JGR Atmospheres

RESEARCH ARTICLE
10.1029/2021JD036065

Special Section:
Applications of Machine Learning Algorithms in Modeling Atmospheric Aerosols, Clouds, and Radiation

Relative Importance of Large-Scale Environmental Variables to the World-Wide Variability of Thunderstorms

Nana Liu[1], Chuntao Liu[2], and Philippe E. Tissot[2,3]

[1]Department of Earth System Science, University of California Irvine, Irvine, CA, USA, [2]Department of Physical and Environmental Sciences, Texas A&M University at Corpus Christi, Corpus Christi, TX, USA, [3]Conrad Blucher Institute, Texas A&M University at Corpus Christi, Corpus Christi, TX, USA

Key Points:

- The geographical distribution of thunderstorms is closely reproduced by a global random forest model
- The regional and global random forest models can be used to investigate the relative importance of different large-scale variables for thunderstorm over different regions
- Convective available potential energy, convective inhibition, and warm cloud depth are confirmed as important variables to parameterize convective intensity at the sub-grid scale in climate model

Correspondence to:
C. Liu,
chuntao.liu@tamucc.edu

Citation:
Liu, N., Liu, C., & Tissot, P. E. (2022). Relative importance of large-scale environmental variables to the world-wide variability of thunderstorms. *Journal of Geophysical Research: Atmospheres*, *127*, e2021JD036065. https://doi.org/10.1029/2021JD036065

Received 18 OCT 2021
Accepted 15 AUG 2022

Author Contributions:
Conceptualization: Chuntao Liu
Data curation: Chuntao Liu
Funding acquisition: Chuntao Liu
Investigation: Chuntao Liu
Methodology: Nana Liu, Philippe E. Tissot
Project Administration: Chuntao Liu
Resources: Chuntao Liu
Software: Nana Liu, Chuntao Liu
Supervision: Chuntao Liu
Validation: Nana Liu
Writing – original draft: Nana Liu
Writing – review & editing: Chuntao Liu, Philippe E. Tissot

Abstract This study uses a 16-yr Tropical Rainfall Measuring Mission (TRMM) Convective Features (CFs) and ERA-Interim reanalysis data to investigate the relative importance of four large-scale environmental variables to thunderstorms with random forest models. These four variables include Convective Available Potential Energy (CAPE), Convective Inhibition (CIN), low-level wind shear, and warm cloud depth (WCD). First, these selected four environmental variables show a distinguished difference between CFs with and without lightning flashes. Specifically, CFs with at least one flash have higher CAPE, CIN, and lower WCD than those without lightning. Then, using these four variables, the geographical distribution of thunderstorms, especially the land-ocean contrast in the occurrence of thunderstorms, is closely reproduced with a global random forest model. Such results suggest that a random forest model with key large-scale environmental variables can be a useful tool to estimate the occurrence of global lightning thunderstorms. The study also investigates the relative importance of the selected variables to the occurrence of thunderstorms regionally. Relatively higher skill scores in the regional random forest model than the global one indicate the variation of roles of large scale environment variables over different regions. Though the data-driven models can be utilized to estimate the occurrence of global thunderstorms, how to link the regional relative importance of these variable to the physical processes of thunderstorms needs further investigation.

1. Introduction

It is well known that most lightning on Earth occurs over land, and that lightning is rare over ocean (Christain et al., 2003; Orvile & Henderson, 1986). Though it is possible to generate lightning through volcanic eruption, the majority of the lightning strikes occurs in thunderstorms, where the electric charge separation is mainly driven by the collision of ice particles in convective clouds. Therefore, many attempts have been made to describe lightning with convective ice mass in numerical models (Barthe et al., 2010; Etten-Bohm et al., 2021; Lopez, 2016; Magi, 2015).

Alternatively, the statistical approach of meteorological covariates can be used to investigate the occurrence of these extreme events (Brown & Murphy, 1996; Brooks et al., 2003). Covariates relate environmental conditions that are well-observed in space and time to weather events of interest. This idea has led to many of our current forecasting approaches for intense thunderstorms. For example, proximity soundings, where observations from radiosondes are taken in the vicinity of tornadoes, have been used to understand environmental conditions associated with tornadoes (Rasmussen & Blanchard, 1998). Parameters calculated from sounding data are widely used to identify the pre-convective environments that favor intense thunderstorms. These parameters and indices generally reflect the potential for the development of intense thunderstorms, making it possible to quantify the uncertainty implicit in any forecast (Kaltenböck et al., 2009; Murphy, 1977).

In the past decades, various thermodynamic and kinematic parameters have been designed to characterize the conditions that could favor the formation and development of thunderstorms. Convective available potential energy (CAPE) is the most frequently used as a forecasting tool for gauging the likelihood of intense thunderstorms. This is because the relative motion of hydrometeors inside the cloud depends on the updraft speeds. The strength of updraft is directly related to the vertical profile of CAPE (Blanchard, 1998; Doswell & Evans, 2003; Doswell & Rasmussen, 1994). CAPE has also been used in cumulus parameterization in general circulation models (e.g., Moncrieff & Miller, 1976; Washington & Parkinson, 2005; Ye et al., 1998) and as a predictor of lightning intensity in deep tropical convection (Williams & Renno, 1993). As an opposite parameter of CAPE, convection inhibition (CIN) is another variable that plays an important role in the development of intense convection (e.g., Davies, 2004; Mapes, 1998; Stensrud, 2007). Convection is often widespread but shallow in the absence of CIN (Bennett et al., 2006). The presence of the CIN (or a lid) can allow for the accumulation of heat and moisture, creating the potential for intense convection (Carey & Buffalo, 2007). Besides CAPE and CIN, low-level shear has also gained wide acceptance by the forecaster community as a parameter to identify potential intense weather. Weisman and Klemp (1986) suggested that the bulk Richardson number, which combines CAPE and surface to 6 km wind shear, can be used to differentiate storm type, organization and lifetime. Since then, many researchers have used proxies that combine CAPE and wind shear to model the occurrence of intense convection (e.g., Allen et al., 2015; Diffenbaugh et al., 2013; Sander et al., 2013).

Williams et al. (2005) present that an elevated cloud base height may increase the cloud water concentration in the mixed-phase region because the presence of ice particles in this region is necessary for cloud electrification (e.g., Houze, 1993; MacGorman & Rust, 1998; Takahashi, 1978). Higher cloud base height could result in less entrainment (McCarthy, 1974), stronger updrafts (Williams et al., 2005), and ultimately, higher liquid water content in the mixed-phased zone. In addition, a higher cloud base height also implies a shallower warm cloud depth (WCD), which is the distance between the cloud base height and freezing height. A shallower WCD allows less time for droplets to interact for coalescence (Pierce, 1958) and results in a higher liquid water content in the mixed-phase and charging zone from the freezing of large raindrops (Rosenfeld & Woodley, 2003). Stolz et al. (2015) also confirmed that lightning density and the average height of 30 dBZ echoes are higher for storms with a shallower WCD, compared to those with a deeper WCD. In their multiple-linear regression model to predict global lightning activity, they suggested that WCD, CAPE, and cloud condensation nuclei are among the most influential predictors accounting for the variation of convective intensity (Stolz et al., 2017). Therefore, WCD is also considered in this study.

In past investigations of intense thunderstorms, satellites have provided valuable information about the global distribution of the intense convection (e.g., Cecil & Blankenship, 2012; Liu & Liu, 2016; Spencer & Santek, 1985; Zipser et al., 2006). The near-uniform global coverage of satellites makes them continue to be the most efficient tools to advance the understanding of extreme weather events. As one of the most successful missions during the past decades, the Tropical Rainfall Measuring Mission (TRMM, Kummerow & Barnes, 1998; Kummerow et al., 1998) has provided us a clear picture of the distribution of thunderstorms and lightning activity across the tropics and subtropics (Houze et al., 2015; Liu & Zipser, 2005; Zipser et al., 2006). In order to describe the regional variation of thunderstorms from the perspective of thermodynamic environments, here we focus on the large-scale thermodynamic environments from reanalysis datasets for intense convective systems observed by TRMM. The objective of this study is to answer the following questions:

- Can we utilize these relationships to interpret the global geographical frequency distribution of thunderstorms?
- What is the relative importance of these thermodynamic variables in the probability of thunderstorms over different regions?

