

Inferring Thunderstorm Occurrence from Vertical Profiles of Convection-Permitting Simulations: Physical Insights from a Physical Deep Learning Model

KIANUSCH VAHID YOUSEFNIA¹, CHRISTOPH METZL,^a AND TOBIAS BÖLLE^a

^a *Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany*

(Manuscript received 30 September 2024, in final form 24 April 2025, accepted 6 June 2025)

ABSTRACT: Thunderstorms have significant social and economic impacts due to heavy precipitation, hail, lightning, and strong winds, necessitating reliable forecasts. Thunderstorm forecasts based on numerical weather prediction (NWP) often rely on single-level surrogate predictors, like convective available potential energy and convective inhibition, derived from vertical profiles of three-dimensional atmospheric variables. In this study, we develop Signature-Based Approach of Identifying Lightning Activity using Machine Learning (SALAMA) 1D, a deep neural network which directly infers the probability of thunderstorm occurrence from vertical profiles of 10 atmospheric variables, bypassing single-level predictors. By training the model on convection-permitting NWP forecasts, we allow SALAMA 1D to flexibly identify convective patterns, with the goal of enhancing forecast accuracy. The model's architecture is physically motivated: Sparse connections encourage interactions at similar height levels while keeping model size and inference times computationally efficient, whereas a shuffling mechanism prevents the model from learning nonphysical patterns tied to the vertical grid. SALAMA 1D is trained over central Europe with lightning observations as the ground truth. Comparative analysis against a baseline machine learning model that uses single-level predictors shows SALAMA 1D's superior skill across various metrics and lead times of up to at least 11 h. Moreover, expanding the archive of forecasts from which training examples are sampled improves skill, even when the training set size remains constant. Finally, a sensitivity analysis using saliency maps indicates that our model relies on physically interpretable patterns consistent with established theoretical understanding, such as ice particle content near the tropopause, cloud cover, conditional instability, and low-level moisture.

SIGNIFICANCE STATEMENT: This work aims to improve thunderstorm forecasting by applying machine learning to vertical atmospheric profiles from numerical weather prediction. We developed a model that incorporates physical considerations, resulting in more accurate yet computationally efficient predictions compared to conventional methods. Additionally, the model provides insights into how it identifies thunderstorm occurrence, fostering interpretability and trust. Our research demonstrates how to enhance the skill of machine learning systems in severe weather forecasting, which is crucial for supporting timely, informed decision-making in situations that impact public safety and the economy.

KEYWORDS: Deep convection; Thunderstorms; Numerical weather prediction/forecasting; Postprocessing; Machine learning; Model interpretation and visualization

1. Introduction

Thunderstorms can have devastating impacts on society and the economy due to accompanying phenomena such as lightning, intense precipitation (including graupel and hail), and strong winds. Severe thunderstorms may lead to flash floods, which can result in significant damage and endanger lives (Ntelekos et al. 2007; Piper et al. 2016). Additionally, thunderstorms cause harm to crops and livestock, resulting in considerable economic losses (Holle 2014). Moreover, thunderstorm occurrence affects aviation and logistics, causing costly delays and heightened safety risks (Gerz et al. 2012; Borsky and Unterberger 2019). Although the global impact of climate change on thunderstorm frequency remains highly uncertain and varies regionally, studies suggest that thunderstorms will become more frequent in many European countries (Diffenbaugh et al. 2013; Rädler et al. 2019; Taszarek et al. 2021). This trend underscores the growing importance of

accurate and timely forecasts of thunderstorm occurrence in the future.

Thunderstorm forecasts based on the extrapolation of remote sensing data (nowcasts) become less reliable beyond an hour (Leinonen et al. 2023); therefore, numerical weather prediction (NWP) is routinely used for the longer lead times required in many critical decision-making processes. Invoking mathematical models and current observations, NWP simulates the future atmospheric state encoded in terms of meteorological variables (Bauer et al. 2015; Palmer 2017). In this study, we use a convection-permitting NWP model, which allows convection to be resolved without convection parameterizations (Yano et al. 2018).

The identification of thunderstorms in NWP model output is complicated by the fact that no single variable directly indicates thunderstorm occurrence. Instead, one simultaneously considers multiple predictors, which act as surrogate indicators of thunderstorm occurrence (Sobash et al. 2011; Kober et al. 2012; Simon et al. 2018). These predictors are motivated by a combination of experience, physical models, and domain knowledge. Examples include convective available potential energy (CAPE),

Corresponding author: Kianusch Vahid Yousefnia, kianusch.vahidyousefnia@dlr.de

precipitation rate, and relative humidity at 700 hPa. The occurrence of thunderstorms is then inferred from these surrogate fields based on the forecasters' expertise.

Recently, machine learning (ML) techniques have become increasingly popular for this purpose, leveraging methods such as fuzzy logic (Lin et al. 2012; Li et al. 2021), random forests (Herman and Schumacher 2018; Loken et al. 2020), and neural networks (Sobash et al. 2020; Geng et al. 2021; Zhou et al. 2022; Jardines et al. 2024), with the latter often proving more effective (Herman and Schumacher 2018; Ukkonen and Mäkelä 2019). ML involves training algorithms to recognize patterns and make predictions based on labeled data, which serve as the ground truth. For thunderstorm forecasting, these labels are often obtained from observational data sources such as lightning detection networks (Ukkonen and Mäkelä 2019; Geng et al. 2021) and radar imagery (Gagne et al. 2017; Burke et al. 2020). We recently introduced the feedforward neural network model Signature-Based Approach of Identifying Lightning Activity using Machine Learning (SALAMA), which infers the probability of thunderstorm occurrence from 21 NWP predictors related to thunderstorm activity, outperforming classification based only on NWP reflectivity for lead times up to at least 11 h (Vahid Yousefnia et al. 2024).

The predictors that have been used by ML models for inferring thunderstorm occurrence are single-level variables; i.e., they assign a single value to each horizontal grid point on the NWP model domain. Most single-level predictors can be derived from the vertical profiles of three-dimensional meteorological variables. For example, CAPE is determined by the vertical profiles of pressure, temperature, and specific humidity (Markowski and Richardson 2010). There are additional processes and effects outside the scope of vertical profiles, e.g., orography, or solar radiation, which are possible sources of lift which forecasters consider when predicting convection initiation. In what follows, however, we restrict ourselves to single-level predictors that are derived from vertical profiles. Importantly, the information contained in a set of such single-level predictors is inherently present in the corresponding vertical profiles, though in a more complex, encoded form. Consequently, an ML model trained directly on vertical profiles should, at a minimum, match the performance of one trained on single-level predictors. Given that a model based on vertical profiles has greater flexibility to detect patterns in NWP data, we hypothesize that such a model would outperform one using only single-level predictors—this is the primary focus of the present study.

