Received: 15 March 2023 | Revised: 26 April 2024 | Accepted: 3 May 2024

DOI: 10.1002/qj.4777

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

A machine-learning approach to thunderstorm forecasting through post-processing of simulation data

Kianusch Vahid Yousefnia | **Tobias Bölle** | **Isabella Zöbisch** | **Thomas Gerz**

Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

Correspondence
Kianusch Vahid Yousefnia, DLR Oberpfaffenhofen, Institut für Physik der Atmosphäre, Münchner Str. 20, D-82234 Wessling, Germany.
Email: kianusch.vahidyousefnia@dlr.de

Abstract

Thunderstorms pose a major hazard to society and the economy, which calls for reliable thunderstorm forecasts. In this work, we introduce SALAMA, a feedforward neural network model for identifying thunderstorm occurrence in numerical weather prediction (NWP) data. The model is trained on convection-resolving ensemble forecasts over central Europe and lightning observations. Given only a set of pixel-wise input parameters that are extracted from NWP data and related to thunderstorm development, SALAMA infers the probability of thunderstorm occurrence in a reliably calibrated manner. For lead times up to 11 h, we find a forecast skill superior to classification based only on NWP reflectivity. Varying the spatiotemporal criteria by which we associate lightning observations with NWP data, we show that the time-scale for skillful thunderstorm predictions increases linearly with the spatial scale of the forecast.

KEYWORDS

convection, ensembles, forecasting (methods), mesoscale, numerical methods and NWP, severe weather, thunderstorms/lightning/atmospheric electricity

1 | INTRODUCTION

Though thunderstorms undoubtedly constitute inspiring natural spectacles that move any human being to a certain extent, their impact in the form of lightning, strong winds, and heavy precipitation (including hail) is hazardous to society and the economy. Besides the small but real chance of being struck by lightning (Holle, 2016), thunderstorms pose a threat to crops and livestock (Holle, 2014) as well, and they are known to trigger wild fires (Veraverbeke *et al.*, 2017). In addition, they constitute a major safety concern for aviation (Borsky & Unterberger, 2019; Gerz *et al.*, 2012). Furthermore, thunderstorms and lightning damage electrical infrastructure such as wind turbines (Yasuda *et al.*, 2012), which jeopardizes the transition to sustainable energy production. Finally, since the number of severe thunderstorms is expected to increase due to climate change (Diffenbaugh *et al.*, 2013; Rädler *et al.*, 2019), accurate thunderstorm forecasts become ever more relevant.

Thunderstorm forecasts with lead times of more than 1 h usually rely on numerical weather prediction (NWP). This method consists of simulating the future atmospheric state by numerically solving equations derived from the laws of physics. The accuracy of NWP has improved with the advent of high-performance computing, the increased availability of observational data through satellite imagery, and advances in data assimilation (Bauer *et al.*, 2015;

Yano *et al.*, 2018). In order to use NWP data for thunderstorm predictions, one needs to know how thunderstorms manifest themselves in terms of the NWP output fields. In a post-processing step, this knowledge is then used to identify signs of thunderstorm occurrence in simulation data.

Various ideas for identifying signs of thunderstorm occurrence have been put forward in recent years. For instance, post-processing of NWP data has been blended with nowcasting methods (Hwang *et al.*, 2015; Kober *et al.*, 2012). Empirical knowledge on convective activity has been translated into expert systems using fuzzy logic (Li *et al.*, 2021; Lin *et al.*, 2012). The fuzzy logic technique allows the construction of decision rules for thunderstorm occurrence based on domain knowledge. Lately, machine-learning (ML) methods based on artificial neural networks have gained popularity. These methods generalize the fuzzy logic approach in the sense that decision rules are constructed by solving a data-driven optimization problem. Previous studies include neural networks with relatively few neurons (Jardines *et al.*, 2021; Kamangir *et al.*, 2020; Sobash *et al.*, 2020; Ukkonen & Mäkelä, 2019), as well as deep neural networks with convolutional layers and millions of trainable parameters (Geng *et al.*, 2021; Zhou *et al.*, 2022). Findings suggest that neural network models are more skillful at predicting thunderstorm occurrence than comparable ML approaches like random forests (Herman & Schumacher, 2018; Ukkonen & Mäkelä, 2019). In order to learn predicting thunderstorm occurrence, supervised ML methods require a ground truth of thunderstorm activity. It may be provided by satellite imagery (Jardines *et al.*, 2021; Zhou *et al.*, 2022), radar data (Burke *et al.*, 2020; Gagne *et al.*, 2017; Leinonen *et al.*, 2022), storm reports (Loken *et al.*, 2020; Sobash *et al.*, 2020), and lightning (Geng *et al.*, 2021; Ukkonen & Mäkelä, 2019).

The promising results in ML have encouraged us to apply neural network methods to historical simulation data of ICON-D2-EPS, an NWP ensemble model for central Europe with a horizontal resolution of ∼2 km (Reinert *et al.*, 2020; Zängl *et al.*, 2015). ICON-D2-EPS is a limited-area model that explicitly resolves convection and is run operationally by the German Meteorological Service. To the best of our knowledge, neural networks have not yet been employed for the identification of thunderstorm occurrence in ensemble data with a comparable horizontal resolution. In this work, we present the neural network model SALAMA (Signature-based Approach of Identifying Lightning Activity Using Machine Learning). It has been trained to predict thunderstorm occurrence through the post-processing of simulation data. In Section 2 we describe how independent datasets for the training, testing, and validation of our model have been compiled from NWP forecasts and lightning data. Details on the ML architecture are provided in Section 3. While thunderstorm occurrence is identified in a pixel-wise manner, we systematically vary the spatiotemporal criteria by which the lightning observations are associated with the NWP data. This enables us to study the effect of different spatial scales on the model identification skill and allows us to estimate the advection speed of thunderstorms. Further results are presented in Section 4 and demonstrate that, for lead times up to at least 11 h, SALAMA is more skillful than a baseline method based only on convective available potential energy. In addition, we show a linear relationship between the spatial resolution scale of our model and the time-scale during which skill decreases with lead time. This is consistent with earlier findings that resolving smaller scales brings faster growing forecast errors about (Lorenz, 1969; Selz & Craig, 2015).

2 | DATA

We collected simulation data from the ICON-D2-EPS ensemble model, as well as lightning observations from the lightning detection network LINET (Betz *et al.*, 2009). The simulations were used to extract predictors of thunderstorm occurrence, and lightning observations serve as ground truth.

2.1 | Study region and period

The model domain of ICON-D2-EPS covers the areas of Germany, Switzerland, Austria, Denmark, Belgium, the Netherlands and parts of the neighboring countries. For our study, we cropped the model domain at its borders by approximately 100 km to reduce boundary computation errors. In a cylindrical projection, our study region corresponds to a rectangle with the southwest corner located at 45° N, 1° E, the northeast corner located at 56° N, 16° E and all sides being either parallels or meridians; see Figure 5.

There are daily model runs every 3 h starting at 0000 UTC. We collected simulation data from June to August 2021 over the entire study region in hourly steps, taking always the latest available forecast for each hour. Following this procedure results in forecasts with lead times of 0, 1, or 2 h.

Each model run has 20 ensemble members that differ from each other in a manner consistent with the NWP uncertainty in the initial conditions, model error, and boundary conditions (Reinert *et al.*, 2020). In Section 4.2, we will relate NWP forecast uncertainty, estimated by ensemble variability, to ML model skill.

2.2 | NWP predictors

The atmospheric fields used as predictors of thunderstorm occurrence in this study are given in Table 1. They have been selected as follows. We considered as candidate predictors all two-dimensional fields provided in ICON-D2-EPS, as well as two ICON-D2-EPS pressure-level fields associated with deep moist convection in the literature; namely, the relative humidity at 700 hPa and the vertical wind speed in pressure coordinates at 500 hPa (Li *et al.*, 2021). In addition, we stipulated that the predictors be available on the open-data server of the German Meteorological Service (https://opendata.dwd.de), such that the trained model can eventually be used in real time. For a given candidate input field, we retrieved values on the grid points and time instants on the study domain and period. We also checked for each value whether a thunderstorm occurred (see Section 2.3). Next, we compared histograms of the distribution of the given field during and in the absence of thunderstorm occurrence and kept only fields that differed significantly in the two distributions.

As shown in Table 1, all predictors can be related to thunderstorm activity through physical mechanisms, like instability and moisture. In particular, our selection process has led to predictors that agree with findings in the literature (Jardines *et al.*, 2021; Leinonen *et al.*, 2022; Ukkonen & Mäkelä, 2019). Conversely, convective inhibition, which is sometimes listed as a convective predictor (Kamangir *et al.*, 2020), has not passed the selection process. This is likely due to the fact that we have checked for predictive power in terms of developed thunderstorms. Convective inhibition, however, correlates with the hours leading up to a thunderstorm and has been removed once the storm reaches its mature stage.

It is worth stressing that we have excluded certain parameters on purpose, namely, the geographical location of a thunderstorm event, the time of the day, and the time of the year. In doing so, we assume the existence of a universal signature shared by all thunderstorms, irrespective of where and when they occur. In addition, the list of predictors does not include the lead time of the forecast. In Section 4 We check whether our model, which has been trained on data with lead times between 0 and 2 h, displays skill on data with longer lead times.

