Review 1

Review of "Machine Learning based investigation of the variables affecting summertime lightning frequency over the Southern Great Plains" by Shan et al.

The authors use observational data from a high-quality ground observing site in the US, along with winds from ERA5 reanalysis, and aerosol information from MERRA2 reanalysis to study the relationship between meteorological variables at lightning occurrence. They select days with deep convective clouds and then employ ML techniques to statistically explain the lightning frequency with the met variables. A good accuracy is found with the random forest, though the different ML methods to not vary in accuracy hugely. This model is used to investigate the potential drivers of lightning in some detail, including identifying square root of CAPE as correlated with intra-cloud fraction of lightning.

Overall, this is nice bit of work. There isn't a huge amount of depth or exploration of the ML methods and their potential, but sufficient analysis for some interesting findings regarding significance of range of relevant meteorological variables, as well as aerosol. The various contexts, e.g., on deep convective days, or high CAPE days, add nuanced angles. This adds value, and is generally well explained. I have a few gripes that I outline below but the major comments should not be difficult to remedy. Whilst the paper doesn't present great new insight it is a useful addition to the literature, so I do recommend the paper for publication if the points below can be addressed.

We greatly appreciate the instructive and constructive comments. We have studied them carefully and have addressed them in the revised manuscript. Below are the point-by-point responses to the reviewer's comments.

Major comments:

General – To what extent did you consider using interpretable ML techniques? For scientific understanding of processes they would seem more appropriate and I think should at least be discussed in the intro or discussion as a potential avenue to move beyond the ML black box.

We employed feature importance, which is an interpretable ML technique. To help better understand this quantity, more elaboration is added: The feature importance is a measure of how much one variable decreases the impurity, i.e., the probability that more than one class of data remains in a node after processing through various decision trees in the forest. The feature

importance is calculated by determining the increase or decrease in error when we permute the values of a feature (input variable). If permuting the values causes a huge change in the error, it means the feature is important for our model. In general, feature importance determines how useful a feature (input variable) in the ML model.

Besides feature importance, we used the traditional comparison between non-lightning hours and frequent-lightning hours in section 4.2, so that the internal physical mechanism of lightning is discussed. According to our comparison, our findings are consistent with many previous researches. In this regard, this part provided additional evidence to support what we have found in section 4.1 using ML model, making it more transparent instead of being black box. Also, with the additional sentences briefly introducing feature importance, the whole article is more comprehensive.

Abstract/conclusion – It must be made clear near the beginning of the abstract that all results are based on pre-selected hours with deep convection occurring (unless I've misunderstood, in which case please make sure that's clear in the text). This is important because it means you can't expect the same level of accuracy if applied to any random hour. This point also needs to be made at the beginning of the conclusions before citing model accuracy.

This is a really good point and we have revised the abstract and conclusion section to emphasize that the analysis is limited to hours when convective clouds are detected: When convective clouds were detected, it nowcast lightning occurrence with an accuracy of 76.9 % and an area under curve (AUC) of 0.850.

Fig4 – I find this approach of fitting to binned data a bit questionable. How is an R2 useful here given that if using the relationship to make a prediction one would only ever apply to an individual point, not a binned mean? I can accept the analysis if you also provide the regression line and R2 for the grey points (without any binning).

Original points (grey) and binned mean points (red stars) are both shown in the plot. The values of R (inter-correlation) are added for the linear fits both before and after binning. We want to point out a potential positive relationship between sqrt(CAPE) and IC ratio instead of making IC ratio prediction directly from sqrt(CAPE), which is not feasible given the noise.

Fig5 – how are the ovals determined, "by eye"? There are statistical techniques for doing this. But I wonder if a simple way to make this more objective would be if you took the top and

bottom terciles of IC fractions and plotted the histograms of their IC height, and of their sqrt(CAPE). This would supplement the existing fig5 which looks a bit vague but has some merit.

We have revised Figure 5 and discussion. and now use a 2 x 2 chi-square calculation rather than an "eye" test to examine any association between the variables. It is a test of a hypothesis used in the analysis of contingency tables to examine whether two categorical variables (two dimensions of the contingency table) are independent in influencing the statistics of the test.

Minor comments:

L10 - "excels" compared to what? If you mean compared to non-ML methods then I think this is purely opinion and should either be stated as such to avoid suggesting that this paper provides any evidence for the statement, or softened to suggest that ML methods can be useful but not suggest they are better than alternatives.

