Received: 8 January 2022 | Revised: 28 November 2022 | Accepted: 10 January 2023 | Published on: 31 January 2023

DOI: 10.1002/joc.8005

RESEARCH ARTICLE

International Journal of Climatology RMetS

A machine learning approach for prediction of seasonal lightning density in different lightning regions of India

Chandrani Chatterjee[1] | **Joyjit Mandal**[2] | **Saurabh Das**[1]

[1]Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology, Indore, Indore, India

[2]Department of Computer Science, Central University of Rajasthan, Ajmer, India

Correspondence
Chandrani Chatterjee, Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology, Indore, Madhya Pradesh 453552, India.
Email: chandrani.chatterjee9@gmail.com

Funding information
Science and Engineering Research Board, Grant/Award Number: MTR/2019/001581

Abstract

Lightning is one of the most severe weather events causing significant loss of human lives and resources. Increasing number of lightning fatalities due to recent climatic changes is emerging out to be a serious concern for India during last few years. Proper characterization and parameterization of the same, therefore, is extremely crucial. However, lightning is an extremely dynamic phenomenon having enormous spatio-temporal inhomogeneity especially over such a vast country like India with varied topographic and climatological features. Therefore, proper parameterization of lightning activity over India needs consideration of different lightning climatologies. This study has attempted to resolve the issue by regionalizing Indian subcontinent in different lightning climatologies based on lightning density and associated atmospheric variables that is, CAPE, specific humidity at different pressure levels, temperature, *k* index and cloud particle size and identified seven distinct lightning climatologies over India. A regression model is proposed for estimating the annual and seasonal (monsoon and pre-monsoon) lightning activities over the seven resulting lightning zones based on the said atmospheric variables using machine learning techniques. Four machine learning models have been tested among which Random forest has shown the best accuracy. The regression model has shown an *R*-squared score of 0.81 during monsoon season and 0.71 during the pre-monsoon. The atmospheric features based on their influences on the lightning activity in these seven climatologies has been ranked which presented the evidences of largely varied interplay between different atmospheric variables and lightning over different parts of the country and during different seasons.

KEYWORDS

k-means, lightning climatologies, lightning parameterization, machine learning based regression

1 | INTRODUCTION

Lightning is a threatening weather extreme that poses a serious challenge for human lives and resources. This can trigger wildfires, disrupt air traffic and cause serious fatalities by power surges. Accurate and timely prediction of such phenomena is crucial in real time applications of sectors like aviation (Price and Rind, 1994; De *et al*., 2005). Lightning activities can also act as a tool for the prediction and modelling of convective systems

(Saylor *et al*., 2005; Sun *et al*., 2019; Chatterjee and Das, 2020). Studies to characterize the lightning features and its relation to different atmospheric variables, however, still needs research attention (Price, 2008).

The prediction of short scale weather events like lightning, rainfall etc. have been attempted all over the globe since last few decades. The scientific basis and better feasibility of such predictions over tropics was explained by Charney and Shukla (1981). The work pioneered the seasonal prediction of Indian summer monsoon. The coupled ocean–atmosphere processes are extremely crucial in such predictions (Wang et al., 2005). Elsner and Widen, 2014 presented a model based on Bayesian formulation to predict the number of tornados over Central great plains during the months of April–June using the sea surface temperature (SST) data of February from the Gulf of Alaska and the western Caribbean Sea (WCA). Dowdy (2016) has studied the influence of large scale atmospheric variability and reported strongest correspondence between ENSO and lightning activities. Both the large and short scale climatic drivers were reported to have visible impact on lightning over north western Venezuela which happens to have the highest annual lightning rate (Munoz *et al*., 2016). Particularly over lightning hot-spots like India, the lightning activity has been reported to be very well predictable by slowly varying global predictors (Mallick *et al*., 2022).

The strong association between charge separation in the cloud and collisions of hydometeors has facilitated for numerical parameterization of lightning activities based on hydrometeor properties (Takahashi, 1978). Numerical weather prediction models to predict lightning activities provided a major breakthrough. Price and Rind (1994) proposed a formulation for estimating both total and cloud-to-ground lightning flash rates over land and ocean based on the convective cloud-top heights. Later, Boccippio (2002) proposed revised parameterizations, based on cloud top height to account for the underestimation in lightning flash density over sea by the previous model. A more elaborated parameterization was proposed by Grewe *et al*. (2001) based on cloud-top height and convective mass flux diagnosed by the ECHAM4 global circulation model. Magnitude of Convective Available Potential Energy (CAPE) helps in the updraft and vertical distribution of hydrometeors helps in the charge generation process in a thundercloud (Williams, 1989). Romps *et al*. (2018) parameterized the lightning trend over the United States based on precipitation and convective available potential energy. Aerosol content in a cloud system has been reported to have a crucial role in lightning severity (e.g. Williams and Stanfill, 2002; Mansell and Ziegler, 2013; Stolz *et al*., 2015). A simple parameterization of lightning activities were proposed by McCaul Jr. *et al*. (2009) depending only on the content of hydrometeors. A new parameterization, in the similar line has been proposed to express the lightning densities as a function of short scale variability like hydrometeors contents, CAPE, and cloud-base height (Lopez, 2016).

India receives significant amount of lightning strikes having the ITCZ (Inter-Tropical Convergence Zone) passing through the central part of the country (Tinmaker and Chate, 2013). Several studies have successfully attempted the forecast and parameterization of lightning activities over Indian region (Chaudhary *et al*., 2021; Rajeevan et al., 2012; Madhulata et al., 2013). Maximum surface air temperature seemed to correspond directly with flash rate density (Tinmaker and Chate, 2013). CAPE seemed to have a similar impact on lightning strikes over the north-eastern India and Indo-Gangetic basin (IGB) region (Saha *et al*., 2017). The correlation between upper tropospheric specific humidity and lightning strikes has been reported to be significant whereas, the boundary layer and surface specific humidity are also seemed to be closely linked with convection and lightning activity over IGB and central and eastern part of Indian subcontinent. Upper tropospheric humidity seems to have a strong association with lightning activity over north-eastern and north-western arc of the Himalayas as well (Saha *et al*., 2017). Mohan *et al*., 2021 evaluated the simulation of lightning flash counts based on different lightning parameterization schemes over Maharashtra, India. The offline diagnostic methods of lightning flash estimation using model generated storm parameters were also assessed. Price and Rind (1992) based on cloud top height and vertically integrated ice water path reported excellent accuracy in recreating the spatial lightning pattern. Vani *et al*., 2022 evaluated lightning parameterization based on cloud top height defined by reflectivity threshold for 16 pre-monsoon storms over Maharashtra. The study reported a false alarm ratio of 0.28, 0.25, 0.29, 0.26 from WSM6, Thompson, Morrison and WDM6 parameterizations respectively. Notable over-estimation in lightning flash was also reported with spatial and temporal shift.

Even though significant work have been carried out on parameterization of lightning activities, the complexity of the process and use of large number of data features makes machine learning, an excellent alternative to the conventional techniques. The large amount of data flow in atmospheric science, demands the use of data driven approaches. Various machine learning algorithms can make the computers learn skills from sets of recorded atmospheric data and to apply it on a new set of data. The forecaster have amalgamated atmospheric science with machine learning in order to improve the prediction of various weather phenomena at varied scales. It is

evident that, the human-computer mix that is, the assimilation of physical understanding of atmospheric processes and machine learning algorithms in place of human entered principals can be of great help in atmospheric science. Recently, the ML based techniques have been gaining high attention in NWP with several researchers describing their advantages over the conventional studies. Manzato (2013) has developed a forecast model based on neural network ensemble to predict the hail event over North-eastern Italy. The successful use of machine learning models to forecast the probability of hail storms and the radar-estimated size distribution of hail showed potential of ML in weather prediction (Gagne *et al*., 2020). Herman and Schumacher (2018) presented physical and statistical insight regarding regression and tree-based models for extreme rainfall prediction. Mostajabi *et al*. (2019) have reported potential of basic atmospheric datasets being used to find out the correlation patterns between lightning incidence and atmospheric data. A machine-learning-based model was proposed to nowcast the lightning over a specific region 30 min in advance, based on four meteorological parameters namely air pressure, air temperature 2 m above ground, relative humidity and wind speed.

However, successful parameterization of lightning activities over a large region, seeks serious consideration of spatial variation as the local topography and associated climatic factors differs largely over different regions. Besides, the relationships among different atmospheric parameters are not essentially unique, particularly if a large geographic area is considered. Therefore, proper identification of homogenous lightning regions is a vital step in this process. This is especially true for a country like India with such enormous topographic and climatic diversity (Williams and Stanfill, 2002; Tinmaker and Chate, 2013; Murugavel *et al*., 2014). The effects of recent climatic changes are evident on the global lightning activities. However, the susceptibility of different regions to it, will vary because of the inhomogeneity of the process (Collier *et al*., 2013).

This study aims to regionalize India in different lightning climatologies. It focuses on developing a regression model for estimating the annual and seasonal (monsoon and pre-monsoon) lightning activities in different lightning climatologies existing over India based on various atmospheric parameters, namely specific humidity at four different pressure levels, Convective Available Potential Energy (CAPE), air temperature (2 m above the surface), K-index, cloud ice particle size and cloud liquid particle size. The current study also aims to provide a ranking of different atmospheric features based on their influences on the lightning activity over different lightning climatologies during monsoon and pre-monsoon.

2 | DATA AND METHODOLOGY

2.1 | Data

The monthly statistics of the lightning strikes and atmospheric variables that is, specific humidity, Convective Available Potential Energy (CAPE), air temperature above 2 m from the surface, k-index, cloud ice particle size and cloud liquid particle size for the period of January 2003 to December 2013 have been used for this study.