In this work, we introduce SALAMA 1D, a deep neural network trained on vertical profiles of three-dimensional meteorological variables from a convection-permitting NWP model. We thereby—and in contrast to the original SALAMA model, to which we refer as SALAMA 0D in this work—bypass the conventional use of single-level predictors. To our knowledge, this is the first study to apply neural networks directly to vertical profiles of convection-permitting forecasts for the purpose of predicting thunderstorm occurrence. Just like for the original model, we use data from the Icosahedral Nonhydrostatic-D2-Ensemble Prediction System (ICON-D2-EPS) (Zängl et al. 2015; Reinert et al. 2020), a convection-permitting NWP ensemble model for central Europe with a horizontal resolution of ~ 2 km and

65 vertical levels, operationally run by the German Meteorological Service (Deutscher Wetterdienst) (DWD). Lightning observations from the lightning detection network (LINET) (Betz et al. 2009) serve as the ground truth. On the other hand, processing vertical profiles instead of single-level predictors requires an adjustment of the SALAMA 0D architecture to account for the increased input dimensionality. To keep model complexity in check, we are guided by physical principles in the design of SALAMA 1D. Specifically, a sparse layer reduces the number of parameters by promoting interactions within the same height levels, while a shuffling mechanism discourages the model from learning patterns related to the vertical grid structure. This shuffling also acts as a form of regularization, helping to prevent overfitting. Our results demonstrate that SALAMA 1D achieves superior performance compared to SALAMA 0D, highlighting the advantage of training directly on vertical profiles.

A major issue of ML models concerns the interpretability of their output (Flora et al. 2024; Yang et al. 2024; Dramsch et al. 2025). In this work, we conduct a sensitivity analysis using saliency maps to explore how the model processes the input vertical profiles. As this analysis will reveal, our model rediscovers patterns that align with established theoretical understandings of thunderstorm development and occurrence.

Our work demonstrates the potential of applying deep neural network models directly to complex, minimally processed input data, as opposed to relying on lower-dimensional, feature-engineered inputs fed into shallow ML models. Furthermore, we illustrate how incorporating physical constraints can enhance the model's robustness and computational efficiency, while also emphasizing the importance of gaining interpretability—and thus trust—through techniques such as saliency maps. These advancements are particularly crucial for enabling ML models to reliably assist in critical decision-making processes in severe weather forecasting.

The structure of this work is as follows: In section 2, we provide details on the NWP data used, the lightning observations, and the methodology for compiling ML training sets. Section 3 describes our ML model, including its architecture and training process. In section 4, we report our results, including the sensitivity analysis, while section 5 summarizes our work and elaborates on its implications with possible future research avenues.

2. Data

Our objective is to train an ML model—SALAMA 1D—for inferring the conditional probability of thunderstorm occurrence given an array $\xi \in \mathbb{R}^{N_t \times N_z}$ of N_t meteorological variables at N_z height levels. As a baseline for comparison, we will also train a separate ML model (SALAMA 0D; section 3c) which infers the probability of thunderstorm occurrence from $N_t = 21$ single-level meteorological variables. The ML task is known as binary classification. ML models for this task are trained with datasets of examples. In our case, an example is a tuple (ξ, y) , where the label $y \in \{0, 1\}$ denotes the class to which the sample ξ belongs (class 1: thunderstorm occurrence and class 0: no thunderstorm occurrence). We first collect an

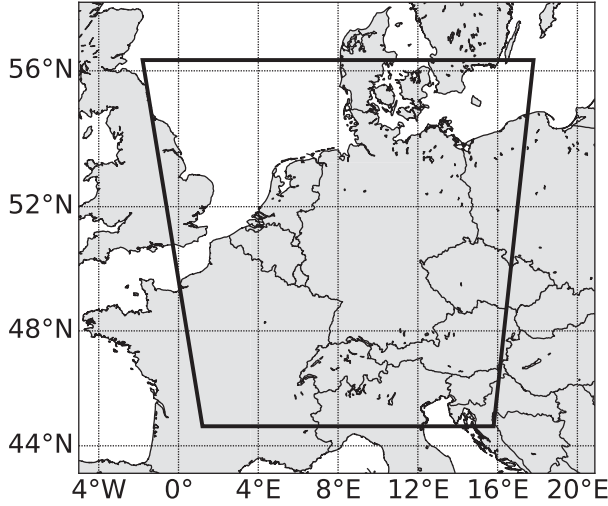


FIG. 1. Study region for this work, shown in a parallel projection. The polygon vertices are listed counterclockwise from the bottom left: (44.7°N, 1.2°E), (44.7°N, 15.8°E), (56.3°N, 17.8°E), and (56.3°N, 1.8°E).

archive of NWP forecasts of ξ (section 2a) and the corresponding labels y from lightning observations (section 2b). We then randomly draw examples from the archive to compile datasets for training, validation, and testing (section 2c).

Data are collected for a region which encompasses Germany, as well as parts of its neighboring countries, as shown in Fig. 1. This region roughly corresponds to the NWP model domain, which we cropped at the borders by approximately 80 km to reduce boundary computation errors.

a. NWP data

The operational runs of ICON-D2-EPS are initialized eight times daily, starting at 0000 UTC, and produce forecasts with a time resolution of 1 h. For each full hour of the day, we collect only the latest (and, therefore, most accurate) available forecast, which is at most 2 h old. We gather forecasts for several summer months in 2021 (June–August), 2022 (May–July), and 2023 (July–August). ICON-D2-EPS comprises 20 ensemble members, which reflect the NWP uncertainty in the initial conditions, model error, and boundary conditions (Reinert et al. 2020). We collect the forecasts for all members of the ensemble system.

For SALAMA 1D, we extract the N_f variables given in Table 1 from the forecasts; these variables correspond to the fields which are operationally available in ICON-D2-EPS on vertical levels. We keep the fields on their native grid (vertically: $N_z = 65$ nonequidistant levels and horizontally: spherical triangles) to avoid interpolation errors. For SALAMA 0D, we extract the corresponding single-level predictors (Vahid Yousefnia et al. 2024).

b. Lightning observations

To reconstruct the timing and location of past thunderstorm occurrences in our study region, we consult observation

TABLE 1. (Instantaneous) vertical ICON-D2-EPS field profiles used in this study.

