

Review 10.5194/egusphere-2026-4363 preprint

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

Abstract

L10: Regarding the finding that the most important predictors depend on the model type, the U-Net has repeatedly been shown to outperform baseline models such as an MLP. To support a feature-importance statement like this one, I would suggest a comparison with state-of-the-art architectures. Alternatively, a simpler approach would be to focus on identifying the most informative input features, either by varying the network's hyperparameters or by making incremental architecture changes.

1 Introduction

L16: While it is true that lightning-ignited wildfires cause ecological damage, biodiversity loss and emissions, it could be argued that fire-management policy, combined with global surface warming, is altering the natural behaviour of lightning-ignited fires and their benefit of clearing dead or diseased tree stands. In that sense, such wildfires could be considered a naturally occurring event. The statement is not misleading for other readers, however.

L26: In “convective convective processes”, the word is duplicated.

2.2 ERA5 data

L111-113: Convective precipitation, which comes from the convection scheme, is a short-range forecast accumulation, so the models partly post-process the IFS scheme. Suggest results with and without it, and stating its accumulation window relative to the lightning hour.

3.1 Single-cell vs spatial models

L133-138: In the comparison of single-cell models with the U-Net, two things change at once: spatial context and knowledge of location, since a fixed domain plus padding lets a CNN encode position (Islam et al. 2020). Suggest three computation schemes:

- a 1x1 kernel convolution network;
- single-cell models given latitude and longitude and orography;
- XGB with neighbourhood statistics (mean, max, gradients over 3x3 to 23x23 windows).

3.2 U-Net architecture

L146-147, Table A2: The Tree-structured Parzen Estimator search chose the minimum allowed depth (two downsampling levels) and the narrowest channels.

Suggest extending the range downward and reporting skill against depth. This is the most direct test of the mesoscale hypothesis.

L153-155: Platt scaling, fitted here on the predicted probability, is defined on logits. By applying the scaling to probabilities, it cannot represent the identity, which may explain the residual overestimation. Suggest logit-based scaling.

L142-143: For the input padded from 101x149 to 112x160 pixels, state the padding type. Zero padding creates artificial edge gradients. Suggest reflection padding or scoring on the interior only, plus skill against distance to the boundary.

3.3 Training pipeline

L173-174: Regarding the held-out years 2008, 2015 and 2023, it is stated the year 2023 introduced inhomogeneity, yet 2023 is a third of the final test. Suggest per-year scores in Table 2 and a homogeneity check.

3.4 Metrics

L186: The Fractions Brier Skill Score used here is non-standard. The established Fractions Skill Score (Roberts and Lean 2008) is computed across neighbourhood sizes.

L188-189: As only threshold-independent metrics are reported, suggest adding a performance diagram that incorporates POD, success ratio, CSI, and frequency bias (Roebber 2009), the standard in convective verification.

4.1 Quantitative results

L224-225, Table 2: The held-out-year scores are reported only for the final U-Net. Add the four baselines too for a fair comparison.

4.2 Calibration

L236-240, Fig. 4b: Among the calibration curves of all models, the XGB curve falls to zero observed frequency above about 0.65 without comment, and Table A1 lists a quantile regression loss for XGB. Explain or switch to a logistic objective.

4.3 Climatology

L256: For the U-Net's much better fit to the diurnal cycle, no mechanism is given and no model receives time information. Suggest giving single-cell models cyclic hour and day-of-year features, and previous-hour predictors, before crediting spatial context.

L269-270: Although the U-Net's over-prediction in mountain regions is attributed to ERA5 resolution, the local models under-predict over the Alps and Pyrenees, so the sign changes with architecture, not data. Suggest static orography channels such as elevation for testing the correctness of the statement.

L258-262: The explanation that the autumn overestimation comes from lightning activity shifting over the Mediterranean is testable with the existing mask. Show the cycles for land and sea separately.

4.4 Feature importance

L275-279: The quantity obtained by retraining each model on all 32 feature subsets is a Shapley value of predictive performance, not SHAP, which attributes individual predictions. Rename it, and report predictor correlations, since Shapley splits credit among correlated inputs and this may drive the architecture dependence.

L297-299: The explanation that convective precipitation matters more for the U-Net because of its spatial clustering is untested. Suggest spatially shuffling one channel at increasing scales and measuring the skill loss.

4.5 Case study

L336-341: The U-Net's failure to capture small-scale patterns, attributed to its convolutional architecture, is expected from pixel-wise cross-entropy.

4.6 Interpretability

L378-379: In the radial perturbation method, random values with a sharp edge are out of distribution, so the error may reflect artefacts, not lost information. Suggest using the same region from another date.