To answer these questions, lightning flashes, a commonly used proxy to recognize intense thunderstorms (e.g., Carey et al., 2003; Lang & Rutledge, 2002; MacGorman & Burgess, 1994), are used to identify these rare events. The lightning flashes are observed by the Lightning Imaging Sensor (LIS) on the TRMM satellite. The LIS provided 16+ years of continuous total lightning observations with high detection efficiency across the tropics and subtropics after its launch in 1997. The 16 years of TRMM data and analysis methodology used in this study are introduced in Section 2. The results are presented in Section 3, which includes the relationships between several selected atmospheric factors and the occurrence of intense thunderstorms, the reconstructed geographical distribution and seasonal variation of those events by these relationships, and the relative importance of these atmospheric variables over selected regions.

2. Data and Method

2.1. Convective Features With and Without Lightning

To differentiate the precipitation systems with and without lightning, we use the Version 7 TRMM Convective Feature (CF) database during 1998–2013 (Liu et al., 2008). In this database, CFs are defined by grouping the contiguous pixels with convective precipitation, identified by the TRMM precipitation radar (Iguchi et al., 2009). Within the area of each feature, the total number of TRMM LIS detected lightning flashes are counted. Because

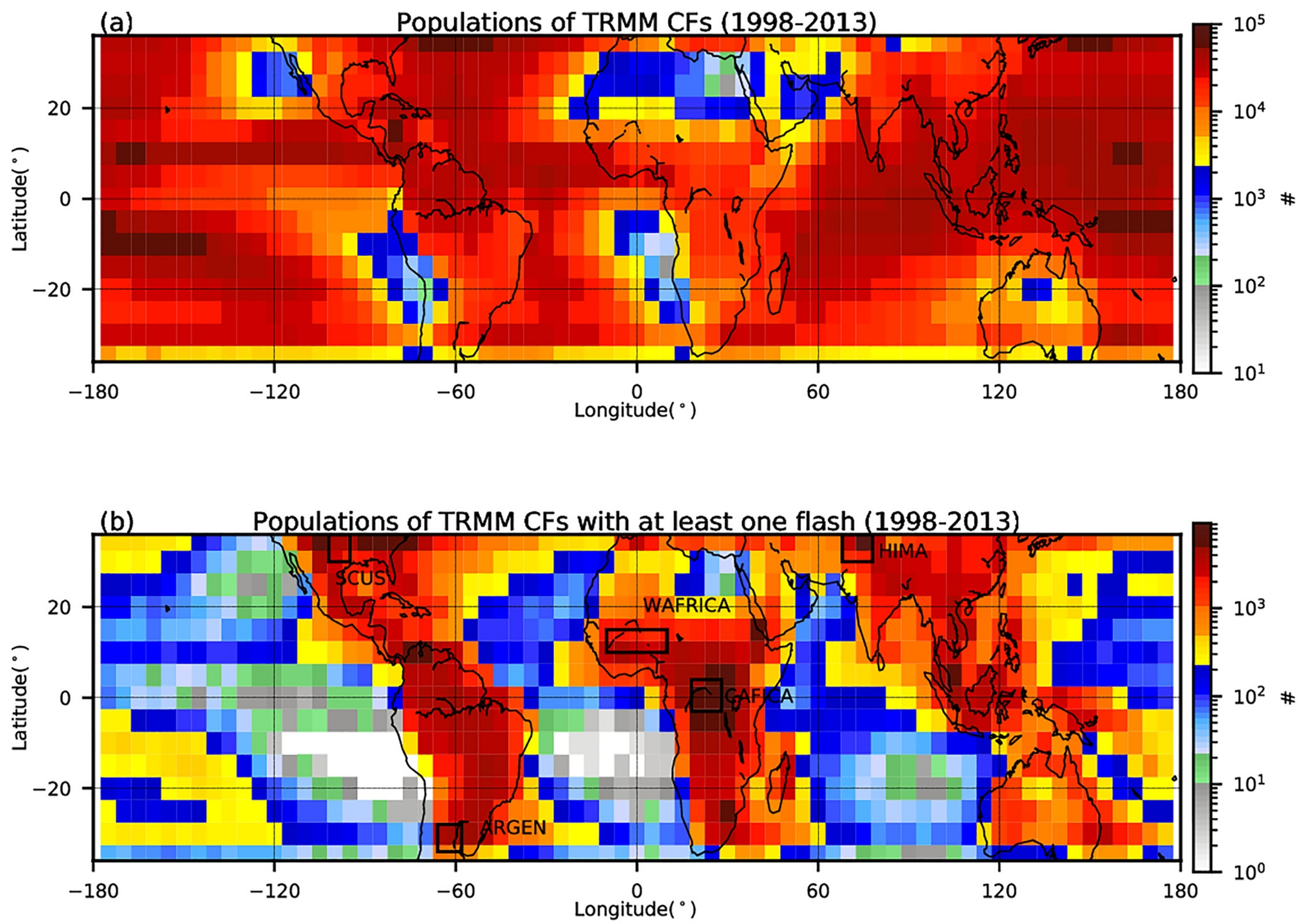

Figure 1. Geographical distribution of (a) Convective features (CFs) and (b) CFs with at least one flash. The distribution is created on a 5° × 5° grid from 16 years (1998–2013) of Tropical Rainfall Measuring Mission (TRMM) observations. The regions of interest are boxed in panel (b).

these two instruments are on the same satellite platform, this provides a concurrent measurement of convective precipitation and lightning (Liu et al., 2011, 2012; Zipser et al., 2006). Since LIS has a staring view time about 80–95 s, this can lead to the minimum detection lightning flash rate of about 0.6 flash per minute, which is used to separate those CFs with at least one flash and without flash. Therefore, in this study, a thunderstorm is really referred to as a CF associated with a lightning flash rate greater than 0.6 flash per minute. Note that here we only consider and focus on the convective region of precipitation systems because the lightning flashes in stratiform region could have more complicated mechanisms (Carey et al., 2005; Peterson & Liu, 2011) and probably have different types of relationships to their large-scale environments. As shown in Figure 1a, the TRMM 16-year CF database provides significant amount samples over 36°S–36°N for this study. To remove noise and make sure we capture the main objective, only CFs with at least four contiguous pixels (with size ~75 km^2) are used. This still provides a total of ~25 million CF samples. Most of these samples are over the ocean (Figure 1a), however, only a small fraction (4%) of these CFs have at least one flash. Most of the one flash CFs are located over land as shown in Figure 1b, consistent with many past studies (Albrecht et al., 2016; Boccippio et al., 2000; Cecil et al., 2014; Christian et al., 2003).

2.2. Thermodynamic Environments From ERA-Interim Reanalysis Data Set

One application of this study is to test the feasibility of estimating thunderstorm distribution using a large-scale environment for convective intensity parameterization in climate models. Most current climate model simulations are running at resolutions around 1° (Eyring et al., 2016; Taylor et al., 2012). There are reanalysis products available at high resolutions, for example, ERA5 (Hersbach et al., 2020) and MERRA-2 (Gelaro et al., 2017).

10.1029/2021JD036065

These products have the scale resolving convective systems and include the impact of these systems on the large-scale environment, make it difficult to pick a representative environment at a high-resolution grid, for example, upstream or downstream, based on only the center locations of CFs. Therefore, in this study, we select to use ERA-Interim product. The large-scale thermodynamic environments are from the 6-hourly ERA-Interim reanalysis data with $0.75° \times 0.75°$ horizontal resolution, at 37 pressure levels (Dee et al., 2011). For each CF, ERA-Interim temperature (T), horizontal wind (u and v), geopotential heights, relative humidity, surface pressure are temporally linearly interpolated to the CF time, and spatially collocated from the closest grid to the CF geographical center location. Then, CAPE and CIN are computed using the algorithms outlined by Emanuel (1994) using the temperature and humidity profiles from levels above the ground surface, which is determined by surface pressure value. The low-level wind shear ($SHEAR_{1-3km}$) is defined as the difference in the horizontal wind between 1 and 3 km above ground level that determined by surface pressure and 37 pressure levels. The WCD is calculated by subtracting the lifting condensation level from the freezing level (height of the 0°C isotherm). Note that all of these variables can be both calculated from observations, if the relevant atmospheric variables are properly measured, or can be obtained from the model outputs.