2.3 | Lightning observations

In supervised learning, ML models are trained on data for which the ground truth is known. For this reason, we required knowledge of thunderstorm occurrence for our study domain and period. By reason of their high detection

TABLE 1 List of the 21 input parameters used in the study ("DIA": including sub-grid scale).

Physical significance	ICON parameter name	Description
Instability	CAPE_ML	Mixed-layer convective available potential energy
	CEILING	Ceiling height
	OMEGA500	Vertical wind speed in pressure coordinates at 500 hPa
	PS	Surface pressure
	PMSL	Surface pressure reduced to mean sea level
Cloud cover	CLCH	High level clouds (0–400 hPa)
	CLCM	Mid-level clouds (400–800 hPa)
	CLCL	Low-level clouds (800 hPa to soil)
	CLCT	Total cloud cover
Precipitation and moisture	DBZ_CMAX	Maximal radar reflectivity
	ECHOTOP	Echotop pressure
	RELHUM700	Relative humidity at 700 hPa
	RELHUM_2M	2 m relative humidity
Column-integrated water quantities	TQC, TQC_DIA	Cloud water
	TQG	Graupel
	TQI, TQI_DIA	Ice
	TQV, TQV_DIA	Water vapor
	TWATER	Total water content

efficiency and spatial accuracy over the entire study region, we employed lightning observations to assess the occurrence of thunderstorms. Specifically, we resorted to the LINET network (Betz *et al.*, 2009), which exploits the radio spectrum to continuously measure strokes of lightning over Europe. The technology achieves a detection efficiency of more than 95% and an average location accuracy of 150 m. Though the technology is able to differentiate between cloud-to-ground and intracloud flashes, we have considered all lightning events as we are only interested in the yes/no occurrence of thunderstorm activity.

Given a set of predictors retrieved from a grid point $\mathbf{x}$ on the study domain at time t during the study period, we considered thunderstorm activity to occur at $(\mathbf{x}, t)$ if a flash was detected at any $(\mathbf{x}_l, t_l)$ with

$$\|\mathbf{x} - \mathbf{x}_l\| < \Delta r \qquad |t - t_l| < \Delta t, \tag{1}$$

where $\|\cdot\|$ denotes the great-circle distance between $\mathbf{x}$ and $\mathbf{x}_l$. We trained our model with different values for the spatial and temporal thresholds Δr and Δt in order to study the relationship between them and classification skill systematically.

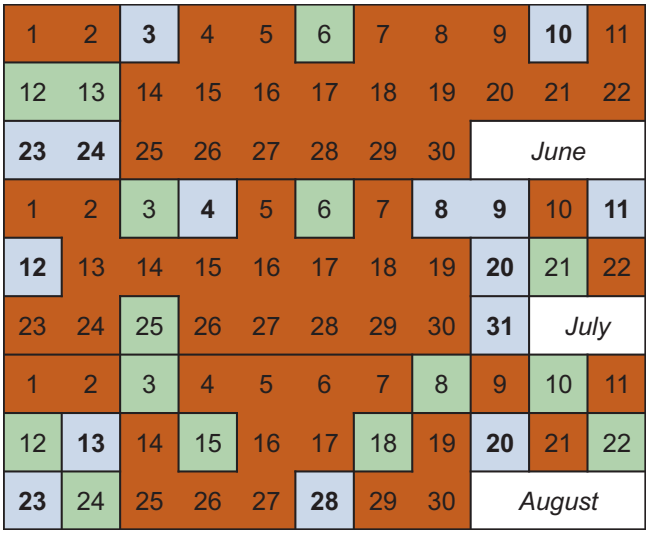

FIGURE 1 Days (from 0800 UTC to 0800 UTC) during the summer of 2021 that were used for compiling the datasets for training (dark background), testing (light background with bold numerals), and validation (light background). The days have been distributed at random among the three sets.

2.4 | Compiling independent datasets

The data obtained from NWP and lightning observations can be considered a set of tuples (ξ, y), where $\xi \in \mathbb{R}^n$ denotes the $n = 21$ input parameters and $y \in \{0, 1\}$ corresponds to a label of the ground truth (1: thunderstorm occurrence; 0: no thunderstorm occurrence). As the input fields were provided on a triangular grid, we first performed an interpolation onto a 0.125° × 0.125° longitude–latitude grid. The labels were produced on the same grid. For each full hour during the study period, for each ensemble member and for each grid point, we fetched the input parameters and the corresponding label, taking always the latest available forecast.

We compiled three statistically independent datasets. A training set was used only for training the neural network model (a precise definition of training is given in Section 3.1), whereas its skill was measured on a test set with data that the model had not seen during training. A third dataset, the validation set, was used to monitor training progress (see Section 3.1). In an attempt to assure statistical independence between the datasets, we took two measures. First, assuming possible day-to-day correlations in the input parameters (e.g., induced by the synoptic scale) to be negligible for convective events with life spans of the order of a few hours, we used separate days for training, testing, and validation. In addition, we took into account that intense thunderstorms that form in the afternoon may well live on after 0000 UTC. We therefore defined days to begin at 0800 UTC, a time of the day chosen by checking when lightning activity in the collected data is minimal. The latter measure prevents data from one thunderstorm at different times appearing in separate datasets. Figure 1 offers an overview of the days contained in each dataset. The days were randomly distributed among the three sets. Additionally, we randomly subsampled the data such that the training set consists of 4×10^5 tuples, and the test and validation sets each contain 10^5 tuples.

The rarity of thunderstorms makes predicting their occurrence more challenging, as ML models tend to struggle with learning from unbalanced datasets (Sun *et al.*, 2009). As a matter of fact, we verified that, when trained on a climatologically consistent dataset, our model would predict the majority class (i.e., no thunderstorm) at every occasion. We therefore undersampled the majority class in the training set, such that both labels appear equally frequently (class balance). On the other hand, the validation and testing sets remain climatologically consistent since we wish to quantify model performance in a realistic setting. Having different sample climatologies in the training and test sets, however, requires model output calibration, which is discussed in Section 3.2.

3 | METHODS

In this section, we provide details on SALAMA, focusing on how it has been trained and calibrated. In addition,

we introduce metrics for the evaluation of model skill and present a baseline model for comparison.

3.1 | Model description

It is worthwhile to introduce some ML terminology. The three datasets used for training, testing, and validation (see Section 2.4) are made up of examples $(\boldsymbol{\xi}, y)$. Each example consists of a pattern $\boldsymbol{\xi} \in \mathbb{R}^n$ of n input features and a label $y \in \{0, 1\}$.

Given a pattern $\boldsymbol{\xi}$, the problem at hand is to infer the probability of thunderstorm occurrence, which constitutes a task known as binary classification. In the following, we consider both the pattern and its corresponding label to originate from a random experiment. Therefore, let $\boldsymbol{\Xi}$ be an n-dimensional random variable for the pattern and let Y be a random variable of thunderstorm occurrence (1: thunderstorm; 0: no thunderstorm). We are interested in $P(Y = 1|\boldsymbol{\Xi} = \boldsymbol{\xi})$; namely, the conditional probability of thunderstorm occurrence if the pattern is known. A feedforward artificial neural network model is a function $f : \mathbb{R}^n \to (0, 1)$ that models the relationship between the input pattern and the corresponding probability of thunderstorm occurrence. We refer to f simply as a neural network. Neural networks use compositions of matrix multiplications, as well as nonlinear operations referred to as activation functions. The architecture of our neural network is presented in Figure 2. It consists of the input and output layers as well as hidden layers, where each layer is a vector of numbers obtained from the previous layer by one matrix multiplication and by applying an activation function to the result in a component-wise manner. The complexity of f is adjustable through the number of hidden layers and the size of each hidden layer; that is, the number of nodes. Our model has three hidden layers and 20 nodes per hidden layer. Moreover, we use rectified linear units for the hidden layers and a sigmoid function to map the output layer to a probability between zero and one.

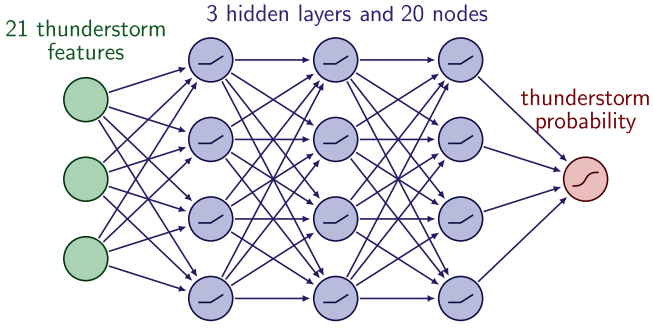

FIGURE 2 The architecture of SALAMA. Input features are scaled to order 1. We use rectified linear units as activation functions in the hidden layers. A sigmoid function maps the output layer to the open interval $(0, 1)$.

