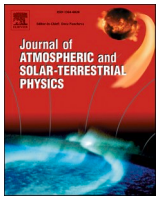
Journal of Atmospheric and Solar–Terrestrial Physics 259 (2024) 106255

Contents lists available at ScienceDirect

Journal of Atmospheric and Solar-Terrestrial Physics

journal homepage: www.elsevier.com/locate/jastp



Research Paper

An explainable machine learning technique to forecast lightning density over North-Eastern India

Joyjit Mandal [a], Chandrani Chatterjee [b,*], Saurabh Das [c]

[a] Department of Computer Science, Central University of Rajasthan, NH-8, Bandar Sindri, Dist-Ajmer-305817, Rajasthan, India
[b] Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India
[c] Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology, Simrol, Indore, Madhya Pradesh-453552, India

ARTICLE INFO

Handling Editor: Dora Pancheva

Keywords:
Lightning prediction
North eastern India
SHAP
Machine learning regression

ABSTRACT

Increasing lightning fatalities over India is a concerning subject. Especially, it is pretty crucial over North-Eastern part of the country where lightning is extremely frequent. Given the complex nature of the problem, machine learning can be an excellent option in such forecasting scenarios. However, such dynamic processes seek proper transparency of the model. The current work attempts to devise a model for short range prediction (one month ahead) of lightning density based on primary atmospheric parameters from satellite data with a lead time of one month over North –Eastern and Eastern part of the country. Random Forest regression seems to outperform other models explored, with a R^2 of 0.86 and an MAE of 0.0071. The interpretation of the model output using SHAP index reveals that 2 m temperature at previous two months and CAPE and K-index at previous month has a positive impact on the output of the model whereas, instantaneous surface heat flux of previous month and two month prior K-index has an inhibiting effect on model's output. The use of machine learning techniques for atmospheric predictions without the shed of the black box can be of importance to the scientific community. Such studies especially over lightning prone tropical regions can be crucial in meteorological forecasting applications.

1. Introduction

Increased lightning fatalities are one of the most serious concerns, India is currently dealing with. Timely prediction of lightning activities is therefore of crucial interest in today's scenario. Even though the numerical parameterization of lightning is quite a few decades old concept now, the history of lightning prediction over India is not very long. Moreover, given the enormous topographic diversity offered by India, it is difficult to forecast such a dynamic weather variable precisely over different parts of the nation.

Prediction of a thunderstorm is one of the most challenging task in weather prediction, because of the small spatial and temporal span and complexity. The process of lightning generation is believed to be extremely dynamic and complex in nature. The conventional way of predicting convective storm's location was to have an extrapolation of radar echoes (Wilson et al., 1998). Whenever, the prediction lead time becomes large, a lot of uncertainties come in. The cause behind such limitations lies in the genesis of lightning itself. The separation of positive and negative charges is reported to be the foundation for lightning activities (Deierling and Petersen, 2008). The deep convection uplifts the water vapour which later condenses and forms various forms of hydrometers with different sizes and shapes. The hydrometeors collide with each other in their subsequent upward motion which gives rise to charge on the particles (Reynolds et al., 1957; Takahashi, 1978). Even though the interactions between hydrometeors are believed to be the primary reason behind the initial charge build up, the association is not that straightforward as several other factors like liquid water amount and rate of riming can have their role in it. After this non-inductive charging, an inductive generation may take place due to the initial build-up of charge (Ziegler et al., 1991). It is to be noted that, the inductive mechanism is again governed by the complex interplay between processes like rebounding collision and presence of super cooled drops in cloud. However, many other processes can play significant role in charge generation on crystals during condensation, evaporation or melting of ice. Reynolds et al., 1957 suggested through laboratory experiment that graupel pellets acquire more charges if

* Corresponding author. Center for Soft Computing Research Indian Statistical Institute, 203 B.T. Road, Kolkata-700108, India.
E-mail addresses: joyjitmandal1@gmail.com (J. Mandal), chandrani.chatterjee9@gmail.com (C. Chatterjee), das.saurabh01@gmail.com, saurabh.das@iti.ac.in (S. Das).

https://doi.org/10.1016/j.jastp.2024.106255
Received 21 October 2023; Received in revised form 30 April 2024; Accepted 9 May 2024
Available online 10 May 2024

graupels grown by riming and collide with ice. Various researchers (e.g., Williams et al., 2002; Yuan et al., 2011; Mansel and Ziegler, 2013; Stolz et al., 2015) have reported that aerosol density may increase the probability of lightning activities as it can act as a cloud condensation nuclei capable of producing smaller cloud droplets. Even though Williams and Stanfill, 2002 have proposed this "aerosol hypothesis'' as the probable reason behind lower lightning activity over oceans and the Amazon region compared to central Africa, it requires more similar association over other geographic region to be accepted as an inherent factor governing lightning process.

