = 0.727). By incorporating previous observational and modeling studies examining the relationship between aerosols and lightning, we demonstrate the significant impact of aerosols on lightning prediction. Aerosols primarily influence lightning through their microphysical and radiative properties. Previous studies have explored various models for lightning occurrence nowcasting, including numerical simulations and machine learning approaches. However, these approaches have not yet achieved the desired level of model performance. Numerous studies have identified increased lightning flash densities in metropolitan areas or downwind regions, partially attributable to the microphysical effects of aerosols[48,49]. Regional case studies, particularly in northern and central Africa, have also reported the influence of different aerosol types on lightning[35]. These studies highlight that the dominant effects of aerosols, whether microphysical or radiative, depends on the aerosol type, ultimately affecting lightning rates through aerosol loading. However, in the CONUS the variability in aerosol loading and composition is much higher than in previous studies, posting challenges in statistically disentangling the effects of aerosol types and loading alone. Interpretable machine learning, capable of capturing complex non-linear relationships within the model, offers a suitable tool for analyzing the contribution of aerosol compositions. In this study, leveraging enriched observations from the GLM and employing an interpretable machine learning model, feature analysis provides insights into the influence of aerosol compositions on lightning occurrence. The analysis consistently reveals a negative impact of black carbon species in aerosols on lightning frequencies, aligning with theoretical studies emphasizing the heating effect of black carbon, leading to convection changes and inhibiting lightning occurrence[35,50,51] . Furthermore, the analysis underscores the significant feature importance of sulfate aerosols, which is in agreement with previous reports[46]. As indicated by Jin et al.[52], sulfate aerosols promote ice-phase microphysical processes, intensifying lightning activity. Regardless of aerosol composition, a negative contribution to lightning occurrence is observed when AOD exceeds a certain threshold, supporting observations by Shi et al.[53]. Our proposed model also identifies sea salt aerosols as stimulants for lightning occurrence, although recent reports suggest that the behavior of sea salt aerosols varies depending on particle modes[37]. This discrepancy may be attributed to the relatively lower sea salt aerosol loading in the CONUS compared to maritime conditions, combined with the predominance of fine-mode aerosols on land. Overall, the feature analysis demonstrates strong agreement with theoretical and modeling studies of lightning occurrence, offering a potential approach for parameterizing lightning occurrence in numerical models in the future.

The Result section has analyzed the applicability of the proposed model based on the meteorological conditions and aerosol information. The model demonstrates strong performance in regions with high lightning densities and moderate and high aerosol loading. Given that the regions with high demand for lightning protection largely coincide with the regions where the model performs well, it can effectively applied in areas where a precise lightning prediction model is urgently needed to mitigate economic losses caused by extreme lightning events. However, the model exhibits limited accuracy in regions with low lightning frequencies or low aerosol loading, primarily observed in the western CONUS. In these regions, many cases have been found to have high uncertainty in predicting lightning occurrence. This reduced applicability can be attributed to the models' limited ability to handle the imbalanced dataset between lightning-active and lightning-inactive cases. Despite efforts made in this study to address this imbalance issue, such as Focal Loss as a replacement for the conventional loss function, this challenge still restricts the application of the prediction model in lightning-sparse regions. To improve the model's applicability in the western CONUS, future enhancements in machine learning models should focus on addressing the imbalance issue. By tacking this challenge, the model's performance and applicability in regions with low lightning frequencies can be enhanced.

The development of data-driven models heavily relies on the quality of observation datasets. In the context of aerosol observations, the current data is obtained from CAMS forecast products, which incorporate real-time satellite observations into

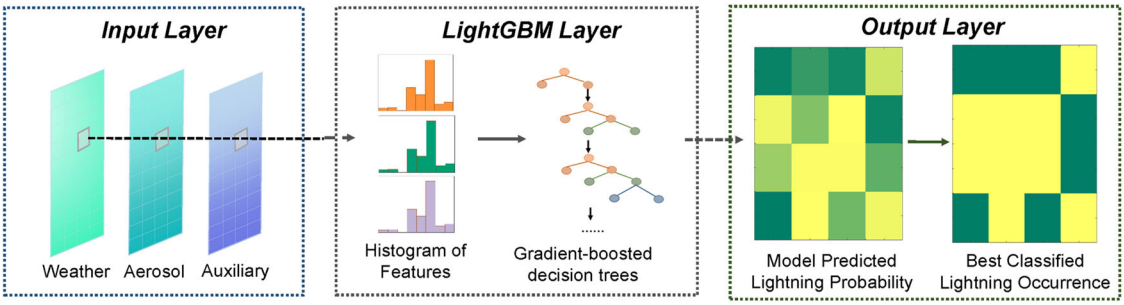

Fig. 7 Flowchart of the LightGBM model. Flow of the model as to integrate the meteorological, aerosol and auxiliary dataset into the LightGBM model, and to generate the prediction of lightning occurrence in the future hour.