ICON variable	Description
U	Zonal wind speed
V	Meridional wind speed
T	Temperature
P	Pressure
QV	Specific humidity
QC	Cloud water mixing ratio
QI	Cloud ice mixing ratio
QG	Graupel mixing ratio
CLC	Cloud cover
W	Vertical wind speed

data from the LINET (Betz et al. 2009). We choose lightning observations as the ground truth for thunderstorm occurrence due to their high and uniform detection efficiency ($\geq 95\%$ for LINET) and spatial accuracy (150 m for LINET), similarly to Ukkonen and Mäkelä (2019) and Vahid Yousefnia et al. (2024).

Given an NWP grid point at horizontal position \mathbf{x} and time t , we consider a thunderstorm to occur at (\mathbf{x}, t) if a flash of lightning is detected at any \mathbf{x}_i, t_i with

$$\|\mathbf{x} - \mathbf{x}_i\| < \Delta r \quad \text{and} \quad |t - t_i| < \Delta t, \quad (1)$$

where $\|\cdot\|$ denotes the great-circle distance between \mathbf{x} and \mathbf{x}_i on a perfect sphere of Earth with a radius of 6371.229 km, as assumed in the ICON-D2-EPS model (Reinert et al. 2020). The spatial and temporal thresholds used in this study read $\Delta r = 15$ km and $\Delta t = 30$ min.

c. ML dataset compilation

At this stage, we have gathered a dataset archive of tuples (ξ, y) . Here, ξ denotes the NWP data for a particular ensemble member at a particular grid point and time and y denotes the corresponding ground truth. We compile datasets for training, validation, and testing, by randomly drawing tuples from the archive. We provide a summary of the compiled datasets in Table 2 (the model configuration in the table will be introduced in this section). When compiling the datasets, there are two issues to be taken into consideration, as we will discuss next.

First, to reduce correlations between the datasets, we ensure temporal separation, which is common practice (Ravuri et al. 2021; Geng et al. 2021; Jardines et al. 2024). To this end, the data of 2021 and 2022 are used for training, even days of 2023 are used for testing and odd days of 2023 are used for validation. To partition the data from 2023, we let each day begin at 0800 UTC, which we identified to be the hour of least lightning activity in our observations. The reason for shifting the start of the day is to minimize the risk of dataset correlations caused by thunderstorms which persist after 0000 UTC (Vahid Yousefnia et al. 2024).

The second issue that arises is high-class imbalance (observing the class “thunderstorm occurrence” is climatologically less likely than observing the opposite class). We estimate the relative frequency g of thunderstorm occurrence expected for

TABLE 2. Summary of the datasets used for training, testing, and validation of the ML model configurations used in this study.

	SALAMA 1D-2021	SALAMA 1D-2022	SALAMA 0D
Training set			
Time period	June–August 2021	June–August 2021, May–July 2022	June–August 2021
Dataset size	4×10^5	4×10^5	4×10^5
Class imbalance	1:1	1:1	1:1
Validation set			
Time period	July–August 2023 (odd days)	July–August 2023 (odd days)	July–August 2023 (odd days)
Dataset size	10^5	10^5	10^5
Class imbalance	Climatologically consistent	Climatologically consistent	Climatologically consistent
Test set			
Time period	July–August 2023 (even days)	July–August 2023 (even days)	July–August 2023 (even days)
Dataset size	10^5	10^5	10^5
Class imbalance	Climatologically consistent	Climatologically consistent	Climatologically consistent

the test set by evaluating lightning observations for July and August from 2018 to 2022 as follows: We first assign labels to the observations by applying Eq. (1) for each full hour and each grid point in the study region. Next, we average over the study region and each month, obtaining 10 samples of g (one for each July and August of the 5 years above). Finally, we generate 200 bootstrap resamples of the 10-sample dataset, which yields $g = 1.93^{+0.23}_{-0.24} \times 10^{-2}$ as estimate of the median and the symmetric 90% confidence interval. The problem with such a small value of g is that the model might not see enough examples of the minority class during training to learn meaningful patterns. To ensure that the model is presented with sufficiently many positive examples, we compile a class-balanced training set, which contains an equal number of positive and negative examples. We arrange class balance by randomly drawing tuples from the archive and keeping a negative example only if the current number of negative examples does not exceed half of the dataset target size. Thereby, we randomly under-sample the majority class, as is common practice (Hasanin and Khoshgoftar 2018; Mohammed et al. 2020; Vahid Yousefnia et al. 2024). The validation and test sets, however, are compiled in a climatologically consistent manner. The reason for this choice is to ensure that we evaluate our model in a realistic setting in which thunderstorms rarely occur. When a model trained on balanced data is used with climatologically consistent datasets, we need to calibrate the raw model output p' using the following formula to obtain a well-calibrated probability p (Vahid Yousefnia et al. 2024):

$$p = \frac{gp'}{gp' + (1 - g)(1 - p')}. \quad (2)$$

Finally, to examine whether extending the study period from which to gather examples enhances skill, we compile two training sets for SALAMA 1D: In addition to a training set consisting of examples from 2021 and 2022 (yielding a model configuration to which we refer as SALAMA 1D-2022), we compile a second training set made up of examples from only 2021 (SALAMA 1D-2021). The models are tested, however, on the same test set. The training set for the baseline model is equally made up of examples from only 2021, making it readily comparable to SALAMA 1D-2021. Since SALAMA 0D is

trained on a different set of atmospheric variables, we cannot use the same test set as for the SALAMA 1D models. To ensure comparability, we compile the baseline model test set such that each sample corresponds to the same forecast, retrieved at the same grid point and from the same ensemble member, as its counterpart in the SALAMA 1D test set. This guarantees that any differences in model performance can be attributed to the choice of input variables.

3. Methods

In this section, we present the architecture used for SALAMA 1D and give training details. In addition, we summarize the main aspects of our baseline model SALAMA 0D, which infers thunderstorm occurrence from derived single-level predictors.

a. Model description

Given an input sample ξ of NWP predictors (for a given member, grid point, and forecast time), we aim to develop a model that predicts the corresponding probability of thunderstorm occurrence. Our model constitutes a lead-time-independent postprocessing framework for NWP forecasts. To obtain, for example, an 8-h forecast of the probability of thunderstorm occurrence, one needs to apply our model to an 8-h forecast of NWP predictors. While we train on forecasts with a lead time of at most 2 h, we examine in section 4 whether the patterns learned by the ML model generalize to longer lead times.

