Machine Learning based investigation of the variables affecting summertime lightning occurrence over the Southern Great Plains

Siyu Shan[1], Dale Allen[1], Zhanqing Li[2], Kenneth Pickering[1], Jeff Lapierre[3]

[1]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA
[2]Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Centre, University of Maryland, College Park, MD, USA
[3]Earth Networks, Inc., Advanced Environmental Monitoring (AEM), Germantown, Maryland, USA

Correspondence to: Zhanqing Li (zhanqing@umd.edu)

Abstract. Lightning is affected by many factors, many of which are not routinely measured, well understood, or accounted for in physical models. Several commonly used machine learning (ML) models have been applied to analyse the relationship between Atmospheric Radiation Measurement (ARM) data and lightning data from the Earth Networks Total Lightning Network (ENTLN) in order to identify important variables affecting lightning occurrence in the vicinity of the South Great Plains (SGP) ARM site during summer months (June, July, August and September) of 2012 to 2020. Testing various ML models, we found that the Random Forest model is the best predictor among common classifiers. When convective clouds were detected, it predicts lightning occurrence with an accuracy of 76.9 % and an area under curve (AUC) of 0.850. Using this model, we further ranked the variables in terms of their effectiveness in nowcasting lightning and identified geometric cloud thickness, rain rate and convective available potential energy (CAPE) as the most effective predictors. The contrast in meteorological variables between no-lightning and frequent-lightning periods was examined on hours with CAPE values conducive to thunderstorm formation. Besides the variables considered for the ML models, surface variables and mid-altitude variables (e.g., equivalent potential temperature and minimum equivalent potential temperature, respectively) have statistically significant contrasts between no-lightning and frequent-lightning hours. For example, the minimum equivalent potential temperature from 700 hPa to 500 hPa is significantly lower during frequent-lightning hours compared with no-lightning hours. Finally, a notable positive relationship between intra-cloud (IC) flash fraction and the square root of CAPE

($\sqrt{CAPE}$) was found suggesting that stronger updrafts increase the height of the electrification zone, resulting in fewer flashes reaching the surface and consequently a greater IC flash fraction.

1 Introduction

Thunderstorms are most common during the warm season when high moisture and buoyant instability are available (Doswell III et al., 1996). The frequency of lightning is related to multiple meteorological variables including convective available potential energy (CAPE), rain rate, geometrical cloud thickness, wind shear, and multiple microphysical variables such as the diameter of ice crystals (Sherwood et al., 2006; Lal et al., 2014) and cloud droplet size (Orville et al., 2001). CAPE plays an important role in lightning activity (Pawar et al., 2012; Romps et al., 2014; Romps et al., 2018) with the magnitude and vertical distribution of CAPE affecting the updraft velocity and vertical distribution of cloud water path and consequently the lightning charge generation process inside deep convective clouds (Williams, 2017). When studying daily records of flashes, Williams et al. (2002) found a CAPE threshold of approximately 1000 $J \cdot kg^{-1}$ above which lightning is likely. Both lightning activity and rainfall in deep convective systems are physically related to mixed-phase cloud processes involving super-cooled water, ice and graupel. Heavy glaciation aloft is essential to produce frequent lightning activity (Williams et al., 1989). Monthly and seasonal correlation coefficients between precipitation and lightning counts were found to vary between 0.81 and 0.98 over the central and eastern Mediterranean Sea during winter time (Price and Federmesser, 2006). The influence of cloud thickness on lightning is complicated. According to Takahashi (1978), the mixed phased zone of convective clouds is crucial for the charge separation mechanism. Warm cloud depth is defined as vertical thickness between the lifting condensation level (LCL) and the freezing level (0 °C). Cold cloud depth is defined as the thickness from the freezing level to the storm top. The depth of the warm cloud region is critical for determining the cloud droplets growth. A larger warm cloud depth is likely to enhance the efficiency of warm rain–collision–coalescence processes, and lower the altitude at which precipitation forms thus lessening the number of droplets available to be lofted into the mixed-phase region. where they can affect electrification in the thunderstorm (Carey and Buffalo, 2007). The mixed phase region includes graupel and ice crystals, so it is closely related to the lightning activity. Price and Rind (1992) showed that the lightning flash rate within a convective cloud is proportional to the fifth power of the cloud-top height. Furthermore, Yoshida et al. (2009) found that the number of lightning flashes per second per convective cloud is proportional to the fifth power of the cold-cloud depth regardless of location. Wind shear's influence on convective systems is mixed. Richardson et al. (2007) found that strong wind shear may weaken the vertical development of an isolated supercell. Wind shear at different levels can play different roles in convective systems. According to Chen et al. (2015), increasing wind shear in the lower troposphere results in a more organized quasi-linear convective system. By increasing wind shear at the upper vertical levels only, the convective intensity is weakened but the structure is not affected much. Bang and Zipser (2016) analysed wind shear in the lowest 200 hPa of the atmosphere and found that the magnitude of the wind shear is a poor discriminator of lightning occurrence. Stolz et al. (2017), based on an analysis over multiple regions, found that total lightning density increases with increasing wind shear, but the signal is relatively weak compared with other variables.

Both natural and anthropogenic aerosols affect lightning activity (Westcott, 1995; Altaratz et al., 2010; Wang et al., 2011; Li et al., 2019; Zhao et al., 2020; Sun et al., 2021). High aerosol loading related to volcanic activity is closely correlated with general lightning activity at different time scales (Yuan et al., 2011), and smoke caused by man-made forest fires increases cloud condensation nuclei (CCN) concentrations during the Amazon dry season, invigorating the electrical activity in the low aerosol loading environment (Altaratz et al., 2010). Weekly cycles in lightning activity are also observed (Bell et al., 2009) and are consistent with cycles in precipitation over the southeast US (Bell et al., 2008). This apparent weekly cycle in afternoon lightning activity, peaking on Wednesday and with a minimum on Saturday and Sunday, can only be explained by aerosol's weekly cycle, given the fact that no significant dynamical or thermal weekly cycle is observed. Enhanced lightning activity is observed over two of the world's busiest shipping lanes in the Indian Ocean and the South China Sea, which cannot be explained by meteorological factors, and is therefore likely due to aerosol particles emitted from the ship engines (Thornton et al., 2017). Wang et al. (2018) found that the type of aerosol affects lightning formation with much higher flash rates in moist central Africa than dry northern Africa. In both regions, the lightning flash rate changes with aerosol optical depth in a boomerang shape: first increasing with aerosol optical depth up to approximately 0.3, and then decreasing for dust and flattening for smoke aerosols.

There are two types of lightning flashes: cloud-to-ground (CG) flashes and intra-cloud (IC) flashes. Many approaches have been used to predict flash types, which involve complicated interactions between atmospheric processes. For example, a new prognostic variable, potential electrical energy, is introduced to the Weather Research and Forecasting (WRF) cloud-resolving model to predict the dynamic contribution of the grid-scale-resolved microphysical and vertical velocity fields, so that it can be used to predict both CG and IC flashes in convection-allowing forecasts (Lynn et al., 2012). Using observations, the product of CAPE and precipitation explains 77 % of the variance in the time series of total CG flashes over the contiguous United States (Romps et al., 2014). Therefore, Tippett and Koshak (2018) used the product of CAPE and rain rate as a proxy to predict CG lightning over the US and produce CG lightning threat forecasts.

Today, lightning prediction remains challenging because lightning production is stochastic involving microphysical and thermodynamic processes. In recent years, machine learning (ML) based predicting or nowcasting of lightning occurrence has become popular. A four-parameter model based on four commonly available surface weather variables (air pressure at station level, air temperature, relative humidity and wind speed) developed by Mostajabi et al. (2019) has considerable predictive skill for lightning occurrence and produces warnings for lead times up to 30 minutes. The importance of the input variables in this model fits with the generally accepted physical understanding of surface processes driving thunderstorms. CG lightning damages infrastructure, leads to the loss of life and ignites forest fires (Cooper et al., 2019). Therefore, ML-based prediction of CG lightning is increasing. For example, La Fata et al. (2021) used ML to nowcast the spatial distribution of CG flashes, while He et al. (2020) used a ML algorithm based on a WRF simulation to predict CG lighting over the Alaskan tundra.