The numerical parameterization of lightning could have been realized because of the connection between charge separation in the cloud and collisions of hydrometeors (Takahashi 1978). Convective cloud top heights were found to be a crucial precursor in the parameterization of lightning flash rates over land and ocean (Price and Rind (1994). A similar parameterization was proposed by Boccippio (2002) which takes care of the underestimation in lightning flash density over sea by the previous model. Grewe et al. (2001) devised a detailed lightning parameterization based on cloud-top height and convective mass flux procured from the ECHAM4 global circulation model. Wilson et al., 1998 presented significant correlation between Convective Available Potential Energy (CAPE) and updraft whereas, the vertical distribution of hydrometeors has been reported to be closely linked with the charge generation process in a convective system. Aerosol content in a cloud system is crucial in deciding the lightning severity (e.g.,Williams et al., 2002; Mansel and Ziegler, 2013; Stolz et al., 2015). The lightning activities in a thundercloud can be represented as a function of hydrometeors contents, CAPE, and cloud-base height (Lopez, 2016).

In Numerical weather prediction models, the uncertainty associated with the prediction increases as the time scale increases. Charney and Shukla (1981) reported the scientific feasibility of seasonal weather predictions over tropical regions. Authors established the scientific basis of Indian summer monsoon prediction one season ahead. The association between oceanic and atmospheric processes are of unavoidable interest in such predictions (Wang et al., 2006). Elsner and Widen (2014) proposed a parameterization using Bayesian formulation to forecast the number of tornados occurring between April–June over Central great plains based on sea surface temperature (SST) data of February from the Gulf of Alaska and the western Caribbean Sea (WCA). A neighbourhood-based equitable threat score (ETS) was proposed for precipitation prediction by Clark et al., (2010). The neighbourhood based approach seemed to relax the criteria for correct forecasts by considering grid points within a specified radius. The index proposed can be used to find out mismatch in precipitation forecast skills between different models as a function of spatial scale. To be specific, both short and long scale catalysts play crucial role in governing the process of seasonal lightning (Dowdy, 2016). Seasonal lightning activities were reported to be successfully predictable by slowly varying global predictors especially over lightning hot-spots like India (Mallick et al., 2002). In recent past, machine learning techniques are being extensively used for prediction of lightning activities. A neural network ensemble model was proposed to predict the hail event over North-eastern Italy (Manzato, 2013). The successful forecast of hail storms established the basis for use of ML in weather prediction (Gagne et al., 2020). The physical and statistical insight regarding regression and tree-based models for extreme rainfall prediction is also defined (Herman et al., 2018; Mostajabi, 2019 presented the feasibility of fundamental atmospheric datasets to be used in lightning prediction. Authors proposed a ML model to nowcast the lightning over a specific region up to 30 min in advance, based on four meteorological parameters namely air pressure, the air temperature 2 m above ground, relative humidity and wind speed.

It is, therefore, understandable that given the complexity of the process machine learning can be an excellent option to deal simultaneously with vast weather data and non-linearity of the process. Also, in case of monthly or seasonal prediction, numerical models may have compromised prediction skills, and ML approaches are meaningful, since they are able to exploit information hidden in the features unlike the numerical models. The uncertainties and spatial dynamics of the process makes it crucial to have sufficient transparency of the prediction process. Especially, over tropics, forecasting any dynamic atmospheric parameter always demands serious consideration of underlying intrinsic properties of the variations. Varotsos et al., (2013) reported notable mismatch between the measured and the modelled temperature anomalies in the tropics and analysed the possible factors behind it. The study reported vertical exaggeration of warming by the models. The detrended actual fluctuations was assumed to exhibit a white noise nature while the modelled values follow a long-range power law correlation. Detrended fluctuation analysis was applied in a similar line to find out the intrinsic properties of fluctuation in Sahel precipitation and rain anomaly (Efstathiou and Varotsos, 2012). Short scale Sahel precipitation anomaly for 1900–2010 has been found to be positively correlated with long scale ones in a power law fashion whereas; the nature of similar relationship presented for 1948–2001 is random. Moreover, it has been reported that the lightning process over different Indian regions have varied dependence on various atmospheric factors (Chatterjee et al., 2023). Therefore, an explainable model for prediction of lightning density is the call of the time. Various places over India are prone to significant amount of lightning especially during the pre-monsoon season. The highest lightning density is observed over the North-eastern part of the country. Increasing weather extremities due to recent climatic changes has made it even more crucial to have a transparent lightning prediction model over such an area of large lightning fatalities (Yadava et al., 2020).

This article proposes a machine learning based model for prediction of lightning density over north-Eastern and Eastern portion of India. The model forecasts monthly lightning activities with a lead time of one month based on atmospheric parameters i.e., air temperature at 2 m, CAPE, K index, surface sensible heat flux values till the previous month. The model output is interpreted through SHAP index for creating a transparent machine learning environment without a shed of an ML black box.

2. Data and methodology

2.1. Data and instrument

The atmospheric parameters namely air temperature at 2 m, CAPE, K index, and surface sensible heat flux, were obtained from ERA 5 monthly dataset provide by ECMWF (European Center for Medium Range Weather Forecast) at a resolution of 0.25° x 0.25° for the period of 2000–2013.

