Answer to Reviewer 2 comments

1 Acknowledgments

We sincerely thank the editor and the three reviewers for their careful reading of our manuscript and for their constructive and insightful comments.

We thank Reviewer 2 for their detailed review, with comments ranging from the overall study design to the code and the bibliography. Their concerns about the fairness of the comparison led us to tune the MLP and XGB with the same nested procedure and objective as the U-Net, to treat the training sample size of the LR, GAM and XGB as a hyperparameter, and to compare the calibration of all models with and without Platt scaling. Their questions also led us to rename our feature importance values to Shapley population variable importance measure (SPVIM), to justify our tuning objective, and to state the limitations related to climate change, to the varying grid-cell area, and to the location information available to the U-Net. Finally, their question on the random values used in the perturbation method led us to find and correct an error in that analysis.

We have revised the manuscript thoroughly to address every point raised. Below we provide a point-by-point response to each comment, indicating which sections of the revised manuscript (main.tex) have been modified. The Reviewer's comments are in italic and prefixed by "LX-Y:"; our responses follow each comment, together with a short summary of the corresponding changes implemented in the manuscript. The lines in the "CHANGES" paragraphs correspond to the revised version of the manuscript.

2 General comment and main changes

This preprint presents a comparison of different machine learning models to predict lightning occurrence across Europe using ERA5 reanalysis data and lightning data from the Met Office. To capture the mesoscale atmospheric patterns that drive convection, the authors employ a spatially-aware U-Net architecture. The performance of this specialized neural network (originating from computer vision tasks) is evaluated against several single-grid baseline models, including Logistic Regression, GAM, XGBoost, and an MLP. Based on a 16-year dataset (2008–2023), the authors conclude that the U-Net significantly outperforms the baseline models, and they employ SHAP values and spatial perturbations to interpret the features and neighborhood sizes driving these predictions. The paper is well-motivated and well-written. It provides valuable insights and tackles an important, relevant topic.

Summary of the Review

However, there are a few critical methodological and structural concerns regarding the manuscript. The primary issues include an uneven baseline comparison that likely handicaps the single-grid models through a lack of spatial/geographic features, alongside a mismatch in probability calibration and in case of GAM/XGB also by severe data subsampling (0.1%). Additionally, the unconventional aggregation of SHAP values over evaluation metrics needs to be justified.

In general, it seems like at a few places "less would be more." For example, five metrics are summed for the hyper parameter tuning, instead of choosing a primary one to optimize against. Furthermore, the training procedure and hyperparameter tuning of the U-Net are quite sophisticated, while little is said about how the baseline models were built. This doesn't invalidate the results, but it makes it difficult to determine whether the performance gap stems from the U-Net's inherent architectural advantages or simply from an imbalance in optimization effort.

Minor but necessary corrections regarding code availability and ERA5 grid distortion (a 0.25° x 0.25° cell covers a different physical spatial extent in Southern Europe compared to Northern Europe) should also be addressed.
While I mark a couple of points as "Major," I am confident they can be successfully addressed by the authors.

ANSWER: We thank the reviewer for this thorough and constructive review and for their positive assessment of the manuscripts. Besides the point-by-point answers below, we summarize here the main changes between the original and the revised manuscript. Following the reviewers' comments, we re-optimized the single grid cell models: the MLP and XGB are now tuned with the same nested cross-validation procedure and objective as the U-Net, the training sample size of the LR, GAM and XGB is a tuned hyperparameter, and XGB is trained with the binary logistic objective. The scores of these models therefore differ slightly from the original version, but the U-Net still outperforms all of them for every metric, and our conclusions are unchanged, although we now clearly state that our methodology does not allow to isolate the individual impacts of location knowledge and spatial context awareness. The confidence intervals and statistical tests now better account for the overlap of the training sets between the leave-one-year-out splits, the feature importance values are renamed Shapley population variable importance measure (SPVIM), and the corrected radial perturbation analysis gives a radius of useful information of about 500 km instead of 320 km (Section 4.6). Appendix A now gives the tuned hyperparameters, optimization method and loss function of every model, and the model schematic (Figure 2) has been updated. We also added several analyses. In the main text, Section 4.3 now discusses the four seasons and the regional biases of each model (with the regions defined in Appendix E), and Section 4.4 separates the SPVIM values into individual and interaction effects. The supplementary material has new sections on the performance diagram (Section S1.1), overfitting (Section S1.2), the calibration of all models (Section S1.3), the seasonal climatology with relative error maps and land and sea cycles (Section S1.4), the separation of the SPVIM values (Section S3.2), the spatial structure of the predictors (Section S3.3) and the variability of the saliency maps (Figure S19).

3 Point by point response to Reviewer 2

Legend for the following concerns

[c] comment

[m] minor

[M] major

[q] question

[r] rephrase

[s] suggestion

[t] typo

Detailed Comments

hyperlinks (L49, L69–70, L127)

> **ANSWER:** The citations are hyperlinks in our source files; the links were lost when the journal typeset the preprint. We will check that they work in the revised version.

L85, L312–313, L567, L573: code archive.

