Review of the preprint:

From logistic regression to deep learning: machine learning modeling of lightning in ERA5 reanalysis data

This preprint presents a well-motivated and carefully constructed comparison of five machine learning models: logistic regression, generalized additive model, extreme gradient boosting, multi-layer perceptron and a U-Net. These shall predict lightning occurrence across Europe using predictors from ERA5, with the central goal of testing the hypothesis whether "*large-scale thunderstorm clusters such as mesoscale convective systems are driven by broad spatial predictor patterns*". The paper is well-written, clearly organized and deals with an important topic. The methodological design is understandable and the interpretability analyses (SHAP-based feature importance, gradient-based saliency and radial perturbation) give valuable insights. The central finding that the U-Net outperforms single-grid-cell models, particularly for extreme events, and that this improvement is associated with a spatial scale on the order of 100–300 km is credible and a very interesting finding.

That said, I still have some comments and concerns, ranging from minor to moderate.

1. Lines 108-110: "*We follow Battaglioli et al. (2023a) and use five variables derived from the ERA5 reanalysis dataset on 137 hybrid-sigma levels (Hersbach et al., 2020): (i) convective precipitation, (ii) most-unstable lifted index, (iii) most-unstable mixing ratio, (iv) average relative humidity between 500 and 850 hPa, and (v) a land-sea mask.*"

This sentence implies all five variables are derived from the 137-level model-level (and not pressure-level, I guess?) output, but this is only true for some of them. Convective precipitation is a surface diagnostic output directly by ERA5's convection scheme, not something computed by processing the vertical profile. The land-sea mask is a static, single-level (2D) field with no vertical dimension at all. Only the lifted index and mixing ratio (and possibly the layer-averaged RH) genuinely require scanning the vertical profile to identify the most-unstable parcel and even then, it should be stated explicitly that this is done via the thundeR package (mentioned later, line 119–120) rather than left implicit here. Suggested revision: clarify which variables are genuinely profile-derived (via thundeR, using the full 137-level model-level resolution) versus which are single-level/surface fields taken directly from ERA5, to avoid implying uniform derivation across all five.

2. Line 111: This is the first place the acronym "AR-CHaMo" appears in the text. A short explanation with the citation here would help readers who jump directly to the Methods section.

3. Line 200: Section 3.5 is too short to stand alone. I would recommend to put the information to the previous section.

4. Figure 3: The caption states *"Higher is better for all metrics"* but does not specify the possible range of each metric. For example, PR-AUC, ROC-AUC are bounded in [0, 1], but for example FBSS and Deviance explained are skill scores relative to a baseline model and are not bounded to [0, 1]. Please state the theoretical range of each metric, either in the Figure 3 caption or in Table 1/Appendix D, so readers can correctly judge how close each model is to its bounds.

5. Figure 3: The paper reports only out-of-sample (test) performance across the 13 LOYO splits, which is of course the correct way to do analyses like these. It would be interesting to briefly discuss whether the in-sample (training) performance for each model deviates significantly from the out-of-sample performance, to give readers a sense of the train-test gap for each model class. This is particularly relevant for the more flexible models (MLP, U-Net), where a large gap would indicate overfitting risk that a single out-of-sample number doesn't reveal.

6. Lines 218–223 *"We confirm this by testing the statistical significance of the difference in performance between the U-Net and the other models for each metric with a Student's t-test, using the 13 test metrics computed on each split of the LOYO procedure as samples. We find that the p-value is below 10^{-5} for every model and metric ..."*

The 13 "samples" used in the t-test are not independent, because each LOYO split's training set overlaps with every other split's by 11 out of 12 years. This means the resulting models and their test-set errors, are correlated rather than i.i.d., which violates a core assumption of the Student's t-test. This likely leads to an "overoptimistically" small p-value reported ($< 10^{-5}$).

I recommend using a paired test instead, since all models are evaluated on the same 13 splits: for each split, compute the difference in performance between the U-Net and the competing model, then test whether the mean of these 13 differences is significantly different from zero (paired t-test), or use the Wilcoxon signed-rank test on the same differences if normality is a concern. Pairing at least removes the confound of some years being harder to predict than others for every model. Note, however, that this does not fully solve the underlying problem that the 13 differences are still not fully independent, since the folds still share overlapping training data. This residual limitation should ideally be acknowledged even after switching to a paired test.

7. Line 238: "*The U-Net is the best calibrated model overall, with a calibration curve closer to the diagonal at every level, ...*"

Being the best-calibrated of the five models does not by itself establish that the calibration is good in an absolute sense. The text itself notes the U-Net *"still overestimates probabilities above 30%"* even after Platt scaling. Please state explicitly whether this residual miscalibration is considered acceptable for your further analyses, and, importantly, clarify which set of results elsewhere in the paper (Figure 3, Table 2, the climatology in Section 4.3, the case studies in Section 4.5) use the Platt-scaled (calibrated) U-Net output versus the raw output.

8. Lines 263–264: "*For the spatial distributions, we show in Figure 6 how the different models compare in the summer season. Winter, spring and autumn are available in the supplementary material (see Figures S2, S3, and S4).*"

Given that mesoscale convective systems occur across all seasons (not just summer), moving three of the four seasons entirely to the supplement without any discussion in the main text is a missed opportunity. At minimum, a sentence or two summarizing whether the U-Net's advantage over single-grid-cell models holds consistently across all four seasons, or whether it is specific to summer convection and works less good in the other seasons (and if so, why), would strengthen the claims made in this section. Further, reporting RMSE in absolute terms (hours of lightning) is difficult to

interpret without knowing the baseline magnitude at each location. An RMSE of 2 hours has a very different meaning in a grid cell with a true climatological total of 2 lightning-hours versus one with 30 lightning-hours. Please express the RMSE as a proportion of the observed climatological mean (or total) at each location (e.g., a normalized RMSE (NRMSE) or percentage error map) in addition to, or instead of, the raw RMSE, so that the spatial error patterns can be properly interpreted.

9. Lines 310–311: *"We present the results for the 15th of August 2008 event in Figure 8, and the results for the other two events are available in Figures S5 and S7 in the supplementary material."*

As with the seasonal spatial distributions, moving two of the three case studies entirely to supplementary material without any summary in the main text weakens the generality of the claims drawn from the single presented case (Aug 2008). A short sentence confirming (or qualifying) whether the same qualitative pattern holds for the 2015 and 2023 events would let readers judge whether the August 2008 case is representative or an outlier.

10. Lines 371–372: *"We defined a radius of useful information R as the radius at which the median of the saliency values is 1/10 of the maximum of the median value across all distances."*

This threshold is not motivated or justified anywhere in the text and the resulting radius estimates (135 km via saliency, 320 km via perturbation) are reported to a precision that implies more confidence than an arbitrary 1/10 cutoff can support. Please either provide a reason for this specific threshold or report how sensitive R is to this choice. For instance, what R would be obtained at 1/5 or 1/20, so readers can judge how robust the conclusions are to this choice.

11. Figure 7: Because the method retrains the model on all 32 possible feature subsets, the resulting Shapley values already account for interactions between predictors, not just each variable's isolated effect. This is a strength of the approach, but it isn't stated explicitly in the text, and the bar charts in Figure 7 could be misread as showing independent, additive "main effects" per variable. Please add one or two sentences clarifying that the procedure actually does capture interaction effects by construction. It would also be worth briefly discussing whether any of the reported patterns, e.g., the U-Net's stronger reliance on convective precipitation, likely reflect a genuine interaction with the other predictors, rather than a standalone effect of that variable alone.