Lightning data has been obtained from Lightning Imaging Sensor (LIS). It is a space-based instrument capable of measuring both cloud to ground and cloud to cloud lightning strikes. It can keep track of lightning rate, and radiant energy as well, both during the day and night. Therefore, monitoring the global lightning activity and identifying the lightning climatologies are possible with this mission. Here, the flash rate information from LIS/OTD Monthly Climatology Time Series (LRMTS) dataset with a resolution of 2.5° × 2.5° (Cecil D LIS/OTD 2.5 Degree Low Resolution Monthly Climatology Time Series, 2003)were used.

2.2. Methodology

2.2.1. Pre-processing of data

The missing values were handled using 5 neighbour KNN imputation (Pedregosa et al., 2011) here. As the atmospheric variables had a resolution of 0.25° x 0.25° the features were interpolated to a similar resolution as that of the lightning data i.e., 2.5° × 2.5°.

2.2.2. Machine learning model for prediction of lightning density

The whole Eastern and North-eastern portion of India (21.25° N-31.25° N, 83.75°E-98.75° E) were considered for the study. Three models were examined for prediction of monthly prediction of lightning density with a lead time of one month based on the atmospheric parameters. The weather variables during 2000–2011 were used for training purpose whereas the data of 2012–13 were kept separate for testing purpose. Samples in training and testing data were 4970 and 840 respectively. The initial predictive feature set was chosen by exhaustive literature study. The initial feature set had Multivariate NINO, CAPE of two previous months, temperature and K index of two previous months, and instantaneous surface sensible heat flux. Further, recursive feature elimination method was used to shortlist the most important features which shortlisted CAPE of two previous months, temperature and K index of two previous months, and instantaneous surface sensible heat flux. The month information was also included in the feature set. Three regression models namely Decision tree (Gladwin, 1989), Random Forest (Breiman, 2001) and XGBoost (Chen and Carlos, 2016) were tested here (Models detailed in Appendix). The model parameters were chosen by hyper-parameter tuning. Here, each hyper parameter was tested with 12 fold cross validation where each fold was a single year. The hyper parameter with best mean score in cross validation was chosen as best hyper parameter set.

The accuracy of prediction has been verified by two metrics i.e. Mean Absolute Error (MAE) and R^2. MAE happens to be the arithmetic mean of the absolute errors between the actual and predicted samples. If x be the actual data with mean $\bar{x}$ and y be the predicted data then

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n} \quad (1)$$

R^2 has been computed as

$$R^2 = 1 \frac{SS_{res}}{SS_{tot}} \quad (2)$$

Where, total sum of square $SS_{tot} = \sum_i (x_i - \bar{x})^2$ and Sum of residual $SS_{res} = \sum_i (x_i - y_i)^2$

2.2.3. Interpretation of model output using SHAP index

Nowadays, more and more complex AI models are being devised. Most of the internal dynamics happens within the shed of a black box. With the development of AI the models will get even more complex and incomprehensible to humans. Here, the output of the model is explained by SHAP value. SHAP (SHapley Additive exPlanations) index was originally discovered by Lundberg and Lee (2017). The method can explain individual predictions and it is actually derived from game theory. SHAP explains each prediction by calculating the role of each feature in the prediction. Each feature here is analogous to a player in game theory. Shapley values describe how to fairly distribute the prediction among the features. The Shapley value of a feature represents its contribution in the resulting prediction, weighted and summed over all possible feature combinations. Suppose, in a model a group N (with n features) predicts a result of v(N). Then, the contribution of each feature is given by its SHapley value i.e.

$$\varnothing_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{j\}) - v(S)] \quad (3)$$

Where, S is a subset of features i.e. the input to the model and v(S) is the prediction for feature values in set S. The total contribution (or Shapley value $\varnothing_j$) of feature a particular feature is computed by the mean of its contribution overall possible permutations.

SHAP is an additive feature attribution method. SHAP defines the model's output as the sum of the real values attributed to each input feature. The explanatory model for additive feature attribution method is shown is equation (4).

$$g(x') = \varnothing_0 + \sum_{j=1}^{M} \varnothing_j x'_j \quad (4)$$

Where $\varnothing_j \in \mathbb{R}$ is the feature attribution for feature j. $x' \in \{0,1\}^M$ is the coalition vector and M is maximum coalition size.

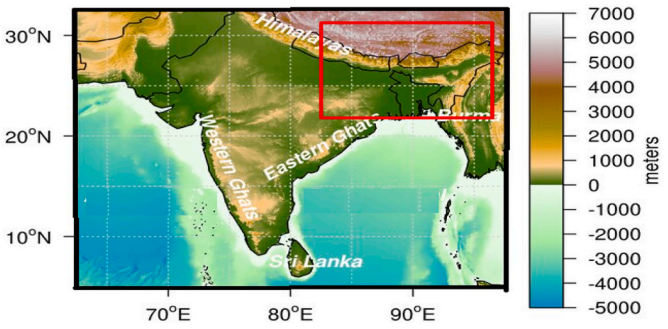


Fig. 1. Position of the area used for this study in Indian map.

