Predicting thunderstorm risk probability at very short time range using deep learning

Mélanie Bosc[1], Adrien Chan-Hon-Tong[2], Aurélie Bouchard[1], and Dominique Béréziat[3]

[1]DPHY, ONERA, Université Paris-Saclay, 91120, Palaiseau, France
[2]DTIS, ONERA, Université Paris-Saclay, 91120, Palaiseau, France
[3]CNRS, Sorbonne Université, 75005, Paris, France

Correspondence: Mélanie Bosc (melanie.bosc@onera.fr)

Abstract. Forecasting electrical activity within the atmosphere remains one of the most challenging predictions, especially due to the chaotic nature of thunderstorms. Lightning strikes are precisely located and occur very quickly, which makes this task particularly difficult. Additionally, these phenomena pose a significant risk to aviation, as they statistically strike each aircraft more than once per year. Over the years, several techniques have been employed for very short-term lightning forecasting (lower than one hour and every five minutes), such as observation-based methods and, more recently, deep learning methods. Previous studies often face difficulties in accurately forecasting lightning probability, and even with AI-driven methods, it is still difficult to obtain calibrated outputs. To address this limitation, we propose a methodology that successfully predicts lightning risk using Convolutional Neural Networks (CNNs) with attention mechanisms. The network is fed with satellite observations and Numerical Weather Prediction (NWP) outputs formatted as a spatio-temporal sequence. Results show a F_1 score of 0.65 for 5-minute predictions and 0.5 for 30-minute predictions with a very low Expected Calibration Error (ECE) of less than 10%. Thanks to the well-calibrated outputs, risk probability maps can be plotted, showing areas with strong to low chances of having electrical activity.

1 Introduction

In the context of aeronautics, one of the main dangers that planes encounter along their routes is the presence of cumulonimbus clouds. Indeed, these clouds are associated with strong updrafts and downdrafts, as well as the potential presence of hail and severe turbulence. In addition, these clouds are the main natural lightning generators on Earth, representing a major hazard for airplanes (Holton, 2004). Commercial aircraft are, on average, struck by lightning once a year (Uman, 1987), which requires grounding and mandatory maintenance operations to ensure that no significant damage has occurred. These operations lead to both time and financial losses for airline companies. Furthermore, when lightning strikes an aircraft, its structure may be damaged depending on the intensity of the strike (Chemartin et al., 2012). Electronic malfunctions can also occur, and lightning-induced arcs may cause fuel tank explosions (Plumer et al., 1982; Laroche et al., 2012), potentially leading to serious accidents, as was the case with Pan Am Flight 214 from San Juan to Baltimore, which resulted in the deaths of 81 passengers (Laroche et al., 2015). Lightning strikes also pose a human risk, as they may hurt pilots and onboard people (EASA, 2018; Bouchard et al., 2024; Mäkelä et al., 2013). Beyond aeronautics, thunderstorms also threaten various sectors such as

25 public safety, insurance, agriculture, and even rocket launches. To this end, certain standards have been established, such as the requirement for commercial aircraft always to remain at least 32 km away from a thunderstorm classified as dangerous or from areas with very intense radar reflectivity signals (FAA, 2013). Other methods have also been developed, such as the implementation of a tactical support system based on statistical analysis of thunderstorm trends, which allows up to 80 % of lightning events to be avoided (Yoshikawa and Ushio, 2019). Finally, lightning detectors can also be directly integrated with
30 the radar systems or lightning mappers located in the aircraft's nose to alert pilots of lightning activity along their route (EASA, 2018; Milani, 2025).

To enhance overall air safety, the ALBATROS project[1] from the European Union Horizon Europe was launched in 2023. The present study aligns with one of ALBATROS's objectives, which is to develop safety risk models to predict and prevent emerging hazards in aviation, and particularly lightning ones. Given that lightning is a dangerous phenomenon, the ability
35 to forecast it is therefore essential. In addition, cumulonimbus clouds, which form when moisture, instability and a triggering mechanism are present, can become very large in the case of multicellular storm systems, whereas isolated storms typically have a characteristic size of approximately 10 km and a duration of about 1 hour. Consequently, forecasting such punctual events is particularly challenging due to the chaotic nature of storm systems and lightning itself, as well as their spatio-temporal scales. Several forecasting methods have been developed to predict thunderstorms and lightning activity for lead times of less than a
40 few hours (Wilson, 1998), which corresponds to a nowcasting task.

On the one hand, numerous prediction methods have been developed based on the use of one type of observation. For example, thunderstorms can be forecast using radar data, which are then extrapolated to predict future electrical activity (Dixon and Wiener, 1993; Johnson et al., 1998; Handwerker, 2002). Extrapolation techniques have also been developed using satellite data alone (Zinner et al., 2008), lightning detection network data (Betz et al., 2008; Pédeboy et al., 2016), or radiosonde data (Sénési
45 and Thepenier, 1997). Techniques based on the fusion of heterogeneous data can also improve forecast performances through extrapolation, as in Meyer (2010); Kober and Tafferner (2009); Burrows et al. (2005), or by using optical flow methods which rely on satellite or ground-based lightning detection data to predict convection (Müller et al., 2022).

On the other hand, statistical methods such as belief functions (Dezert, 2021; Bouchard et al., 2022), fuzzy logic-based approaches (James et al., 2018), or integro-differential modeling (North et al., 2020) can be used to estimate thunderstorm threat
50 probability based on storm-related predictors. Other methods rely directly on Numerical Weather Prediction (NWP) models, which simulate and forecast the convective processes responsible for the initiation and maintenance of thunderstorms (Lynn et al., 2012; Dafys et al., 2018). The outputs of these models may be combined with ensemble techniques to produce probabilistic weather forecasts, such as in Bouttier and Marchal (2020).

However, extrapolation-based forecasts become unreliable beyond a 30-minute lead time due to a significant drop in per-
55 formance (Wilson, 1998). In parallel, statistical approaches are often highly complex, requiring a deep understanding of thunderstorm physics and involving high computational costs. These limitations have motivated the development of alternative forecasting strategies.

---

[1]https://www.albatros-horizon.eu

In recent years, the advent of Artificial Intelligence (AI) has enabled significant progress in weather prediction methods. Meteorological sensors are now widespread, and a vast amount of data is available from observations (satellite-based, ship-based, airborne, ground-based platforms), as well as from reanalysis derived from NWP outputs. Neural Networks (NNs) have made it possible to leverage this data to produce long-term forecasts on a global scale that can compete with traditional NWP methods (Bodnar et al., 2024; Andrychowicz et al., 2023; Lam et al., 2023; Price et al., 2024; Couairon et al., 2024).

IA methods have shown promising results in forecasting mesoscale meteorological phenomena such as convective cells or intense convection (Pan et al., 2021; Bouget et al., 2021). In addition, various studies have demonstrated the feasibility of predicting thunderstorms or lightning activity using NN and different types of input data, such as ground-based lightning detection networks, satellite observations, radar data (Collins et al., 2016; Brodehl et al., 2022; Geng et al., 2021; Bosc et al., 2024; Zhou et al., 2020; Cintineo et al., 2022) and even NWP outputs (Creswick, 2025; Korpinen et al., 2024; Leinonen et al., 2023). These approaches have been applied successfully across a wide range of forecasting horizons, extending up to 2, 3, or even 6 hours. However, AI-based methods tend to be poorly calibrated and often rely on radar data, which limits their suitability for forecasting lightning along aircraft flight paths.

In this work, we developed an NN-based method to forecast lightning risk probability at mesoscale, every 5 minutes up to 1 hour, providing calibrated output maps without relying on radar data, which leverages spatio-temporal sequences to generate forecasts. Several AI-based architectures are already well adapted to this type of input, such as ConvLSTM (Shi et al., 2015) and PredRNN (Wang et al., 2022), making them particularly suitable for weather forecasting tasks. More recent approaches, such as attention mechanisms, have shown improvements in time series prediction (Janny et al., 2022; Yu et al., 2024; Lin et al., 2019), which motivated the use of an attention-based model, ED-DRAP (Che et al., 2022), in our work. Beyond architectural choices, the design of the loss function plays a critical role in producing well-calibrated outputs. While the issue of calibration in neural networks has been extensively studied (Guo et al., 2017; Nixon et al., 2019; Wang, 2024), it is still rarely addressed explicitly in the context of weather forecasting applications. For this reason, a specific study was carried out on the choice of the loss function to ensure output calibration and enable the generation of physically interpretable lightning forecast maps.

This paper is organized as follow. Input data are described in Sect. 2, and are then fed as spatio-temporal sequences to an adapted version of ED-DRAP, detailed in Sect. 3, to generate probabilistic risk maps of lightning activity. The outputs are successfully calibrated and are then analysed in Sect. 4, followed by a discussion in Sect. 5.

2 DATA

To train the prediction model, relevant data must be collected to extract features related to lightning and thunderstorms. These data are collected from satellite's sensors and outputs of NWP models over a specific area.

2.1 Studied area

Data is collected from different meteorological sources over a precise area to restrict the study. The selected area detailed in Figure 1 is named Continental United States (CONUS), and it was chosen due to the availability of sensors with high spatial

and temporal resolution. The area is also near the Intertropical Convergence Zone (ZCIT), which has a well-known hot and humid climate that favors the occurrence of thunderstorms (Bouchard et al., 2022; Hobbs, 1987; Virts et al., 2013).
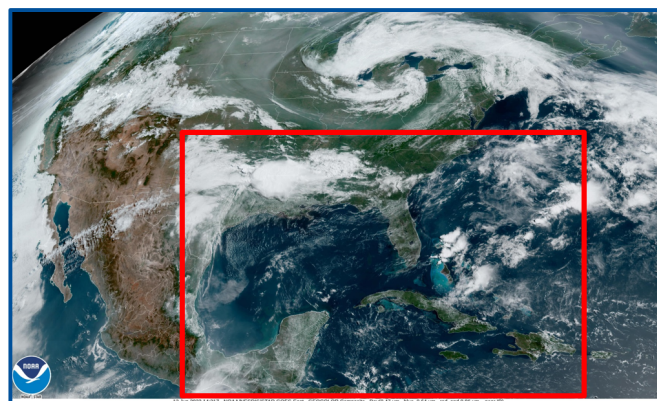
Figure 1. The blue rectangle represents the geographical studied area CONUS, and the red rectangle represents the final smaller chosen area centered over the Gulf of Mexico and Florida. https://www.star.nesdis.noaa.gov/GOES/conus.php?sat=G16

As CONUS area is very large, a smaller area delimited by the red rectangle centered over the Gulf of Mexico and Florida has been chosen to reduce the study's dimensions and the training computation cost. The exact coordinates of the studied area are [15 °N ; 40 °N] degrees for latitude and [100 °W ; 65 °W] degrees for longitude.