The entries, also referred to as weights, of the matrices that connect the layers are adjusted according to the data in the training set. We therefore add a subscript $\mathbf{w} \in \mathbb{R}^d$ to f to express the dependence on the d weights. If $f_{\mathbf{w}}$ constitutes an accurate representation of the conditional probability of thunderstorm occurrence—that is, $f_{\mathbf{w}}(\boldsymbol{\xi}) \approx P(Y = 1|\boldsymbol{\Xi} = \boldsymbol{\xi})$—then the likelihood of observing a label y for a given input feature $\boldsymbol{\xi}$ reads

$$L(\mathbf{w}|\boldsymbol{\xi}, y) = \begin{cases} f_{\mathbf{w}}(\boldsymbol{\xi}), & y = 1, \\ 1 - f_{\mathbf{w}}(\boldsymbol{\xi}), & y = 0. \end{cases} \tag{2}$$

Denote by $(\boldsymbol{\xi}^{(i)}, y^{(i)})_{i=1\ldots N}$ the training set with N examples. The most likely configuration of weights, given the training set, is then obtained by minimizing the negative logarithm of the likelihood function,

$$-\log \mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{N} \log L(\mathbf{w}|\boldsymbol{\xi}^{(i)}, y^{(i)}), \tag{3}$$

with respect to the weights. The expression in Equation (3) is referred to as the binary cross-entropy loss function in ML terminology. The process of determining the weights that minimize loss is called training. We trained SALAMA using the robust iterative stochastic method Adam (Kingma & Ba, 2014). However, if one used the configuration of weights that minimizes Equation (3) exactly, a neural network would likely suffer from overfitting (i.e., learning parts of the noise in the data as well). To this end, we implemented an early stopping procedure, in which loss was monitored on the validation set during training. Once the validation loss no longer decreased, training was stopped.

Before training, each input feature has been scaled in a way that its sample standard deviation in the training set is of the order of unity. In addition, we trained not only on the architecture presented in Figure 2 but also varied the number of hidden layers, as well as the number of nodes per layer. We found that, once a certain complexity was reached in terms of the size of the network, adding new nodes or layers had no effect on the validation loss at the end of training. The architecture in Figure 2 constitutes the smallest network for which this complexity threshold has been exceeded.

3.2 | Analytic model calibration

In order to address the climatological rarity of thunderstorm occurrence, we have artificially increased the

fraction of positive examples in the dataset used for the training of our neural network (see Section 2.4). In this section, we explain why this dataset modification causes our model to be miscalibrated and derive an analytic correction for model output calibration.

It is crucial to understand that if the trained model were naively applied to a test set with a different fraction of positive examples than in the training set, the produced probabilities would be inconsistent with the observed relative frequency of thunderstorm occurrence. In order to see this, we use Bayes' theorem to expand the conditional probability of thunderstorm occurrence given a pattern $\boldsymbol{\xi}$, which yields

$$P(Y=1|\boldsymbol{\Xi}=\boldsymbol{\xi}) = \frac{P(\boldsymbol{\Xi}=\boldsymbol{\xi}|Y=1)P(Y=1)}{P(\boldsymbol{\Xi}=\boldsymbol{\xi})}. \tag{4}$$

The denominator can be expressed as

$$\begin{aligned} P(\boldsymbol{\Xi}=\boldsymbol{\xi}) = {} & P(\boldsymbol{\Xi}=\boldsymbol{\xi}|Y=1)P(Y=1) \\ & + P(\boldsymbol{\Xi}=\boldsymbol{\xi}|Y=0)P(Y=0). \end{aligned} \tag{5}$$

Let $P(Y=1) = 1 - P(Y=0) = g$, where g denotes the climatological probability of thunderstorm occurrence with no prior knowledge. Then,

$$P(Y=1|\boldsymbol{\Xi}=\boldsymbol{\xi}) = \frac{1}{1+(1-g)R(\boldsymbol{\xi})/g}, \tag{6}$$

where the residual function $R(\boldsymbol{\xi}) = P(\boldsymbol{\Xi}=\boldsymbol{\xi}|Y=0)/P(\boldsymbol{\Xi}=\boldsymbol{\xi}|Y=1)$ is not expected to depend on g. Equation (6) shows that the conditional probability of thunderstorm occurrence carries an implicit g dependence. The training set contains an increased fraction $\tilde{g}$ of positive examples (in our work, $\tilde{g}=1/2$), whereas the corresponding fraction in the test set is (up to fluctuations due to the finite sample size) equal to the climatological value g. During training, therefore, the neural network learns to produce the following model output:

$$f_{\mathbf{w}}(\boldsymbol{\xi},\tilde{g}) = \frac{1}{1+(1-\tilde{g})R(\boldsymbol{\xi})/\tilde{g}}. \tag{7}$$

When we want to apply our neural network to a dataset with $g \neq \tilde{g}$, the correct probability output reads

$$f_{\mathbf{w}}(\boldsymbol{\xi},g) = \frac{f_{\mathbf{w}}(\boldsymbol{\xi},\tilde{g})}{f_{\mathbf{w}}(\boldsymbol{\xi},\tilde{g}) + \frac{1-g}{g}\frac{\tilde{g}}{1-\tilde{g}}(1-f_{\mathbf{w}}(\boldsymbol{\xi},\tilde{g}))}, \tag{8}$$

which can be derived by formulating Equation (7) with g and $\tilde{g}$ and using one equation to eliminate $R(\boldsymbol{\xi})$ in the other one. On the other hand, if the sample climatologies of the training set and test set are equal ($\tilde{g}=g$), Equation (8) yields $f_{\mathbf{w}}(\boldsymbol{\xi},g) = f_{\mathbf{w}}(\boldsymbol{\xi},\tilde{g})$; that is, no probability correction is needed.

If the model probability output is consistent with the observed relative frequency of thunderstorm occurrence, the model forecasts are referred to as reliable. In order to check whether our neural network provides reliable forecasts, we used the test set to produce a reliability diagram. For this purpose, one partitions the interval $(0,1)$ of possible forecast probabilities into bins. For each bin, one considers all examples whose model probability falls into the bin. Then, one computes the relative frequency of thunderstorm occurrence and plots it against the bin-averaged model probability per bin. The resulting curve is referred to as the calibration function. An example for one configuration of lightning labels is shown in Figure 3a, for which 10 equidistant bins have been used. Shown are two calibration functions: The light gray line corresponds to a calibration function "without" any probability correction, whereas the solid black line results from applying Equation (8) to our model output. The uncertainty on the observed frequency spans the 5th and 95th percentiles of fluctuations and has been estimated through a bootstrap resampling procedure similar to Bröcker and Smith (2007a): By drawing with replacement, one produces variations of the original test set and considers the sample-to-sample fluctuations of observed relative frequencies. The uncalibrated line severely overestimates the relative frequency of thunderstorm occurrence at all model probabilities. As has been worked out, this is not a result of faulty training but stems from having different sample climatologies in the training and test sets. After calibration, however, our model produces reliable forecasts for probabilities close to 0 and 1. On the other hand, our model underestimates the relative frequency of thunderstorm occurrence for forecast probabilities below 0.6. Further calibration could be done using statistical methods like isotonic regression (Niculescu-Mizil & Caruana, 2005), which is beyond the scope of this work. Instead, we consider our model sufficiently reliable and appreciate that the level of reliability has been attained by means of the analytical correction, Equation (8), alone.

In addition to calibration curves, binning the forecast probabilities allows the introduction of two useful metrics of classification skill. Of the N examples in the test set, let N_i fall into bin i with bin width Δp_i, bin-averaged model probability p_i and observed relative frequency $\overline{o}_i$ of thunderstorm occurrence. We then define the following two bin-wise terms:

$$\mathrm{RES}_i = \frac{1/\Delta p_i}{g(1-g)}\frac{N_i}{N}(p_i-g)^2, \tag{9}$$

$$\mathrm{REL}_i = \frac{1/\Delta p_i}{g(1-g)}\frac{N_i}{N}(p_i-\overline{o}_i)^2. \tag{10}$$

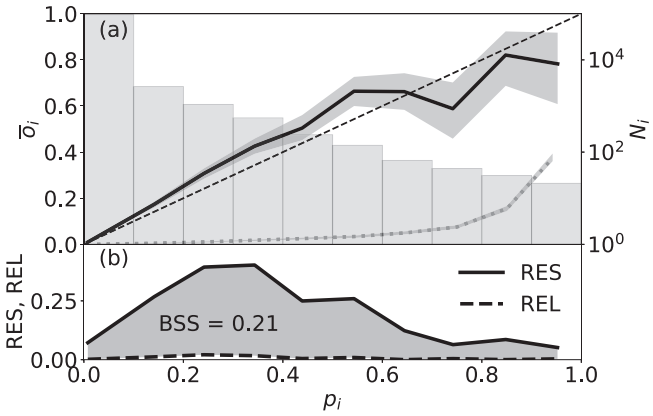

FIGURE 3 Reliability diagram of SALAMA, evaluated for the test set with the label configuration $\Delta r = 15$ km, $\Delta t = 30$ min (see Section 2.3). (a) Calibration curve after applying probability correction, Equation (8) (black solid line), and before (gray light dotted line), and histogram of examples per bin. Perfect reliability is indicated by a dashed diagonal. Shaded band corresponds to the symmetric 90% confidence interval obtained by 200 bootstrap resamples. (b) Bin-wise resolution (RES) and reliability (REL)—see Equations 9 and 10—and their relation to the Brier skill score (BSS; see Section 3.3) as a function of model probability.