> **ANSWER:** We thank the reviewer for their suggestion. The code and a data sample are now archived on Zenodo. Every link to the GitHub repository is replaced by a citation to this archive (Burq et al., 2026): in the introduction, Section 4.5, and the code and data availability statements. We also added a .python-version file.

L24-27, L409, L436, L481: typos.

> **ANSWER:** They are now corrected.

L552, L556 : formula formatting.

> **ANSWER:** We have applied the suggestion : `\operatorname` is now used for BS, J, Dice, FBSS and Deviance, and `\text` for the subscripts.

L45–53: [r] Be more precise about "similar tasks" in L50. I understand the message: former work often uses the atmospheric conditions at one specific location (without its surroundings) to model the probability of a severe weather event (hail, lightning, etc.) at that specific location. However, this section interchangeably uses "lightning" and "severe weather events in a broader sense," making it hard to follow whether the given references are about lightning specifically or other phenomena. I suggest first zooming out to discuss literature on convective weather events in general, and then zooming in to specify which parts of the literature cover lightning.

ANSWER: We agree that this paragraph used "lightning" and "convective hazards" interchangeably. The studies cited model lightning, hail or tornadoes with the same single grid cell approach, so the paragraph now describes this approach for convective hazards in general. Studies specific to lightning are discussed in the following paragraphs.

CHANGES: Line 42: "[...] to model the occurrence of convective hazard [...]"
Line 47: "[...] the output is a probability of hazard occurrence [...]"
Line 49: "[...] to predict the occurrence of convective hazards."

L56: [r] Be more precise. ERA5 has a resolution of 0.25° x 0.25°, which is less than 800 km^2 near the equator. Moving toward the poles, the covered area (in terms of square kilometers) gets even smaller.

ANSWER: The sentence was meant to give an order of magnitude. It now gives the actual area of an ERA5 grid cell over our domain: from about 400 km^2 at 60°N to 600 km^2 at 35°N, against about 100 km^2 for an individual convective cell.

CHANGES: Line 55-56: "(between 400 and 600 km2 in the case of ERA5 over Europe)."

L57–58: [m] Can you provide a reference for this?

ANSWER: The sentence was unclear and has been reformulated. Individual convective cells have a horizontal scale of about 10–30 km, and when they organize into clusters they reach the mesoscale (about 100–300 km) (Houze, 2004). Our hypothesis is that giving this mesoscale information to the model helps predict these systems.

CHANGES: Lines 57-58: "When convective systems (with typical horizontal scale $\sim$ 10 to 30 km) organize in clusters they reach the synoptic mesoscale ($\sim$ 100 to 300 km) (Houze, 2004)."

L88: [q] Should this be "hybrid sigma-pressure levels"?

> **ANSWER:** Both terms refer to the same vertical coordinate, but "hybrid sigma-pressure levels" is the more precise one and is now used.

L102: [m] Your study domain is Europe. A 0.25° horizontal resolution does not correspond to ∼30 km (at least for longitude, this varies quite a lot across Europe).

> **ANSWER:** The text now gives the grid spacing in both directions: about 28 km along a meridian, and between 14 km (60°N) and 23 km (35°N) along a parallel.
>
> **CHANGES**: Lines 107-108: "[...] which corresponds to about 28 km between two points along a meridian and between 14 and 23 km along a parallel."

L116–117: [m] Where is the land-sea mask taken from? Are (larger) lakes marked as land? What is the threshold ratio for a grid cell to be marked as land or sea?

> **ANSWER:** The land-sea mask is the ERA5 land-sea mask. It is not binary: it gives the fraction of land in each grid cell, from 0 (water only) to 1 (land only). Lakes and large rivers count as water, so the cells that contain them take intermediate values (for example, about 0.4 over Lake Geneva). Over our domain, 40% of the cells have the value 0, 16% have the value 1 and 44% have a value in between. The models use this continuous field without a threshold. A 0.5 threshold is only used to define the land and sea regions of the analysis (Appendix E).
>
> **CHANGES**: Lines 116-117: "We also use two other variables downloaded directly from the ERA5 reanalysis dataset: (iv) convective precipitation and (v) a land-sea mask."

L151: [m] Be more specific about the weights. How is the weight of a sample chosen? I see that you add more information about this in Appendix B, but a brief note on this should be included in the main text.

> **ANSWER:** Section 3.2 now states that w is a hyperparameter tuned in the inner loop of the LOYO procedure. Its search range and selected values are given in Tables A4 (MLP) and A5 (U-Net), and Appendix B describes the training.

CHANGES: Line 167: "w is a hyperparameter that is adjusted during the optimization phase."

L164: [m] Add a reference for the autocorrelation claim.

ANSWER: We added references for this claim.

CHANGES: Lines 182-184: ". Leinonen et al. (2022b) found that randomly splitting thunderstorm samples led to overfitting and therefore assigned each day to a single subset to account for the autocorrelation of thunderstorm environments. Under persistent blocking, favorable conditions can even last one to three weeks (Piper et al., 2016; Mohr et al., 2020)"

L185: [m] I did not find the definition of FBSS in Roberts and Lean (2008). I cannot even find the term "Brier" in this reference. Are you referring to FSS? If so, it appears to be based on MSE (not RMSE, as stated in Table 1). Please add more information or refer to your Appendix D.