2.3. Statistical Models to Describe the Regional Variation of Thunderstorms

Because of the non-linear nature of the relationship between the large-scale environments and lightning, we applied a statistical classifier – the random forest model. Random forest method uses bootstrap samples to construct multiple decision trees (Breiman, 2001). Each decision tree is grown with a random subset of predictors to its maximum size. Random forests are trained by fitting a decision tree to each randomly sampled subset of the training data and aggregating the predictors. The random sampling adds more diversity and leads to a more robust model. In the process of each tree growing in the random forest model, the relative importance of each predictor is calculated based on how much the accuracy decreases when this predictor is dropped. In other words, the larger the decrease in accuracy, the more important the variable is in the prediction. Then, the relevance is normalized so that the sum of all the relative importance values equals to 1. To perform the prediction using the trained random forest model, we pass the test features through each randomly created tree. Each predicted target vote is calculated for the same feature. Then, the most frequently predicted class by the individual trees, also known as a majority vote, is considered the final prediction. One drawback of random forest method is the requirement of large samples, which is well-satisfied by the amount of samples from 16-year TRMM CF database.

To build the random forest models, the open-source Python package from scikitlearn (https://scikit-learn.org/stable) is used. Here, we label CFs with at least one flash as class 1 and CFs with no flashes as class 2. 16-year data set of TRMM CFs and their four thermodynamic variables (CAPE, CIN, Shear, and WCD) is used to build the random forest models. 80% of the data set is used as training data to calibrate the random forest models. In addition, CFs with lightning are relatively rare events, making up only ~4% of all the CFs over the whole TRMM domain. To minimize the impact of the unbalanced data set, the model uses the values of the target to automatically adjust weights inversely proportional to class frequencies to balance the data set. Then, the environmental variables from the remaining 20% of the ERA-Interim reanalysis data are used to classify the CFs with lightning or not and assess the performance of the random forest model. To verify the robustness of the random forest model, we randomly split samples and rerun the model 50 times and derived the standard deviations from results.

3. Results

3.1. Relationships Between Thermodynamic Variables and Probability of Thunderstorms

The histograms of selected thermodynamic and dynamic environments in CFs with or without lightning flashes are shown in Figure 2. Most of CFs with lightning have higher CAPE, CIN, and lower WCD than those without lightning. The CFs with lightning over land tend to have larger CAPE and CIN and lower WCD than over ocean. The CFs without lightning have slight differences between land and ocean, with lower CAPE, CIN, and low-level wind shear over ocean. This is consistent with the literature showing that the magnitude of CIN is relatively smaller over ocean than over land (Williams & Renno, 1993). Compared to other three variables, low level wind shear shows a relatively less significant difference between lightning and non-lightning CFs (Figure 2c). All these suggest that these variables are related to the lightning generation mechanisms, which could be different between land and ocean.

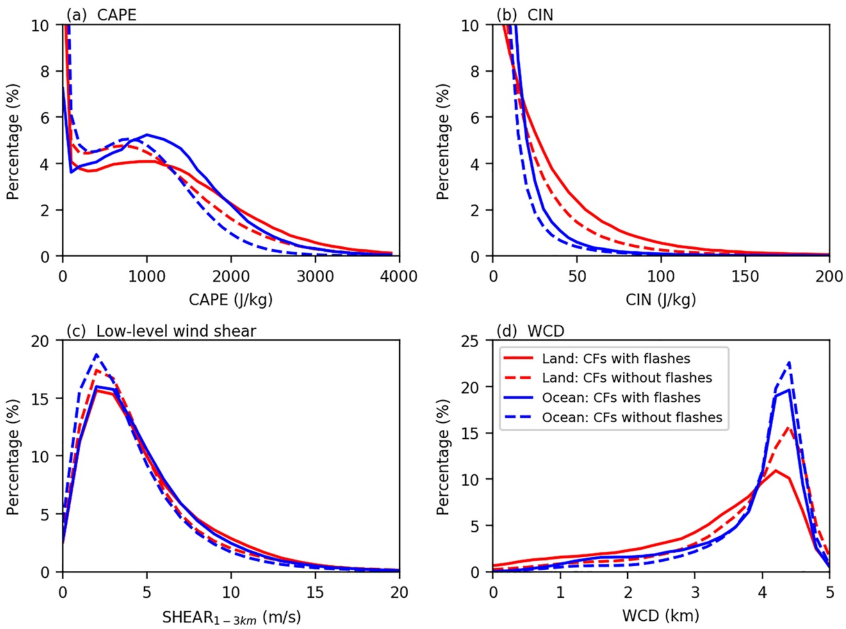

Figure 2. Histogram of thermodynamic environments for convective features (CFs) with and without flashes over land and ocean.

3.2. A Random Forest Model for Thunderstorms Using the Four Thermodynamic Variables

The relative importance of these four atmospheric variables is further investigated using global and local random forest models. In the global random forest model, the whole 16-year TRMM CFs and their environmental variables, derived from ERA-Interim reanalysis data are used to train and test the random forest model. Note that in the global random forest model, the land and ocean CFs are equally treated without a scaling factor, meaning that there are more CFs with lightning from land but more CF samples from over ocean in the training data set.

Figure 3a presents the geographical distribution of thunderstorms with at least one flash from the testing data set, which is randomly selected from all CFs as shown in Figure 1b. Even with 20% of the whole data set, the general geographical distribution of thunderstorms is consistent with the well-documented climatology of thunderstorms (e.g., Boccippio et al., 2000; Christian et al., 2003; Liu et al., 2012). Figure 3b shows the geographical distribution of number of CFs predicted with lightning based only on four ERA-Interim variables using the random forest global model. The geographical distribution pattern of predicted and observed number of CFs are close with spatial correlation coefficients of 0.88 between the observed and the random forest model predicted CFs (Figures 3a and 3b). The differences between the predicted and the observed are shown in Figure 3c. Over most of regions, the random forest model underestimates the total CFs with lightning by 10%–50% in tropics, especially over the tropical ocean. Over dry regions with fewer CF samples (e.g., Sahel), the random forest model overestimates the CFs with lightning.

The land dominance of thunderstorms has been related to the surface properties and aerosol effects (Williams & Sátori, 2004). However, without considering the aerosols, the strong land and ocean contrast in the frequency of thunderstorms can be closely reproduced with the global random forest model only based on the four thermodynamic variables from the reanalysis data (Figure 3b). This suggests that the land-ocean contrast in convective intensity can be largely interpreted by the fundamental differences between the thermodynamic conditions over land and ocean.

3.3. Regional Variation of Relative Importance of Thermodynamic Variables for Thunderstorms

The relative influence of the four environmental variables for thunderstorms is investigated further over selected regions (shown in Figure 1b). As listed in Table 1, the large sample size of training and testing datasets makes it possible to build a robust model not only globally, but also individual models for the selected regions. The