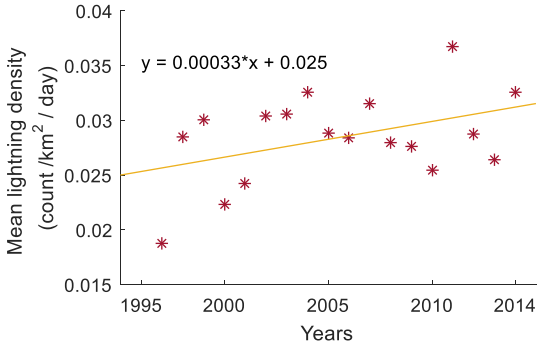


Fig. 2. Mean annual lightning density during 1995–2014 over Eastern and North-Eastern part of India.

3. Result and discussion

The article focuses on the Eastern and North-Eastern part of the country, which is extremely prone to lightning activities especially because of the frequent cyclonic movements in BOB resulting from the large temperature difference between the sea and Eastern coast (Tinmaker and Chate, 2013).

3.1. Recent trend of lightning activities over NE India

The recent climatic changes have increased the frequency of lightning activities due to its role in altering the boundary layer heating pattern (Price, 2009). According to the IPCC report (2021) the regions already prone to tropical turbulences, weather extremities are more susceptible to the increasing severities whereas, the dry and calm places are being at an increasing risk of droughts. The recent trend of lightning activities over the NE and Eastern part of the country (Fig. 1) has been investigated in order to understand the vulnerability of this area to such threats. The annual average lightning activity over this region (21.25° N-31.25° N, 83.75°E-98.75° E) has showed a significantly increasing trend in last 20 years (Fig. 2). Therefore, it can be assessed that proper prediction of weather extremities over the region is crucial.

3.2. Prediction of lightning densities using ML based regression

Three machine learning architecture were developed using three regression models i.e., Decision tree, Random forest and XGBoost. The three models were cross validated with eleven years' of data i.e.,

Table 1
Cross-validation scores with best found hyper parameter of each model.

ML model	Score	Year											Mean	Std	Hyper-parameter
		2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011			
Decision Tree	MAE	0.0099	0.0087	0.0080	0.0095	0.0089	0.0090	0.0100	0.0083	0.0085	0.0085	0.0140	0.0071	0.0092	0.0016	{'max_depth': 27, 'min_samples_split': 55}
	R^2	0.7287	0.8161	0.8411	0.7826	0.7768	0.7340	0.7463	0.8285	0.7812	0.7361	0.6419	0.8697	0.7736	0.0590	
Random Forest	MAE	0.0085	0.0077	0.0070	0.0079	0.0080	0.0081	0.0073	0.0082	0.0071	0.0080	0.0134	0.0059	0.0081	0.0017	{'estimator': 800, 'max_depth': 23, 'max_sample': 0.9, 'min_sample_split': 5}
	R^2	0.8510	0.8739	0.9008	0.8600	0.8553	0.8238	0.8716	0.8472	0.8550	0.8260	0.6631	0.9128	0.8450	0.0602	
XGBoost	MAE	0.0088	0.0078	0.0068	0.0071	0.0084	0.0086	0.0083	0.0080	0.0076	0.0083	0.0131	0.0062	0.0083	0.0016	{'n_estimators': 400, 'max_depth': 9, 'eta': 0.1, 'subsample': 0.9, 'colsample_bytree': 0.9}
	R^2	0.8246	0.8679	0.9009	0.8898	0.8287	0.7927	0.8363	0.8390	0.8335	0.7843	0.6868	0.9025	0.8323	0.0572	

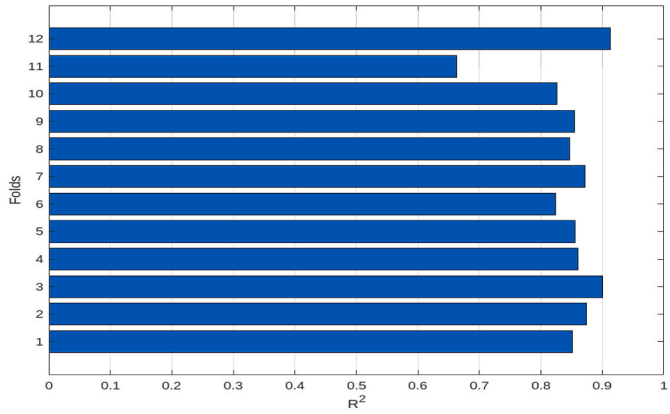


Fig. 3. Performance (R^2) of the model in 12 fold cross validation (2000–2011).

Table 2
Test scores (2012–2013) of each model.