ANSWER: We thank the reviewer for this comment. Roberts and Lean (2008) indeed define the fractions skill score (FSS, their Eq. 6) as:

$$\mathrm{FSS} = 1 - \frac{\mathrm{MSE}_{\mathrm{model,\,neighborhood}}}{\mathrm{MSE}_{\mathrm{ref,\,neighborhood}}}$$

The two differences in our FBSS score are :

- our reference is the climatology (instead of the worst-case model as in Roberts and Lean (2008));
- we use continuous forecast fields instead of binary fields. This is exactly how the Brier score (Brier, 1950) generalizes the MSE to classification tasks.

So our score is :

$$\mathrm{FBSS} = 1 - \frac{\mathrm{Brier}_{\mathrm{model,\,neighborhood}}}{\mathrm{Brier}_{\mathrm{clim,\,neighborhood}}}$$

The reviewer is also right that the score is based on MSE and not RMSE. We have corrected this typo in Table 1 and have clarified the FBSS definition in Section 3.4 and Appendix D.

CHANGES: Lines 208-209: "[...] the Fractions Brier Skill Score (FBSS) – which measures the Brier score (Brier, 1950), averaged over a neighborhood (Roberts and Lean, 2008)"

Lines 657-662: "The Fractions Skill Score (Roberts and Lean, 2008) is defined as $\mathrm{FSS} = 1 - \frac{\mathrm{MSE}_{\mathrm{model,\,neighborhood}}}{\mathrm{MSE}_{\mathrm{ref,\,neighborhood}}}$. In Roberts and Lean (2008), the MSE is computed between two binary fields: the observations and the prediction, and the reference is the worst-case scenario where the two fields never overlap.

The Brier Score (Brier, 1950) extends the mean squared error (MSE) to classification tasks by measuring the mean squared error between the predicted probabilities and the true binary labels: $\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$. We use this to avoid thresholding the predictions."

Line 666: "$\mathrm{FBSS} = 1 - \frac{\mathrm{BS}_{\mathrm{model,\,neighborhood}}}{\mathrm{BS}_{\mathrm{ref,\,neighborhood}}}$"

L190: [r] The formula's formatting looks odd.

ANSWER: We have edited the formating of the formula.

L211: [q] Which set the 49 days refer to?

ANSWER: These are the extreme days of the three held-out years (2008, 2015 and 2023), which are used to test the final models. Section 4.1 now defines the extreme days first and then gives their number: between 6 and 35 in each LOYO test year, and 49 in the three held-out years.

CHANGES: Line 233: "There are between 6 and 35 extreme days in each LOYO test set."

Line 246: "There are 49 extreme days in the 3 held-out years."

L234: [r] "80%" is probably not exact; this should be "Approx. 80%" (or something similar).

ANSWER: We now give the exact value: 93% of the samples have a predicted probability below 0.01, which is the first bin of Figure 4a.

CHANGES: Line 256: "93% of all samples are in the first bin."

L230–235: [q] Do you have an intuition for the huge spike in the purple line near a 1.0 predicted probability? This should be addressed in the text.

ANSWER: The spike came from the bins with high predicted probabilities, which contained very few samples, so their observed frequency was very noisy. Figure 4 now uses adaptive bin sizes that each contain at least 50 samples, and the spike is gone.

CHANGES: Figure 4 caption: "To ensure each point is statistically meaningful, we use adaptive bin sizes such that every bin contains at least 50 samples."

L236–241: [M] I am not fully convinced by this comparison, as you compare plain model outputs (LR/GAM/XGB) to the calibrated curve of U-Net/MLP. A calibration layer is not a new invention and could be applied to any of those baseline models. It would be fairer to compare all models either with or without a calibration layer. Comparing the uncalibrated U-Net/MLP curves to the uncalibrated curves of the other models is much less impressive. State clearly which models are recalibrated and either calibrate all or not. Sidenote: When applying Platt scaling to every model, keep in mind that if a model is already perfectly calibrated, Platt scaling would worsen the calibration since it does not include the identity function.

ANSWER: We agree that the manuscript must state clearly which models are recalibrated and why. Only the MLP and the U-Net use Platt scaling, because they are trained with a weighted loss.

- Reason why these two models need it: weighting the positive samples by w shifts the optimal logit by ln w. The raw output of these models is therefore the probability of lightning under a distribution in which lightning is w times more frequent. This raw output overestimates the observed frequency by a factor of 7 to 8.5. Platt scaling corrects this, and the fitted intercepts match $b \approx -\ln w$. The calibration layer is therefore part of these two models, and their uncalibrated output cannot be read as a probability of lightning.
- Why the other models don't: LR, GAM and XGB minimize the unweighted log-loss, which is a proper scoring rule, so they are calibrated by construction. To test this, we fitted Platt scaling to LR, GAM and XGB in each of the 13 LOYO splits. The coefficients are close to the identity (a between 0.97 and 0.99, b between -0.09 and 0). The calibration curves with and without scaling overlap, and the ratio of predicted to observed lightning hours changes by less than 0.03. We therefore keep these three models without scaling.