Revised. "Several commonly used machine learning (ML) models have been applied to analyse …" We have softened the statements, as ML method is an efficient method to demonstrate the feature importance ranking. As we know, the lightning is a complicated outcome of multiple convective variables, such as CAPE, cloud thickness, etc. ML technique helps us rank the importance of these variables for lightning formation, nowcasting, and/or prediction.

L16 - "predicted" whilst not necessarily incorrect, it is a bit misleading in this case. The relationships can "explain" that level of lightning occurrence. But since the input variables are taken from the same time as the lightning occurrence (please make sure this is clear in the methods), I don't really consider them to be predicting.

Yes, now-casting is a better verb here as it can mean prediction of the very recent past, the present (which is what is done here), or the near future. On the other hand, "prediction" is more commonly used in the ML field to define an output that comes from the model, as well as outside of the weather community. We replaced "predict" by "nowcast" when appropriate.

L28 - "affected by" is doing quite a bit of work here. There is evidence that these variables could be statistically used to explain lightning variability. But it seems a stretch to say rain rate

affects the frequency of lightning via a physical mechanism. I think using the term "related to" in some way would be more precise.

Revised to "related to".

L115 – Why have you only taken the closest era5 grid cell for wind shear, but aggregated MERRA AOT to 1degree?

*The website we acquired the MERRA-2 AOT from includes software that provides us with a distance-weighted average remapping to the GPCC1.0 grid (Level 3 and 4 Regridder and Subsetter Information), see L128. This software is not provided with the ERA5 data set so we used the value in the grid cell containing the SGP site. Also, we found that the averaged 1 deg by 1 deg wind shear is very similar to the 0.25*0.25 shear.*

L125 – How different are the winds between ERA5 and MERRA over the site and surrounding region. It seems that using these two datasets for different variables is somewhat conflicting.

The temporal resolution of the archived MERRA-2 wind dataset is too coarse to be useful here. Thus, we use ERA5 winds when calculating the wind shear.

Sec3.1 - can we have some more information about what parameters were chosen for the model development (e.g. how many trees, their depth, criterion for quality of split)

We have tried multiple sets of Random Forest parameters, but found that the default set whose parameters are specified in: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) performs best. The values of the parameters are too numerous to list here. To address your questions, the number of trees in the forest is 100, the maximum depth of the tree is "none", so the nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples, which is the minimum number of samples required to split an internal node. Varying the number of trees (e.g., 10, 50 and 100 as default), the tree depth (e.g., 5, 10 and "none" as default) and other criteria have also been tested.

Fig1 correlations – If N approximately 800 then several of these may still be significant correlations, which I think should probably be noted.

In the revised Fig1, the asterisks denote correlations that are significant values at $p<0.05$.

Methods – Please include a short summary of each classifier method tested. Highlight key differences and what about them warranted testing them here.

An additional paragraph is added to introduce each classifier in the table.

The Support Vector Machine (SVM) algorithm fits a hyperplane in space. The dimensions of the hyperplane are equal to the number of features. This approach results in a distinct classification of data points. The kernel uses a set of mathematical functions to process the data. Linear and radial basis function (RBF) kernels are two different kernels used in SVM. Logistic Regression is a classification algorithm used to predict a binary outcome based on a set of independent variables and the Sigmoid function. Decision Tree is a tree-like structure where each internal node tests an attribute, each branch corresponds to an attribute value and each leaf node represents the final decision or prediction. Gaussian Naive Bayes is based on the probabilistic approach and Gaussian distribution, which assumes that each parameter has an independent capacity of predicting the output variable.

L218 – Please include a definition of feature importance.

Added.

The feature importance is a measure of how much one variable decreases the impurity, i.e., the probability that more than one class of data remains in a node after processing through various decision trees in the forest. An impurity of zero means that all of the cases in the node fall into the same class or category.

L221 - "importance" is such a loaded name for this metric, but that can't be helped. My understanding is that it is calculated from which variables occur most in the trees. To me that is a very vague definition of important. So I would encourage you to spell out what "feature importance" is and then I would avoid interchanging it with the assumption that those are actually the most important features, given the metric's limitations. I would for instance consider changing this sentence to say '"Cloud Thickness", "Rain Rate" and "CAPE" are the most frequently used variables in the model'. That is upfront about what has been calculated. A better metric to judge how important a variable is would be to remove it from the model and see how successful the remaining variables are at producing a model.