Up to a factor $g(1-g)$, known as the uncertainty term, the sums $\sum_i \Delta p_i \mathrm{RES}_i$ and $\sum_i \Delta p_i \mathrm{REL}_i$ are respectively called the resolution and reliability of the model. Resolution measures forecast variance, with higher values of resolution indicating a better ability of the model to differentiate between thunderstorm and non-thunderstorm patterns (Toth *et al.*, 2003). Reliability quantifies the mean-squared deviation of the calibration curve from the diagonal. The bin-wise terms defined in Equations 9 and 10 offer an overview of how much each probability bin contributes to reliability and resolution. For instance, both resolution and reliability are most impacted by examples with model probabilities of ~0.25.

3.3 | Skill evaluation metrics

Metrics for evaluating classification skill using a test set with N examples include the Brier score (BS),

$$\mathrm{BS} = \sum_{k=1}^{N} (p^{(k)} - y^{(k)})^2, \quad p^{(k)} = f_{\mathbf{w}}(\boldsymbol{\xi}^{(k)}, g), \tag{11}$$

which is known for being strictly proper (Bröcker & Smith, 2007b). Normalization with a reference BS, $\mathrm{BS}_{\mathrm{ref}} = \sum_{k=1}^{N} (g - y^{(k)})^2$, of a random climatological model yields the Brier skill score (BSS):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}. \tag{12}$$

TABLE 2 Contingency matrix for binary classification.

Forecast thunderstorm	Observed thunderstorm	
	True	False
True	Hit	False alarm
False	Miss	Correct reject

Murphy (1973) showed that BSS can be written as the difference between resolution and reliability (see Section 3.2). Thus, in terms of Equations (9) and (10), BSS is given by the area between RES and REL as functions of p. This is illustrated in Figure 3b.

Though the BSS directly acts on the probability outputs $p^{(k)}$ of the model—see Equation (11)—a large class of classification metrics requires the conversion of probabilities to binary output first. This is done by introducing a decision threshold $\tilde{p}$. If $p > \tilde{p}$, thunderstorm occurrence for the corresponding example is deemed "true," otherwise it is "false." In combination with the two options from the label, there are four possible outcomes for each example. They are presented as a contingency matrix in Table 2.

Though there are an infinite number of options to combine the four possible outcomes to a single skill score, we selected the scores in this study based on their suitability for tasks with significant class imbalance. Namely, we do not wish to reward our model for correctly classifying the majority class. This amounts to dismissing scores that explicitly use correct rejects.

Given a test set and a fixed decision threshold, the probability of detection (POD) and false-alarm ratio (FAR) are defined by

$$\mathrm{POD} = \frac{\mathrm{Hits}}{\mathrm{Hits} + \mathrm{Misses}}, \tag{13}$$

$$\mathrm{FAR} = \frac{\mathrm{False\ alarms}}{\mathrm{Hits} + \mathrm{False\ alarms}}. \tag{14}$$

Here, "Hits" refers to the number of examples in the test set that qualify as a "hit" according to Table 2. POD is often referred to as recall in the ML literature, whereas 1 − FAR is also known as precision.

Precision and recall need to be simultaneously optimized for a useful classifier. For instance, perfect recall is easily achieved by predicting the thunderstorm class at every occasion. For problems with class imbalance, a popular choice of combining the two scores consists of taking the harmonic mean, which yields the F_1 score:

$$F_1 = \frac{2}{\mathrm{POD}^{-1} + (1 - \mathrm{FAR})^{-1}} = \frac{2 \times \mathrm{Hits}}{2 \times \mathrm{Hits} + \mathrm{Misses} + \mathrm{False\ alarms}}. \tag{15}$$

Another option of combining the contingency matrix elements is given by the critical success index (CSI):

$$\mathrm{CSI} = \frac{\mathrm{Hits}}{\mathrm{Hits} + \mathrm{Misses} + \mathrm{False\ alarms}}. \tag{16}$$

A modification of the CSI consists of subtracting as many hits as a model randomly classifying according to climatology would obtain. The equitable threat score (ETS) reads

$$\mathrm{ETS} = \frac{\mathrm{Hits} - \mathrm{Hits\ by\ accident}}{\mathrm{Hits} - \mathrm{Hits\ by\ accident} + \mathrm{Misses} + \mathrm{False\ alarms}}, \tag{17}$$

where the hits by accident amount to $g \times (\mathrm{Hits} + \mathrm{False\ alarms})$.

3.4 | Baseline model

As thunderstorms are accompanied by convective precipitation, radar reflectivity constitutes a natural surrogate for thunderstorm occurrence in the nowcasting community (Dixon & Wiener, 1993; Turner *et al.*, 2004; Wilson *et al.*, 1998). ICON-D2-EPS outputs the column-maximal radar reflectivity (see DBZ_CMAX in Table 1), which we refer to as reflectivity in what follows. In order to construct a baseline for comparison with SALAMA, we repeat training our model, but use only reflectivity as input. The architecture of the baseline model is identical to the one presented in Figure 2 except for the input layer, which now has only a single node.

Figure 4 shows the resulting reliability diagram. The light dotted line corresponds to the uncorrected calibration curve, whereas the dash-dotted line results from applying probability correction, Equation (8). The baseline model produces well-calibrated output for small model probabilities, whereas the model displays underconfidence above probabilities of approximately 0.2. As examples with higher probabilities than 0.2 make up less than 1% of the examples in the test set, we therefore assume that these examples did not contribute sufficiently to the loss function, which instead favored well-calibrated small probabilities. In an effort to construct a competitive baseline model, we use the validation set to fit a linear function to the part of the dash-dotted calibration curve with probabilities higher than 0.15. Then, if the output of the baseline model after application of probability correction, Equation (8), is denoted by p, the calibrated output reads $C(p)$ for $p > 0.15$, and p otherwise. The resulting well-calibrated calibration curve is given by the solid line in the reliability diagram. The histogram and the lower panel in Figure 4a refer to the latter calibration curve. One can see that BSS is essentially determined by the baseline resolution. Both SALAMA (see Figure 3) and the baseline model receive most contributions to the reliability term from model probabilities around 0.2. As a matter of fact, the baseline scores better than SALAMA in terms of reliability. On the other hand, the baseline resolution is significantly worse, which results in a lower BSS.

Figure 4b shows the learned and calibrated relationship between NWP reflectivity and the corresponding probability of thunderstorm occurrence. The herein observed monotonously increasing relationship implies that thunderstorms become more likely as reflectivity increases. A typical threshold for defining thunderstorms in nowcasting is 35 dBZ (Dixon & Wiener, 1993; Mueller *et al.*, 2003), for which the probability of thunderstorm occurrence reads 0.22.

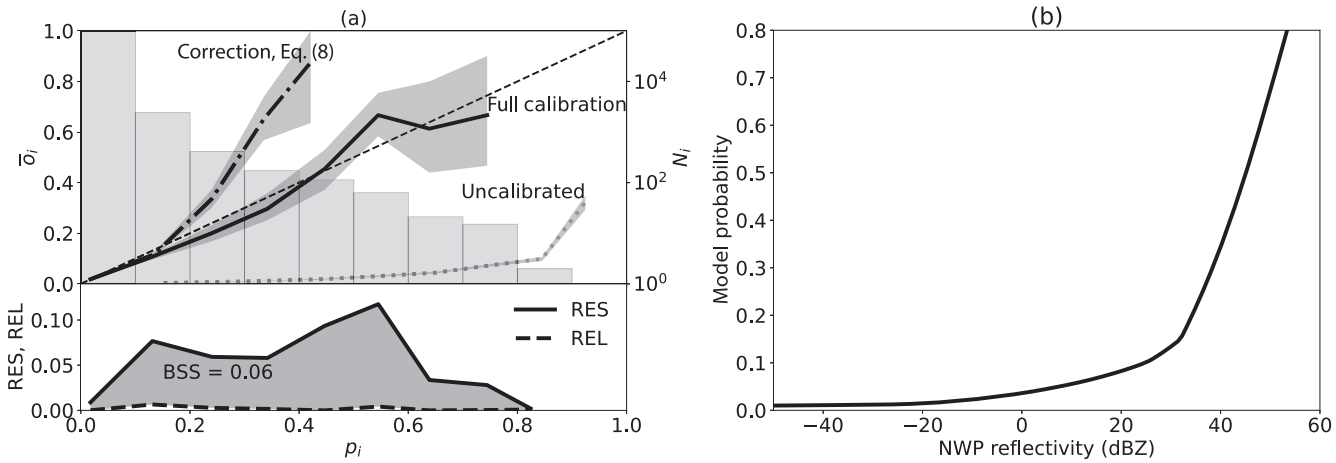

FIGURE 4 Training of the baseline model. (a) Reliability diagram panels as in Figure 3, but for the baseline model. (b) Learned relationship between the baseline input field and the corresponding probability of thunderstorm occurrence. BSS: Brier skill score; NWP: numerical weather prediction; RES: resolution; REL: reliability.

