Supplementary material for *From logistic regression to deep learning: machine learning modeling of lightnings in reanalysis data*

Adrien Burq[1,2], Jean Jouhaud[2], Victor Xing[2], Victor Bouvier[2], Vincent Forcadell[2], Mateusz Taszarek[3,4], Mathieu Vrac[1], and Davide Faranda[1,5,6]

[1]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CEA-CNRS-UVSQ, Université Paris-Saclay, IPSL, 91191 Gif-sur-Yvette, France
[2]Descartes Underwriting, 148 rue de Courcelles, 75017 Paris, France
[3]Department of Meteorology and Climatology, Adam Mickiewicz University, Poznań, Poland
[4]Skywarn Poland, Warsaw, Poland
[5]London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK
[6]Laboratoire de Météorologie Dynamique/IPSL, École Normale Supérieure, PSL Research University, Sorbonne Université, École Polytechnique, IP Paris, CNRS, Paris, 75005, France

Correspondence: Adrien Burq (adrien.burq@lsce.ipsl.fr)

Contents

1 Season climatology

Figure S1 shows that the U-Net better captures the large-scale spatial structure of the summer lightning climatology than the single grid cell models, with its predicted power spectrum lying closer to the observations at small wavenumbers. At large wavenumbers, however, the U-Net's spectrum falls below the others, reflecting the spatial smoothing induced by predicting a probability with a convolutional architecture.

This is confirmed by the spatial maps (panels b–g): the U-Net better reproduces the intensity of the main climatological hot spots, such as the Alps, the Apennines, the Pyrenees, and the Mediterranean Sea, while producing smoother spatial fields than the single grid cell models. Among the single grid cell models, performance varies considerably, with LR showing the largest RMSE and a tendency to strongly overestimate (over the Adriatic Sea) or underestimate (over mountainous regions) lightning activity, while GAM and MLP perform more comparably to each other. The U-Net's RMSE of 6.43 is a substantial improvement: over 20% of relative improvement compared to the GAM which is the best single grid cell model in terms of RMSE.

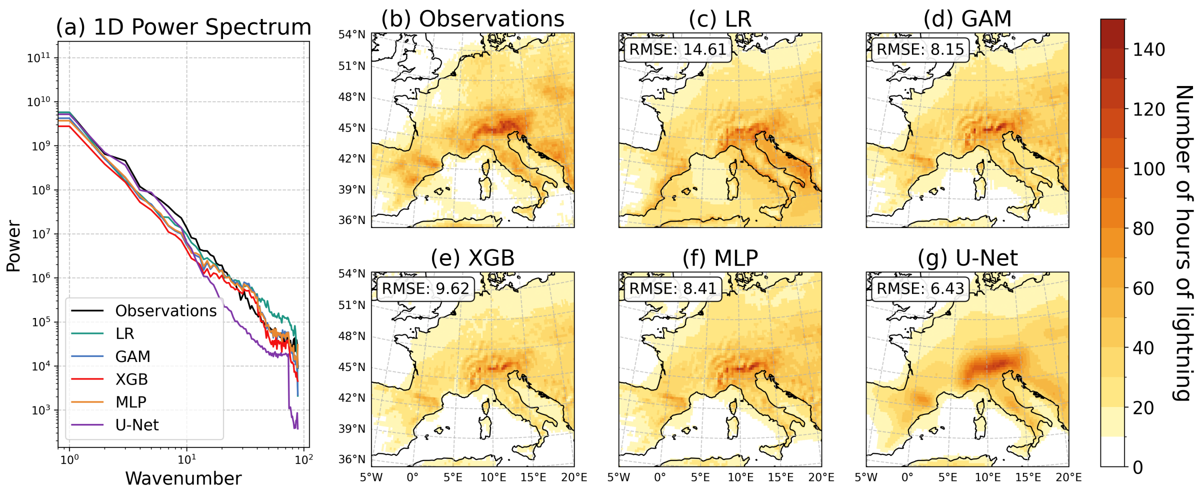

Figure S1. Yearly average number of hours of lightning in Summer on the test set (2008, 2015 and 2023). On (a) we assess the 1-D power spectrum of the images in (b) to (g). (c) to (g) feature the raw predictions of the models with respect to the observations (b). Modeled climatology and observations are compared with the RMSE between the two images (lower is better).

Figures S2, S3 and S4 show that the U-Net achieves the lowest RMSE across all seasons, consistent with the results for summer presented in the main text. The nature and spatial structure of the biases, however, vary across seasons.