In this study, we use a ML model to investigate the meteorological variables affecting lightning occurrence over the Southern Great Plains during summer. Then, the contrast in variables between no-lightning and frequent-lightning hours is shown for strong convective environments. Lastly, the IC fraction's relationship with the square root of CAPE ($\sqrt{CAPE}$) and its potential physical mechanism is discussed.

The scientific questions we address are which variables are the most important to predict the lightning occurrence. Previous research focused on one or two variables, or one class of variables to determine their impact on lightning. What is new here is that we develop a systematic approach to narrow and choose the variables.

2 Data

2.1 Earth Networks Total Lightning Network (ENTLN)

ENTLN is a total lightning detection system and consists of over 1800 sensors deployed in over 100 countries. It detects wideband (1 Hz to 12 MHz) electric field signals emitted by both IC and CG lightning. In addition, for each flash, exact time, geolocation and peak current are recorded as well (Zhu et al., 2022). ENTLN records the flash type, IC or CG, and also provides an estimation of the source height of IC flashes. Typically, signal timing measurements from at least 5 sensors are able to determine the latitude, longitude, height and time that define the source location. The more sites that are used the smaller the uncertainty becomes (Heckman, 2014).

In this study, we use ENTLN flashes within the 1° × 1° grid box (36°-37° N, 97°-98° W) that includes the Atmospheric Radiation Measurement (ARM) South Great Plains (SGP) site. Hourly flash records of summer months (June, July, August and September) from 2012 to 2020 are used.

2.2 ARM

Multiple datasets are collected at the US Department of Energy ARM program SGP site, which is located at 36.6° N, 97.5° W. The SGP atmospheric observatory was the first field measurement site established by the ARM user facility, and it is currently one of the world's largest and most extensive climate research facilities. Variables including convective cloud thickness, rain rate, and >10 dBz vertical extent are downloaded or calculated from various ARM SGP datasets, and are considered to be representative of the entire 1° × 1° region. We discuss the detailed processing method in Section 3.3.

2.3 Other data sources

The wind shear values used in this study are calculated using fields from the "ERA5 hourly data on pressure levels from 1959 to present" dataset (Hersbach et al., 2023). ERA5 is the 5th generation ECMWF reanalysis for global climate and weather. The analysis is produced at a 1-hour time resolution using an advanced 4D-var assimilation scheme (Hersbach et al., 2020). The eastward and northward components of the wind with a 0.25° × 0.25° spatial resolution (centred at 36.5° N, 97.5° W) are downloaded for 750, 500 and 250 hPa levels. The hourly wind shear is then calculated between 750 and 500 hPa, and between 750 and 250 hPa:

$$Wind\ Shear\ (750/250\ \text{hPa}) = \sqrt{(u_{750\ \text{hPa}} - u_{250\ \text{hPa}})^2 + (v_{750\ \text{hPa}} - v_{250\ \text{hPa}})^2}$$

$$Wind\ Shear\ (750/500\ \text{hPa}) = \sqrt{(u_{750\ \text{hPa}} - u_{500\ \text{hPa}})^2 + (v_{750\ \text{hPa}} - v_{500\ \text{hPa}})^2}$$

Fine particulate matter ($PM_{2.5}$) concentrations are obtained from the US Environmental Protection Agency Air Quality System database. We have taken the average value of hourly surface $PM_{2.5}$ concentrations measured in the nearby counties of Kay (in Oklahoma, 36.7° N, 97.1° W), Sedgwick (in Kansas, 37.7° N, 97.3° W) and Sumner (in Kansas, 37.5° N, 97.4° W). One measurement is available in each county.

The column aerosol optical thickness (AOT) used in this study comes from the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2), which is the latest version of global atmospheric reanalysis for the satellite era produced by NASA Global Modeling and Assimilation Office using the Goddard Earth Observing System Model (GEOS) version 5.12.4. The dataset covers the period of 1980-present. M2T1NXAER (or tavg1_2d_aer_Nx) is an hourly time-averaged 2-dimensional data collection in MERRA-2. This collection consists of assimilated aerosol diagnostics, and the data field is time-stamped with the central time of hours starting from 00:30. We aggregated the 0.625° × 0.500° MERRA-2 data onto a 1° × 1° grid using an on-line function. The function used distance-weighted averaging to remap to the GPCC1.0 grid (Level 3 and 4 Regridder and Subsetter Information). For aerosol extinction, we took the 36°-37° N, 97°-98° W grid box value of the total aerosol extinction AOT at 550 nm.

3 Methods

In this section, the Random Forest classifier is introduced and various ML related terms are defined. It will be shown that the Random Forest classifier has the best performance among all common classifier ML models.

3.1 Area under curve (AUC) calculation

The receiver operating characteristics (ROC) curve was first used in signal detection theory to represent the trade-off between hit rates and false alarm rates (Green and Swets, 1966). For a ML classifier model, a positive or negative prediction for a certain threshold will be made for a given set of input variables. A confusion matrix is then made that records the frequency of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) predictions. The true positive rate (TPR, TPR=TP/(TP+FN)) and false positive rate (FPR, FPR= FP/(FP+TN)) can be calculated accordingly. TPR and FPR vary with threshold and we can put (FPR, TPR) points on the ROC space as the threshold changes. Both FPR and TPR range in value from 0 to 1, and we connect the points to get an ROC curve. The area under the ROC curve integrating from 0 to 1 is called AUC, which measures the discriminatory power of the predictive classification model.

3.2 Random Forest classifier and 10-fold cross validation

The Random Forest classifier is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by the most trees.

Ten-fold cross-validation is a resampling procedure commonly used to evaluate ML models. First, the dataset is shuffled randomly and split into 10 groups, which is a necessary step. Nine of the groups are used for training and the other group for evaluation. In this application, we predict the occurrence of lightning (Yes vs No) using 9 training groups and evaluate the prediction using the remaining group. The RepeatedStratifiedKFold classifier includes multiple adjustable parameters including number of trees (n_estimators, defaults to 100) and maximum depth of the tree (max_depth, defaults to “none”). Setting the maximum depth to “none” ensures that the nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split (default to 2) samples, which is the minimum number of samples required to split an internal node. The parameter min_samples_leaf (default to 1) is the minimum number of samples required to be at a leaf node. We tried each parametrization option in Table 1 for 50 times but found that the default parameters provided the best or nearly the best performance. Since random shuffling can affectthe performance we have chosen to retain the default parameters. Our model only simulated the convective hours over SGP, when convective clouds are detected from ARM SGP site, which won't cause temporal auto-correlation since convective clouds do not occur frequently (817 hours in total among 9 summers).

Parameters		Accuracy	AUC
Default: n_estimators = 100, max_depth = None, min_samples_split = 2, min_samples_leaf = 1		76.9 % ± 0.3 %	0.850 ± 0.002
Changing n_estimators only	n_estimators = 10	74.4 % ± 0.4 %	0.812 ± 0.003
	n_estimators =50	76.7 % ± 0.3 %	0.846 ± 0.002
	n_estimators = 200	76.9 % ± 0.3 %	0.853 ± 0.001
Changing max_depth only	max_depth = 5	74.6 % ± 0.2 %	0.832 ± 0.001
	max_depth = 10	76.5 % ± 0.3 %	0.848 ± 0.001
	max_depth = 50	77.0 % ± 0.3 %	0.851 ± 0.002
Changing min_samples_split only	min_samples_split = 3	76.8 % ± 0.3 %	0.849 ± 0.002
	min_samples_split = 4	76.7 % ± 0.3 %	0.850 ± 0.001
Changing min_samples_leaf only	min_samples_leaf = 2	76.5 % ± 0.2 %	0.847 ± 0.001
	min_samples_leaf = 3	76.1 % ± 0.3 %	0.845 ± 0.001

Table 1: Random Forest classifier performances using different parameters.