numerical models to simulate atmospheric composition. In this study, we propose the utilization of real-time aerosol observations as predictors for lightning occurrence. Ideally, direct observations of aerosols from satellites would accurately capture aerosol composition and enable precise lightning prediction. However, existing aerosol products face limitations due to incomplete satellite imagery and inadequate coverage of valid aerosol information. As advancements in satellite retrievals for aerosols continue to evolve, it is expected that real-time aerosol-based lightning prediction models will exhibit improved performance.

METHODS

Lightning occurrence observations from GLM

In this study, the dataset of lightning occurrence observations is retrieved from the GLM onboard GOES-16 geostationary satellite, which is a satellite-borne single channel, near-infrared optical transient detector[54–56]. The GLM is the sensor onboard a geostationary satellite that can be used to map total lightning flashes with continuous regional observations and fine spatial resolution. The high-resolution GLM sensor (1372 × 1300 pixels) is equipped with a Charge Coupled Device (CCD) with a narrow band interference filter operating in the near infrared range (777.4 nm), with wide field of view (FOV) covering most of western hemisphere[57]. The spatial resolution of GLM is 8 km in the nadir, reaching 14 km at the edges, and it records lightning every ~2 ms and delivers compiled data file every 20 s[58,59]. GLM provides three levels of observations: events, groups and flashes, representing the individual illumination with resolution of 2 ms, lightning events in the same 2-ms time window and lightning groups that overlap within 15 km in space 330 ms in time, respectively (Goodman et al., 2013). In this study, we utilized flash level of GLM products to label the occurrence of lightning within a pixel. The GLM flash products detect all forms of lightning continuously, with a fine spatial resolution and detection efficiency of over 70%[30]. In this study, we retrieved the GOES-R Series GLM L2+ Data Product (GRGLMPROD) for our analysis. Owing to the above advantages of GLM, observation of the concentrated region that includes the contiguous United States with lightning detection with high temporal resolution (20 s) and full spatial coverage is possible for enriching the valid samples for observation-based data-driven lightning prediction model. Figure 3 depicts the summertime distribution of lightning density across CONUS. The majority of the lightning flashes occur in the southeastern CONUS, while the western CONUS does not observe frequent lightning flashes over the study period. The coastal regions in the southeastern CONUS show a higher record of lightning occurrences than the inland regions. The highest lightning flash density lies in Florida (mean value of 6.40 flash/km^2 and standard deviation of 5.79 flash/km^2).

Aerosol observation

To best characterize aerosols with full spatial coverage, the aerosol information utilized in this model consists of the aerosol optical depth of five aerosol components (including black carbon, dust, organic carbon, sulfate and sea salt), and the surface $PM_{2.5}$ concentration which represent the lower aerosol level as supplementary of the aerosol vertical information. In order to fulfil nowcasting of lightning occurrence, forecast products of aerosols are obtained in this study. The optical depth of individual aerosol component is obtained from Copernicus Atmosphere Monitoring Service (CAMS) global atmospheric composition forecast products[60,61]. The dataset is an hourly-level product provided by real-time forecasting service from the assimilation by combining a previous forecast with current satellite observations.

The real-time spatially continuous and hourly-level $PM_{2.5}$ dataset is obtained following a published method by Zeng[62], which uses machine learning models to estimate hourly-level spatially continuous surface $PM_{2.5}$ based on meteorological conditions and auxiliary information. In this method, the fundamental in-situ measurements are obtained from Air Quality System (AQS) monitoring network operated by United States Environmental Protection Agency. The datasets are validated by a 10-fold cross-validation method with R^2 of 0.791 and RMSE of 4.33 μg/m^3, shown in Supplementary Fig. 6a. Supplementary Figure 6b shows the mapped distribution of $PM_{2.5}$ with spatial continuity as a result of the model estimation.