In winter, all models exhibit a strong negative bias over the Mediterranean Sea, which the U-Net partially alleviates; the single grid cell models, particularly LR and GAM, show the largest negative biases in this region. In spring, the picture is more nuanced: while the U-Net reduces the negative bias over the southern part of the domain, it introduces stronger positive biases over orographically complex regions such as the Alps, the Pyrenees, and the Apennines, suggesting that the U-Net may

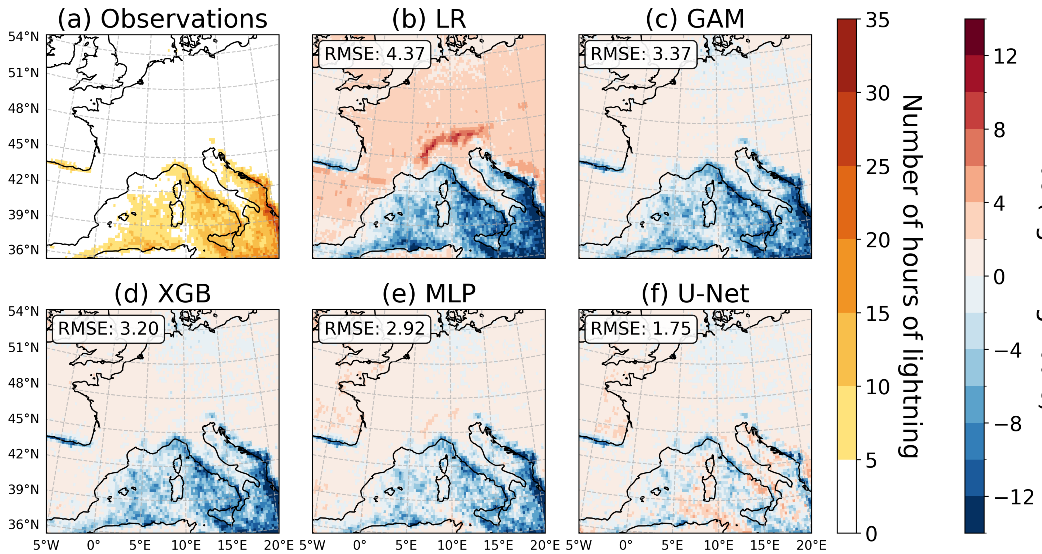

Figure S2. Yearly average number of hours of lightning in Winter on the test set (2008, 2015 and 2023). We assess the error of the models (b) to (f) with respect to the observations (a). Modeled climatology and observations are compared with the RMSE between the two images (lower is better).

overestimate the contribution of orographic forcing to climatological lightning in this season. In autumn, the U-Net reduces the negative bias over the Ionian region but increases the positive bias over Italy and the Adriatic Sea.

Overall, while the U-Net consistently achieves better RMSE, its bias patterns are spatially heterogeneous and season-dependent, in contrast to the single grid cell models which tend toward a more uniform negative bias across the domain.

2 Case studies

2.1 2015-08-11 event

The severe weather event impacting southern Italy on the 11th and 12th of August 2015 was driven by a cut-off low originating over the Atlantic. As this low-pressure system reached the warm central Mediterranean, it triggered extended deep convection over southern Italy. More intense cells developed over the seas surrounding Italy because of the large sea-tropospheric temperature gradient which is, at the end of the summer season, stronger than over land in this region. The stationarity of the cut-off low and the availability of heat and moisture contributed to the regeneration of the convective cells, enhancing the intensity of the system. This, combined with coastal orography, triggered extreme precipitation rates and lightning activity. Furnari et al. (2020) presents the event in more detail.

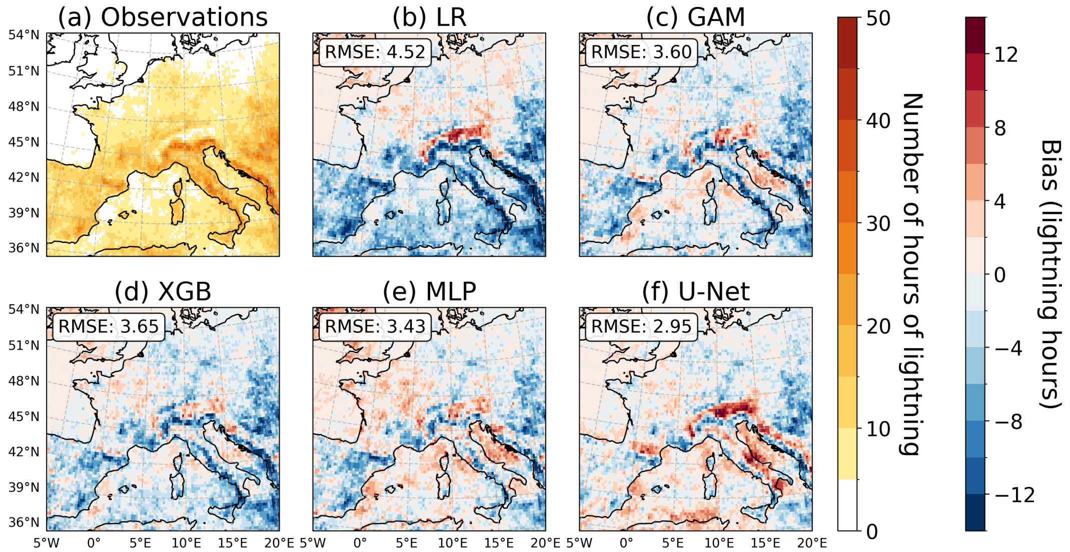

Figure S3. Same as S2 but for Spring.