Added in the manuscript.

The importance of variables can also be estimated by removing them from the model and seeing how successful the remaining variables are at predicting the true outcome. By removing "Cloud Thickness", "Rain Rate" and "CAPE" separately, the accuracy dropped from 76.9 %

to 72.1 %, 75.6 % and 74.3 %, and the AUC dropped from 0.850 to 0.797, 0.830 and 0.821. The impact of removing other variables was smaller compared with these three variables.

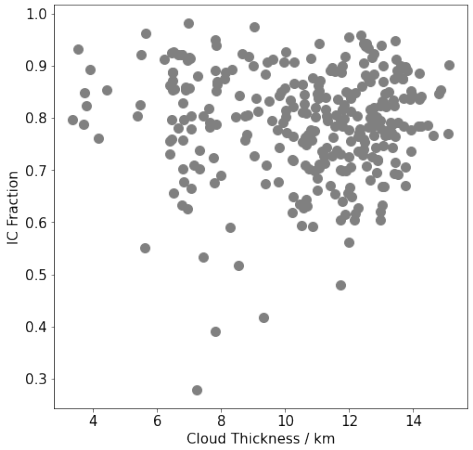
L270 – previous studies tend to find a boomerang relationship, so which part of that boomerang are you aerosol values in?

Wang et al. (2018) binned the data and used the 3-month averaged AOD mean to obtain the boomerang curve. The turning point is about 0.3, and the maximum AOD can be up to more than 0.9, which is a severely polluted environment. However, aerosol concentration was low at SGP (the average and median value of PM2.5 concentration are 9.1 and 8.2 ug/m^3. Also, in table 6 we showed that the average MERRA-2 AOT is less than 0.3 in no-lightning hours and frequent-lightning hours.

Table6 – is there any difference in the LCL temperature?

Slightly different although $p < 0.05$, no-lightning hours: 24.1 ℃ vs frequent lightning hours: 23.0 ℃.

Fig4 – Can you make an equivalent plot using the geometric cloud depth? That is the main metric I've have seen used to predict IC fraction before, so it would provide an interesting comparison.



The figure is shown here and R-squared = 0.0014 in this situation, suggesting that the variables are uncorrelated.

L352 – Could you give a sentence or so in the paper to comment on what the IC fraction is like in non-plentiful cases? Is it highlight variable, or generally low, or high?

In non-plentiful cases, IC fraction can be problematic to do statistical analysis because there may be less than 10 flashes in an hour. The 25th percentile count is 21.0 (when count > 0 and convective clouds detected). When flashes are below 25th percentile, the IC ratio is 0.662 ± 0.316, but the value for plentiful flash hours (what we have done in our MS) is 0.789 ± 0.103, respectively. We can see the standard deviation divided by the mean is large during non-plentiful cases, so we only considered the plentiful flash hours.

Technical comments:

Fig1 caption – is this a "pearson" correlation?

Fig1 caption – are based on the 817 hours or something else? Please state in the caption.

They are the Pearson correlations based on 817 hours of convective clouds detected and where it was possible to calculate all 8 variables. We also added this in the caption of Figure 1.

Fig3 caption – please include in the caption the total number of hours in the histogram. And ideally what % of all model data that this was.

Added is the total number of hours in the histogram. There are 608 hours with both flashes and convective clouds detected, accounting for 2.3% among all 26352 hours in the summer months (June, July, August and September) from 2012 to 2020.

Review 2

This manuscript tried to identify meteorological factors that may affect the lightning occurrence using random forest (RF) and composite analysis based on the observations at ARM SGP site. The results show that the RF model was superior to several other frequently-used machine-learning methods, with the geometric cloud thickness, rain rate and CAPE as the most important predictors. The composite analysis shows that other variables such as equivalent potential temperature and mid-altitude humidity may affect the lightning occurrence frequency and that the square root of CAPE is positively correlated with the fraction of intra-cloud flashes in the lightning occurrence. I think the results are interesting but not well organized. It is unclear what are the new findings of this study compared with previous studies. Some issues in the analysis are also ambiguous and need further clarification. Therefore, I suggest a major revision before this manuscript can be accepted. The detailed comments are given blow.

Many thanks for the constructive comments. We have studied them carefully and have addressed them in the revised manuscript. Below are the point-to-point responses to the reviewer's comments.