3.3 ARM dataset processing

Cloud top height, cloud base height and cloud type are obtained from CLDTYPE data product with a temporal resolution of 1 minute. For the SGP site, deep convective clouds are identified as clouds with cloud base height lower than 3.5 km and cloud top height higher than 6.5 km (Flynn et al., 2017). We use this product to identify convective clouds and calculate the convective cloud thickness from the cloud base to the cloud top. For each hour, the variable "Cloud Thickness" is obtained by averaging thicknesses for each minute during the hour with convective clouds. Rain rate is measured with a temporal resolution of 1 minute (Bartholomew, 2016) and contained in the VDIS product. The variable "Rain Rate" is the hourly sum. The ARSCLKAZR1KOLLIAS data product provides us with zenith-pointing radar reflectivity profiles at Ka-band (35 GHz) every 4 seconds with a vertical resolution of 30 meters. According to Seo and Liu (2005), the relationship between radar reflectivity and ice water content for the six ice particle types near the ARM SGP site show that ice water content for each vertical layer is proportional to the 0.79th power of radar reflectivity. We have set a threshold of 10 dBz for layers, so that each layer will have at least 0.5 $g \cdot m^{-3}$ ice water content when its radar reflectivity exceeds this threshold. We only take measurements of radar reflectivity at altitudes higher than 3 km as temperatures at altitudes lower than this are always too low to support ice in clouds during the summer at the ARM SGP site. The hourly average extent and centroid of radar reflectivity exceeding 10 dBz are recorded as variables "Radar Reflectivity > 10 dBz Extent" and "Radar Reflectivity > 10 dBz Centroid", respectively. These variables are chosen because they are closely associated with mixed-phase clouds extent and height. Fifty-four environmental variables (Jensen et al., 1998) are measured every minute and recorded in INTERPOLATEDSONDE. We primarily use the profiles of pressure, temperature and dew point and calculate the meteorological variables listed in Table 2 for this product. We use the 30th minute profile of the hour to calculate these variables, except for CAPE. We calculate the average CAPE based on the 15th minute and 45th minute profiles of the hour. We do not calculate values for each minute due to computational expense. The AOSCCN1COL and AOSCCN2COLAAVG data sets include cloud CCN concentrations every minute. We use both datasets because the AOSCCN1COL data set ended in September of 2017 and the AOSCCN2COLAAVG data set started from April of 2017. The data sets both measure the CCN concentrations at different supersaturation levels by manipulating the supersaturation in the instruments from 0.1 % to 1.2 %, although there are some minor changes in technique. According to Politovich and Cooper (1988), the maximum supersaturation is usually smaller than 0.5 % in cumulus clouds. Thus, we have selected all CCN concentration measurements at supersaturation in the range from 0.4 % to 0.6 %, and calculated the average value for each hour. Measurements of planetary boundary layer height (PBLH) are conducted every 30 seconds using micropulse lidar (MPL) and recorded in PBLHTMPL1SAWYERLI. We take the average value of PBLH for each hour and record them as variable "PBLH".

Because of differences in temporal resolution and formatting, each variable from the ARM SGP site is merged into a database at a temporal resolution of 1 hour (Table 2).

Data Product Name	Variables Obtained or Derived from Dataset
CLDTYPE	Convective cloud type, Cloud Thickness

VDIS	Rain Rate
ARSCLKAZR1KOLLIAS	Radar Reflectivity > 10 dBz Extent, Radar Reflectivity > 10 dBz Centroid
INTERPOLATEDSONDE	CAPE, Surface Equivalent Potential Temperature, 0 °C Freezing Level Height
AOSCCN1COL, AOSCCN2COLAAVG	CCN Concentration
PBLHTMPL1SAWYERLI	Planetary Boundary Layer Height

Table 2: Data sets containing the variables considered for use in the lightning parameterization.

4 Results

4.1 ML based investigation of the variables affecting lightning occurrence

First, we identified convective hours using the CLDTYPE product at the ARM SGP site. This product provides an automated cloud type classification based on microphysical quantities derived from vertically pointing lidar and radar. Twenty-four hours were checked each day.

Numerous meteorological variables were considered for use in the ML based analysis. Eight mostly independent variables (i.e., variables with inter-correlations |R| of 0.5 or less) were selected for further analysis. These variables and their inter-correlations are shown in Figure 1. In total, there were 817 hours with detectable deep convective clouds and measurements of all eight variables available during JJAS of 2012-2020. Lightning was observed in the 1° × 1° grid box (36°-37° N, 97°-98° W) encompassing the ARM site in 509 of those hours.

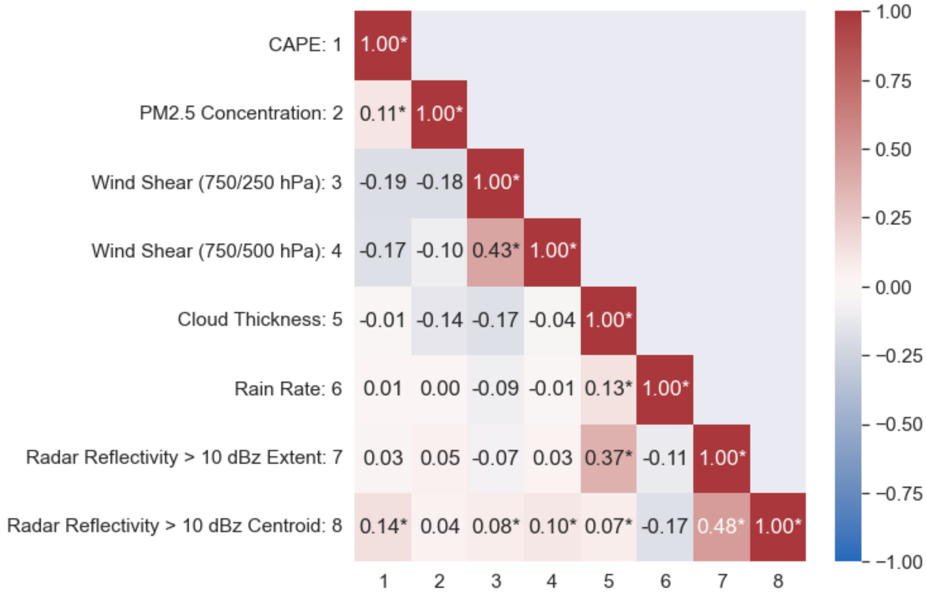

Figure 1: Pearson correlation coefficients between variables selected for use in the ML analysis. Asterisks indicate that correlations are significant at the 95% level. The relatively low correlations between the pairs make them good candidates for the analysis.

In addition to the Random Forest method, five other ML classifier schemes were tested. The Support Vector Machine (SVM) algorithm fits a hyperplane in space. The dimensions of the hyperplane are equal to the number of features. This approach results in a distinct classification of data points, by using different kernels containing a set of mathematical functions to massage the data. Linear and radial basis function (RBF) kernels are two different kernels using in SVM. Logistic Regression is a classification algorithm used to predict a binary outcome based on a set of independent variables and the Sigmoid function. Decision Tree is a tree-like structure where each internal node tests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction. Gaussian Naive Bayes is based on the probabilistic approach and Gaussian distribution, which assumes that each parameter has an independent capacity of predicting the output variable. Our goal is to use the 8 input variables to predict the occurrence of lightning in a convective hour. We repeat 10-fold cross-validation 50 times in order to estimate the overall performance of different ML models. Based on our 50 simulations with the 10-fold cross-validation, the Random Forest Classifier was identified as the best classifier among the common classifiers shown below, because of its highest accuracy and its AUC value (Table 3).

Classifier Name	Accuracy	AUC
SVM with linear kernel	72.1 % ± 0.1 %	0.797 ± 0.001
SVM with RBF kernel	74.0 % ± 0.2 %	0.821 ± 0.001
Random Forest	76.9 % ± 0.3 %	0.850 ± 0.002
Logistic Regression	72.3 % ± 0.1 %	0.800 ± 0.001
Decision Tree	69.8 % ± 0.5 %	0.679 ± 0.004
Gaussian Naive Bayes	74.2 % ± 0.2 %	0.812 ± 0.002

Table 3: Mean accuracy and AUC with standard deviation for each ML classifier method. Each method was run 50-times using 10-fold cross-validation. The accuracy is defined as the ratio of correct predictions of lightning occurrence (Yes vs. No) to total predictions. AUC provides an aggregate measure of performance across all classification thresholds and can have values ranging from 0.5 to 1.0. Models with higher values of AUC do a better job of distinguishing between convective hours with and without lightning.

After choosing the Random Forest classifier model, we split the dataset randomly into training and test sets with split percentages of 75 % to 25 % and performed 1000 simulations with the Random Forest Classifier to evaluate its overall performance, as shown in the confusion matrix (Table 4). This classifier predicts lightning occurrence with an accuracy of 77 % using these 8 input variables.

