https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

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

Abstract. Most lightning parameterization schemes rely on local approaches where the predictors are the atmospheric variables in the same grid cell as the output. To validate the hypothesis that large-scale thunderstorm clusters – such as mesoscale convective systems – are driven by broad spatial predictor patterns, we model lightning occurrence using architectures capable of processing surrounding grid-cell data rather than relying solely on local point-based inputs. This study develops a deep convolutional neural network (U-Net) to model lightning occurrence across Europe using ERA5 reanalysis data. The model is trained on 13 years of data and evaluated with a leave-one-year-out cross-validation strategy. We compare the performance of the U-Net to several local machine learning models of increasing complexity, including logistic regression, generalized additive model, extreme gradient boosting, and multi-layer perceptron. We find that the U-Net outperforms all single grid cell models in overall performance and on the most extreme events. Through a feature importance study, we find that the most important predictors depend on the model type. Finally we show with a spatial sensitivity study that the U-Net captures mesoscale patterns driving lightning occurrence.

1 Introduction

Lightning is a major natural hazard that poses significant risks to human life, ecosystems, and critical infrastructure. Globally, lightning is responsible for between 6,000 and 24,000 fatalities annually (Holle, 2016), including about 60 per year in Europe (Kühne et al., 2025). Beyond its direct impact on human safety, lightning is the leading natural cause of wildfire ignition, contributing to a large portion of the total burned area and resulting in significant ecological damage, loss of biodiversity, and release of toxic fumes and carbon emissions (Pérez-Invernón et al., 2023; Huntrieser et al., 2016; Aguilera et al., 2021). In addition, lightning poses a threat to the energy infrastructure, from transmission grids to renewable energy installations and

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

gas pipelines (Haseeb and Krawczuk, 2025; Li et al., 2024; Charalambous et al., 2026). Understanding the climatology of lightning with its past and future trends is therefore crucial for risk assessment and adaptation planning in a changing climate (Surowiecki et al., 2024). As thunderstorm development is a small-scale phenomenon, its representation in numerical weather prediction (NWP) requires the use of convective-allowing models (Kain et al., 2008; Sobash and Kain, 2017; Loken et al., 2017) with spatial grid resolution typically below 4 km, which requires a large amount of computational resources. Such resolution can be prohibitive for long-term reanalysis or climate projection datasets, especially on the global domain. Weather and climate agencies therefore often use coarser-resolution models where parameterization schemes to approximate are applied to approximate convective convective processes (Hong and Lim, 2006; Lopez, 2016). alternatively, they use convective parameters proxies that are linked to the occurrence of lightning (Westermayer et al., 2017; Battaglioli et al., 2023a; Matczak et al., 2026)

The simulation of lightning in NWP dates back to the 1990s with the development of a parameterization scheme based on cloud top height in Price and Rind (1992). Since then, more complex microphysics parameterization schemes have been developed to represent the processes leading to lightning occurrence in NWP models (Thompson et al., 2004; Hong and Lim, 2006; Morrison et al., 2009; Yair et al., 2010; Kumar et al., 2026), sometimes running a nested higher-resolution convective-allowing model to better capture the convective processes (Lopez, 2016). Some studies have focused on using sounding-derived convective parameters associated with severe convective events. For example, Diffenbaugh et al. (2013) studied the convective available potential energy (CAPE), convective inhibition (CIN) and low-level wind shear in climate projection models (CMIP5). In general, many studies agree that the environments associated with severe convective storms over certain areas are projected to increase in frequency and intensity with climate change (Diffenbaugh et al., 2013; Feng et al., 2016; Prein et al., 2017), although more recent studies (Taszarek et al., 2021b, a) have shown that local trends may not be consistent with the globally projected changes as increasing convective inhibition and reductions in relative humidity can affect the process of storm development and offset increases resulting from warming climate. Moreover, it was also found that observed trends were not always consistent with modeled trends (Taszarek et al., 2021a; Pilguj et al., 2022; Manzato et al., 2025).

With the rise of machine learning, new approaches have been developed to replace physical parameterization schemes and to model the occurrence of lightning and other convective hazards in climate models directly from the convective parameters (McGovern et al., 2023; Wang et al., 2023). Initial approaches relied on single grid cell models that independently predicted the probability of hazard occurrence for each grid cell, with input predictors derived from atmospheric variables in the same grid cell. For example, Rädler et al. (2018) and Battaglioli et al. (2023a) developed generalized additive models that take as input predictors of convection such as measure of atmospheric buoyancy (lifted index) and mid-level relative humidity at a given location. In such models, the output is a probability of lightning occurrence at the same location where environment is evaluated. Mazurek et al. (2025); Czernecki et al. (2019) and Saha et al. (2025) developed random forest models and Cavaiola et al. (2024) a neural network to perform similar tasks. Those models are trained with the convective parameters as input and evaluated against the corresponding lightning observations. The models are then used at inference on reanalysis data to reconstruct the historical climatology of severe convective storms, or on climate projection data to derive future trends (Battaglioli et al., 2026). Those single grid cell models parameterize lightning for each grid cell locally and independently,

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

meaning that they only use the predictors at the pixel of interest to estimate the probability of lightning occurrence at this pixel. This is a common approach in the literature and is justified by the fact that the typical spatial scale of individual convective cells ($\sim$100 km^2) is smaller than the resolution of the reanalysis data usually used for the modeling ($\sim$1000 km^2 in the case of ERA5).

The hypothesis that leads to the present study is that large-scale convective clusters such as mesoscale convective systems are associated with large, mesoscale patterns (over 100 km in our study) in the convective parameters. Those mesoscale patterns may be better captured with a model that has access to the spatial information of the surrounding grid cells than with a single grid cell model. Vision-based deep learning approaches such as U-Net architectures (Ronneberger et al., 2015) provide a framework to capture these spatial patterns. For weather forecasting and nowcasting, the spatial and temporal resolution of the input data is higher than for climatological studies (up to about 5 minutes and 100 meters resolution for convection-permitting simulations, radar or high-resolution satellite data). Weather forecast models can therefore better resolve the spatial and temporal scales of convective processes. Lin et al. (2019) and Geng et al. (2021) were among the first studies to use convolutional recurrent neural networks to nowcast lightning from a fusion of NWP forecasts and recent weather station observation data. Then, Brodehl et al. (2022); Leinonen et al. (2022, 2023); Bosc et al. (2025a) and Bosc et al. (2025b) developed models to nowcast lightning from radar and satellite data in addition to the NWP and station observations. More recently, generative models have also been developed to nowcast an ensemble of possible scenarios of severe weather (Sha et al., 2024). Inspired by natural language processing models, attention-based models have also been introduced (Popick, 2025). We now also see the emergence of similar models for forecasting with longer lead times (McClung et al., 2026).

While vision-based deep learning models have been widely adopted for severe weather forecasting, to our knowledge, only three studies have applied this approach so far for lightning climatology. Cheng et al. (2024) studied the parameterization of lightning over the continental United States of America with a convolutional neural network that they compared to a random forest model and a multi-layer perceptron. Kalashnikov et al. (2024) developed a similar parameterization over the Western United States with a convolutional neural network that they compared to a logistic regression model. Yin et al. (2026) compared physical parameterization schemes, a tree-based model and a U-Net at the global level. We complement their study by looking at lightning parameterization over Europe with a focus on the most extreme events, which are defined in this study as the events with the largest flash count over Europe during one day. We implement and compare several single grid cell models of increasing complexity: a logistic regression, a generalized additive model, an extreme gradient boosting and a multi-layer perceptron. We then compare their performance to a U-Net architecture that incorporates spatial information from the surrounding grid cells.

The main goal of this study is to (1) develop a model that has access to spatial information to better represent lightning activity, (2) compare it to single grid cell models, and (3) evaluate key characteristics of the new model, such as feature importance and spatial sensitivity. We also provide a GitHub repository to reproduce some of the results presented.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

2 Data

We rely on two sources of data: (1) lightning observations from the Met Office's ATDnet lightning detection network (Gaffard et al., 2008); and (2) convective parameters derived from the ERA5 reanalysis dataset on 137 hybrid-sigma levels (Hersbach et al., 2020). Lightning observation data is gridded to a 0.25° horizontal spatial resolution and 1-h temporal resolution to match the ERA5 reanalysis data spatio-temporal resolution.

2.1 Lightning data

Over Europe, lightning observations have been consistently reported since the late 2000s (Taszarek et al., 2020). The Met Office has operated a very-low-frequency (VLF) lightning detection network since 1987 (Lee, 1986). Major upgrades in 2007 and 2023 (Gaffard et al., 2008; Marlton et al., 2022) improved detection efficiency but introduced temporal inhomogeneities in the record. To ensure consistency, this study uses data from 2008 to 2023, providing 16 years of homogeneous, good quality-lightning observations (Poelman et al., 2013b). The system primarily detects cloud-to-ground flashes, with a detection efficiency of about 89% over Europe, and to a lesser extent intra-cloud flashes which are detected with an efficiency of ∼24% (Enno et al., 2016, 2020). This dataset has been used in several previous studies on lightning modeling (Taszarek et al., 2020; Battaglioli et al., 2023a; Matczak et al., 2026). These studies provide further details on lightning climatology over Europe derived from these observations. The location uncertainty of detected flashes is about 1–2 km across Europe (Poelman et al., 2013a), which is sufficient for this study since the data are subsequently gridded onto the ERA5 reanalysis dataset grid with an hourly temporal resolution and a 0.25° horizontal resolution (∼30 km). Following Rädler et al. (2018); Taszarek et al. (2020); DiGangi et al. (2022) and Battaglioli et al. (2023a), a "lightning hour" or "lightning occurrence" at a given grid cell is defined as an hour during which at least two flashes are detected within the grid cell. Such grid cells are designated as positive samples. The threshold of 2 flashes helps filter out potential false alarms while preserving most of the signal. This processing results in a data imbalance, with only about 0.4% of the data being positive.

2.2 ERA5 reanalysis data

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. These variables were selected as the best combined predictors of lightning within their generalized additive model AR-CHaMo. Convective precipitation is a direct indicator of convective activity in ERA5, though it is a result of convective parameterization schemes as ERA5 does not explicitly simulate convection due to its coarse resolution. The lifted index (Galway, 1956) is a measure of atmospheric buoyancy (thermodynamic instability), which is a key factor for convection and lightning. The mixing ratio and relative humidity provide information on the moisture content of the atmosphere, which also influences convection and lightning. Too dry atmosphere and entrainment can often contribute to failure modes in convective development (Mulholland et al., 2021). Finally, the land-sea mask is included because lightning occurrence can differ significantly between land and sea due to differences in surface heating

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

and moisture availability (Romps et al., 2018; Schaeffler, 2025; Abbott et al., 2025). Similarly to previous literature focusing on lightning environments across Europe, the parameters used in this work were computed using the thundeR rawinsonde processing package (Czernecki et al., 2025).

The region of interest for this study is Europe, which experiences a significant number of lightning events and has been the focus of previous studies on lightning climatology (Piper and Kunz, 2017; Taszarek et al., 2020; Enno et al., 2020; Battaglioli et al., 2023a). The region is defined by the following latitude and longitude bounds: [35°N, 60°N] and [12°W, 25°E]. This region includes a variety of climate zones and topographical features, which can influence lightning occurrence.

3 Method

3.1 From single grid cell models to vision-based deep learning approaches

Thunderstorm climatologies often utilize local, single grid cell approaches (Battaglioli et al., 2023a, b; Saha et al., 2025; Battaglioli et al., 2026; Ehrensperger et al., 2025; Campos et al., 2024). By modeling each grid cell independently, these frameworks estimate marginal probabilities of severe weather occurrence based on the assumption that individual convective cells are smaller than the data resolution. These models summarize the atmospheric column through key ingredients: moisture, instability, lift and vertical wind shear (McNulty, 1985; Doswell et al., 1996; NOAA, 2023; Kaltenböck et al., 2009). We schematically represent this column-based approach in Figure 1.

However, the independent pixel assumption fails to capture large-scale organized events like mesoscale convective systems, where neighboring conditions drive storm evolution (NOAA). To address this, we move beyond single grid cell modeling to incorporate surrounding spatial information. We evaluate several single grid cell models — logistic regressions (LR), generalized additive models (GAM) (Hastie and Tibshirani, 1986), extreme gradient boosting (XGB) (Chen and Guestrin, 2016), and multi-layer perceptrons (MLP) (Rumelhart et al., 1986) — and compare them against a U-Net architecture (Ronneberger et al., 2015). This vision-based deep learning approach is designed to capture the complex spatial patterns inherent in organized convective systems. A brief overview of the hyperparameters of the models, optimization methods, and loss functions is provided in Table A1 and the selected hyperparameters for the final optimized model are provided in Table A2.