Section 3.2 now states this, and a new section of the supplementary material (Section S1.3 "Re-calibration across models") reports the comparison. Regarding the reviewer's sidenote: our Platt scaling acts on the logit (there was a typo in the manuscript that indicated "probability" instead of "logit") : $\sigma(az + b)$, so it includes the identity (a = 1, b = 0).

CHANGES: Line 171: "[...] which fits a logistic regression where the input is the logit predicted [...]"

Lines 172-174: "Note that we do not apply this scaling to the other single grid cell models since they are not trained with a weighted loss function. Therefore they are by construction already calibrated – see Section S1.3 of the supplementary material for more details."

L243: [q] Are 16 years enough to call this a climatology?

ANSWER: We agree that 16 years is short for a climatology. The text now calls these quantities "statistics of lightning occurrence over 2008–2023", and says that we refer to them as "climatology" in the rest of the paper for simplicity.

CHANGES: Line 265: "We present here the statistics of lightning occurrence over the 2008-2023 period – which we will refer to as climatology"

Figure 5: [q] Are these graphs showing the results on the test set? Because otherwise the graphs could (in theory) simply show that the U-Net is overfitting on the training data.

ANSWER: Yes, these are test-set results. Each year is predicted by the LOYO model for which it is the test year, that is, a model that did not see this year during training or hyperparameter tuning.

L274–285: [M] This appears to be a highly unusual way to use/calculate SHAP values and requires a strong explanation and/or reference. Why not just look at individual events and aggregate over them?

ANSWER: We thank the reviewer for this very relevant comment. Calling these quantities "SHAP values" was a misuse of the term, and we have revised Section 4.4 accordingly. What we do is in fact the Shapley population variable importance measure (SPVIM) defined by Williamson and Feng (2020). With five predictors, all $2^5 = 32$ subsets are evaluated, so the values are exact. Our use differs from Williamson and Feng (2020) in two ways: the values are computed separately for each architecture, and the evaluation scores differ from the training loss. We explain scores rather than individual predictions because the question we address is how much skill each predictor brings, which is how predictors are usually selected in convective-environment modeling. We did not aggregate local SHAP values over events, for two reasons:

1. Four of our five scores do not split into per-sample terms. The AUCs depend on the ranking of all predictions, Dice is a ratio of sums over the domain, and FBSS uses neighborhood fractions. Aggregated local SHAP values would measure how much a predictor moves the predicted probability, not how much skill it adds.
2. Local attributions would not be comparable across architectures. For the single grid cell models they apply to five scalar inputs; for the U-Net they apply to every input pixel.

We replaced the "SHAP" mentions by "SPVIM" throughout (Section 4.4, figure captions, conclusion and carbon footprint statement). We also added the definition of the SPVIM value (Eq. 1), the corresponding

references (Williamson and Feng (2020) and Lipovetsky and Conklin (2001)), and the two differences from Williamson and Feng (2020).

CHANGES: Lines 308-320: "We assess here to what extent having each feature as input to the model changes the performance of the model for each evaluation metric. To do this, we compute the Shapley population variable importance measure (SPVIM) (Williamson and Feng, 2020) separately for each model and evaluated exactly on all 25 = 32 subsets of the 5 features. Therefore, we train 32 instances of each model (LR, GAM, XGB, MLP and U-Net) with the same architecture but different input predictors, yielding one SPVIM value per feature, model and score function. We compute the SPVIM for each score function (PR-AUC, ROC-AUC, Dice, FBSS and Deviance), such that the values describe the importance of a predictor for a given model family and a given score. This extends what the Shapley Net Effects do for linear regression (Lipovetsky and Conklin, 2001). Our implementation of SPVIM differs slightly from Williamson and Feng (2020): the score functions on which we perform the evaluation differ from the loss function used to train the models. Let N be the set of the n = 5 predictors and, for T $\subseteq$ N , define t = |T | and let v(T) be the test-set score (PR-AUC, ROC-AUC, Dice, FBSS or Deviance) of a model of a given architecture trained with only the predictors in T as input. The importance of predictor i is defined as:

$$\phi_{\mathrm{Sh}}(i) = \sum_{T \subseteq N \setminus \{i\}} \frac{(n-t-1)!\,t!}{n!} \big[v(T \cup \{i\}) - v(T) \big], \qquad i \in N.\text{"} \tag{1}$$

L338: [r] I disagree with "is expected." I would not expect this per se, since the U-Net has skip connections (I assume this is what you mean by "residual connection") and would therefore be able to "see" small-scale patterns as well. However, if it happens (as it does in your study), I agree with your interpretation. I would assume the smooth appearance of your graph indicates instead that the problem the model has to solve is highly difficult.

ANSWER: We agree that "is expected" was too quick. Our U-Net has skip connections between the encoder and the decoder, and residual layers within each block (Appendix A). Both give the decoder access to high-resolution features, so the architecture does not by itself force smooth outputs. We agree with the reviewer that the smoothness reflects how difficult the task is. The 0.25° ERA5 predictors, in which convection is parameterized, only weakly constrain the exact location of convective cells. A model trained with a pixel-wise loss is therefore rewarded for spreading probability over all plausible locations, rather than for predicting a sharp pattern that may be misplaced. The end of the paragraph has been rewritten accordingly (Section 4.5).