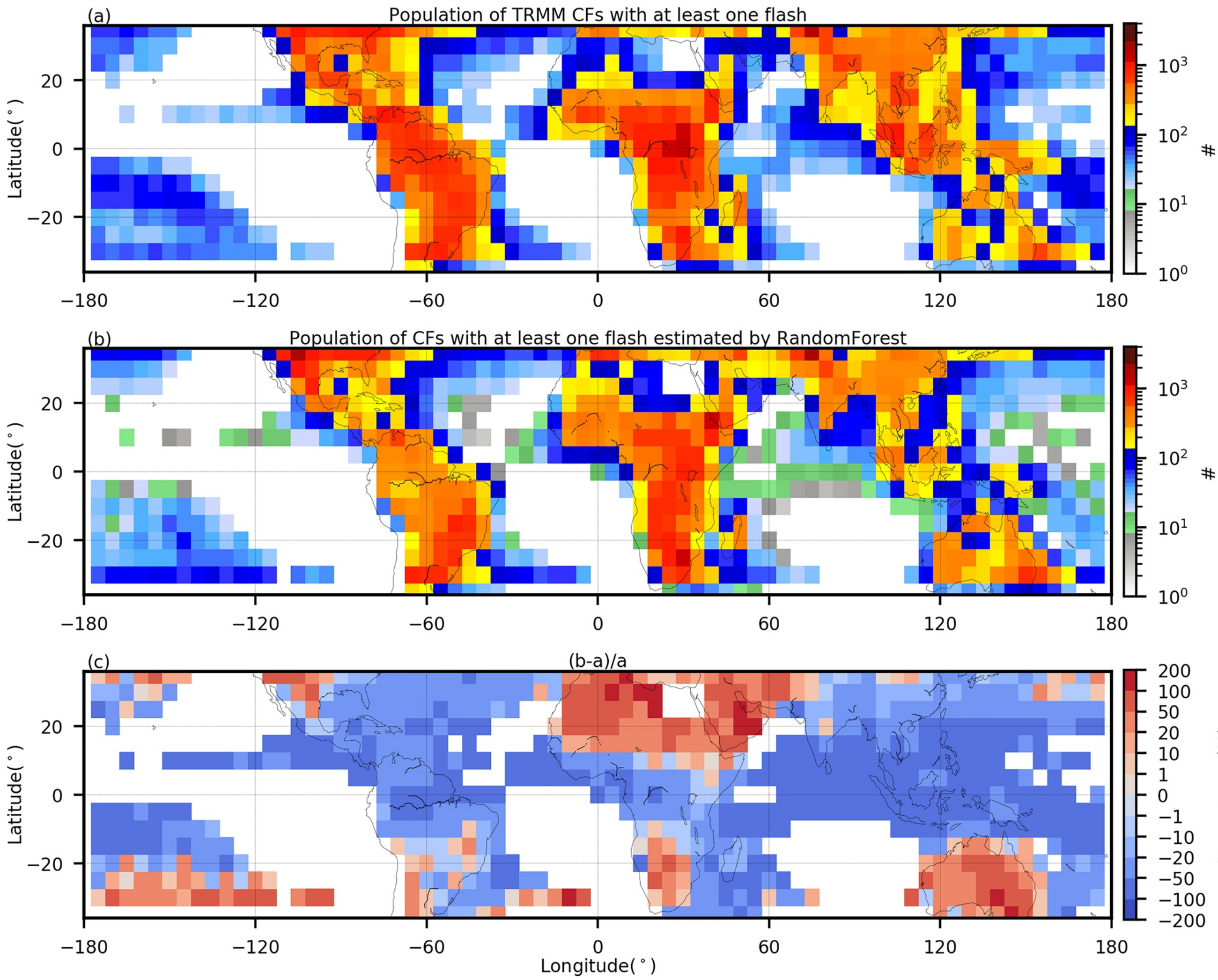

Figure 3. (a) Geographical distribution of the population of intense convective features (CFs) observed by Tropical Rainfall Measuring Mission (TRMM), (b) Estimated geographical distribution of the population of intense CFs by global random forest model with four atmospheric variables (Convective Available Potential Energy (CAPE), convective inhibition, $SHEAR_{1\text{–}3km}$ and warm cloud depth). (c) Normalized difference between the estimation and observations. The geographical distribution of CFs is created in 5° × 5° grids. Note that the random forest model used here includes all samples over land and ocean.

prediction skill scores of the models over different regions are listed in Table 1, including the probability of detection, false alarm rate, Critical Success Index (CSI), and Heidke Skill Score (HSS). The values of these scores range from 0 to 1, with 1 being ideal score. As a commonly used skill score for assessing the accuracy of thunderstorms forecasts (e.g., Huntrieser et al., 1997; Mazur et al., 2009; Mitchell et al., 1998), the CSI measures the ratio of hits to the total number of hits, false alarms and misses. The HSS compares the proportion of correct forecasts to a no skill forecast. While not impressive, the regional models have HSS values ranging between 0.11 for CAFRICA and 0.29 for Himalayas (HIMA). Note that the current models still have difficulties in predicting individual cases with correct time and locations, which results in relatively low skill scores. However, the well-reproduced geographical distribution of the occurrence of thunderstorms with four selected variables, especially the land-ocean contrast, implies that these models are still useful tools to improve the understanding of the mechanism of thunderstorms. The interpolation of 6 hourly ERA-Interim data could lead to mismatch to the convective system and lead to a relatively low CSI and HSS scores. However, given a large enough size of samples, the random forest model could capture the main relationship between the environment and thunderstorms. Therefore, the geographical distribution of the occurrence of thunderstorms is well reproduced.

Table 1
Populations of Training and Testing CFs With at Least One Flash and Skill Scores of Random Forest Models Over Global and Selected Regions

		Global	ARGEN	CAFRICA	SCUS	HIMA	WAFRICA
Training: 0.8	CFs (#)	18,154,483	21,188	35,636	21,145	28,455	19,625
	>0 flashes (#)	723,368	5,291	12,344	5,139	11,404	5,541
Testing: 0.2	CFs (#)	4,538,612	5,297	8,909	5,287	7,114	4,907
	>0 flashes (#)	147,578	1,389	3,047	1,295	2,966	1,139
Skill Scores	POD	0.26	0.44	0.50	0.46	0.67	0.59
	FAR	0.68	0.61	0.60	0.63	0.44	0.61
	CSI	0.17	0.26	0.29	0.26	0.44	0.31
	HSS	0.26	0.19	0.11	0.19	0.29	0.21

Note. ARGEN, Argentina; CFs, convective features; CSI, critical success index; FAR, false alarm rate; HIMA, Himalayas; HSS, Heidke Skill Score; SCUS, South Central US; sPOD, probability of detection.

For example, higher CSI values from the selected regions than the global one indicate the importance of local effects of the large-scale environment on the occurrence of thunderstorms. Therefore, the relative importance of the four variables is further explored with the global and regional random forest models.

Figure 4 illustrates the relative importance of the four environmental variables for the prediction of thunderstorms over different regions. In the global random forest model, the relative importance of CAPE is the highest among these four variables. However, the relative importance of these four atmospheric variables varies by region. For example, Argentina (ARGEN) and CAFRICA have CAPE as the most important predictor; WCD is more important over South Central US (SCUS) and HIMA. Over the whole TRMM domain, the relative influence of CIN and $SHEAR_{1-3km}$ on the prediction of thunderstorms is relatively weak compared to CAPE and WCD. Though an important factor in the development of intense convection (e.g., Diffenbaugh et al., 2013; Sander et al., 2013; Weisman & Klemp, 1982; Weisman & Trapp, 2003), $SHEAR_{1-3km}$ is found to have the lowest relative importance

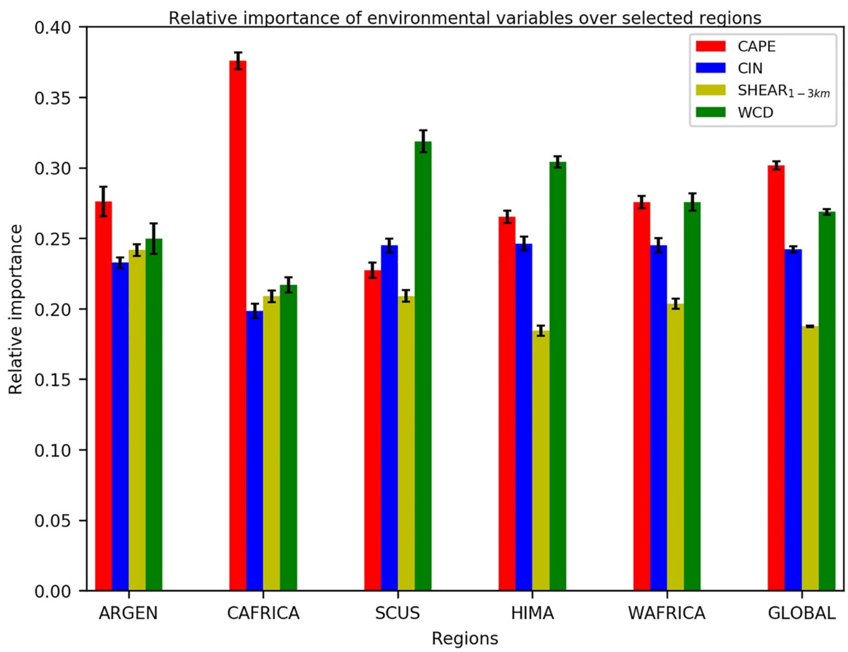

Figure 4. Random forest relative importance of four environmental variables (Convective Available Potential Energy (CAPE), convective inhibition (CIN), $SHEAR_{1-3km}$, warm cloud depth (WCD)) in the probability of thunderstorms over different regions. For each local region, 80% of convective features are used to train the model, and rest 20% are used to test the performances.