Lightning data were obtained from lightning imaging sensor (LIS). It is a space-based instrument which is used to detect both cloud to ground and cloud to cloud lightning strikes. It also measures the amount, rate, and radiant energy of lightning during both day and night. The mission, therefore, facilitates for understanding the global lightning activities and thunderstorm climatology. The study used LIS/OTD Monthly Climatology Time Series (LRMTS) dataset with a resolution of 2.5°×2.5°(Cecil, 2006). Specific humidity at four different pressure levels (300 mbar, 500 mbar, 850 mbar, 1,000 mbar), CAPE, temperature (2 m above the surface) and *k*-index data were obtained from ERA5 monthly data product (Hersbach *et al*., 2019). Cloud particle size data has been procured from MODIS level 3 (Aqua/Terra) monthly dataset at a spatial resolution of 1°×1° (Webb *et al*., 2017).

2.2 | Methodology

2.2.1 | Pre-processing of data

The datasets for different variables considered are of different spatial resolutions. ERA5 reanalysis datasets have a resolution of 0.25°×0.25°, LRMTS lightning dataset is 2.5°×2.5° gridded data whereas; MODIS dataset has 1°×1° spatial resolution. All these datasets were interpolated to a resolution of 1.5°×1.5° using linear interpolation method for grid matching purpose. The consistency of original and re-sampled data was ascertained for example, Figures 1a,b and 2a,b show the distribution of one down-sampled and one up-sampled data. Even though up sampling has to be done in lightning measurements, it is evident that the distributions before and after the interpolation are consistent. Missing data imputation has been done by K-nearest neighbours (KNN) imputation method using scikit-learn (Pedregosa *et al*., 2011) method. This method is based on KNN approach which fills the missing values with the weighted average of five nearest neighbours. The nearest neighbour computation is done using nan-Euclidean distance metric.

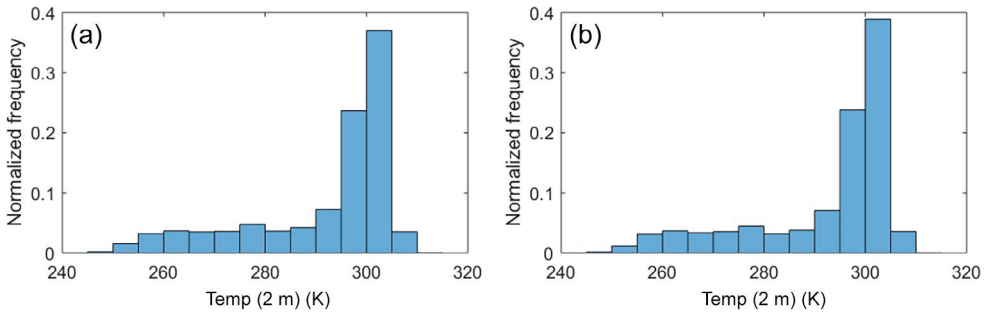

FIGURE 1 Distribution of (a) atmospheric temperature (K) at a height of 2 m above the ground (a) before and (b) after the interpolation [Colour figure can be viewed at wileyonlinelibrary.com]

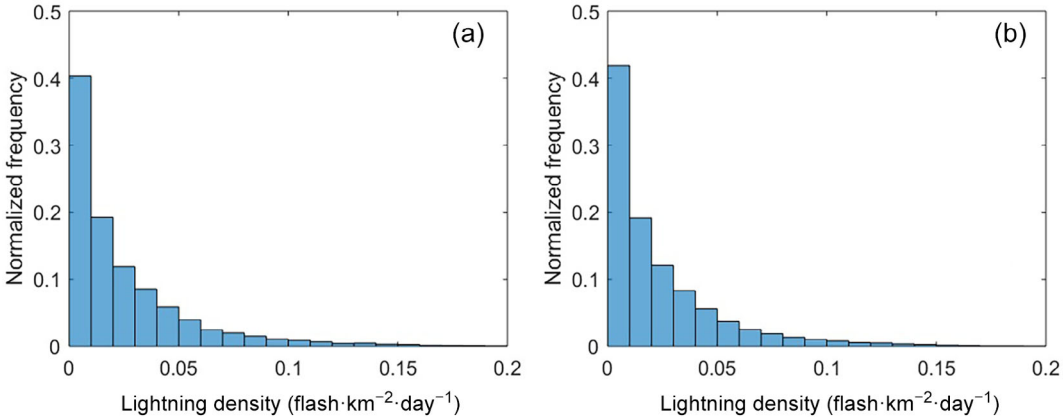

FIGURE 2 Distribution of lightning density (flash·km^{-2}·day^{-1}) (a) before and (b) after the interpolation [Colour figure can be viewed at wileyonlinelibrary.com]

2.2.2 | Identification of lightning climatologies

Parameterization of such a dynamic weather phenomenon like lightning needs proper identification of different lightning climatologies existing over different parts of the country. The entire dataset have been divided into two time periods that is, 2003–2011 and 2012–2013. The regionalization of lightning climatologies has been carried out using the first portion of data. The regionalization was performed based on lightning density along with all the atmospheric variable using K-means (Hartigan and Wong, 1979) clustering algorithm. Here, K-means (mentioned in details in Appendix S1) was attempted for its simplicity and efficient generalization capability to clusters of any shape and size which is important for handling natural data. The feature set was normalized before clustering to avoid computational disproportion.

2.2.3 | Regression models for estimation of lightning activities

Four machine learning based regression models, that is, Ridge regression (Hoerl and Kennard, 1970), Lasso regression (Santosa and Symes, 1986), Decision tree (Gladwin, 1989) and Random forest (Breiman, 2001) have been tested here for estimation of lightning activities (details of the models are mentioned in Appendix S1). As mentioned above, the entire dataset has been divided in two time periods that is, 2003–2011 and 2012–2013. Regression model has been developed using the first period of data as training set whereas; the second period of data has been used as test set. Total number of data points in the training and test set were 47,628 and 10,584 respectively. Here, all the above mentioned atmospheric variables along with the cluster and month information have been used as predictors. Each time the one hot encoding was used on categorical variables (month and cluster), the first column of both encoded categorical variables were dropped to avoid a dummy variable trap. One hot encoding was performed for month and cluster information. Data with degree 3 polynomial features have been provided as an input to the models.

The performances of the models were evaluated using *R*-squared scores as a metric. The hyper-parameter tuning has been done using ninefold cross-validation for each of the models where each fold contains a single year of data. The hyper-parameters providing the best mean score in this case were selected as the best set of hyper-parameters.

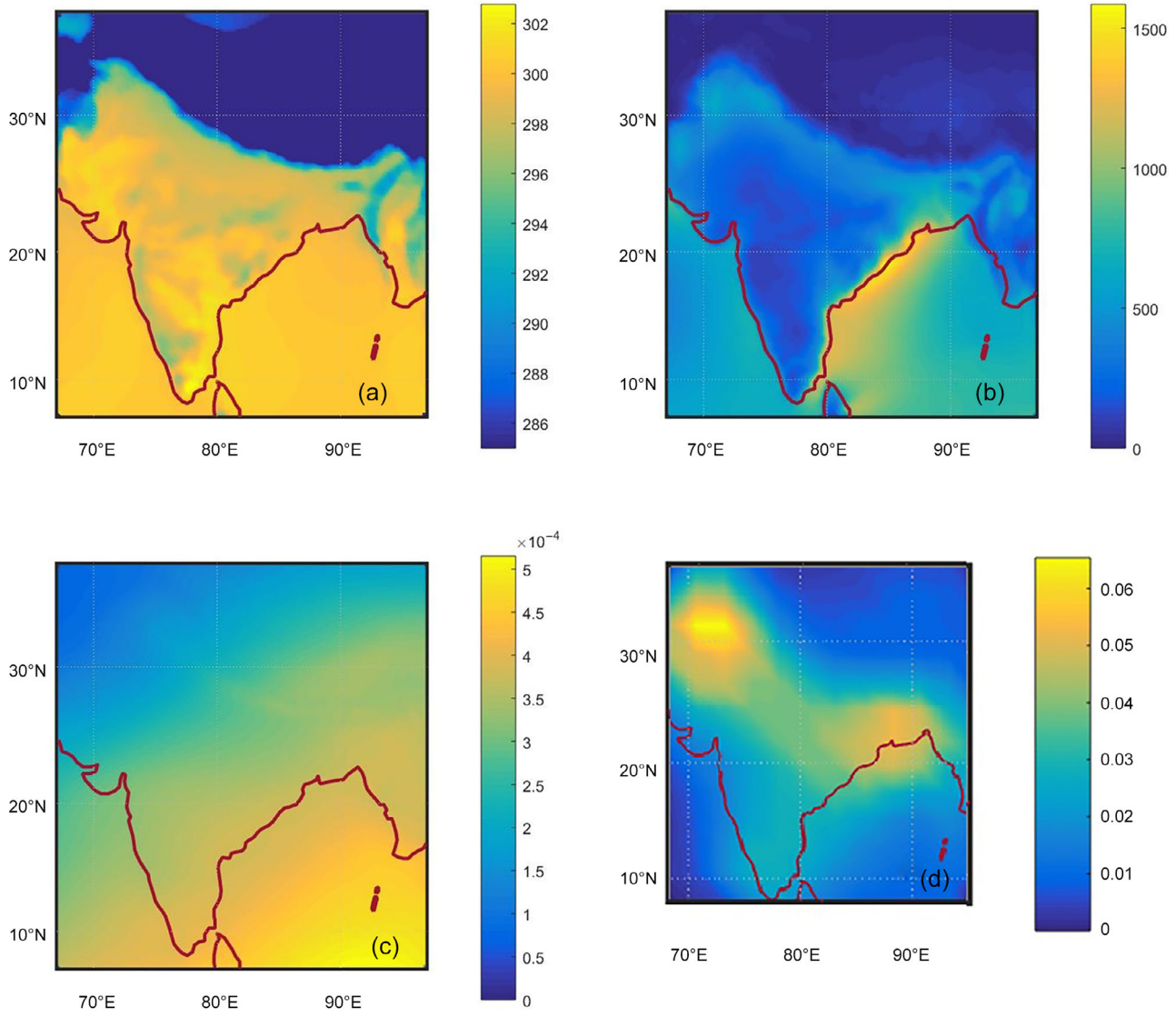

FIGURE 3 Spatial distribution of average (a) atmospheric temperature (K), (b) CAPE ($J \cdot kg^{-1}$), (c) specific humidity (at 300 hpa) ($kg \cdot kg^{-1}$) and (d) number of lightning flash ($flash \cdot km^{-2} \cdot day^{-1}$) over the Indian subcontinent during the year 2003–2014 [Colour figure can be viewed at wileyonlinelibrary.com]

2.2.4 | Feature ranking to understand the atmospheric influence on lightning activities

The predictive features that is, atmospheric variables have been ranked in order of their influence on lightning activities in each of the clusters separately during monsoon and pre-monsoon. Pearson correlation was used for this purpose to understand the association between each of the atmospheric parameter and lightning density over different clusters.