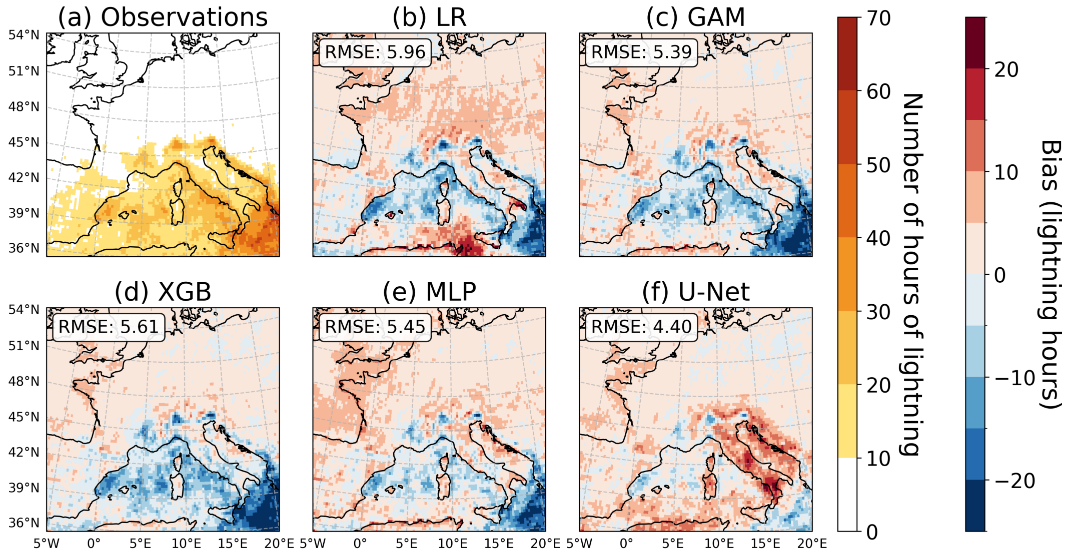

Figure S4. Same as S2 but for Autumn.

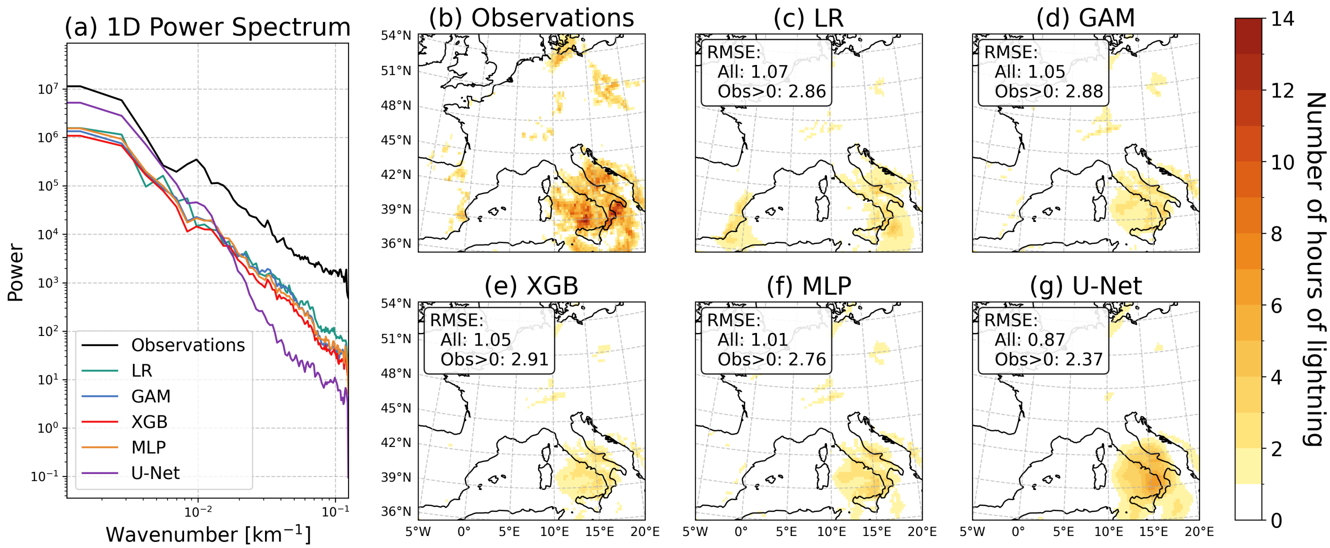

Figure S5. 11th of August 2015 event modeling across the different models. We compute the RMSE between the images of the observations and the predictions aggregated over the day. We also compute the ROC-AUC, PR-AUC, Dice, FBSS and Deviance metrics for this particular case study based on the hourly observations and predictions (see Figure S6).

