Reply to anonymous Reviewer 1

> 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.

We appreciate and want to thank for your exceptionally careful and detailed review. We have revised the manuscript based on your comments and suggestions and are confident that it has significantly improved in clarity and overall quality.

Major comments

MajC1

> 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.

Thank you for highlighting the need to emphasize the novelty and our objectives more clearly. We have thoroughly revised the abstract, Section 1 (Introduction), Section 4.1 (Performance of the Deep Learning Approach), Section 5 (Discussion and Conclusions) to enhance clarity. In summary, our work directly utilizes raw model-level data instead of relying on expert-selected variables derived from vertical profiles. This on one hand makes the methodology easier transferable to other regions with less expert knowledge and on the other hand showcases the capability of AI to extract meaningful patterns. Due to the high number of correlated input features, commonly used plots for visualizing SHAP values are not feasible for interpretation. Therefore, we aggregated the results for a more global understanding, introduced scaled SHAP values for improved explainability, and visualized the median and quantiles of the results as vertical profiles to aid interpretation.

MajC2

> 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?

Thanks for pointing out that we missed to explain the topography variable. Like you assume, it is the geopotential height at model level 137 (adjacent to surface) of the given grid cell. We have added a footnote. The choice for the Neural Network was made (among other reasons) because it is able to handle a large number of input features, is comparably fast to train and performs well in many complex classification tasks. We included a note about the number of input features in the first paragraph of Section 3 (Methods).

We did not include surface fields, since we solely work with "raw" model level data. This sets our study apart from other studies. Pressure on the model level would indeed be an interesting additional variable, which we will consider for future research. In this study, we had to keep the number of overall parameters low to ensure that computing time and memory usage remain managable.

While slope, sub grid variability and similar properties would be important for physical modelling, the neural network will basically learn to "fingerprint". We provided longitude and latitude as input, thus the model is able to grasp that different "atmospheric patterns" are important at different locations (aka sub-grid topography). Also, since we are using a fully connected neural network which basically allows for all linear combinations of any inputs, the model is able to access the derivatives by learning the parameters (weights and biases) accordingly.

MajC3

> Another example of choices is the various parameters related to the training (e.g. dropout, early stopping patience) which are just given without justification.

The justification for adding dropout and early stopping patience is to prevent the model from overfitting. We decided not to add more description regarding the concrete choices of parameters, as these are commonly known best practices in the machine learning community, and the specific parameter values are within the commonly used ranges. However, we would like to provide a more detailed explanation here:

- Hidden nodes per layer: The first hidden layers approximately match the size of the input layer to capture the full complexity initially. The last hidden layers are smaller to save computational power.

- Number of layers: Sufficiently deep to fit more complex patterns.

- Leaky ReLU: Computationally fast; the gradient does not vanish with negative input.

- Sigmoid function on output layer: Ensures the output is between 0 and 1 and is commonly used for binary classification.

- Mean-standard scaling: Input needs scaling. We have experimented with mean-standard and min-max scaling; the former performed notably better.

- Dropout: Used to prevent overfitting, especially with highly imbalanced data. We experimented with values between 0.1 and 0.6. The specific value did not significantly affect performance, but the model benefited from using dropout.

- Early stopping patience: Due to the vast amount of data, the model converges in a few epochs. Each epoch took around one hour. Early stopping ensures that training continues as long as performance improves and stops once it plateaus, preventing overfitting.

- Binary cross-entropy loss function: Commonly used for binary classification problems. Weighting positive events proportionally to their relative occurrence is best practice for imbalanced classification tasks.

While these being the final parameter choices, we of course experimented with number of layers, number of nodes, different activation functions and other parameters to ensure that the model's training is stable and performs well.

MajC4

> 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?

The assumption that inputs are independent, the consequences when this condition is not met and which approach for sampling Shapley values should be the preferred one, are sources of many debates in the literature [1, 2].

Scott Lundberg (creator of the SHAP library) writes in a discussion [3]: __________ The original SHAP paper proposed pure conditional expectations for measuring the value of a set of input features, and then proposed using the Shapley values to reduce this exponential number of values down to a single number for each feature. To make things more tractable we can assume feature independence. This is of course never true in practice, and so may seem like a terrible approximation. But it turns out that you can look at this assumption from a very different perspective, where you break feature dependence not because of an independence assumption, but because of arguments based on causal inference. __________

To summarize: There are two main approaches to approximate Shapley values which fundamently differ in the way they sample left-out (dropped) features to account for feature attribution. The interventional approach (used by Deep SHAP, which we employed in our work) treats inputs as independent and thereby identifying which inputs are genuinely used by the model. In the case of correlated inputs the trained model might not give equal contribution to all correlated variables and since the interventional Shapley values are "true to the model", also the Shapley values will only reflect the input variables the model actually uses. In the context of our work, we believe that using Deep SHAP is the right choice. We agree that the current paragraph in the manuscript is rather confusing and thus will replace the second and third paragraph of Section 3.3 (Explainability) with a clearer argumentation. Following up on the previous explanations, the noise observed in CIWC in Figure 2 is indeed real to the model. This figure indicates that the model relies more on the model levels of CIWC where SHAP peaks compared to model levels with no or lower peaks.

[1] Chen, Hugh, et al. "True to the model or true to the data?." arXiv preprint arXiv:2006.16234 (2020).

[2] Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." International Conference on artificial intelligence and statistics. PMLR, 2020.

[3] christophM. discussion. GitHub. https://github.com/christophM/interpretable-ml-book/issues/142#issuecomment-564681746

MajC5

> 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.

Data is split based on distinct days. 20% of these distinct days are used for validation, while the remaining 80% serve as training dataset. We will add this information in the revised version (second paragraph in Section 2 (Data)).

