Answer to Reviewer 3 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 3 for their constructive comments, which challenged several of our interpretations. They pointed out that the U-Net gains both spatial context and knowledge of location over the single grid cell models, which the manuscript now states explicitly. Their suggestions led us to add performance diagrams and the held-out scores of all models, and to test our land–sea explanation of the autumn overestimation. They also led us to relate the importance of convective precipitation for the U-Net to the spatial structure of the predictors, and to revise our explanations of the smoothness of the U-Net predictions and of its over-prediction over the Alps.

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 "LX-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 Main changes

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 2

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.

> **ANSWER:** We agree that the statement was too general. The U-Net is the only spatial model in our comparison, so the difference in feature importance may come from this particular architecture rather than from spatial models in general. We have rephrased the abstract and have included to the discussion (Section 5) that adding other state-of-the-art spatial architectures to the comparison would make the results more robust.
> Within the models we test, the ranking does not depend on model capacity or hyperparameters. The single grid cell models range from a logistic regression to gradient-boosted trees and an MLP tuned in each LOYO split, and all four have similar SPVIM values. The shift towards convective precipitation as the main predictor appears only for the U-Net. The SPVIM values of the U-Net come from 13 networks tuned separately in each LOYO split, so the values are not tied to one hyperparameter set.
>
> **CHANGES**: Lines 9-10: "the most important predictors differ between the U-Net and the single grid cell models."
> Lines 520-521: "Repeating the study over different sets of data, predictors and state-of-the-art spatial models can improve the robustness of the results."

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 behavior 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.

ANSWER: We agree with the reviewer: fire is a natural part of many ecosystems, and fire suppression and warming are changing fire regimes. So that we do not present every lightning-ignited wildfire as a disaster, we have replaced "resulting in" with "which can result in".

CHANGES: Lines 15-16: "lightning is the leading natural cause of wildfire ignition,15 contributing to a large portion of the total burned area and which can result in [...]"

L26: In "convective convective processes", the word is duplicated.

ANSWER: Correction applied.

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.

ANSWER: Indeed, convective precipitation comes from the IFS convection scheme, which is a short-range forecast accumulation. All our variables are hourly so we use the 1h accumulation window of convective precipitation output from ERA5. Therefore the models link "how much convective precipitation occurred during one hour" to "was there occurrence of lightning within that same hour". Section 2.2 now clearly states that we use the ERA5 output. The SPVIM analysis trains every model on all 32 predictor subsets, including the subset without convective precipitation. We find that it is the main predictor for the CNN-based model.

CHANGES: Lines 116-117: "We also use two other variables downloaded directly from the ERA5 reanalysis dataset: (iv) convective precipitation [...]"

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 neighborhood statistics (mean, max, gradients over 3x3 to 23x23 windows).*

ANSWER: We thank the reviewer for their very interesting comment. Indeed we agree that here the added spatial information encodes two things at once: spatial context and knowledge of location, and that our current experiments do not allow to separate each contribution. Our aim is to measure the gain from the spatial information as a whole over the single grid cell schemes used in the literature (e.g. Czernecki et al., 2019; Battaglioli et al., 2023; Cavaiola et al., 2024; Mazurek et al., 2025; Saha et al., 2025). We agree that separating the impact of spatial context knowledge and location knowledge is a very interesting addition and keep this as a future direction. We have reformulated the introduction and Section 3.1 to clarify that we do not separate the two contributions, and have added this as a limitation and a perspective in the discussion.

CHANGES: Lines 82-84: "This spatial information is two-fold: the U-Net has access to both the location in the domain of each pixel and the spatial context surrounding each pixel. This study assesses the overall impact of the additional spatial information and does not isolate the individual contribution of spatial context and location."