3 | RESULTS

The spatial distributions of some key predictors for lightning activities over the country were studied at first to

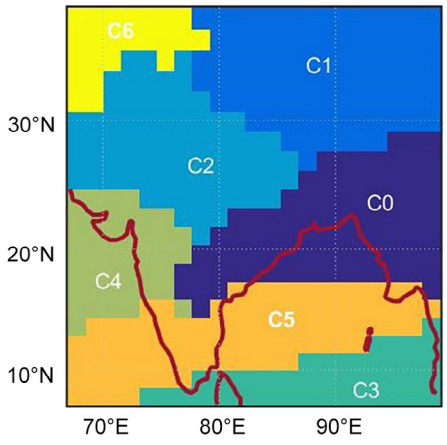

FIGURE 4 Clusters over Indian subcontinent based on lightning activities [Colour figure can be viewed at wileyonlinelibrary.com]

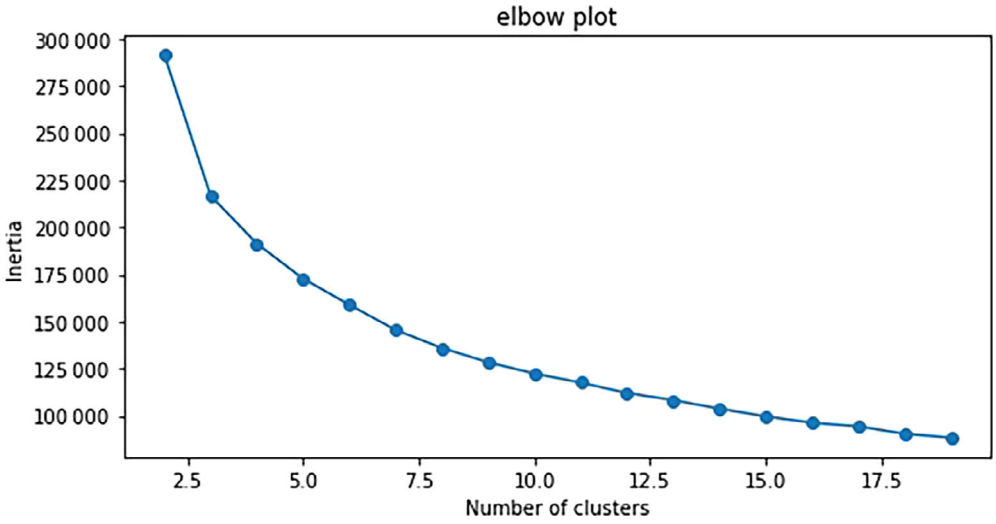

FIGURE 5 Elbow plot for the clustering process [Colour figure can be viewed at wileyonlinelibrary.com]

have an idea about the general climatologies existing over different parts of the country.

The average atmospheric temperature seemed to be low over the Himalayan region with slightly higher values over the Western Ghats. These parts are notably distinct in temperature from the rest of the Indian subcontinent (Figure 3a). The southern and western India seemed to be hotter than the northern and specially the eastern regions. High values of CAPE were noticed in the north-eastern and coastal part of the country whereas; the western part seems to have a little lower value (Figure 3b). Murugavel *et al.* (2014) also reported similar distribution. The specific humidity over Indian subcontinent showed a continuous increase from the north-western parts towards the south-eastern portions (Figure 3c). Highest lightning activities were observed in the Himalayan and north-eastern regions (Figure 3d) which finds good agreement with previous studies (Unnikrishnan et al. 2021). The western parts seemed to have a lesser lightning density than the rest of the subcontinent.

The above spatial inhomogeneity observed in lightning strikes and the associated atmospheric catalysts implicates the requirement of identifying different lightning climatologies over the country.

The clustering process has divided India in seven different clusters (Figure 4). The neighbouring pixels indicated in similar colour identify a single cluster. It is evident that all the clusters are distinctly separable. The clusters are represented as C0–C6 for easy interpretability. The number and consistency of the clusters have been verified with Elbow method (Figure 5).

The distributions of the features have been studied next to confirm the separability of the atmospheric variables in different clusters (Figure 6). The distribution of the variables in all the seven clusters showed visible differences which finds good agreement with the consistency of the clustering process indicated by Elbow method.

For example, distribution of specific humidity at 850 mbar is unimodal for clusters 1 and 6 where cluster 6 showed much larger values. On the other hand, for cluster 2 the distribution is bimodal. Distributions for clusters 3 and 5 are left skewed and unimodal with dominance of larger values in cluster 5.

Table 1 depicts the performance of each fold (year) in the ninefold cross validation for each four algorithms used. Random forest showed the best mean score although in some of the folds Ridge or Lasso regression showed better results. Decision Tree, however, seemed to under-perform in this case. The *R*-squared scores for individual ML models are presented for the test set in Table 2.

Random forest regression showed the best *R*-squared score on the test set that is, 0.81, whereas; decision tree showed the lowest *R*-squared score of 0.71. Therefore, estimation of lightning density was carried out using the best performing algorithm that is, Random forest. The model uncertainty has been investigated within 95% confidence level to have an idea of the stability in prediction. The uncertainty interval on *R*-squared score was found to be 0.80 ± 0.01.

The prediction model has been tested with the monthly lightning data of 2012–2013. Figures 7 and 8a–g present a comparison between actual and estimated lightning density in each of the clusters formed during the months of the two test years that is, 2012 and 2013 respectively. Good matching was observed in all the clusters except cluster 3 in 2012 (Figure 7a–g). In the year 2013, the actual and calculated lightning density showed good match in clusters 1 and 2 whereas; it was decent in clusters 5 and 6. The performance accuracy was medium in clusters 0 and 4.

The time series of the comparison between the actual and predicted values are studied further to investigate whether there is any seasonal bias in the model performance (Figure 9a–g).

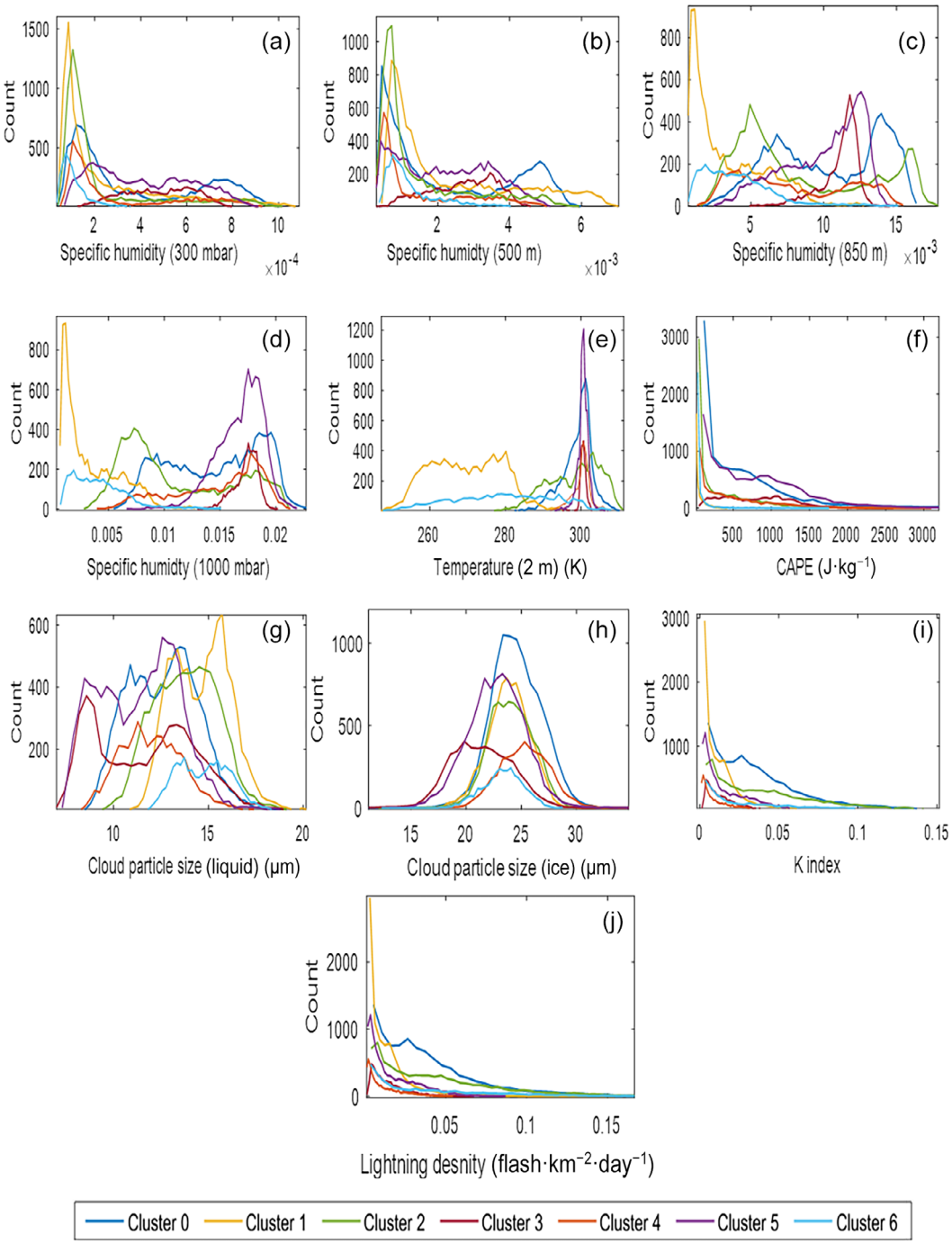

FIGURE 6 Distribution of various atmospheric features in different clusters [Colour figure can be viewed at wileyonlinelibrary.com]