CHANGES: Lines 391-399: "On the other hand, the small-scale features – which are associated with large wavenumbers on the power spectrum – are not well captured by any model, the U-Net being worse than the others. This smoothness is not imposed by the architecture, since the skip connections of the U-Net give its decoder access to high-resolution features; we rather attribute it to the difficulty of the task. The exact location of convective cells is only weakly constrained by the 0.25° ERA5 predictors, in which convection is parameterized, so a model trained with a pixel-wise loss is rewarded for spreading probability over all plausible locations rather than predicting a sharp but possibly misplaced pattern. Summed over the day, these probabilities give an expected number of lightning hours rather than one possible realization, so simulating events with realistic small-scale patterns would require a sampling step that draws spatially coherent binary fields from the predicted probabilities."

Figure 2 has been updated.

Figure 11: [s] This is very interesting. I would love to see another row of graphs showing the standard deviation of these values.

ANSWER: We thank the reviewer for this suggestion. We added the pixel-wise standard deviation of the saliency masks to the supplementary material (Figure S19). The standard deviation has the same spatial structure as the mean and is centered on the pixel of interest and decreases with distance. It also reveals a difference between the two groups of pixels. For lightning pixels, the standard deviation is about half the mean, whereas for non-lightning pixels it is twice the mean. The saliency of non-lightning pixels is strongly skewed: most of them have almost no saliency, while a small minority is as sensitive as lightning pixels. We hypothesize that these are pixels in convective environments where lightning was possible but not observed.

CHANGES: Line 430: "We look at the standard deviation of these saliency masks in Section S4.2."

L379: [m] Be more specific about the random values. "Random" can take many different forms, and if you are using random values that are out of sample, ML models often tend to behave unexpectedly.

ANSWER: We agree that this needed to be specified. For each predictor, we fit a normal distribution to its values over the training years and draw the random values from it, independently at each pixel. The draws are clipped to the physical range of the predictor (non-negative mixing ratio and precipitation; relative humidity and land-sea mask in [0, 1]). Every perturbed value therefore stays within the range seen in training. Section 4.6 now describes this.

While addressing this comment, we found that the previous version applied noise drawn in physical units to inputs that the model receives in standardized form. The perturbed values were therefore out of sample, as the reviewer suspected. We corrected this and repeated the analysis for radii up to 1500 km. The radius of useful information is now about 500 km, compared with 320 km in the previous version. At this radius, the median relative change of the prediction is about 10% (Figure 12b).

CHANGES: Lines 442-444: "For each predictor, we fit a normal distribution to its values over the training years, and we sample the random values from this distribution, independently at each pixel. We clip the samples to the physical range of the predictor (e.g. non-negative precipitation, relative humidity between 0 and 1), so that each value remains within the range seen during training."

L392: [r] Use `\operatorname{argmin}` *and* `\operatorname{error}` *for nicer typesetting.*

ANSWER: We removed this definition because the 10% threshold was arbitrary. Both radii (saliency and perturbation) are now defined with the Kneedle variant of the elbow method (Satopää et al., 2011), so the formula is no longer needed.

CHANGES: Lines 456-458: "Similarly to the saliency maps, we define a radius of useful information R with the "kneedle" variant of the Elbow method. We find that R $\sim$ 500 km, which corresponds to a (37 $\times$ 37) kernel centered on the pixel of interest. At this radius, the relative error is about 10%."

L454–458: [m] Due to climate change, I would be concerned about out-of-sample issues with this trained model. This limitation should be explicitly noted and discussed.

ANSWER: We fully agree. Under warming, the predictors (in particular mixing ratio, lifted index and convective precipitation) may leave the range sampled in 2008–2023, and none of our models is guaranteed to extrapolate correctly. Tree-based models such as XGB saturate beyond the training range, and nothing constrains the extrapolation of splines and neural networks to be physical. The U-Net may also rely on spatial patterns learned in the present climate. The conclusion now states this limitation explicitly, and we will perform two checks before any application to climate projections in future work:

- measuring how much of the projected predictor distribution falls outside the training range;
- testing how the model responds to inputs perturbed with a climate change signal of increasing amplitude, to verify that it stays physically consistent.

CHANGES: Lines 532-533: “Note that climate change signals may change the distributions of the predictors and we have no guarantee of generalization to out-of-distribution samples yet. Assuming this issue is addressed, [...]”

Table A1: [m] Optimization method: Can you be more specific? Which variant of gradient descent is used? Especially for MLPs and U-Nets, Adam is typically used nowadays. Regarding the loss functions, can you elaborate on the weighting strategy, as this is highly relevant for imbalanced datasets? You are sure that you used Quantile Regression for the binary classification task for XGB?

ANSWER: We thank the reviewer for raising these interesting points :

- Optimization method: the MLP and the U-Net are trained with Adam, the LR with L-BFGS and the GAM with P-IRLS. Table A1 now states this.
- Weighting: the positive class is weighted by w in the binary cross-entropy, and w is tuned in each LOYO split (6–19 for the MLP, 4–20 for the U-Net). In addition, a weighted random sampler draws training days with a probability proportional to the logarithm of their flash count (Appendix B). Platt scaling then corrects the resulting shift (see our answer to L236–241).
- XGB: “quantile regression” referred to the previous model. We re-optimized the model : it is now trained with the binary logistic objective (binary cross-entropy), and the table is corrected.