2.2 Satellite data

Satellite data are relevant in the context of detecting and predicting thunderstorms. The selected data come from the Geostationary Operational Environmental Satellite (GOES-R/GOES-16), which belongs to the National Oceanic and Atmospheric Administration (NOAA) and is equipped onboard with six sensors. In the case of this study, the focus is on two Earth-pointing sensors: the Advanced Baseline Imager (ABI) (Schmit et al., 2017) and the Geostationary Lightning Mapper (GLM) (Goodman et al., 2013).

The ABI sensor is a multi-channel passive imaging radiometer with a spatial resolution of 0.5 km in the visible spectrum and 2 km in the infrared. It acquires data from 16 wavelength bands and has a temporal resolution of 5 minutes over the CONUS area (Schmit et al., 2017; NOAA, 2020). From the ABI sensor, the 13th band was selected because it is more sensitive to cloud classification. This band is located in the infrared at $10.3\,\mu$m and provides radiance data. The product selected from this sensor is the Brightness Temperature (BT), which has been derived from the radiance using Planck's law. Indeed, the BT is correlated with the presence of thunderstorms because a low BT value corresponds to a high cloud top altitude, which often indicates the presence of a cumulonimbus cloud.

Regarding the GLM, it is a space-based camera that observes electrical activity in the atmosphere with a nadir spatial resolution of 8 km and a lightning detection rate of 70-90 %. This sensor operates day and night, but performs better at night due to the contrast effect. It provides three different products every 20 seconds. The first product is the events, which correspond

to sensor pixels whose luminosity value has exceeded a threshold, indicating an electrical activity. The second product is the groups, which represent a cluster of adjacent events occurring within the same integration time, illustrating lightning and its spatial extent. The last product is the flashes that correspond to a cluster of groups occurring within a 16.5 km area and within a 330 ms duration, from which the centroid is determined. For all products, the type of lightning (cloud-to-ground or intra-cloud) cannot be determined, so all the electrical activity is taken into account (Goodman et al., 2012, 2013). From the GLM sensor, the groups were chosen because they provide larger spatial information about lightning, which is beneficial for predicting the lightning risk probability areas. As the groups were collected and transferred every 20 seconds, they were aggregated over 5-minute intervals around the ABI's acquisition time to match the temporal resolution of the ABI's data.

Both datasets were undersampled using a common mesh of $0.08° \times 0.08°$ to achieve the same spatial resolution of approximately 8.8 km depending on the latitude and longitude coordinates. In Figure 2a representing the BT, darker pixels correspond to lower brightness temperature values, indicating higher cloud top altitudes. Additionally, in Figure 2b representing the groups, white pixels correspond to lightning activity while black pixels indicate the absence of lightning, referred to here as the background. These pictures highlight that the groups appear to be strongly spatially correlated with low BT values. If the BT is low and lightning occurs within a cloud, the cloud is likely a cumulonimbus, and this is the information that should be taken into account by a model.

(a) (b)
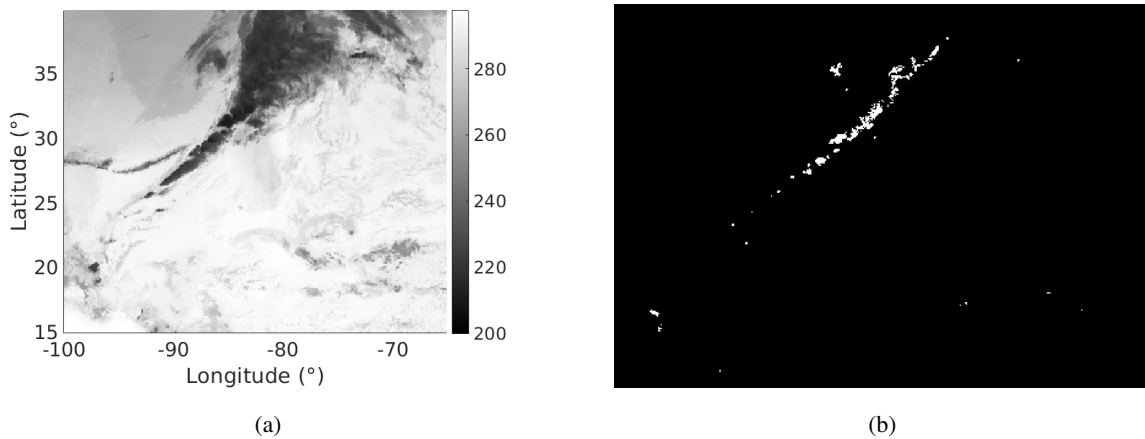
Figure 2. BT image from ABI with grayscale colormap in Kelvins and darker pixels corresponding to higher BT (a). Groups product from GLM where white pixels correspond to lightning and black ones to background (b). These data are acquired on 01/13/23 at 00:06 UTC from GOES-R ABI and GLM sensors.

2.3 NWP Data

In addition to satellite data, NWP output data were used to enhance the information on thunderstorms. Indeed, also using these data in addition to satellite ones can help to have better forecasts (Geng et al., 2021; Leinonen et al., 2022). The Global Forecast System (GFS) (White et al., 2019) developed by the National Centers for Environmental Prediction (NCEP) was

130 chosen because it provides global predictions for many different meteorological parameters with a spatial resolution of 0.25°
× 0.25° and a temporal resolution of 3 hours in the archives. NWP models rely on the numerical resolution of fluid mechanics
equations, incorporating data assimilation and physical parametrizations to forecast many meteorological parameters.

In this study, two meteorological parameters related to lightning activity have been selected as inputs for the network.
The Lifted Index (LI) is a variable that indicates the presence of atmospheric instability, which is a key condition for the

135 development of cumulonimbus clouds (Haklander and Van Delden, 2003; Galway, 1956). A negative LI value indicates that
the air parcel is warmer than its surroundings and can continue to rise, demonstrating the unstable nature of the atmosphere and
the presence of convection, which allows cumulonimbus clouds to form. In this study, we defined bestLI value as the lowest
LI computed for different altitude levels. In Figure 3a, representing a map of bestLI value, darker pixels correspond to lower
best LI values, indicating areas with a very unstable atmosphere.

140 The second selected parameter is derived from the Relative Humidity (RH) (Malardel, 2009). Here, the maximum RH value
across all altitude levels (referred to as maxRH) is chosen to enable the network to identify areas with the highest concentration
of water vapor, indicating the presence of clouds in the atmosphere (Price, 2000). The maxRH is illustrated in Figure 3b, where
whiter pixels represent areas with the highest maxRH, indicating locations where clouds are very thick. Finally, the areas with
the lowest bestLI and the highest maxRH are strongly correlated with the location of lightning in Figure 2b.

(a)   (b)
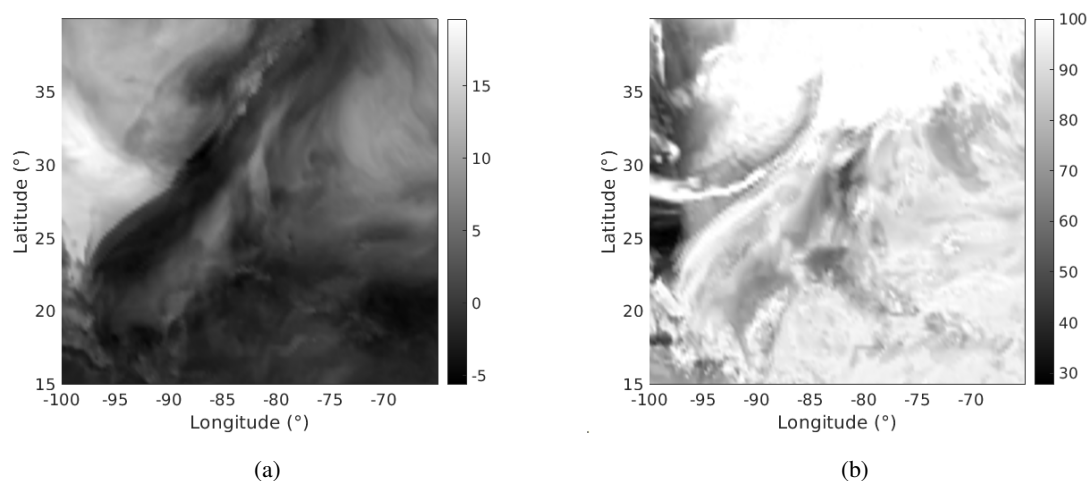
Figure 3. Map of bestLI (in Kelvin) with arker pixels corresponding to lower values of LI so higher chances of convection (a) and map of
maxRH (in %) with darker pixels corresponding to higher maxRH so to the presence of clouds (b). These data are derived from the 00:00
UTC forecast of the 01/12/2023 18:00 UTC GFS run.

145 As the GFS model provides predictions at an average resolution of 25 km, oversampling was performed using Lanczos inter-
polation (Duchon, 1979) to obtain maps of the same studied area with the same spatial resolution as the brightness temperature
and groups maps. Lanczos interpolation is a method for resampling signals such as images, preserving details and reducing
artifacts, although it is more complex than simple bilinear interpolation.

In addition, GFS provides outputs every 3 hours, which is less frequent than the 5 minutes resolution of the rest of the dataset.

150 To compensate, the same GFS data were reused across multiple 5-minute timesteps, introducing some redundancy. Specifically, the following configuration was adopted: 00:00 UTC forecasts were applied from 00:00 UTC to 01:30 UTC, 03:00 UTC forecasts from 01:30 UTC to 04:30 UTC, and 06:00 UTC forecasts from 04:30 UTC to 05:00 UTC.