3.2 Optimized U-Net Architecture

In our case, the input image of our U-Net is a 5-channel image of size 101×149 pixels, which is subsequently padded to 112×160 pixels, where each channel corresponds to one of the 5 environmental variables at a given hour. The output image is a single-channel image of size 101×149 pixels, where the values correspond to the probabilities of lightning occurrence at each grid cell for the same timestep as the input. We use a standard U-Net architecture, inspired by Holderrieth and Erives (2025). We empirically selected the loss function and optimized the other hyperparameters (see Table A2) using a Tree-structured Parzen Estimator search strategy during the inner loop of the Leave-One-Year-Out (LOYO) procedure (see Section 3.3) with

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

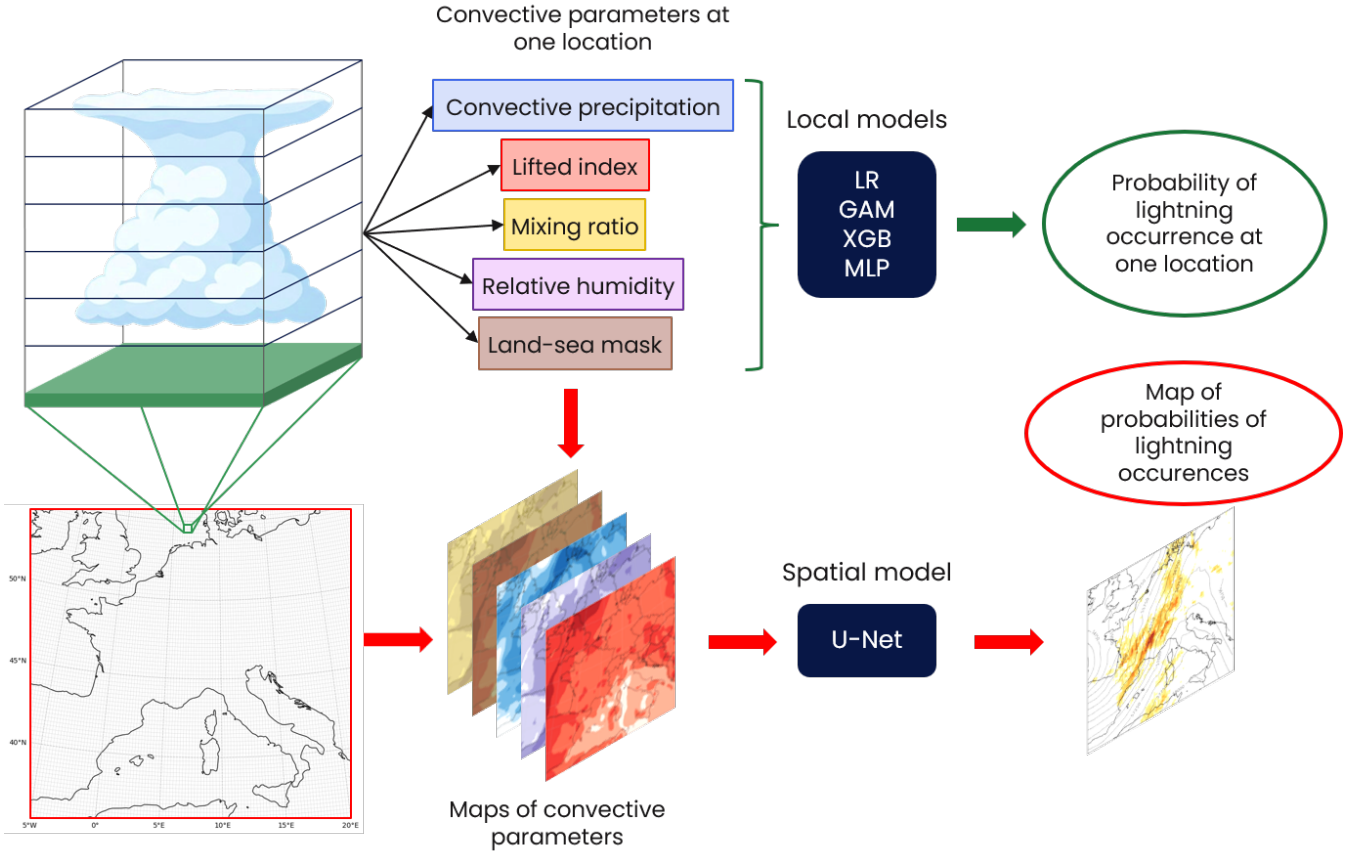

Figure 1. Lightning prediction pipeline for both local and spatial models. The atmospheric column at a given location is summarized in a few variables. Those variables are fed as input of Machine Learning models. The single grid cell models take the local convective parameters and output one probability of lightning occurrence at one location. The spatial models take the entire maps of convective parameters as input and output the corresponding map of probabilities of lightning occurrence. The models are trained with ERA5 reanalysis data for the inputs and ATDnet observations for the ground truth.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

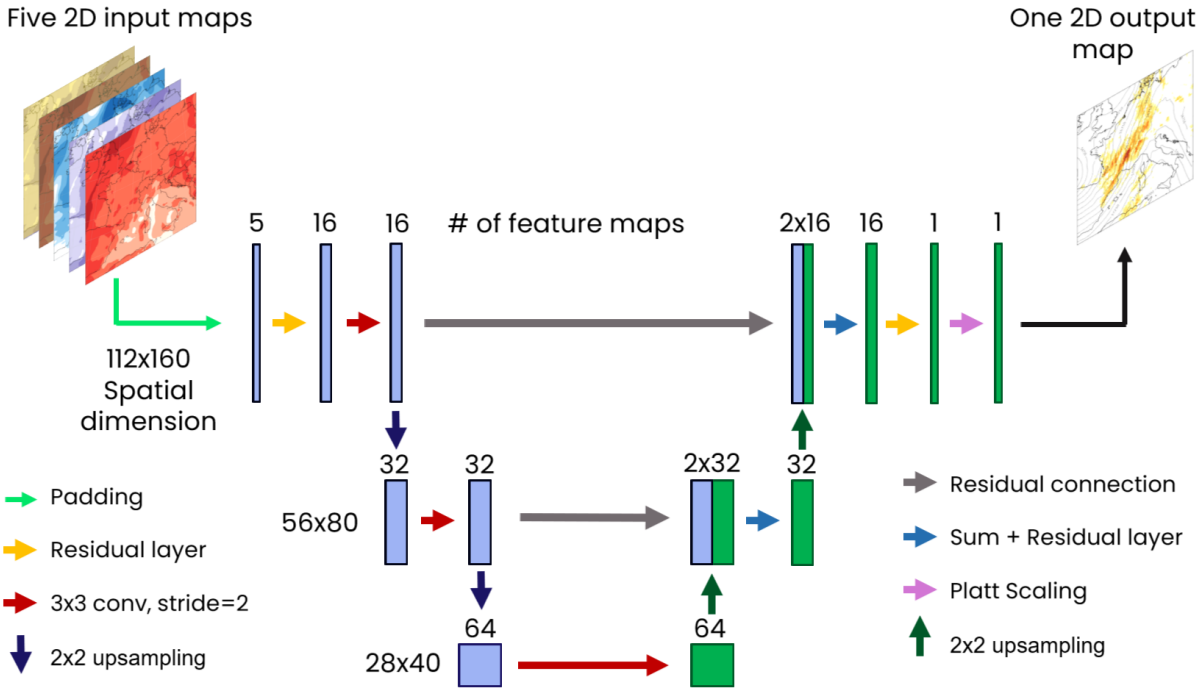

Figure 2. Schematic diagram of the architecture of the U-Net model. The model is trained with five environmental variables as input and outputs the corresponding map of lightning probabilities. To maintain well-calibrated probabilities, we add a Platt scaling layer to the U-Net architecture.

the Optuna python package (Akiba et al., 2019) by optimizing a score made of the sum of the five metrics described in Section 3.4. Our final architecture is shown in Figure 2.

To address class imbalance, the U-Net and MLP are trained with weighted loss functions: the error made at positive pixels is multiplied by a weight $w \geq 1$. Because mispredicting positives is penalized more heavily than missing negatives, the model learns to over-predict probabilities. Therefore, the models are not guaranteed to be calibrated, meaning that the predicted probabilities may not correspond to the true probabilities of lightning occurrence. To address this issue, we apply a post-hoc calibration method called Platt scaling (Platt et al., 1999; Guo et al., 2017), a post-hoc calibration method which fits a logistic regression where the input is the predicted probability from the U-Net (or MLP) and the output is the true binary label of lightning occurrence.

3.3 Training pipeline

This section presents a brief overview of the training pipeline and modeling choices we implement. More details are provided in Appendix B.

We need our training, validation and test datasets to be independent and identically distributed (iid). To this end, we decide not to break down temporality into smaller chunks than years. Indeed, thunderstorms feature a strong seasonal cycle (see Figure

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

5(b)), meaning that each dataset must contain a balanced representation of all four seasons (i.e. equal numbers of winters, springs, summers and autumns within each set). To mitigate data leakage, we split the data into as few temporal chunks as possible. Thunderstorm environments exhibit autocorrelation on the order of 24 hours, and intra-annual climate variability — such as teleconnections and slow surface variable dynamics — can have long-lasting effects that propagate leakage across seasons. Splitting at a finer scale than years would therefore multiply the number of temporal boundaries and the associated leakage risk. We place each boundary during the period of lowest thunderstorm activity (i.e. during winter) so that the 1st of January cut falls near the trough of the seasonal cycle and minimizes leakage between adjacent years (see Figure 5(b)). This breakdown is a deliberate trade-off to obtain data as close to iid as possible. In the rest of the paper, we treat these 16 yearly blocks as iid samples.

In order to evaluate the robustness of the different algorithms to the choice of training and test data, we choose to use a cross-validation strategy to quantify the uncertainty of the performance linked to the choice of model and data split. We keep 3 years aside for final evaluation and we apply leave-one-year-out (LOYO) cross-validation on the remaining 13 years. The 3 held-out years are 2008, 2015 and 2023 such that we test our final model on the beginning, middle and end of the record. We add a comparison of uncertainty quantification methods in Appendix C.

The LOYO pipeline yields 13 models and we report the corresponding results in Section 4.1. More details can be found in Appendix B on how the models are trained, how the U-Net's hyperparameters are optimized and how the uncertainties are computed.

Once the LOYO procedure is done, we perform a final optimization round with 3-fold cross-validation on the 13 years used for the LOYO, and once the best hyperparameters are found we train the final optimal model on the 13 years. We then test the final performance of the model on the 3 held-out years. These results are reported in Table 2.

3.4 Metrics

We use 5 different metrics to evaluate the performance of our models: the area under the precision-recall curve (PR-AUC) (Davis and Goadrich, 2006), the area under the receiver operating characteristic curve (ROC-AUC) (Hanley and McNeil, 1982), a probabilistic version of the intersection over union metric: the continuous Dice Coefficient (Dice) (Shamir et al., 2018), the Fractions Brier Skill Score (FBSS) (Roberts and Lean, 2008; Brier, 1950), and the Deviance explained (Deviance) (McFadden, 1974). We propose a summary of the metrics used in Table 1 and provide additional details on the definition and interpretation of each metric in the Appendix D. These metrics are complementary, capture different aspects of model performance and are all threshold-independent. We choose to use threshold-independent metrics to avoid the issue of selecting a threshold to convert predicted probabilities into binary predictions, which can be arbitrary and may not reflect the true performance of the model across different applications. Threshold-dependent metrics such as accuracy - Accuracy $= \frac{\text{True Positive+True Negatives}}{Positives+Negatives}$ require selecting a threshold, which can lead to misleading conclusions if not done carefully. For instance, in the case of imbalanced data, a high accuracy can be achieved by simply predicting the majority class, while the model may perform poorly on the minority class.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Metric	What it measures	Why we use it
PR-AUC	Quality of positive predictions.	Focuses on the positive class. Good for imbalanced data.
ROC-AUC	Discrimination and sample ranking ability.	Threshold-independent. Widely used but is driven by true negatives for imbalanced datasets.
Dice	Spatial overlap between predictions and observations.	Evaluates the spatial accuracy of predictions. Used in image-to-image tasks.
FBSS	Improvement of the Brier score over a baseline model.	Adapts the RMSE score for classification tasks. Evaluates the calibration and the sharpness of the predicted probabilities.
Deviance	Improvement in log-likelihood over a baseline model. Generalizes the R^2 metric for classification tasks.	Evaluates calibration and sharpness with a focus on overconfident predictions. Used for model selection in previous studies.

Table 1. Summary of the metrics used for model evaluation, what they measure and why they are used.

In our case of lightning prediction, the model should be particularly good at predicting the positive class (lightning occurrence), which is the minority class. As a result, we promote models that produce true positives, while not predicting false positives and false negatives. In other words, we aim at maximizing simultaneously Precision and Recall. The PR-AUC is a metric that captures this trade-off and measures the quality of the positive predictions. This metric is therefore particularly relevant for our case, as it has been argued in Sobash and Ahijevych (2024).

3.5 Interpretability and explainability

We 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.

4 Results

In this section, we compare the performance across the different models that we implemented in this study. We first present the quantitative results for all the metrics, then the physical aspects with the seasonal and diurnal cycles of the predictions, as well as the yearly and seasonal spatial distributions of the predicted probabilities. We compare the feature contribution across the different models and finally we present case studies of particularly intense thunderstorm events. We compare the lightning occurrence predicted by the models for these events.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

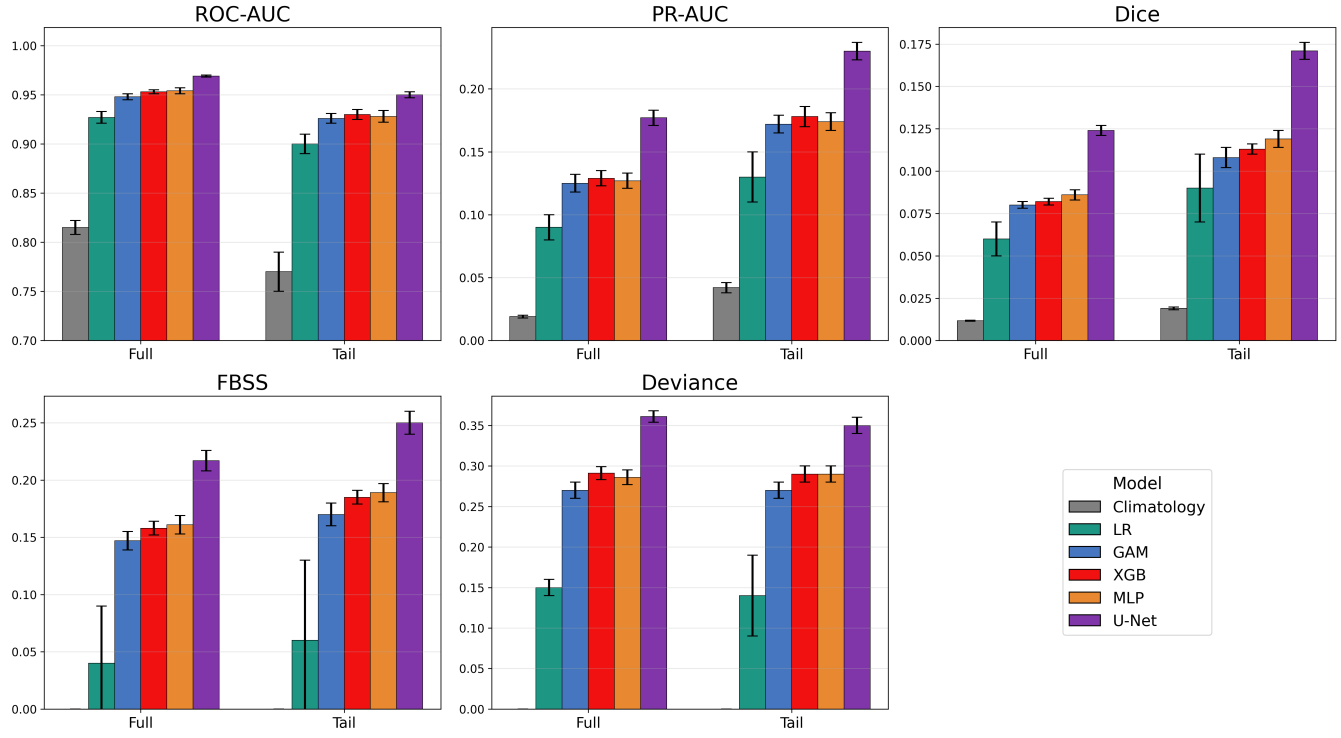

Figure 3. Scores for all models. The "Full" column corresponds to the performance on the whole test set (the left out year for each LOYO split). The "Tail" column corresponds to the performance on the subset of days in the test set where the total number of lightning flashes summed over the domain exceeds the 95th percentile of the daily flash count distribution computed over the full dataset. The values are reported as (mean ± margin) where the whole interval is the 95% confidence interval. All the U-Net scores are significantly better at the 99% confidence level (as evaluated with a Student's t-test). Higher is better for all metrics.