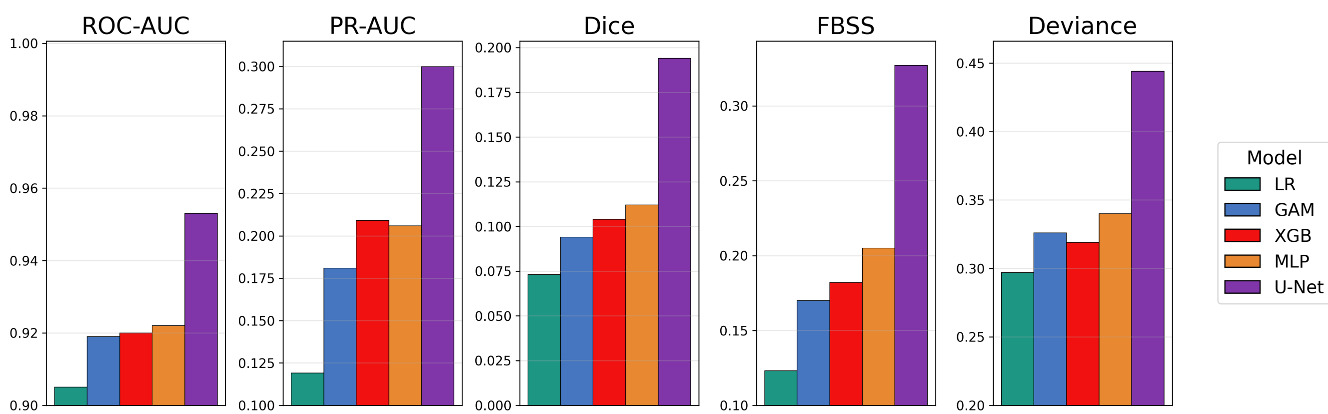

Figure S6. Scores of each model for the 11th of August 2015 event.

Similarly to the 15th of August 2008 event, we note from Figure S6 that the U-Net outperforms all other models across every evaluation metric for this extreme event. Figure S5 shows that the U-Net better captures the spatial extent and location of the lightning activity than the other models. While all models tend to underestimate the intensity of the event, the single grid cell models do so more severely, also struggling to reproduce the full spatial extent of the active region over the south of Italy. The power spectrum further confirms that the U-Net better captures the large-scale features of the event, with its predicted spectrum lying closer to the observations at small wavenumbers. On the other hand, the small-scale features are not well captured by any model. The U-Net performs worse than the other models on this task, probably because of its convolutional architecture. As mentioned in the main text, this smooth behavior is expected as we predict the expected number of hours of lightning during the day with a convolutional architecture.

2.2 2023-08-05 event

From the 3rd to the 6th of August 2023, the northern and central Mediterranean was dominated by a high-altitude cold Atlantic trough, which led to the formation of a shallow cyclone that stalled a weather front over Slovenia. The air mass went through an exceptionally warm Mediterranean Sea, an inexhaustible reservoir of moisture. The combination of atmospheric instability, moisture and wind shear generated persistent convective cells, which triggered heavy downpours and widespread lightning activity throughout the region. ARSO (2023) presents the event in more detail.