ML Models	Scores	
	R^2	MAE
Decision Tree	0.7879	0.0086
Random Forest	0.8636	0.0071
XGboost	0.8657	0.0072

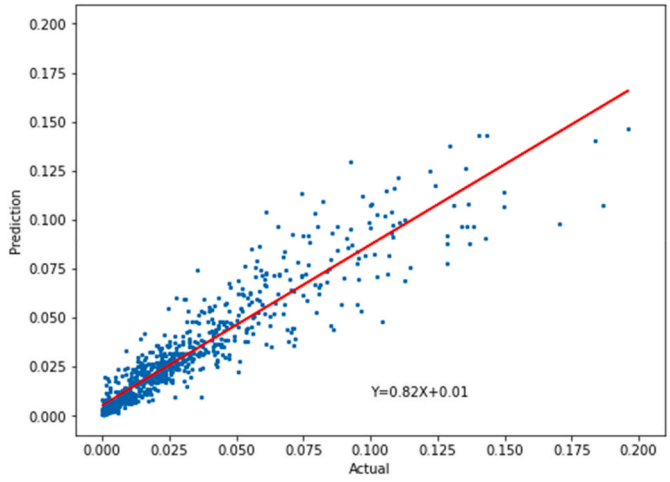


Fig. 4. The actual and predicted lightning density during 2012–2013. (Blue dots represents observed data whereas; the red line is the line of fitting). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2000–2011. Table 1 depicts the best hyperparameter sets as found for each of the model. While the R^2 of validation were found to be comparable and good for Random forest and XGBoost i.e., 0.84 and 0.83, it was found to be decent (0.77) for Decision tree. Performance of the best performing model here, i.e. RF during the cross validation has been presented in Fig. 3 In each fold. It is to be noted here each fold consist of a single year between 2000 and 2011. Further, the models were tested with two years' of data during 2012–2013. The testing R^2 were found to be 0.78, 0.86 and 0.86 for Decision tree, Random Forest and XGBoost (Table 2). Lowest MAE was observed for Random Forest model.i.e. 0.0071. It shows the efficiency of the RF model implemented here, in case of unknown data. In practice, even though decision tree is a much simpler algorithm to be implemented, the capability of RF in minimize the overfitting makes it very effective for all the real world problem especially with large datasets. Even though the testing accuracy showed

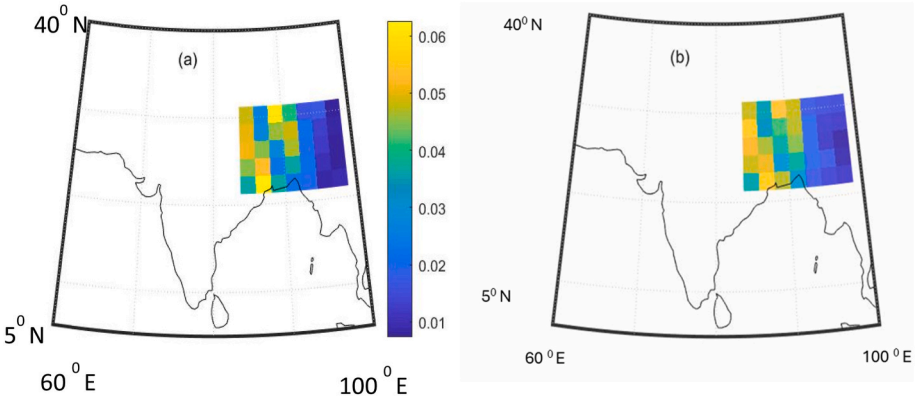


Fig. 5. Spatial pattern of actual and precited mean lightning densities.

Table 3
Season wise performance of the final model.

	Pre-monsoon:	Monsoon:
R^2_score	0.8027	0.7781
MAE_score	0.0128	0.0077

comparable results for Random Forest and XGBoost, the stability of model was found to be better in case of RF which showed a higher cross validation accuracy. For the said reason, the further analysis was carried out with Random Forest.

The actual and predicted values of lightning for the testing data i.e., for the months of 2012–2013 over the North-Eastern India shown very good matching with a 'goodness of fit' of 0.82 (Fig. 4).

Further, given the absolute dynamic nature of lightning process the prediction has been verified with the spatial variability as well. The actual and predicted spatial pattern of lightning averaged over the period of testing (2012–2013) over the area of interest has been presented in Fig. 5. It is evident that the pattern is nicely captured by the model as it closely resembles the actual lightning densities over the 2.5° × 2.5° grids over the Eastern and North-Eastern part of the country. It is to be remembered that even though this is a very highly active lightning zone, the spatial inhomogeneity is large. But it is to be noted that the relative spatial diversity is reflected successfully in the model output. However, there are very slight underestimation of lightning densities in case of very high values.

3.3. Investigation of seasonal bias on output

The output of the model was further investigated to check whether there is any seasonal bias in the model. The performance of the model was explored separately for Pre-monsoon and Monsoon which are the two seasons with majority of lightning activities over India. Table 3 presents the R^2 and Mean absolute error for the two said seasons. The R^2 for Pre-monsoon and monsoon season were found to be 0.80 and 0.77. It is evident that even though the Pre-monsoon scores are a little better, there is no large variation between the performance of the model during Monsoon and Pre-monsoon which confirms the stability of the proposed model.