General comments:

It is unclear what scientific questions the manuscript is trying to tackle. It seems most findings can be found in previous studies reviewed in Introduction. The authors have to clarify this in Introduction.

*The key scientific question of our study is which variables are the most important to predict the lightning occurrence. Previous research was focused on one or two variables, or one class of variables to determine their impact on lightning following a systematic approach to narrow and choose the variables. We don't need to control the rest of the environment, enabling us to demonstrate hourly lightning occurrence research with small region (1*1 degree) instead of discussing daily/monthly lightning averaging over a large region. A regressed methodology to rank the variables that affect lightning is provided, which enables us to identify the importance of various variables in the complex convective systems. As reviewer one noted, the various contexts, e,g, on deep convective days, or high CAPE days, add nuanced angles. Also, the IC fraction relationship with sqrt(CAPE) is newly discovered. While tenuous, this relationship may spur follow-up studies.*

The RF analysis (Section 4.1) and the composite analysis (Sections 4.2 and 4.3) seem apart and not connected at all. The importance of cloud thickness, rain rate and CAPE has been revealed by previous studies, so Section 4.1 is not necessary for the composite analysis. Meanwhile, the RF analysis focuses on the lightning occurrence (lightning vs. no lightning) while the composite analysis focuses on the lightning frequency and type, where the variable importance could be different. The authors better clarify the logical relationship between the sections for better understanding of the results.

This is a good point and we have revised the title focusing on "occurrence" instead of "frequency". Also, we added sentences to connect section 4.1 and 4.2 before 4.2 saying: ML provides robust feature importance rankings, which are useful for determining the relative importance of each variable. To aid in physical interpretation, we compare the meteorological variables' difference with or without the existence of lightning.

In section 4.1, we provided a systematic method to support previous research to show the importance and rank them via the powerful ML technique. We followed up by addressing the physical mechanism in section 4.2 which complemented the ML based data processing analysis in section 4.1. Previous researches indeed showed the importance of different meteorological variables, but the rank can be difficult to determine given the fact that lightning production is complicated and involving with multiple variables. ML techniques enable us to solve this problem mathematically. In order to make ML more transparent rather than being a black box, section 4.2 elaborated the traditional comparison between non-lightning hours and frequent-lightning hours in section 4.2, so that the internal physical mechanism of lightning is discussed. According to our comparison, our findings are consistent with many previous researches, which provided us with additional evidence to support what we have found in section 4.1 using ML model.

Some issues about data and analysis method are ambiguous. Examples are given in the specific comments.

Specific comments:

Line 10, the sentence 'Machine learning … atmospheric conditions and clouds' is not necessary and can be deleted. 'a site that ..' is not a complete sentence.

Besides telling some variables are important to the lightning occurrence, Abstract better tell how these variables may affect the lightning occurrence (e.g., drier middle troposphere leads to higher lightning frequency).

We have modified some texts of the abstract: Besides the variables considered for the ML models, surface variables and mid-altitude variables (e.g., equivalent potential temperature and minimum equivalent potential temperature, respectively) have statistically significant contrasts between no-lightning and frequent-lightning hours. For example, the minimum equivalent potential temperature from 700 hPa to 500 hPa is significantly lower during frequent-lightning hours than no-lightning.

Lines 40-46, the discussions about cloud depth are too verbose and can be shortened.

We shortened the discussion albeit slightly.

Line 79, it is unclear what are the challenges. Giving some examples can help the readers understand what is new in the manuscript.

The erratic performance is resulted from multiple aspects including mixed-phase convective clouds, limited information of vertical profiles (temperature, dew points). Given the plenty of variables are involved, we need to demonstrate their relative importance before doing more accurate prediction, which is the main question we want to address in this paper.

The challenges and the new findings are added in the last paragraph, "Today, lightning prediction remains challenging because lightning production is stochastic involving microphysical and thermodynamic processes."

Line 90, better give a short description about why the square root of CAPE is discussed.

The reason we use square root of CAPE is discussed in the first paragraph of Section 4.3. Specifically, Holton (1973) found that CAPE plays an important role in determining maximum parcel updraft velocity, which is proportional to $\sqrt{CAPE}$ *based on parcel theory.*

Line 100, how many ENTLN sensors are available in the 1 deg* 1 deg domain? Any reasons for choosing the size?