4.1 Quantitative results

We report the results of all models for all the metrics in Figure 3. For each metric, we report the score on the whole test set and the score on the most extreme days (49 days). The most extreme days are defined as days for which the total number of lightning flashes summed over the domain exceeds the 95th percentile of the daily flash count distribution computed over the full dataset. Those extreme days are often associated with large mesoscale convective systems for which we expect the U-Net to perform particularly well compared to the single-column models. We also report the 95% confidence interval for each metric, which is computed across the 13 splits of the LOYO procedure (see Section 3.3).

According to Figure 3, the U-Net outperforms all the other models for all the metrics by a significant margin, both on the "full" test set and on the "tail" of the distribution. 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 so the

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

difference in performance between the U-Net and the other models is statistically significant at the 99% confidence level for all metrics, both on the "full" test set and on the "tail" of the distribution. This means that we can reject the null hypothesis that there is no difference in performance between the U-Net and the other models, and conclude that the U-Net performs significantly better than the other models for all metrics.

Metric	**ROC-AUC**		**PR-AUC**		**Dice**		**FBSS**		**Deviance**	
Model	Full	Tail	Full	Tail	Full	Tail	Full	Tail	Full	Tail
Final U-Net	0.970	0.953	0.179	0.241	0.125	0.177	0.222	0.26	0.366	0.36

Table 2. Scores for the final U-Net model, computed over the 3 held out years: 2008, 2015 and 2023. Tail and Full columns are the same as described in Figure 3.

For the final model, optimized over the 13 years of the LOYO procedure, all metrics (except the explained deviance for extreme days) are slightly better than the average scores of the U-Net across the LOYO splits (see Table 2). They are either in the top of the confidence intervals or slightly above it, which suggests that the final model performs better than most of the U-Net models trained during the LOYO procedure, but not significantly better than the best ones. This behavior was expected since we have optimized the hyperparameters with one additional year of data.

4.2 Calibration

In Figure 4(a), we compare the calibration curve of the U-Net before and after applying Platt scaling. Without recalibration, the model overestimates the probability of lightning occurrence at all levels, which is expected since our loss function is a weighted binary cross-entropy that puts more weight on the positive class and thus encourages the model to predict higher probabilities. After recalibration, the model is much better calibrated overall but still overestimates probabilities above 30%. Note the logarithmic scale for the proportion of total samples per bin which shows the heavy data imbalance: 80% of all samples are in the first bin.

In Figure 4(b), we compare the calibration curves of all the models. We see that all models tend to underestimate the small probabilities (below 20% chance of occurrence) and overestimate the higher probabilities (above 30% chance of occurrence). The U-Net is the best calibrated model overall, with a calibration curve closer to the diagonal at every level, and especially above 50% chance of occurrence where the other models tend to overestimate the probabilities more. The models tend to overestimate more the very high probabilities (above 80% chance of occurrence) although the reliability of the curve at these points suffers from a small sample size.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

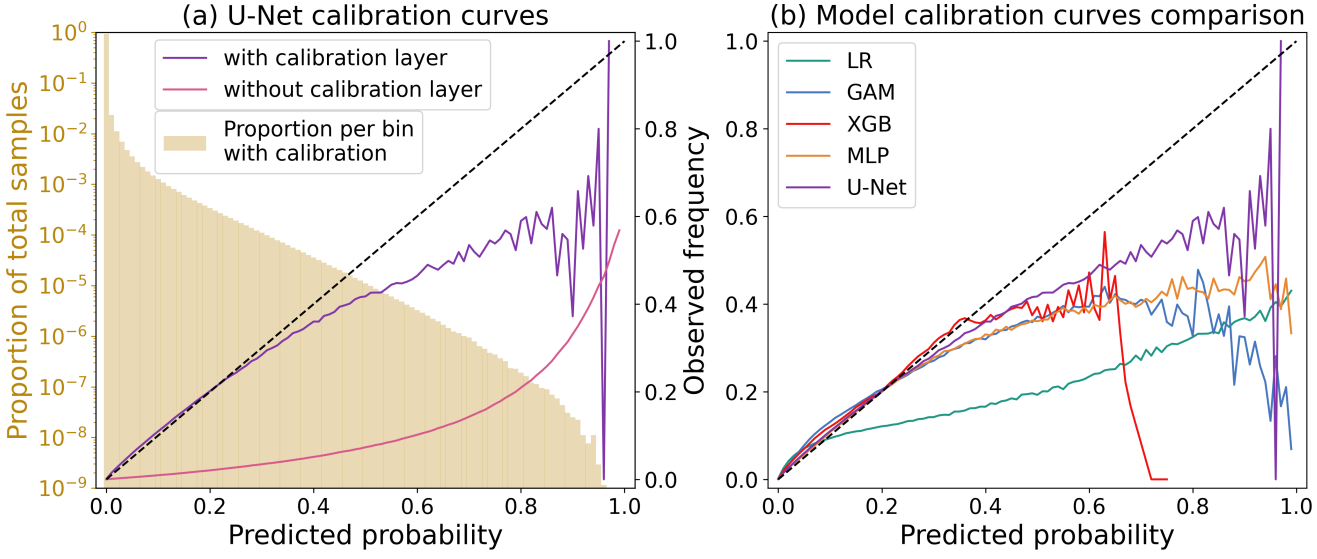

Figure 4. Calibration curves. Panel (a) shows the calibration curve of the U-Net before and after applying Platt scaling. The gold shaded bar plot shows the number of samples in the test set for each predicted probability bin (for the re-calibrated model). Panel (b) compares the calibration curves of all the models.

4.3 Climatology modeling

We present here the climatology of lightning occurrence over the 2008-2023 period modeled by the U-Net model, compared to the other models and to the observations. We look at the diurnal and seasonal cycles of the predicted lightning occurrence, as well as the spatial distribution of the predicted probabilities of lightning occurrence across seasons (see Figures 5 and 6).

In Figure 5(a) we evaluate the ability of the models to capture the diurnal cycle of thunderstorm occurrence, and in Figure 5(b) their ability to capture the yearly cycle. To do this, we build a probability density function of lightning occurrence over the day: for each hour, we count the total number of pixels with lightning occurrence across the 13 years of the cross-validation pipeline, then divide by the total number of pixels with occurrence over the period. We build the yearly cycle in the same way, counting pixels with lightning occurrence for each calendar day and dividing by the total across the 13 years. This analysis aims to determine if (1) the model predicts the right proportion of lightning for a given hour compared to the other hours of the day; and if (2) the model predicts the right proportion of lightning for a given day compared to the other days of the year.

We quantify the difference between the observed and modeled cycles with the Jensen-Shannon divergence (JSD) (Lin, 1991), a measure of the distance between two probability density functions. We use a base 2 logarithm such that a divergence of 0 means that the two distributions are identical and a divergence of 1 means that the two distributions do not overlap at all. We find that the U-Net captures the diurnal cycle much better than the other models (+62% improvement compared to the logistic regression). For the yearly cycle, all models feature comparable performance, with a good overall fit (the logistic regression being the worst option). We note a slight overestimation of all models in October and November. Since all models feature

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

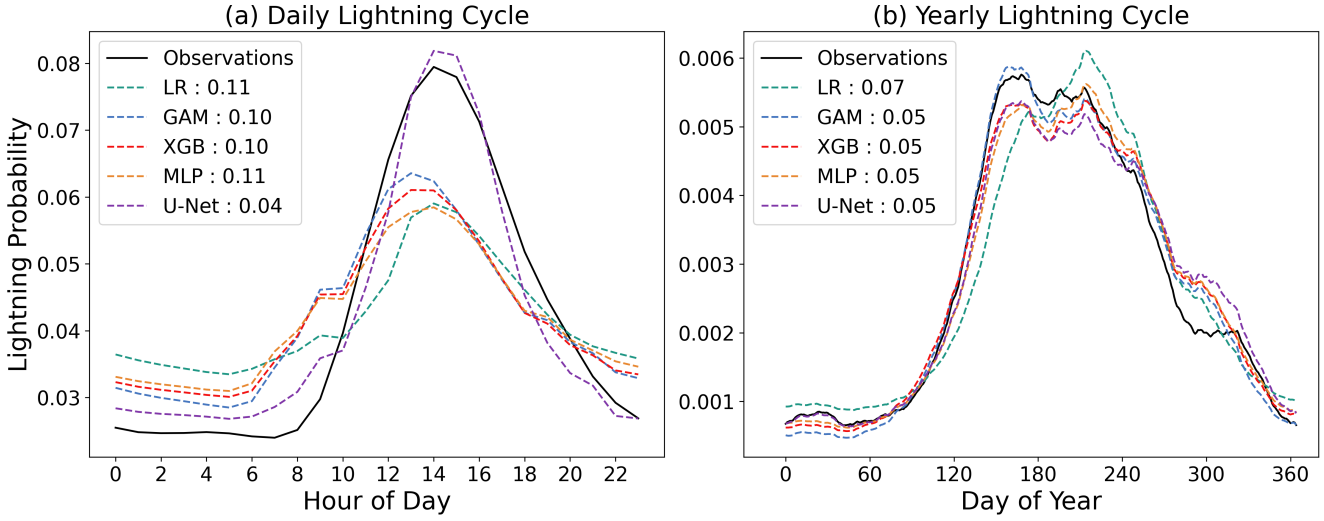

Figure 5. Daily (a) and yearly (b) cycles as observed and modeled presented as probability density functions of the number of hours of lightning. We compute those cycles with the LOYO procedure on all years except 2008, 2015 and 2023. In the legend we report the Jensen-Shannon divergence between the modeled and observed distributions. Lower is better.

this overestimation, the reasons probably lie in the predictor data. It may be related to lightning activity shifting across the Mediterranean sea in October and November while activity over land significantly drops, as convective parameters are not handled as well over the oceans and lightning parameterizations tend to overestimate lightning activity over oceans (Matsui et al., 2016; Pérez-Invernón et al., 2022).

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). We compare the observed climatology with the predicted climatology for each model by looking at the RMSE and the bias for each model. Overall, all models feature biases, although the U-Net features the lowest RMSE. The U-Net also alleviates some of the biases that the other models have, such as the underestimation of lightning in mountainous regions (the Alps, the Pyrenees, the Apennines, the Serra Calderona and the Dinaric Alps) and the overestimation of lightning over the Mediterranean Sea (in the case of the logistic regression). The error made by the U-Net is overall smaller but it still over-predicts over the Alps and the Pyrenees. This result may be related to resolution of ERA5 that may not be capable of properly reflecting complex orography gradients that contribute to lightning development. We also note that the U-Net tends to make smoother predictions which is visible if we look at the raw predicted climatology (i.e. before computing the bias). We present this in Figure S1 in the supplementary material.

4.4 Feature importance

We assess here to what extent having each feature as input to the model changes the performance of the model for each evaluation metric. To do this, we compute Shapley Additive Explanations (SHAP) values (Lundberg and Lee, 2017; Molnar,

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

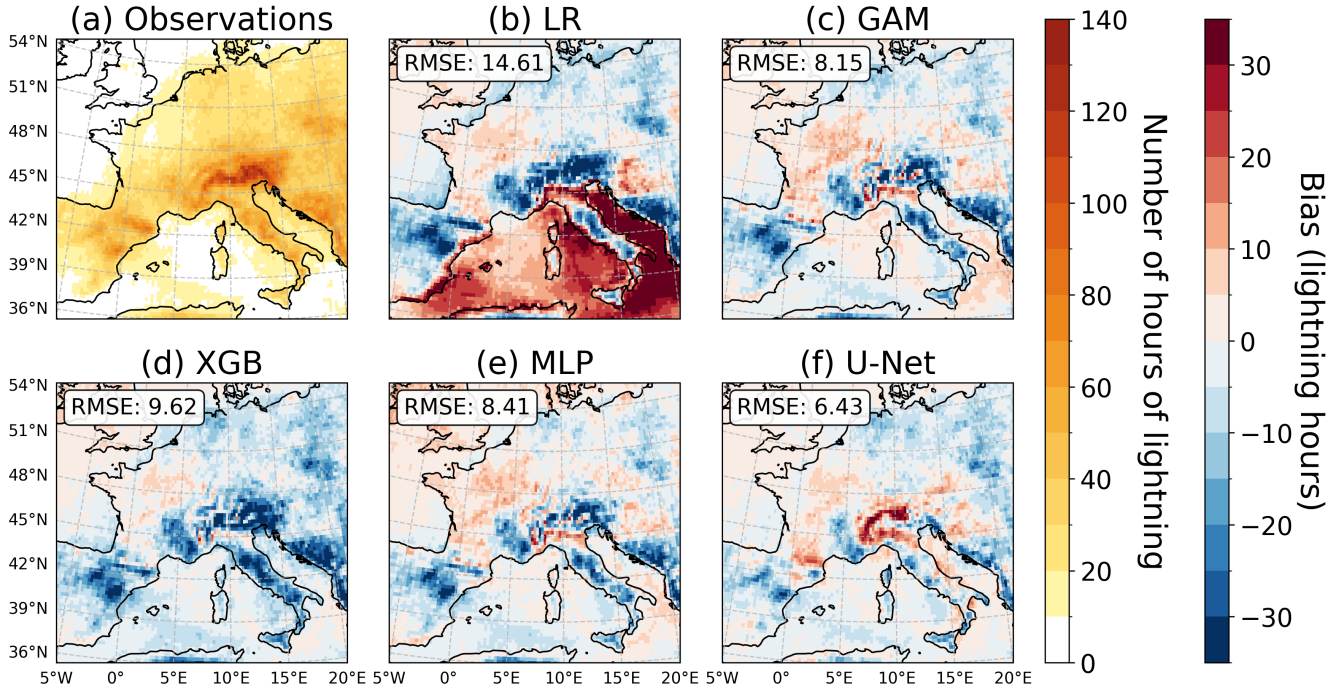

Figure 6. Yearly average number of hours of lightning in Summer on the test set (2008, 2015 and 2023). We assess the error of the models (b) to (f) with respect to the observations (a). Modeled climatology and observations are compared with the RMSE between the two images (lower is better).