We use an artificial neural network model $\mathbb{R}^{N_f \times N_z} \rightarrow (0, 1)$ to describe the relationship between the input sample ξ and the corresponding probability of thunderstorm occurrence. The architecture of SALAMA 1D, as illustrated in Fig. 2, combines dense layers with a sparse layer strategically designed to reduce the number of parameters. This approach addresses challenges such as overfitting and the high computational demands typically associated with large ML models. Instead of using the pruning technique (LeCun et al. 1989; Frankle and Carbin 2019), we incorporate physical aspects and symmetry considerations to achieve a reduction in parameters. Because translational symmetry is broken along the z direction, weight sharing, as in convolutional layers, cannot be applied effectively. Instead, we implement sparse connections, allowing interactions

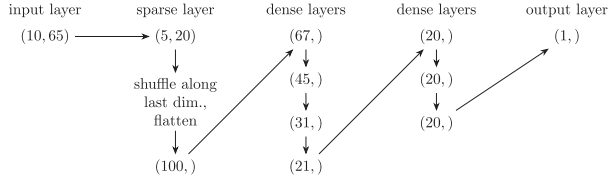


FIG. 2. Change in input size during a forward pass in SALAMA 1D. The sparse layer (section 3b) reduces dimensionality and shuffles the data to prevent the model from learning dependencies tied to the vertical grid structure. Additionally, the shuffling acts as a regularization technique, helping to limit overfitting. Input fields are scaled to order 1. We use rectified linear units as activation functions after the flattened sparse layer and each dense layer and a sigmoid function to map the output layer to the open interval (0, 1). The sparse layer has 8100 trainable parameters, and the other layers add 13 226 parameters. SALAMA 1D is lightweight with a computational complexity (Sovrasov 2018) of roughly 22kMAC (multiply accumulate operations). SALAMA 0D (section 3c) requires 1.3kMAC.

only between field values at similar height levels (section 3b). Dense layers further downstream then construct dependencies between more distant field values. Additionally, we introduce a shuffling mechanism to ensure that the model does not rely on the vertical grid structure, forcing it to infer vertical orientation from the data itself. This design allows the model to, for instance, associate the formation of ice particles with the height of the tropopause temperature inversion rather than a fixed height level, such as level 11. It turns out that shuffling also regularizes the model, limiting overfitting issues further.

Training is then performed analogously to Vahid Yousefnia et al. (2024): Evaluating the training set, we scale the input fields to have zero mean and unit variance before minimizing binary cross-entropy loss via the Adam optimizer (Kingma and Ba 2014). Using minibatches of size 1000, we train for

300 epochs. After training, we inspect the validation loss as a function of epoch and select the smallest epoch for which the loss no longer decreases. Equation (2) is applied with the sample climatology value found in section 2c whenever the model is used on climatologically consistent datasets.

b. Sparse connections

In this section, we provide technical details on the sparse layer implementation, an illustration of which is provided by Fig. 3. The input layer is given by an array of shape (N_f, N_z) with a field dimension (iterating over the N_f field profiles) and a height dimension (iterating over the N_z vertical levels). Now, we consider a block of shape (N_f, k) , where k is the size of the block along the height dimension. We densely connect the nodes within this block to h nodes in the following layer. Next, we slide the block by s nodes along the height dimension and, again, densely connect the corresponding nodes to h subsequent nodes of the following layer. Starting with a block at the bottom of the input layer, we repeat this procedure until reaching the top of the input layer. Provided that $N_z - k$ is divisible by s , this procedure leads to $N_k = (N_z - k + s)/s$ blocks and produces $N_k \times h$ nodes in the following layer. We incorporate a shuffling mechanism that randomly permutes the order of the blocks for each example during training.

In contrast to a convolutional layer with a sliding kernel, all N_k blocks in our sparse layer have their own set of free parameters. In total, the sparse layer contributes $N_k \times (N_f \times k + 1) \times h$ parameters to the model. We have studied a large variety of (k, s, h) combinations and found that skill depends barely on the particular sliding block configuration as long as a sufficiently large number of parameters are exceeded. Setting $k = 8$, $s = 3$, and $h = 5$, which corresponds to the smallest model configuration with saturating skill, we

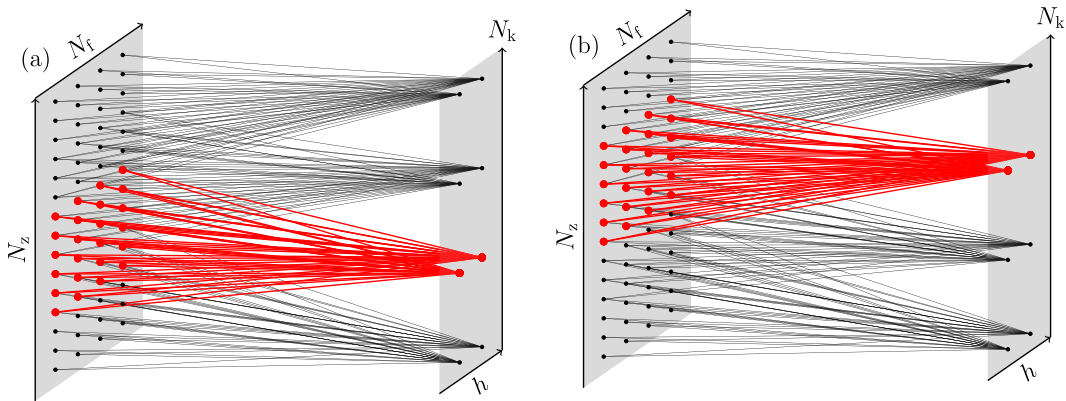


FIG. 3. Illustration of the connections (lines between bold dots) between the input layer of shape (N_f, N_z) and the following layer of shape (h, N_k) . (a) A block of nodes of shape (N_f, k) in the input layer is connected to a row of h nodes in the following layer. (b) Then, the block is shifted upward by s nodes and rewired with the next row of h nodes of the following layer. The input layer is shown in two dimensions to help visualize each vertical field profile, but there is no spatially extended structure beyond the vertical z direction. Equally, we show the following layer in two dimensions to illustrate the yield of each group of connections in a separate row; however, this layer is flattened further downstream (Fig. 2). For better readability, the layer sizes and hyperparameter settings used in this illustration do not correspond to those in the actual model.

obtain $N_k = 20$ blocks and 8100 trainable parameters. In comparison with a fully dense layer, the parameter size is reduced by around 90%.

c. Baseline model

To study the potential benefit of postprocessing vertical NWP profiles instead of derived single-level predictors, we implement the ML model from [Vahid Yousefnia et al. \(2024\)](#), which we refer to as SALAMA 0D in the following (in contrast to SALAMA 1D, which processes one-dimensional vertical profiles). SALAMA 0D uses a set of 21 established predictors related to thunderstorm activity to infer calibrated probabilities of thunderstorm occurrence. The predictors are mapped to a single node via three dense layers with 20 nodes each. The architecture of SALAMA 0D is the result of a hyperparameter study with increasingly many layers and nodes per layer until we observed saturation in terms of validation loss ([Vahid Yousefnia et al. 2024](#)). Therefore, SALAMA 0D skill is not limited by model architecture, and any difference in skill between SALAMA 1D and SALAMA 0D can be attributed to input data differences. Furthermore, both models operate point by point, which allows for a direct and fair comparison. We train SALAMA 0D with training data from 2021 which matches the NWP model, the study region, and study period, used for SALAMA 1D ([section 2](#)). In contrast to the original paper, we feed the NWP data to SALAMA 0D on its native horizontal grid, reducing potential interpolation errors.