10.1029/2021JD036065

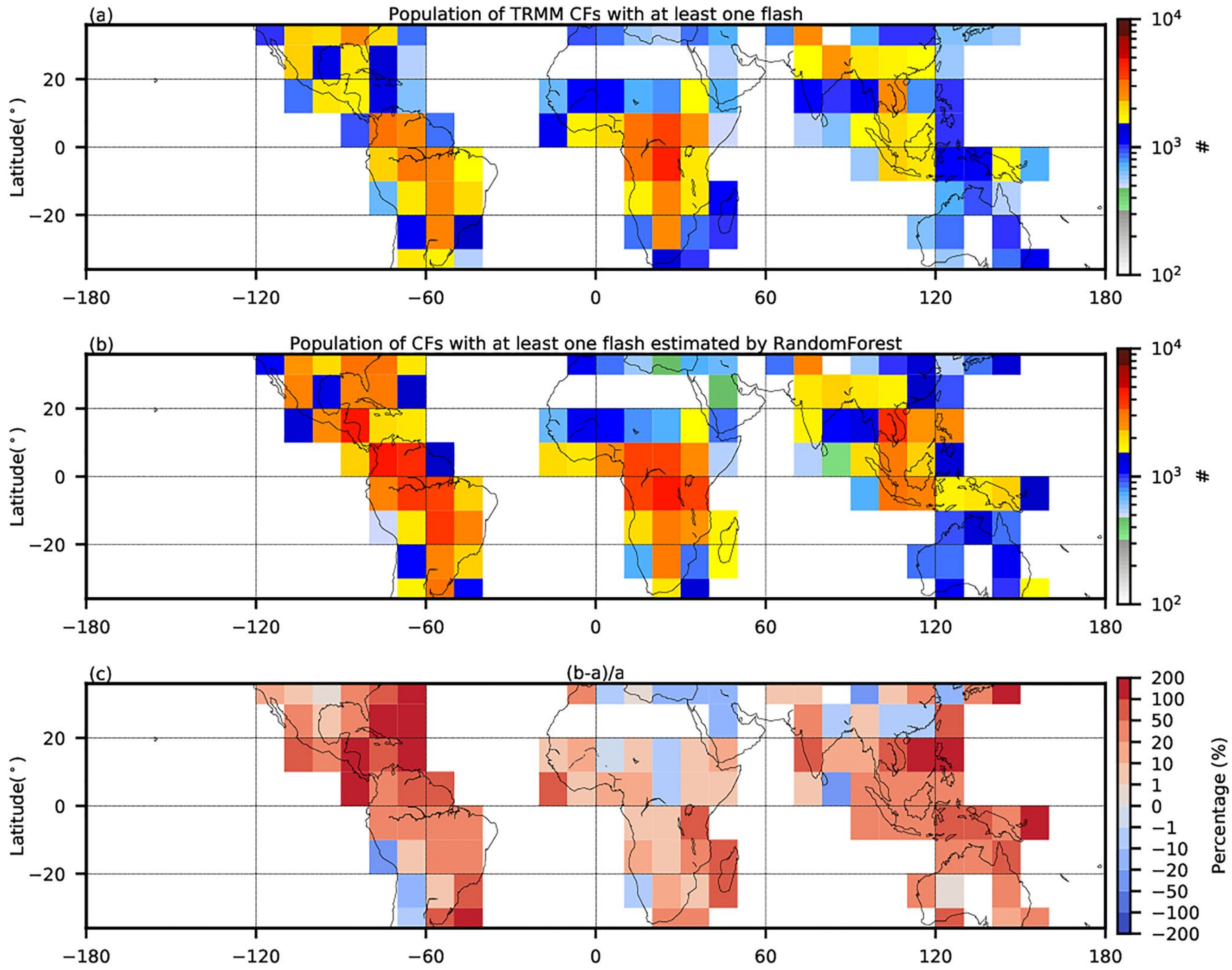

Figure 5. (a) Geographical distribution of the population of convective features (CFs) with at least one flash observed by Tropical Rainfall Measuring Mission (TRMM), (b) Estimated geographical distribution of the population of CFs with at least one flash by local random forest model with four atmospheric variables (Convective Available Potential Energy, convective inhibition, $SHEAR_{1-3km}$ and warm cloud depth). (c) Normalized difference between the estimation and observations. The geographical distribution of CFs is created on a 10° × 10° grid. The local random forest model is created from 16-years (1998–2013) of TRMM CFs. In each grid, 80% of the data is used to build the model, while the remaining 20% of the 16-year CF data set is used to test the model. Boxes with less than 500 CFs with at least one flash are left blank.

to predict thunderstorms among these four variables. Note that here all the CFs with at least one flash are the targets, which include numerous weak thunderstorms with low lightning rates. The low-level shear is known to be important for organized thunderstorms. Therefore, it is not a surprise that shear does not stand out in this case. Etten-Bohm et al. (2021) also suggested that there is significant difference between lightning and non-lightning environments.

To investigate the regional variation of the relative importance of these four variable, we reconstructed the geographical distribution of thunderstorms using local random forest model in 10° × 10° grids. In each grid, 80% of the data is used as a training data set while the remaining 20% is considered as a testing data set. In this process, only the grids with sufficient target samples (>500 CF with lightning samples) are analyzed and shown. The 20% of the observations and prediction of thunderstorms using 10° × 10° local random forest models are shown in Figure 5. Overall, there is a slight overestimation of thunderstorms by the local models. In addition to the general pattern of the geographical distribution, the hotspot regions of intense thunderstorms, such as CAFIRCA, WAFICA, HIMA, SCUS and ARGEN, can also be closely reproduced. The overestimation of thunderstorms is found to be more pronounced over coastal regions.

