Answer to Reviewer 1 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 1 for their relevant review, which improved both the rigor and the clarity of the manuscript. Their comments led us to better account for the dependence between the leave-one-year-out splits in our statistical tests and confidence intervals, to check the U-Net for overfitting, and to replace the arbitrary threshold that defined the radius of useful information. They also led us to summarize in the main text the seasons and case studies that were previously only in the supplementary material, to add relative error maps, and to separate the Shapley values into individual and interaction effects.

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 "Lines X-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 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.

ANSWER: We thank the reviewer for their positive assessment of our work and for their constructive comments, which helped us improve the manuscript. Besides the point-by-point answers below, we summarize here the main changes between the original and the revised manuscripts. 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 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 pressurelevel, 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.

ANSWER: We agree that the sentence implied the five variables were all derived the same way. Section 2.2 now separates them into two groups. The most-unstable lifted index, the most-unstable mixing ratio and the mean relative humidity between 500 and 850 hPa are computed with the thundeR package from the vertical profiles on the 137 hybrid sigma-pressure levels of ERA5. Convective precipitation (the hourly accumulation from the ERA5 convection scheme) and the land-sea mask are single-level fields taken directly from ERA5.

CHANGES: Lines 114-117: "We follow Battaglioli et al. (2023a) and use three variables derived from the ERA5 reanalysis dataset on 137 hybrid sigma-pressure levels (Hersbach et al., 2020): (i) most-unstable lifted index, (ii) most-unstable mixing ratio and (iii) average relative humidity between 500 and 850 hPa. We also use two other variables downloaded directly from the ERA5 reanalysis dataset: (iv) convective precipitation and (v) a land-sea mask."

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.

ANSWER: We agree that the acronym should be explained. We removed it completely as it appeared only once in the manuscript. The text now gives the full name of the model, "Additive Regressive Convective Hazard Model", in the same sentence as the citation to Battaglioli et al. (2023a).

CHANGES: Lines 117-118: "These variables were selected as the best combined predictors of lightning within the Additive Regressive Convective Hazard Model in Battaglioli et al. (2023a)."

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

ANSWER: Section 3.5 has been removed, and its content is now the last paragraph of Section 3.3 (Training pipeline).

CHANGES: Lines 202-203: "We also perform a feature importance study for each model along with an interpretability study of the predictions of the U-Net. The methodology associated with each study is detailed in the corresponding paragraphs of 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.

ANSWER: We have updated Table 1 to add a "Range" column in which the bounds of each metric are reported.

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.

ANSWER: We thank the reviewer for this suggestion. We evaluated the final U-Net, the most flexible of our models, on its 13 training years (new Section S1.2 and Table S1 of the supplementary material). The training and test scores are very close. On the full period, the training-year and held-out-year scores are:

- ROC-AUC: 0.969 against 0.970
- PR-AUC: 0.180 against 0.179
- Dice: 0.126 against 0.125
- FBSS: 0.223 against 0.222
- Deviance: 0.349 against 0.360

Some scores are slightly higher on the training years and others on the test years, and all the differences are within the uncertainty of the LOYO scores (Figure 3). We therefore find no evidence that the U-Net overfits the training data. Section 4.1 now mentions this result.

CHANGES: Lines 243-244: "We also evaluate the final U-Net on its training years in the supplementary material (Table S1) and find no evidence that it overfits the training data."

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 testset 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.

ANSWER: We thank the reviewer for this very relevant remark. The test we performed was already a paired Student's t-test (scipy.stats.ttest_rel) on the 13 per-split differences between the U-Net and each other model, but the manuscript did not say so. Section 4.1 now states it.

We did not switch to the Wilcoxon signed-rank test because it would not change the conclusion. The U-Net outperforms every other model on all 13 splits, for every metric, both on the full test sets and on the extreme days. The exact Wilcoxon p-value is therefore always the smallest possible with 13 pairs: $2/2^{13} \approx 2.4 \cdot 10^{-4}$ (Wilcoxon, 1945).