The model estimated the lightning activities significantly well in cluster 0 throughout the time range except April–June in the year 2013 (Figure 9a). The model under-performed during the pre-monsoon months over the regions of cluster 1 (Figure 9b) whereas; significant good match between estimated and actual values were observed in clusters 2, 3, 5 and 6 (Figure 9c,d,f,g) throughout the time range. Performance of the model was decent in cluster 4 except little deviation during pre-monsoon (Figure 9e). This seasonal impact on model performance pointed towards the need of studying the model performance separately during the two most lightning prone seasons in India that is, pre-monsoon and monsoon. The spatial distribution of the estimated and actual lightning values retrieved from LIS have been compared to have a look at the performance variability of the model at annual and seasonal (monsoon and pre-monsoon) scale. Figure 10a shows the estimated annual lightning density. It is evident from the LIS retrieved (Figure 10b) values that spatial pattern of the lightning

TABLE 1 Hyper-parameters and performance of the regression models

ML algorithm	Best hyper-parameter	*R*-squared score of each fold (year)									Mean score	Standard deviation
		2003	2004	2005	2006	2007	2008	2009	2010	2011		
Ridge	Alpha: 10	0.85	0.83	0.81	0.85	0.83	0.83	0.80	0.73	0.83	0.82	0.03
Lasso	Alpha: 1 e-07	0.85	0.82	0.80	0.83	0.82	0.81	0.79	0.74	0.82	0.81	0.03
Decision tree	Max depth: 22 Min samples split: 55	0.77	0.70	0.68	0.72	0.74	0.75	0.71	0.68	0.78	0.72	0.03
Random forest	*n* estimators: 600 Max. Depth: 27 Min samples split: 5 Max. Samples: 0.9	0.84	0.84	0.85	0.87	0.84	0.84	0.85	0.73	0.85	0.83	0.04

TABLE 2 Performance of different regression models

Models	*R*-squared score
Ridge regression	0.79
Lasso regression	0.79
Decision tree	0.71
Random Forest	0.81

density were well captured by the model. However, there was some underestimation of values in very high lightning zones that is, in North-eastern and Northern Himalayan part.

The lightning density during monsoon is very high over Northern Himalayan region (Figure 11a). The model identified the spatial pattern including the very high lightning zone in northern Himalayan. The central portion of this region, however, showed little lower lightning density than in actual data (Figure 11b). Pre-monsoon also showed similar performance with little underestimation of values in North-eastern region (Figure 12a,b).

The model under-performed in regions of very high lightning density both at seasonal and annual scale (Figures 10–12) but the relative spatial pattern of lightning is captured successfully.

The mean and variability of actual and estimated lightning density in each of the clusters showed good matching (Table 3).

3.1 | Atmospheric variables and their degree of influence on lightning activities over different climatologies during monsoon and pre-monsoon season

Several researchers have revealed the dependence of lightning activities on atmospheric variable like CAPE, specific humidity and temperature (Williams, 1995; Saha *et al*., 2017). But, the degree of dependence has been reported to vary widely in different spatial scales (Dewan *et al*., 2017). Therefore, this study has attempted to rank the above mentioned atmospheric variables based on their influences on lightning activities in different lightning climatologies derived over the country based on Pearson correlation (Figure 13a–n). The Eastern coast of India belongs to cluster 0. The region (Figure 13a) reported strong correlation of CAPE and surface temperature with lightning activities during monsoon whereas; during pre-monsoon specific humidity at 850 mbar and cloud ice particle size were identified as major influential factors for lightning (Figure 13b). Murugavel *et al*., 2014 reported strong correlation between CAPE and lightning activities during monsoon. Lightning is not controlled alone by CAPE over Indian region during pre-monsoon whereas; it plays a major role in monsoon convection. During pre-monsoon other factor like moisture availability and orographic nature plays crucial role (Murugavel *et al*., 2014) which finds good agreement with the results. Cluster 1 (Figure 13c,d) covering the Tibetan plateau showed best correlation between specific humidity at 1,000 and 850 mbar and lightning strikes during both the season which supports the previously reported results over this region (Li *et al*., 2020). The presence of mountain chains is crucial in convection forming processes over this part than the influence of CAPE. The central and north-western region (cluster 2) also showed good association of CAPE and cloud particle size with lightning activities during monsoon region (Figure 13e). The pre-monsoon lightning is observed to be well correlated with k-index and specific humidity at 300 mbar (Figure 13f). During pre-monsoon these regions maintains a very high temperature and in spite of low CAPE value over central and northern India, lightning is observed to be high over these regions compared to coastal regions where CAPE is observed to be higher. So, CAPE is not the sole cause of lightning over these regions during pre-monsoon (Murugavel *et al*., 2014). Cluster 3 covers the lower part of Indian Ocean. Figure 13h showed good association between CAPE and lightning activities over this region during pre-monsoon.

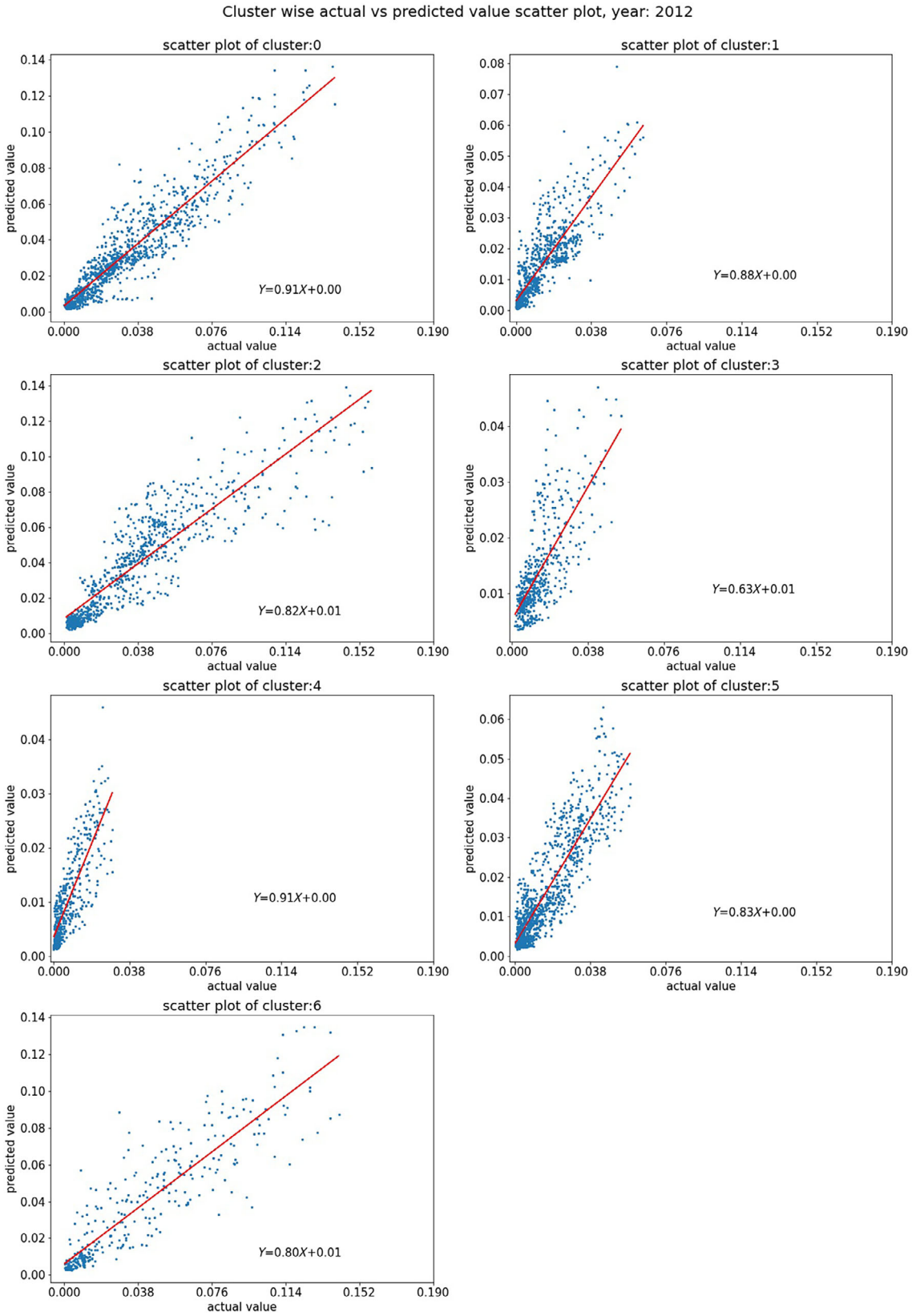

FIGURE 7 The comparison between actual and predicted lightning density (flash·km^{-2}·day^{-1}) in the seven clusters estimated by random forest model for the months in 2012 [Colour figure can be viewed at wileyonlinelibrary.com]