4 | RESULTS

SALAMA provides a general post-processing framework for NWP ensemble forecasts. Whereas we trained SALAMA on lead times up to 2 h, we apply the same model to all lead times and all ensemble members individually, using neither the lead time nor the ensemble member index as input feature. Working with ensemble data, our framework readily allows us to study the ensemble spread of thunderstorm occurrence. For example, if we have, for a given location, a 3 h forecast of ICON-D2-EPS at hand, it consists of 20 input feature tuples (one tuple for each ensemble member). One can now compute a thunderstorm probability according to Equation (8) for each member. As we will discuss in Section 4.2, the ensemble spread of thunderstorm probability is linked to the NWP forecast uncertainty of the input features. In the following, we compare SALAMA with the baseline model based on reflectivity (see Section 3.4) and move on to investigating how the spatiotemporal thresholds of the lightning label configuration (see Section 2.3) influence the classification skill of SALAMA as a function of lead time.

4.1 | Comparison with baseline model

In this section, we keep the thresholds of the lightning label configuration (see Section 2.3) fixed to the particular choice $\Delta r = 15$ km, $\Delta t = 30$ min. The climatological fraction of thunderstorm examples in the test set amounts to $g = 0.021$ in this configuration. The results of this section, however, do not change qualitatively if another configuration is used.

As a first step, we visually compare the performance of SALAMA and the baseline model in a case study. For this purpose, we run SALAMA for three consecutive hours of an evening with thunderstorm occurrence over southern Germany. This day has not been used for the training of SALAMA. In Figure 5 we plot the probability of

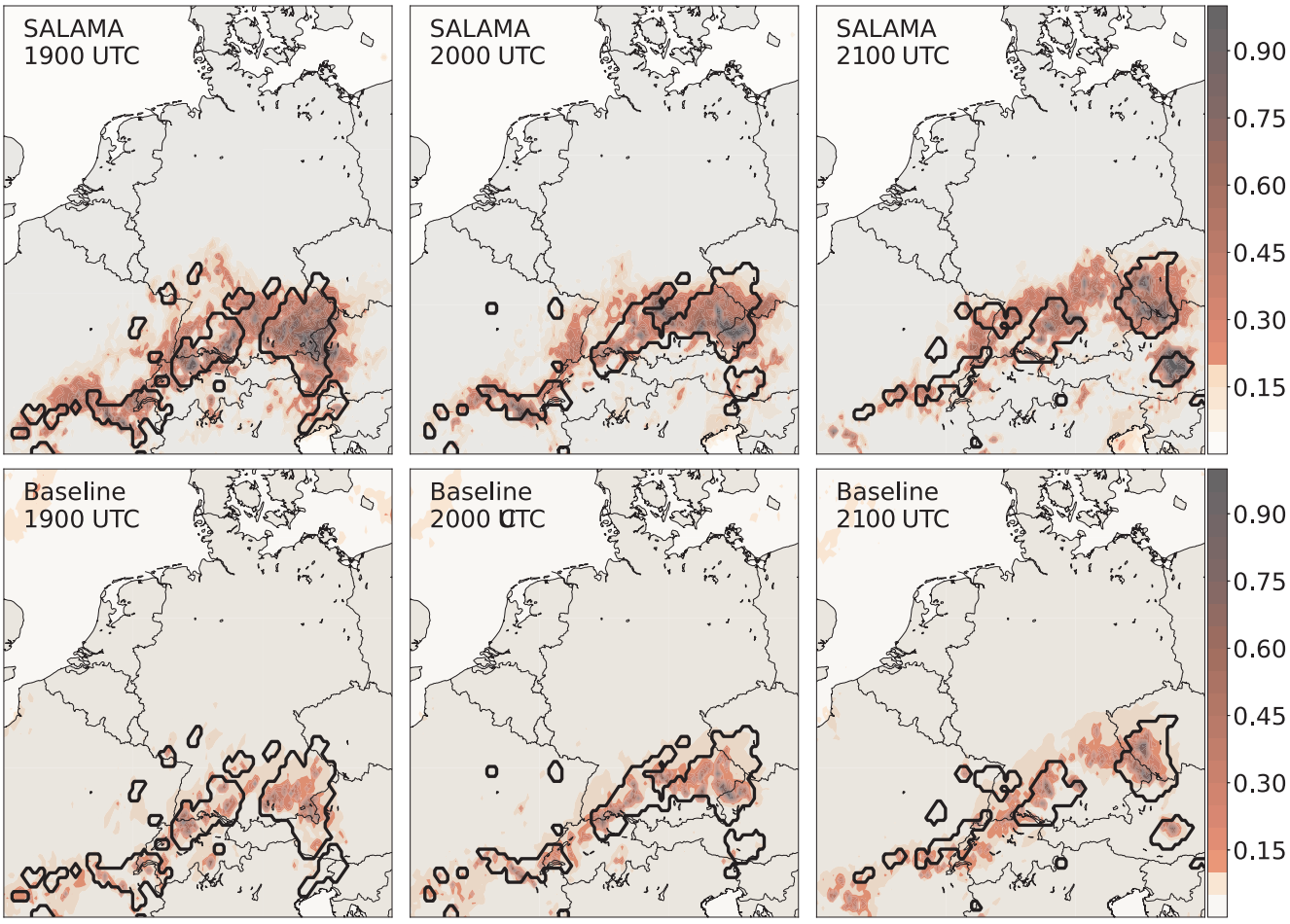

FIGURE 5 Probability of thunderstorm occurrence for June 23, 2021, from 1900 UTC on, for SALAMA (upper row) and the baseline model (lower row). The model lead times for the three hours are 1 h, 2 h, and 0 h, respectively. The filled contours display the result for the first ensemble member of ICON-D2-EPS, whereas lightning labels ($\Delta r = 15$ km, $\Delta t = 30$ min; see Section 2.3) are shown as black contours. A jump in the color maps indicates the decision thresholds used for the evaluation of the skill scores in Table 3.

thunderstorm occurrence for an arbitrary member of the NWP ensemble for the entire study domain. Observed thunderstorm occurrence is given by black contours. The corresponding plots for the baseline model are added below the panels of SALAMA. In this particular case study, SALAMA tends to detect far more thunderstorm pixels than the baseline model does. On the other hand, SALAMA seems to produce more false alarms.

In order to compare the skill of SALAMA and our baseline quantitatively for the entire study period, we evaluate the skill scores introduced in Section 3.3. We use for this purpose the test set introduced in Section 2.4, which consists of examples of the entire summer of 2021. For some of the scores, we need to set a decision threshold. As a criterion, we demand that forecasts be unbiased (average fraction of examples classified as thunderstorms is equal to the observed fraction of thunderstorm examples), yielding thresholds of 0.148 (SALAMA) and 0.126 (baseline). The thresholds are also indicated in the color bars of Figure 5. The threshold found for reflectivity corresponds to 28 dBZ and is slightly below the typical literature threshold cited in Section 3.4.

The performance of SALAMA and the baseline is summarized in Table 3. Irrespective of the skill score under consideration, SALAMA scores better than the baseline model. The uncertainties are computed here, as well as for the subsequent evaluations, by the bootstrap resampling method introduced in Section 3.2. Note that we obtain POD = 1 − FAR = F_1 for both models. This is a result from our choice of decision threshold: recall generally equals precision for unbiased forecasts (Wilks, 2011).

Drawing (POD, 1 − FAR) for different decision thresholds into one diagram, one obtains the precision–recall (PR) diagram in Figure 6. A random model with no skill corresponds to the dashed horizontal curve 1 − FAR = g, where g denotes the climatological fraction of positive examples in the test set. Models with skill display PR curves above the horizontal line, with higher areas under the curve (AUCs) indicating higher classification skill. Both models considered in this study display higher skill than a random model following climatology would. SALAMA, however, has higher classification skill than the baseline, as can be seen from the higher AUC in the PR curve in Figure 6. The enhanced skill of SALAMA with respect to the baseline model illustrates that a multiparameter approach to thunderstorm forecasting is superior to employing a single input feature.

TABLE 3 Scores for classification skill, evaluated on the test set, for SALAMA and the baseline model. The probability thresholds used for evaluation are chosen such that the forecast is unbiased and amount to 0.148 (SALAMA) and 0.126 (baseline). Uncertainties are obtained from 200 bootstrap resamples and show the symmetric 90% confidence interval.

Skill score	SALAMA	Baseline
PR-AUC	0.358 (18)	0.141 (12)
BSS	0.209 (10)	0.063 (7)
POD	0.403 (16)	0.189 (12)
1 − FAR	0.402 (17)	0.188 (13)
F_1	0.403 (15)	0.189 (12)
CSI	0.252 (12)	0.104 (7)
ETS	0.241 (12)	0.093 (7)

Abbreviations: BSS, Brier skill score; CSI, critical success index; ETS, equitable threat score; FAR, false-alarm ratio; POD, probability of detection; PR-AUC, area under the precision–recall curve.