2.4 Dataset creation & characteristics

Using the four datatypes (BT, groups, bestLI, and maxRH) defined in Subsect. 2.2 and 2.3, a dataset has been created using sev-
155 eral days in Winter over January, February, and December. The data was collected only in the morning, between 00:00 UTC and 05:00 UTC, for the years between 2020 and 2023. This period was selected to restrict the study and to observe thunderstorms at a specific time of the year. In total, data was collected for 154 days, with an average of 50 % of days with thunderstorms and lightning and 50 % of days without, to allow the neural network to adapt to every situation, even when there is no lightning activity, thus avoiding bias.

160 Regarding the satellite data, with images available every 5 minutes between 00:00 UTC and 05:00 UTC, there is a total of 30 images per day for each data type, amounting to 60 images per day, which corresponds to a total of 18,480 images. For the GFS output data, with outputs every 3 hours, we have 3 images per day for each data type, totalling 924 images. All of the images are about 313×438 pixels and during the training and testing phase, images are centered and cropped to reach the shape of 256×256 pixels corresponding to an area between [17.3 °N ; 37.7 °N] degrees in latitude and [93 °W ; 72 °W] degrees
165 in longitude. The database was randomly split once by day into 70 % for training and 30 % for testing.

3 MODEL

This section outlines the architecture of the proposed model. Given that all input data has been converted into images and the objective is to perform semantic segmentation to predict the location of electrical activity, Convolutional Neural Networks (CNN) have been selected for this task (LeCun et al., 1998; Long et al., 2015).

170 ### 3.1 Sequencial input

The objective of this study is to predict maps indicating areas with the probability of having lightning for various timesteps. To achieve this, we deploy a semantic segmentation method, which involves generating an image or mask where each pixel obtains a confidence score of belonging to a class. By applying different thresholds over the confidence scores, pixels are classified in several bins corresponding to the probability of having electrical activity on the forecasted map.

175 To accomplish this task, a sequence of images was selected as the input for the neural network. To determine the optimal number of timesteps to consider, a comparative study was conducted to evaluate the performance of models using 2, 4, 6, and 8 timesteps as inputs for predicting lightning occurrences at intervals of 5, 10, 15, ⋯ up to 60 minutes. The study revealed that the best performance was achieved using 6 timesteps separated by 5 minutes in average, corresponding to 30 minutes for each

prediction horizon. Using 6 timesteps as input instead of 4 showed an increase of 2 % of the F_1 score (presented in Sect. 4.1) for predictions at 30 minutes.
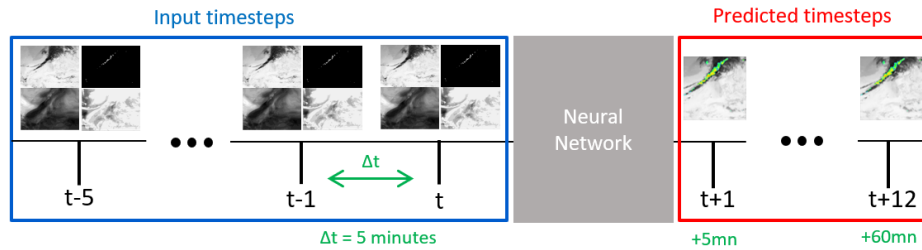
Figure 4. 6 input timesteps using the four datatypes separated by 5 minutes are taken to predict the lightning risk probability maps every 5 minutes ahead up to 1 hour by using a neural network.

180

As illustrated in Figure 4, the input tensor consists of four dimensions $[c, t, H, W]$: the first dimension c represents the channels, which correspond to the number of different data types, totalling four. The second dimension t corresponds to the number of selected timesteps, which is six. The last two dimensions H and W represent the spatial dimensions of the images, set at 256×256 pixels, as the network is trained using tiles from the input images. Using these spatio-temporal sequences, the

185 goal is to predict risk probability maps to have lightning at intervals of 5 minutes, up to one hour.

3.2 ED-DRAP

The network used in this study is named ED-DRAP (Che et al., 2022). It was first introduced for a precipitation prediction task and employs an encoder-decoder architecture with both spatial and sequential attention mechanisms. Several studies have already shown that attention mechanisms help computer vision tasks and particularly time series prediction (Guo et al., 2022;

190 Archambault et al., 2024; Vaswani et al., 2017).

For this study, the original ED-DRAP architecture was adapted to better fit our data and methodology. A schematic representation of the architecture is given in Figure 5. In particular, skip connections were removed after initial experiments suggested they had no positive impact on performance with our settings. In the encoder part, a first convolution is used and is then followed by encoding blocks containing 3-D convolutions, batch normalization layers, and ReLU layers. In the decoder part, some

195 attention modules have been added in addition to upconvolution blocks. These attention blocks are composed of a superposition of 3-D sequence attention modules and 3-D spatial attention modules and are well-described in (Che et al., 2022). Sequence attention modules called SEA are composed of an average pooling layer, followed by two convolutions and a sigmoid activation function. Then, the spatial attention module, called SPA, is composed of a 3-D convolution followed by a sigmoid layer. In both blocks, an attention mask is computed and then applied to the extracted features using a Hadamard product. The SEA

200 module allows the network to capture the most important time-related features, and the SPA module helps to capture the most important spatial features. Then, they are used in a residual block called SSAB, which is composed of two 3-D convolutions

8
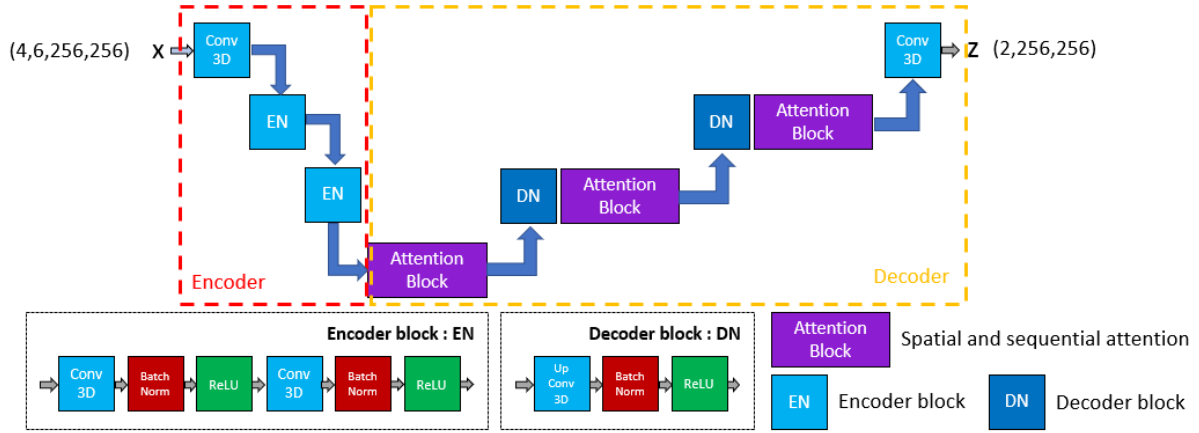
Figure 5. Adapted ED-DRAP architecture schematized: the encoder part is represented in red, the decoder one in yellow, and the block descriptions below. Input tensor has a shape of $[c, t, H, W] = [4, 6, 256, 256]$.

followed by the SEA module and the SPA one. In addition, the attention block defined in Figure 5 is also a residual block which uses 2 SSAB blocks and then a 3-D convolution at the end.

The model returns a tensor of shape $(2, 256, 256)$. It corresponds to a map of lightning confidence scores for both classes,
205 lightning and non-lightning. These scores will be used to generate the probability maps shown in Sect. 4.3.

3.3 Calibration

As mentioned earlier, at the end of the network, a mask with values is generated. This output is then passed through a softmax layer to obtain confidence scores for each pixel, indicating the likelihood of lightning occurrence, with values ranging between 0 and 1. At this stage, machine learning studies typically stop and use these values as probabilities. However, these confidence
210 scores cannot be interpreted as real probabilities at this step. In reality, the predicted confidence scores need to be compared to the actual frequency of events in the ground truth data to be sure that outputs are calibrated (Guo et al., 2017; Nixon et al., 2019; Wang, 2024).

To achieve this, a reliability diagram is plotted with confidence scores on the x-axis and the ratio of real lightning occurrences on the ground truth to the number of pixels with the corresponding confidence score on the y-axis. This diagram is interpreted
215 using Figure 6a. For example, if all pixels with confidence scores between 0.35 and 0.45 are considered, an average of 40 % of these pixels should correspond to actual lightning strikes when compared to the ground truth. But if it only corresponds to 20 %, it means that the network has overestimated the probability, and if it corresponds to 60 %, it means that the network has underestimated the probability. Each reliability diagram in this article is generated from a single model trained on the full training dataset and evaluated on the entire test dataset. The selected model corresponds to the one whose performance is
220 closest to the average over the five training runs.

(a)   (b)
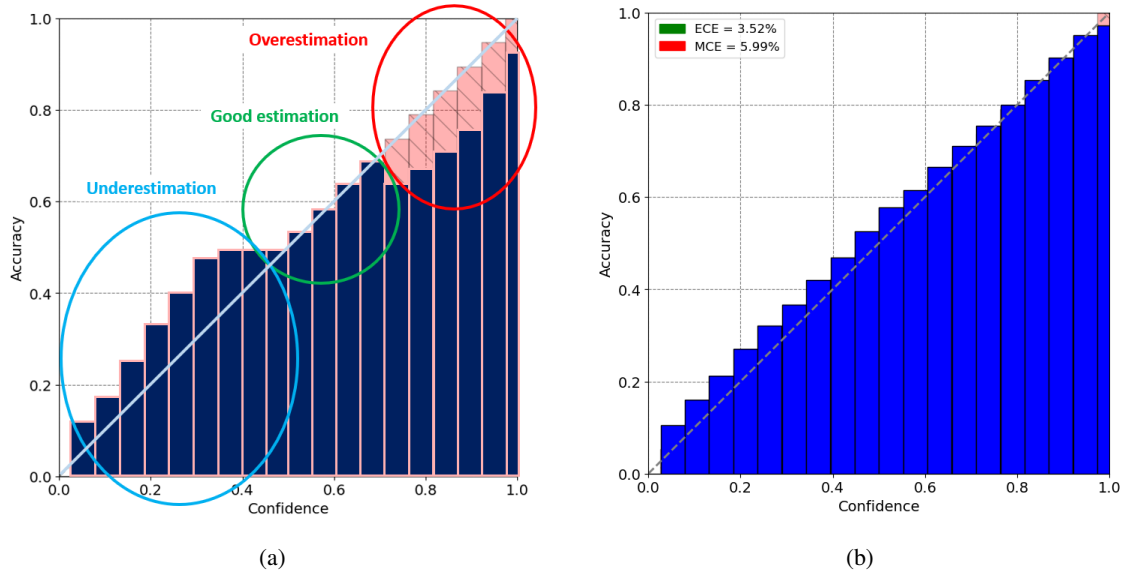
Figure 6. (a) Example of reliability diagram where the light blue circle corresponds to the underestimation area, the green circle corresponds to the correct estimation area, and the red circle corresponds to the overestimation area. The perfect calibration is represented by the diagonal line. (b) Reliability diagram plotted using ED-DRAP results for 5-minute predictions averaged on the entire test set. The x-axis corresponds to confidence scores and the y-axis to accuracy (the number of real lightning over the total number of pixels classified in the bin).