2025) and report them in Figure 7. We compute the SHAP values at the global level — not over individual predictions, but over aggregate evaluation metrics (PR-AUC, ROC-AUC, Dice, FBSS and Deviance). We compute SHAP values by retraining each model on every possible subset of the 5 features and evaluating its performance. Since there are $2^5 = 32$ subsets, we train 32 instances of each model (LR, GAM, XGB, MLP and U-Net), yielding one SHAP value per feature and per model. To compute uncertainties, similarly as in Section 3.3, we use a LOYO pipeline. For each model, we repeat the experiment 13 times, each time changing the training and test datasets. For each model, we get 13 sets of SHAP values, based on which we derive confidence intervals, similarly as in Appendix B. Assuming that the SHAP values are normally distributed across models, we compute the 95% confidence interval as $(x \pm t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}})$, where x is the mean SHAP value, $t_{\alpha/2,n-1}$ is the critical value of the t-distribution with $n-1$ degrees of freedom, σ is the standard deviation across models, and $n = 13$ is the number of models.

Figure 7 shows the SHAP values for the ROC-AUC and PR-AUC metrics. Results for the remaining metrics are provided in the supplementary material (Figure S9). Each model is assessed individually. For the single grid cell models, the lifted index is the dominant predictor under both metrics, in line with the existing literature Battaglioli et al. (2023a); Matczak et al. (2026). Under the PR-AUC, however, convective precipitation rises to comparable importance alongside the lifted index. This

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

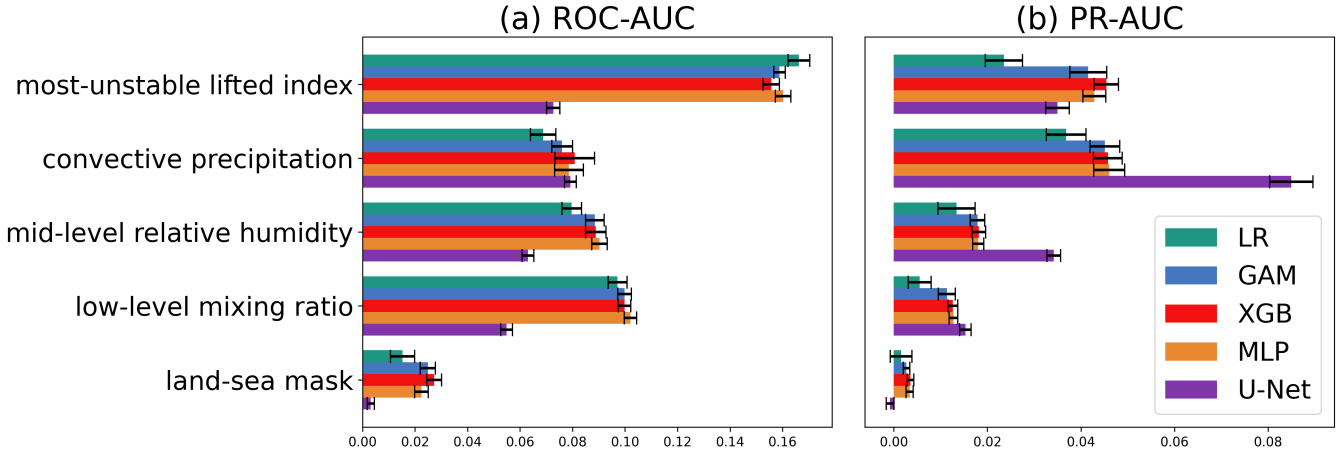

Figure 7. SHAP values evaluated for two metrics: (a) ROC-AUC and (b) PR-AUC. Confidence intervals are computed with the LOYO procedure by computing the SHAP values for each of the 13 splits. Bars indicate the average improvement induced by including each feature over every possible combination of the other features.

distinction reflects the different sensitivities of the two metrics: ROC-AUC captures overall discriminative skill across both classes, so it rewards predictors that are informative about the dominant negative class (non-lightning), while PR-AUC focuses exclusively on the positive class (lightning occurrence) and therefore amplifies the contribution of predictors directly tied to convective activity. This has a practical implication: studies targeting the occurrence of thunderstorms should prioritize PR-AUC (or equivalent positive-class-focused metrics) when selecting predictors, since optimizing for ROC-AUC or deviance can favor predictors that are good at identifying non-events rather than events.

For the U-Net, the relative importance of convective precipitation is substantially larger than for the single grid cell models, while the lifted index, though still relevant, is comparatively less dominant. This likely reflects the U-Net's capacity to exploit local spatial structure: convective precipitation has pronounced spatial clustering that a convolutional architecture can leverage, whereas the lifted index varies more smoothly. Under the PR-AUC, there is also a substantial increase in the relative importance of mid-level relative humidity, compared to its importance for single grid cell models. An example of the input fields is presented in the supplementary material on Figure S10. Taken together, these results suggest that predictor importance is not an intrinsic property of the variable itself, but depends jointly on the model architecture and the evaluation metric. Both architecture and metric should therefore be chosen in accordance with the scientific goal of the study.

4.5 Extremes representation: a case study

We show in this section that the U-Net improves the representation of extreme events compared to the other models through case studies. The goal is to compare how well the spatial extent and details of the events are captured by the different models. We use the final models, trained all years except 2008, 2015 and 2023. We choose to focus on the most extreme event of each year in the test dataset (2008, 2015 and 2023) in terms of daily total lightning count over the domain. These events

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

occurred on the 15th of August 2008, 11th of August 2015 and 5th of August 2023 respectively. They were associated with large-scale convective systems that affected different parts of the domain. 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. We also provide the code to reproduce these results on all the extreme events of the test dataset in the git repository https://github.com/AdrienBq/lightning_modelling.

On the 15th of August 2008, a severe convective event unfolded across central and eastern Europe. The synoptic situation was characterized by an upper-level trough with a strong mid-tropospheric temperature gradient and southwesterly winds. These conditions favored the uptake of moisture and temperature from the Mediterranean basin. This air, mixed with cold air masses from northern and central Europe, was at the origin of a squall line extending from the Po Valley to the Baltic Sea, resulting in widespread lightning activity and a tornado outbreak in Poland (Chmielewski et al., 2013).

Figure 8 presents the prediction aggregated over the entire day during which the event occurred ((c) to (g)). This gives us the expected number of hours of lightning occurrence during the day at each pixel. We do the same for the observations in Figure 8(b). We show the 1-D power spectrum (van der Schaaf and van Hateren, 1996) of the spatial distributions of lightning occurrence for observations and models in Figure 8(a). This power spectrum is computed by taking the Fourier transform of the images and by averaging the power of pixels with the same wavenumber in the Fourier domain. This power spectrum gives us a measure of how well spatial patterns at different scales are captured by the models: low wavenumbers correspond to large-scale patterns, while high wavenumbers correspond to small-scale patterns.

We compare the predicted expected number of hours with lightning occurrence between each model and the observations by computing the evaluation metrics for this particular case study, and we report them in Table 9. We also compute the RMSE between the images. Since our images are very sparse, the RMSE is computed both on the full images (All) and on the subset of pixels where lightning was observed (Obs>0).

We first note from Table 9 that the U-Net model outperforms all the other models on the evaluation metrics for this particular case study. We also see from Figure 8 that the U-Net better captures the large-scale spatial patterns of the event than the other models. The other models manage to capture regions with the most lightning activity (over the Alps and across central Europe), but they do not capture as accurately the maximum intensity and spatial extent of the event, underestimating the total area of the event. The power spectrum of the U-Net's prediction also shows that the U-Net captures the large-scale features of the event better than the other models as the power spectrum is closer to that of the observations for small wavenumbers. 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. The failure of the U-Net to capture these small-scale patterns is expected: its convolutional architecture favors a smooth, large-scale structure, and the daily aggregation of the hourly probability of lightning yields an expected number of hours of lightning. It shows, however, that to be able to simulate realistic events with small-scale patterns, we need to have a sampling step to sample realistic binary predictions from the predicted probabilities.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

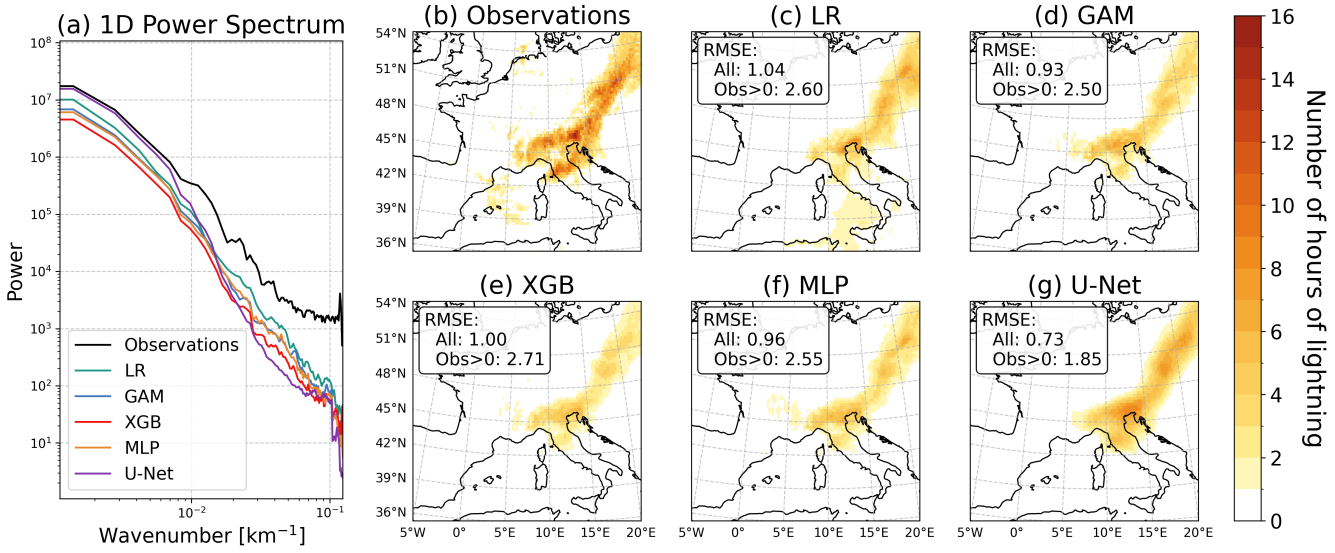

Figure 8. 15th of August 2008 event modeling across different models ((c) to (g)). We compute the RMSE between the images of the observations (b) and the predictions aggregated over the day. We do this both for the full image (All) and for the subset of pixels where lightning was observed (Obs>0). On (a) we present the 1-D power spectrum of the spatial distributions of lightning occurrence for observations and models. 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 Table 9).

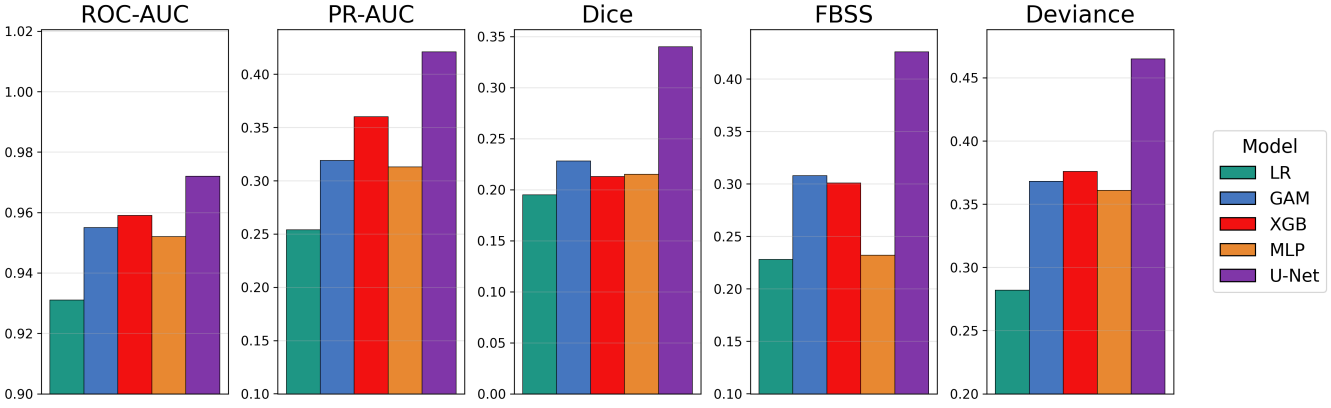

Figure 9. Scores of each model for the 15th of August 2008 event.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

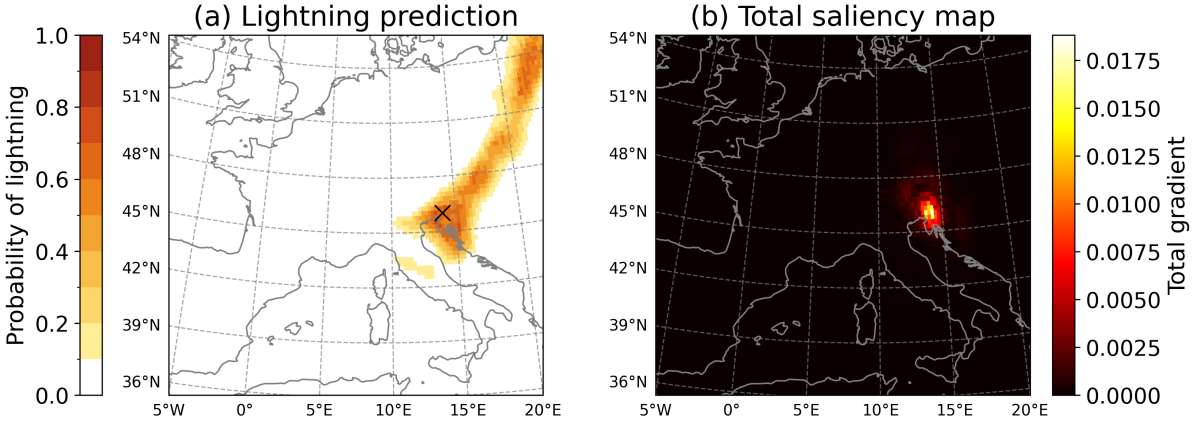

Figure 10. Saliency map on the 15th of August 2008 at 21:00 UTC for the prediction at (14°E, 46°N). (a) shows the predicted probability of lightning occurrence and (b) the norm of the backpropagated gradient of the output at (14°E, 46°N) with respect to the input image. The brighter the pixel, the more sensitive the model is to changes in the predictors at this pixel.