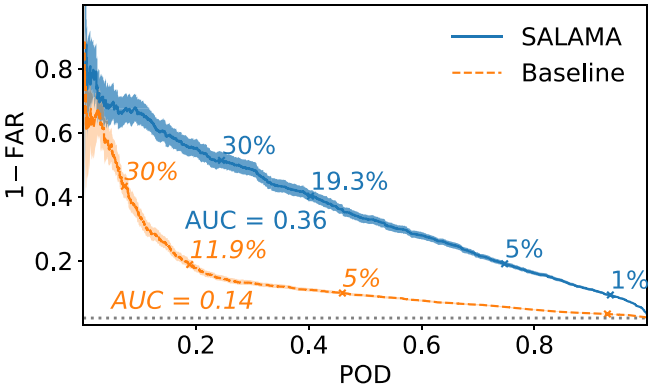

FIGURE 6 Precision–recall curve for SALAMA (solid) and the baseline model (dashed), evaluated on the test set. The annotations added to the curves (for baseline in italics) correspond to different decision thresholds; see Section 3.3. Gray dotted line denotes models with no identification skill. Uncertainties are obtained from 200 bootstrap resamples and show the symmetric 90% confidence interval. AUC: area under the curve; FAR: false-alarm ratio; POD: probability of detection.

4.2 | Lead time dependence of classification skill

The datasets for training, testing, and validation introduced in Section 2.4 and used in Section 4.1 are comprised of NWP forecasts with lead times up to 2 h. The reason for this choice was to train and evaluate our model in a setting of minimal NWP forecast uncertainty. On the other hand, this procedure raises the question whether the thunderstorm signature learned by the model generalizes to NWP data with higher lead times (and higher forecast uncertainty). For this purpose, we generate test sets in which the examples come from NWP forecasts with fixed lead time. Each set contains 10^5 examples. We use the same dates as for the test sets introduced in Section 2.4. In Figure 7 we plot the SALAMA classification skill, measured in terms

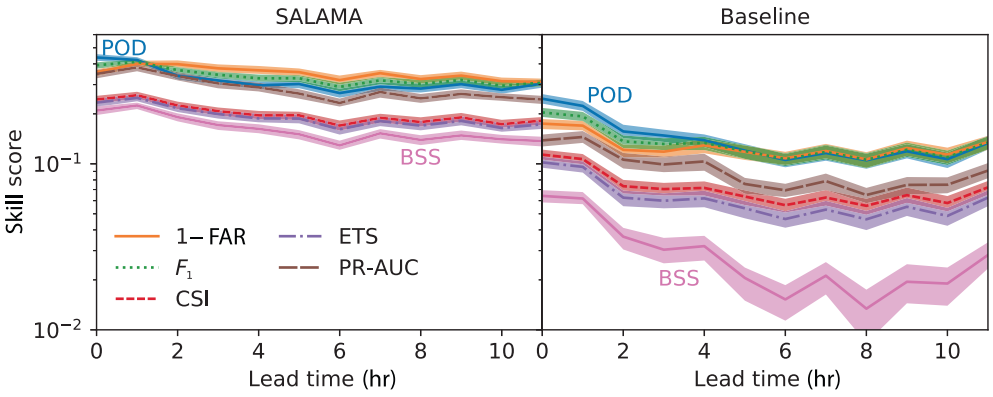

FIGURE 7 Classification skill as a function of lead time for SALAMA (left) and the baseline model (right). The probability thresholds used for evaluation are chosen such that the forecast is unbiased and amount to 0.148 (SALAMA) and 0.126 (baseline). Uncertainties are obtained from 200 bootstrap resamples and show the symmetric 90% confidence interval. BSS: Brier skill score; CSI: critical success index; ETS: equitable threat score; FAR: false-alarm ratio; POD: probability of detection; PR-AUC: area under the precision–recall curve.

of the skill scores introduced in Section 3.3 as a function of lead time and compare it with the dependence obtained for the baseline model. Figure 7 shows that, for SALAMA, classification skill decreases approximately exponentially (note the log-scaling of the y-axis) for lead times longer than 1 h, irrespective of the skill score under consideration. The classification skill of SALAMA at a lead time of 1 h is actually higher than at 0 h, which is likely a spin-up effect from the NWP model (Sun *et al.*, 2014). On the other hand, SALAMA skill is systematically superior to baseline skill for all lead times. In fact, even the 11-h lead-time skill of SALAMA is higher than the baseline skill for any of the lead times considered. For both models, the curves of POD, 1 − FAR, and F_1 start close to one another but then diverge from each other as lead time grows. As the equality of precision and recall is indicative of unbiased forecasts, it follows that forecasts become biased for higher lead times. For small lead times, the forecasts are essentially unbiased. This is consistent with the fact that the decision thresholds have been chosen such that a test set containing lead times up to 2 h yields unbiased forecasts.

It is tempting to assume that the decrease in skill with lead time originates from an increasing NWP forecast uncertainty for longer lead times. We can use ensemble data to check this hypothesis. Let q be either one of the 21 input features or the model thunderstorm probability; that is, a quantity that is given for each ensemble member and for all lead times. Then, define the ensemble spread σ'_q of q as the ensemble standard deviation of q:

$$\sigma'_q(t_{\text{lead}}) = \sqrt{\langle q(t_{\text{lead}})^2\rangle - \langle q(t_{\text{lead}})\rangle^2}, \tag{18}$$

where we make the dependence on the lead time t_{lead} explicit. The angle brackets denote the average over all 20 ensemble members. Denote by $\overline{\sigma'_q}(t_{\text{lead}})$ the expression obtained by performing an average of σ'_q over the entire study region and all times associated with the test set. Lastly, we define the normalized ensemble spread of q,

$$\sigma_q(t_{\text{lead}}) = \frac{\overline{\sigma'_q}(t_{\text{lead}})}{\overline{\sigma'_q}(0\text{ hr})}, \tag{19}$$

as a function of lead time. It quantifies ensemble spread in such a way that different input features can be directly compared with each other. In Figure 8, the normalized ensemble spread of each of the 21 input features is shown as thin solid lines and the corresponding curve for the model output of SALAMA is drawn in thick and dashed lines. One can see that the ensemble spread does indeed increase with lead time for most input features, with the increase being approximately linear. The ensemble spread of the SALAMA output increases in line with the majority of the input features and with a similar slope. This suggests that the decrease in classification skill observed in Figure 6 is solely due to the increasing variance in the simulation data.

4.3 | Effect of the label size

So far, the temporal and spatial thresholds of the label configuration have been fixed to $\Delta r = 15$ km and $\Delta t = 30$ min

(henceforth referred to as the default configuration). In this section, we study how varying the spatiotemporal thresholds affects the classification skill of SALAMA.

As a first step, we compute reliability diagrams for different label configurations. In Figure 9a we study a configuration with smaller thresholds than for the configuration studied so far. Figure 9b displays a configuration with reduced Δt and increased Δr. In Figure 9c, both thresholds are increased with respect to the default configuration. The exact choice of Δt and Δr for the three panels is somewhat arbitrary but still allows for qualitative insight. Irrespective of the configuration, forecasts are well calibrated for small and large model probabilities. In addition, model skill, quantified in terms of BSS, increases from left to right. The diagrams show that the increase in BSS is mainly due to enhanced contribution to resolution from probabilities between 0.3 and 0.6, whereas reliability stays approximately constant.

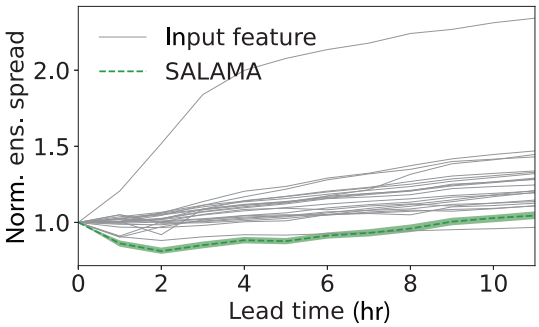

FIGURE 8 Normalized ensemble spread—see Equation (19)—of input features in comparison with spread of model thunderstorm probability as a function of lead time. Each thin solid line refers to one of the 21 input features. The thick dashed line is associated with SALAMA probability output. A shaded band represents the symmetric 90% confidence interval of uncertainty, estimated with 200 bootstrap resamples.

As we have seen in Section 4.2 that the qualitative lead time dependence of SALAMA skill does not depend on the skill score, we consider from now on only PR-AUC for further investigations. We start by computing PR-AUC for several label configurations, which is shown in Figure 10. The color pattern in the figure suggests that the two thresholds are not independent variables of classification skill. Instead, one can find a parameter c with the units of a velocity such that classification skill is nearly constant along lines:

$$s = \Delta r + c\Delta t = \text{const.} \tag{20}$$

Indeed, s corresponds to a spatial resolution scale; it determines the minimal spatial accuracy that can be expected from a model trained with a given label configuration. We expect the parameter c to roughly quantify the speed at which regions of thunderstorm occurrence are advected in the atmosphere. A fit to the data provides $c = 5.2(3)\text{m}\cdot\text{s}^{-1}$, which is similar to typical low- to mid-tropospheric wind speeds in central Europe. Lines of constant spatial scale appear as dashed lines in Figure 10. Classification skill increases with s. This is in line with the displayed observation of increased BSS in the reliability diagrams. This is also consistent with the work of Roberts (2008), which investigates the spatial variation of precipitation forecast skill. Note that the spatiotemporal thresholds for the reliability diagram in the middle panel have been chosen such that s takes on the same value as the default configuration ($\Delta t = 30$ min, $\Delta r = 15$ km).