November to April is the convectively active season in Indian Ocean. Lightning in tropical storms is nearly muted over the ocean when CAPE is small. In the high-CAPE areas, on the other hand, oceanic storms can

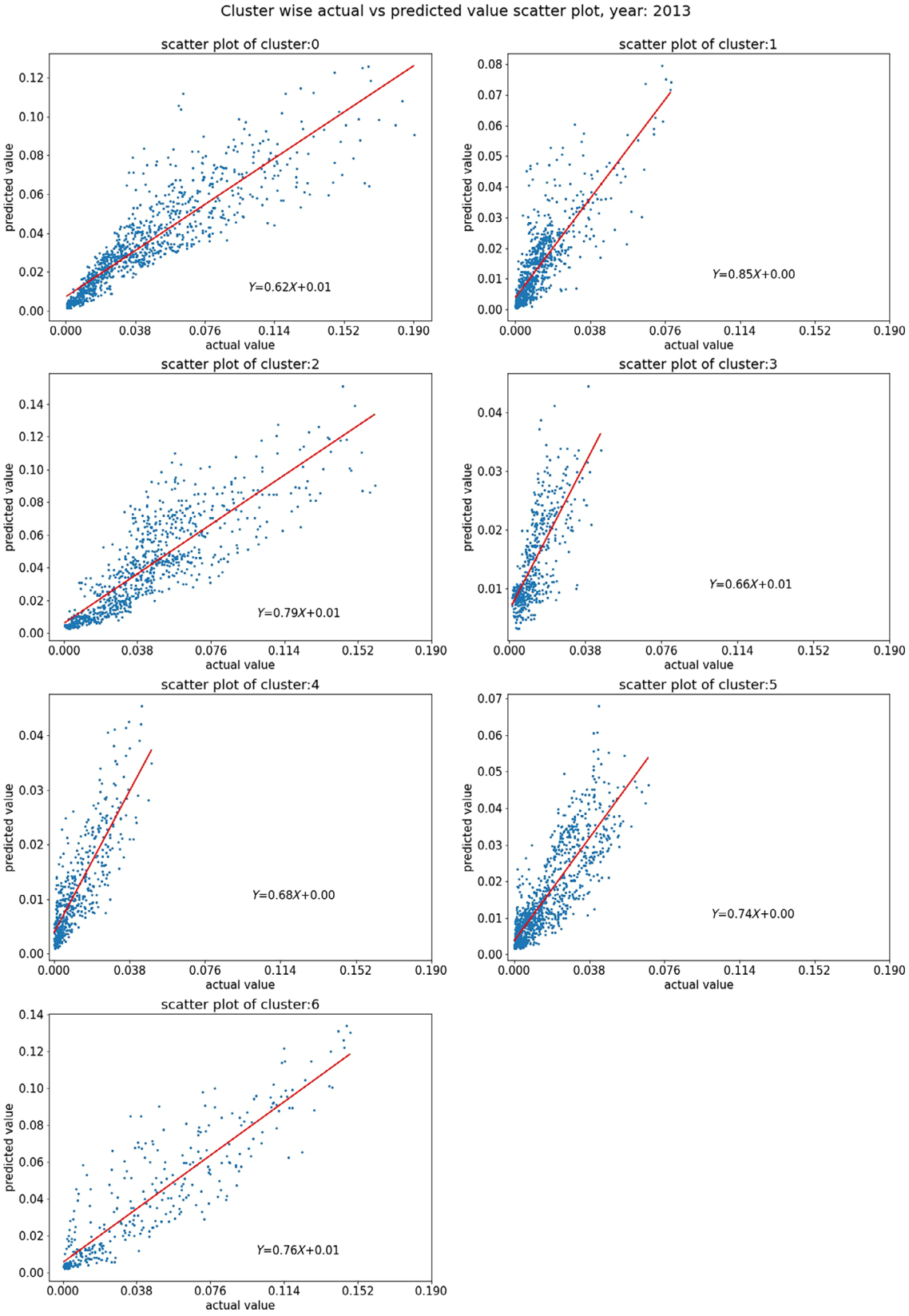

FIGURE 8 The comparison between actual and predicted lightning density (flash·km^{-2}·day^{-1}) in the seven clusters estimated by random forest model for the months in 2013 [Colour figure can be viewed at wileyonlinelibrary.com]

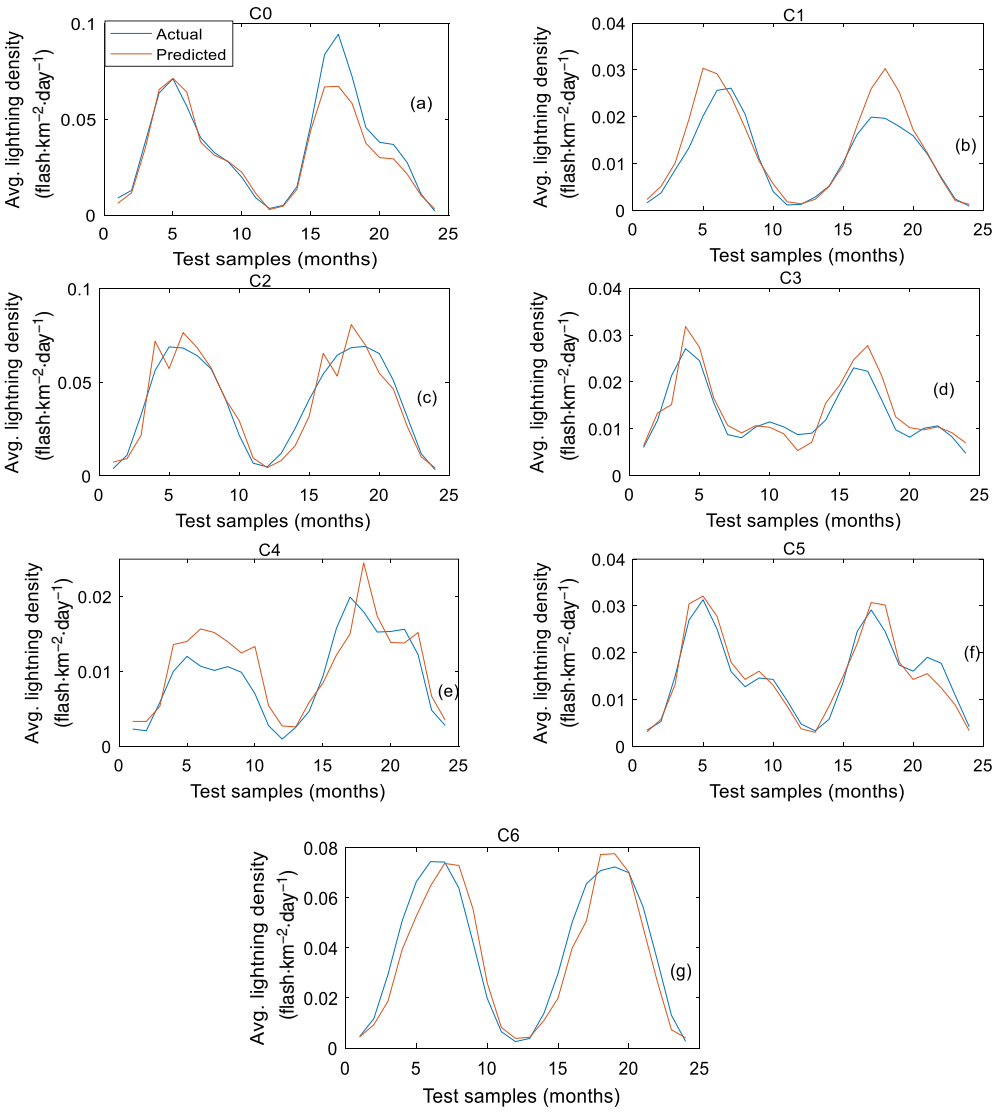

FIGURE 9 Time series (for consecutive months in 2012–2013) for the actual and predicted values of lightning strikes in the seven clusters by random forest model [Colour figure can be viewed at wileyonlinelibrary.com]

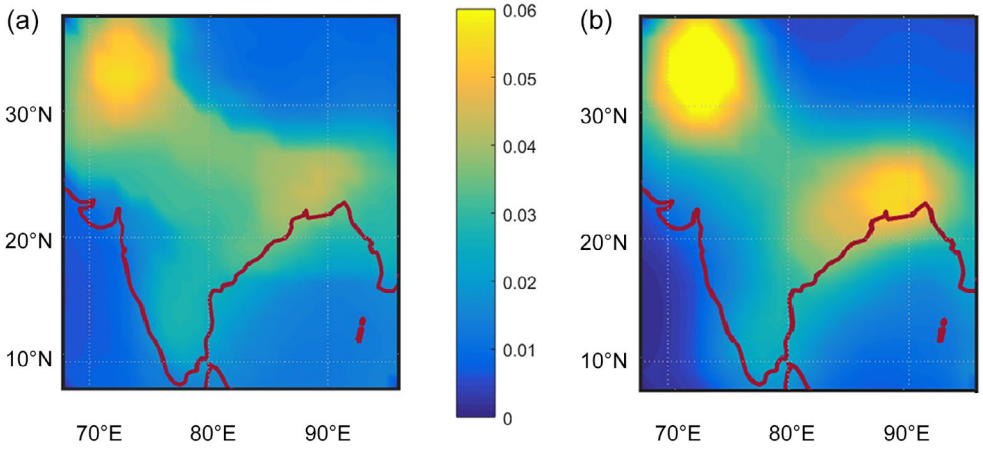

FIGURE 10 Spatial distribution of annual lightning density (flash·km^{-2}·day^{-1}) over Indian region (a) estimated and (b) LIS retrieved actual value [Colour figure can be viewed at wileyonlinelibrary.com]

result into as many lightning flashes as land storms (Cheng *et al.*, 2021). During, JJAS, lightning over this regions showed greater dependence on cloud particle size and specific humidity at 850 mbar (Figure 13g). The lightning activity over Arabian sea (cluster 4) seemed to be strongly correlated with the specific humidity at different pressure levels and thermal instability indicated by *k*-index during both the seasons (Figure 13i,j). This