	Prediction: No Lightning	Prediction: Lightning
Truth: No Lightning	24.5 % ± 2.5 %	13.0 % ± 2.6 %
Truth: Lightning	10.4 % ± 2.5 %	52.1 % ± 2.9 %

Table 4: This confusion matrix shows the accuracy of the Random Forest Classifier. The frequency (percentage and standard deviation) of the binary prediction that fell into each of the 4 categories is shown. The overall accuracy, sum of the true negatives (24.5 %) and true positives (52.1 %), is about 77 %.

In addition to the confusion matrix, an overall ranking of feature importance is also generated from the ML model, as shown in Figure 2. The feature importance is a measure of how much one variable decreases the impurity, i.e., the probability that more than one class of data remains in a node after processing through various decision trees in the forest. This figure shows the percent of the 1000 simulations that each variable was identified as the most important feature (column #1) to the least important feature (column #8). For example, the variable "Cloud Thickness" was identified as the most important feature in 57.4 % of the 1000 runs, while it is the second (#2), the third (#3) and the fourth (#4) most important feature in 33.4 %, 8.9 % and 0.3 % of the runs. From the ranking distribution, we can identify that "Cloud Thickness", "Rain Rate" and "CAPE" are the top 3 important variables determining lighting occurrence in the model. The sum of the 1st, 2nd and 3rd place percentages for each of these variables exceed 90 %. The next 3 most important variables are "Radar Reflectivity > 10 dBz Extent", "Wind Shear (750/250 hPa)" and "Radar Reflectivity > 10 dBz Centroid". The least important variables are "$PM_{2.5}$ Concentration" and "Wind Shear (750/500 hPa)". The low sensitivity to $PM_{2.5}$ concentrations could be due to its small range of variability, especially compared with other variables' relatively large variations. The differentiation of the variables into most, modest and least important categories is distinct according to the robust ranking distribution, as can be seen in Figure 2. The importance of variables can also be estimated by removing them from the model and seeing how successful the remaining variables are at predicting the true outcome. In section 3.2, our model was found to have an overall accuracy of 76.9 % and AUC of 0.850. By removing "Cloud Thickness", "Rain Rate" and "CAPE" separately, the accuracy dropped from 76.9 % to 72.1 %, 75.6 % and 74.3 %, and the AUC dropped from 0.850 to 0.797, 0.830 and 0.821. The impact of removing other variables was smaller. Based on these metrics, the "Cloud Thickness" is most important followed by "Rain Rate" and then "CAPE".

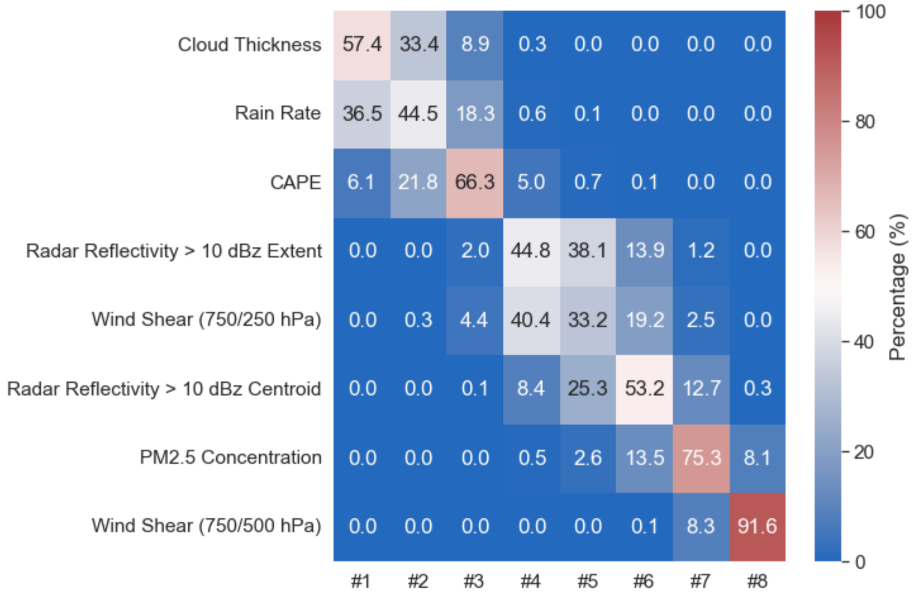

Figure 2: Variables importance ranking distribution. Probability distribution function showing the frequency that each variable was rated as the most-to-least important after running the Random Forest Classifier with random splitting 1000 times.

4.2 Contrast in meteorological variables between no-lightning and frequent-lightning hours in strong convective environments

ML provides robust feature importance rankings, which are useful for determining the importance of each variable. To aid in physical interpretation, we compare the meteorological variables' difference with or without the existence of lightning. In convective hours with lightning, the hourly flash count distribution is shown in Figure 3. The average and median number of ENTLN flashes per hour in the 1° × 1° grid box containing the SGP site are 864.9 and 162.5 respectively, with the large difference indicating the distribution is skewed by hours with very frequent lightning.

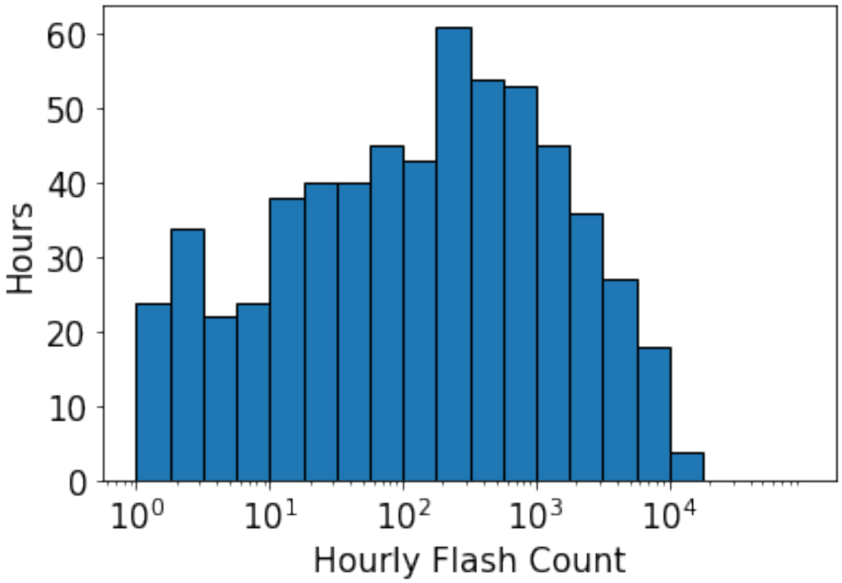

Figure 3: Hourly flash count distribution for flashing hours in the 1° × 1° grid box containing the SGP site. Note that the x-axis is logarithmic. There are 608 hours with both flashes and convective clouds detected, accounting for 2.3% among all 26352 hours in the summer months (June, July, August and September) from 2012 to 2020.

To ensure the environment is favourable for lightning, we have set a threshold of CAPE = 2000 $J\cdot kg^{-1}$ and only selected hours with convective clouds when CAPE is larger than 2000 $J\cdot kg^{-1}$, where 2000 $J\cdot kg^{-1}$ is chosen as the threshold for a strong convective environment following several studies (Rutledge et al., 1992; Chaudhuri, 2010; Chaudhuri and Middey, 2012; Hu et al., 2019). Overall, there were 175 hours satisfying the CAPE threshold. Of these hours, 41 had no lightning in a three-hour period centred on the CAPE observation and were labelled as “no-lightning hours”. Seventy-five of the hours had three-hour mean flash rates exceeding the median flash rate of 162.5 and were classified as “frequent-lightning hours” while the remainder of the hours (59) were deemed intermediate lightning hours.
The contrast in meteorological variables between the no-lightning and frequent-lightning hours is shown in Table 5.