Next, we investigate how the decrease of classification skill with lead time depends on the spatial scale. Motivated by the observed decay of classification skill with lead time (see Section 4.2), we fit an exponential function $\exp(-t_{\text{lead}}/\tau)$ to the lead time dependence of classification skill (measured again by the PR-AUC). The skill decay time τ then provides a characteristic time-scale for the decrease of classification skill. For each label

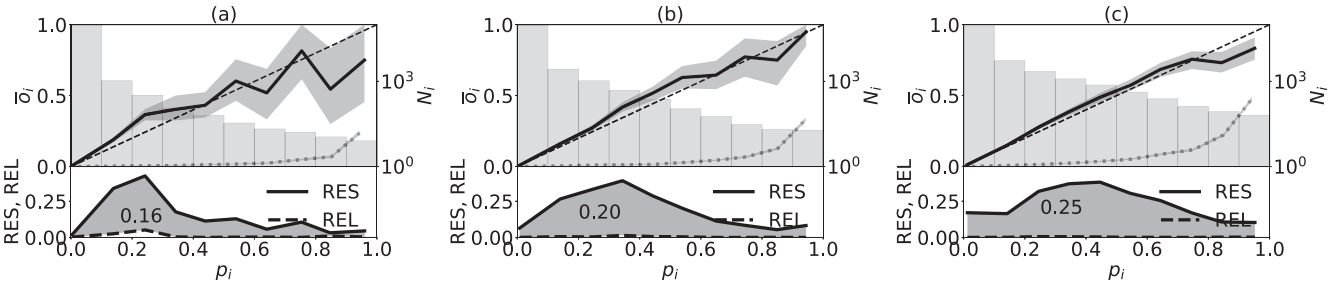

FIGURE 9 Reliability diagrams as in Figure 3, but with label configurations (a) $\Delta t = 15$ min, $\Delta r = 9$ km ($s = 14$ km), (b) $\Delta t = 10$ min, $\Delta r = 21$ km ($s = 24$ km), (c) $\Delta t = 40$ min, $\Delta r = 24$ km ($s = 36$ km). The spatial scale s is introduced in Equation (20). RES: resolution; REL: reliability.

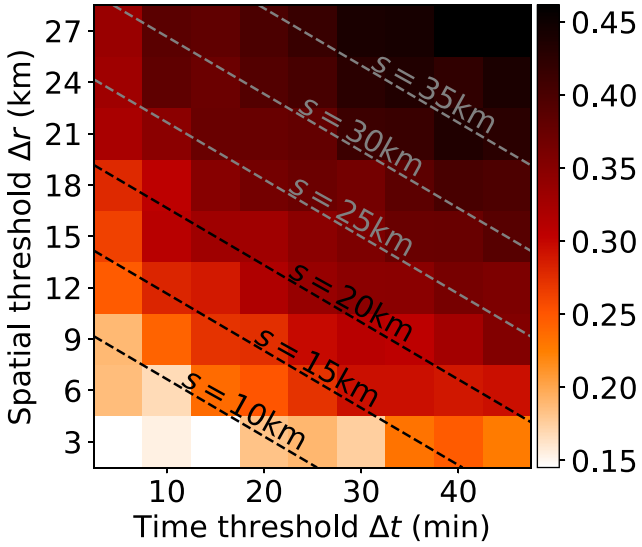

FIGURE 10 Classification skill of SALAMA, expressed in terms of the area under the precision–recall curve, as a function of the label configuration (see Section 2.3). The slope of the dashed lines is chosen such that classification skill is approximately constant along the lines. Each line corresponds to a specific spatial scale s; see Equation (20).

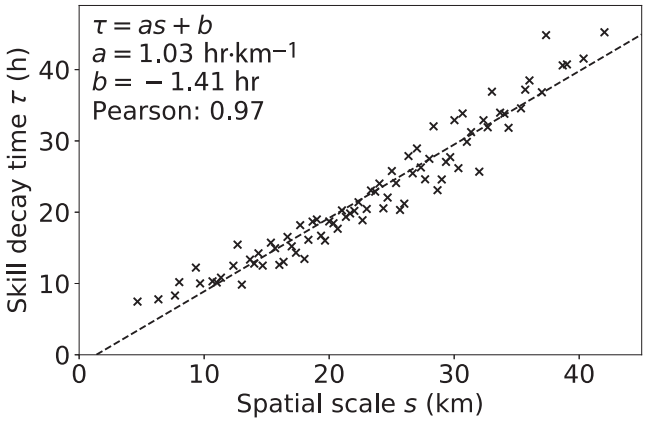

FIGURE 11 Decay time of classification skill (quantified by the area under the precision–recall curve) as a function of the spatial scale. Each data point corresponds to one label configuration in Figure 10. The parameters of a linear fit are also shown, as well as the Pearson coefficient of correlation.

configuration in Figure 10, we compute the corresponding spatial scale as well as τ. In Figure 11 we present a scatter plot of τ and s. The figure shows a tight positive linear correlation between the two quantities, which means that classification skill decreases more slowly for coarser label configurations. This is in agreement with the anticipation (Lorenz, 1969) that the ability to resolve smaller scales in NWP models results in forecast errors growing more rapidly. Our finding is complementary to convection studies involving a scale-dependent skill score (Roberts, 2008) and high-resolution simulations (Selz & Craig, 2015).

5 | CONCLUSION AND PERSPECTIVES

Addressing the need for accurate thunderstorm forecasting and leveraging advances in high-resolution NWP and ML, we have presented SALAMA, a feedforward neural network model that identifies thunderstorm occurrence in NWP forecasts up to 11 h in advance in a pixel-wise manner. The inference of the probability of thunderstorm occurrence is based on input parameters that are physically related to thunderstorm activity and do not explicitly feature information on location, time, or forecast range. This gives reason to expect that the signature learned by the model generalizes to thunderstorms outside the study region of this work and remains valid in a changing climate. In addition, the availability of all input features in real time makes SALAMA readily available for operational use.

We have addressed the technical challenge caused by the rarity of thunderstorms and the corresponding small fraction of positive examples by increasing this fraction during training and analytically accounting for the increase when testing. This approach has allowed us to ensure reasonable reliability without calibration fits. Furthermore, we have proposed a novel visualization of reliability and resolution as a function of bin-wise model probability. The visualization arguably proves useful for evaluating how examples with a certain model probability contribute to classification skill.

Working with ensemble data, we have studied how the NWP forecast uncertainty depends on the lead time of the forecast and related it to the classification skill decrease of SALAMA. This has suggested that the decrease in skill is the result of an increasing uncertainty in the input feature forecasting.

During the training process, we have systematically varied the spatiotemporal criteria by which we associate lightning observations with NWP data. This has allowed us to test SALAMA with different spatial scales and to estimate the order of magnitude of the speed at which thunderstorms are advected in the atmosphere. We have shown that classification skill increases with the spatial scale of the forecast and is higher than for a baseline model based on NWP reflectivity alone. Furthermore, we have found that the decay time of classification skill is proportional to the spatial scale. In combination with the result that the SALAMA classification skill

is correlated with the NWP forecast uncertainty, our findings have indicated that resolving thunderstorms at smaller scales reduces the predictability of thunderstorm occurrence.

In a future study, it will be useful to check the universality of the thunderstorm signature learned by SALAMA by, for example, testing it on data outside of central Europe or for a different time period than the summer of 2021. Moreover, one may explore whether classification skill can be improved by shifting from a pixel-wise consideration of input features to taking their spatiotemporal structure into account as well.

ACKNOWLEDGEMENTS

We thank George Craig and Tobias Selz for helpful discussions. We gratefully acknowledge the computational and data resources provided through the joint high-performance data analytics (HPDA) project "terrabyte" of the DLR and the Leibniz Supercomputing Center (LRZ). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

The Python code for SALAMA will be made available upon reasonable request.

ORCID

Kianusch Vahid Yousefnia https://orcid.org/0000-0003-2644-2539
Tobias Bölle https://orcid.org/0000-0003-3714-6882
Isabella Zöbisch https://orcid.org/0000-0003-2035-7931
Thomas Gerz https://orcid.org/0000-0003-2923-3126

REFERENCES

Bauer, P., Thorpe, A. & Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55.

Betz, H.D., Schmidt, K., Laroche, P., Blanchet, P., Oettinger, W.P., Defer, E. et al. (2009) Linet—an international lightning detection network in Europe. *Atmospheric Research*, 91, 564–573.