4.6 U-Net interpretability

In this section, we use gradient-based and perturbation-based methods to investigate the U-Net's receptive field. Our gradient-based method is a pixel-wise attribution method that highlights the input pixels that were relevant for a neural network's prediction (Molnar, 2025). It outputs maps called saliency maps which are computed by taking the gradient of the output of the model at a given pixel with respect to the input image. The method is similar to the backpropagation technique that is used during training. The difference is that we backpropagate the value of the output at one given pixel instead of the error between the predicted image and the true image. The pixels with the highest absolute gradients are the ones that the model is the most sensitive to. This is a proxy to compute the most relevant pixels for the prediction of the model. This method is simple but has some limitations. For example, vanishing gradients and activation function saturation lead to an underestimation of the receptive field of the model (Sundararajan et al., 2017). There are more sophisticated gradient-based methods to compute similar attribution maps (Alber et al., 2019), but the simplicity of the one we use makes it a suitable first approach to investigate the spatial information the U-Net uses for its predictions.

An example of saliency maps for the U-Net is shown in Figure 10. We take the output of the model on the 15th of August 2008 at 21:00 UTC for the prediction at (14°E, 46°N) (see plot (a)) and take the gradient of the output at this pixel with respect to the input image. We get an image with 5 channels (one per input) and we show the image of the norm of the 5-D vector at each pixel in Figure 10(b). The feature-wise gradient is presented in the supplementary material (see Figure S10). We can see that the U-Net is using information in an area that is roughly 250 km wide, which is larger than the pixel itself but much smaller than the full domain.

We can extract a mask centered on the pixel of interest to then compare multiple pixel-attributions. We repeat the process for all pixels, all hours and all days of our dataset of extremes (the top 5% days with the most lightning strikes). We only

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

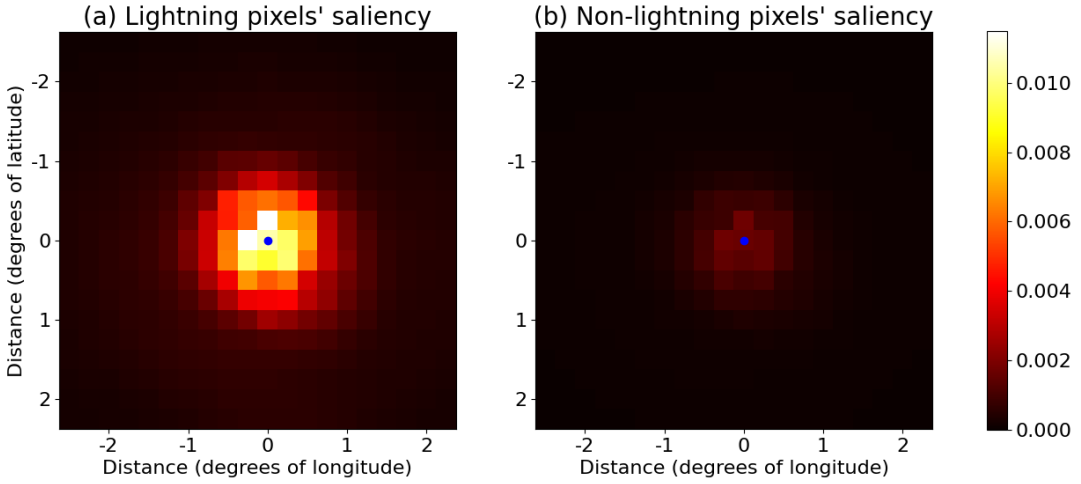

Figure 11. Average saliency mask across the 2023 most extreme events, for (a) pixel where lightning was observed and (b) pixel where it was not.

look at the extremes for computational reasons. We report in Figure 11 the average mask for pixels with observations of lightning and pixels without. We observe that the two resulting masks are spatially very similar, with the difference that the lightning-occurrence mask is 10 times more intense than the non-lightning-occurrence mask. This suggests that the U-Net is more sensitive to the predictors when lightning is observed than when it is not: in lightning-prone environments, a small change in the predictors can lead to a larger change in the predicted probabilities of lightning occurrence than in non-lightning-prone environments.

From the saliency mask of each pixel we then derive its radial profile by averaging the saliency values across pixels with the same distance to the pixel of interest. We report the median, 10th and 90th quantile values of the radial profiles across every distance to the pixel of interest in Figure 12(a). We separate positive and negative pixels. We can see that the saliency values are the highest around the pixel of interest, and then decrease as we move away from it. We defined a radius of useful information R as the radius at which the median of the saliency values is $\frac{1}{10}$ of the maximum of the median value across all distances. In Figure 12(a), for pixels where lightning was observed, we obtain a max median value of 0.01 so our sensitivity threshold is 0.001 and the radius for which the median sensitivity values falls below this threshold is $R \sim 135$ km. Therefore, the region the U-Net uses for its predictions is approximately a (9×9) kernel centered on the pixel of interest. We also find the same discrepancy between lightning and non-lightning pixels as in Figure 11, the model being much more sensitive to surrounding pixels when lightning is observed.

Our second method is a radial perturbation-based method. For a prediction at a given time and location, we perturb the input image by replacing the values of the predictors outside a given radius R from the pixel of interest with random values. We then evaluate the difference between the prediction with the perturbed input and the prediction with the original input. By doing this for different values of R, we can evaluate how sensitive the predictions of the model are to perturbations at increasing distance.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

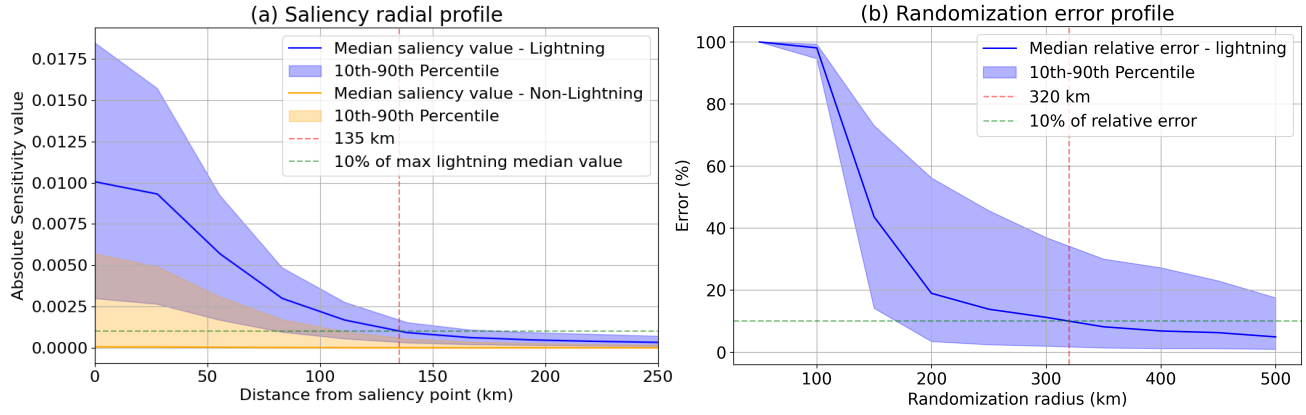

Figure 12. U-Net's receptive field size. (a) features the saliency radial profiles for pixels with lightning occurrence and pixels without. The radius of useful information is around 135 km. (b) features the relative error profile of the randomization-based method, which suggests a radius of useful information of around 320 km.

This gives us an estimate of the radius of useful information for the predictions of the model. This method is complementary to the saliency maps because it does not suffer from vanishing gradients and activation function saturation, but it is more computationally expensive since it requires re-evaluating the model on multiple perturbed images for each prediction (i.e. for each pixel, each hour and each day of the test set). This method can be seen as a radial variant of the occlusion method presented in (Zeiler and Fergus, 2013), where instead of occluding a square patch anywhere in the input image, we occlude a circular patch centered on the pixel of interest.

Our main interest here is understanding the receptive field of the U-Net to predict lightning during the most extreme events. For computational reasons, we perform our radial perturbation analysis on the same set of extreme events as the saliency maps, but only for pixels where lightning was observed. We report the relative error of the performance of the model on the perturbed images compared to the original images as a function of the radius of the perturbation in Figure 12(b). Similarly to the saliency maps, we define a radius of useful information as $R = argmin_r\{\forall d > r, \text{median}(error(d)) < 10\%\}$. We find that $R \sim 320$ km, which corresponds to a (23×23) kernel centered on the pixel of interest. This is larger than the radius derived from the saliency maps, but it is consistent with the fact that saliency maps are known to underestimate the attention radius of the model.

Our interpretation is that pixels between 135 and 300 km from the pixel of interest give the mesoscale context of the event, which is useful for the predictions of the model. Small perturbations there do not greatly affect the predictions of the model, which may explain why the saliency is small. However, large perturbations in this large-scale region will completely break the context and therefore the possibility of severe weather, regardless of the local conditions. This may explain why performance drops significantly at these distances under the perturbation method. On the other hand, pixels within 135 km of the pixel of interest give the local context of the event. Small variations for these pixels can have a large effect on the occurrence or non-occurrence of lightning.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Finally we also study the isotropy of the saliency maps by comparing the saliency maps' radial profiles in specific directions (North, South, East, West). We did not find any significant differences among the directions. The Figure can be found in the supplementary material (see Figure S11).

5 Conclusion and discussion

While vision-based deep learning is now common for lightning nowcasting and forecasting, its use for lightning climatology remains rare, and the local single grid cell approach still dominates. The added value of this study is threefold. First, we provide a controlled and statistically rigorous comparison — over Europe and with a specific focus on the most extreme events — between a spatially-aware U-Net and a four single grid cell models of increasing complexity (logistic regression, generalized additive model, extreme gradient boosting and multi-layer perceptron), isolating the contribution of spatial context. Our leave-one-year-out cross-validation pipeline is designed to minimize temporal leakage and enables us to quantify uncertainties on the evaluation metrics. We show that the U-Net architecture outperforms single grid cell models for lightning modeling in reanalysis data, confirming past results from Kalashnikov et al. (2024); Cheng et al. (2024) and Yin et al. (2026). This holds in particular for the most extreme and impactful events, which are often associated with large-scale convective systems. For such events, the U-Net's access to spatial context is expected to be most advantageous relative to single grid cell models. We find that the U-Net reproduces the large-scale spatial patterns of the most extreme events more accurately than the other models, though it does not resolve their small-scale structure. These results are consistent with the U-Net having access to the spatial context of the predictors, which the other models lack. Incorporating such information is thus beneficial, particularly as the additional training computational cost remains limited (see the carbon footprint disclaimer below).

Second, we show through the SHAP values analysis that the most important predictors depend on both the model architecture and the evaluation metric. For the single grid cell models, we find that the lifted index is the most important predictor for overall performance (predicting both positive and negative samples), confirming the previous study from Battaglioli et al. (2023a). However we also find that convective precipitation is as important a predictor as the lifted index for predicting the positive class. For the U-Net, we find that convective precipitation is the most important predictor for predicting the positive class (PR-AUC), although the differences are not as clear for overall performance (ROC-AUC). This suggests that the selection of predictors should depend on the goal of the study and be done in conjunction with the selection of the model architecture and the evaluation metrics. This result has direct practical consequences for predictor and metric selection in future convective-hazard studies.

Third, through saliency mapping and a radial perturbation method, we quantify the spatial scale that the model exploits, showing that lightning involved in mesoscale convective systems require mesoscale predictors for most accurate prediction. The characteristic radius established in this study is of 300 km around the location of interest. Taken together, these results validate the hypothesis that mesoscale convective clusters are driven by broad spatial predictor patterns that single grid cell models cannot capture.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

While the results are promising, there are several limitations to this study that should be addressed in future work. First, we only predict a probability of occurrence of lightning at each timestep, but it would be interesting to also predict the lightning flash density or the maximum peak current of the strikes, which are may be more informative target variables for risk assessment. Second, we note that in terms of realism of the events, the observations are not directly comparable to the predictions as we are comparing binary events to probabilities, which are intrinsically smoother. Therefore a sampling step is needed to sample binary predictions from the predicted probabilities. To do this sampling step, simple methods such as thresholding or Bernoulli sampling are not satisfactory because they break the spatial and temporal dependencies between neighboring grid cells which are crucial for simulating realistic events. Future work should explore more sophisticated sampling methods that take into account the spatial and temporal correlation of the predictions such as generative models (Goodfellow et al., 2014; Ho et al., 2020). Third, as we train our model on a fixed domain, it cannot generalize to other regions, contrary to the single grid cell models which can in theory be applied anywhere – assuming that the models can generalize to unseen localizations. A vision-based model that is not trained on a fixed domain could be useful to infer lightning probabilities in regions where lightning observations are scarce. Finally, we rely only on a single reanalysis and a single set of lightning observations. Repeating the study over different sets of data and predictors can improve the robustness of the results.