Figure 6b shows a reliability diagram for 5-minute predictions using ED-DRAP, and it indicates that the network is well-calibrated because for each bin of confidence scores, the number of actual lightning corresponds to the predicted values. These results have been obtained by selecting the right input data, an adapted architecture, and by finding the best α coefficient in the total loss function 4, as described in Sect. 4.2.

225   In addition to the reliability diagram, calibration metrics such as the Expected Calibration Error (ECE) and the Maximum Calibration Error (MCE), defined in Eq. 1, are used to quantitatively assess how well the forecasts are calibrated.

$$ECE = \sum_{m=1}^{M} |acc(B_m) - conf(B_m)| \qquad MCE = max_{m \in 1,...,M} |acc(B_m) - conf(B_m)| \qquad (1)$$

with M the total number of bins, B_m the mth bin, $acc(B_m)$ the prediction accuracy for B_m that corresponds to the total number of correct predictions over the total number of elements in B_m and $conf(B_m)$ the confidence score for B_m.

230   These quantities are defined in Eq. 2.

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i) \qquad conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \qquad (2)$$

with $\hat{y}_i$ the prediction, y_i the ground truth and $\hat{p}_i$ the confidence for the ith example.

10

ECE measures how closely the probabilistic predictions match the actual frequencies of events on average, while MCE indicates the highest error found across all the bins. They both need to be as low as possible for a well-calibrated network's output. Calibration's results are discussed in Sect. 5.2.

3.4 Imbalanced problem and training

The choice of the loss function is critical for every machine learning project (Ciampiconi et al., 2021). As a matter of fact, this function will determine how well the model learns the values of its parameters and how well it can fit the data to make correct predictions.

Here, one of the biggest difficulties is that lightning pixels are less present compared to the background. This issue frequently happens and has been addressed as the imbalanced problem (He and Garcia, 2009; Wang et al., 2016). In our dataset, they represent only 1 % on average in the images of groups. Given this, using only the Cross-Entropy function as the loss function will lead to predicting only the background, as it is the majority class. This function evaluates the classification error of every individual pixel independently of their class, so if a class is overrepresented, the focus will be put on minimizing the error for this class. A solution implemented in this study is to add another term to the training loss, which is the diceloss function (Sudre et al., 2017). It can be described as follows in Eq. 3:

$$Dice = 1 - \frac{1}{2}\left(\frac{2\sum(y_0 z_0) + \epsilon}{\sum y_0 + \sum z_0 y_0 + \epsilon} + \frac{2\sum(y_1 z_1) + \epsilon}{\sum y_1 + \sum z_1 y_1 + \epsilon}\right) \tag{3}$$

with $\epsilon = 10^{-5}$ to avoid 0-division, y_0 and y_1 the ground truths for lightning and background classes, and z_0 and z_1, the predictions for the same classes. This function calculates the Intersection over Union (IoU) for both classes, then averages them and returns the opposite. It helps the model focus on detecting more lightning because, unlike Cross-Entropy, it evaluates the percentage of each class that is correctly predicted, regardless of the class proportion in the image. Finally, the total loss function can be written as in Eq. 4:

$$loss = CrossEntropy(y, z) + \alpha Dice(y, z) \tag{4}$$

with z the prediction made by the model and y the ground truth, which is the image of groups taken at the forecasted horizon. This method can be considered self-supervised because the labels are derived from one of the input data types used to train the network, and the predictions correspond to this input at a future time step. Here, the α coefficient has been chosen to be equal to 0.001 following a study detailed in Sect. 4.2 showing that this coefficient value gave the best calibration scores. This choice allows the model to best balance between precision and detection and to provide well-calibrated outputs.

To address the class imbalance problem in our dataset, we also tested the Focal Loss (FL) function (Lin et al., 2015) to train the network. Specifically, it aids in predicting the minority class by imposing a higher penalty on errors related to this class, unlike the cross-entropy function, which treats all classes and pixels with equal importance. FL writes as in Eq. 5:

$$FocalLoss = -(1 - p_t)^\gamma log(p_t) \tag{5}$$

with γ being the coefficient that enables the network to focus on hard or underrepresented examples, and p_t representing the model's estimated probability of belonging to the lightning class. If $\gamma > 0$, the emphasis is placed on detecting hard examples.

265 We tested several values of γ such as $1, 2, 3$, and 4. Score results using the FL have been compared to those using the total loss defined in Eq. 4. In every configuration, performance was similar or lower than when using our loss function, and the model's output was not calibrated, showing an average of 30% of ECE and 100% of MCE, which is significantly worse than our results.

4 RESULTS

4.1 Metrics

270

To evaluate the performance of our model on the task of predicting lightning probabilistic areas for different forecast horizons, various metrics have been employed. These metrics are based on quantities such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), which are the components of the confusion matrix. Here, TP corresponds to well-identified lightnings, TN corresponds to well-identified background, FP corresponds to pixels predicted as lightning instead

275 of background so to false alarms, and FN corresponds to missed lightnings. Using these four quantities, some metrics can be computed such as Probability Of Detection (POD) or Recall, Precision, and F_1 score as described in Eq. 6.

$$recall = \frac{TP}{TP + FN} \qquad precision = \frac{TP}{TP + FP} \qquad F_1 = 2\frac{recall * precision}{recall + precision} \qquad (6)$$

Here, recall corresponds to the number of well-identified lightning over the total number of real lightning, precision corresponds to the number of well-identified lightning over the total number of predicted lightning and the F_1 score corresponds

280 to the harmonic mean between recall and precision. Precision, recall, and F_1 score should be as high as possible for the network to achieve good prediction performance and are computed here for the positive class, so lightning. These metrics will be computed to compare the performances of different models in Sect. 5.1.

4.2 Choice of the α coefficient

To choose the right α coefficient in front of the dice loss function, a study of its impact on calibration has been made using

285 ED-DRAP to make a forecast at 30 minutes, and results can be seen on Figure 7a and 7b.

ECE and MCE values are minimal when the chosen α coefficient is 0.001. Here, the test has only been plotted for a forecasted horizon of 30 minutes, but the results are the same at different forecasting horizons. Consequently, $\alpha = 0.001$ is always the best choice to have a calibrated output, so this value is kept for the entire study.

To be sure that the well-calibrated output of the network ED-DRAP is a specificity of the combination of the architecture,

290 the methodology, and the adapted loss function, the α coefficient in the loss function for the training of ConvLSTM network has also been modified. The tests are also plotted for a 30-minute forecast, and results can be observed in Figure 7.

Figure 7 highlights the fact that the best scores are achieved when $\alpha = 0.001$ for both architectures, and that ConvLSTM never reaches ECE and MCE scores lower than that of ED-DRAP. It is a strong clue that a great calibration score may be obtained by combining an optimal α value in the loss function and a suitable architecture.

(a)   (b)
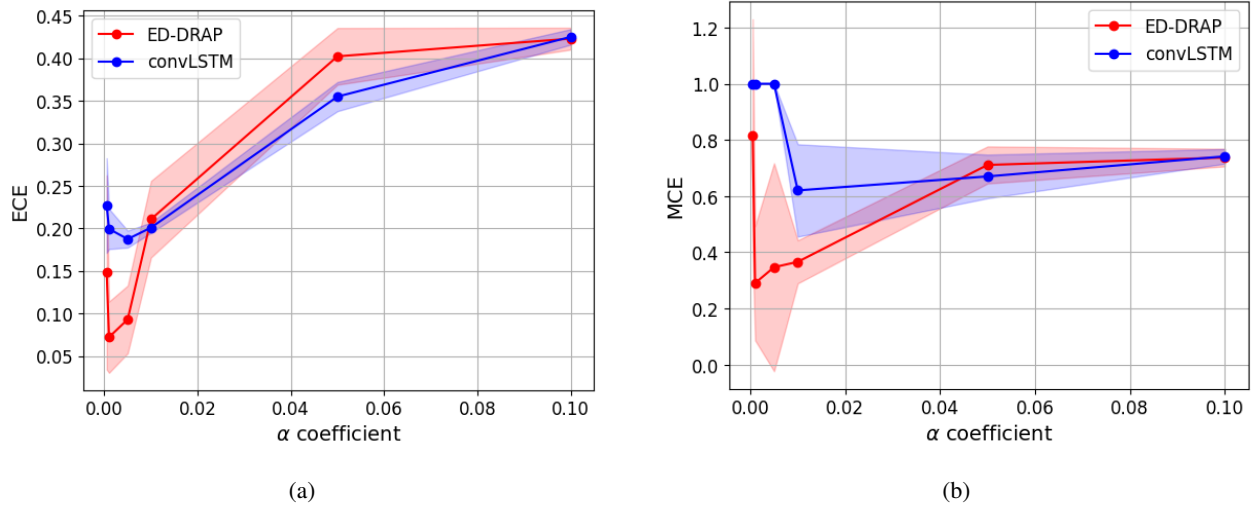
Figure 7. ECE (a) and MCE (b) means (solid lines) with their standard deviations shown as shaded areas, plotted against different α values for 30-minute forecasts using ConvLSTM and ED-DRAP networks.