*ENTLN detects wideband radio waves (1 Hz to 12 Mhz) emitted by lightning. The location is determined by triangulation using results from multiple sensors that can be several hundred kms apart. Because of the long propagation of the waves, the number of ENTLN sensors in 1 deg * 1 deg domain is not important Also, US continent has pretty high density of sensors,*

*ensuring a relatively high detection efficiency. The MERRA-2 (0.625*0.5) and ERA-5 (0.25*0.25) have different spatial resolutions, and we merge them into the same resolution (1*1), then we assume the ARM SGP measurements represent the whole domain of 1*1.*

The data and method sections are not well organized. For example, Section 2.2 talked about the ARM SGP site, but it is unclear what instruments observed the variables of cloud thickness, rain rate and radar echo strength, what are the data spatial and temporal resolutions. I did not get the information until Section 3.3. I suggest rearrange the contents in Sections 2 and 3 for better understanding.

We have added a pointer in Section 2.2 to show that the data processing discussion is in Section 3, so that the structure is improved. "We discuss the detailed processing method in Section 3.3." this pointer has been added at the end of 2.2.

Line 112, delete 'providing several improvements compared with ERA-I' if the authors do not want to tell what the improvements are.

Revised.

Line 119, 'the average value of hourly surface PM2.5', how long is the averaging window?

The average is not temporal. It is simply the average hourly PM2.5 from 3 sites.

Line 128, GPCC.10 grid, is this information necessary here? If not, delete it to avoid misleading.

Remapping is performed to this grid. It is a particular grid and is useful here

Line 161, 0.5 g m-3 or more ice …. In addition, how is the value 0.5 calculated?

*According to Seo and Liu (2005)'s formula. Ice Water Content (IWC) = 0.078*Z^0.79, where Z is the radar reflectivity. When Z = 10, IWC =0.48 g/m^3.*

Line 162, 'less' to 'lower'

Revised.

Line 168, why use the profile at the 30th minute rather than the hourly mean?

We choose to use the 30th minute rather than the hourly mean to save computational cost.

Line 170-175, the meanings of AOSCCN1COL and AOSCCN2COLAVG are unclear. How do they differ between each other?

The AOSCCN1COL data set ended on 2017-09-29 while the AOSCCN2COLAAVG data set started on 2017-04-12. So, we need both of them to get a continuous observation of CCN. The data sets provide two estimates of the same variable. AOSCCN2COLAAVG is Dual Column

with ramping mode averaged, while AOSCCN1COL is only Single Column. Single column measurement means that it uses only one cylinder allowing aerosol flow together with sheath flow to blow in and then detected by optical particle counter. Dual column measurement has two identical cylinders to take the measurement.

Line 179, the word 'homogenous' is not used properly.

Have removed this word.

Line 187-188, the total sample account is 817 with 509 having lightning? It seems too small. How many trainable parameters in the RF and other machine-learning models? Is it possible that the models are overfitted?

The sample size is small, and that's why we have used 10-fold cross-validation, which is a resampling procedure used to evaluate ML models applied to a limited data sample. Also, we repeated the cross validation 50 times. It is not likely to be overfitted because each group is used as both a training set and a test set, which is a big advantage of using the cross-validation method.

Line 189, any reason to choose 8 independent variables? I don't think it is required for the input of RF or any other machine-learning models. Using independent variables for the input does not necessarily yield the best result.

We tried many variables, but only chose those variables because their correlations with respect to each other were less than 0.5. Using multiple mostly independent variables also allows us to answer questions about whether the effect of one variable depends on the level of another. And this is why we are using (mostly) independent variables.

Table 2, how are hyperparameters configurated in these models? What is the criterion for these configurations?

We have tried multiple sets of Random Forest parameters, but found that the default set whose parameters are specified in: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) performs best. The values of the parameters are too numerous to list here. To address your questions, the number of trees in the forest is 100, the maximum depth of the tree is "none", so the nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples, which is the minimum number of samples required to split an internal node. Varying

the number of trees (e.g., 10, 50 and 100 as default), the tree depth (e.g., 5, 10 and "none" as default) and other criteria have also been tested. We have added these information in the manuscript.

Line 209, I don't understand the meanings of 'performed 1000 simulations' and 'ran the Random Forest Classifier to evaluate …'

Revised into a clearer way to express:

The feature importance is a measure of how much one variable decreases the impurity (the presence of more than one class in a subset of data) through various decision trees in the forest. This figure shows the percent of the 1000 simulations that each variable was identified as the most important feature (column #1) to the least important feature (column #8). For example, the variable "Cloud Thickness" was identified as the most important feature in 57.4 % of the 1000 runs, while it is the second (#2), the third (#3) and the fourth (#4) most important feature in 33.4 %, 8.9 % and 0.3 % of the runs.