We agree that the paired test still assumes independent splits. To account for the overlap of the training sets, we now apply the correction of Nadeau and Bengio (2003), which was designed for repeated cross-validation. It inflates the variance of the mean difference from $\hat{\sigma}^2/k$ to $(1/k + n_{\text{test}}/n_{\text{train}})\,\hat{\sigma}^2$, with $k = 13$ and $n_{\text{test}}/n_{\text{train}} = 1/12$. This multiplies the t-statistic by about 0.69. With the correction, the p-value remains below 10^{-6} for every model and metric, so our conclusions are unchanged. The caption of Figure 3 and Section 4.1 now refer to this corrected test.

We applied the same correction to the confidence intervals of Figure 3 and of the SPVIM values (Section 4.4, Figure 7 and the corresponding supplementary figures). It widens these intervals by a factor $\sqrt{1 + 13/12} \approx 1.44$. Appendix B describes the correction and acknowledges the limitation the reviewer raised: the correction was derived for repeated random train/test splits, so for LOYO cross-validation it remains an approximation and does not fully account for the dependence between splits. It does, however, give more conservative intervals than assuming that the splits are independent.

CHANGES: Lines 250-251: "[...] the corrected paired Student's t-test from Nadeau and Bengio (2003) [...]".

Figure 3 caption : "Bars show the mean across splits and error bars the 95% confidence interval, corrected for the overlap of the training sets (Nadeau and Bengio, 2003) (see Appendix B)."

Lines 580-583: "However, the training sets of any two splits share 11 of their 12 years, so the test metrics of the different splits are correlated rather than independent, and the usual standard error $\frac{\sigma}{\sqrt{n}}$ underestimates the uncertainty of their mean. We therefore apply the correction of Nadeau and Bengio (2003), which inflates the variance of the mean from $\frac{\sigma^2}{n}$ to $\left(\frac{1}{n} + \frac{n_{\text{test}}}{n_{\text{train}}}\right)\sigma^2$, and we compute the 95% confidence interval as $\left(\bar{x} \pm t_{\alpha/2,n-1}\sigma\sqrt{\frac{1}{n} + \frac{n_{\text{test}}}{n_{\text{train}}}}\right)$"

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.

ANSWER: We thank the reviewer for this comment. We have revised Section 4.2 on three points.

- First, the XGB is now the best-calibrated model below a predicted probability of 0.5, with the U-Net close behind. Above 0.5, the U-Net is the best-calibrated model. This result comes from the re-optimization of the models following the comments from Reviewers 2 and 3.
- Second, Section 4.2 now states explicitly that the residual miscalibration is acceptable for our analyses.
- Third, Platt scaling is applied only to the MLP and the U-Net, because they are the only models trained with a weighted loss, which inflates their raw output (see our answer to Reviewer 2, L236–241). The scaling layer is part of these two models, so every result reported for them uses the calibrated output. The uncalibrated output appears only in Figure 4a and Figure S2, for comparison. LR, GAM and XGB are trained with the unweighted log-loss and are therefore calibrated by construction. Section 3.2 now states this, and the new Section S1.3 of the supplementary material shows that Platt scaling leaves these three models almost unchanged.

CHANGES: 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."

Lines 257-263: "The XGB model is the best calibrated model at predicted probabilities below 0.5, with the U-Net close behind. For probabilities higher than 0.5, the U-Net is the best calibrated model. We see that all models tend to overestimate the higher probabilities – starting at 15% chance for LR and up to 40% chance for XGB. The models overestimate more the very high probabilities (above 70% chance of occurrence). The residual miscalibration is considered acceptable for our further analyses. Figure S2 in the supplementary shows similar results for the calibration on the most extreme days: XGB being the best calibrated model, followed by the U-Net. Section S1.3 features additional information on the calibration of the models."

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 singlegrid-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.

ANSWER: We thank the reviewer for these suggestions.

On the other seasons, Section 4.3 now summarizes them. The U-Net's advantage is not specific to summer convection: it has the lowest RMSE in every season. It reduces the RMSE of the best single grid cell model by about 36% in winter, 5.5% in spring, 15% in summer and 16% in autumn. The gain is largest in winter, when lightning is concentrated over the Mediterranean Sea and the U-Net largely corrects the strong underestimation that all single grid cell models show there. It is smallest in spring, when the single grid cell models come closest to the U-Net. Section S1.4 of the supplementary material analyses the four seasons in more detail.