4. Results

We intend to investigate two aspects concerning the model skill of SALAMA 1D. First, we compare SALAMA 1D to the introduced baseline model, studying the potential benefit of considering vertical profiles (instead of derived single-level predictors) for the prediction of thunderstorm occurrence. On the other hand, we are interested in examining whether extending the study period from which to gather training examples enhances skill. Therefore, we show the results for two configurations of SALAMA 1D:

- SALAMA 1D-2021: SALAMA 1D, trained with data from 2021.
- SALAMA 1D-2022: SALAMA 1D, trained with data from 2021 and 2022.

The baseline model, SALAMA 0D, has been trained with data from 2021; it can, therefore, be readily compared with SALAMA 1D-2021. Note that while SALAMA 1D-2022 is trained on data from a longer study period than the two other models, the size of the training set does not change. Potential improvements in skill with respect to SALAMA 1D-2021 can, therefore, be unambiguously attributed to increased data variability. The results from the model comparison are given in [section 4a](#). Furthermore, we study how sensitively SALAMA 1D-2022 reacts to small input changes. The results, shown in [section 4b](#), offer insight into how the model infers thunderstorm occurrence from the input.

a. Model comparison

To get a first idea of the skill of the three models, we consider two cases with thunderstorm activity in central Europe, namely, 1700 UTC 24 July 2023 (case A) and 1200 UTC 2 August 2023 (case B). These two cases were chosen since they display multiple simultaneous convective regions of varying size. In [Fig. 4](#), we show maps of the probability of thunderstorm occurrence for central Europe as produced by the three models and compare them with lightning observations. The probability maps have been computed by retrieving the latest NWP forecast for each target time (case A: the 2-h forecast of the 1500 UTC model run and case B: 0-h forecast of the 1200 UTC run) and applying the SALAMA models to them. The ML models can be applied to all NWP ensemble members individually, producing separate probability output for each member. We show the results for only a single member. For both cases, we also show raw NWP output to see where the NWP model likely produces convection. To this end, we consider the column-maximal radar reflectivity product of ICON-D2-EPS. Specifically, for a given pixel, we compute the fraction of pixels within a radius of 15 km which exceed a threshold of 37 dBZ (e.g., [Theis et al. 2005](#); [Roberts and Lean 2008](#)). Exceedance probabilities of reflectivity with thresholds between 30 and 40 dBZ have also been used in previous studies to identify thunderstorm occurrence (e.g., [Mueller et al. 2003](#); [Leinonen et al. 2022](#); [Ortland et al. 2023](#)).

Case A is characterized by intense thunderstorm activity from the Alps to northern Germany, with roughly 10 convective objects of different sizes. Most lightning contours are predicted by all three models. However, SALAMA 0D produces a significant number of false alarms. SALAMA 1D-2021 corrects many of them, especially in southern Germany. SALAMA 1D-2022 tends to make its predictions more confidently than the other models, resulting in more contours that are filled out with high-probability pixels. On the other hand, the model seems to produce slightly more false alarms than SALAMA 1D-2021.

Thunderstorm activity in case B occurs primarily over the Benelux, while two smaller thunderstorms are observed over the North Sea. The latter two events are missed by the three models, though SALAMA 1D-2022 only misplaces the storms toward the south. The thunderstorm over the Benelux is captured to some extent by all the models. However, the SALAMA 1D models are more confident in their predictions, producing high-probability pixels almost everywhere within the thunderstorm contour. On the other hand, they overestimate the size of the thunderstorm, resulting in false alarms directly outside the contour. SALAMA 0D predicts a wide band of thunderstorm activity over France and southwestern Germany, which was not confirmed by lightning observations. This region of false alarms is significantly reduced by the two SALAMA 1D models, with SALAMA 1D-2022 reducing the region to essentially zero.

The ML models, overall, align with raw NWP structures, with the highest ML probability output being collocated with a high likelihood of exceeding 37 dBZ. On the other hand, the ML models tend to correct the areal size of simulated convection, with SALAMA 1D-2022 producing the least false alarms. Remarkably, the SALAMA 1D models can also produce