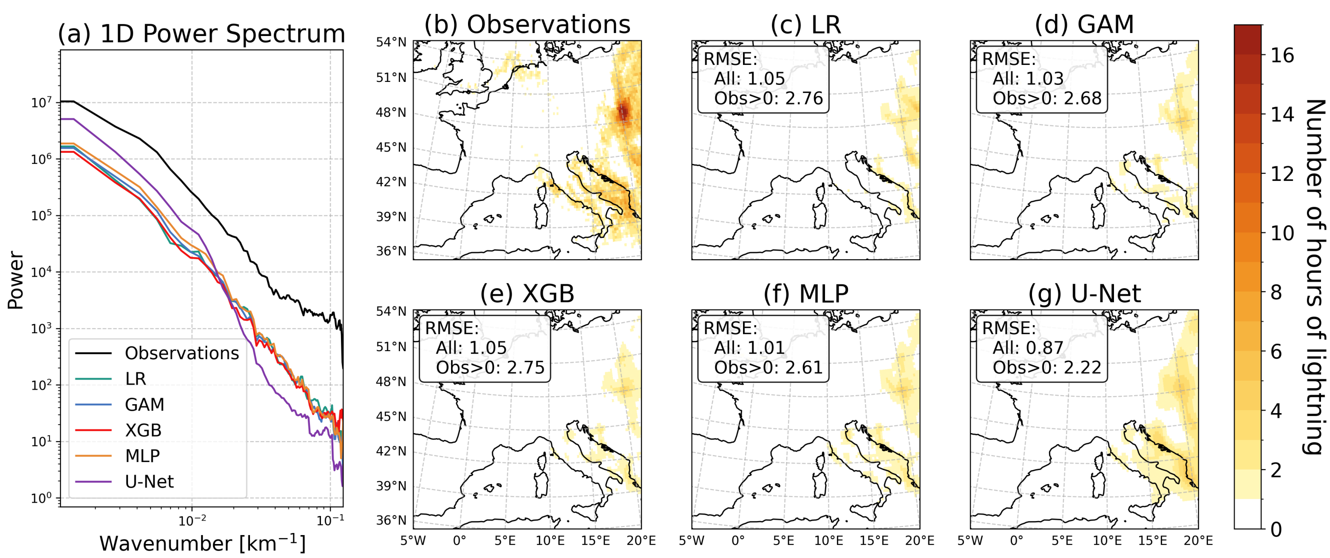

Figure S7. 5th of August 2023 event modeling across the different models. We compute the RMSE between the images of the observations and the predictions aggregated over the day. We also compute the ROC-AUC, PR-AUC, Dice, FBSS and Deviance metrics for this particular case study based on the hourly observations and predictions (see Figure S8).

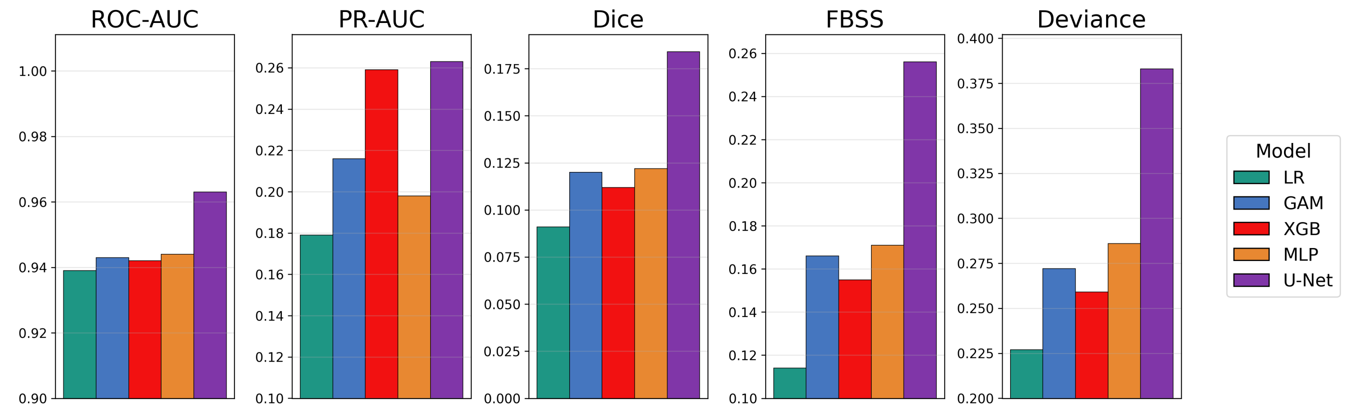

Figure S8. Scores of each model for the 5th of August 2023 event.

As with the previous events, Figure S8 shows that the U-Net outperforms all other models across all evaluation metrics, though the gap on PR-AUC with respect to XGB is notably smaller than for other metrics. Figure S7 confirms that the U-Net better captures the spatial extent and location of the lightning activity, which spans from the Adriatic Sea to Poland. All models underestimate the total intensity of the event, but the single grid cell models do so more severely and fail to reproduce the full northward extent of the active region. The power spectrum shows a consistent pattern: the U-Net's prediction lies closer to the observed spectrum at small wavenumbers. At large wavenumbers, as expected from the smoothness of the predicted expected number of lightning hours, no model reproduces the observed small-scale variability, with the U-Net being the smoothest of all due to its convolutional architecture.