4.3 Probability maps

Thanks to the well-calibrated outputs, physically interpretable lightning risk probability maps have been plotted. Here, the red color has been chosen for probabilities between 90 and 100 %, down to blue for probabilities around 5 %. Regions without coloration are considered to have a negligible or zero probability of being hazardous. Using this method, risk probability maps have been created, as shown in Figure 8, for different lead times and for a specific day and time in the test database.

These figures illustrate that high-risk probabilities are less frequent when forecasts are made for longer lead times, and they are also less precise, as expected. For instance, in Figure 8d, no risk probabilities exceed 65 %, indicating that the algorithm is less confident than for shorter lead times, such as in Figure 8b. To create probability maps, the output from the inference step is used. This enables the application of different probability thresholds to color-code the various lightning risk levels based on the pixel values in the output mask. Each pixel is assigned to a bin according to its confidence score, which lies within the [0, 1] interval and serves as the basis for classification.

An example of a probability map is plotted in Figure 9 with the lightning ground truth overlaid on it in order to visually see how well the real lightning fits on the prediction risk areas. This map was determined using the ED-DRAP model for predictions 30 minutes ahead on 13/01/2023 at 00:56 UTC. It shows that nearly all lightning strikes are predicted within the colored areas. In this example, only 5 % of lightning strikes are missed, and this result remains consistent across all forecast horizons, as shown in Fig. 10c.

13
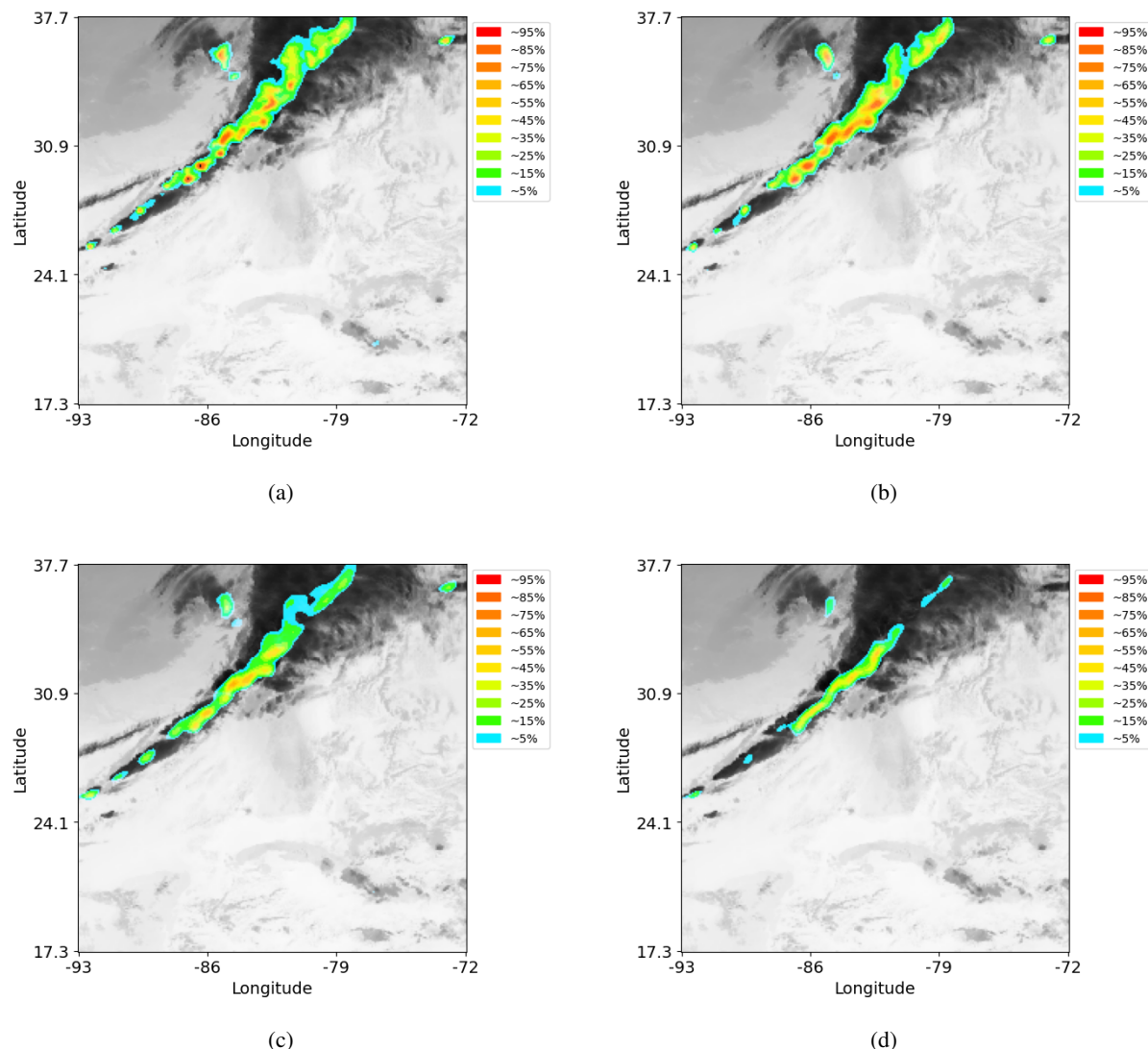
Figure 8. Lightning risk probability maps for various forecast horizons obtained using ED-DRAP and the developed methodology. Forecasts for 5 minutes at 00:31 UTC (a), 20 minutes at 00:46 UTC (b), 45 minutes at 01:11 UTC (c), 60 minutes at 01:26 UTC (d) for the 13/01/2023.

4.4 Threshold modification

A threshold τ is used to determine whether a pixel with a given confidence score should be classified as lightning or as background at the output of the model. Typically, a threshold of 0.5 is applied, meaning that pixels are classified as lightning if they are more likely to be so than background. However, this threshold can be adjusted depending on the application. For instance, lowering the threshold results in more pixels being classified as lightning, which increases recall but decreases precision. In this study, since the primary objective is to detect lightning activity while minimizing false alarms, the threshold
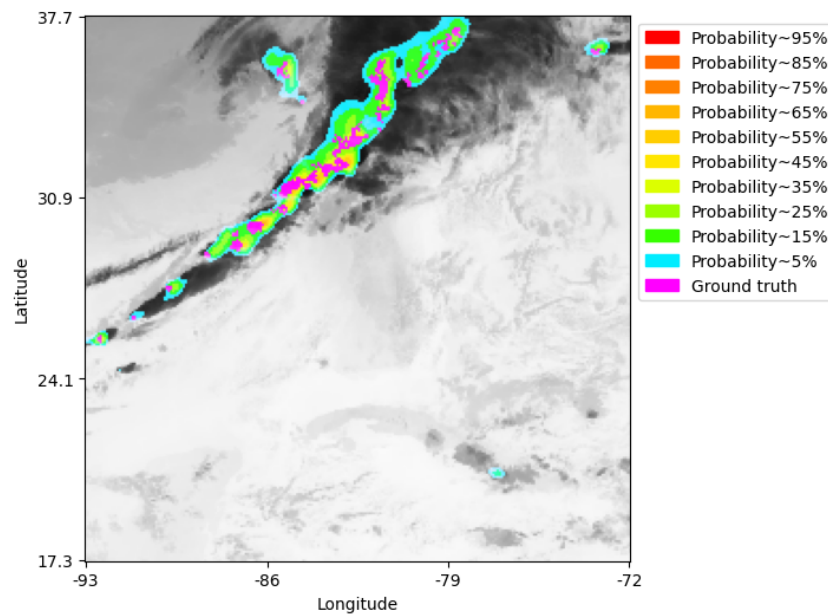
Figure 9. Risk probability map using ED-DRAP for 30-minute forecast horizon with ground truth plotted in magenta on 13/01/2023 at 00:56 UTC.

can be lowered. Nevertheless, the choice of threshold can be left to the user, who may prefer to prioritize detecting lightning activity or reducing false alarms. Figure 10 shows the different metrics obtained across all forecast periods for various threshold values.
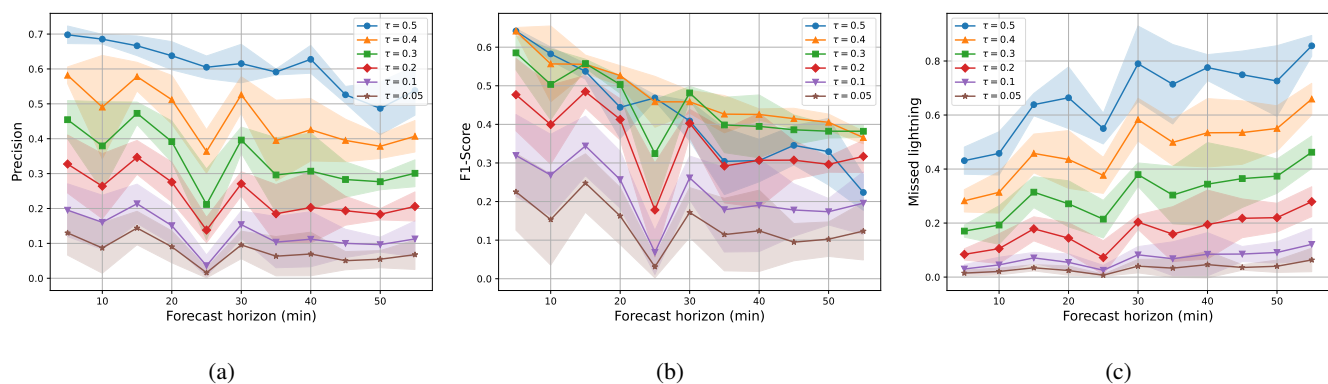
Figure 10. Evaluation metrics means (solid lines) with their standard deviations shown as shaded areas, plotted against several forecast horizons using different probability thresholds τ. On the x-axis, the lead times in minutes, and on the y-axis, the precision (a), the F_1 score (b), and the missed lightning ratio (c).