Meteorological Variables	No-Lightning Hours	Frequent-Lightning Hours	p-value

CAPE ($J \cdot kg^{-1}$)	2627 ± 712	2669 ± 585	> 0.05
Rain Rate ($mm \cdot h^{-1}$)	0.16 ± 0.51	8.57 ± 15.25	< 0.001
Cloud Thickness (km)	6.22 ± 2.13	10.66 ± 3.04	< 0.001
Wind Shear (750/250 hPa) ($m \cdot s^{-1}$)	14.63 ± 5.17	12.04 ± 6.22	< 0.05
Wind Shear (750/500 hPa) ($m \cdot s^{-1}$)	9.45 ± 4.17	7.65 ± 3.85	< 0.05
Radar Reflectivity > 10 dBz Extent (km)	0.13 ± 0.24	1.31 ± 1.60	< 0.001
Radar Reflectivity > 10 dBz Centroid (km)	4.80 ± 1.23	5.73 ± 1.96	< 0.01
$PM_{2.5}$ Concentration ($\mu g \cdot m^{-3}$)	10.92 ± 3.83	6.57 ± 3.89	< 0.001

Table 5: The contrast (mean, standard deviation, and significance of difference) in meteorological variables put into ML Random Forest model between no-lightning and frequent-lightning hours. For example, a p-value of "< 0.01" indicates that the difference is significant at the 99 % level.

After limiting the analysis to hours with CAPE > 2000 $J \cdot kg^{-1}$, we do not see a significant difference in CAPE between the no-lightning and frequent-lightning hours, indicating that once CAPE reaches a high value, it is no longer a good predictor for lightning occurrence. The rain rate and convective cloud thickness are much larger when lightning is frequent (p-value less than 0.001), indicating that lightning is associated with high rain rates and deep convective clouds. This finding is consistent with the ML results. The "Radar Reflectivity > 10 dBz Extent" variable is an order of magnitude larger when lightning occurs, indicating that total ice water path, which is associated with high values of radar reflectivity is also much higher. In addition, the centroid altitude of radar reflectivity is higher by about 19 %, a difference that is significant at the 99 % confidence interval (CI). Mean vertical wind shear is smaller when flashes are present with decreases of about 18 % in 750 to 250 hPa shear (significant at 95 % CI) and 20% in 750 to 500 hPa shear (significant at 95 % CI). Perhaps surprisingly, differences in $PM_{2.5}$ between the non-flashing and frequent-flashing hours are significant. Specifically, hours with frequent lightning have 40 % less $PM_{2.5}$ than no-lightning hours. This result is seemingly inconsistent with the ML analysis discussed earlier, which showed that $PM_{2.5}$ had little effect on lightning occurrence and with previous studies finding that enhanced lightning activity is related to higher aerosol loading.

CCN concentrations are also more than 30 % smaller during frequent-lightning hours than no-lightning hours (Table 6). Similarly, values of MERRA-2 Total Aerosol Extinction AOT at 550 nm simultaneous with convective hour are lower in frequent-lightning hours than in no-lightning hours (significant at 99 % CI). One plausible explanation for this is aerosol wet removal, given the fact that lightning occurrence is closely related to precipitation. Therefore, we examined the $PM_{2.5}$ and CCN concentrations during the convective hour and also during the hours preceding the convective hour as shown in Table 6. During all these hours, we still notice less $PM_{2.5}$ concentration when flashes are frequent, but differences in CCN concentrations are small and insignificant statistically.

Another possible explanation is mixing of pollutants throughout the planetary boundary layer (PBL). A higher PBLH is associated with greater vertical mixing and often a larger CAPE and higher surface temperature (Zhang et al., 2013). Sun and Liang (2020) found that higher PBLHs were common during extreme precipitation. Both higher CAPE and higher precipitation rates are related to lightning occurrence. We calculated the product of PBLH and $PM_{2.5}$ or CCN concentration, assuming that pollutants are distributed homogeneously within the PBL and compared the values between no-lightning and frequent-lightning hours. Even though differences in the product of CCN concentration and PBL height between no-lightning and frequent-lightning periods were nearly 50 %, the differences were insignificant at the 95 % CI due to large variability. Thus, mixing through a deeper PBL could be the cause of the differences in $PM_{2.5}$ and CCN concentrations.

Meteorological Variables	No-Lightning Hours	Frequent-Lightning Hours	p-value
$PM_{2.5}$ Concentration in convective hours ($\mu g \cdot m^{-3}$)	10.92 ± 3.83	6.57 ± 3.89	< 0.001
$PM_{2.5}$ Concentration 1 hour before convective hours ($\mu g \cdot m^{-3}$)	10.94 ± 3.82	7.08 ± 4.28	< 0.001
CCN Concentration in convective hours (cm^{-3})	1640.13 ± 401.55	1094.22 ± 1205.45	< 0.05

CCN Concentration 1 hour before convective hours (cm^{-3})	1616.69 ± 384.77	1249.85 ± 1222.03	> 0.05
MERRA-2 Total Aerosol Extinction AOT at 550 nm in convective hours	0.28 ± 0.10	0.22 ± 0.10	< 0.05
$PM_{2.5}$ Concentration in convective hours × PBLH ($\mu g \cdot m^{-3} \cdot km$)	10.88 ± 6.31	9.45 ± 8.36	> 0.05
CCN Concentration in convective hours × PBLH ($cm^{-3} \cdot km$)	1377.31 ± 493.49	1999.69 ± 3402.52	> 0.05

Table 6: The contrast (mean, standard deviation, and significance of difference) in aerosol-related variables between no-lightning and frequent-lightning hours.

Some additional meteorological variables are calculated from the INTERPSONDE data set at the ARM SGP site. This value-added product provides us with profiles of pressure, temperature and dew point. The contrast of these meteorological variables between no-lightning and frequent-lightning hours is shown in Table 7.

Meteorological Variables	No-Lightning Hours	Frequent-Lightning Hours	p-value
LCL Height (km)	1.11 ± 0.41	1.11 ± 0.47	> 0.05
0 °C Freezing Level Height (km)	4.78 ± 0.22	4.64 ± 0.20	< 0.001
Surface Equivalent Potential Temperature (K)	356.64 ± 4.81	353.39 ± 6.42	< 0.01
Minimum Equivalent Potential Temperature from 700 hPa to 500 hPa (K)	333.30 ± 4.18	328.84 ± 4.21	< 0.001
Average Specific Humidity (SH) from Surface to LCL ($g \cdot kg^{-1}$)	15.59 ± 1.68	14.92 ± 1.47	< 0.05

Average Relative Humidity (RH) from Surface to LCL (%)	69.8 ± 11.5	69.1 ± 12.4	> 0.05
Average Mid-tropospheric SH from 700 hPa to 500 hPa ($g \cdot kg^{-1}$)	6.09 ± 1.05	5.10 ± 1.12	< 0.001
Average Mid-tropospheric RH from 700 hPa to 500 hPa (%)	73.2 ± 13.1	62.0 ± 13.1	< 0.001

Table 7: The contrast (mean, standard deviation, and significance of difference) between no-lightning and frequent-lightning hours in variables derived from the INTERPSONDE data product.

According to the table, the LCL Height and vertically-integrated RH from the surface to LCL do not vary between no-lightning and frequent-lightning hours. These variables affect warm cloud depth (Medina et al., 2022). Differences in the height of the 0 °C freezing level (0.14 km), mean SH from surface to LCL (0.67 $g \cdot kg^{-1}$) and surface equivalent potential temperature (3.25 K) are relatively small but significant statistically. The mid-tropospheric SH and RH are much lower during hours with thunderstorm activity (0.99 $g \cdot kg^{-1}$, 11.2 %, respectively), which is consistent with the analysis of convective profiles in the Amazon by Wall et al. (2014). They speculated that the increased lapse rate of humidity associated with a dry mid-troposphere increased the lapse rate of equivalent potential temperature and increased the severe storm threat when abundant moisture was present in the lower troposphere. Finally, the minimum equivalent potential temperature in mid-troposphere is lower in frequent-lightning hours (4.46 K), which is consistent with Scala et al. (1990) who found that cells with a less pronounced equivalent potential temperature minimum are less likely to produce vigorous vertical transport than those developing in environments with a relatively strongly pronounced minimum. The low equivalent potential temperature region is considered as a source of cool dry air which feeds penetrating downdrafts helping to maintain an intense storm (Pickering et al., 1993).

4.3 IC flash fraction relationship with $\sqrt{CAPE}$

Holton (1973) found that CAPE plays an important role in determining maximum parcel updraft velocity, which is proportional to $\sqrt{CAPE}$ based on parcel theory. We have noticed a positive relationship between IC fraction and $\sqrt{CAPE}$, as shown in Figure 4. This analysis is based on 304 hours with convective clouds detected at ARM SGP site from the CLDTYPE product and plentiful flashes (Hourly Flash Count > Median Value of flashes during convective hours = 162.5, which doesn't have the same definition with "frequent-lightning hours" in section 4.2), to ensure the statistics are meaningful.