CHANGES: Table A1 has been updated.

Table A2: [s] Learning rate: I suggest either using scientific notation for the entire row or not using it at all.

ANSWER: Done: the learning rate is now in scientific notation for the entire row.

L481: [s] Also add the version of the Optuna package.

ANSWER: Added: Optuna 4.4.0 (Appendix B).

L483–484: [c] I am not convinced that this mix of metrics is the best choice, because it gives the model the flexibility to perform well on four of the metrics while performing poorly on the fifth. The choice of the optimization target is rarely easy and heavily influences model convergence. I am not arguing that using a sum of five metrics is necessarily a bad choice for this particular study, but it feels as though you haven’t fully decided what you want to optimize for. Also, looking at Figure 3, the metrics

come on different scales (ROC-AUC is at ∼0.97 while Dice is at ∼0.13), so an unweighted sum implicitely weights the metrics differently.

ANSWER: We agree that the choice of objective matters, and that an unweighted sum weights the metrics implicitly. What drives the selection, however, is not the level of each metric but how much it varies between hyperparameter sets. We checked this on the MLP tuning, for which the five metrics of every trial were stored (30 trials in each of the 14 tuning runs). The standard deviation is 0.06 for ROC-AUC, 0.03 for PR-AUC, 0.02 for Dice, 0.04 for FBSS and 0.08 for Deviance. This means that Deviance and ROC-AUC have indeed a larger implicit weight, but the order of magnitude stays the same across all metrics. The metrics largely agree on how they rank the trials. The mean Spearman correlation with the sum ranges from 0.82 (Dice) to 0.97 (PR-AUC and FBSS). For each run, the selected trial has its median rank between 1 and 4 out of 30 on every metric. The sum therefore does not select configurations that do well on four metrics and poorly on the fifth. We chose a multi-metric objective deliberately. The models serve both climatological uses, which require calibrated probabilities (FBSS, Deviance), and the detection of lightning events (PR-AUC, Dice). All models are tuned with the same objective, so this choice does not favor any of them.

L494: [m] Training GAM/XGB on only 0.1% of the data must be mentioned in the main text. Report the subsampling in the main text, and show that GAM/XGB scores are insensitive to the subsample size, since "sufficient for convergence" is asserted without evidence.

ANSWER: We thank the reviewer for this comment. We agree that subsampling is an important part of how these models are trained, and that the statement "sufficient for convergence" was not supported by evidence. The number of training samples is now a tuned hyperparameter of the LR, GAM and XGB (Tables A1–A3), and Appendix B describes how it is selected.

Figure 1 (in the response below) shows the scores as a function of the number of training samples, for models trained on the 13 LOYO years and evaluated on the 3 held-out years:

- LR: apart from ROC-AUC, the scores do not improve beyond 10^5 samples.
- XGB: the scores level off beyond $2 \cdot 10^7$ samples.
- GAM: the scores still increase slightly between 10^6 and $3 \cdot 10^6$ samples, but the memory used by pyGAM limits us to $2 \cdot 10^6$ samples. Our GAM reproduces the model of Battaglioli et al. (2023) and reaches a similar explained deviance with $2 \cdot 10^6$ samples, so we consider this sample size sufficient.

These sizes are 0.006%, 0.13% and 1.3% of the training set of a LOYO split. The samples are spread over the training period by drawing one random hour of every training day. Appendices A and B now present this.

CHANGES: Lines 542-545: "The number of training samples for LR and GAM is limited by the memory used for matrix inversion. For XGB, the number of training samples is a trade-off between performance and the time required to fit the model. For each model, we selected the lowest number of training samples for which there was no more improvement in terms of total score or the maximum number of samples that could fit in memory (see Appendix B)."

We added Tables A1, A2 and A3.

Lines 592-601: "The LR, GAM and XGB are constrained by memory and training time so they cannot use the entire training set. Therefore, we fit them on subsets of the data of size 105 samples for the LR, 2·106 samples for the GAM and 2·107 samples for XGB, i.e. 0.006%, 0.13% and 1.3% of the training set of a LOYO split. These numbers are selected as the lowest number of samples n_{opt} for which the optimization objective does not increase for $n_{\mathrm{samples}} > n_{\mathrm{opt}}$, or the maximum number of samples that fit into memory. To cover as many meteorological situations as possible, we spread these samples over all the training days: we draw one random hour of every training day, and a random subset of its grid cells. The subset size is adjusted such that the sum is equal to n_{opt}. The MLP is trained with mini-batches drawn from all the training days: each batch contains 1 to 3 days (a tuned hyperparameter, see Table A4), 6 random hours of each day and 20% of the grid cells of each hour. We evaluate the performance of all models on the entire test set."

References: [m] There are a couple of references pointing to arXiv that have most likely been published in peer-reviewed journals by now. Additionally, there are a few references missing a DOI, and the formatting style (e.g., uppercase vs. Lowercase letters) is inconsistent.