As we focused on the isolation of the contribution of spatial inputs, we only used a limited set of predictors which were selected from past works. We show in our study that changing the architecture of the model can change the importance of the predictors, and that the convective precipitation is a much more important predictor for the U-Net than for the other models. This suggests that there might be other predictors that are particularly useful for the U-Net that we did not include in our study. To improve this model, future work should explore a larger set of predictors, tailored for this kind of architecture to obtain even better predictions.

This work can also lead to multiple weather and climate-oriented applications. The first is applying the model to climate projections such as CMIP6 models. As lightning is an indicator of deep convective activity, projecting future changes in lightning activity under different climate scenarios is relevant for risk assessment. The framework developed here could be applied to climate model outputs to study how the frequency, intensity and spatial distribution of lightning events may evolve at different warming levels. It could help study worst-case scenarios across different sectors such as insurance or energy. The methodology could also be extended to other convective hazards, such as hail and tornadoes, although it would require adaptation because the data associated with those hazards is usually of lower quality since it still relies on crowd-sourced reports (Dotzek et al., 2009; Edwards et al., 2018; Allen et al., 2020; Taszarek et al., 2020; Battaglioli et al., 2023a). Finally, this framework can be adapted to weather forecasts at the medium range (2 to 10 days), for which convection-permitting models are not usually run. It would provide automated medium range convective outlooks for local communities, similarly to framework developed by Taszarek and Matczak (2025).

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Appendix A: Models description

Model	Hyperparameters	Optimization method	Loss function
LR	Learning rate	Gradient descent	Binary cross-entropy
GAM	Spline family (exponential or smoothing splines), smoothing parameter if smoothing splines family	Maximum likelihood estimation	Log-likelihood
XGB	Nb. of trees, learning rate, max depth	Gradient boosting optimization	Quantile regression loss
MLP	Nb. of layers, nb. of neurons per layer, nb. of epochs, learning rate, scheduler, positive-class weight, batch size, nb. of batches per epoch	Gradient descent with backpropagation	Weighted binary cross-entropy
U-Net	Nb. of residual layers, nb. of downsampling layers, nb. of channels, nb. of epochs, learning rate, scheduler, positive-class weight, batch size, nb. of batches per epoch	Gradient descent with backpropagation	Weighted binary cross-entropy

Table A1. Models implemented in this study and their hyperparameters, optimization method and loss function.

A1 U-Net architecture

The convolutional neural network that we use in this study is a U-Net architecture (Ronneberger et al., 2015), which is a widely used architecture for image-to-image tasks. The U-Net architecture consists of an encoder, a bottleneck and a decoder with skip connections between the corresponding layers of the encoder and decoder. The encoder is composed of several convolutional layers and downsampling layers (max pooling), which extract features from the input image and reduce its spatial resolution. The decoder is composed of several convolutional layers and upsampling layers (transposed convolution), which reconstruct the output image from the encoded features and increase its spatial resolution. The skip connections allow the decoder to access high-resolution features from the encoder, which helps improve the output image. We also add residual layers between the convolutional and downsampling / upsampling layers. Table A2 summarizes the hyperparameters of the U-Net that we tune.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Hyperparameter	Values	Optimized model values
Loss function	{Weighted Binary cross-entropy, Focal loss, Dice loss}	Weighted Binary cross-entropy
Nb. of residual layers	{1, 2}	2
Nb. of downsampling / upsampling layers	{2, 3, 4}	2
Number of channels per convolutional layer	{16, 32, 64, 128, 256}	{16, 32, 64}
Learning rate	$[1e-4, 1e-2]$	0.001
Scheduler	$[0.8, 1]$	0.968
Positive-class weight	$[1, 20]$	4
Batch size	{24, 48, 72}	48
Nb. of batches per epoch	$[20, 60]$	47
Nb. of epochs	$[20, 60]$	47

Table A2. U-Net hyperparameter tuning. Each parameter is optimized within a given range of possible values. All hyperparameters are optimized using a Tree-structured Parzen Estimator, except for the loss function which was selected manually.

Appendix B: Leave-One-Year-Out nested Cross-Validation

Our entire pipeline is a "nested cross-validation": inside each LOYO split, we do a 3-fold cross-validation for selecting the best hyperparameters of the models that need hyperparameter tuning. We split the 12 years of training into 3 folds of 4 consecutive years. Then we train the model 3 times, each time using 2 folds for training and 1 fold for validation, and we select the hyperparameters that maximize the average performance on the validation sets across the 3 folds. We use a Tree-structured Parzen Estimator search strategy with the Optuna python package (Akiba et al., 2019) to optimize the hyperparameters. The objective function of this search is the sum of the five metrics described in Section 3.4. We choose to optimize the sum of the metrics rather than one single metric to ensure that we select hyperparameters that perform well across all the different aspects of model performance captured by the different metrics. Once the hyperparameters are tuned, we retrain the model on the whole training set (12 years) and evaluate it on the test set (1 year).

With 13 years in total, each LOYO split uses 12 years for training and 1 year for testing. We can then derive confidence intervals on the test sets performance metrics, which are reported in Section 4.1. We assume that each test metric is normally distributed across splits, and we compute the 95% confidence interval as $(\bar{x} \pm t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}})$, where $\bar{x}$ is the mean test metric across splits, $t_{\alpha/2,n-1}$ is the critical value of the t-distribution with $n-1$ degrees of freedom, σ is the standard deviation across splits, and n is the number of splits (13 in our case).

In the case of the single grid cell models (LR, GAM, XGB and MLP), the input data is a 5-dimensional vector of the 5 environmental variables at a given pixel, and at a given hour. The output is a single value: the probability of lightning occurrence at this pixel and hour (see Figure 1). Due to RAM limitations, we train the GAM and XGB models on a random subset of 2 million samples (0.1% of the training set), which is sufficient for convergence. The LR and MLP are trained on the whole training set with gradient descent. We evaluate the performance of all models on the entire test set.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

In the case of the U-Net, the input is a 5-channel image of size 101×149 pixels, which is subsequently padded to 112×160 pixels and where each channel corresponds to one of the 5 environmental variables at a given hour. The output image is a single-channel image of size 101×149 pixels, where the values correspond to the probabilities of lightning occurrence at each grid cell for the same timestep as the input. The U-Net is trained on the whole training set, and we use a batch size of 48 samples, which corresponds to 48 hours of data. To save time when loading the data, we load full days (24 hours) at a time, and we randomly sample 2 days per batch. To handle the class imbalance for the U-Net and MLP, we use a weighted random sampler in the dataloader, which samples days with a probability proportional to the logarithm of the number of lightning strikes during the day. We also use a weighted binary cross-entropy loss function, where the positive class (lightning occurrence) is weighted with a weight w that is optimized as a hyperparameter during the inner loop of the LOYO procedure (see Table A2).

Appendix C: Uncertainty quantification

Uncertainty quantification is a crucial aspect of model evaluation, as it provides insights into the reliability and robustness of the predictions of the model. However, depending on the context and the question we want to answer, different methods of uncertainty quantification are more or less relevant. We provide below a brief overview of the different methods of uncertainty quantification and their use cases.

The **classical train-test split** is the most straightforward approach for evaluating model performance. Its primary goal is to determine how a specific, single model performs after being trained on a designated subset of data. This method is typically reserved for the final evaluation phase of a project before deployment. However, its reliability depends heavily on scale. It requires a very large test dataset to be able to derive meaningful confidence intervals on the performance metrics.

Bootstrapping's focus is similar but tries to alleviate the need of a very large dataset for evaluating the performance of the model. By repeatedly sampling from the test data, it measures how much the performance varies when the evaluation set changes. This is particularly useful for a final evaluation when the model training process is computationally expensive, as it allows uncertainty quantification without having to retrain the model multiple times. It still requires a large test dataset to be able to derive meaningful confidence intervals on the performance metrics.

Unlike the previous two methods, **cross-validation** is designed to evaluate the effectiveness of an algorithm as a whole rather than a specific trained model instance. It quantifies how well the chosen algorithm performs on unseen data on average. The trade-off is computational: it is best used when training the model k times is not prohibitively expensive, providing a robust average performance metric across multiple "folds."

Appendix D: Detailed metrics

The **PR-AUC** (Davis and Goadrich, 2006) evaluates the Precision and Recall across all possible thresholds and reports the area under the corresponding curve. It quantifies the reliability of the model's positive predictions. The **F1-score** is also a metric that captures the trade-off between precision and recall as it is the harmonic mean of Precision and Recall. However, as it is

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

a threshold-dependent metric and some information is lost in the threshold selection process, we retain the PR-AUC in this study.

The **ROC-AUC** (Hanley and McNeil, 1982) is based on the True Positive Rate and False Positive Rate. It evaluates the ability of the model to discriminate between the positive and negative classes across all possible thresholds, and rank each sample from the least likely to the most likely to belong to the positive class. The issue with the ROC-AUC is that it is very sensitive to class imbalance through the true negatives (TN) value in the denominator of the FPR. Thus when the negative class is largely dominant (as is the case in our study), the FPR can be very low even if the model makes a lot of false positive errors, which can lead to an overestimation of the performance of the model (if we care about minimizing the false positives). It has therefore been argued (Sobash and Ahijevych, 2024) that the PR-AUC is more informative than the ROC-AUC for imbalanced data because it is not influenced by the true negative predictions, which account for most of the data in this case, but are not the main interest of the model. It has however been used widely in previous severe weather modeling studies (Czernecki et al., 2019; Battaglioli et al., 2023a; Sobash and Ahijevych, 2024; Saha et al., 2025; Matczak et al., 2026) and we use it as well for completeness and to compare with previous studies.

The **continuous Dice Coefficient** (Shamir et al., 2018) is a measure of similarity between two images. It is a probabilistic adaptation of the Dice coefficient: $Dice = \frac{2*|A \cap B|}{|A|+|B|}$, where A and B are respectively the predicted and observed binary maps. The Dice coefficient requires a thresholding step, such that A is a binary map. The continuous Dice Coefficient adapts this idea to continuous predictions by translating the union of pixels into a product between predictions and labels at every pixel and the cardinality of A into a sum of probabilities: $Dice = \frac{2*\sum_i (p_i * y_i)}{\sum_i (p_i + y_i)}$. It is an improper scoring rule and is used mostly for probabilistic image segmentation tasks because it mostly captures the sharpness of the predicted probabilities, and is not so sensitive to calibration. The Dice coefficient measures the spatial overlap between the model's high-probability regions and the observed lightning occurrences.

The **FBSS** is a metric with 3 parts. The Fractions part (Roberts and Lean, 2008) refers to the fact that contrary to the other metrics, for all samples, we average the predictions and observations in a neighborhood around the sample to avoid the double penalty problem (McGovern et al., 2023). Here we have selected a (3×3) kernel, centered on the sample, for the neighborhood. The Brier Score (Brier, 1950) measures the mean squared error between the predicted probabilities and the true binary labels: $BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2$. Here in our fractional context we define p_i as the average predicted probability in the neighborhood of sample i, and y_i is the average observed lightning occurrence in the same neighborhood. The Brier Score evaluates both the calibration and the sharpness of the predicted probabilities and is sensitive to both false positives and false negatives. The Skill Score part refers to the fact that we compare the Brier Score of our model to the Brier Score of a baseline model: $FBSS = 1 - \frac{BS_{model}}{BS_{baseline}}$. The baseline model here is a model that does not use any of the predictors, and simply predicts, for each location and each season, the average probability of lightning occurrence in the training set. The FBSS measures, over a spatial neighborhood, how close the predicted probabilities are to the observed outcomes relative to a baseline model.

The **explained Deviance** (also referred to as McFadden's pseudo R^2 (McFadden, 1974)) is a metric that compares the log-likelihood of a model to the log-likelihood of a baseline model: $Deviance = 1 - \frac{LL_{model}}{LL_{baseline}}$. The baseline model is the same as for the FBSS. This metric generalizes the R^2 metric for classification tasks and has also been used in previous studies for

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

model selection (Allen et al., 2015; Battaglioli et al., 2023a). Like the FBSS, the deviance explained is a proper scoring rule that evaluates both calibration and sharpness, but penalizes more large errors (overconfident predictions), for example predicting 99% chance of occurrence for an event that does not occur. The deviance therefore differs slightly from the FBSS: it quantifies the reduction in overconfident errors relative to a baseline model. Keeping both metrics is therefore interesting to have a more complete picture of the model performance.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Code availability. The code to reproduce results from this study is available at the following url: https://github.com/AdrienBq/lightning_modelling.

Data availability. ERA5 postprocessed data used in this work was provided by Mateusz Taszarek under contribution from a grant from the Polish National Science Center (2020/39/D/ST10/00768) and can be made available. Contact him (mateusz.taszarek@amu.edu.pl) for usage information. The lightning data used in this work was provided by Graeme Marlton from the Met Office. Contact him (graeme.marlton@metoffice.gov.uk) for license and usage information.

Sample availability. A sample of the data is available on the same github repository as the code : https://github.com/AdrienBq/lightning_modelling.

Carbon footprint. The carbon footprint of this study is estimated to be 6.5 kg CO_2, which is equivalent to the carbon footprint of 30 km driven by an average car. This estimation is based on the energy consumption of training the models (on V100 GPUs for the U-Net and AMD EPYC 7302 16-Core CPUs for the other models), running the analyses and storing the data. We used the French average carbon intensity of electricity for the energy consumption estimation. Training and evaluating once the U-Net model takes 20 minutes on a V100 GPU which has a carbon footprint of around 25 gCO_2/h, so the footprint of training the U-Net model once is around 8.3 gCO_2. The bulk of the footprint comes from the optimization of the U-Net's hyperparameters and SHAP values analysis which combines 52 trainings (20 for hyperparameter tuning 32 for the SHAP analyses) for each of the 13 LOYO splits. This amounts to around 5.8 kg CO_2 for the U-Net alone. The remaining 0.7 kg comes from the other analyses (0.5 kg) and data storage (0.2 kg).