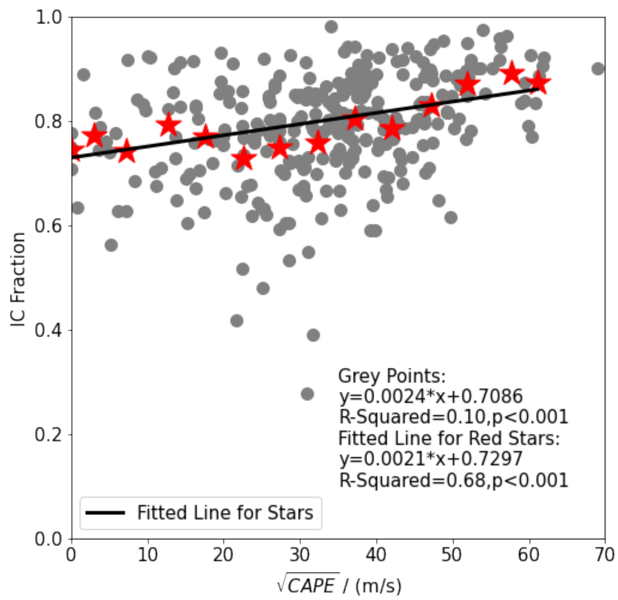

Figure 4: The Relationship between IC Fraction and $\sqrt{CAPE}$. The grey points show the IC fraction for convective hours with plentiful flashes while the red stars show the mean IC fraction for 5 m/s $\sqrt{CAPE}$ bins. The fitted line for the binned data and its equation are shown in the figure.

From Figure 5, as $\sqrt{CAPE}$ increases from 0 to 60 m/s, the IC fraction increases from 0.7 to about 0.9. A hypothesis for the relationship is that higher $\sqrt{CAPE}$ represents a stronger convective environment with stronger updrafts. The stronger updrafts bring the electrification zone further above the surface, resulting in fewer flashes reaching the surface and consequently a greater IC flash fraction. This hypothesis is supported by the fact that higher IC flash fractions are associated with higher IC heights, as shown in Figure 5. This relationship has not been widely discussed in previous studies, as they have focused on land-ocean contrast (Lapp and Saylor, 2007) or cloud vertical development (Williams et al., 1989). We tested that association between the $\sqrt{CAPE}$ and IC fraction using a Chi-Square calculator for a 2 × 2 contingency table and found that the relationship was significant at $p < 0.001$.

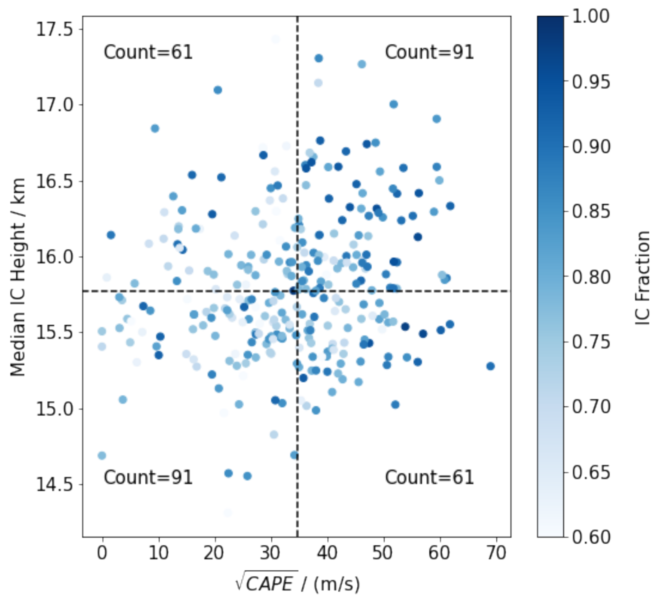

Figure 5: IC Height and Fraction with $\sqrt{CAPE}$. Median IC height is plotted against the $\sqrt{CAPE}$ for hours when flash rates exceed the median of flashing hours. The intensity of the colours shows the fraction of flashes that are IC. The dashed lines show the median value of $\sqrt{CAPE}$ and IC Height, respectively. The counts were used in the Chi-Square calculator and show the number of data points in each region of the figure.

Price and Rind (1993) found that the ratio of CG to IC lightning is related to the cold cloud thickness rather than the height of the freezing level. The cold cloud thickness method has been applied to models to estimate the production of nitrogen oxides by lightning (e.g., Price and Rind, 1994; Goldberg et al., 2022; Pérez-Invernón et al., 2023). The relationship found here between CG fraction and $\sqrt{\text{CAPE}}$ if verified with additional lightning data sets over a broader area would provide an alternative approach for parameterizing the ratio of CG to IC lightning in chemistry and climate models.

5 Conclusion

Previous ML-based studies of lightning frequency focus on larger regions, have coarser time resolution, or focus on CG lightning only. Here, we take advantage of rich measurements of atmospheric and cloud properties at the ARM SGP site and ENTLN flash counts to explore the factors affecting flash rates on an hourly time resolution using ML models. We limit the analysis to hours when convective clouds are detected at the SGP site and then nowcast the occurrence of lightning and examine the conditions under which lightning occurs. We begin by inputting eight mostly-independent meteorological variables into a Random Forest ML model to predict lightning occurrence. The ML model has an accuracy of more than 76% and AUC of 0.850, and the top, middle and least important variables sorting is significant according to the robust ranking distribution. The most important variables affecting lightning occurrence turn out to be cloud thickness, rain rate and CAPE.

In strong convective environments (CAPE > 2000 $J \cdot kg^{-1}$), several variables including rain rate and cloud thickness vary significantly between no-lightning and frequent-lightning periods. In addition, our analysis indicates that values of mid-tropospheric humidity are typically lower during frequent-flashing hours with low values of mid-tropospheric humidity indicative of greater convective instability. Both the 0 °C freezing level height and the surface equivalent potential temperature have small but statistically significant differences between no-lightning and frequent-lightning hours. Minimum equivalent potential temperatures in the mid-troposphere are typically 4.46 K lower in frequent-lightning hours, suggesting that a source of cool dry air from penetrating downdrafts is helpful for maintaining intense storms.

A positive relationship is found between $\sqrt{CAPE}$ and IC fraction in convective hours with plentiful flashes, which can provide models with another alternative parameterization option of the ratio of CG to IC. It may be explained by the fact that higher $\sqrt{CAPE}$ represents a stronger convective environment, which can bring the electrification zone further above the surface, resulting in a greater IC flash fraction. This hypothesis is supported by the variation in median IC heights with $\sqrt{CAPE}$ although more analysis is needed to confirm the preliminary finding due to uncertainties in IC heights from ENTLN and the limited sample size. Lightning Mapping Array (LMA) data with more accurate flash heights could be used together with ENTLN flash type information to verify the positive relationship between $\sqrt{CAPE}$ and IC fraction.

As lightning processes are complicated, better time resolution is needed to better understand the mechanism. This study focuses on hourly time resolution. ML can provide a quick and efficient result when dealing with multiple variables, while subsequent analysis and discussion are essential to understand the physical meaning behind the result. This study only focuses on the region around the ARM SGP site, and we simply assumed that those measurements are representative of the entire 1° × 1° grid, which adds uncertainty because the scale of convection is typically smaller than this. Future analysis over other regions is desired to enrich the data volume, in order to train the ML model and get more reliable and robust results.

Data availability

All ARM SGP datasets can be found at the ARM archive (https://adc.arm.gov/discovery/#/results/site_code::sgp) for the AOSCCN1COL (https://doi.org/10.5439/1256093), AOSCCN2COLAAVG (https://doi.org/10.5439/1323894), ARSCLKAZR1KOLLIAS (https://doi.org/10.5439/1393437 and https://doi.org/10.5439/1228768), CLDTYPE (https://doi.org/10.5439/1349884), INTERPOLATEDSONDE (https://doi.org/10.5439/1095316), PBLHTMPL1SAWYERLI (https://doi.org/10.5439/1637942) and VDIS (https://doi.org/10.5439/1025315). ERA5 hourly data on pressure levels from 1940 to present (https://doi.org/10.24381/cds.bd0915c6), US EPA Air Quality Data (https://www.epa.gov/outdoor-air-quality-data/download-daily-data) and MERRA-2 tavg1_2d_aer_Nx (https://doi.org/10.5067/KLICLTZ8EM9D) are also publicly available.