10.1029/2021JD036065

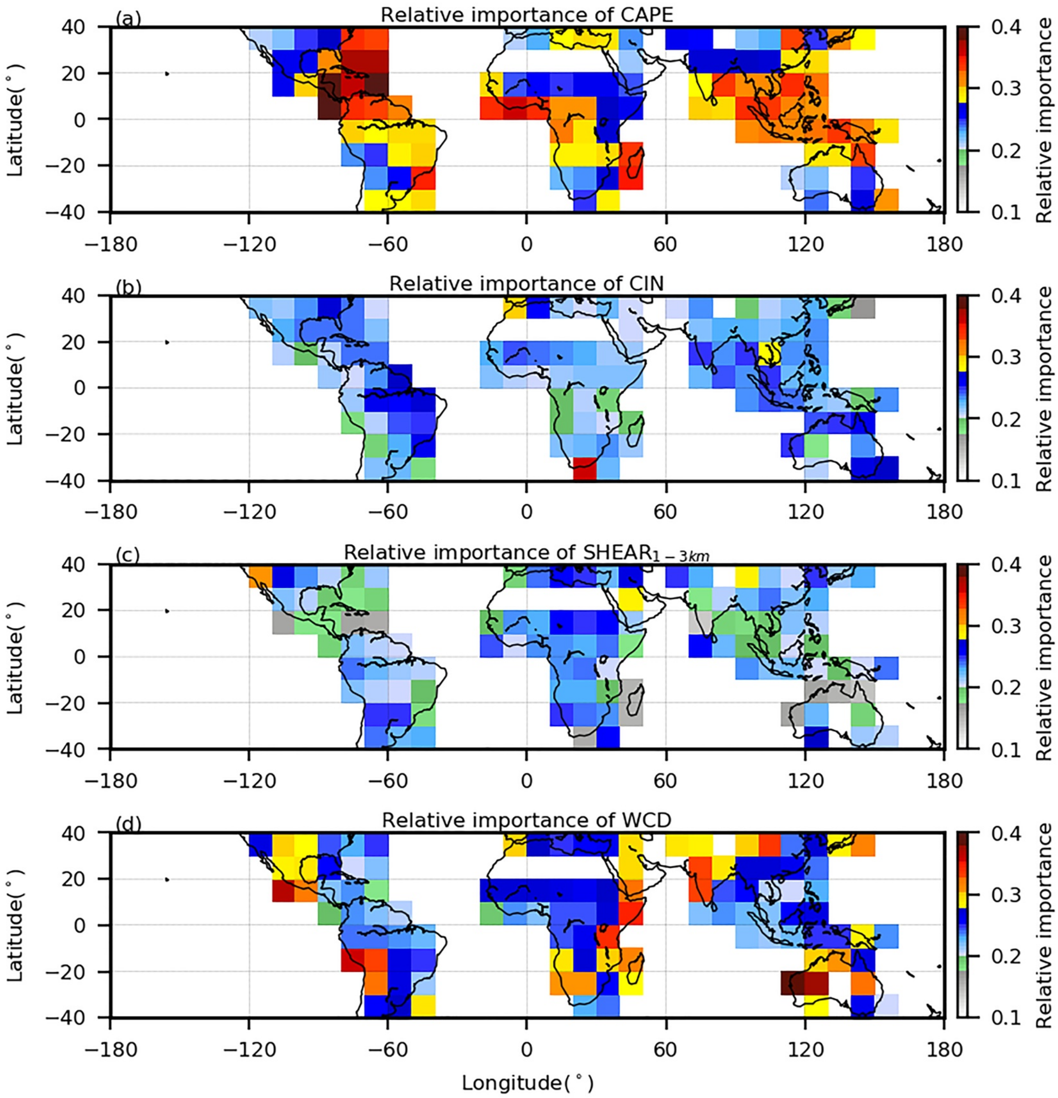

Figure 6. The geographical distribution of the relative importance of environmental variables in the process of predicting convective features with lightning using random forest model within 10° × 10° grid. (a) Convective Available Potential Energy (CAPE), (b) convective inhibition (CIN), (c) SHEAR$_{1-3km}$, (d) warm cloud depth (WCD).

The regional variation of the relative importance of the four environmental variables is shown in Figure 6. Higher relative importance values of CAPE are found over the Amazon, central Africa and the Maritime Continents, as well as the coastal regions (Figure 6a). For example, the Gulf of Guinea is characterized by higher relative importance of CAPE than that of the other three variables. The Bay of Bengal, which is the site of the highest mean precipitation of the entire Asian monsoon region (Zuidema, 2003), is also found to have the highest relative importance of CAPE. In general, the relative importance of CIN is smaller than other variables (Figure 6b). However, higher relative importance of CIN over the Amazon, southeast United States and the Maritime Continents indicates its influence on thunderstorms is more pronounced over these regions. These land regions have more abundant moisture and are close to oceanic scenario (e.g., Martin et al., 2016). CIN is needed to accumulate the moisture potential energy. Hotspot regions of thunderstorms, such as SCUS, ARGEN and the Sahel (Zipser et al., 2006), are characterized by higher relative importance of SHEAR$_{1-3km}$ than other regions (Figure 6c). WCD is relatively more important over mountain regions (Figure 6d). It is interesting to note that compared to CIN and low-level shear, WCD is relatively more important for convection to have lightning over many land regions, such as South Central US (Figure 6d), west Indian, northern Australian, and Japan. We speculate that these regions have convective systems of different types of weather regimes that WCD can be used as a good indicator. All

10.1029/2021JD036065

above regional variations indicate that the role of these environmental variables in the development of thunderstorms can vary in a small-scale regionally.

4. Summary

The analysis of a 16-yr TRMM CFs data set has shown that the probability of thunderstorms is related to their thermodynamic and kinematic environmental variables. Those environmental variables, derived from the ERA-Interim reanalysis data, include CAPE, CIN, $\mathrm{SHEAR}_{1\text{–}3\mathrm{km}}$, WCD. Each of them corresponds to certain physical processes. In order to examine the relationships between thermodynamic environments and thunderstorms, random forest models are built to reconstruct the global and regional distribution of thunderstorms based only on the variables derived from the reanalysis data. The major findings are following:

- There is significant difference in CAPE, CIN, $\mathrm{SHEAR}_{1\text{–}3\mathrm{km}}$, WCD between CFs with and without lightning, as well as between land and ocean. With these four large-scale parameters, the geographical distribution of thunderstorms is well-represented utilizing a global random forest model. This implies the crucial role of in CAPE, CIN, $\mathrm{SHEAR}_{1\text{–}3\mathrm{km}}$, WCD in thunderstorms and the value of random forest model as a useful tool to estimate the occurrence of global thunderstorms.
- The well-reproduced land-ocean contrast in thunderstorms suggests that the fundamental differences between the thermodynamic conditions over land and ocean are largely responsible for the differences in producing lightning.
- The relative importance of the four environmental variables is examined further over several hotspot regions as well as over $10° \times 10°$ grids over land. The importance of the four variables varies significantly in different regions and likely in seasons. CAPE plays an important role in the occurrence of thunderstorms over Amazon, central Africa, the maritime continents, and the coastal regions. The WCD defined by the difference between the cloud bottom and freezing level height, is helpful in the prediction of thunderstorms over mountain and dry regions.

Caution must be taken in interpreting some of the results here. First, ERA-Interim reanalysis data used in this study have uncertainties resulting from the forecast model, data assimilation, and data sources used (Dee et al., 2011). Second, only four environmental variables have been selected as inputs of the models. Other relevant factors, such as aerosol, column cloud fraction, and synoptic ascent are not considered in this study. Moreover, the current models in this study still have difficulties in predicting individual case correctly. Such difficulty results in a relatively poor skill score of the model. All of these factors mentioned above, as well as the limited resolution of ERA-Interim reanalysis data (both temporally and spatially), make further investigation with higher resolution datasets necessary to understand the favorable conditions for intense thunderstorms over different regions. Even with these limitations, the well-represented geographical distribution and the land-ocean contrast of thunderstorms suggest that these models of this study are valuable to estimate the occurrence of a thunderstorm with the help of longer-record reanalysis and model simulations under the changing climate. The next step of this study is to include additional observations from the Global Precipitation Mission (Hou et al., 2014) and to use ERA5 reanalysis data set (Hersbach et al., 2020). With more data samples and a higher temporal and spatial resolution environmental data set, the relative importance of other environment variables can be further explored, and better statistical models could be constructed.

Data Availability Statement

ERA-Interim reanalysis products (ECMWF 2009) are freely available online in archives hosted by the University Corporation for Atmospheric Research (UCAR, https://rda.ucar.edu/datasets/ds627.0/) and European Centre for Medium-Range Weather Models (ECMWF, https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim). The TRMM data products are freely available online in archives hosted by NASA Precipitation Process System (PPS) at https://gpm.nasa.gov/data/sources/pps-research. The TRMM Precipitation Feature (PF) databases are freely available online hosted by Texas A&M University at Corpus Christi at https://atmos.tamucc.edu/trmm/data/.

10.1029/2021JD036065

Acknowledgments

This research was supported by NASA Precipitation Measurement Mission grant 80NSSC22K0591 under the direction of Dr. Gail Jackson. This material is also based upon work supported by the National Science Foundation under Grant No. ICER-2019758 and AGS-2219639. Thanks to the Precipitation Processing System (PPS) team at NASA Goddard Space Flight Center, Greenbelt, MD, for data processing assistance.