MajC6

> 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).

Division in subdomains was only needed to illustrate the effect of two different thresholding methods (based on F1 score or calibrated on the actual number of lightning events) on the diurnal cycle of lightning. These subdomains had been chosen for differences in the diurnal cycle due to topographic differences between them. Since we have deleted the comparison in Section 4.1 (Performance of the deep learning approach) between the two methods as too much of a tangent topic in response to your questioning its usefulness, subdomains are no longer needed. We want to note that we did not retune the output threshold based on subdomains or locations, as doing so would limit the universality of this method, as you pointed out.

MajC7

> 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.

We acknowledge that introducing a second threshold to discuss a tangent topic is rather confusing. As mentioned in our response to point 6, we have removed this comparison altogether.

MajC8

> 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?

Like you assume, True Positives are split into “less confident” and “very confident” based on threshold phi. True Positives are given by model outputs greater than phi. The True Positive category is further subdivided into less and very confident True Positives. Less confident True Positives are given by correctly classified samples with a model output smaller than (1 + phi) / 2 and very confident True Positives by outputs greater than or equal to (1 + phi) / 2. We have added this information to the third paragraph of Section 4.2 (Identifying patterns exploited by the deep learning model) of the revised manuscript.

MajC9

> 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?

A single sample can be attributed to one, multiple or even none of the three categories. Approximately 39.8% of the True Positives belong to the cloud-dominant, 2.6% to the mass-dominant and 7.9% to the wind-dominant class. We have supplemented the manuscript (Section 4.2 (Identifying patterns exploited by the deep learning model)) with this information. A more comprehensive list is provided here but not added to the paper to avoid bloating the manuscript:

- Number of samples in cloud-dominant TPs: 5726 (39.8% of TPs)
- Number of samples in mass-dominant TPs: 380 (2.6% of TPs)
- Number of samples in wind-dominant TPs: 1133 (7.9% of TPs)
- Number of samples TPs without dominance: 7331 (51% of TPs)
- Number of samples being cloud and mass dominant at the same time: 7
- Number of samples being cloud and wind dominant at the same time: 191
- Number of samples being cloud, wind, and mass dominant at the same time: 0
- Number of samples being mass and wind dominant at the same time: 0

Regarding the split of cloud into cloud-wind and cloud-mass: I apologize for having overlooked this part of your question. I have not yet had the opportunity to discuss it with the co-authors, some of whom are on holiday until the end of September. We will address this question as soon as we have the chance to discuss it.

MajC10

> 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.

Thank you for suggesting the use of dashed lines to mark the boundaries. We have adapted the plots accordingly. Additionally, we removed the shaded areas as they bloated the graphics and slightly affected the colors due to the transparency settings. We also removed Figure 4a (old numbering), since it was quite hard to read and also did not add a lot of value. We used http://hclwizard.org:3000/cvdemulator/ to check on the colors and they should be suitable for color-blind persons.

MajC11

> 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.

The idea was to demonstrate the model’s ability to pick up on regional differences, but we agree that the figure itself does not add a lot of value. Thus we removed the figure in the revised version.

MajC12

> 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.

We have added a more detailed explanation and thoroughly revised Section 4.3 (Sample case study), and added paragraph 4 in Section 5 (Discussion and Conclusions).

The reason why we added the case study is because researchers interested in applying the method are usually also interested in seeing concrete examples / case studies. If desired we could also move this part into the Appendix.

MajC13

> 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.

This is a very interesting thought. We actually performed experiments on generalisation and presented the results in a poster session at the EGU 23 [4] (the poster itself is downloadable as supplement material). To summarize: We have trained a model on the same area using the summer months, but without longitude/latitude and without the day of the year. We then used this model to classify cells with lightning activity on a much bigger area (Continental Europe). The model clearly learned the patterns on landcovered areas, but like we also observe in the case study, overestimates lightning activity. This is mainly due to the choice of threshold and the difficulty of accurately finding the spatial and temporal extend of the lightning event. We have also added this discussion to the manuscript (4th paragraph in Section 5 (Discussion and Conclusions).

[4] G. Ehrensperger, T. Hell, G. J. Mayr, and T. Simon, “Evaluating the generalization ability of a deep learning model trained to detect cloud-to-ground lightning on raw ERA5 data,” Copernicus Meetings, Feb. 2023. doi: 10.5194/egusphere-egu23-15817.

Minor comments

MinC1, MinC2, MinC4, MinC10, MinC11

> 1. Abstract, line 2. “wind shears” -> “wind shear”
>
> 2. Abstract, line 13. “as physically meaningful” -> “as a physically meaningful”
>
> 3. p2, line 8. “perform reasonably good” -> “perform reasonably well”
>
> 4. p13, line 6. “cyrstals” -> “crystals”
>
> 5. p15, line 12. “MGT-I” -> “MTG-I”

Thanks for the list of typos. We have corrected the manuscript accordingly.

MinC3

> 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.”

Thanks. We followed your suggestion.

MinC5

> 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.

Thanks. This will be updated in the revised version.

MinC6

> Table 1. Units should be in roman not italic font by convention.

Thanks. This is fixed.

MinC7

> 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.

We including additional details on the standardization process in the revised manuscript (2nd paragraph of Section 3.1 (Deep Learning Approach)).

MinC8

> 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.

We have already referred to Deep SHAP and cited Lundberg (2017) in line 2 of section 3.3. However, we have revised parts of the text and the footnote to improve readability.

MinC9

> 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?

Sorry for the confusion. Like you suspect, the shaded areas show the interquartile range (25%-75%) of the data. We corrected this in the revised version. Regarding your second question it is the second option (interquartile range of scaled SHAP values across all cells).