Meteorological variables

Same as the previous studies[14,17], six meteorological factors that have certain indications on lightning are selected in the prediction model, including surface pressure (SP), temperature at 500 hPa (T500), relative humidity at 500 hPa (SH), 10 m U-component wind speed at 500 hPa (UW), 10 m V-component wind speed at 500 hPa (VW)[63]. To be consistent with the AOD, the selected meteorological factors are obtained from the same dataset of CAMS global atmospheric composition forecasts to avoid any error caused by the heterogeneity of data sources.

Both CAMS aerosol composition forecast and meteorological forecast interpolated to grids of 0.25° × 0.25° and one-hour temporal level by bilinear interpolation. The statistics of all variables included in the dataset are shown as Supplementary Table 2.

With the completion of data collection, the GLM dataset and CAMS products are pre-processed by gridding to 0.25°, followed by data filtering where suspicious noises and outliers are removed from the dataset. The noises and outliers are defined as the lightning occurrence whose flash is recorded less than five times in 5 min (approximately in 15 files), considering the intrinsic spatial and temporal continuity of lightning flash.

Lightning prediction model

In this study, we selected the LightGBM model to forecast the occurrence of lightning in the next hour (Fig. 7), considering its excellent performance in classification tasks while maintaining low computation cost.

As a highly efficient ML-model based on the Gradient Boosting Decision Tree, the LightGBM has been widely applied owing to its advantages of low computation cost and high learning accuracies, especially when processing large and complex datasets[64]. Given the large hourly and spatially continuous dataset for lightning prediction, LightGBM is considered the most suitable tool due to greatly reduced computation processing time and high accuracy.

The hyperparameters of the LightGBM model were optimized using a grid search strategy, where various combination of hyperparameters were tested in batches. The best combination of hyperparameters was selected based on the results of these tests. The optimized hyperparameter settings can be found in Supplementary Table 3.

To address the potential data unbalance problem (low fraction of lightning-active cases in the total cases), focal loss function is implemented in the lightGBM model, with the expression of loss function in Eq. (1). The weighting hyperparameters α and γ were introduced in the LightGBM layer to emphasize the misclassified positive classes. In our optimization process, we set $\alpha = 0.75$ and $\gamma = 0$ (in Supplementary Fig. 5) to achieve the desired balance between precision and recall in the model's predictions.

$$L = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where:

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{otherwise} \end{cases}$$

The parameters y denotes the ground-truth class and p denotes the model's estimated probability for the class with label $y = 1$.

The LightGBM model can be expressed as Eq. (2), where the subscript t represents the current moment, while the $t + 1$ represents the lightning status in the subsequent hour, which is the target of prediction. The model prediction result is a binary classification result, where 0 represents no lightning occurrence and 1 represents lightning occurrence within the next hour. The temporal information is captured through the inclusion of the day of year (DOY) and local hour (HH) as features. To address the significance of aerosols, the Eq. (3) removes the information on aerosols to predict the lightning occurrence and compares with Eq. (2).

$$[LN_{t+1} = \text{LightGBM}(DOY, HH, T500_t, SP_t, UW_t, VW_t, SH_t, BC_t, OC_t, DUST_t, SS_t, SO4_t, PM2.5_t, Flash_t)] \tag{2}$$

$$[LN_{t+1} = \text{LightGBM}(DOY, HH, T500_t, SP_t, UW_t, VW_t, SH_t, Flash_t)] \tag{3}$$

Model evaluation schemes

The study period is 2020 summer (June, July and August) when lightening is the most frequent across a year, and there are 37,415,530 records for training and testing the LightGBM model. We also evaluated the model's transferability by testing it on the 2021 summertime dataset, even though it was trained solely on the 2020 summertime dataset.

In this study, the model performance is evaluated by 10-fold day-based cross-validation method, which is a common evaluation approach to assess the model's overall performance. In each fold, datasets are divided into consecutive days spanning approximately 1/10 of total study period and subsequently) assigned as testing set while other data samples are assigned as training set. Then, the LightGBM machine learning model is trained with the training set and its performance is evaluated on the testing set. The process repeats for 10 times until all samples have been assigned as testing set for once. The overall performance of the model is determined by taking the average of all 10 runs.