320   POD, F_1 score, and the percentage of missed lightning have been plotted in Figure 10 for several lead times and thresholds. If a threshold of 0.5 or 0.4 is used, the F_1 score is maximized with a high precision score but a higher percentage of missed lightning. As the threshold decreases, precision and the F_1 score also decrease, but the percentage of missed lightning significantly decreases. Therefore, we selected a threshold of 0.05 to plot blue areas on the risk probability maps to capture the maximum number of lightning strikes. Since lightning is a hazardous and punctual phenomenon, the primary objective is to

325   achieve a high recall while maintaining an acceptable level of precision. A suitable trade-off can be obtained by selecting a threshold of 0.2 or 0.3, which yields a recall of approximately 80 % and a precision of around 35 %. Given that the False Alarm Rate (FAR) is defined as $1 -$ precision, this corresponds to a FAR of about 65 %. Such a value remains acceptable, considering that the model aims to identify large areas at risk to issue warnings and minimize the occurrence of undetected lightning events.

5   Discussion

330   ### 5.1   Scores comparison between models and forecast horizons

To ensure that ED-DRAP truly outperforms other networks, it is necessary to compare the scores obtained by testing various other spatio-temporal predictor networks using different metrics. We selected two well-known networks for comparison with ED-DRAP: ConvLSTM (Shi et al., 2015) and PredRNN (Wang et al., 2022). The ConvLSTM model used here is the vanilla version, with 2 layers of convolutional LSTM and 64 hidden dimensions. The same parameters were used for the

335   classical PredRNN in this study. ED-DRAP model is also compared to these networks, as well as to persistence and to a U-Net model (Ronneberger et al., 2015). Results for U-Net are not plotted here because the network only provides results for a 5-minute forecast period and is not able to forecast to longer horizons. This can be explained by the fact that it is a segmentation network, but it has not been created to catch spatio-temporal dependencies and to make multiple timesteps predictions using it. For 5-minute forecasts, U-Net achieves an F_1 score of 0.44, which remains lower than those obtained by the other

340   spatio-temporal neural networks, as it can be seen in Figure 11.

Here, each network was trained with 100 epochs, the loss function described in Subsect. 3.4, the Adam optimizer, batches of 2, and a learning rate of 0.0001. The computed metrics, described in Subsect. 4.1 for recall, precision, and F_1 score, and in Subsect. 3.3 for ECE and MCE, were calculated using a threshold of 0.5. This means that if a pixel's confidence score is above 0.5, it is forecasted as lightning. We chose to use this threshold first to compare scores with other methods, but this choice is

345   discussed in Subsect. 4.4.

A distinct model is trained for each forecast horizon, allowing predictions to be optimally adapted to each lead time. All metrics were computed over an ensemble of 5 trainings for each forecast horizon and were then averaged to ensure robustness. Train a model on our dataset takes an average of 76 minutes on a NVIDIA RTX A5000. However, the inference phase only takes 10 seconds to generate a map at a chosen lead time on a CPU. The results are presented in Figure 11.

350   First, for each metric, all performances decrease with the forecast horizon, which is commonly observed in meteorological predictions. Then, in Figure 11a, both PredRNN and ConvLSTM do not achieve better scores than persistence, except for ED-DRAP, which yields a higher average recall. This may be due to the fact that PredRNN and ConvLSTM lack the capacity

(a)   (b)   (c)
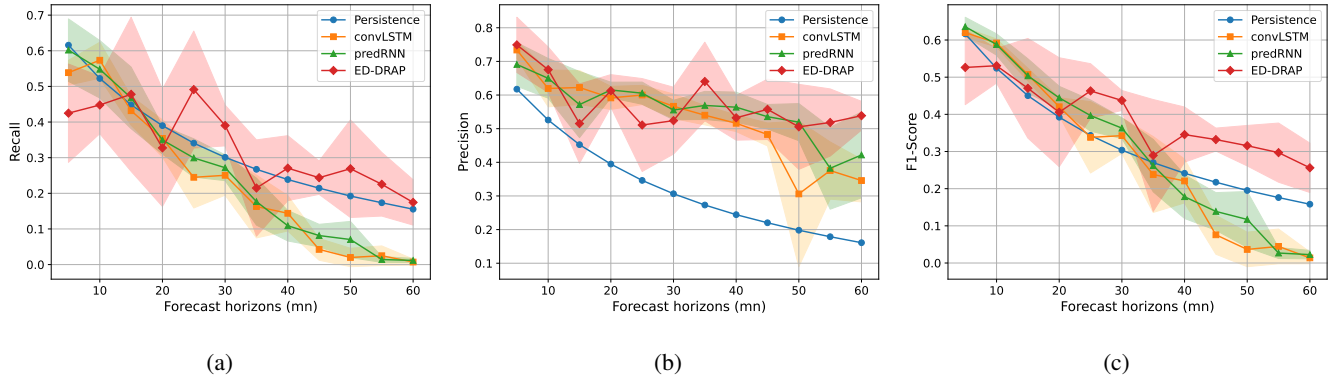
Figure 11. Recall (a), precision (b), and F_1 score (c) means (solid lines) with their standard deviations shown as shaded areas, plotted against several lead times for different tested models. The x-axis represents the lead times (minutes), and y-axis represents the metric value.

to capture the complex spatio-temporal correlations required to accurately predict electric activity, especially in the presence of rare and highly localized events. Additionally, in Figure 11b, persistence precision score is the lowest for each forecast period, and all the other networks give similar results. Finally, in Figure 11c, the F_1 score is better using ED-DRAP, especially for longer forecast periods. This demonstrates that ED-DRAP is better suited to make predictions using our dataset and methodology because it succeeds in achieving great scores, such as a recall of 0.4, a precision of 0.52, and an F_1 score of 0.44 for 30-minute forecasts, which are all better scores than with the other networks.

5.2 Calibration results comparison between models and forecast horizons

Here, calibration results are compared thanks to the ECE and the MCE metrics. First, as shown in Figures 12a and 12b, persistence cannot achieve good calibration results in this case because it merely compares one timestep to another, without involving confidence scores.

Then, the last three networks are compared over different forecast periods. The ECE and MCE are minimal when using ED-DRAP, with an average calibration error of only $10\,\%$ even for a 50-minute prediction. At this forecast period, ConvLSTM and PredRNN have a minimum calibration error of $30\,\%$. Regarding the maximum calibration error, it reaches $100\,\%$ for every network except ED-DRAP.

To understand where the calibration errors occur, reliability diagrams can be plotted. On Figure 13, we compared the reliability diagrams of ConvLSTM, PredRNN, and ED-DRAP for a forecast period of 30 minutes.

ED-DRAP is the only model to achieve great calibration scores and for ConvLSTM and PredRNN, low-risk probabilities are well-calibrated, but these networks are not well-calibrated for high confidence scores when it comes to high-risk probabilities. In fact, if all pixels with a predicted confidence score of 0.9 are considered, $0\,\%$ of them are actual lightning strikes, leading to poor network calibration. ED-DRAP successfully gives confidence scores, which are calibrated when compared to the ground truth. All these results explain why ED-DRAP has been chosen for this electrical activity risk probability forecasting task.

(a)   (b)
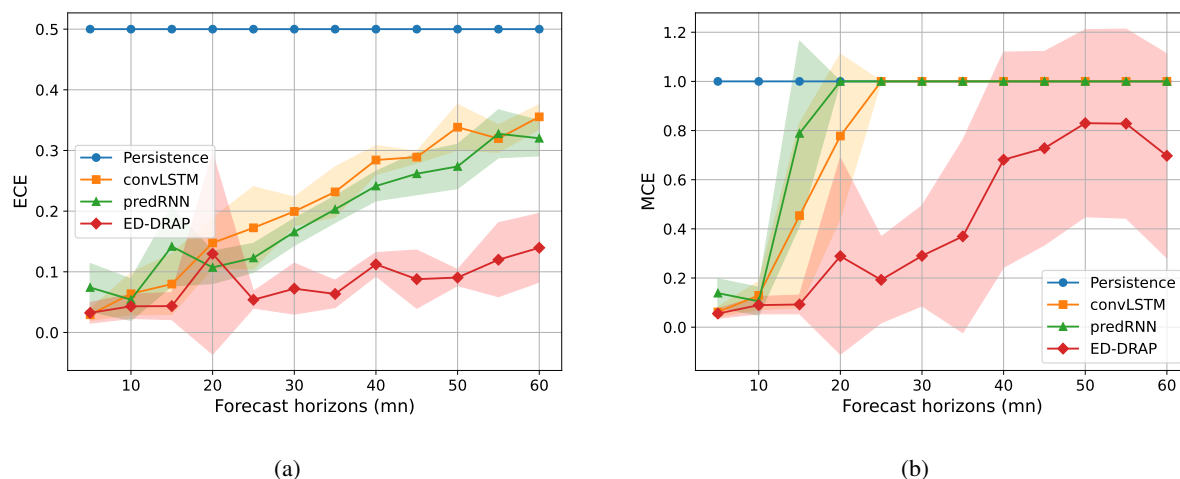
Figure 12. ECE (a) and MCE (b) means (solid lines) with their standard deviations shown as shaded areas, plotted against several lead times for the different tested models. On x-axis are the forecast periods (minutes) and on y-axis are the ECE and the MCE metrics.

(a)   (b)   (c)
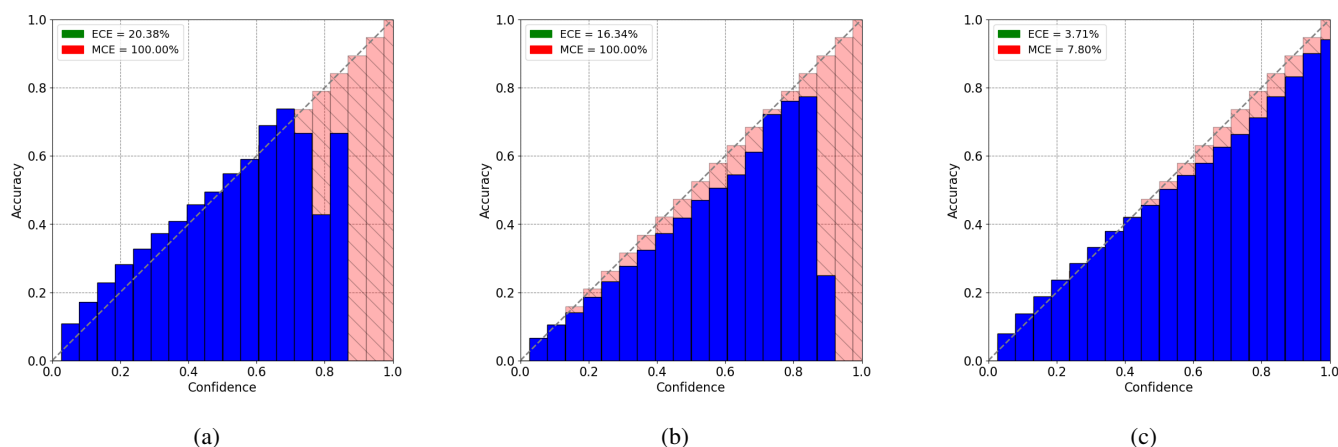
Figure 13. Reliability diagrams for 30-minute predictions using convLSTM (a), predRNN (b), and ED-DRAP (c). The x-axis represents the calibration bins and the y-axis represents accuracy. ECE and MCE scores are printed on the left top of the plot.