Author contribution

SS, DA, ZL and KP designed the experiments and SS carried them out. JL provided the ENTLN data used in this research. SS prepared the manuscript with contributions from all co-authors.

Competing interests

Some authors are members of the editorial board of ACP. The peer-review process was guided by an independent editor, and the authors have also no other competing interests to declare.

Acknowledgements

This research was supported by NASA (Grant No. 80NSSC20K0131) and National Science Foundation (Grant No. AGS2126098). The authors thank Earth Networks, Inc. for providing ENTLN data. The authors also thank Melody Avery, Maureen Cribb, Pengguo Zhao and Mengyu Sun who provided comments on a draft version of this manuscript.

References

Level 3 and 4 Regridder and Subsetter Information: https://disc.gsfc.nasa.gov/information/documents?keywords=grid&title=Level%203%20and%204%20Regridder%20and%20Subsetter%20Information, last access: March 2023.

Altaratz, O., Koren, I., Yair, Y., and Price, C.: Lightning response to smoke from Amazonian fires, Geophysical Research Letters, 37, https://doi.org/10.1029/2010GL042679, 2010.

Bang, S. D. and Zipser, E. J.: Seeking reasons for the differences in size spectra of electrified storms over land and ocean, Journal of Geophysical Research: Atmospheres, 121, 9048-9068, https://doi.org/10.1002/2016JD025150, 2016.

Bartholomew, M. J.: Impact Disdrometers Instrument Handbook, DOE Office of Science Atmospheric Radiation Measurement (ARM) Program …, https://doi.org/10.2172/1251384, 2016.

Bell, T. L., Rosenfeld, D., and Kim, K. M.: Weekly cycle of lightning: Evidence of storm invigoration by pollution, Geophysical Research Letters, 36, https://doi.org/10.1029/2009GL040915, 2009.

Bell, T. L., Rosenfeld, D., Kim, K. M., Yoo, J. M., Lee, M. I., and Hahnenberger, M.: Midweek increase in US summer rain and storm heights suggests air pollution invigorates rainstorms, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD008623, 2008.

Carey, L. D. and Buffalo, K. M.: Environmental control of cloud-to-ground lightning polarity in severe storms, Monthly weather review, 135, 1327-1353, https://doi.org/10.1175/MWR3361.1, 2007.

Chaudhuri, S.: Convective energies in forecasting severe thunderstorms with one hidden layer neural net and variable learning rate back propagation algorithm, Asia-Pacific Journal of Atmospheric Sciences, 46, 173-183, https://doi.org/10.1007/s13143-010-0016-1, 2010.

Chaudhuri, S. and Middey, A.: A composite stability index for dichotomous forecast of thunderstorms, Theoretical and Applied Climatology, 110, 457-469, https://doi.org/10.1007/s00704-012-0640-z, 2012.

Chen, Q., Fan, J., Hagos, S., Gustafson Jr, W. I., and Berg, L. K.: Roles of wind shear at different vertical levels: Cloud system organization and properties, Journal of Geophysical Research: Atmospheres, 120, 6551-6574, https://doi.org/10.1002/2015JD023253, 2015.

Cooper, M. A. and Holle, R. L.: Reducing lightning injuries worldwide, Springer2019.

Doswell III, C. A., Brooks, H. E., and Maddox, R. A.: Flash flood forecasting: An ingredients-based methodology, Weather and forecasting, 11, 560-581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2, 1996.

Flynn, D., Shi, Y., Lim, K., and Riihimaki, L.: Cloud Type Classification (cldtype) Value-Added Product, DOE Office of Science Atmospheric Radiation Measurement (ARM) Program …, https://doi.org/10.2172/1377405, 2017.

Goldberg, D. L., Harkey, M., de Foy, B., Judd, L., Johnson, J., Yarwood, G., and Holloway, T.: Evaluating NO x emissions and their effect on O 3 production in Texas using TROPOMI NO 2 and HCHO, Atmospheric Chemistry and Physics, 22, 10875-10900, https://doi.org/10.5194/acp-22-10875-2022, 2022.

Green, D. M. and Swets, J. A.: Signal detection theory and psychophysics, Wiley New York1966.

He, J. and Loboda, T. V.: Modeling cloud-to-ground lightning probability in Alaskan tundra through the integration of Weather Research and Forecast (WRF) model and machine learning method, Environmental Research Letters, 15, 115009, https://doi.org/10.1088/1748-9326/abbc3b, 2020.

Heckman, S.: ENTLN status update, XV international conference on atmospheric electricity, Norman, Oklahoma, U.S.A., 15-20 June 2014, 15-20,

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., and Schepers, D.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

Holton, J. R.: An introduction to dynamic meteorology, American Journal of Physics, 41, 752-754, https://doi.org/10.1119/1.1987371, 1973.

Hu, J., Rosenfeld, D., Ryzhkov, A., Zrnic, D., Williams, E., Zhang, P., Snyder, J. C., Zhang, R., and Weitz, R.: Polarimetric radar convective cell tracking reveals large sensitivity of cloud precipitation and electrification properties to CCN, Journal of Geophysical Research: Atmospheres, 124, 12194-12205, https://doi.org/10.1029/2019JD030857, 2019.

Jensen, M., Giangrande, S., Fairless, T., and Zhou, A.: interpolatedsonde, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States). Atmospheric …, https://doi.org/10.5439/1095316, 1998.

La Fata, A., Amato, F., Bernardi, M., D'Andrea, M., Procopio, R., and Fiori, E.: Cloud-to-Ground lightning nowcasting using Machine Learning, 2021 35th International Conference on Lightning Protection (ICLP) and XVI International Symposium on Lightning Protection (SIPDA), 1-6, https://doi.org/10.1109/ICLPandSIPDA54065.2021.9627428,

Lal, D. M., Ghude, S. D., Singh, J., and Tiwari, S.: Relationship between size of cloud ice and lightning in the tropics, Advances in Meteorology, 2014, https://doi.org/10.1155/2014/471864, 2014.

Lapp, J. and Saylor, J.: Correlation between lightning types, Geophysical Research Letters, 34, https://doi.org/10.1029/2007GL029476, 2007.

Li, Z., Wang, Y., Guo, J., Zhao, C., Cribb, M. C., Dong, X., Fan, J., Gong, D., Huang, J., and Jiang, M.: East Asian study of tropospheric aerosols and their impact on regional clouds, precipitation, and climate (EAST-AIRCPC), Journal of Geophysical Research: Atmospheres, 124, 13026-13054, https://doi.org/10.1029/2019JD030758, 2019.

Lynn, B. H., Yair, Y., Price, C., Kelman, G., and Clark, A. J.: Predicting cloud-to-ground and intracloud lightning in weather forecast models, Weather and forecasting, 27, 1470-1488, https://doi.org/10.1175/WAF-D-11-00144.1, 2012.

Medina, B. L., Carey, L. D., Bitzer, P. M., Lang, T. J., and Deierling, W.: The Relation of Environmental Conditions with Charge Structure in Central Argentina Thunderstorms, Earth and Space Science, 9, e2021EA002193, https://doi.org/10.1029/2021EA002193, 2022.

Mostajabi, A., Finney, D. L., Rubinstein, M., and Rachidi, F.: Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques, Npj Climate and Atmospheric Science, 2, 41, https://doi.org/10.1038/s41612-019-0098-0, 2019.

Orville, R. E., Huffines, G., Nielsen-Gammon, J., Zhang, R., Ely, B., Steiger, S., Phillips, S., Allen, S., and Read, W.: Enhancement of cloud-to-ground lightning over Houston, Texas, Geophysical Research Letters, 28, 2597-2600, https://doi.org/10.1029/2001GL012990, 2001.

Pawar, S., Lal, D., and Murugavel, P.: Lightning characteristics over central India during Indian summer monsoon, Atmospheric research, 106, 44-49, https://doi.org/10.1016/j.atmosres.2011.11.007, 2012.

Pérez-Invernón, F. J., Gordillo-Vázquez, F. J., Huntrieser, H., and Jöckel, P.: Variation of lightning-ignited wildfire patterns under climate change, Nature communications, 14, 739, https://doi.org/10.1038/s41467-023-36500-5, 2023.