Lines 147-154: "Compared to the single grid cell models, the U-Net has access to two distinct sources of spatial information. The first is the spatial context: the prediction at a grid cell depends on the predictors in the surrounding grid cells, within the receptive field of the convolutional layers. This is the information that we expect to characterize organized convective systems. The second is the location of the grid cell. Since the U-Net always processes the same fixed domain, it can infer where a grid cell lies from static features of its input: the coastline pattern of the land-sea mask around the grid cell, the persistent geographic imprint of the other predictors (e.g. orography) and distance to the image boundaries (Islam et al., 2020; Semih Kayhan and van Gemert, 2020). Our comparison measures the combined effect of both sources and does not separate the contribution of the spatial context from that of the location information."

Lines 513-519: "The fixed domain also means that the gain of the U-Net over the single grid cell models has two sources that our setup does not separate: the spatial context provided by the surrounding grid cells, and the location of each grid cell, which the U-Net can infer from static features of its input. Future work could isolate the contribution of the spatial context by giving the single grid cell models the latitude and longitude of each grid cell as additional inputs, so that both model families have access to the location and differ only by the spatial context. Conversely, training the U-Net on randomly positioned sub-domains would remove its knowledge of the absolute location while keeping the spatial context. Such a model, not tied to a fixed domain, could also be used to infer lightning probabilities in regions where lightning observations are scarce"

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.

> **ANSWER:** We agree that depth is linked to the use of spatial context, since it sets the theoretical receptive field. However, with two downsampling levels, the theoretical receptive field of our U-Net is already 137×137 pixels. This is 60% of the domain on average: 30% for a pixel in a corner and 90% for a pixel in the center. The model therefore already has access, in theory, to far-away mesoscale information. The relevant question is whether the model actually uses this information, i.e. what its effective receptive field is. We evaluate it with the saliency maps and the radial perturbation method of Section 4.6. The effective receptive field is much smaller than the theoretical one: 37×37 pixels (a radius of about 500 km), against 137×137. Note that the receptive field does not depend on depth alone. It is also controlled by the kernel sizes and on the number of convolutional layers at each level.

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.

> **ANSWER:** We thank the reviewer for noticing this. Platt scaling is indeed defined on logits, and this is what we have done; our phrasing did not describe the methodology accurately. We have changed "the input is the predicted probability from the U-Net (or MLP)" to "the input is the logit predicted by the U-Net (or MLP) – i.e. the value before the final sigmoid activation –". Section S1.3 of the supplementary material now details the scaling and its fitted coefficients.
>
> **CHANGES**: Lines 170-172: "To address this issue, we apply a post-hoc calibration method called Platt scaling (Platt, 1999; Guo et al., 2017), which fits a logistic regression where the input is the logit predicted by the U-Net (or MLP) – i.e. the value before the final sigmoid activation – and the output is the true binary label of lightning occurrence."

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.

ANSWER: We agree with the reviewer that the padding type was missing from the manuscript. We apply reflection padding, and the model is scored only on the interior, i.e. the initial 101×149 domain. We have added this information to Section 3.2.

CHANGES: Lines 158-159: "which is subsequently padded to 112×160 pixels with reflective padding. "

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.

ANSWER: We thank the reviewer for pointing out this ambiguity. The inhomogeneity comes from the Met Office's new lightning detection network, LEELA, introduced at the beginning of 2023. The former ATDnet network kept running operationally until the end of 2023. We only use ATDnet data and do not combine them with LEELA data, to avoid the inhomogeneity introduced by a new detection technology. This is also why our data stop at the end of 2023. The data we use are therefore homogeneous. We have clarified this in Section 2.1.

CHANGES: Lines 98-99: "To ensure consistency, we use a single lightning detection technology (ATDnet) that has run homogeneously from 2008 to the end of 2023. It was replaced in 2023 – with data overlapping for one year – by the new LEELA system."

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 neighborhood sizes.