In addition, compared to the diagram in Figure 13c, the outputs are slightly less well-calibrated. Specifically, the bins do not
375   perfectly align with the diagonal and are positioned above it, indicating that the network tends to overestimate the presence of lightning. For instance, among the pixels with a confidence score of 0.5, only 43% actually correspond to lightning when compared to the ground truth. While the output for the 30-minute forecast remains reasonably well-calibrated, it is less so than the 5-minute forecast, which is expected as the prediction task becomes more challenging over longer timeframes.

6 Conclusions

380 In the framework of air safety enhanced by the ALBATROS project, the study has focused on developing a lightning risk probability methodology to allow airplanes to avoid dangerous stormy areas. To do that, we have created a database using meteorological parameters, such as satellite data products and outputs from Numerical Weather Prediction (NWP) models, to train neural networks for generating very short-term lightning strike risk forecast maps. The study highlights a methodology that involves using multiple timesteps as input to feed a modified version of a neural network named ED-DRAP, which is an
385 encoder-decoder model utilizing spatial and sequential attention.

The study also emphasizes the importance of selecting an appropriate loss function tailored to the problem, specifically addressing the imbalanced nature of the dataset. The article demonstrates that using the Dice Loss function can be beneficial in such imbalanced scenarios and can aid in calibrating the network's outputs.

Moreover, combining useful meteorological data with an adapted neural network and a suitable loss function can result
390 in well-calibrated network outputs. This allows for the creation of lightning activity probability risk maps using different probability thresholds and colors. These maps can be generated for various forecast periods, ranging from 5 minutes to 1 hour, while remaining well-calibrated. They can be adapted to user preferences based on the desired precision and recall scores.

Future work will focus on several directions. First, we plan to incorporate radar data into the input, as it has been shown to improve prediction performance (Leinonen et al., 2022). Second, since our current methodology requires training separate
395 models for each forecast horizon, we aim to develop a multi-horizon approach by transforming the architecture into an autoregressive model. Finally, given the promising results of diffusion models in meteorological forecasting using AI, we intend to explore their application to our task.

400 *Competing interests.* The authors declare that they have no conflict of interest.

405 **References**

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., Agrawal, S., and Kalchbrenner, N.: Deep Learning for Day Forecasts from Sparse Observations, https://doi.org/10.48550/arXiv.2306.06079, 6 July 2023.

Archambault, T., Filoche, A., Charantonis, A. A., and Béréziat, D.: Pre-training and Fine-tuning Attention Based Encoder Decoder Improves Sea Surface Height Multi-variate Inpainting, VISAPP 2024 - 19th International Conference on Computer Vision Theory and Applications,
410 https://doi.org/10.5220/0012357400003660, 2024.

Betz, H. D., Schmidt, K., Oettinger, W. P., and Montag, B.: Cell-tracking with lightning data from LINET, in: Advances in Geosciences, 9th EGU Plinius Conference on Mediterranean Storms (2007) - 9th Plinius Conference on Mediterranean Storms, Varenna, Italy; 13 September 2007, 55–61, https://doi.org/10.5194/adgeo-17-55-2008, 2008.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Vaughan, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J.
415 K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: A Foundation Model for the Earth System, https://doi.org/10.1038/s41586-025-09005-y, 21 November 2024.

Bosc, M., Chan-Hon-Tong, A., Bouchard, A., and Béréziat, D.: StrikeNet: A Deep Neural Network to predict pixel-sized lightning location, https://doi.org/10.5220/0013110700003912, 2024.

Bouchard, A., Buguet, M., Chan-Hon-Tong, A., Dezert, J., and Lalande, P.: Comparison of different forecasting tools for short-range lightning
420 strike risk assessment, Nat Hazards, 115, 1011–1047, https://doi.org/10.1007/s11069-022-05546-x, 2023.

Bouchard, A., Kutyla, C., Lalande, P., and Flourens, F.: Study of an Atypical In-service Lightning Strike on a Commercial Jetliner, in: ICOLSE 2024, Sao Paulo, Brazil, 2024.

Bouget, V., Béréziat, D., Brajard, J., Charantonis, A., and Filoche, A.: Fusion of Rain Radar Images and Wind Forecasts in a Deep Learning Model Applied to Rain Nowcasting, Remote Sensing, 13, 246, https://doi.org/10.3390/rs13020246, 2021.

425 Bouttier, F. and Marchal, H.: Probabilistic thunderstorm forecasting by blending multiple ensembles, Tellus A: Dynamic Meteorology and Oceanography, 72, https://doi.org/10.1080/16000870.2019.1696142, 2020.

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.

Burrows, W. R., Price, C., and Wilson, L. J.: Warm Season Lightning Probability Prediction for Canada and the Northern United States,
430 Weather and Forecasting, 20, 971–988, https://doi.org/10.1175/WAF895.1, 2005.

Che, H., Niu, D., Zang, Z., Cao, Y., and Chen, X.: ED-DRAP: Encoder–Decoder Deep Residual Attention Prediction Network for Radar Echoes, IEEE Geoscience and Remote Sensing Letters, 19, 1–5, https://doi.org/10.1109/LGRS.2022.3141498, 2022.

Chemartin, L., Lalande, P., Peyrou, B., Chazottes, A., Elias, P. Q., Delalondre, C., Cheron, B. G., and Lago, F.: Direct Effects of Lightning on Aircraft Structure: Analysis of the Thermal, Electrical and Mechanical Constraints, Aerospace Lab, 1–15, 2012.

435 Ciampiconi, L., Elwood, A., Leonardi, M., Mohamed, A., and Rozza, A.: A survey and taxonomy of loss functions in machine learning, https://doi.org/10.48550/arXiv.2301.05579, 18 November 2024.

Cintineo, J. L., Pavolonis, M. J., and Sieglaff, J. M.: ProbSevere LightningCast: A Deep-Learning Model for Satellite-Based Lightning Nowcasting, https://doi.org/10.1175/WAF-D-22-0019.1, 2022.

Collins, W. G., Tissot, P., Collins, W. G., and Tissot, P.: Thunderstorm Predictions Using Artificial Neural Networks, in: Artificial Neural
440 Networks - Models and Applications, IntechOpen, https://doi.org/10.5772/63542, 2016.

Couairon, G., Lessig, C., Charantonis, A., and Monteleoni, C.: ArchesWeather: An efficient AI weather forecasting model at 1.5° resolution, https://doi.org/10.48550/arXiv.2405.14527, 3 July 2024.

Creswick, A.: A deep learning approach for probabilistic forecasts of cumulonimbus clouds from NWP data, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-9783, https://doi.org/10.5194/egusphere-egu25-9783, 2025.

445   Dafis, S., Fierro, A., Giannaros, T. M., Kotroni, V., Lagouvardos, K., and Mansell, E.: Performance Evaluation of an Explicit Lightning Forecasting System, Journal of Geophysical Research: Atmospheres, 123, 5130–5148, https://doi.org/10.1029/2017JD027930, 2018.

Dezert, J., Bouchard, A., and Buguet, M.: Multi-Criteria Information Fusion for Storm Prediction Based on Belief Functions, in: 2021 IEEE 24th International Conference on Information Fusion (FUSION), 2021 IEEE 24th International Conference on Information Fusion (FUSION), 1–8, https://doi.org/10.23919/FUSION49465.2021.9626835, 2021.

450   Dixon, M. and Wiener, G.: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A Radar-based Methodology, https://doi.org/10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2, 1993.

Duchon, C. E.: Lanczos Filtering in One and Two Dimensions, https://doi.org/10.1175/1520-0450(1979)018<1016:LFIOAT>2.0.CO;2, 1979.

EASA document: Weather Information to Pilots Strategy Paper, an outcome of the all weather Operations Project, Air Traffic Management
455   Air Navigation Services Development, p 45, (2018).

FAA AC 00-24C: Advisory circular, thunderstorms, US Department of Transportation, Federal Aviation Administration, 2013.

Galway, J. G.: The Lifted Index as a Predictor of Latent Instability, 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.
460   org/10.1002/qj.4167, 2021.

Goodman, S., Mach, D., Koshak, W., and Blakeslee, R.: GLM Lightning Cluster-Filter Algorithm, 2012.

Goodman, S. J., Blakeslee, R. J., Koshak, W. J., Mach, D., Bailey, J., Buechler, D., Carey, L., Schultz, C., Bateman, M., McCaul, E., and Stano, G.: The GOES-R Geostationary Lightning Mapper (GLM), Atmospheric Research, 125–126, 34–49, https://doi.org/10.1016/j.atmosres.2013.01.006, 2013.

465   Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q.: On Calibration of Modern Neural Networks, https://doi.org/10.48550/arXiv.1706.04599, 3 August 2017.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., and Hu, S.-M.: Attention Mechanisms in Computer Vision: A Survey, Comp. Visual. Med., 8, 331–368, https://doi.org/10.1007/s41095-022-0271-y, 2022.

Haklander, A. J. and Van Delden, A.: Thunderstorm predictors and their forecast skill for the Netherlands, Atmospheric Research, 67–68,
470   273–299, https://doi.org/10.1016/S0169-8095(03)00056-5, 2003.

Handwerker, J.: Cell tracking with TRACE3D—a new algorithm, Atmospheric Research, 61, 15–34, https://doi.org/10.1016/S0169-8095(01)00100-4, 2002.

He, H. and Garcia, E. A.: Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, 21, 1263–1284, https://doi.org/10.1109/TKDE.2008.239, 2009.