References

Albrecht, R. I., Goodman, S. J., Buechler, D. E., Blakeslee, R., & Christian, H. J. (2016). Where are the lightning hotspots on Earth? *Bulletin America Meteorology Social*, *97*(11), 2051–2068. https://doi.org/10.1175/bams-d-14-00193.1

Allen, J. T., Tippett, M. K., & Sobel, A. H. (2015). An empirical model relating U.S. monthly hail occurrence to large-scale meteorological environment. *Journal of Advances in Modeling Earth Systems*, *7*(1), 226–243. https://doi.org/10.1002/2014MS000397

Barthe, C., Deierling, W., & Barth, M. C. (2010). Estimation of total lightning from various storm parameters: A cloud-resolving model study. *Journal of Geophysical Research*, *115*, D24202. https://doi.org/10.1029/2010JD014405

Bennett, L. J., Browning, K. A., Blyth, A. M., Parker, D. J., & Clark, P. A. (2006). A review of the initiation of precipitating convection in the United Kingdom. *Quarterly Journal of the Royal Meteorological Society*, *132*(617), 1001–1020. https://doi.org/10.1256/qj.05.54

Blanchard, D. (1998). Assessing the vertical distribution of CAPE. *Weather and Forecasting*, *13*(3), 870–877. https://doi.org/10.1175/1520-0434(1998)013<0870:atvdoc>2.0.co;2

Boccippio, D. K., Goodman, S. J., & Heckman, S. (2000). Regional differences in tropical lightning distributions. *Journal of Applied Meteorology*, *39*(12), 2231–2248. https://doi.org/10.1175/1520-0450(2001)040<2231:rditld>2.0.co;2

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brooks, H. E., Lee, J. W., & Craven, J. P. (2003). The spatial distribution of intense thunderstorm and tornado environments from global reanalysis data. *Atmospheric Research*, *67–68*, 73–94. https://doi.org/10.1016/s0169-8095(03)00045-0

Brown, B. G., & Murphy, A. H. (1996). Verification of aircraft icing forecasts: The use of standard measures and meteorological covariates. In *13th Conference on Probability and Statistics in the Atmospheric Sciences* (pp. 251–252). The American Meteor Society. Preprints.

Carey, L. D., & Buffalo, K. M. (2007). Environmental control of cloud-to-ground lightning polarity in severe storms. *Monthly Weather Review*, *135*(4), 1327–1353. https://doi.org/10.1175/mwr3361.1

Carey, L. D., Murphy, M. J., McCormick, T. L., & Demetriades, N. W. S. (2005). Lightning location relative to storm structure in a leading-line, trailing-stratiform mesoscale convective system. *Journal of Geophysical Research*, *110*, D03105. https://doi.org/10.1029/2003JD004371

Carey, L. D., Petersen, W. A., & Rutledge, S. A. (2003). Evolution of cloud-to-ground lightning and storm structure in the Spencer, South Dakota, tornadic supercell of 30 May 1998. *Monthly Weather Review*, *131*(8), 1811–1831. https://doi.org/10.1175//2566.1

Cecil, D. J., & Blankenship, C. B. (2012). Toward a global climatology of intense hailstorms as estimated by satellite passive microwave imagers. *Journal of Climate*, *25*(2), 687–703. https://doi.org/10.1175/jcli-d-11-00130.1

Cecil, D. J., Buechler, D. E., & Blakeslee, R. J. (2014). Gridded lightning climatology from TRMM-LIS and OTD: Dataset description. *Atmospheric Research*, *135*, 404–414. https://doi.org/10.1016/j.atmosres.2012.06.028

Christian, H. J. (2003). Global frequency of lightning as observed from space by the optical transient detector. *Journal of Geophysical Research*, *108*, 4005. https://doi.org/10.1029/2002JD002347

Davies, J. M. (2004). Estimations of CIN and LFC associated with tornadic and non-tornadic supercells. *Weather Forecasting*, *19*(4), 714–726. https://doi.org/10.1175/1520-0434(2004)019<0714:eocala>2.0.co;2

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Diffenbaugh, N. S., Scherer, M., & Trapp, R. J. (2013). Robust increases in intense thunderstorm environments in response to greenhouse forcing. *Proceedings of the National Academy of Sciences*, *110*(41), 16361–16366. https://doi.org/10.1073/pnas.1307758110

Doswell, C. A., III, & Evans, J. S. (2003). Proximity sounding analysis for derechos and supercells: An assessment of similarities and differences. *Atmospheric Research*, *67–68*, 117–133. https://doi.org/10.1016/s0169-8095(03)00047-4

Doswell, C. A., III, & Rasmussen, E. N. (1994). The effect of neglecting the virtual temperature correction on CAPE calculations. *Weather Forecasting*, *9*(4), 625–629. https://doi.org/10.1175/1520-0434(1994)009<0625:teontv>2.0.co;2

ECMWF. (2009). ERA-Interim Project [Dataset]. Research Data Archive at the National Center for Atmospheric Research. Retrieved from https://rda.ucar.edu/datasets/ds627.0/

Emanuel, K. A. (1994). *Atmospheric convection*. Oxford University Press.

Etten-Bohm, M., Yang, J., Schumacher, C., & Jun, M. (2021). Evaluating the relationship between lightning and the large-scale environment and its use for lightning prediction in global climate models. *Journal of Geophysical Research: Atmospheres*, *126*, e2020JD033990. https://doi.org/10.1029/2020jd033990

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective analysis for research and applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/jcli-d-16-0758.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *The Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., et al. (2014). The global precipitation measurements mission. *The Bulletin of the American Meteorological Society*, *95*(5), 701–722. https://doi.org/10.1175/BAMS-D-13-00164.1

Houze, R. A. (1993). *Cloud dynamics* (p. 573). Academic Press.

Houze, R. A., Jr., Rasmussen, K. L., Zuluaga, M. D., & Brodzik, S. R. (2015). The variable nature of convection in the tropics and subtropics: A legacy of 16 years of the tropical rainfall measuring mission satellite. *Reviews of Geophysics*, *53*(3), 994–1021. https://doi.org/10.1002/2015RG000488

Huntrieser, H., Schiesser, H. H., Schmid, W., & Waldvogel, A. (1997). Comparison of traditional and newly developed thunderstorm indices for Switzerland. *Weather Forecasting*, *12*(1), 108–125. https://doi.org/10.1175/1520-0434(1997)012<0108:cotand>2.0.co;2

Iguchi, T., Kozu, T., Kwiatkowski, J., Meneghini, R., Awaka, J., & Okamoto, K. (2009). Uncertainties in the rain profiling algorithm for the TRMM precipitation radar. *Journal of the Meteorological Society of Japan*, *87*, 1–30. https://doi.org/10.2151/jmsj.87A.1

Kaltenböck, R., Diendofer, G., & Dotzek, N. (2009). Evaluation of thunderstorm indices from ECMWF analyses, lightning data and intense storms reports. *Atmospheric Research*, *93*, 381–396.

Kummerow, C., & Barnes, W. (1998). The tropical rainfall measuring mission (TRMM) sensor package. *Journal of Atmospheric and Oceanic Technology*, *15*, 809–817.

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., & Simpson, J. (1998). The tropical rainfall measuring mission (TRMM) sensor package. *The Journal of Atmospheric and Oceanic Technology*, *15*(3), 809–817. https://doi.org/10.1175/1520-0426(1998)015<0809:ttrmmt>2.0.co;2

Lang, T. J., & Rutledge, S. A. (2002). Relationships between convective storm kinematics, precipitation, and lightning. *Monthly Weather Review*, *130*(10), 2492–2506. https://doi.org/10.1175/1520-0493(2002)130<2492:rbcskp>2.0.co;2

Liu, C., Cecil, D., & Zipser, E. J. (2011). Relationships between lightning flash rates and passive microwave brightness temperatures at 85 and 37 GHz over the tropics and subtropics. *Journal of Geophysical Research*, *116*, D23108. https://doi.org/10.1029/2011JD016463

Liu, C., Cecil, D. J., Zipser, E. J., Kronfeld, K., & Robertson, R. (2012). Relationships between lightning flash rates and radar reflectivity vertical structures in thunderstorms over the tropics and subtropics. *Journal of Geophysical Research*, *117*, D06212. https://doi.org/10.1029/2011jd017123

Liu, C., & Zipser, E. (2005). Global distribution of convection penetrating the tropical tropopause. *Journal of Geophysical Research*, *110*, D23210. https://doi.org/10.1029/2005JD00006063

Liu, C., Zipser, E. J., Cecil, D. J., Nesbitt, S. W., & Sherwood, S. (2008). A cloud and precipitation feature database from 9 years of TRMM observations. *Journal of Applied Meteorology and Climatology*, *47*(10), 2712–2728. https://doi.org/10.1175/2008JAMC1890

Liu, N., & Liu, C. (2016). Global distribution of deep convection reaching tropopause in 1 year GPM observations. *Journal of Geophysical Research: Atmospheres*, *121*(8), 3824–3842. https://doi.org/10.1002/2015JD024430

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

MacGorman, D. R., & Burgess, D. W. (1994). Positive cloud-to-ground lightning in tornadic storms and hailstorms. *Monthly Weather Review*, *122*(8), 1671–1697. https://doi.org/10.1175/1520-0493(1994)122<1671:pctgli>2.0.co;2

MacGorman, D. R., & Rust, W. D. (1998). *The electrical nature of storms* (p. 422). Oxford University Press.