On the relative error, we computed (model $-$ observations) / observations at each grid cell for every season and model, and added these maps to Section S1.4 (Figures S5, S7, S9 and S11). In every season, the U-Net has the largest share of grid cells with a relative error between -25% and $+25\%$: 41%, 54%, 56% and 39% of the cells in winter, spring, summer and autumn, against at most 25%, 51%, 49% and 31% for the single grid cell models. The relative error maps also show that all models overestimate lightning in autumn over the north of the domain, where little lightning is observed. This is consistent with the yearly cycle in Figure 5.

We kept the absolute bias maps in the main text, for two reasons:

- Many grid cells have very little observed lightning. Over the three test years, the share of grid cells with at most one lightning hour per year is 70% in winter, 30% in spring, 13% in summer and 25% in autumn. The share with no lightning hour at all is 39%, 13%, 4% and 8%. Where lightning is scarce, the relative error is very volatile, because a small absolute error gives a large relative error. Predicting 4 lightning hours in a cell with 2 observed gives a relative error of 100%, whereas the same 2-hour error in a cell with 30 observed gives about 7%.
- For impact studies such as wildfire ignition or damage to infrastructure, the regions of interest are those with high lightning activity. The absolute bias maps give the most weight to these regions. Each figure also shows the observed climatology in panel (a), so the reader can relate the bias at each location to the local amount of lightning.

We report the RMSE as a quick check to compare the models at a glance. We show the observed climatology in panel (a) to give the baseline magnitude at each location. We agree that this metric is not perfect as it aggregates

the error over all pixels and this is why the true evaluation on which we rely is the scores that we report in Section 4.1.

CHANGES: Lines 299-304: "The advantage of the U-Net is not specific to summer convection: it also achieves the lowest RMSE in winter, spring and autumn (see Figures S6, S8, and S10), reducing the RMSE of the best single grid cell model by about 36% in winter, 5.5% in spring, 15% in summer and 16% in autumn. The gain is largest in winter, when lightning is concentrated over the Mediterranean Sea and the U-Net largely corrects the strong underestimation shown there by all single grid cell models, and smallest in spring, when the performance of the single grid cell models is closer to the performance of the U-Net. We also show in Figure S4 that the U-Net tends to make smoother predictions."

Lines 305-306: "We show absolute bias maps here because some regions have very few lightning hours which make relative bias maps unstable and less reliable, but we include those additional maps in the supplementary material (see Figures S5 to S11)."

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.

ANSWER: As with the seasonal maps, Section 4.5 now ends with a short summary of the two other events.

CHANGES: Lines 400-404: "The same qualitative behavior is found for the two other extreme events of the test set, on the 11th of August 2015 over southern Italy and on the 5th of August 2023 from the Adriatic Sea to Poland (Figures S12, S13, S14, S15): the U-Net obtains the best scores on all evaluation metrics and it better reproduces the spatial extent of the active region, which the single grid cell models underestimate. In both cases, the U-Net's power spectrum is again the closest to the observations at small wavenumbers, while it remains the smoothest of all models at large wavenumbers."

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.

ANSWER: The 1/10 threshold was indeed arbitrary. Both radii are now defined with the "Kneedle" variant of the elbow method (Thorndike, 1953; Satopää et al., 2011), which finds the point where the decay of the profile levels off without requiring a threshold (Section 4.6 and Figure 12). To show how sensitive the radii are to the cutoff, we also computed them with fixed thresholds:

- Saliency maps: the elbow is at about 140 km. Thresholds of 1/5, 1/10 and 1/20 of the maximum median saliency would give radii of about 110, 135 and 190 km.
- Radial perturbation: while addressing a comment from Reviewer 2, we found and corrected an error in the perturbation. The random values were drawn in physical units, whereas the model receives standardized inputs. We repeated the analysis for radii up to 1500 km (see our answer to Reviewer 2, L379). With the corrected method, the elbow is at about 500 km, where the median relative change of the prediction is about 10%. Thresholds of 20%, 10% and 5% on this median change would give radii of about 400 km, 500 km, and between 500 and 750 km. The change from 320 km in the previous version therefore comes from this correction, not from the new definition of the radius.