3 SHAP

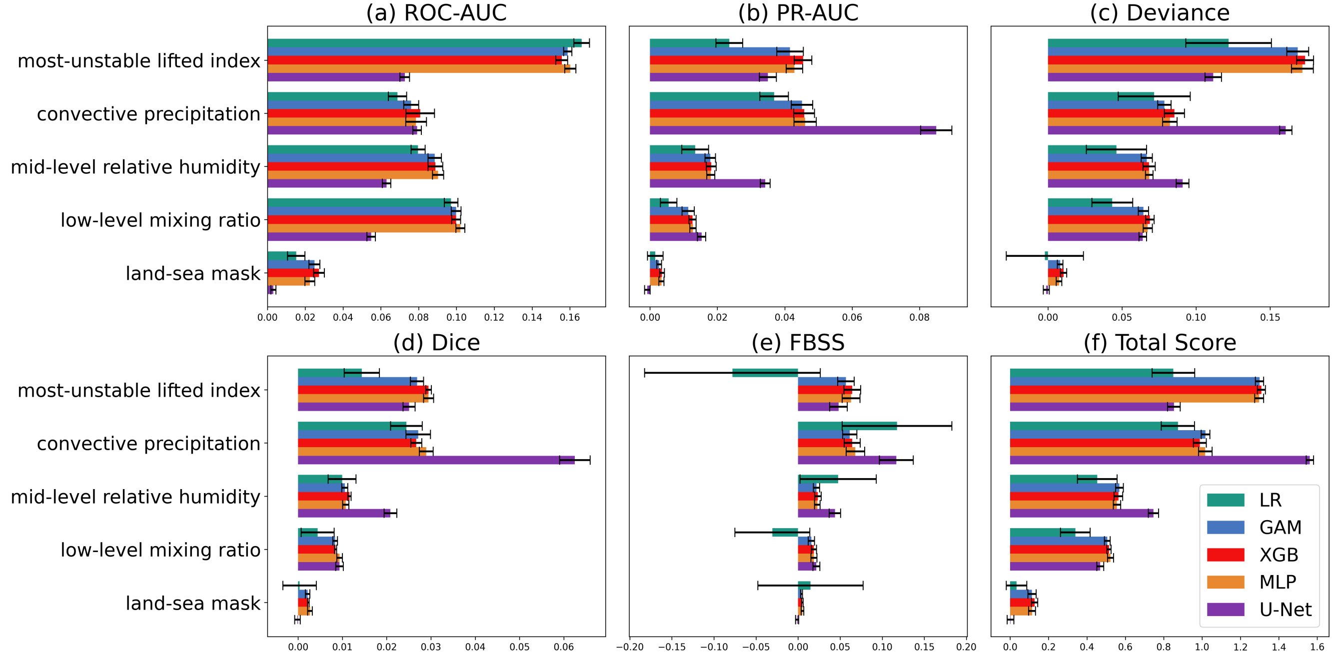

Figure S9. SHAP value analysis for all metrics: (a) ROC-AUC, (b) PR-AUC, (c) Deviance, (d) Dice, (e) FBSS and (f) Total score which is the sum of the five previous metrics. Confidence intervals are computed with the LOYO procedure by computing the SHAP values for each of the 13 splits.

Figure S9 shows the SHAP values for all evaluation metrics: (a) ROC-AUC, (b) PR-AUC, (c) Deviance, (d) Dice, (e) FBSS and (f) Total score. The total score is the score used for optimizing the hyperparameters of the U-Net and is the sum of the five other metrics. The ROC-AUC and PR-AUC panels are discussed in the main text, and we focus here on the remaining four.

The Deviance panel is close to the ROC-AUC results: the lifted index is the dominant predictor for all single grid cell models, with large and consistent SHAP values, while convective precipitation, relative humidity and mixing ratio play a secondary but non-negligible role. This parallel is expected, since both metrics integrate performance over the full distribution of predicted probabilities and both are therefore dominated by performance on the majority class in this heavily imbalanced setting. The U-Net again shows a comparatively stronger contribution from convective precipitation, consistent with the pattern observed in the PR-AUC panel.

The Dice panel is similar to the PR-AUC panel in its overall structure, which is consistent with the fact that both metrics focus on the positive class. Like the PR-AUC, the Dice coefficient is relatively insensitive to true negatives and penalizes models that fail to predict observed events, so the importance of predictors tied to convective activity is expected to increase. The most striking feature of this panel is the markedly larger SHAP value of convective precipitation for the U-Net, substantially exceeding those of all single grid cell models. As discussed for the PR-AUC, this likely reflects the U-Net's ability to exploit the spatial clustering of convective precipitation, which aligns well with the spatial overlap measure that the Dice coefficient quantifies. As for the PR-AUC panel, the lifted index and convective precipitation have similar contributions for the single grid cell models.

The FBSS panel shows the most heterogeneous behavior across models and predictors, and the logistic regression results are particularly striking. Unlike all other models, LR exhibits negative SHAP values for the lifted index and mixing ratio, meaning that including these predictors is associated with degraded performance in terms of FBSS. Equally notable are the wide confidence intervals for LR on this metric, which are substantially larger than those observed for any other model-metric combination. These two features (i.e. negative importance and high variance) do not appear for LR on any other metric, including the Dice coefficient — which shares the sensitivity to predicted event frequency of the FBSS — and the deviance explained — which shares its skill-score structure. This specificity points toward the neighborhood-averaging component of the FBSS as the likely distinguishing factor. By smoothing predicted and observed fields spatially before scoring, the FBSS introduces a sensitivity to the spatial sharpness and local calibration of the forecast that point-wise metrics do not capture. A tentative interpretation is that the lifted index, being a spatially smooth field, may lead the LR model — which is constrained to a linear decision boundary — to produce overly diffuse probability forecasts that are specifically penalized under this spatial averaging, while non-linear models may learn locally sharper mappings that avoid this degradation. Why the other models are apparently unaffected, and whether the wide confidence intervals reflect a genuine instability of the FBSS optimum cannot be established from the SHAP values alone. These observations should therefore be treated as hypotheses that motivate dedicated follow-up analyses, such as inspecting the spatial structure of LR probability forecasts or comparing neighborhood-averaged calibration curves across model families.