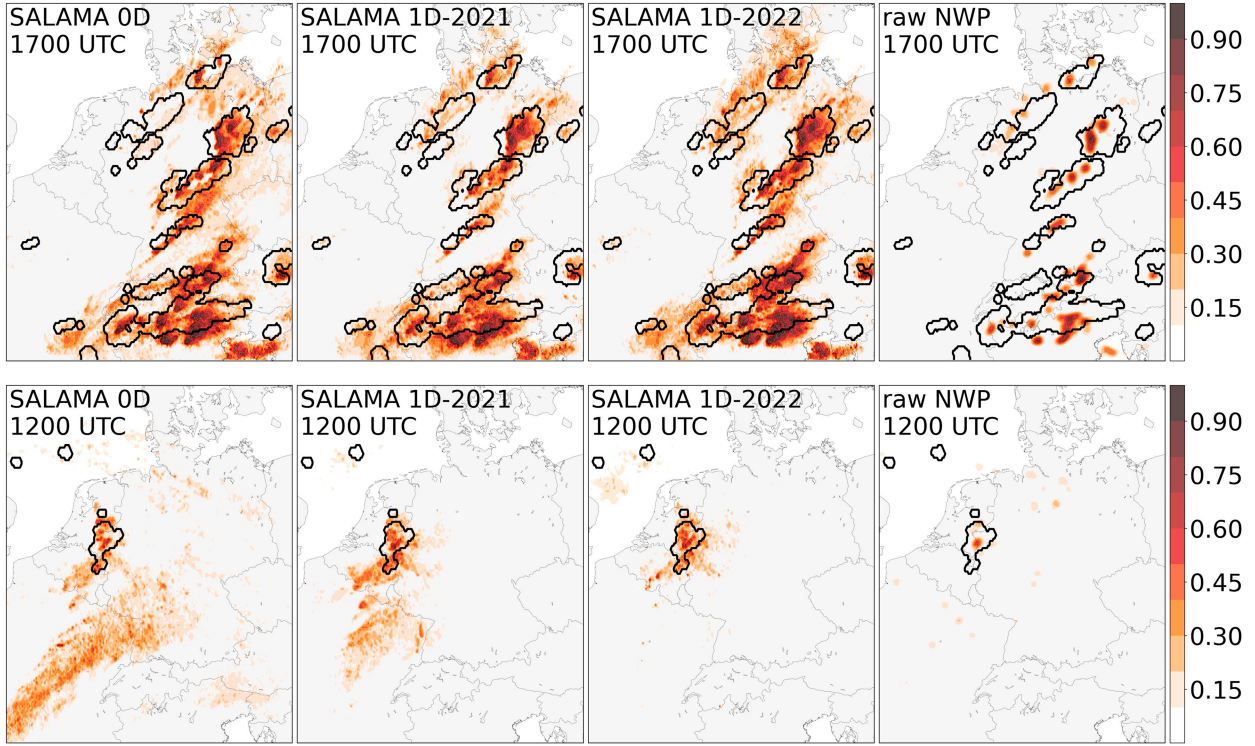


FIG. 4. Model probability of thunderstorm occurrence for the three models of this study, evaluated for (top) 1700 UTC 24 Jul 2023 and (bottom) 1200 UTC 2 Aug 2023. The filled contours with varying shading display the result for the first ensemble member of ICON-D2-EPS, whereas lightning labels (section 2b) are shown as black contours. None of the dates have been used for training. NWP forecast lead times are 2 h for (top) and 0 h for (bottom) (noting that 0-h forecasts have limited operational utility, as they become available only after the valid time has passed). (fourth column) The probability of exceeding a reflectivity threshold of 37 dBZ for the first ensemble member of ICON-D2-EPS. To this end, we compute for each pixel the fraction of pixels within a radius of 15 km at which the column-maximal radar reflectivity product of ICON-D2-EPS exceeds the threshold.

high-probability output when lightning occurred but no convection has been triggered in the NWP model, as can be seen for the lightning regions over France for case A, which suggests that our ML models, SALAMA 1D-2022 in particular, may be able to correct for NWP model biases.

To compare the models quantitatively, we use the test set from section 2c. The first instrument which we use for comparing model skill is a reliability diagram (Wilks 2011; Bröcker and Smith 2007). Partitioning the range (0, 1) of possible model probabilities into N_b equidistant bins, we distribute the test examples among the bins according to the model probability they have been assigned. For each bin $i = 1, 2, \dots, N_b$, we extract the observed relative frequency \bar{o}_i of thunderstorm occurrence, the bin-averaged model probability p_i , and the number N_i of examples per bin. A reliability diagram then consists of a calibration function and a refinement distribution. The calibration function is a plot of \bar{o}_i against p_i and measures whether the model probabilities are consistent with the observed relative frequency of thunderstorm occurrence, a characteristic known as reliability. A well-calibrated model exhibits a calibration function close to the 1:1 diagonal. The refinement distribution corresponds to the distribution of model probabilities. Skillful models are capable of producing well-calibrated model probabilities larger than climatology, which is referred to as resolution.

In the upper panels in Fig. 5, we show for each of our models the corresponding reliability diagram with $N_b = 10$ bins. All models display a similar degree of high reliability. We reiterate here that it is important to apply the analytic model calibration (2); otherwise, high reliability could not be expected. The refinement distributions, as well, look similar. However, the model resolutions differ significantly: Following Vahid Yousefnia et al. (2024), we introduce the binwise contributions to resolution and reliability:

$$\text{RES}_i = \frac{1/\Delta p}{g(1-g)} \frac{N_i}{N} (p_i - g)^2, \quad (3)$$

$$\text{REL}_i = \frac{1/\Delta p}{g(1-g)} \frac{N_i}{N} (p_i - \bar{o}_i)^2, \quad (4)$$

where $\Delta p = 1/N_b$ denotes bin width. The binwise contributions of RES_i to resolution are positively oriented (“the higher, the better”), while the contributions of REL_i to reliability are negatively oriented. In consequence, the area between RES_i and REL_i as a function of p_i serves as a positively oriented measure of skill. According to Murphy (1973), this area is equivalent to the Brier skill score (BSS) if a random model based on climatology (section 2c) is applied as a reference. The lower

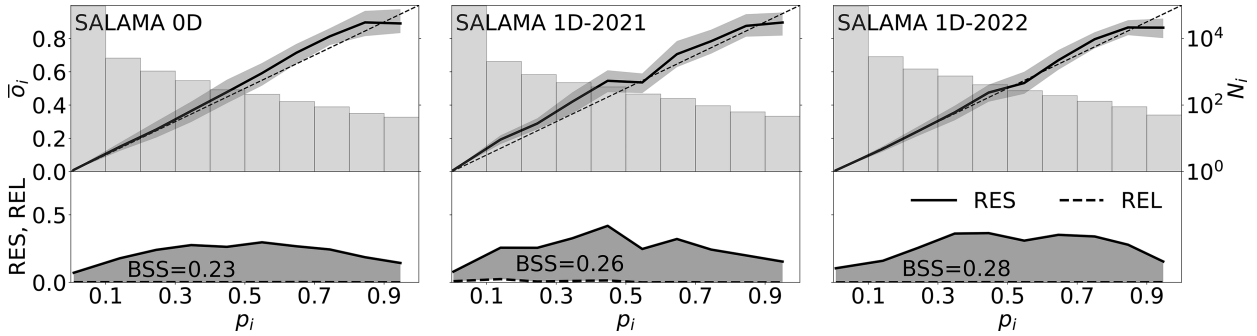


FIG. 5. Reliability diagram for (left) SALAMA 0D, (center) SALAMA 1D-2021, and (right) SALAMA 1D-2022. (top) The calibration curve and the refinement distribution and (bottom) binwise resolution and reliability [Eqs. (3) and (4)]. The uncertainty on the calibration curve is obtained from 10^4 bootstrap resamples, using daywise block resampling, and shows the symmetric 90% confidence interval. The area enclosed by binwise resolution and reliability corresponds to the BSS with climatology as reference. The BSS differences between the models are significant on a 90% confidence level, as we check in Table 3.

panels in Fig. 5 display binwise reliability and resolution, revealing that the increase in skill, measured by the BSS, for the two SALAMA 1D models is attributed to improved resolution. For SALAMA 1D-2021, both low- and high-probability examples enhance resolution, while for SALAMA 1D-2022, all bins with $p_i > 0.3$ contribute additional improvements to resolution.