The radii therefore vary by a few tens of percent with the cutoff, but their order of magnitude does not change, and neither does our interpretation. The U-Net is most sensitive to the predictors within about 100–200 km of the pixel of interest, and it uses mesoscale context up to about 400–750 km. We report both radii as approximate values ($R \sim 140$ km and $R \sim 500$ km).

CHANGES: Lines 435-436: "We define a radius of useful information R using the "kneedle" variant of the Elbow method (Thorndike, 1953; Satopaa et al., 2011)."

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."

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 × 37) kernel centered on the pixel of interest. At this radius, the relative error is about 10%."

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

105 *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.*

ANSWER: We thank the reviewer for this very relevant comment. Following the comments of Reviewers 2 and 3, we now call these values SPVIM (Shapley population variable importance measure; Williamson and Feng, 2020) rather than SHAP values, because they attribute the performance of a model, not individual predictions. As the reviewer points out, retraining each model on all 32 subsets means that each value includes the predictor's interactions with the other predictors. The caption of Figure 7 now states this.

To quantify the role of these interactions, we decomposed each SPVIM value into two terms (new Section S3.2 and Figure S17):

- the individual effect of the predictor, i.e. the skill of the model trained on this predictor alone compared with no predictor;
- an interaction effect, which gathers its interactions with all subsets of the other predictors. A positive interaction effect means the predictor is complementary to the others; a negative one means it is redundant with them.

A new paragraph in Section 4.4 summarizes the results:

- The single grid cell models have positive interaction terms for every predictor except convective precipitation, so they exploit the synergy between predictors. The U-Net has negative interaction terms for every predictor: once the model has access to their spatial patterns, the predictors carry redundant information.
- The U-Net's stronger reliance on convective precipitation comes from these interactions, not from the standalone effect of this variable. For every model, the individual effects rank convective precipitation first, with about twice the effect of the lifted index, and the U-Net increases the individual effects of both predictors by similar factors. The difference comes from the lifted index. In the single grid cell models, it gains skill by combining with the other predictors (its interaction effect is +65% to +79% of its individual effect), whereas in the U-Net it is partly redundant with them. Convective precipitation is the least redundant predictor for the U-Net.
- We relate this to the spatial structure of the predictors (Section S3.3, Table S4). Convective precipitation is the only predictor clustered at a scale close to that of lightning: its e-folding distance is 90 km, against 60 km for lightning and 490 to 920 km for the other predictors. It therefore carries local information that the smooth fields of the other predictors do not.

CHANGES: Lines 341-356: "In Figure 7, each SPVIM value comprises the individual effect of the predictor and the effect of the interactions of that predictor with the other predictors. We provide more details on the decomposition between the individual effect and the interaction effect in the supplementary material in Section

S3.2. We show an example of such decomposition for the PR-AUC score in Figure S17. We find that contrary to the single grid cell models, the U-Net has negative interaction terms for every variable, which means that the predictors have redundant information and that the SPVIM value is driven mostly by the standalone effect of each predictor rather than by the interaction effect. The single grid cell models however have positive interaction terms for all predictors, except convective precipitation, which means that they take advantage of the synergy between predictors. We also find that the shift in the most important feature – from lifted index to convective precipitation – does not come from the individual effects, but from the interactions. The individual effects rank convective precipitation first for every model, with about twice the effect of the lifted index. The interactions however show a different picture: the lifted index is complementary to the other predictors in single grid cell models, and partly redundant with them in the U-Net. Relative to its standalone effect, convective precipitation is the least redundant predictor for the U-Net. This interaction difference results in a global importance discrepancy between the U-Net and the single grid cell models. We show in Table S4 of the supplementary material that convective precipitation is the only predictor that carries local information on where convection occurs: it is clustered at a scale close to that of the lightning observations, whereas the other predictors vary on scales 8 to 15 times longer, which can explain why it is the least redundant field for the U-Net."