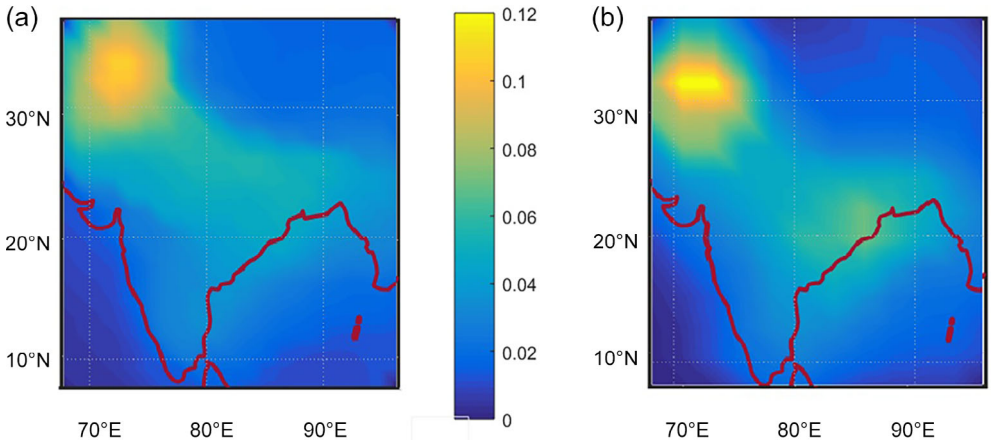

FIGURE 11 Spatial distribution of monsoon lightning density (flash·km^{-2}·day^{-1}) over Indian region (a) estimated and (b) LIS retrieved actual value [Colour figure can be viewed at wileyonlinelibrary.com]

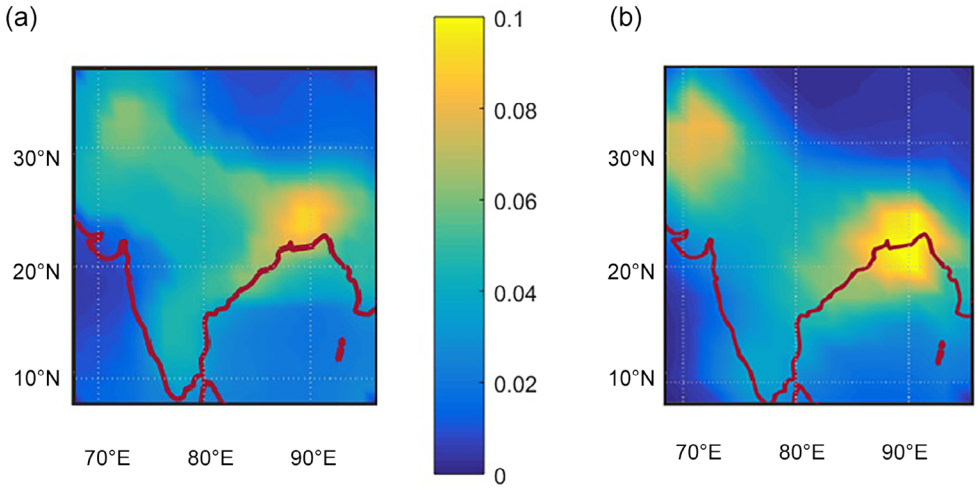

FIGURE 12 Spatial distribution of pre-monsoon lightning density (Flash/km^2/day) over Indian region (a) estimated and (b) LIS retrieved actual value [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Actual and predicted mean and variability of lightning density in different clusters

	Annual				Monsoon				Pre-monsoon			
	Actual		Estimated		Actual		Estimated		Actual		Estimated	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
C0	0.07	0.02	0.07	0.01	0.09	0.03	0.08	0.01	0.13	0.05	0.12	0.03
C1	0.02	0.01	0.03	0.01	0.04	0.02	0.04	0.01	0.03	0.03	0.04	0.02
C2	0.08	0.03	0.08	0.02	0.12	0.06	0.12	0.04	0.11	0.04	0.14	0.01
C3	0.03	0.01	0.03	0.01	0.02	0.01	0.03	0.01	0.05	0.02	0.05	0.01
C4	0.02	0.01	0.02	0.01	0.03	0.02	0.03	0.01	0.02	0.02	0.02	0.01
C5	0.03	0.02	0.03	0.01	0.04	0.02	0.04	0.02	0.05	0.03	0.05	0.02
C6	0.08	0.04	0.07	0.02	0.13	0.07	0.14	0.05	0.10	0.05	0.07	0.02

supports the previous reporting of lightning over this area being majorly controlled by temperature induced effects (Qie et al. 2020). Cluster 5 majorly covers southern Indian region along with some portion of Arabian sea and Bay of Bengal. It is well observed that moisture content is strongly correlated with lightning activity over these regions of southern India (Murugavel *et al.*, 2014). Figure 13k–l supports the previous findings during both the seasons. Some parts of Pakistan belong to cluster 6. This region showed good correlation between specific humidity at 850 and 1,000 mbar and lightning activities during pre-monsoon season

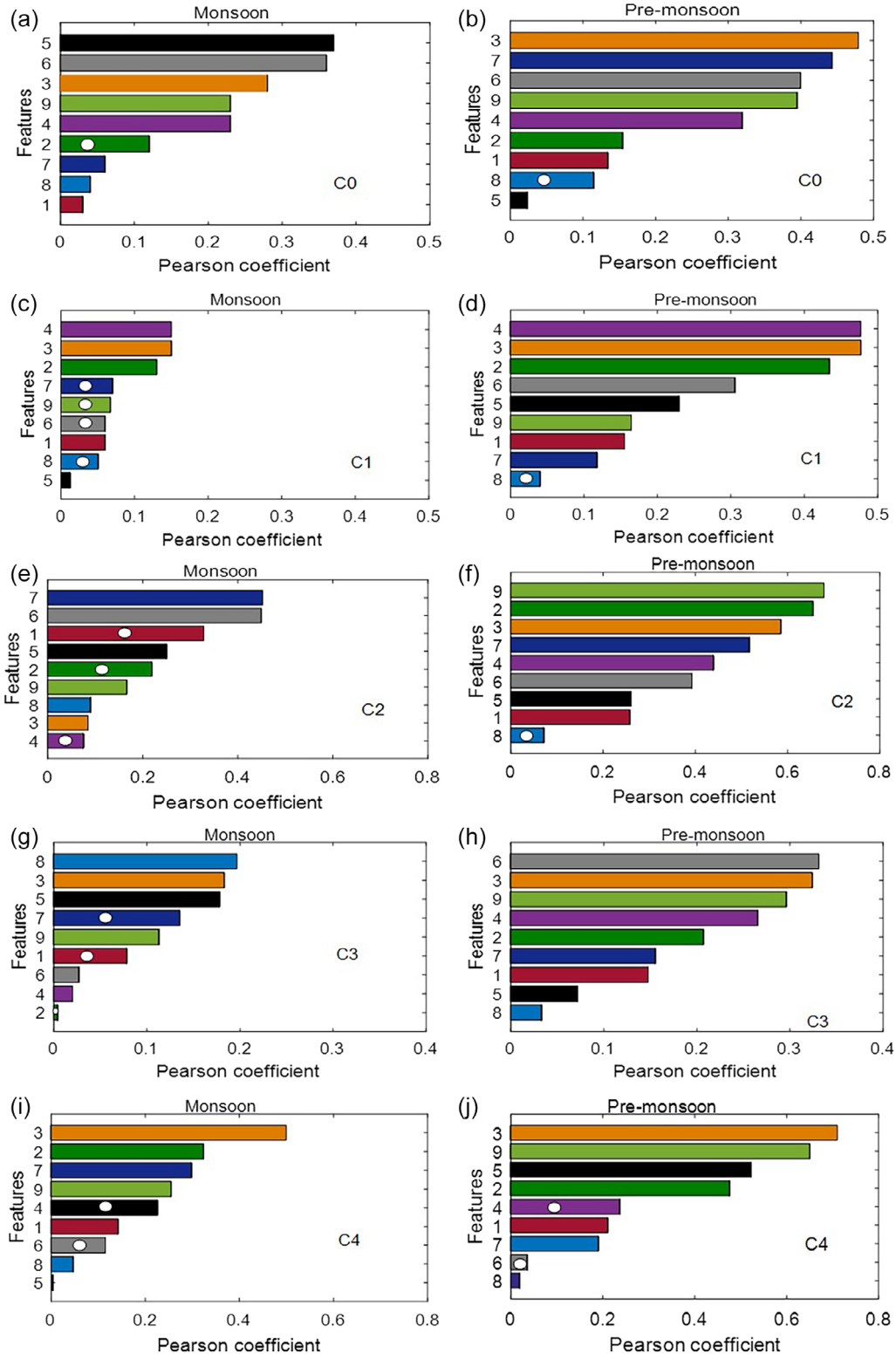

FIGURE 13 Rank of features for the seven clusters. Absolute value of Pearson coefficient used for ranking. The white filled dot represents negative correlation [Colour figure can be viewed at wileyonlinelibrary.com]

(Figure 13n). CAPE, however, did not show any strong influence during pre-monsoon probably because of the convective inhibition in this region (Ahmad *et al.*, 2019). On the contrary, both CAPE and humidity seemed to be well-correlated with lightning activity in this region during monsoon (Figure 13m). It is to be noted here,