In Table 3, we summarize the performance of the three models using several additional skill scores established for problems with class imbalance (Wilks 2011; Saito and Rehmsmeier 2015; Vahid Yousefnia et al. 2024). These scores are positively oriented and bounded by unity. Across all skill scores, the SALAMA 1D models consistently outperform SALAMA 0D, with SALAMA 1D-2022 showing higher skill than SALAMA 1D-2021.

So far, we have worked with a test set that consists of examples from NWP forecasts with a lead time of at most 2 h (section 2a). Next, we examine systematically how model skill depends on NWP forecast lead time. For this purpose, we generate test sets in which the examples result from NWP forecasts with a fixed lead time. Just like before, these test sets with fixed lead times are sampled in a climatologically consistent manner

and consist of 10^5 examples. The target times are drawn from the same test days as in section 2c.

The lead-time dependence of skill is shown in the upper panel of Fig. 6. While we measure skill in terms of the BSS, we have checked that the results of this section do not qualitatively change when considering a different skill score. All three models exhibit an approximately exponential decrease in skill. The rate at which skill decreases is very similar for the three models. This suggests that the decrease in skill is not model-specific but results from an increasing NWP forecast uncertainty, which is consistent with previous work (Vahid Yousefnia et al. 2024). As a consequence, the SALAMA 1D models' superior skill for low lead times is passed on to longer lead times. In the lower panel of Fig. 6, we show the difference in skill as a function of lead time for all model pairs. Again, we find that the SALAMA 1D models consistently outperform SALAMA 0D at a confidence level of 90%, with SALAMA 1D-2022 showing higher skill than SALAMA 1D-2021.

It is worth noting that the decrease in skill of SALAMA 0D is stronger than reported in Vahid Yousefnia et al. (2024). There, the initial skill decreased by at most 30% after 11 h,

TABLE 3. Scores for classification skill, as defined in, e.g., Vahid Yousefnia et al. (2024), evaluated on the test set. All scores except BSS and area under the precision–recall curve (PR-AUC) require setting a decision threshold to convert probabilities to binary output. The threshold is chosen for each model such that the average fraction of examples classified as thunderstorms is equal to the observed fraction of thunderstorm examples. For this threshold, recall equals precision, and the F_1 score, such that only recall is reported here. Uncertainties are obtained from 10^4 bootstrap resamples, using daywise block resampling, and show the symmetric 90% confidence interval. The last three columns evaluate the distribution of difference in skill between the models [score(A) – score(B) for model pair (A, B)], obtained from the bootstrap resamples, and show that all differences are significant on a 90% confidence level.

Skill score	0D	1D-21	1D-22	1D-21, 0D	1D-22, 1D-21	1D-22, 0D
BSS with climatology as reference	$0.234^{+0.037}_{-0.048}$	$0.261^{+0.035}_{-0.044}$	$0.281^{+0.034}_{-0.044}$	$0.027^{+0.015}_{-0.013}$	$0.020^{+0.008}_{-0.008}$	$0.047^{+0.014}_{-0.013}$
PR-AUC	$0.397^{+0.055}_{-0.071}$	$0.439^{+0.051}_{-0.065}$	$0.452^{+0.047}_{-0.061}$	$0.043^{+0.021}_{-0.019}$	$0.012^{+0.011}_{-0.011}$	$0.055^{+0.021}_{-0.017}$
Recall	$0.414^{+0.045}_{-0.059}$	$0.452^{+0.041}_{-0.054}$	$0.465^{+0.039}_{-0.050}$	$0.038^{+0.017}_{-0.016}$	$0.014^{+0.009}_{-0.009}$	$0.052^{+0.020}_{-0.018}$
Critical success index (CSI)	$0.261^{+0.037}_{-0.045}$	$0.292^{+0.035}_{-0.043}$	$0.303^{+0.034}_{-0.041}$	$0.031^{+0.013}_{-0.013}$	$0.012^{+0.007}_{-0.007}$	$0.043^{+0.015}_{-0.014}$
Equitable threat score (ETS)	$0.250^{+0.035}_{-0.043}$	$0.281^{+0.033}_{-0.042}$	$0.293^{+0.032}_{-0.039}$	$0.031^{+0.013}_{-0.013}$	$0.012^{+0.007}_{-0.007}$	$0.043^{+0.015}_{-0.014}$

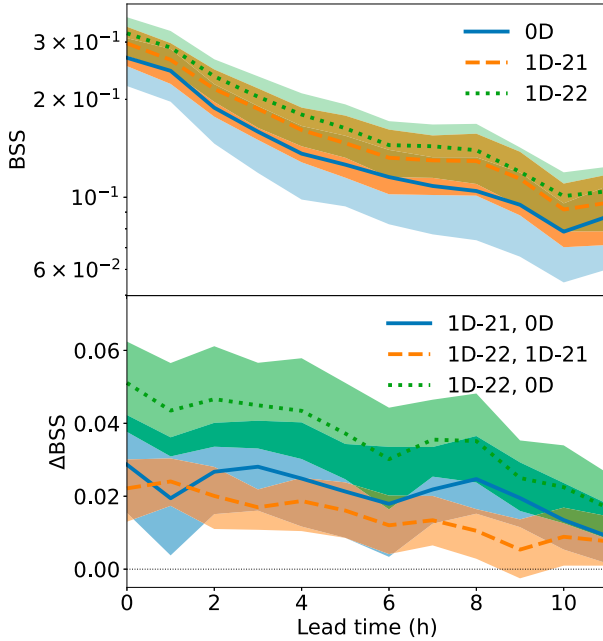


FIG. 6. Lead-time dependence of model skill, quantified by the BSS with climatology as reference. (top) The BSS for the individual models and (bottom) the skill difference $\Delta\text{BSS} = \text{BSS}(A) - \text{BSS}(B)$ between model pair (A, B) . Uncertainties are obtained from 10^4 bootstrap resamples, using daywise block resampling, and show the symmetric 90% confidence interval.

while the decrease here is approximately twice as much. This may result from using a more diverse training set in this study (we use twice as many days to compile the training set).

b. Interpretability study

For the remainder of this section, we focus on SALAMA 1D-2022, referring to it simply as SALAMA 1D. Our goal is to gain insight into how our model classifies input. We start by inspecting how the vertical profiles of the SALAMA 1D input fields look like on average for test set samples to which our model assigns a particularly high or low probability of thunderstorm occurrence. Figure 7 shows the average vertical profiles for the top and bottom probability percentiles of test set samples. Shaded bands denote the symmetric 50% confidence interval. Note that we converted specific humidity to dewpoint temperature T_d for a more straightforward comparison with temperature. For better orientation within the panels, we also plot the average tropopause height, which we compute according to the WMO definition (lowest level where the lapse rate drops to $\leq 2 \text{ K km}^{-1}$, with the average lapse rate within 2 km above remaining $\leq 2 \text{ K km}^{-1}$; WMO 1957).