ANSWER: We agree with the reviewer that the Fractions Brier Skill Score is non-standard. Roberts and Lean (2008) define the fractions skill score (FSS, their Eq. 6) as:

$$\mathrm{FSS} = 1 - \frac{\mathrm{MSE}_{\mathrm{model,\,neighborhood}}}{\mathrm{MSE}_{\mathrm{ref,\,neighborhood}}}$$

The two differences in our score are:

- our reference is the climatology (instead of the worst-case model as in Roberts and Lean (2008)), so that we use the same baseline as for the deviance
- we use continuous forecast fields instead of binary fields. This is exactly how the Brier score (Brier, 1950) generalizes the MSE to classification tasks

So our score is:

$$\mathrm{FBSS} = 1 - \frac{\mathrm{Brier}_{\mathrm{model,\,neighborhood}}}{\mathrm{Brier}_{\mathrm{clim,\,neighborhood}}}$$

We have clarified the FBSS definition in Section 3.4 and Appendix D.

CHANGES: Lines 208-209: "[...] the Fractions Brier Skill Score (FBSS) – which measures the Brier score (Brier, 1950), averaged over a neighborhood (Roberts and Lean, 2008)"

Lines 657-662: "The Fractions Skill Score (Roberts and Lean, 2008) is defined as $\mathrm{FSS} = 1 - \frac{\mathrm{MSE}_{\mathrm{model,\,neighborhood}}}{\mathrm{MSE}_{\mathrm{ref,\,neighborhood}}}$. In Roberts and Lean (2008), the MSE is computed between two binary fields: the observations and the prediction, and the reference is the worst-case scenario where the two fields never overlap. The Brier Score (Brier, 1950) extends the mean squared error (MSE) to classification tasks by measuring the mean squared error between the predicted probabilities and the true binary labels: $\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$. We use this to avoid thresholding the predictions."

Line 666: "$\mathrm{FBSS} = 1 - \frac{\mathrm{BS}_{\mathrm{model,\,neighborhood}}}{\mathrm{BS}_{\mathrm{ref,\,neighborhood}}}$"

80 *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.*

ANSWER: We thank the reviewer for this suggestion and agree that the performance diagram (Roebber, 2009) is a standard tool in convective verification. Our decision to report only threshold-independent metrics was deliberate. The threshold used to turn probabilities into binary warnings depends on the application and on each end user's costs and losses. Reporting POD, SR, CSI and frequency bias at one threshold would tie the evaluation to that arbitrary choice.

To connect with this verification framework while keeping our choice, we have added performance diagrams (Figure S1). Each curve is traced by varying the probability threshold from 0 to 1, rather than evaluated at a single operating point. Readers can see where each model sits in POD–SR space and which CSI and bias values are reachable, without our committing to one threshold. We analyze this figure in Section S1.1 of the supplementary material.

CHANGES: We have added Section S1.1 and Figure S1.

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.

ANSWER: We have added the scores of the four single grid cell models, as well as those of the climatology, to Table 2. The U-Net remains the best model for all metrics.

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.

ANSWER: We thank the reviewer for this comment and agree that the calibration curves needed more explanation. Table A1 described a previous model. We have re-optimized the XGB model following the second reviewer's comments: XGB is now trained with the binary logistic objective (binary cross-entropy), not a quantile regression loss, and we have corrected the table. The drop of the XGB curve came from the highest probability bins, which contained very few samples. XGB predicts a probability above 0.65 for only 238 of the $\sim 1.7 \times 10^9$ test samples, and never above 0.78. The observed frequency in these bins was therefore not statistically meaningful: it was zero whenever none of their few samples had lightning. Figure 4 now uses adaptive bin sizes so that every bin contains at least 50 samples, and the XGB curve no longer drops to zero.

CHANGES: We have added updated Table A1.

Figure 4 caption : "To ensure each point is statistically meaningful, we use adaptive bin sizes such that every bin contains at least 50 samples."

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.

ANSWER: We want to test how well each model reproduces the cycles without any additional information. No model, the U-Net included, has access to the hour of the day or the day of the year. Therefore, we argue that giving this information only to the single grid cell models would make the comparison unfair.