The Total Score panel aggregates across all metrics and recovers the patterns seen individually. The lifted index remains the dominant predictor for all single grid cell models, and convective precipitation is again more important for the U-Net. The land-sea mask contributes little across all models and metrics, suggesting it provides limited marginal information once the thermodynamic and moisture predictors are included. The Total Score panel also reinforces the finding that the U-Net has a qualitatively different predictor importance profile from the four single grid cell models, which cluster together. This supports the conclusion that architecture choice and predictor selection are coupled decisions that should be made jointly with respect to the evaluation objective.

4 Saliency maps

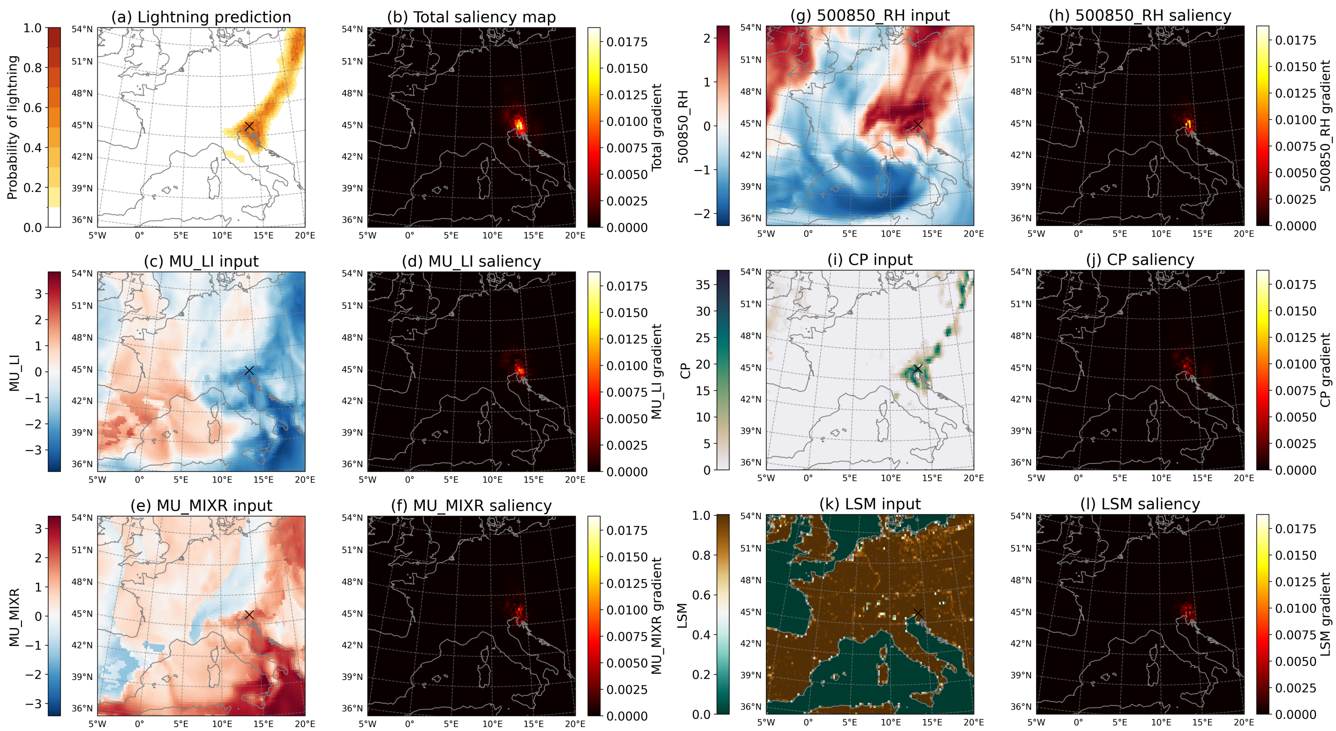

Figure S10. Feature-wise saliency maps for the 15th of August 2008 event. (a) shows the predicted probability of lightning at (14°E, 46°N). (c), (e), (g), (i), (k) display the five input fields (most-unstable lifted index, most-unstable mixing ratio, mean relative humidity between 500 and 850 hPa, convective precipitation and land-sea mask). (d), (f), (h), (j), (l) show the gradient of the predicted probability at (14°E, 46°N) with respect to each of these inputs. (b) is the pixel-wise L^2 norm of the gradient across the five input channels.