Line 118-119, 'This figure …', confusing sentence

We modified the sentence to improve clarity. The sentence now reads: This figure shows the percent of the 1000 simulations that each variable was identified as the most important feature (column #1) to the least important feature (column #8). For example, the variable "Cloud Thickness" was identified as the most important feature in 57.4 % of the 1000 runs, while it is the second (#2), the third (#3) and the fourth (#4) most important feature in 33.4 %, 8.9 % and 0.3 % of the runs.

Line 225, how small is the variability of PM2.5? How to define 'small'?

SGP has a really low concentration of aerosols (the average and median value of PM2.5 concentration are 9.1 and 8.2 ug/m^3, the standard deviation value is 5.8 ug/m^3). Thus, by small variability we mean small absolute variability.

Line 225, 'top, middle' to 'most, modest'

Revised.

Figure 2, 800 samples with 1000 splitting, seems overfitted

We used 75 % to 25 % splitting, which means that about 600 samples (75% of the sample size) are used to train the model and the rest are used to test the performance of the model. There

are so many combinations to pick 600 from 800 randomly, so it is not subject to the overfitted problem.

Line 244-245, confusing sentence.

Line 244-254, these sentences are confusing and better be revised.

Yes that section could be improved. This section now reads … To ensure the environment is favourable for lightning, we have set a threshold of CAPE = 2000 J·kg-1 and for this analysis only selected those hours with convective clouds when CAPE is larger than 2000 J·kg-1, given the fact that this threshold of CAPE is considered as a strong convective environment in several studies (Rutledge et al., 1992; Chaudhuri, 2010; Chaudhuri and Middey, 2012; Hu et al., 2019). Overall, there were 175 hours satisfying the CAPE threshold. Of these hours, 41 had no lightning in a three-hour period centered on the CAPE observation and were labelled as no-lightning "hours". Seventy-five of the hours had three-hour mean flash rates exceeding the median flash rate of 162.5 and were classified as frequent-lightning "hours" while the remainder of the hours (59) were deemed intermediate lightning hours.

Table 5, the difference in CCN * PBLH is not statistically significant and thus cannot support the statement in Line 284-285.

Revised to emphasize that the difference of the products is not significant.

Line 300, 'vertically integrated' to 'mean SH from surface to LCL"

Revised.

What is the rational for the analysis in Table 6?

Focusing on strong CAPE environment (> 2000 J/kg), we are examining which variables that we obtain from vertical profiles vary the most between no- and frequent- lightning hours.

Section 4.3, is the vertical velocity observed at the SGP site? If yes, why not use it instead of using CAPE? Can it be used in the RF model?

There is a dataset measuring vertical velocity at the SGP site. We checked this dataset and didn't choose this dataset for two reasons: 1. It is not a quality assured dataset, and is not a recommended dataset by ARM; 2. More importantly, it is limited to low altitude only (less than 4 km), which is not suitable for deep convection.

Line 315, 'hours with plentiful flashes' to 'frequent-lightning hours', be consistent with the definition in Line 250.

Actually, they have different thresholds. 'Hours with plentiful flashes requires that the hourly flash count exceeds the median flash count of 162.5, and the CAPE value is not confined in order to see the relationship between IC fraction and sqrt(CAPE). However, 'frequent-lightning hours' require the mean flash rate during the three-hour period exceed the median hourly flash rate and also require CAPE to exceed 2000 J/kg. Thus, they have different definitions. We have added an additional sentence to emphasis the difference to remind the readers.

Since CAPE is positively correlated with the fraction of IC flashes (Section 4.3) while it does not differ between no-lightning and frequent-lightning hours (Table 4), does it imply that CAPE may be negatively corelated with the fraction of CG flashes?

Yes. The flashes consist of IC and CG, and of course: IC fraction + CG fraction = 1. Thus, a negative correlation is assured.

What are the rational for composite analysis in Section 4.2 and 4.3?

The physical understanding obtained from a machine learning analysis can be limited. Thus, we added section 4.2 to increase physical understanding about what variables are most different between no-lightning and lightning hours. In Section 4.3, we have shown additional interesting results which may intrigue more follow-up research to see how robust this finding is.