Should the U-Net manage to infer the time of day or the day of year based on the spatial information it has access to, then it is relevant to credit the gain – or part of it – to this spatial information.
Note that following the reviewer's comment on L133–138, we now refer to "spatial information", i.e. spatial context and location, rather than to "spatial context" alone.

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.

ANSWER: Indeed our attribution to the ERA5 resolution was too quick, and the analysis was not precise enough. The U-Net over-predicts over the main Alpine ridge, which is known to be a relative minimum of lightning activity (Bertram et al., 2004; Manzato et al., 2022; Feldmann et al., 2023). The single grid cell models under-predict south of this ridge, where the U-Net performs better. The U-Net's over-smoothing seems to prevent it from distinguishing the subtle boundary between the southern slopes and the main ridge. We observe a similar but less intense pattern over the Pyrenees. We have reformulated our analysis accordingly (Section 4.3 and Section S1.4). We have added Figure E1 to define the regions precisely.

CHANGES: Lines 289-290: "We use regions defined in Figure E1 in this paragraph."

Lines 292-295: " One such bias is the underestimation of lightning in mountainous regions: over the southern Alpine foothills (+2h for the U-Net against -9h to -20h for the single grid cell models), over the Pyrenees (0h against -2h to -11h), over the Apennines (-5h against -11h to -15h), over the Dinaric Alps (-12h against -16h to -19h), and over the Iberian chain (-16h against -18h to -26h)."

Lines 296-298: " The error made by the U-Net is overall smaller but it still over-predicts over the main alpine ridge compared to the other models (+23h against -20h to +2h) and North of the Pyrenees (+6h)."

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.

ANSWER: We thank the reviewer for this suggestion. The explanation is indeed testable. We have added the yearly cycles over land and sea separately (Figure S3) and the relative error over land and sea in September, October and November (Table S3, Section S1.4). Lightning activity does shift from land to sea in autumn, but the overestimation is made over land, not over sea. In October, the models overestimate the number of lightning hours by +48% to +74% over land, against -16% to +18% over sea: the predictions over land do not drop as

fast as the observations. Our statement was therefore incorrect, and we have reformulated it in Section 4.3.

CHANGES: Lines 281-287: "We note an overestimation of all models in September and October. We show the relative error between the total predicted lightning hours and the observed lightning hours in September, October and November over: all pixels, land pixels and sea pixels in Table S3 of the supplementary material. We find that the autumn overestimation is driven by an overestimation of lightning over land. Lightning activity shifts towards the Mediterranean Sea and predictions do not drop as fast. We show in Figure S3 that the U-Net represents the yearly cycles over land and sea as well as or better than the GAM, XGB and MLP models. However these models have over- and underestimations over land and sea respectively that cancel each other out. This explains their smaller overall Jensen-Shannon distance that we see in Figure 5(b).

105 **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.

ANSWER: We thank the reviewer for this comment and interesting suggestion. The "SHAP" naming was indeed incorrect: the quantity we compute is the Shapley Population Variable Importance Measure (SPVIM; Williamson and Feng, 2020), a Shapley value of predictive performance. We have corrected the naming throughout the manuscript.

We agree that Shapley values split credit among correlated inputs and that this could drive the architecture dependence. However, the correlations between predictors are a property of the data and are the same for all models. They therefore cannot, by themselves, explain why the importance of a predictor differs between architectures; what differs is how each model exploits the information that the predictors share. Moreover, pairwise correlations only measure linear overlap and cannot tell redundancy from complementarity. Two weakly correlated predictors can be strongly complementary, as instability and moisture are for convection.

We therefore quantified directly how the credit is shared in each model, by decomposing the SPVIM values into individual and interaction effects (new Section S3.2 and Figure S17). The individual effect is the skill a model gains from the predictor alone, so it involves no credit splitting. The interaction effect gathers the credit that the predictor shares with all subsets of the other predictors. It is positive when they are complementary and negative when they are redundant.