Author contributions. All authors designed the study. AB implemented the models, ran the analyses and wrote the article. DF, MV, VX, JJ, VF and VB provided technical advice and guidance during the study, and contributed to writing and reviewing the article.

Competing interests. The authors have no other competing interests to declare.

Acknowledgements. We acknowledge the financial support of Descartes Underwriting, along with the useful discussions within the R&D team at Descartes. We also acknowledge the ESTIMR team at LSCE, the PowdDev consortium (ANR-22-PETA-0016), as well as the TRACCS PC 4 (ANR-22-EXTR-0005) and 10 (ANR-22-EXTR-00011) consortia. We also thank the Polish National Science Center (grant 2020/39/D/ST10/00768). We also acknowledge Graeme Marlton from the Met Office who provided the lightning data. We give special thanks to Mikel Legasa, Flavio Pons and Maximilien Burq for the insightful discussions and suggestions during this research. Artificial intelligence tools were utilized to assist in code development and proofread the manuscript.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

References

Abbott, T. H., Jeevanjee, N., Cheng, K., Zhou, L., and Harris, L.: The Land-Ocean Contrast in Deep Convective Intensity in a Global Storm-Resolving Model, Journal of Advances in Modeling Earth Systems, 17, https://doi.org/10.1029/2024ms004467, 2025.

Aguilera, R., Corringham, T., Gershunov, A., and Benmarhnia, T.: Wildfire smoke impacts respiratory health more than fine particles from other sources: observational evidence from Southern California, Nature Communications, 12, https://doi.org/10.1038/s41467-021-21708-0, 2021.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., and Kindermans, P.-J.: iNNvestigate Neural Networks!, Journal of Machine Learning Research, 20, 1–8, http://jmlr.org/papers/v20/18-540.html, 2019.

Allen, J. T., Tippett, M. K., and Sobel, A. H.: An empirical model relating U.S. monthly hail occurrence to large-scale meteorological environment, Journal of Advances in Modeling Earth Systems, 7, 226–243, https://doi.org/10.1002/2014MS000397, 2015.

Allen, J. T., Giammanco, I. M., Kumjian, M. R., Jurgen Punge, H., Zhang, Q., Groenemeijer, P., Kunz, M., and Ortega, K.: Understanding Hail in the Earth System, Reviews of Geophysics, 58, https://doi.org/10.1029/2019rg000665, 2020.

Battaglioli, F., Groenemeijer, P., Púčik, T., Taszarek, M., Ulbrich, U., and Rust, H.: Modeled Multidecadal Trends of Lightning and (Very) Large Hail in Europe and North America (1950–2021), Journal of Applied Meteorology and Climatology, 62, https://doi.org/10.1175/JAMC-D-22-0195.1, 2023a.

Battaglioli, F., Groenemeijer, P., Tsonevsky, I., and Púčik, T.: Forecasting large hail and lightning using additive logistic regression models and the ECMWF reforecasts, Natural Hazards and Earth System Sciences, 23, 3651–3669, https://doi.org/10.5194/nhess-23-3651-2023, 2023b.

Battaglioli, F., Taszarek, M., Groenemeijer, P., Púčik, T., and Rädler, A.: Contrasting trends in very large hail events and related economic losses across the globe, Nature Geoscience, 19, 52–58, https://doi.org/10.1038/s41561-025-01868-0, 2026.

Bosc, M., Chan-Hon-Tong, A., Bouchard, A., and Béréziat, D.: Predicting thunderstorm risk probability at very short time range using deep learning, https://doi.org/10.5194/egusphere-2025-2893, 2025a.

Bosc, M., Chan-Hon-Tong, A., Bouchard, A., and Béréziat, D.: StrikeNet: A Deep Neural Network to Predict Pixel-Sized Lightning Location:, in: Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 299–306, SCITEPRESS - Science and Technology Publications, Porto, Portugal, ISBN 978-989-758-728-3, https://doi.org/10.5220/0013110700003912, 2025b.

Brier, G. W.: VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY, Monthly Weather Review, 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2, 1950.

Brodehl, S., Müller, R., Schömer, E., Spichtinger, P., and Wand, M.: End-to-End Prediction of Lightning Events from Geostationary Satellite Images, Remote Sensing, 14, 3760, https://doi.org/10.3390/rs14153760, 2022.

Campos, C., Couto, F. T., Santos, F. L. M., Rio, J., Ferreira, T., and Salgado, R.: ECMWF Lightning Forecast in Mainland Portugal during Four Fire Seasons, Atmosphere, 15, 156, https://doi.org/10.3390/atmos15020156, 2024.

Cavaiola, M., Cassola, F., Sacchetti, D., Ferrari, F., and Mazzino, A.: Hybrid AI-enhanced lightning flash prediction in the medium-range forecast horizon, Nature Communications, 15, 1188, https://doi.org/10.1038/s41467-024-44697-2, 2024.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Charalambous, C. A., Dimitriou, A., Kokkinos, N., Kioupis, N., and Manolis, T.: Lightning strikes and buried gas pipelines: Mechanisms, risk assessment, mitigation, and case insights from the hellenic gas transmission system, Electric Power Systems Research, 255, 112 809, https://doi.org/10.1016/j.epsr.2026.112809, 2026.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, p. 785–794, ACM, https://doi.org/10.1145/2939672.2939785, 2016.

Cheng, W.-Y., Kim, D., Henderson, S., Ham, Y.-G., Kim, J.-H., and Holzworth, R. H.: Machine Learning–Based Lightning Parameterizations for the CONUS, Artificial Intelligence for the Earth Systems, 3, e230 024, https://doi.org/10.1175/AIES-D-23-0024.1, 2024.

Chmielewski, T., Nowak, H., and Walkowiak, K.: Tornado in Poland of August 15, 2008: Results of post-disaster investigation, Journal of Wind Engineering and Industrial Aerodynamics, 118, 54–60, https://doi.org/10.1016/j.jweia.2013.04.007, 2013.

Czernecki, B., Taszarek, M., Marosz, M., Półrolniczak, M., Kolendowicz, L., Wyszogrodzki, A., and Szturc, J.: Application of machine learning to large hail prediction - The importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5, Atmospheric Research, 227, 249–262, https://doi.org/10.1016/j.atmosres.2019.05.010, 2019.

Czernecki, B., Taszarek, M., and Szuster, P.: thunder: Computation and Visualisation of Atmospheric Convective Parameters, https://bczernecki.github.io/thundeR/, r package version 1.1.5, 2025.

Davis, J. and Goadrich, M.: The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd international conference on Machine learning - ICML '06, ICML '06, p. 233–240, ACM Press, https://doi.org/10.1145/1143844.1143874, 2006.

Diffenbaugh, N. S., Scherer, M., and Trapp, R. J.: Robust increases in severe thunderstorm environments in response to greenhouse forcing, Proceedings of the National Academy of Sciences, 110, 16 361–16 366, https://doi.org/10.1073/pnas.1307758110, 2013.

DiGangi, E. A., Stock, M., and Lapierre, J.: Thunder Hours: How Old Methods Offer New Insights into Thunderstorm Climatology, Bulletin of the American Meteorological Society, 103, E548–E569, https://doi.org/10.1175/BAMS-D-20-0198.1, 2022.

Doswell, C. A., Brooks, H. E., and Maddox, R. A.: Flash Flood Forecasting: An Ingredients-Based Methodology, Weather and Forecasting, 11, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2, 1996.

Dotzek, N., Groenemeijer, P., Feuerstein, B., and Holzer, A.: Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD, Atmospheric Research, 93, 575–586, https://doi.org/10.1016/j.atmosres.2008.10.020, 2009.

Edwards, R., Allen, J. T., and Carbin, G. W.: Reliability and Climatological Impacts of Convective Wind Estimations, Journal of Applied Meteorology and Climatology, 57, 1825–1845, https://doi.org/10.1175/jamc-d-17-0306.1, 2018.

Ehrensperger, G., Simon, T., Mayr, G. J., and Hell, T.: Identifying lightning processes in ERA5 soundings with deep learning, Geoscientific Model Development, 18, 1141–1153, https://doi.org/10.5194/gmd-18-1141-2025, 2025.

Enno, S.-E., Anderson, G., and Sugier, J.: ATDnet Detection Efficiency and Cloud Lightning Detection Characteristics from Comparison with the HyLMA during HyMeX SOP1, Journal of Atmospheric and Oceanic Technology, 33, 1899–1911, https://doi.org/10.1175/JTECH-D-15-0256.1, 2016.

Enno, S.-E., Sugier, J., Alber, R., and Seltzer, M.: Lightning flash density in Europe based on 10 years of ATDnet data, Atmospheric Research, 235, 104 769, https://doi.org/10.1016/j.atmosres.2019.104769, 2020.

Feng, Z., Leung, L. R., Hagos, S., Houze, R. A., Burleyson, C. D., and Balaguru, K.: More frequent intense and long-lived storms dominate the springtime trend in central US rainfall, Nature Communications, 7, 13 429, https://doi.org/10.1038/ncomms13429, 2016.

Gaffard, C., Nash, J., Atkinson, N., Bennett, A., Callaghan, G., Hibbett, E., Turp, M., and Schulz, W.: Observing Lightning Around the Globe from the Surface, The Preprints, 20th Interna-tional Lightning Detection Conference, 2008.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Galway, J. G.: The Lifted Index as a Predictor of Latent Instability, Bulletin of the American Meteorological Society, 37, 528–529, https://doi.org/10.1175/1520-0477-37.10.528, 1956.

Geng, Y., Li, Q., Lin, T., Yao, W., Xu, L., Zheng, D., Zhou, X., Zheng, L., Lyu, W., and Zhang, Y.: A deep learning framework for lightning forecasting with multi-source spatiotemporal data, Quarterly Journal of the Royal Meteorological Society, 147, 4048–4062, https://doi.org/10.1002/qj.4167, 2021.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative Adversarial Networks, https://doi.org/10.48550/ARXIV.1406.2661, 2014.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q.: On Calibration of Modern Neural Networks, https://doi.org/10.48550/arXiv.1706.04599, arXiv:1706.04599 [cs], 2017.

Hanley, J. A. and McNeil, B. J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve., Radiology, 143, 29–36, https://doi.org/10.1148/radiology.143.1.7063747, 1982.

Haseeb, S. A. and Krawczuk, M.: A State-of-the-Art Review of Structural Health Monitoring Techniques for Wind Turbine Blades, Journal of Nondestructive Evaluation, 45, https://doi.org/10.1007/s10921-025-01296-5, 2025.

Hastie, T. and Tibshirani, R.: Generalized Additive Models, Statistical Science, 1, https://doi.org/10.1214/ss/1177013604, 1986.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Ho, J., Jain, A., and Abbeel, P.: Denoising Diffusion Probabilistic Models, https://doi.org/10.48550/ARXIV.2006.11239, 2020.

Holderrieth, P. and Erives, E.: Introduction to Flow Matching and Diffusion Models, https://diffusion.csail.mit.edu/, 2025.

Holle, R. L.: A Summary of Recent National-Scale Lightning Fatality Studies, Weather, Climate, and Society, 8, 35–42, https://doi.org/10.1175/wcas-d-15-0032.1, 2016.

Hong, S. and Lim, J.-O. J.: The WRF Single-Moment 6-Class Microphysics Scheme (WSM6), Asia-pacific Journal of Atmospheric Sciences, 42, 129–151, https://api.semanticscholar.org/CorpusID:120362377, 2006.

Huntrieser, H., Lichtenstern, M., Scheibe, M., Aufmhoff, H., Schlager, H., Pucik, T., Minikin, A., Weinzierl, B., Heimerl, K., Pollack, I. B., Peischl, J., Ryerson, T. B., Weinheimer, A. J., Honomichl, S., Ridley, B. A., Biggerstaff, M. I., Betten, D. P., Hair, J. W., Butler, C. F., Schwartz, M. J., and Barth, M. C.: Injection of lightning-produced NOx, water vapor, wildfire emissions, and stratospheric air to the UT/LS as observed from DC3 measurements, Journal of Geophysical Research: Atmospheres, 121, 6638–6668, https://doi.org/10.1002/2015jd024273, 2016.

Kain, J. S., Weiss, S. J., Bright, D. R., Baldwin, M. E., Levit, J. J., Carbin, G. W., Schwartz, C. S., Weisman, M. L., Droegemeier, K. K., Weber, D. B., and Thomas, K. W.: Some Practical Considerations Regarding Horizontal Resolution in the First Generation of Operational Convection-Allowing NWP, Weather and Forecasting, 23, 931–952, https://doi.org/10.1175/waf2007106.1, 2008.

Kalashnikov, D. A., Davenport, F. V., Labe, Z. M., Loikith, P. C., Abatzoglou, J. T., and Singh, D.: Predicting Cloud-To-Ground Lightning in the Western United States From the Large-Scale Environment Using Explainable Neural Networks, Journal of Geophysical Research: Atmospheres, 129, e2024JD042 147, https://doi.org/10.1029/2024JD042147, 2024.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Kaltenböck, R., Diendorfer, G., and Dotzek, N.: Evaluation of thunderstorm indices from ECMWF analyses, lightning data and severe storm reports, Atmospheric Research, 93, 381–396, https://doi.org/https://doi.org/10.1016/j.atmosres.2008.11.005, 4th European Conference on Severe Storms, 2009.

Kühne, T., Antonescu, B., Groenemeijer, P., and Púčik, T.: Lightning Fatalities in Europe (2001–20), Weather, Climate, and Society, 17, 205–215, https://doi.org/10.1175/wcas-d-24-0038.1, 2025.

Kumar, S., N., U., Gautam, A. S., Potdar, S. S., Biswasharma, R., Singh, K., Siingh, D., and Singh, R. P.: Evaluation of WRF Microphysics Schemes for Simulating Lightning and Rainfall over the Complex Terrain of the Western Himalayan Region, https://doi.org/10.22541/essoar.177265325.51198794/v1, 2026.

Lee, A. C. L.: An Operational System for the Remote Location of Lightning Flashes Using a VLF Arrival Time Difference Technique, Journal of Atmospheric and Oceanic Technology, 3, 630 – 642, https://doi.org/10.1175/1520-0426(1986)003<0630:AOSFTR>2.0.CO;2, 1986.