ANSWER: We updated the bibliography. Preprints that have since been published now cite the published version. DOIs are added wherever they exist; the remaining entries without a DOI are conference proceedings, books, theses, software or reports. Capitalization is harmonized.

Open Questions After Reading the Paper

1. [m] Is the varying spatial resolution in kilometers (comparing a 0.25°x0.25° cell in southern Europe with one in northern Europe) accounted for in any way, and how does this variation influence the results?

ANSWER: No, the varying grid-cell area is not accounted for. Over our domain, a 0.25° cell covers about 630 km^2 at 35°N and 390 km^2 at 60°N. A lightning hour is defined per grid cell, so the same flash density gives a higher probability of occurrence in a larger cell. Part of the north–south gradient in lightning occurrence may therefore reflect cell size. The impact may be twofold :

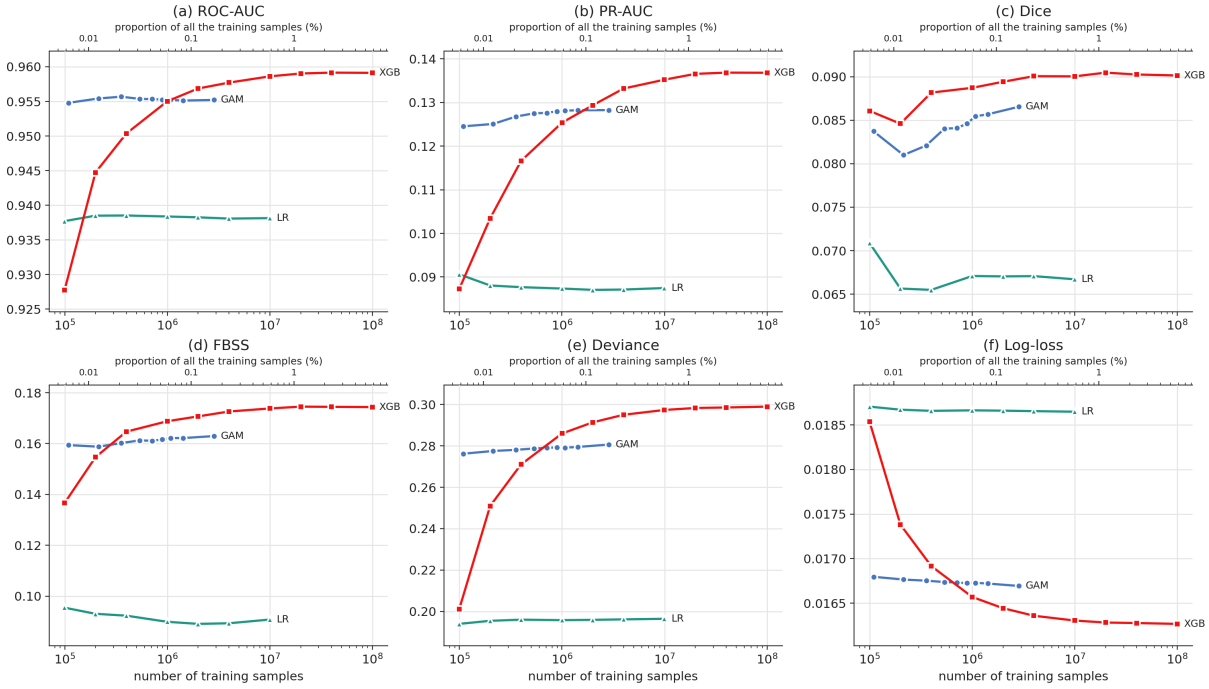


Figure 1. Impact of the number of samples used to fit the LR, GAM and XGB model for (a) ROC-AUC, (b) PR-AUC, (c) Dice, (d) FBSS, (e) Deviance (f) Log-loss. Higher is better for (a)-(e) and lower is better for (f).

- The impact is different across models : single grid cell models have no direct information about cell size, although predictors correlated with latitude may capture part of it. The U-Net however can infer the location of each cell (Section 3.1), so it may partly learn this effect.
- It may impact the precise value of the radii defined in Section 4.6, but not their order of magnitude, which is the important result here.

The conclusion now acknowledges this limitation and two remedies: predicting a flash density (per km^2) rather than an occurrence, and adding latitude and longitude as inputs.

CHANGES: Lines 503-504: "We also do not take into account the spatial resolution variation induced by the (latitude, longitude) grid. This could be solved by moving to lightning flash density estimation."

Lines 515-517: "Future work could isolate the contribution of the spatial context by giving the single grid cell models the latitude and longitude of each grid cell as additional inputs, so that both model families have access to the location and differ only by the spatial context."

2. [q] You argue that neighboring cells are important to capture large-scale organized events. Did you consider also utilizing the 137 vertical levels of ERA5 data (e.g., using all or a subset of these layers as additional input channels for the U-Net)? What about temporal neighbors, such as the previous and next hours?