Feature importance by interpretable machine learning module

To interpret the machine learning model and address the insight of the features, the Shapley Additive ExPlanation Approach (SHAP) method is applied on the LightGBM model (M1). SHAP has been widely used in recent studies to interpret the neural-network-based and tree-based machine learning models[65–67]. The SHAP approach distributes the total gains among the players based on coalitional game theory[68]. The SHAP can retrieve the quantitative contribution of each feature in each sample for a well-trained machine learning model, which can explain the machine learning model and interpret the importance of feature to a sample-specific view. In the SHAP theory, the different of a model prediction by a variable is contributed by its marginal contribution. Considering the interactive effects between the variables, every possible variable combination of each sample is computed[69]. Thus, the results can be interpreted as a linear summation of feature attributions, as expressed in Eq. (4). By interpreting the LightGBM model with SHAP, we can obtain the individual contribution of each feature to the occurrence of lightning, and the relative importance of each variable can be derived from that.

$$LN_{prob} = SHAP_o(M, x) + SHAP_i(M, x_i) \tag{4}$$

where

$SHAP_i$ represents the SHAP value of the i variable,

$SHAP_0$ represents the expected value of the model output for the dataset, LN_{prob} shows the predicted lightning occurrence probability, a continuous value between 0–1.

DATA AVAILABILITY

The lightning observation by Geostationary Lightning Mapper is retrieved via https://www.avl.class.noaa.gov/. The lightning observation by Lightning Mapping Array is retrieved via https://search.earthdata.nasa.gov/. The AQS PM2.5 observations based on ground-level sites are retrieved from https://aqs.epa.gov/aqsweb/airdata/. The CAMS aerosol and meteorological data are retrieved by https://ads.atmosphere.copernicus.eu/. Derived data supporting the findings of this study are available from the corresponding author upon reasonable request. Demo datasets for model training are available in Zenodo[70].

CODE AVAILABILITY

Important source codes and relative processing codes for the analysis of this study are available in Github[71].

Received: 22 February 2023; Accepted: 10 August 2023;
Published online: 24 August 2023

REFERENCES

1. Borden, K. A. & Cutter, S. L. Spatial patterns of natural hazards mortality in the United States. *Int. J. Health Geogr.* **7**, 1–13 (2008).
2. Cooper, M. A. & Holle, R. L. *Reducing Lightning Injuries Worldwide* 1st edn (Springer International Publishing, 2019).
3. Levy, H., Moxim, W. & Kasibhatla, P. A global three-dimensional time-dependent lightning source of tropospheric NOx. *J. Geophys. Res. Atmos.* **101**, 22911–22922 (1996).
4. Carey, L. & Rutledge, S. A multiparameter radar case study of the microphysical and kinematic evolution of a lightning producing storm. *Meteor. Atmos. Phys.* **59**, 33–64 (1996).