CHANGES: Lines 341-347: "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."

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.

ANSWER: We thank the reviewer for this comment and agree that this explanation was not tested in the original manuscript. The decomposition of the SPVIM values into individual and interaction effects (Section S3.2, Figure S17) shows that, compared with the GAM, XGB and MLP, the U-Net increases the individual effects of the lifted index and of convective precipitation by similar factors (about $\times 2$). The larger relative importance of convective precipitation therefore may not come from a better use of its spatial structure, but from the interactions. In the GAM, XGB and MLP, the lifted index gains skill by combining with the other predictors: its interaction effect is +65% to +79% of its individual effect. In the U-Net, all predictors are redundant, convective precipitation the least.

To understand why, we characterized the spatial structure of each predictor on the 49 extreme days of the test set (new Section S3.3, Table S4). We used two metrics: the e-folding distance of the spatial autocorrelation, and the frequency-matched Fractions Skill Score (Roberts and Lean, 2008) against neighborhood size (note that here we apply thresholding on the predictors to use the standard FSS and not the FBSS discussed above). Here are the two main findings:

- Convective precipitation is the only predictor clustered at a scale close to that of the lightning: its e-folding distance is 90 km, against 60 km for the lightning and 490 to 920 km for the other predictors.
- It is also the only predictor that locates the lightning at the mesoscale: its FSS is useful from 230 km, against 960 km or more for the other predictors.

Convective precipitation therefore carries local information on where convection occurs. The other predictors are smooth fields that describe where storms can occur. Once the U-Net sees their spatial patterns, these fields likely describe the same air masses and overlap, which makes them redundant. We have rephrased the sentence

accordingly.

CHANGES: Lines 347-356: "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."

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.

ANSWER: We thank the reviewer for this comment and agree. We have reformulated Section 4.5. The smoothness of the U-Net predictions is not imposed by the convolutional architecture, since the skip connections give the decoder access to high-resolution features. It is rather expected from a pixel-wise loss. Since the ERA5 predictors only weakly constrain the exact location of convective cells, the model is rewarded for spreading probability over all plausible locations.

CHANGES: Lines 391-399: "On the other hand, the small-scale features – which are associated with large wavenumbers on the power spectrum – are not well captured by any model, the U-Net being worse than the others. This smoothness is not imposed by the architecture, since the skip connections of the U-Net give its decoder access to high-resolution features; we rather attribute it to the difficulty of the task. The exact location of convective cells is only weakly constrained by the 0.25° ERA5 predictors, in which convection is parameterized, so a model trained with a pixel-wise loss is rewarded for spreading probability over all plausible locations rather than predicting a sharp but possibly misplaced pattern."

4.6 Interpretability

120 *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.*

ANSWER: We thank the reviewer for this comment. We first clarify the protocol, which was not described precisely enough. The random values are drawn independently at each pixel from the normal distribution fitted to each predictor over the training years. They are then clipped to the physical range of the predictor (e.g. non-negative precipitation, relative humidity between 0 and 1), so each value remains within the range seen during training. We have added this detail to Section 4.6.

We agree that the perturbed images, with spatially uncorrelated values and a sharp edge at the radius R, are not realistic. This is inherent to occlusion-based attribution methods, of which our method is a radial variant (Zeiler and Fergus, 2014). Replacing the pixels beyond R with the same region on another date would also create a discontinuity at R, between two unrelated weather situations, which the model could equally react to.

The purpose of the method is to determine how far from the pixel of interest the inputs still influence the prediction. Whether the prediction changes because information is lost or because the model reacts to the perturbation, the change shows that the model is sensitive to these pixels. The method therefore gives an upper estimate of the radius of useful information. It complements the saliency maps, which use realistic inputs but are known to underestimate this radius.

CHANGES: 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. "