3.4. Investigation of the influence of the used feature on lightning activities using SHAP index

SHAP can infer the importance of each feature used for the prediction. Fig. 6 depicts the role of the features in the prediction. It is evident that the air temperature of the previous two months have important role to play in deciding the lightning activity over a region during the next month and the relationships are pretty linear i.e. low values of the parameters affected the output (lightning density) in a negative way whereas; positive influence was observed for higher values (Fig. 6(a–b)). It is to be noted, that both the inhibition and favourability on the lightning density were intense for the higher temperature of previous month and the peak SHAP value reached a figure of 0.025 whereas; it was around 0.015 for another one month prior 2 m temperature. Similar association were observed for CAPE (Fig. 6(c–d)). CAPE of previous month (Fig. 6(c)) found to be correlating positive only after its crosses a certain threshold. However, a two months prior CAPE has presented mostly a negative impact on the output. The proportionality of surface temperature and CAPE over the lightning activities of this region fall in line with the previously reported results (Murugavel et al., 2014).

The k index of the previous month, on the other hand, have shown a linear influence on the lightning density of the region (Fig. 6(e)). Interestingly, the high k index during two month (Fig. 6 (f)) before the lightning event showed inhibiting effect on the lightning densities of the region. The k index which is based on the vertical temperature lapse rate was shown to effect the lightning of next month directly. The K index can be assumed to be a thunderstorm potential. So, the linear relationship between k index and lightning is an expected scenario. Probably, the orographic influences have a role to play especially over the NE part of the region considered here. The surface heat flux of previous month (Fig. 6(g)), however, had a pretty linear inhibition effect on the output of the model. This parameter is the transfer of heat between the Earth's surface and the atmosphere, through the effects of turbulent air motion. The turbulence in air can be a good measure of lightning possibilities instantaneously (Tinmaker et al., 2021). Most probably, the previous month's successful radiative transfer has a crucial positive effect on reducing the lightning severities during the coming month.

The study proposes a machine learning based simple approach for forecasting monthly lightning density one month ahead of time based on easily available satellite data. This makes the process suitable for any region on this earth. However, the lightning data available from satellite based sensors have moderate resolution. Here, the data used, has a spatial resolution of 2.5° × 2.5°. Nowadays, there are a few grounds based lightning networks available, but such data are neither freely accessible nor has long time span to train any model. Therefore, most of the lightning prediction being carried out has been focusing at a very confined area of interest. This study attempts to overcome that scenario by using a seamless method using satellite data. However, there has to be a compromise in the part of spatial resolution. Also, slight underestimation was noted by the model in case of the very high value grids as observed in spatial pattern.

4. Conclusion

Precise prediction of lightning densities over such lightning prone areas like North-Eastern and Eastern India is crucial in saving both