ANSWER: We thank the reviewer for this suggestion. Using the ERA5 model levels directly, or the previous and next hours, could improve the scores; with derived predictors, the way the vertical profile is summarized is fixed when the predictors are chosen. However, the aim of this study is to isolate the effect of spatial information, not to discover predictors from data. We therefore kept a small set of predictors with demonstrated skill in previous studies (Battaglioli et al., 2023), and gave the same inputs to all models. The extension to other predictors (which can be either the full 137 vertical levels or other post-processed variables) is now mentioned in the "future work" part of the discussion, and we have already started working on a model that uses the previous and next hours.

CHANGES: Lines 524-527: "[...] This suggests that there might be other predictors that are particularly useful for the U-Net that we did not include in our study. To improve this model, future work should explore a larger set of predictors, tailored for this kind of architecture to obtain even better predictions"

3. [M] I suspect that the baseline comparisons are not entirely fair: Much effort was put into hyperparameter tuning for the U-Net. Did you perform similar hyperparameter tuning for the baseline models? You argue that single-grid models suffer from not knowing mesoscale patterns. However, the U-Net has an additional "unfair" advantage: each pixel inherently knows its absolute location, and each location has its own static characteristics (e.g., mountainous regions are more prone to lightning). Thus, the U-Net knows both the location of each pixel and its surroundings. This could largely be compensated for by also feeding the location (latitude/longitude) to the single-grid models. Doing so would isolate exactly what you intend to test: the only difference would be that the U-Net sees spatial surroundings, while both architectures have access to absolute geographic location.

ANSWER: We thank the reviewer for this comment and we agree on both points.
Although we did some tuning on the single grid-cell models, it was not as exhaustive as for the U-Net. In the revised version, the MLP and XGB are tuned with the same nested procedure and objective as the U-Net, in the inner 3-fold cross-validation of each LOYO split (Appendix A).

- LR : serves as a simple baseline.
- GAM: we reproduce a model from the literature (Battaglioli et al., 2023), so we only tuned their training sample size and, for the GAM, the smoothing parameter (Appendix A, Table A2).
- XGB: a grid search over tree depth and number of trees (Table A4).

- MLP: a TPE search (30 trials) over its architecture, regularization, learning schedule and positive-class weight (Table A4).

We also agree that the U-Net has access both to the location of each cell and to its surroundings, and that our setup does not separate the two. We kept the inputs of Battaglioli et al. (2023) for all models so that the baselines remain comparable to the literature, so the study measures the combined effect. The introduction, Section 3.1 and the conclusion now state this explicitly. The conclusion also proposes the reviewer's experiment as future work (latitude and longitude as additional inputs to the single grid cell models), together with the converse experiment (training the U-Net on randomly positioned sub-domains).

CHANGES: Lines 82-84: "This spatial information is two-fold: the U-Net has access to both the location in the domain of each pixel and the spatial context surrounding each pixel. This study assesses the overall impact of the additional spatial information and does not isolate the individual contribution of spatial context and location."

Lines 147-154: "Compared to the single grid cell models, the U-Net has access to two distinct sources of spatial information. The first is the spatial context: the prediction at a grid cell depends on the predictors in the surrounding grid cells, within the receptive field of the convolutional layers. This is the information that we expect to characterize organized convective systems. The second is the location of the grid cell. Since the U-Net always processes the same fixed domain, it can infer where a grid cell lies from static features of its input: the coastline pattern of the land-sea mask around the grid cell, the persistent geographic imprint of the other predictors (e.g. orography) and distance to the image boundaries (Islam et al., 2020; Semih Kayhan and van Gemert, 2020). Our comparison measures the combined effect of both sources and does not separate the contribution of the spatial context from that of the location information."

Lines 513-519: "The fixed domain also means that the gain of the U-Net over the single grid cell models has two sources that our setup does not separate: the spatial context provided by the surrounding grid cells, and the location of each grid cell, which the U-Net can infer from static features of its input. Future work could isolate the contribution of the spatial context by giving the single grid cell models the latitude and longitude of each grid cell as additional inputs, so that both model families have access to the location and differ only by the spatial context. Conversely, training the U-Net on randomly positioned sub-domains would remove its knowledge of the absolute location while keeping the spatial context. Such a model, not tied to a fixed domain, could also be used to infer lightning probabilities in regions where lightning observations are scarce"

We added Tables A2, A3 and A4:

Lines 546-551: "Table A2 features the hyperparameters we tune for the GAM. The smoothing parameter λ of the GAM is tuned to 0.6 for all splits as the test scores are unchanged for λ between 0.06 and 0.6 and decrease for larger values. Table A3 features the hyperparameters we tune for the XGB model. The max depth and the number of trees are tuned with a grid search in the inner 3-fold cross-validation of each LOYO

split, with the same objective as the U-Net. For each depth, one model is fitted with the largest number of trees (1500) and scored at every number of trees of the grid. Depth 3 is the best depth in every split, and its scores are flat beyond 1000 trees. We therefore use depth 3 and 1000 trees for all splits."

Lines 553-556: "Table A4 features the hyperparameters we tune for the MLP. All hyperparameters are optimized with a Tree-structured Parzen Estimator (30 trials) in the inner 3-fold cross-validation of each LOYO split, with the same objective as the U-Net. The selected architecture varies between the LOYO splits (1 to 3 hidden layers of 64 or 128 neurons, positive-class weight between 6 and 19)."