5. Kamra, A. K. & Ramesh Kumar, P. Regional variability in lightning activity over South Asia. *Int. J. Climatol.* **41**, 625–646 (2021).
6. Kotroni, V. & Lagouvardos, K. Lightning in the Mediterranean and its relation with sea-surface temperature. *Environ. Res. Lett.* **11**, 034006 (2016).
7. Xiong, Y. J., Qie, X. S., Zhou, Y. J., Yuan, T. & Zhang, T. L. Regional responses of lightning activities to relative humidity of the surface. *Chin. J. Geophys.* **49**, 311–318 (2006).
8. Griffiths, R. & Phelps, C. The effects of air pressure and water vapour content on the propagation of positive corona streamers, and their implications to lightning initiation. *Q. J. R. Meteorol. Soc.* **102**, 419–426 (1976).
9. Yair, Y. et al. Predicting the potential for lightning activity in Mediterranean storms based on the Weather Research and Forecasting (WRF) model dynamic and microphysical fields. *J. Geophys. Res. Atmos.* **115** (2010).
10. Lopez, P. A lightning parameterization for the ECMWF integrated forecasting system. *Mon. Weather Rev.* **144**, 3057–3075 (2016).
11. Gharaylou, M., Farahani, M. M., Mahmoudian, A. & Hosseini, M. Prediction of lightning activity using WRF-ELEC model: Impact of initial and boundary conditions. *J. Atmos. Sol. Terr. Phys.* **210**, 105438 (2020).
12. Zepka, G., Pinto, O. Jr & Saraiva, A. Lightning forecasting in southeastern Brazil using the WRF model. *Atmos. Res.* **135**, 344–362 (2014).
13. Giannaros, T. M., Kotroni, V. & Lagouvardos, K. Predicting lightning activity in Greece with the Weather Research and Forecasting (WRF) model. *Atmos. Res.* **156**, 1–13 (2015).
14. Mostajabi, A., Finney, D. L., Rubinstein, M. & Rachidi, F. Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *npj Clim. Atmos. Sci.* **2**, 1–15 (2019).
15. Pakdaman, M., Naghab, S. S., Khazanedari, L., Malbousi, S. & Falamarzi, Y. Lightning prediction using an ensemble learning approach for northeast of Iran. *J. Atmos. Sol. Terr. Phys.* **209**, 105417 (2020).
16. Leinonen, J., Hamann, U., Germann, U. & Mecikalski, J. R. Nowcasting thunderstorm hazards using machine learning: The impact of data sources on performance. *Nat. Hazards Earth Syst. Sci.* **22**, 577–597 (2022).
17. Moon, S.-H. & Kim, Y.-H. Forecasting lightning around the Korean Peninsula by postprocessing ECMWF data using SVMs and undersampling. *Atmos. Res.* **243**, 105026 (2020).
18. Essa, Y., Ajoodha, R. & Hunt, H. G. A LSTM recurrent neural network for lightning flash prediction within Southern Africa using Historical Time-series Data. In: *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (IEEE, 2020).
19. Leal, A. F., Rakov, V., Alves, E. R. & Lopes, M. N. Estimation of–CG lightning distances using single-station E-field measurements and machine learning techniques. In: *2019 International Symposium on Lightning Protection (XV SIPDA)* (IEEE, 2019).
20. Coombs, M. L. et al. Short-term forecasting and detection of explosions during the 2016–2017 eruption of Bogoslof volcano, Alaska. *Front. Earth Sci.* **6**, 122 (2018).
21. Cecil, D. J., Buechler, D. E. & Blakeslee, R. J. TRMM LIS climatology of thunderstorm occurrence and conditional lightning flash rates. *J. Clim.* **28**, 6536–6547 (2015).
22. Rudlosky, S. D. & Shea, D. T. Evaluating WWLLN performance relative to TRMM/LIS. *Geophys. Res. Lett.* **40**, 2344–2348 (2013).
23. Honda, T., Sato, Y. & Miyoshi, T. Potential impacts of lightning flash observations on numerical weather prediction with explicit lightning processes. *J. Geophys. Res. Atmos.* **126**, e2021JD034611 (2021).
24. Silva, S. J., Keller, C. A. & Hardin, J. Using an explainable machine learning approach to characterize earth system model errors: application of SHAP analysis to modeling lightning flash occurrence. *J. Adv. Model. Earth Syst.* **14**, e2021MS002881 (2022).
25. Heuscher, L., Liu, C., Gatlin, P. & Petersen, W. A. Relationship between lightning, precipitation, and environmental characteristics at mid-/high latitudes from a GLM and GPM perspective. *J. Geophys. Res. Atmos.* **127**, e2022JD036894 (2022).
26. Rodríguez-Pérez, J. R., Ordóñez, C., Roca-Pardiñas, J., Vecín-Arias, D. & Castedo-Dorado, F. Evaluating lightning-caused fire occurrence using spatial generalized additive models: a case study in central Spain. *Risk Anal.* **40**, 1418–1437 (2020).
27. Schultz, C. J., Andrews, V. P., Genareau, K. D. & Naeger, A. R. Observations of lightning in relation to transitions in volcanic activity during the 3 June 2018 Fuego Eruption. *Sci. Rep.* **10**, 1–12 (2020).
28. Murphy, K. M., Bruning, E. C., Schultz, C. J. & Vanos, J. K. A spatiotemporal lightning risk assessment using lightning mapping data. *Weather Clim. Soc.* **13**, 571–589 (2021).
29. Montanya, J. et al. Potential use of space-based lightning detection in electric power systems. *Electr. Power Syst. Res.* **213**, 108730 (2022).
30. Bateman, M. & Mach, D. Preliminary detection efficiency and false alarm rate assessment of the Geostationary Lightning Mapper on the GOES-16 satellite. *J. Appl. Remote Sens.* **14**, 032406 (2020).