Pickering, K. E., Thompson, A. M., Tao, W. K., and Kucsera, T. L.: Upper tropospheric ozone production following mesoscale convection during STEP/EMEX, Journal of Geophysical Research: Atmospheres, 98, 8737-8749, https://doi.org/10.1029/93JD00875, 1993.

Politovich, M. K. and Cooper, W. A.: Variability of the supersaturation in cumulus clouds, Journal of the atmospheric sciences, 45, 1651-1664, https://doi.org/10.1175/1520-0469(1988)045<1651:VOTSIC>2.0.CO;2, 1988.

Price, C. and Federmesser, B.: Lightning-rainfall relationships in Mediterranean winter thunderstorms, Geophysical Research Letters, 33, https://doi.org/10.1029/2005GL024794, 2006.

Price, C. and Rind, D.: A simple lightning parameterization for calculating global lightning distributions, Journal of Geophysical Research: Atmospheres, 97, 9919-9933, https://doi.org/10.1029/92JD00719, 1992.

Price, C. and Rind, D.: What determines the cloud-to-ground lightning fraction in thunderstorms?, Geophysical Research Letters, 20, 463-466, https://doi.org/10.1029/93GL00226, 1993.

Price, C. and Rind, D.: Modeling global lightning distributions in a general circulation model, Monthly Weather Review, 122, 1930-1939, https://doi.org/10.1175/1520-0493(1994)122<1930:MGLDIA>2.0.CO;2, 1994.

Richardson, Y. P., Droegemeier, K. K., and Davies-Jones, R. P.: The influence of horizontal environmental variability on numerically simulated convective storms. Part I: Variations in vertical shear, Monthly weather review, 135, 3429-3455, https://doi.org/10.1175/MWR3463.1, 2007.

Romps, D. M., Charn, A. B., Holzworth, R. H., Lawrence, W. E., Molinari, J., and Vollaro, D.: CAPE times P explains lightning over land but not the land-ocean contrast, Geophysical Research Letters, 45, 12,623-612,630, https://doi.org/10.1029/2018GL080267, 2018.

Romps, D. M., Seeley, J. T., Vollaro, D., and Molinari, J.: Projected increase in lightning strikes in the United States due to global warming, science, 346, 851-854, https://doi.org/10.1126/science.1259100, 2014.

Rutledge, S. A., Williams, E. R., and Keenan, T. D.: The down under Doppler and electricity experiment (DUNDEE): Overview and preliminary results, Bulletin of the American Meteorological Society, 73, 3-16, https://doi.org/10.1175/1520-0477(1992)073<0003:TDUDAE>2.0.CO;2, 1992.

Scala, J. R., Garstang, M., Tao, W. k., Pickering, K. E., Thompson, A. M., Simpson, J., Kirchhoff, V. W., Browell, E. V., Sachse, G. W., and Torres, A. L.: Cloud draft structure and trace gas transport, Journal of Geophysical Research: Atmospheres, 95, 17015-17030, https://doi.org/10.1029/JD095iD10p17015, 1990.

Seo, E. K. and Liu, G.: Retrievals of cloud ice water path by combining ground cloud radar and satellite high-frequency microwave measurements near the ARM SGP site, Journal of Geophysical Research: Atmospheres, 110, https://doi.org/10.1029/2004JD005727, 2005.

Sherwood, S. C., Phillips, V. T., and Wettlaufer, J.: Small ice crystals and the climatology of lightning, Geophysical Research Letters, 33, https://doi.org/10.1029/2005GL025242, 2006.

Stolz, D. C., Rutledge, S. A., Pierce, J. R., and van den Heever, S. C.: A global lightning parameterization based on statistical relationships among environmental factors, aerosols, and convective clouds in the TRMM climatology, Journal of Geophysical Research: Atmospheres, 122, 7461-7492, https://doi.org/10.1002/2016JD026220, 2017.

Sun, C. and Liang, X.-Z.: Improving US extreme precipitation simulation: dependence on cumulus parameterization and underlying mechanism, Climate Dynamics, 55, 1325-1352, https://doi.org/10.1007/s00382-020-05328-w, 2020.

Sun, M., Liu, D., Qie, X., Mansell, E. R., Yair, Y., Fierro, A. O., Yuan, S., Chen, Z., and Wang, D.: Aerosol effects on electrification and lightning discharges in a multicell thunderstorm simulated by the

WRF-ELEC model, Atmospheric Chemistry and Physics, 21, 14141-14158, https://doi.org/10.5194/acp-21-14141-2021, 2021.

Takahashi, T.: Riming electrification as a charge generation mechanism in thunderstorms, Journal of Atmospheric Sciences, 35, 1536-1548, https://doi.org/10.1175/1520-0469(1978)035<1536:REAACG>2.0.CO;2, 1978.

Thornton, J. A., Virts, K. S., Holzworth, R. H., and Mitchell, T. P.: Lightning enhancement over major oceanic shipping lanes, Geophysical Research Letters, 44, 9102-9111, https://doi.org/10.1002/2017GL074982, 2017.

Tippett, M. K. and Koshak, W. J.: A baseline for the predictability of US cloud-to-ground lightning, Geophysical Research Letters, 45, 10,719-710,728, https://doi.org/10.1029/2018GL079750, 2018.

Wall, C., Zipser, E., and Liu, C.: An investigation of the aerosol indirect effect on convective intensity using satellite observations, Journal of the atmospheric sciences, 71, 430-447, https://doi.org/10.1175/JAS-D-13-0158.1, 2014.

Wang, Q., Li, Z., Guo, J., Zhao, C., and Cribb, M.: The climate impact of aerosols on the lightning flash rate: is it detectable from long-term measurements?, Atmospheric Chemistry and Physics, 18, 12797-12816, https://doi.org/10.5194/acp-18-12797-2018, 2018.

Wang, Y., Wan, Q., Meng, W., Liao, F., Tan, H., and Zhang, R.: Long-term impacts of aerosols on precipitation and lightning over the Pearl River Delta megacity area in China, Atmospheric Chemistry and Physics, 11, 12421-12436, https://doi.org/10.5194/acp-11-12421-2011, 2011.

Westcott, N. E.: Summertime cloud-to-ground lightning activity around major Midwestern urban areas, Journal of Applied Meteorology and Climatology, 34, 1633-1642, https://doi.org/10.1175/1520-0450-34.7.1633, 1995.

Williams, E., Rosenfeld, D., Madden, N., Gerlach, J., Gears, N., Atkinson, L., Dunnemann, N., Frostrom, G., Antonio, M., and Biazon, B.: Contrasting convective regimes over the Amazon: Implications for cloud electrification, Journal of Geophysical Research: Atmospheres, 107, LBA 50-51-LBA 50-19, https://doi.org/10.1029/2001JD000380, 2002.

Williams, E. R.: Meteorological aspects of thunderstorms, in: Handbook of atmospheric electrodynamics, CRC Press, 27-60, 2017.

Williams, E. R., Weber, M., and Orville, R.: The relationship between lightning type and convective state of thunderclouds, Journal of Geophysical Research: Atmospheres, 94, 13213-13220, https://doi.org/10.1029/JD094iD11p13213, 1989.

Yoshida, S., Morimoto, T., Ushio, T., and Kawasaki, Z.: A fifth-power relationship for lightning activity from Tropical Rainfall Measuring Mission satellite observations, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2008JD010370, 2009.

Yuan, T., Remer, L. A., Pickering, K. E., and Yu, H.: Observational evidence of aerosol enhancement of lightning activity and convective invigoration, Geophysical Research Letters, 38, https://doi.org/10.1029/2010GL046052, 2011.

Zhang, Y., Seidel, D. J., and Zhang, S.: Trends in planetary boundary layer height over Europe, Journal of climate, 26, 10071-10076, https://doi.org/10.1175/JCLI-D-13-00108.1, 2013.

Zhao, P., Li, Z., Xiao, H., Wu, F., Zheng, Y., Cribb, M. C., Jin, X., and Zhou, Y.: Distinct aerosol effects on cloud-to-ground lightning in the plateau and basin regions of Sichuan, Southwest China, Atmospheric Chemistry and Physics, 20, 13379-13397, https://doi.org/10.5194/acp-20-13379-2020, 2020.

Zhu, Y., Stock, M., Lapierre, J., and DiGangi, E.: Upgrades of the Earth networks total lightning network in 2021, Remote sensing, 14, 2209, https://doi.org/10.3390/rs14092209, 2022.