Figure S10 shows the feature-wise saliency analysis corresponding to the 15th of August 2008 event discussed in the main text. This is an example for a single lightning prediction at 21:00 UTC for a target pixel located near the head of the Adriatic Sea at (14°E, 46°N). It shows the predicted probability of lightning in panel (a), which forms an elongated band of high lightning probabilities extending from Slovenia into Central Europe, up to Poland. Lightning probability is about 0.6 at the target location. Panels (c), (e), (g), (i), (k) display the five input fields (most-unstable lifted index, most-unstable mixing ratio, mean relative humidity between 500 and 850 hPa, convective precipitation and land-sea mask), while panels (d), (f), (h), (j), (l) show the gradient of the predicted probability at (14°E, 46°N) with respect to each of these inputs. Panel (b) is the pixel-wise L^2 norm of the gradient across the five input channels.

The most striking feature across all saliency panels is the strong spatial localization of the gradients: sensitivity is concentrated within a few grid cells of the target pixel. This is true even where the input fields exhibit large-scale coherent structures, such as the synoptic-scale humidity band in panel (g) or the broad mixing ratio pattern in panel (e). The model's prediction at a given location therefore appears to depend almost exclusively on input values in a small neighborhood around that location, suggesting a limited effective receptive field.

Among the five inputs, relative humidity shows the largest gradient magnitude and is the clear leading contributor, followed by lifted index. The remaining three channels (mixing ratio, convective precipitation and land-sea mask) show smaller and broadly comparable saliency magnitudes, forming a second tier of contributors for this particular case. The ordering is physically plausible in that mid-level humidity and instability are precursors of deep convection, although these magnitudes are read from colorbars and should be treated as approximate rather than exact.

One noticeable point is that the spatial pattern of the CP input closely matches the shape of the predicted lightning probability, and CP emerges as the dominant feature in our SHAP analysis. Yet here CP appears only as a second-tier contributor in the saliency map. Such results are not necessarily contradictory: gradient saliency measures local sensitivity to infinitesimal perturbations at the current input value, whereas the SHAP analysis measures the marginal contribution of including CP relative to models trained without it. When CP is large and the model has learned a strong, near-deterministic mapping from CP to output, the local gradient can be small (a saturation-like effect) even though removing CP entirely would substantially degrade the prediction, which is what the global SHAP analysis captures. It should also be emphasized that Figure S10 shows a single prediction and is therefore illustrative rather than robust. Firm conclusions about the relative importance of CP would require aggregating saliency over many cases, which is what we do for the total gradient in the main text. The SHAP results presented in the main text remain the more reliable indicator of overall feature importance.

5 Attention Isotropy

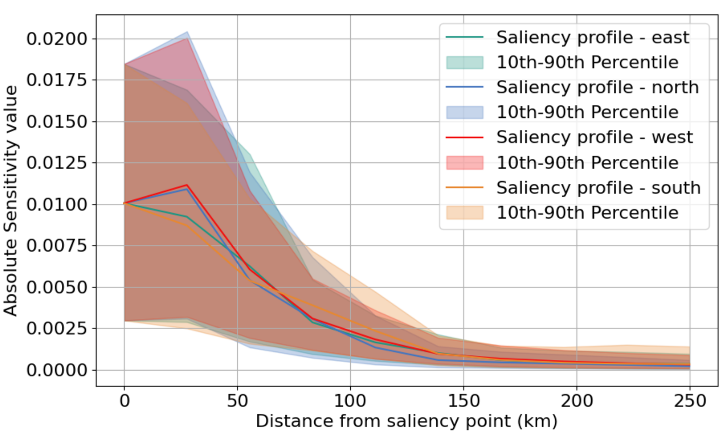

Figure S11. Isotropy analysis of the saliency maps.

Figure S11 is an extension of the Saliency radial profile presented in the main text. Here we look at the radial profiles for the locations where lightning was observed in the four cardinal directions: North, West, South and East. The observed decay of the saliency value with the distance to the location where the backpropagation is performed is similar to the one described in the main text.

We observe that the four median curves along with the four inter-quantile spreads are very similar across all directions. We only see small, non-significant increases in saliency for the Northern and Western profiles. This suggests the model's effective receptive field is approximately isotropic.

References

ARSO: Report about the extreme rainfall between 3rd and 6th of August 2023, https://meteo.arso.gov.si/uploads/probase/www/climate/text/sl/weather_events/padavine_3-6avg2023-preliminarno.pdf, [Accessed 07-05-2026], 2023.

Furnari, L., Mendicino, G., and Senatore, A.: Hydrometeorological Ensemble Forecast of a Highly Localized Convective Event in the Mediterranean, Water, 12, 1545, https://doi.org/10.3390/w12061545, 2020.