

Review 10.5194/egusphere-2026-4363 preprint

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 $0.25^\circ \times 0.25^\circ$ 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.

Legend for the following concerns:

- [c] comment
- [m] minor
- [M] major
- [q] question
- [r] rephrase
- [s] suggestion
- [t] typo

Detailed Comments:

- **L24-27** [t] “approximate” is duplicated; “alternatively” should start with an upper case letter
- **L49:** [t] “Cavaiola et al. (2024)” is not a hyperlink.
- **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.
- **L56:** [r] Be more precise. ERA5 has a resolution of $0.25^\circ \times 0.25^\circ$, 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.
- **L57-58:** [m] Can you provide a reference for this?
- **L69-70:** [t] “Sha et al., 2024” is not a hyperlink.
- **L85:** [m] GitHub repositories are not permanent. I recommend publishing on a repository like Zenodo and providing a proper citation to the code here. Kudos on the GitHub repo, which even includes a `pyproject.toml` and `uv.lock`. I would additionally suggest providing a `.python-version` file just to be sure.
- **L88:** [q] Should this be “hybrid sigma-pressure levels”?
- **L102:** [m] Your study domain is Europe. A 0.25° horizontal resolution does not correspond to $\sim 30 \text{ km}$ (at least for longitude, this varies quite a lot across Europe).

- **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?
- **L127:** [r] “Battaglioli et al., 2023a” is not a hyperlink.
- **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.
- **L164:** [m] Add a reference for the autocorrelation claim.
- **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.
- **L190:** [r] The formula’s formatting looks odd.
- **L211:** [q] Which set the 49 days refer to?
- **L234:** [r] “80%” is probably not exact; this should be “Approx. 80%” (or something similar).
- **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.
- **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.
- **L243:** [q] Are 16 years enough to call this a 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.
- **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?
- **L312–313:** [m] As mentioned earlier, replace the Git repository with a permanent, citable link (e.g., published on Zenodo).
- **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.
- **Figure 11:** [s] This is very interesting. I would love to see another row of graphs showing the standard deviation of these values.
- **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.
- **L392:** [r] Use `\operatorname{argmin}` and `\operatorname{error}` for nicer typesetting.
- **L409:** [t] “and a four” -> “and four”
- **L436:** [t] “which are may” -> “which may”
- **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.
- **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?
- **Table A2:** [s] Learning rate: I suggest either using scientific notation for the entire row or not using it at all.
- **L481:** [t] “python” -> “Python”
- **L481:** [s] Also add the version of the Optuna package.

- **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 implicitly weights the metrics differently.
- **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.
- **L552:** [t] Use `\operatorname{BS}` for better typesetting.
- **L556:** [t] Use `\operatorname{error}` for every text part.
- **L567:** [m] See previous comments regarding the GitHub link; please make your repository citable.
- **L573:** [m] Same as L567.
- **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.

Open Questions After Reading the Paper:

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