31. Thornton, J. A., Virts, K. S., Holzworth, R. H. & Mitchell, T. P. Lightning enhancement over major oceanic shipping lanes. *Geophys. Res. Lett.* **44**, 9102–9111 (2017).
32. Altaratz, O., Kucienska, B., Kostinski, A., Raga, G. B. & Koren, I. Global association of aerosol with flash density of intense lightning. *Environ. Res. Lett.* **12**, 114037 (2017).
33. Altaratz, O., Koren, I., Yair, Y. & Price, C. The impact of aerosols on lightning activity in thunderstorms. In: *12th Plinius Conference on Mediterranean Storms* (2010).
34. Liu, Y. et al. Aerosol effects on lightning characteristics: a comparison of polluted and clean regimes. *Geophys. Res. Lett.* **47**, e2019GL086825 (2020).
35. Wang, Q., Li, Z., Guo, J., Zhao, C. & Cribb, M. The climate impact of aerosols on the lightning flash rate: is it detectable from long-term measurements? *Atmos. Chem. Phys.* **18**, 12797–12816 (2018).
36. Sun, M. et al. Aerosol effects on electrification and lightning discharges in a multicell thunderstorm simulated by the WRF-ELEC model. *Atmos. Chem. Phys.* **21**, 14141–14158 (2021).
37. Pan, Z. et al. Coarse sea spray inhibits lightning. *Nat. Commun.* **13**, 1–7 (2022).
38. Engel-Cox, J. A., Holloman, C. H., Coutant, B. W. & Hoff, R. M. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmos. Environ.* **38**, 2495–2509 (2004).
39. Hu, X., Waller, L., Lyapustin, A., Wang, Y. & Liu, Y. 10-year spatial and temporal trends of PM 2.5 concentrations in the southeastern US estimated using high-resolution satellite data. *Atmos. Chem. Phys.* **14**, 6301–6314 (2014).
40. Fuchs, B. R. & Rutledge, S. A. Investigation of lightning flash locations in isolated convection using LMA observations. *J. Geophys. Res. Atmos.* **123**, 6158–6174 (2018).
41. López, J. A. et al. Spatio-temporal dimension of lightning flashes based on three-dimensional Lightning Mapping Array. *Atmos. Res.* **197**, 255–264 (2017).
42. Tao, W. K., Chen, J. P., Li, Z., Wang, C. & Zhang, C. Impact of aerosols on convective clouds and precipitation. *Rev. Geophys.* **50** (2012).
43. Altaratz, O., Koren, I., Remer, L. & Hirsch, E. Cloud invigoration by aerosols—coupling between microphysics and dynamics. *Atmos. Res.* **140**, 38–60 (2014).
44. Shen, C. Analysis of detrended time-lagged cross-correlation between two non-stationary time series. *Phys. Lett. A* **379**, 680–687 (2015).
45. Larkin, N. K., Raffuse, S. M. & Strand, T. M. Wildland fire emissions, carbon, and climate: US emissions inventories. *Ecol. Manag.* **317**, 61–69 (2014).
46. Zhao, P. et al. Distinct aerosol effects on cloud-to-ground lightning in the plateau and basin regions of Sichuan, Southwest China. *Atmos. Chem. Phys.* **20**, 13379–13397 (2020).
47. Proestakis, E. et al. Aerosols and lightning activity: the effect of vertical profile and aerosol type. *Atmos. Res.* **182**, 243–255 (2016).
48. Orville, R. E. et al. Enhancement of cloud-to-ground lightning over Houston, Texas. *Geophys. Res. Lett.* **28**, 2597–2600 (2001).
49. Van Den Heever, S. C. & Cotton, W. R. Urban aerosol impacts on downwind convective storms. *J. Appl. Meteorol. Climatol.* **46**, 828–850 (2007).
50. Zhao, P. et al. Potential relationship between aerosols and positive cloud-to-ground lightning during the warm season in Sichuan, southwest China. *Front. Environ. Sci.* 1112 (2022).
51. Menon, S., Hansen, J., Nazarenko, L. & Luo, Y. Climate effects of black carbon aerosols in China and India. *Science* **297**, 2250–2253 (2002).
52. Jin, Q., Grandey, B. S., Rothenberg, D., Avramov, A. & Wang, C. Impacts on cloud radiative effects induced by coexisting aerosols converted from international shipping and maritime DMS emissions. *Atmos. Chem. Phys.* **18**, 16793–16808 (2018).
53. Shi, Z., Wang, H., Tan, Y., Li, L. & Li, C. Influence of aerosols on lightning activities in central eastern parts of China. *Atmos. Sci. Lett.* **21**, e957 (2020).
54. Edgington, S., Tillier, C. & Anderson, M. Design, calibration, and on-orbit testing of the geostationary lightning mapper on the GOES-R series weather satellite. In: *International Conference on Space Optics—ICSO 2018* (International Society for Optics and Photonics, 2019).
55. Mach, D. M. Geostationary Lightning Mapper clustering algorithm stability. *J. Geophys. Res. Atmos.* **125**, e2019JD031900 (2020).
56. Rudlosky, S. D., Goodman, S. J., Virts, K. S. & Bruning, E. C. Initial geostationary lightning mapper observations. *Geophys. Res. Lett.* **46**, 1097–1104 (2019).
57. Goodman, S. J. et al. The GOES-R geostationary lightning mapper (GLM). *Atmos. Res.* **125**, 34–49 (2013).
58. Oda, P. S., Enoré, D. P., Mattos, E. V., Gonçalves, W. A. & Albrecht, R. I. An initial assessment of the distribution of total Flash Rate Density (FRD) in Brazil from GOES-16 Geostationary Lightning Mapper (GLM) observations. *Atmos. Res.* **270**, 106081 (2022).
59. Peterson, M. J. et al. New World Meteorological Organization certified megaflash lightning extremes for flash distance (709 km) and duration (16.73 s) recorded from space. *Geophys. Res. Lett.* **47**, e2020GL088888 (2020).