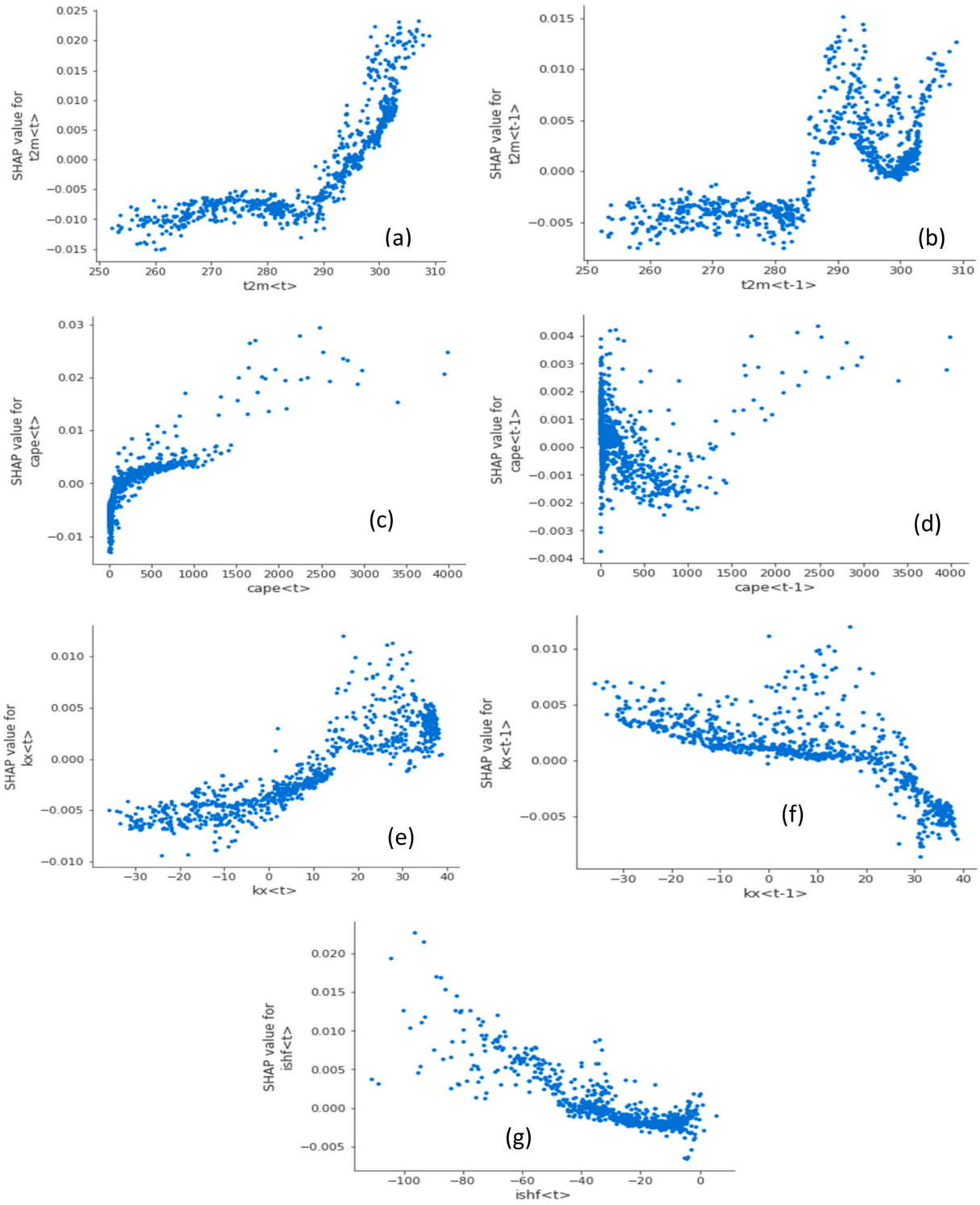


Fig. 6. SHAP values for different features used.

human lives and resources. Month ahead predictions in cases of such complex dynamic weather phenomenon require taking care of both the intricacy and transparency of the process. The current study has proposed a simple model based on Random forest regression using very easily available satellite data to forecast lightning density with a lead time of one month over eastern and North-Eastern part of India. The R^2 score of the model was found to be pretty good (0.86). The model output has been explained with a game theoretic approach called SHAP index. Temperature of two previous months and CAPE of the previous month were found to be governing the process of lightning strongly over this region. The recent climate change scenario has made India extremely vulnerable to weather extremities like lightning. Such studies especially with a transparent machine learning model, capable of showing the role of individual feature on the output can be a good option in the extreme weather prediction over lightning pone areas.

CRediT authorship contribution statement

Joyjit Mandal: Conceptualization, Data curation, Investigation, Methodology. **Chandrani Chatterjee:** Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Saurabh Das:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

data are publicly available and properly cited in text

Acknowledgement

Authors thank the official website of NASA (National Aeronautics and Space Administration) and ECMWF (European Center for Medium Range Weather Forecasting) for the datasets used in the article. One of the authors (SD) also thankfully acknowledges the financial support received under SERB grant MTR/2019/001581.

References

Boccippio, D.J., 2002. Lightning scaling relations revisited. J. Atmos. Sci. 59, 1086–1104. https://doi.org/10.1175/1520-0469(2002)059<1086:LSRR>2.0.CO;2.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

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

Charney, J., Shukla, J., 1981. Predictability of monsoons. Monsoon dynamics 99, 109.

Chatterjee, C., Mandal, J., Das, S., 2023. A machine learning approach for prediction of seasonal lightning density in different lightning regions of India. Int. J. Climatol. 43, 2862–2878. https://doi.org/10.1002/joc.8005.

Chen, T., Carlos, G., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Clark, A.J., Gallus, W.A., Weisman, M.L., 2010. Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. Weather Forecast. 25, 1495–1509. https://doi.org/10.1175/2010WAF2222404.1.

Deierling, W., Petersen, W., 2008. Total lightning activity as an indicator of updraft characteristics. J. Geophys. Res. 113 https://doi.org/10.1029/2007JD009598.

Dowdy, A., 2016. Seasonal forecasting of lightning and thunderstorm activity in tropical and temperate regions of the world. Sci. Rep. 6, 2087. https://doi.org/10.1038/srep20874.

Efstathiou, M., Varotsos, C., 2012. Intrinsic properties of Sahel precipitation anomalies and rainfall. Theor. Appl. Climatol. 109 https://doi.org/10.1007/s00704-012-0605-2.

Elsner, J., Widen, K., 2014. Predicting spring tornado activity in the central Great Plains by 1 March. Mon. Weather Rev. 142, 259–267.

Gagne, D., Christensen, H., Subramanian, A., Monahan, A., 2020. Machine learning for stochastic parameterization: generative adversarial networks in the lorenz '96 model. J. Adv. Model. Earth Syst. 12, e2019MS001896 https://doi.org/10.1029/2019MS001896.

Gladwin, C.H., 1989. Ethnographic Decision Tree Modeling 19. Sage.