Borsky, S. & Unterberger, C. (2019) Bad weather and flight delays: The impact of sudden and slow onset weather events. *Economics of Transportation*, 18, 10–26. Available from: https://www.sciencedirect.com/science/article/pii/S2212012218300753

Bröcker, J. & Smith, L.A. (2007a) Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22, 651–661. Available from: https://journals.ametsoc.org/view/journals/wefo/22/3/waf993.1.xml

Bröcker, J. & Smith, L.A. (2007b) Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22, 382–388. Available from: https://journals.ametsoc.org/view/journals/wefo/22/2/waf966.1.xml

Burke, A., Snook, N., Gagne, D.J., II, McCorkle, S. & McGovern, A. (2020) Calibration of machine learning–based probabilistic hail predictions for operational forecasting. *Weather and Forecasting*, 35, 149–168. Available from: https://journals.ametsoc.org/view/journals/wefo/35/1/waf-d-19-0105.1.xml

Diffenbaugh, N.S., Scherer, M. & Trapp, R.J. (2013) Robust increases in severe thunderstorm environments in response to greenhouse forcing. *Proceedings of the National Academy of Sciences*, 110, 16361–16366. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1307758110

Dixon, M. & Wiener, G. (1993) Titan: Thunderstorm identification, tracking, analysis, and nowcasting—a radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, 10, 785–797. Available from: https://journals.ametsoc.org/view/journals/atot/10/6/1520-0426.1993.010.0785.ttitaa.2.0.co.2.xml

Gagne, D.J., McGovern, A., Haupt, S.E., Sobash, R.A., Williams, J.K. & Xue, M. (2017) Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and Forecasting*, 32, 1819–1840. Available from: https://journals.ametsoc.org/view/journals/wefo/32/5/waf-d-17-0010.1.xml

Geng, Y.-a., Li, Q., Lin, T., Yao, W., Xu, L., Zheng, D. et al. (2021) A deep learning framework for lightning forecasting with multi-source spatiotemporal data. *Quarterly Journal of the Royal Meteorological Society*, 147, 4048–4062. Available from: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4167

Gerz, T., Forster, C. & Tafferner, A. (2012) *Mitigating the Impact of Adverse Weather on Aviation*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 645–659. Available from: https://doi.org/10.1007/978-3-642-30183-4.39

Herman, G.R. & Schumacher, R.S. (2018) Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Monthly Weather Review*, 146, 1571–1600. Available from: https://journals.ametsoc.org/view/journals/mwre/146/5/mwr-d-17-0250.1.xml

Holle, R. L. (2014) Some aspects of global lightning impacts. In 2014 International Conference on Lightning Protection (ICLP), 1390–1395.

Holle, R.L. (2016) A summary of recent national-scale lightning fatality studies. *Weather, Climate, and Society*, 8, 35–42. Available from: https://journals.ametsoc.org/view/journals/wcas/8/1/wcas-d-15-0032.1.xml

Hwang, Y., Clark, A.J., Lakshmanan, V. & Koch, S.E. (2015) Improved nowcasts by blending extrapolation and model forecasts. *Weather and Forecasting*, 30, 1201–1217. Available from: https://journals.ametsoc.org/view/journals/wefo/30/5/waf-d-15-0057.1.xml

Jardines, A., Soler, M., Cervantes, A., García-Heras, J. & Simarro, J. (2021) Convection indicator for pre-tactical air traffic flow management using neural networks. *Machine Learning with Applications*, 5, 100053 Available from: https://www.sciencedirect.com/science/article/pii/S2666827021000256

Kamangir, H., Collins, W., Tissot, P. & King, S.A. (2020) A deep-learning model to predict thunderstorms within 400 km^2

south Texas domains. *Meteorological Applications*, 27, e1905 Available from: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.1905

Kingma, D.P. & Ba, J. (2014) ADAM: A Method for Stochastic Optimization. Available from: https://arxiv.org/abs/1412.6980

Kober, K., Craig, G.C., Keil, C. & Dörnbrack, A. (2012) Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138, 755–768. Available from: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.939

Leinonen, J., Hamann, U., Germann, U. & Mecikalski, J.R. (2022) Nowcasting thunderstorm hazards using machine learning: the impact of data sources on performance. *Natural Hazards and Earth System Sciences*, 22, 577–597. Available from: https://nhess.copernicus.org/articles/22/577/2022/

Li, J., Forster, C., Wagner, J. & Gerz, T. (2021) Cb-fusion–forecasting thunderstorm cells up to 6 hours. *Meteorologische Zeitschrift*, 30, 169–184.

Lin, P.-F., Chang, P.-L., Jou, B.J.-D., Wilson, J.W. & Roberts, R.D. (2012) Objective prediction of warm season afternoon thunderstorms in northern Taiwan using a fuzzy logic approach. *Weather and Forecasting*, 27, 1178–1197. Available from: https://journals.ametsoc.org/view/journals/wefo/27/5/waf-d-11-00105.1.xml

Loken, E.D., Clark, A.J. & Karstens, C.D. (2020) Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Weather and Forecasting*, 35, 1605–1631. Available from: https://journals.ametsoc.org/view/journals/wefo/35/4/wafD190258.xml

Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289–307. Available from: https://doi.org/10.3402/tellusa.v21i3.10086

Mueller, C., Saxen, T., Roberts, R., Wilson, J., Betancourt, T., Dettling, S. et al. (2003) Ncar auto-nowcast system. *Weather and Forecasting*, 18, 545–561. Available from: https://journals.ametsoc.org/view/journals/wefo/18/4/1520-0434.2003.018.0545.nas.2.0.co.2.xml

Murphy, A.H. (1973) A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12, 595–600. Available from: https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450.1973.012.0595.anvpot.2.0.co.2.xml

Niculescu-Mizil, A. & Caruana, R. (2005) Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*. New York, NY, USA: Association for Computing Machinery. URL, pp. 625–632. Available from: https://doi.org/10.1145/1102351.1102430

Rädler, A.T., Groenemeijer, P.H., Faust, E., Sausen, R. & Púčik, T. (2019) Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. *npj Climate and Atmospheric Science*, 2, 30.

Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraff, C. et al. (2020) DWD database reference for the global and regional icon and icon-eps forecasting system. Technical report Version 2.1. 8, Deutscher Wetterdienst Available from: https://www.dwd.de/DWD/forschung/nwv/fepub/icon.database.main.pdf

Roberts, N. (2008) Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, 15, 163–169. Available from: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.57

Selz, T. & Craig, G.C. (2015) Upscale error growth in a high-resolution simulation of a summertime weather event over Europe. *Monthly Weather Review*, 143, 813–827. Available from: https://journals.ametsoc.org/view/journals/mwre/143/3/mwr-d-14-00140.1.xml

Sobash, R.A., Romine, G.S. & Schwartz, C.S. (2020) A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Weather and Forecasting*, 35, 1981–2000. Available from: https://journals.ametsoc.org/view/journals/wefo/35/5/wafD200036.xml

Sun, J., Xue, M., Wilson, J.W., Zawadzki, I., Ballard, S.P., Onvlee-Hooimeyer, J. et al. (2014) Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95, 409–426. Available from: https://journals.ametsoc.org/view/journals/bams/95/3/bams-d-11-00263.1.xml

Sun, Y., Wong, A.K. & Kamel, M.S. (2009) Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23, 687–719.

Toth, Z., Talagrand, O., Candille, G. & Zhu, Y. (2003) Probability and ensemble forecasts. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Vol. 137, p. 163. Chichester: Wiley.

Turner, B.J., Zawadzki, I. & Germann, U. (2004) Predictability of precipitation from continental radar images. Part III: Operational nowcasting implementation (maple). *Journal of Applied Meteorology*, 43, 231–248. Available from: https://journals.ametsoc.org/view/journals/apme/43/2/1520-0450.2004.043.0231.popfcr.2.0.co.2.xml

Ukkonen, P. & Mäkelä, A. (2019) Evaluation of machine learning classifiers for predicting deep convection. *Journal of Advances in Modeling Earth Systems*, 11, 1784–1802. Available from: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001561

Veraverbeke, S., Rogers, B.M., Goulden, M.L., Jandt, R.R., Miller, C.E., Wiggins, E.B. et al. (2017) Lightning as a major driver of recent large fire years in north American boreal forests. *Nature Climate Change*, 7, 529–534.

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, 3rd edition. Amsterdam, Heidelberg: Elsevier Acad. Press.

Wilson, J.W., Crook, N.A., Mueller, C.K., Sun, J. & Dixon, M. (1998) Nowcasting thunderstorms: A status report. *Bulletin of the American Meteorological Society*, 79, 2079–2100. Available from: https://journals.ametsoc.org/view/journals/bams/79/10/1520-0477.1998.079.2079.ntasr.2.0.co.2.xml

Yano, J.-I., Ziemiański, M.Z., Cullen, M., Termonia, P., Onvlee, J., Bengtsson, L. et al. (2018) Scientific challenges of convective-scale numerical weather prediction. *Bulletin of the American Meteorological Society*, 99, 699–710. Available from: https://journals.ametsoc.org/view/journals/bams/99/4/bams-d-17-0125.1.xml

Yasuda, Y., Yokoyama, S., Minowa, M. & Satoh, T. (2012) Classification of lightning damage to wind turbine blades. *IEEJ Transactions on Electrical and Electronic Engineering*, 7, 559–566.

Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/tee.21773

Zängl, G., Reinert, D., Rípodas, P. & Baldauf, M. (2015) The ICON (ICOsahedral non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, 141, 563–579. Available from: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2378

Zhou, K., Sun, J., Zheng, Y. & Zhang, Y. (2022) Quantitative precipitation forecast experiment based on basic NWP variables using deep learning. *Advances in Atmospheric Sciences*, 39, 1472–1486.

How to cite this article: Vahid Yousefnia, K., Bölle, T., Zöbisch, I. & Gerz, T. (2024) A machine-learning approach to thunderstorm forecasting through post-processing of simulation data. *Quarterly Journal of the Royal Meteorological Society*, 150(763), 3495–3510. Available from: https://doi.org/10.1002/qj.4777