Magi, B. I. (2015). Global lightning parameterization from CMIP5 climate model output. *Journal of Atmospheric and Oceanic Technology*, *32*(3), 434–452. https://doi.org/10.1175/jtech-d-13-00261.1

Mapes, B. (1998). The large-scale part of tropical mesoscale convective system circulations. *Journal of the Meteorological Society of Japan Series II*, *76*(1), 29–55. https://doi.org/10.2151/jmsj1965.76.1_29

Martin, S. T., Artaxo, P., Machado, L. A. T., Manzi, A. O., Souza, R. A. F. D., Schumacher, C., et al. (2016). Introduction: Observations and modeling of the Green Ocean Amazon (GoAmazon2014/5). *Atmospheric Chemistry and Physics*, *16*(8), 4785–4797. https://doi.org/10.5194/acp-16-4785-2016

Mazur, R. J., Weaver, J. F., & Vonder Haar, T. H. (2009). A preliminary statistical study of correlations between inflow feeder clouds, supercell or multicell thunderstorms, and severe weather. *Weather Forecasting*, *24*(4), 921–934. https://doi.org/10.1175/2009waf2222149.1

McCarthy, J. (1974). Field verification of the relationship between entrainment rate and cumulus cloud diameter. *Journal of the Atmospheric Sciences*, *31*(4), 1028–1039. https://doi.org/10.1175/1520-0469(1974)031<1028:fvotrb>2.0.co;2

Mitchell, E. D., Vasiloff, S. V., Stumpf, G. J., Witt, A., Eilts, M. D., Johnson, J. T., & Thomas, K. W. (1998). The national severe storms laboratory tornado detection algorithm. *Weather Forecasting*, *13*(2), 352–366. https://doi.org/10.1175/1520-0434(1998)013<0352:tnsslt>2.0.co;2

Moncrieff, M. W., & Miller, M. (1976). The dynamics and simulation of tropical cumulonimbus and squall lines. *Quarterly Journal of the Royal Meteorological Society*, *102*(432), 373–394. https://doi.org/10.1002/qj.49710243208

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. *Monthly Weather Review*, *105*(7), 803–816. https://doi.org/10.1175/1520-0493(1977)105<0803:tvocca>2.0.co;2

Orvile, R. E., & Henderson, R. W. (1986). Global distribution of midnight lightning: September 1977 to August 1978. *Monthly Weather Review*, *114*(12), 2640–2653. https://doi.org/10.1175/1520-0493(1986)114<2640:gdomls>2.0.co;2

Peterson, M., & Liu, C. (2011). Global statistics of lightning in anvil and stratiform regions over tropics and subtropics observed by TRMM. *Journal of Geophysical Research*, *116*, D23201. https://doi.org/10.1029/2011JD015908

Pierce, E. T. (1958). Some topics in atmospheric electricity. In *Recent advances in atmospheric electricity* (pp. 5–16).

Rasmussen, E. N., & Blanchard, D. O. (1998). A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Weather Forecasting*, *13*(4), 1148–1164. https://doi.org/10.1175/1520-0434(1998)013<1148:abcosd>2.0.co;2

Rosenfeld, D., & Woodley, W. L. (2003). Closing the 50-year circle: From cloud seeding to space and back to climate change through precipitation physics. In *Cloud systems, Hurricanes, and the tropical Rainfall measuring mission (TRMM), meteor. Monogr* (pp. 59–80). American Meteor Society.

Sander, J., Eichner, J. F., Faust, E., & Steuer, M. (2013). Rising variability in thunderstorm-related U.S. losses as a reflection of changes in large-scale thunderstorm forcing. *Weather, Climate, and Society*, *5*(4), 317–331. https://doi.org/10.1175/WCAS-D-12-00023.1

Spencer, R. W., & Santek, D. A. (1985). Measuring the global distribution of intense convection over land with passive microwave radiometery. *Journal of Applied Meteorology*, *24*(8), 860–864. https://doi.org/10.1175/1520-0450(1985)024<0860:mtgdoi>2.0.co;2

Stensrud, D. J. (2007). *Parameterization schemes: Keys to understanding numerical weather prediction models* (p. 459). Cambridge University Press.

Stolz, D. C., Rutledge, S. A., & Pierce, J. R. (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

Stolz, D. C., Rutledge, S. A., Pierce, J. R., & van den Heever, S. C. (2017). 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*(14), 7461–7492. https://doi.org/10.1002/2016jd026220

Takahashi, T. (1978). Riming electrification as a charge generation mechanism in thunderstorms. *Journal of the Atmospheric Sciences*, *35*(8), 1536–1548. https://doi.org/10.1175/1520-0469(1978)035<1536:reaacg>2.0.co;2

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/bams-d-11-00094.1

Washington, W. M., & Parkinson, C. L. (2005). *An introduction to three-dimensional climate modeling*. University Science Books.

Weisman, M. L., & Klemp, J. B. (1982). The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Monthly Weather Review*, *110*(6), 504–520. https://doi.org/10.1175/1520-0493(1982)110<0504:tdonsc>2.0.co;2

Weisman, M. L., & Klemp, J. B. (1986). In P. S. Ray (Ed.), *Characteristics of isolated convective storms. Mesoscale Meteorology and Forecasting* (pp. 331–358). American Meteor Society.

Weisman, M. L., & Trapp, R. J. (2003). Low-level mesovortices within squall lines and bow echoes. Part I: Overview and dependence on environmental shear. *Monthly Weather Review*, *131*(11), 2779–2803. https://doi.org/10.1175/1520-0493(2003)131<2779:lmwsla>2.0.co;2

Williams, E., & Renno, N. (1993). An analysis of the conditional stability of the tropical atmosphere. *Monthly Weather Review*, *121*(1), 21–36. https://doi.org/10.1175/1520-0493(1993)121<0021:aaotci>2.0.co;2

Williams, E. R., Mushtak, T. V., Rosenfeld, D., Goodman, S., & Boccippio, D. (2005). Thermodynamic conditions favorable to superlative thunderstorm updraft, mixed phase microphysics and lightning flash rate. *Atmospheric Research*, *76*(1–4), 288–306. https://doi.org/10.1016/j.atmosres.2004.11.009

Williams, E. R., & Sátori, G. (2004). Lightning, thermodynamic and hydrological comparison of the two tropical continental chimneys. *Journal of Atmospheric and Solar-Terrestrial Physics*, *66*(13–14), 1213–1231. https://doi.org/10.1016/j.jastp.2004.05.015

Ye, B., Del Genio, A. D., & Lo, K. K. (1998). CAPE variations in the current climate and in a climate change. *Journal of Climate*, *11*(8), 1997–2015. https://doi.org/10.1175/1520-0442-11.8.1997

Zipser, E., Liu, C., Cecil, D., Nesbitt, S. W., & Yorty, S. (2006). Where are the most intense thunderstorms on Earth? *Bulletin of the American Meteorological Society*, *87*(8), 1057–1071. https://doi.org/10.1175/bams-87-8-1057

Zuidema, P. (2003). Convective clouds over the Bay of Bengal. *Monthly Weather Review*, *131*(5), 780–798. https://doi.org/10.1175/1520-0493