Leinonen, J., Hamann, U., and Germann, U.: Seamless Lightning Nowcasting with Recurrent-Convolutional Deep Learning, Artificial Intelligence for the Earth Systems, 1, e220 043, https://doi.org/10.1175/AIES-D-22-0043.1, 2022.

Leinonen, J., Hamann, U., Sideris, I. V., and Germann, U.: Thunderstorm Nowcasting With Deep Learning: A Multi-Hazard Data Fusion Model, Geophysical Research Letters, 50, e2022GL101 626, https://doi.org/10.1029/2022GL101626, 2023.

Li, M., Cheng, S., Wang, J., Cai, L., Fan, Y., Cao, J., and Zhou, M.: Thunderstorm total lightning activity behavior associated with transmission line trip events of power system, npj Climate and Atmospheric Science, 7, https://doi.org/10.1038/s41612-024-00697-z, 2024.

Lin, J.: Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory, 37, 145–151, https://doi.org/10.1109/18.61115, 1991.

Lin, T., Li, Q., Geng, Y.-A., Jiang, L., Xu, L., Zheng, D., Yao, W., Lyu, W., and Zhang, Y.: Attention-Based Dual-Source Spatiotemporal Neural Network for Lightning Forecast, IEEE Access, 7, 158 296–158 307, https://doi.org/10.1109/ACCESS.2019.2950328, 2019.

Loken, E. D., Clark, A. J., Xue, M., and Kong, F.: Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble, Weather and Forecasting, 32, 1403–1421, https://doi.org/10.1175/waf-d-16-0200.1, 2017.

Lopez, P.: A Lightning Parameterization for the ECMWF Integrated Forecasting System, Monthly Weather Review, 144, 3057–3075, https://doi.org/10.1175/MWR-D-16-0026.1, 2016.

Lundberg, S. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, https://doi.org/10.48550/ARXIV.1705.07874, 2017.

Manzato, A., Fasano, G., Cicogna, A., Sioni, F., and Pucillo, A.: Relationships between Environmental Parameters and Storm Observations in Po Valley: Are They Climate Change Invariant?, Journal of Applied Meteorology and Climatology, 64, 267–298, https://doi.org/10.1175/jamc-d-24-0034.1, 2025.

Marlton, G., Potts, M., Twelves, S., Prust, S., Stone, E., and O'Sullivan, D.: LEELA: The Met Offices next generation lightning location system, https://doi.org/10.5194/egusphere-egu22-2541, 2022.

Matczak, P., Taszarek, M., Sobisiak, A., and Kolendowicz, L.: Environments Associated With Lightning Occurrence Based on Pre- and Post-Convective Rawinsonde Measurements in Central Europe, Geophysical Research Letters, 53, https://doi.org/10.1029/2025gl119901, 2026.

Matsui, T., Chern, J.-D., Tao, W.-K., Lang, S., Satoh, M., Hashino, T., and Kubota, T.: On the Land–Ocean Contrast of Tropical Convection and Microphysics Statistics Derived from TRMM Satellite Signals and Global Storm-Resolving Models, Journal of Hydrometeorology, 17, 1425–1445, https://doi.org/10.1175/jhm-d-15-0111.1, 2016.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Mazurek, A. C., Hill, A. J., Schumacher, R. S., and McDaniel, H. J.: Can Ingredients-Based Forecasting Be Learned? Disentangling a Random Forest's Severe Weather Predictions, Weather and Forecasting, 40, 237–258, https://doi.org/10.1175/WAF-D-23-0193.1, 2025.

McClung, B., McGovern, A., Hill, A. J., Schvartzman, D., and Stock, M.: BoltCast: Medium-range Lightning Prediction with Neural and Long-Short Term Memory Networks, Artificial Intelligence for the Earth Systems, https://doi.org/10.1175/aies-d-25-0044.1, 2026.

McFadden, D.: Conditional logit analysis of qualitative choice behavior, in: Fontiers in Econometrics, edited by Zarembka, P., pp. 105–142, Academic press, New York, 1974.

McGovern, A., Chase, R. J., Flora, M., Gagne, D. J., Lagerquist, R., Potvin, C. K., Snook, N., and Loken, E.: A Review of Machine Learning for Convective Weather, Artificial Intelligence for the Earth Systems, 2, e220 077, https://doi.org/10.1175/AIES-D-22-0077.1, 2023.

McNulty, R. P.: A CONCEPTUAL APPROACH TO THUNDERSTORM FORECASTING, National Weather Digest, 10, 1985.

Molnar, C.: Interpretable Machine Learning, 3 edn., ISBN 978-3-911578-03-5, https://christophm.github.io/interpretable-ml-book, 2025.

Morrison, H., Thompson, G., and Tatarskii, V.: Impact of Cloud Microphysics on the Development of Trailing Stratiform Precipitation in a Simulated Squall Line: Comparison of One- and Two-Moment Schemes, Monthly Weather Review, 137, 991–1007, https://doi.org/10.1175/2008MWR2556.1, 2009.

Mulholland, J. P., Peters, J. M., and Morrison, H.: How Does Vertical Wind Shear Influence Entrainment in Squall Lines?, Journal of the Atmospheric Sciences, 78, 1931–1946, https://doi.org/10.1175/jas-d-20-0299.1, 2021.

NOAA: Mesoscale Convective Systems, https://forecast.weather.gov/glossary.php?word=MESOSCALE, [Accessed 09-03-2026].

NOAA: Ingredients for a Thunderstorm, https://www.noaa.gov/jetstream/thunderstorms/ingredients-for-thunderstorm, [Accessed 09-03-2026], 2023.

Pilguj, N., Taszarek, M., Allen, J. T., and Hoogewind, K. A.: Are Trends in Convective Parameters over the United States and Europe Consistent between Reanalyses and Observations?, Journal of Climate, 35, 3605–3626, https://doi.org/10.1175/jcli-d-21-0135.1, 2022.

Piper, D. and Kunz, M.: Spatiotemporal variability of lightning activity in Europe and the relation to the North Atlantic Oscillation teleconnection pattern, Natural Hazards and Earth System Sciences, 17, 1319–1336, https://doi.org/10.5194/nhess-17-1319-2017, 2017.

Platt, J. et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers, 10, 61–74, 1999.

Poelman, D. R., Honoré, F., Anderson, G., and Pedeboy, S.: Comparing a Regional, Subcontinental, and Long-Range Lightning Location System over the Benelux and France, Journal of Atmospheric and Oceanic Technology, 30, 2394–2405, https://doi.org/10.1175/JTECH-D-12-00263.1, 2013a.

Poelman, D. R., Schulz, W., and Vergeiner, C.: Performance Characteristics of Distinct Lightning Detection Networks Covering Belgium, Journal of Atmospheric and Oceanic Technology, 30, 942–951, https://doi.org/10.1175/JTECH-D-12-00162.1, 2013b.

Popick, B. W. F.: Spatiotemporal Prediction of Atmospheric Events Through Recurrent Deep Learning Model, 2025.

Prein, A. F., Liu, C., Ikeda, K., Trier, S. B., Rasmussen, R. M., Holland, G. J., and Clark, M. P.: Increased rainfall volume from future convective storms in the US, Nature Climate Change, 7, 880–884, https://doi.org/10.1038/s41558-017-0007-7, 2017.

Price, C. and Rind, D.: A simple lightning parameterization for calculating global lightning distributions, Journal of Geophysical Research: Atmospheres, 97, 9919–9933, https://doi.org/10.1029/92JD00719, 1992.

Pérez-Invernón, F. J., Huntrieser, H., Jöckel, P., and Gordillo-Vázquez, F. J.: A parameterization of long-continuing-current (LCC) lightning in the lightning submodel LNOX (version 3.0) of the Modular Earth Submodel System (MESSy, version 2.54), Geoscientific Model Development, 15, 1545–1565, https://doi.org/10.5194/gmd-15-1545-2022, 2022.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Pérez-Invernón, F. J., Gordillo-Vázquez, F. J., Huntrieser, H., and Jöckel, P.: Variation of lightning-ignited wildfire patterns under climate change, Nature Communications, 14, https://doi.org/10.1038/s41467-023-36500-5, 2023.

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Monthly Weather Review, 136, 78–97, https://doi.org/10.1175/2007mwr2123.1, 2008.

Romps, D. M., Charn, A. B., Holzworth, R. H., Lawrence, W. E., Molinari, J., and Vollaro, D.: CAPE Times P Explains Lightning Over Land But Not the Land-Ocean Contrast, Geophysical Research Letters, 45, https://doi.org/10.1029/2018gl080267, 2018.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://arxiv.org/abs/1505.04597, 2015.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, Nature, 323, 533–536, https://doi.org/10.1038/323533a0, 1986.

Rädler, A. T., Groenemeijer, P., Faust, E., and Sausen, R.: Detecting Severe Weather Trends Using an Additive Regressive Convective Hazard Model (AR-CHaMo), Journal of Applied Meteorology and Climatology, 57, 569–587, https://doi.org/10.1175/JAMC-D-17-0132.1, 2018.

Saha, K., Bisht, D. S., and Ashrit, R.: Prediction of lightning events over Bangladesh: A machine learning perspective, Journal of Atmospheric and Solar-Terrestrial Physics, 268, 106 448, https://doi.org/10.1016/j.jastp.2025.106448, 2025.

Schaeffler, L.: Differences of lightning over land and sea in the Mediterranean region, https://ulb-dok.uibk.ac.at/ulbtirolhs/content/titleinfo/12635776, [Accessed 10-03-2026], 2025.

Sha, Y., Sobash, R. A., and Gagne, D. J.: Generative Ensemble Deep Learning Severe Weather Prediction from a Deterministic Convection-Allowing Model, Artificial Intelligence for the Earth Systems, 3, e230 094, https://doi.org/10.1175/AIES-D-23-0094.1, 2024.

Shamir, R. R., Duchin, Y., Kim, J., Sapiro, G., and Harel, N.: Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations, https://doi.org/10.1101/306977, 2018.

Sobash, R. A. and Ahijevych, D. A.: Evaluating Machine Learning–Based Probabilistic Convective Hazard Forecasts Using The HRRR: Quantifying Hazard Predictability and Sensitivity to Training Choices, Weather and Forecasting, 39, 1399–1415, https://doi.org/10.1175/WAF-D-23-0221.1, 2024.

Sobash, R. A. and Kain, J. S.: Seasonal Variations in Severe Weather Forecast Skill in an Experimental Convection-Allowing Model, Weather and Forecasting, 32, 1885–1902, https://doi.org/10.1175/waf-d-17-0043.1, 2017.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, CoRR, abs/1703.01365, http://arxiv.org/abs/1703.01365, 2017.

Surowiecki, A., Pilguj, N., Taszarek, M., Piasecki, K., Púčik, T., and Brooks, H. E.: Quasi-Linear Convective Systems and Derechos across Europe: Climatology, Accompanying Hazards, and Societal Impacts, Bulletin of the American Meteorological Society, 105, E1619 – E1643, https://doi.org/10.1175/BAMS-D-23-0257.1, 2024.

Taszarek, M. and Matczak, P.: Automated Severe Thunderstorm Oulooks from thundeR Package (ASTORP), https://doi.org/10.5194/ecss2025-226, 2025.

Taszarek, M., Allen, J. T., Groenemeijer, P., Edwards, R., Brooks, H. E., Chmielewski, V., and Enno, S.-E.: Severe Convective Storms across Europe and the United States. Part I: Climatology of Lightning, Large Hail, Severe Wind, and Tornadoes, Journal of Climate, 33, 10 239–10 261, https://doi.org/10.1175/JCLI-D-20-0345.1, 2020.

Taszarek, M., Allen, J. T., Brooks, H. E., Pilguj, N., and Czernecki, B.: Differing Trends in United States and European Severe Thunderstorm Environments in a Warming Climate, Bulletin of the American Meteorological Society, 102, E296–E322, https://doi.org/10.1175/BAMS-D-20-0004.1, 2021a.

https://doi.org/10.5194/egusphere-2026-4363
Preprint. Discussion started: 17 August 2026

Taszarek, M., Allen, J. T., Marchio, M., and Brooks, H. E.: Global climatology and trends in convective environments from ERA5 and rawinsonde data, npj Climate and Atmospheric Science, 4, 35, https://doi.org/10.1038/s41612-021-00190-x, 2021b.

Thompson, G., Rasmussen, R. M., and Manning, K.: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part I: Description and Sensitivity Analysis, Monthly Weather Review, 132, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2, 2004.

van der Schaaf, A. and van Hateren, J.: Modelling the Power Spectra of Natural Images: Statistics and Information, Vision Research, 36, 2759–2770, https://doi.org/10.1016/0042-6989(96)00002-8, 1996.

Wang, X., Hu, K., Wu, Y., and Zhou, W.: A Survey of Deep Learning-Based Lightning Prediction, Atmosphere, 14, 1698, https://doi.org/10.3390/atmos14111698, 2023.

Westermayer, A. T., Groenemeijer, P., Pistotnik, G., Sausen, R., and Faust, E.: Identification of favorable environments for thunderstorms in reanalysis data, Meteorologische Zeitschrift, 26, 59–70, https://doi.org/10.1127/metz/2016/0754, 2017.

Yair, Y., Lynn, B., Price, C., Kotroni, V., Lagouvardos, K., Morin, E., Mugnai, A., and Llasat, M. d. C.: Predicting the potential for lightning activity in Mediterranean storms based on the Weather Research and Forecasting (WRF) model dynamic and microphysical fields, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2008jd010868, 2010.

Yin, M., Wang, Y., Fang, Y., Li, F., Qie, X., Wang, Y., Wei, D., Yuan, H., Ji, Y., Zhang, Y., and Liu, J.: A Two-Dimensional Deep Learning Scheme With One Predicator Only to Parametrize Global Lightning, Geophysical Research Letters, 53, https://doi.org/10.1029/2025gl120502, 2026.

Zeiler, M. D. and Fergus, R.: Visualizing and Understanding Convolutional Networks, CoRR, abs/1311.2901, http://arxiv.org/abs/1311.2901, 2013.