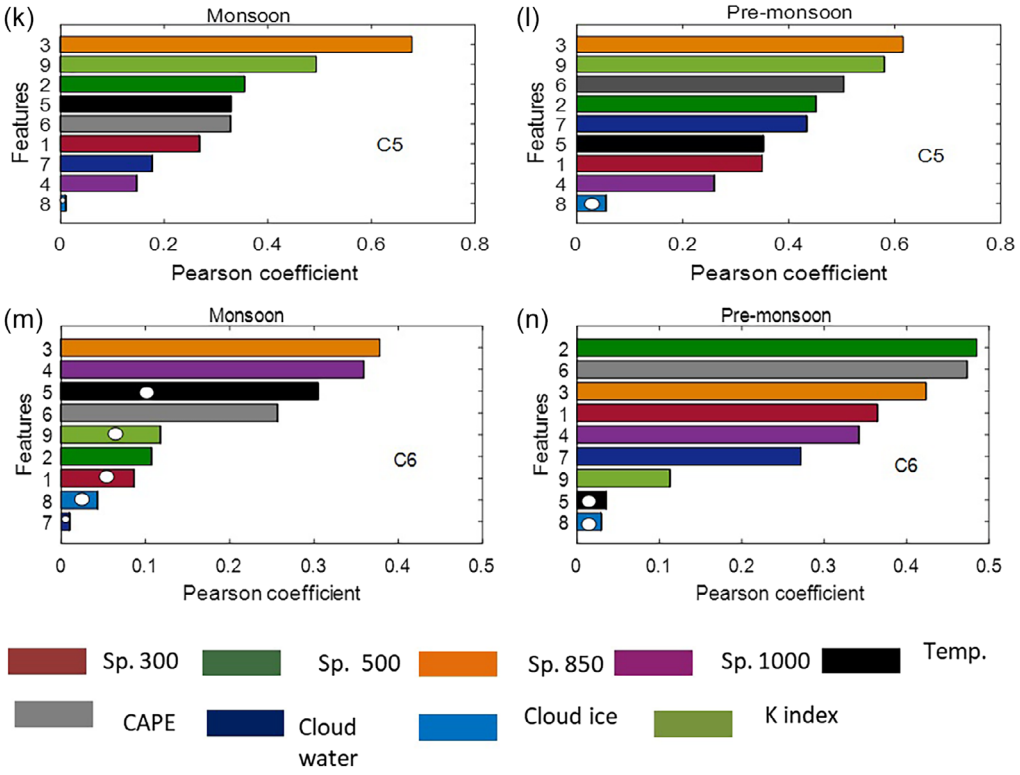

FIGURE 13 (Continued)

the absolute values have been presented here to emphasize on the degree of correlation for ranking and not on the nature. However, the information on the nature of the correlation has also been preserved in the figures.

4 | DISCUSSION

The current study is a novel approach presented for estimating lightning density using an ML based regression model for the entire Indian region. The model has used easily available satellite retrieved parameters. It groups the Indian region in seven lightning zones which information has been used further in the prediction model developed. The use of satellite data makes this approach flexible to be used over any geographic area of interest. However, on the other hand, it limits the time length of data. The study uses very simple machine learning regression models without any computational complexity. The accuracy achieved by the proposed model is decent. The model performed better during monsoon (R-squared score = 0.81) than in pre-monsoon (R-squared score = 0.71) probably because of its limitation in predicting very high values. The spatial comparison between predicted and actual lightning density shows some under-prediction in areas of very high lightning values whereas; it was good in areas of low to moderately high lightning activity. The study attempts to implement a simple lightning prediction model with a special focus on Indian region with available resources. The lightning data used here is obtained from LIS satellite at a spatial resolution of $2.5^\circ \times 2.5^\circ$ which was interpolated to a $1.5^\circ \times 1.5^\circ$ gridding. Even though the said process can bring in some under-estimation, the originality of the approach can be useful in the field of lightning prediction. The future aim of this study is to implement the method using ground based lightning data with longer and finer measurements. This can also facilitate for district level forecast of lightning over India which is currently not achievable by the model. However, for highly accurate real-time forecast a hybrid approach combining dynamical and machine learning techniques is required. Such models have gained high popularity in recent past for their accuracy and interpretability at the same time, that is, performance without being confined within the black box of ML (Johari *et al.*, 2009; Silva *et al.*, 2022). Studies (Lynn and Yair, 2010) have reported good prospects of WRF coupled models for accurate real-time forecast of lightning activities for example, Vani *et al.*, 2022 proposed a model with a POD of 0.90 and FAR of 0.64. The proposed model attempts to device a single model for the first time over entire Indian region. Even though the current form is not a complete structure to be used in

real-time operational forecast of lightning, such ML based approaches, when used with WRF outputs can serve as an excellent hybrid technique towards the real-time lightning forecast goal. Besides, the use of ML provides the flexibility of transfer learning that is, the model can also be used with slightly changed input structure when needed. Therefore, using a sub-set of predictive features (available) in real time operational purpose should not destroy the originality of the approach. Accurate hybrid predictive modelling that is, combining ML with dynamic approaches can be crucial for a country loosing almost ~2,500–3,000 lives/year due to lightning (ncrb.gov.in).

5 | CONCLUSION

This study has proposed an ML based regression model to estimate lightning density over the entire Indian region. The country has been regionalized based on lightning and associated atmospheric variables and this information has been further utilized in the prediction to realize a single model for this estimating lightning over such a vast geographic area. Four simple ML regression models were tested among which random forest has shown the best performance. However the model shown better accuracy during monsoon (R-squared score = 0.81) than that in pre-monsoon (R-squared score = 0.71). The accuracy provided by the model was decent but it under-performed in case of very high values. The study attempted this problem with a special focus on Indian region with available resources of satellite data. Authors believe the simple ML based approach presented in this study can be very useful in successful prediction of lightning density especially from Indian point of view. Authors aim to implement this approach with deeper and more sophisticated ML models using ground lightning measurements with better temporal and spatial resolution in near future.

AUTHORS CONTRIBUTIONS

Chandrani Chatterjee: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing. Joyjit Mandal: data curation, investigation, methodology, software, visualization, writing. Saurabh Das: conceptualization, funding acquisition, methodology, resources, supervision, writing.

ACKNOWLEDGEMENTS

This work was partially funded by SERB under the project grant no. MTR/2019/001581. The assistance is thankfully acknowledged. The authors also acknowledge European Centre for Medium-Range Weather Forecasts and Land Information Systems and Moderate Resolution Imaging Spectro radiometer by NASA for providing access to the datasets used for this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Chandrani Chatterjee https://orcid.org/0000-0002-4180-1694
Saurabh Das https://orcid.org/0000-0003-4373-1631

REFERENCES

Ahmad, R., Latif, M., Adnan, S. and Abuzar, M. (2019) Thunderstorm frequency distribution and associated convective mechanisms over Pakistan. *Theoretical and Applied Climatology*, 137, 755–773. https://doi.org/10.1007/s00704-018-2619-x.

Boccippio, D.J. (2002) Lightning scaling relations revisited. *Journal of the Atmospheric Sciences*, 59(6), 1086–1104. https://doi.org/10.1175/1520-0469(2002)059<1086:LSRR>2.0.CO;2.

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cecil, D. (2006) *LIS/OTD 2.5 Degree Low Resolution Monthly Climatology Time Series (LRMTS) [2003-2013]*. Dataset available online from the. Huntsville, Alabama, USA: NASA Global Hydrology Resource Center DAAC. https://doi.org/10.5067/LIS/LIS-OTD/DATA309.

Charney, J. and Shukla, J. (1981) Predictability of monsoons. *Monsoon dynamics.*, 99, 109.

Chatterjee, C. and Das, S. (2020) On the association between lightning and precipitation microphysics. *Journal of Atmospheric and Solar-Terrestrial Physics*, 207, 105350. https://doi.org/10.1016/j.jastp.2020.105350.

Chaudhari, H., Pokhrel, S., Pawar, S., Konwar, M., Saha, S., Das, S., Deshpande, S., Ghude, S., Barth, M., Rao, S., Nanjundiah, R. and Rajeevan, M. (2021) Evaluating different lightning parameterization schemes to simulate lightning flash counts over Maharashtra. *India. Atmospheric Research*, 255, 105532. https://doi.org/10.1016/j.atmosres.2021.105532.

Cheng, W., Kim, D. and Holzworth, R. (2021) CAPE threshold for lightningover the tropical ocean. *Journal of Geophysical Research: Atmospheres*, 126, e2021JD035621. https://doi.org/10.1029/2021JD035621.

Choudhury, B., Goswami, B., Zahan, Y. and Rajesh, P. (2021) Seasonality in power law scaling of convective and stratiform rainfall with lightning intensity over Indian Monsoon regions. *Atmospheric Research*, 248, 105265. https://doi.org/10.1016/j.atmosres.2020.105265.

Collier, A., Bürgesser, R. and Ávila, E. (2013) Suitable regions for assessing long term trends in lightning activity. *Journal of Atmospheric and Solar-Terrestrial Physics*, 92, 100–104.

De, U., Dube, R. and Prakasa, G. (2005) Extreme weather events over India in the last 100 years. *The Journal of Indian Geophysical Union*, 9, 173–187.

Dewan, A., Hossain, M., Rahman, M., Yamane, Y. and Holle, R. (2017) Recent lightning-related fatalities and injuries in Bangladesh. *Weather, Climate, and Society*, 9, 575–589. https://doi.org/10.1175/WCAS-D-16-0128.1.

Dowdy, A. (2016) Seasonal forecasting of lightning and thunderstorm activity in tropical and temperate regions of the world. *Scientific Reports*, 6, 2087. https://doi.org/10.1038/srep20874.

Elsner, J. and Widen, H. (2014) Predicting spring tornado activity in the central Great Plains by 1 March. *Monthly Weather Review*, 142(1), 259–267.

