Review of egusphere-2024-1718 : “Identifying lightning processes in ERA5 soundings with deep learning”

This manuscript presents a simple deep neural network model which predicts lightning location based on vertical profiles from ERA5 reanalysis data and evaluates its performance compared to a refence model (GAM – generalised additive model). SHAP values are then used to identify which particular features of the input profiles are important in determining the occurrence of lightning. Since the physical processes which generate lightning are not explicitly represented in weather or climate models

The work is a nice demonstration of the use of machine learning for both prediction and process understanding, although the motivation of the paper and the novelty of the findings is not well highlighted in the current manuscript. There are also a number of aspects which need clearer description and some of the results either require more in depth analysis, or just removing. Overall, the manuscript needs revision before it could be considered for publication. I include more specific comments below.

Major comments:

1) The rationale for the work is not completely clear, in particular I think you could do a better job of identifying the novelty of the research compared to previous studies. What new do you learn here compared to previous studies by the co-authors? Or is the novelty really in the method? Is it better (figure 1 / table 3 suggests it is slightly better in some cases, but not hugely – it still misses > half of cases of lightning, but no lightning was seen in about 80% of the modelled “yes” cases). I would be clear about what you are aiming to do up front. I wasn’t clear what you meant by a “holistic description of lightning” (top of p3). The physical insight offered by explainable AI could be really interesting, but I didn’t feel you really took this very far in terms of discussing your results, certainly not much beyond confirming what we already know.
2) A number of decisions are made (e.g. choice of ML method, choice of model variables to include) without a clear justification. I think the variables you have chosen are useful, but there are other things you have neglected (e.g. the height / pressure on the model levels and surface fields) which could be relevant. In particular, as I understand it you have chosen a method which does not know about the links between adjacent levels. Does this mean the model does not know about derivatives? This could be particularly important when thinking about things like stability. I wasn’t quite clear what “topography” means. Is this just the height of the grid cell? What about some measure of slope / sub grid variability? Would this not be relevant too for convection?
3) Another example of choices is the various parameters related to the training (e.g. dropout, early stopping patience) which are just given without justification.
4) I appreciate that there is a formal requirement for inputs to be independent for SHAP values to be calculated using Deep SHAP. Since this is clearly not the case here, I think you need to be more careful in explaining why it is ok to use this algorithm, but also more fundamentally explain what the SHAP values will tell you when the variables are highly correlated (as adjacent points in a profile are likely to be). Looking at some of your results they are very noisy (e.g. CIWC in figure 2). Is this “real” or is this a feature of the implementation of SHAP you are using?

5) Training / validating / test data sets. As I read the paper you use 2010-2018 for training the model and for validating the model (tuning and preventing overfitting). How do you split this data between training and validating? Then 2019 is reserved as a truly independent test data set for evaluating the overall model. Please be clear on this at the start.
6) Subdomains: having described the training of the model and the calculation of an appropriate threshold for producing a binary lighting / no lightning output, you then go on to say that you subdivide the data into four subdomains. There is no clear description of where these subdomains are. If you use them then I think you need to say how they are defined / include a map. Do you retune the output threshold for each subdomain separately? I am not quite clear from the manuscript. If so, why is this necessary and what difference does it make? It would seem to limit the universality of the method if you need a different threshold for different regions. It also makes the precise choice of region quite important (and another arbitrary decision).
7) Actually, re-reading you say "In this case, the model's threshold is calibrated to align the average predicted and observed lightning frequencies of the validation set". This is very different to what you did on the previous page and makes the results of figure 1 look much better. This is very misleading. Please explain clearly why you need to calculate thresholds in two different ways and avoid doing so if it is at all possible.
8) How do you calculate model confidence? A number of figures show categories split into less confident / very confident and it is mentioned in the text, but you don't actually explain how you calculate this? Is it also based on the threshold?
9) On p8 you categorise the data based on the most important group of features (cloud, mass, wind). Identification of a lightning prediction only requires the sum of the scaled SHAP values to exceed 1, so in theory all three subgroups could exceed a summed SHAP value of >0.5. What do you do in that case? What about the case where all 3 groups are less than <0.5 but sum to >1.0? What are the relative frequencies of occurrence of these different categories? Why later do you go and split cloud into cloud-wind and cloud-mass?
10) Figures and use of colour. I found it hard to read many of the figures. E.g. in figure 2 I could not see the pale green mean line for "TP less confident". I also really struggled to see the boundary where pale and dark green shading overlapped. Please consider whether it might be better to use e.g. dashed lines to mark the boundaries of the shaded regions. Similarly in figures 3-5. Also consider the choice of colours in figures 3-5. For those who are colour blind it would be impossible to distinguish these overlapping colours. Red/green is a particularly bad combination for many people.
11) Figure 6 – spatial distribution by category. I wasn't quite sure what the take-home message here was. There is only a very short paragraph which describes the figure but does not really discuss the results at all. Either this needs more analysis, or if it doesn't add anything just remove the figure.
12) Case study (figure 7). Again, I wasn't quite sure what I was supposed to take away from looking at a single case study. The model gets some bits right but tends to predict lightning over too wide an area. Without any meteorological context it is hard to know why. Is this a general result? This case study either needs better justification and more discussion if there is an important result to take from it, or otherwise just remove it.
13) In the introduction you mentioned the potential for using ML models for parametrisation. This would require the model to be generalisable. I wonder if you can revisit this in the discussion and conclusions. How widely applicable is your model? It seems to respond to different features in different regions. Have you tried applying to other unseen regions /

seasons? What would you need to do to make it more generalisable? At least this is worth discussing.

Minor comments:

1) Abstract, line 2. “wind shears” -> “wind shear”
2) Abstract, line 13. “as physically meaningful” -> “as a physically meaningful”
3) p2, lines 4-5. The phrase “numerical computations” sounds odd. How about something like “The term *proxy* is commonly used for quantities derived from model output *after* the simulations has run. *Parametrizations* diagnose lightning *while* the model is running and hence can feed back on the simulation.”
4) p2, line 8. “perform reasonably good” -> “perform reasonably well”
5) p3, 3rd paragraph. I found the sentence ending with “(Sect 3)” a bit confusing. It is ambiguous what “(Sect 3)” refers to. Perhaps delete and then change the following sentence to read “Section 3 describes the two modelling approaches and additionally illustrates …” This makes it clearer what each section is about.
6) Table 1. Units should be in roman not italic font by convention.
7) p5, section 3.1, lines 6-7. “.. are standardized by considering the 74 levels altogether, prior training”. I do not understand what this means. Please explain.
8) P6, line 1. I guess you are using the implementation of DeepSHAP from Lunberg available on GitHub? If so, perhaps say so explicitly / provide a reference. The footnote on p8 mentions *DeepExplainer*.
9) Figure 2 caption “The coloured areas highlighted the 50% quantiles.” This is a bit unclear. Do you mean the shaded are shows the interquartile ranges (25%-75%)? If so, is this of the mean SHAP values calculated at each cell point, or is it the interquartile range of all SHAP values across all cells?
10) p13, line 6. “cyrstals” -> “crystals”
11) p15, line 12. “MGT-I” -> “MTG-I”