60. Flemming, J. et al. Tropospheric chemistry in the Integrated Forecasting System of ECMWF. *Geosci. Model Dev.* **8**, 975–1003 (2015).
61. Courtier, P., Thépaut, J. N. & Hollingsworth, A. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**, 1367–1387 (1994).
62. Zeng, Z. et al. Estimating hourly surface PM2. 5 concentrations across China from high-density meteorological observations by machine learning. *Atmos. Res.* **254**, 105516 (2021).
63. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc*. **146**, 1999–2049 (2020).
64. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Proc. Adv. Neural Inf. Process. Syst*. **30**, 3146–3154 (2017).
65. Wu, Y., Lin, S., Shi, K., Ye, Z. & Fang, Y. Seasonal prediction of daily PM2. 5 concentrations with interpretable machine learning: a case study of Beijing, China. *Environ. Sci. Pollut. Res*. **29**, 45821–836 (2022).
66. Hou, L. et al. Revealing divers of haze pollution by explainable machine Learning. *Environ. Sci. Technol. Lett*. **9** (2022).
67. García, M. V. & Aznarte, J. L. Shapley additive explanations for NO2 forecasting. *Ecol. Inf.* **56**, 101039 (2020).
68. Shapley, L. S. In *Contributions to the Theory of Games II* (eds. Kuhn, H. & Tucker, A.) (Princeton University Press, 1953).
69. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Proc. Adv. Neural Inf. Process. Syst*. **30** (2017).
70. Song, G., Li, S. & Xing, J. *Machine-learning based lightning nowcasting data archive*, Zenodo https://doi.org/10.5281/zenodo.8141921 (2023).
71. Song, G. *lightning-nowcast-model*, Github https://github.com/ARSgesong/lightning-nowcast-model (2023).

ACKNOWLEDGEMENTS

This study was funded by The National Natural Science Foundation of China [No: 41975022] and The Foundation for Innovative Research Groups of the Hubei Natural Science Foundation [No: 2020CFA003].

AUTHOR CONTRIBUTIONS

S.G.: Methodology, Investigation, Model Development, Formal analysis, Writing—Original draft preparation, Writing—Reviewing and Editing. L.S.: Conceptualization, Investigation, Funding acquisition, Supervision, Writing—Reviewing and Editing. X.J.: Supervision, Writing—Reviewing and Editing.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-023-00451-x.

Correspondence and requests for materials should be addressed to Siwei Li.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.