Gagne, D., Christensen, H., Subramanian, A. and Monahan, A. (2020) Machine learning for stochastic parameterization: generative adversarial networks in the lorenz '96 model. *Journal of Advances in Modeling Earth Systems*, 12, e2019MS001896. https://doi.org/10.1029/2019MS001896.

Gladwin, C.H. (1989) *Ethnographic Decision Tree Modeling 19*. London: Sage.

Grewe, V., Brunner, D., Dameris, M., Grenfell, J., Hein, R., Shindell, D. and Staehelin, J. (2001) Origin and variability of upper tropospheric nitrogen oxides and ozone at northern mid-latitudes. *Atmospheric Environment*, 35(20), 3421–3433. https://doi.org/10.1016/S1352-2310(01)00134-0.

Hartigan, J. and Wong, M. (1979) Algorithm AS 136: a K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 1080–1108. https://doi.org/10.2307/2346830.

Herman, G. and Schumacher, R. (2018) "Dendrology" in numerical weather prediction: what random forests and logistic regression tell us about forecasting extreme precipitation. *Monthly Weather Review*, 146, 1785–1812. https://doi.org/10.1038/s41612-019-0098-0.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz, S., Nicolas, J., Peubey, J., Radu, C., Rozum, R., Schepers, I., Simmons, D., Soci, A., Dee, C. and Thépaut, D. (2019) ERA5 Monthly Averaged Data on Pressure Levels from 1979 to Present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.6860a573.

Hoerl, A. and Kennard, R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Johari, D., Rahman, T., Musirin, I. and Nurul, A. (2009) Hybrid meta-EP-ANN technique for lightning prediction under Malaysia environment. *Atmospheric Sciences*, 224, 229.

Li, J., Wu, X., Yang, J., Jiang, R., Yuan, T., Lu, J. and Sun, M. (2020) Lightning activity and its association with surface thermodynamics over the Tibetan Plateau. *Atmospheric Research*, 245, 105118.

Lopez, P. (2016) A lightning parameterization for the ECMWF integrated forecasting system. *Monthly Weather Review*, 144(9), 3057–3075. https://doi.org/10.1175/MWR-D-16-0026.1.

Lynn, B. and Yair, Y. (2010) Prediction of lightning flash density with the WRF model. *Advances in Geosciences*, 23, 11–16. https://doi.org/10.5194/adgeo-23-11-2010.

Madhulatha, A., Rajeevan, M., Venkat Ratnam, M., Bhate, J. and Naidu, C.V. (2013) Nowcasting severe convective activity over southeast India using ground-based microwave radiometer observations. *Journal of Geophysical Research: Atmospheres*, 118, 1–13. https://doi.org/10.1029/2012JD018174.

Mallick, C., Hazra, A., Saha, S., Chaudhari, H., Pokhrel, S., Konwar, M., Dutta, U., Mohan, G. and Vani, K. (2022) Seasonal predictability of lightning over the global hotspot regions. *Geophysical Research Letters*, 49, e2021GL096489. https://doi.org/10.1029/2021GL096489.

Mansell, E. and Ziegler, C. (2013) Aerosol effects on simulated storm electrification and precipitation in a two-moment bulk microphysics model. *Journal of the Atmospheric Sciences*, 70, 2032–2050. https://doi.org/10.1175/JAS-D-12-0264.1.

Manzato, A. (2013) Hail in Northeast Italy: a neural network ensemble forecast using sounding-derived indices. *Weather and Forecasting*, 28, 3–28.

McCaul, E., Jr., Goodman, S., LaCasse, K. and Cecil, D. (2009) Forecasting lightning threat using cloud-resolving model simulations. *Weather Forecasting*, 24, 709–729. https://doi.org/10.1175/2008WAF2222152.1.

Mohan, G., Vani, K., Hazra, A., Mallick, C., Chaudhari, H., Pokhrel, S., Pawar, S., Konwar, M., Saha, S., Das, S., Deshpande, S., Ghude, S., Barth, M., Rao, S., Nanjundiah, R. and Rajeevan, M. (2021) Evaluating different lightning parameterization schemes to simulate lightning flash counts over Maharashtra, India. *Atmospheric Research*, 255(105532), 105532. https://doi.org/10.1016/j.atmosres.2021.105532.

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

Munoz, A., Daz-Lobat´ ´, O., Chourio, J. and Stock, M. (2016) Seasonal prediction of lightning activity in north western Venezuela: large-scale versus local drivers. *Atmospheric Research*, 172-173, 147–162. https://doi.org/10.1016/j.atmosres.2015.12.018.

Murugavel, P., Pawar, S. and Gopalakrishnan, V. (2014) Climatology of lightning over Indian region and its relationship with convective available potential energy. *International Journal of Climatology*, 34, 3179-3187.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

Price, C. (2008) Lightning sensors for observing, tracking and nowcasting severe weather. *Sensors*, 8, 157–170.

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

Price, C. and Rind, D. (1994) 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.

Qie, K., Tian, W., Wang, W., Wu, X., Yuan, T., Tian, H., Luo, J., Zhang, R. and Wang, T. (2020) Regional trends of lightning activity in the tropics and subtropics. *Atmospheric Research*, 242, 104960. https://doi.org/10.1016/j.atmosres.2020.104960.

Rajeevan, M., Madhulatha, A. and Rajasekhar, M. (2012) Development of a perfect prognosis probabilistic model for prediction of lightning over south-east India. *Journal of Earth System Science*, 121, 355–371. https://doi.org/10.1007/s12040-012-0173-y.

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

Saha, U., Kamra, A., Galanaki, E., Maitra, A., Singh, R.P., Singh, A., Chakraborty, S. and Singh, R. (2017) Onthe association of lightning activity and projected change in climate over the Indian sub-continent. *Atmospheric Research*, 183, 173–190. https://doi.org/10.1016/j.atmosres.2016.09.001.

Santosa, F. and Symes, W. (1986) Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4), 1307–1330. https://doi.org/10.1137/0907087.

Saylor, J.R., Ulbrich, C.W., Ballentine, J.W. and Lapp, J.L. (2005) The correlation between lightning and DSD parameters. *IEEE Transactions on Geoscience and Remote Sensing*, 43(8), 1806–1815.

Silva, Y., França, G., Ruivo, H. and Velho, H. (2022) Forecast of convective events via hybrid model: WRF and machine learning algorithms. *Applied Computing and Geosciences*, 16, 100099.

Stolz, D., Rutledge, S. and Pierce, J. (2015) Simultaneous influences of thermodynamics and aerosols on deep convection and lightning in the tropics. *Journal of Geophysical Research – Atmospheres*, 120, 6207–6231. https://doi.org/10.1002/2014JD023033.

Sun, E., Che, H., Xu, X., Wang, Z., Lu, C., Gui, K., Zhao, H., Zheng, Y., Wang, Y., Wang, H. and Sun, T. (2019) Variation in MERRA-2 aerosol optical depth over the Yangtze River Delta from 1980 to 2016. *Theoretical and Applied Climatology*, 136, 363–375. https://doi.org/10.1007/s00704-018-2490-9.

Takahashi, T. (1978) Riming electrification as a charge generation mechanism in thunderstorms. *Journal of Atmospheric Sciences*, 35(8), 1536–1548.

Tinmaker, I.R. and Chate, D.M. (2013) Lightning activity over India: a study of east–west contrast. *International Journal of Remote Sensing*, 34(16), 5641–5650. https://doi.org/10.1080/01431161.2013.794987.

Unnikrishnan, C.K., Pawar, S. and Gopalakrishnan, V. (2021) Satellite-observed lightning hotspots in India and lightning variability over tropical South India. *Advances in Space Research*, 68(4), 1690–1705.

Vani, K.G., Mohan, G.M., Hazra, A., Pawar, S.D., Pokhrel, S., Chaudhari, H.S., Konwar, M., Saha, S.K., Mallick, C., Das, S.K., Deshpande, S., Ghude, S.D., Domkawale, M., Rao, S.A., Nanjundiah, R.S. and Rajeevan, M. (2022) Evaluation and usefulness of lightning forecasts made with lightning parameterization schemes coupled with the WRF model. *Weather and Forecasting*, 37(5), 709–726.

Wang, B., Ding, Q., Fu, X., Kang, I.-S., Jin, K., Shukla, J. and Doblas-Reyes, F. (2005) Fundamental challenge in simulation and prediction of summer monsoon rainfall. *Geophysical Research Letters*, 32, L15711. https://doi.org/10.1029/2005GL022734.

Webb, M.J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R, Chepfer, H, Douville, H, Good, P., Kay, J.E, Klein, S.A, Marchand, R., Medeiros, B, Siebesma A.P, Skinner, C.B, Stevens, B, Tselioudis, G, Tsushima, M, and Watanabe, M. (2017) The cloud feedback model Intercomparison project (CFMIP) contribution to CMIP6 *Geoscientific Model Development* 101, 359–384. https://doi.org/10.5067/MODIS/MCD06COSP_M3_MODIS.061

Williams, E. and Stanfill, S. (2002) The physical origin of the land-ocean contrast in lightning activity. *Comptes Rendus-Physique*, 3, 1277–1292. https://doi.org/10.1016/S1631-0705(02)01407-X.

Williams, E.R. (1989) The tripolar structure of thunderstorms. *Journal of Geophysical Research*, 94(13), 13151–13167.

Williams, E.R. (1995) Meteorological aspects of thunderstorms. In: Volland, H. (Ed.) *CRC Handbook on Atmospheric Electrodynamics 1*. Boca Raton: CRC Press, pp. 27–60.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chatterjee, C., Mandal, J., & Das, S. (2023). A machine learning approach for prediction of seasonal lightning density in different lightning regions of India. *International Journal of Climatology*, *43*(6), 2862–2878. https://doi.org/10.1002/joc.8005