475   Hobbs, P. V.: The Gulf Stream rainband, Geophysical Research Letters, 14, 1142–1145, https://doi.org/10.1029/GL014i011p01142, 1987.

Holton, J. R.: An introduction to dynamic meteorology, 4th edition., Elsevier Academic Press, Burlington, MA, 2004.

James, P. M., Reichert, B. K., and Heizenreder, D.: NowCastMIX: Automatic Integrated Warnings for Severe Convection on Nowcasting Time Scales at the German Weather Service, https://doi.org/10.1175/WAF-D-18-0038.1, 2018.

Janny, S., Bénéteau, A., Nadri, M., Digne, J., Thome, N., and Wolf, C.: EAGLE: Large-scale Learning of Turbulent Fluid Dynamics with
480   Mesh Transformers, The Eleventh International Conference on Learning Representations, 2022.

Johnson, J. T., MacKeen, P. L., Witt, A., Mitchell, E. D. W., Stumpf, G. J., Eilts, M. D., and Thomas, K. W.: The Storm Cell Identification and
Tracking Algorithm: An Enhanced WSR-88D Algorithm, https://doi.org/10.1175/1520-0434(1998)013<0263:TSCIAT>2.0.CO;2, 1998.

Kober, K. and Tafferner, A.: Tracking and nowcasting of convective cells using remote sensing data from radar and satellite, Meteorologische
Zeitschrift, 18, 75–84, https://doi.org/10.1127/0941-2948/2009/359, 2009.

485   Korpinen, A., Hieta, L., and Partio, M.: Thunder probability nowcast, EMS Annual Meeting 2024, Barcelona, Spain, 1–6 Sep 2024,
EMS2024-471, https://doi.org/10.5194/ems2024-471, 2024.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W.,
Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: GraphCast: Learning skillful medium-
range global weather forecasting, https://doi.org/10.1126/science.adi2336, 4 August 2023.

490   Laroche, P., Blanchet, P., Delannoy, A., and Issac, F.: Experimental Studies of Lightning Strikes to Aircraft, Aerospace Lab, 1–13, 2012.

Laroche, P., Lalande, P., Parmantier, J.-P., Issac, F., and Chemartin, L.: Foudroiement en aéronautique, Systèmes aéronautiques et spatiaux,
https://doi.org/10.51257/a-v1-trp4001, 2015.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86,
2278–2324, https://doi.org/10.1109/5.726791, 1998.

495   Leinonen, J., Hamann, U., Germann, U., and Mecikalski, J. R.: Nowcasting thunderstorm hazards using machine learning: the impact of data
sources on performance, Natural Hazards and Earth System Sciences, 22, 577–597, https://doi.org/10.5194/nhess-22-577-2022, 2022.

Leinonen, J., Hamann, U., and Germann, U.: Seamless lightning nowcasting with recurrent-convolutional deep learning, Artificial Intelli-
gence for the Earth Systems, 1, e220043, 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
500   Model, Geophysical Research Letters, 50, e2022GL101626, https://doi.org/10.1029/2022GL101626, 2023.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P.: Focal Loss for Dense Object Detection, https://doi.org/10.1109/ICCV.2017.324, 7
February 2018.

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, 1–1, https://doi.org/10.1109/ACCESS.2019.2950328, 2019.

505   Long, J., Shelhamer, E., and Darrell, T.: Fully Convolutional Networks for Semantic Segmentation, https://doi.org/10.1109/TPAMI.2016.
2572683, 8 March 2015.

Lynn, B. H., Yair, Y., Price, C., Kelman, G., and Clark, A. J.: Predicting Cloud-to-Ground and Intracloud Lightning in Weather Forecast
Models, https://doi.org/10.1175/WAF-D-11-00144.1, 2012.

Mäkelä, A., Saltikoff, E., Julkunen, J., Juga, I., Gregow, E., and Niemelä, S.: Cold-Season Thunderstorms in Finland and Their Effect on
510   Aviation Safety, https://doi.org/10.1175/BAMS-D-12-00039.1, 2013.

Malardel Sylvie : Fondamentaux de météorologie : à l'école du temps, 2e édition., Cépaduès-éditions, Toulouse, xiv+710 pp., 2009.

Meyer, V.: Thunderstorm Tracking and Monitoring on the Basis of Three Dimensional Lightning Data and Conventional and Polarimetric
Radar Data, PhDThesis, Ludwig-Maximilians-Universität München, 2010.

Milani, Z., Nichman, L., Matida, E., Fleury, L., Wolde, M., Bruning, E., McFarquhar, G. M., and Kollias, P.: In-flight measurements of
515   lightning locations using an aircraft-mounted lightning mapper, Aerospace Science and Technology, 160, 110038, https://doi.org/10.1016/
j.ast.2025.110038, 2025.

Müller, R., Barleben, A., Haussler, S., and Jerg, M.: A Novel Approach for the Global Detection and Nowcasting of Deep Convection and Thunderstorms, Remote Sensing, 14, 3372, https://doi.org/10.3390/rs14143372, 2022.

Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., and Tran, D.: Measuring Calibration in Deep Learning, Proceedings of
520   the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 38-41, https://doi.org/10.48550/arXiv.1904.01685, 2019.

NOAA: GOES-R Series Product Definition and Users' Guide (PUG) Volume 5 – Level 2+ Products, version 2.1, available at: https://www.goes-r.gov/products/docs/PUG-L2+-vol5.pdf, 2020.

North, J., Stanley, Z., Kleiber, W., Deierling, W., Gilleland, E., and Steiner, M.: A statistical approach to fast nowcasting of lightning potential
525   fields, Advances in Statistical Climatology, Meteorology and Oceanography, 6, 79–90, https://doi.org/10.5194/ascmo-6-79-2020, 2020.

Pan, X., Lu, Y., Zhao, K., Huang, H., Wang, M., and Chen, H.: Improving Nowcasting of Convective Development by Incorporating Polarimetric Radar Variables Into a Deep-Learning Model, Geophysical Research Letters, 48, https://doi.org/10.1029/2021GL095302, 2021.

Pédeboy, S., Barneoud, P., and Berthet, C.: First results on severe storms prediction based on the French national Lightning Locating System, 2016.

Plumer, J. A. and Robb, J. D.: The Direct Effects of Lightning on Aircraft, IEEE Transactions on Electromagnetic Compatibility, EMC-24,
530   158–172, https://doi.org/10.1109/TEMC.1982.304010, 1982.

Price, C.: Evidence for a link between global lightning activity and upper tropospheric water vapour, Nature, 406, 290–293, https://doi.org/10.1038/35018543, 2000.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R.,
535   and Willson, M.: GenCast: Diffusion-based ensemble forecasting for medium-range weather, https://doi.org/10.48550/arXiv.2312.15796, 1 May 2024.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.1007/978-3-319-24574-4_28, 18 May 2015.

Schmit, T. J., Griffith, P., Gunshor, M. M., Daniels, J. M., Goodman, S. J., and Lebair, W. J.: A Closer Look at the ABI on the GOES-R
540   Series, https://doi.org/10.1175/BAMS-D-15-00230.1, 2017.

Sénési, S. and Thepenier, R.-M.: Indices d'instabilité et occurrence d'orage : le cas de l'Île-de-France, La Météorologie, 18–33, https://doi.org/10.4267/2042/47022, 1997.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., and Woo, W.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, https://dl.acm.org/doi/10.5555/2969239.2969329, 2015.

545   Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations, vol. 10553, 240–248, https://doi.org/10.1007/978-3-319-67558-9-28, 2017.

Uman, M. A.: The Lightning Discharge, 1st ed., vol. 39. Orlando: Academic Press, 377 pages, 1987.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, https://doi.org/10.48550/arXiv.1706.03762, 2017.

550   Virts, K. S., Wallace, J. M., Hutchins, M. L., and Holzworth, R. H.: Highlights of a New Ground-Based, Hourly Global Lightning Climatology, https://doi.org/10.1175/BAMS-D-12-00082.1, 2013.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., and Kennedy, P. J.: Training deep neural networks on imbalanced data sets, in: 2016 International Joint Conference on Neural Networks (IJCNN), 2016 International Joint Conference on Neural Networks (IJCNN), 4368–4374, https://doi.org/10.1109/IJCNN.2016.7727770, 2016.

555   Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P. S., and Long, M.: PredRNN: A Recurrent Neural Network for Spatiotemporal Predictive Learning, https://doi.org/10.1109/TPAMI.2022.3165153, 9 April 2022.

Wang, C.: Calibration in Deep Learning: A Survey of the State-of-the-Art, https://doi.org/10.48550/arXiv.2308.01222, 10 May 2024.

Wilson, J. W., Crook, N. A., Mueller, C. K., Sun, J., and Dixon, M.: Nowcasting Thunderstorms: A Status Report, https://doi.org/10.1175/1520-0477(1998)079<2079:NTASR>2.0.CO;2, 1998.

560   White, G. Yang, F. Tallapragada, V.: The Development and Success of NCEP's Global Forecast System, NCEP Office Note, 2019, AMS, 99th Annual Meeting, 6-10 January 2019. See https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs

Yoshikawa, E. and Ushio, T.: Tactical Decision-Making Support Information for Aircraft Lightning Avoidance —Feasibility Study in Area of Winter Lightning, Bulletin of the American Meteorological Society, 100, https://doi.org/10.1175/BAMS-D-18-0078.1, 2019.

Yu, M., Huang, Q., and Li, Z.: Deep learning for spatiotemporal forecasting in Earth system science: a review, International Journal of Digital
565   Earth, 17, 2391952, https://doi.org/10.1080/17538947.2024.2391952, 2024.

Zhou, K., Zheng, Y., Dong, W., and Wang, T.: A Deep Learning Network for Cloud-to-Ground Lightning Nowcasting with Multisource Data, Journal of Atmospheric and Oceanic Technology, 37, https://doi.org/10.1175/JTECH-D-19-0146.1, 2020.

Zinner, T., Mannstein, H., and Tafferner, A.: Cb-TRAM: Tracking and monitoring severe convection from onset over rapid development to mature phase using multi-channel Meteosat-8 SEVIRI data, Meteorol Atmos Phys, 101, 191–210, https://doi.org/10.1007/
570   s00703-008-0290-y, 2008.