Grewe, V.D., Brunner, M., Dameris, J.L., Grenfell, R., Hein, R.D., Staehelin, J., 2001. Origin and variability of upper tropospheric nitrogen oxides and ozone at northern mid-latitudes. Atmos. Environ. 35, 3421–3433. https://doi.org/10.1016/S1352-231000134-0.

Lopez, P.A., 2016. Lightning parameterization for the ECMWF integrated forecasting system. Mon. Weather Rev. 144, 3057–3075. https://doi.org/10.1175/MWR-D-16-0026.1.

Lundberg, S., Lee, S., 2017. A Unified Approach to Interpreting Model Predictions arXiv.

Mallick, C., Hazra, A., Saha, S., Chaudhari, H., Pokhrel, S., Konwar, M., Dutta, U., Mohan, G., Vani, K., 2002. Seasonal predictability of lightning over the global hotspot regions. Geophys. Res. Lett. 49, e2021GL096489 https://doi.org/10.1029/2021GL096489.

Mansel, E.R., Ziegler, C.L., 2013. Aerosol effects on simulated storm electrification and precipitation in a two-moment bulk microphysics model. J. Atmos. Sci. 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 Forecast. 28, 3–28.

Mostajabi, A, et al., 2019. Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. Npj Climate and Atmospheric Science 2, 41.

Murugavel, P., Pawar, S.D., Gopalakrishan, V., 2014. Climatology of lightning over Indian region and its relationship with convective available potential energy. Int. J. Climatol. 34, 3179–3187. https://doi.org/10.1002/joc.3901.

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., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Price, C., 2009. Thunderstorms, lightning and climate change. In: Betz, H.D., Schumann, U., Laroche, P. (Eds.), Lightning: Principles, Instruments and Applications. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-9079-0_24.

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

Reynolds, S., Brook, M., Gourley, M.F., 1957. Thunderstorm charge separation. J. Meteorol. 14, 426–436. https://doi.org/10.1175/1520-0469(1957)014,0426:TCS.2.0.CO;2.

Stolz, D., Rutledge, S.A., Pierce, J.R., 2015. Simultaneous influences of thermodynamics and aerosols on deep convection and lightning in the tropics. J. Geophys. Res. Atmos. 120, 6207–6231. https://doi.org/10.1002/2014JD023033.

Takahashi, T., 1978. Riming electrification as a charge generation mechanism in thunderstorms. J. Atmos. Sci. 35, 1536–1548. https://doi.org/10.1175/1520-0469(1978)035,1536:REAACG.2.0.CO;2.

Tinmaker, I.R., Chate, D., 2013. Lightning activity over India : a study of east – west contrast. International Journal of Remote Sesning 34, 5641–5650.

Tinmaker, M.I.R., Jena, C.K., Ghude, S.D., Dwivedi, A.K., Islam, S., Kulkarni, S.H., Khare, M.K., Chate, D.M., 2021. Relationship of lightning with different weather parameters during transition period of dry to wet season over Indian region. J. Atmos. Sol. Terr. Phys. 220.

Varotsos, C.A., Efstathiou, M.N., Cracknell, A.P., 2013. Plausible reasons for the inconsistencies between the modeled and observed temperatures in the tropical troposphere. Geophys. Res. Lett. 40, 4906–4910. https://doi.org/10.1002/grl.50646.

Williams, E., Stanfill, S., 2002. The physical origin of the land–ocean contrast in lightning activity.Comptes Rendus Physique , 3.10 : 1277-1292.

Wang, B., Ding, Q., Fu, X., Kang, I.S., Jin, K., Shukla, J., Doblas-Reyes, F., 2006. Fundamental challenge in simulation and prediction of summer monsoon rainfall. Geophys. Res. Lett. 32, L15711 https://doi.org/10.1029/2005GL022734.

Williams, E., et al., 2002. Contrasting convective regimes over the Amazon: implications for cloud electrification. J. Geophys. Res. 107, 8082. https://doi.org/10.1029/2001JD000380.

Wilson, J.W., Crook, N.A., Mueller, C.K., Sun, J., Dixon, M., 1998. Nowcasting thunderstorms: a status report. Bull. Am. Meteorol. Soc. 79, 2079–2209.

Yadava, P.K., Soni, M., Verma, S., et al., 2020. The major lightning regions and associated casualties over India. Nat. Hazards 101, 217–229. https://doi.org/10.1007/s11069-020-03870-8, 2020.

Yuan, T.L., Remer, L.A., Pickering, K.E., Yu, H., 2011. Observational evidence of aerosol enhancement of lightning activity and convective invigoration. Geophys. Res. Lett. 38, L04701 https://doi.org/10.1029/2010GL046052.

Ziegler, C.L., MacGorman, D.R., Dye, J.E., et al., 1991. A model evaluation of noninductive graupel-ice charging in the early electrification of a mountain thunderstorm. J. 840 Geophys. Res.-Atmos. 96, 12833–12855. https://doi.org/10.1029/91JD01246,301991.