The first column of panels in Fig. 7 shows temperature and dewpoint temperature T_d for the 2 percentiles. The top percentile troposphere displays more moisture than the bottom percentile, in particular in the 2–5-km height. This is consistent with the documented importance of moisture for thunderstorm development, as the buoyancy of rising air parcels is otherwise reduced by dry-air entrainment (Zhang et al. 2003; Peters et al. 2023; Marquis et al. 2023). In the second column

in Fig. 7, we show average profiles of mixing ratios of cloud water (QC), cloud ice (QI), and graupel (QG). The column suggests that the model uses nonvanishing profiles of QI and QG to discriminate between the top and the bottom percentiles. For high-probability samples, ice particle content peaks close to the tropopause at a height of 10–12 km, consistent with measurements of vertical hydrometeor distributions (e.g., Vivekanandan et al. 1999; Hubbert et al. 2018). The third column in Fig. 7 displays the average profiles of cloud cover (CLC). In general, our model associates nonvanishing CLC with a high probability of thunderstorm occurrence. In particular, close to the tropopause, CLC tends to be 100%. This is consistent with anvil cloud-top levels (Markowski and Richardson 2010). In the fourth column in Fig. 7, we show vertical profiles of the three wind components U , V , and W . It is noteworthy that high-probability samples tend to have southwesterly wind profiles, whereas samples from the bottom percentile display northwesterly winds. This is consistent with studies on the typical propagation direction of thunderstorms in central Europe (Hagen et al. 1999). On the other hand, vertical profiles of W vanish for both percentiles. We presume that due to convection displacement errors in the NWP model, the updraft regions within simulated deep convection rarely match observed lightning observations. Therefore, SALAMA 1D may have learned not to rely on W for inferring thunderstorm occurrence. The last column of Fig. 7 shows the average profiles of pressure. Both profiles appear to be essentially hydrostatic. Surface pressure tends to be lower for the top percentile than for the bottom percentile.

Comparing the average profiles of the SALAMA 1D input fields for the 2 percentiles is useful to get a first idea of whether our model separates the thunderstorm class from the majority class in a physically interpretable manner. On the other hand, this analysis does not inform about the relative importance of the individual atmospheric variables. Therefore, we conduct a linear sensitivity analysis of the conditional probability f of thunderstorm occurrence (SALAMA 1D model output) with respect to the input. The general idea is to consider for a given input sample $\xi = (\xi_{ij}) \in \mathbb{R}^{N_i \times N_z}$ the partial derivatives of f :

$$S_{ij}(\xi) \equiv \frac{\partial f(\xi)}{\partial \xi_{ij}}. \quad (5)$$

The term $S_{ij}(\xi)$ constitutes a measure of how much f reacts to changes in ξ_{ij} and, therefore, quantifies the importance of ξ_{ij} to the outcome. The term $S_{ij}(\xi)$ is commonly referred to as saliency in the ML literature (Simonyan et al. 2014; Li et al. 2022), while meteorologists might know it as adjoint sensitivity (Errico 1997; Warder et al. 2021).

In order for saliency values $S_{ij}(\xi)$ to be comparable across all indices i and j , we need to scale the input fields appropriately. This scaling accounts for the fact that the fields have different units and vary differently from one sample to another. Consider an unscaled input field $\xi(z)$ (e.g., pressure), which we take as a function of height z above ground. We define the corresponding scaled fields as

$$\tilde{\xi}(z) = \frac{\xi(z) - \mu_\xi(z)}{\sigma_\xi(z)}, \quad (6)$$

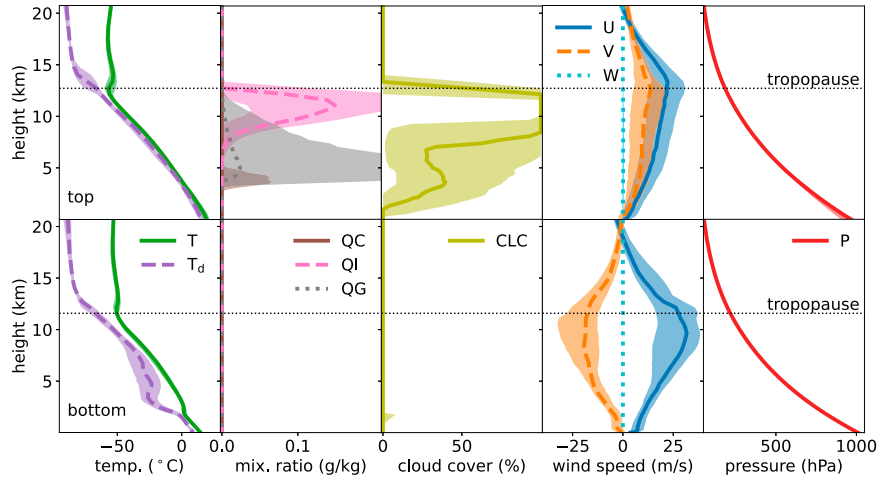


FIG. 7. Average vertical field profiles for the (top) top probability percentile of test set samples and (bottom) the bottom percentile, with annotated levels of the tropopause. We convert average specific humidity to dewpoint temperature T_d for a comparison with T . Shaded bands correspond to the symmetric 50% confidence interval.

where $\mu_{\xi}(z)$ and $\sigma_{\xi}(z)$ encode a characteristic background climatology for $\tilde{\xi}(z)$. We define them as

$$\mu_{\xi}(z) = P_{50}[\tilde{\xi}(z)], \quad (7)$$

$$\sigma_{\xi}(z) = \frac{P_{99}[\tilde{\xi}(z)] - P_1[\tilde{\xi}(z)]}{2}, \quad (8)$$

where P_n stands for the n th percentile of the variable in the brackets, evaluated for the training set. We then compute saliency with respect to the scaled fields.

While saliency varies from one sample to another and can be used to interpret individual predictions, we propose averaging over the top probability percentile to obtain more robust insight into the input fields and height levels which contribute most to high-probability output. To this end, we compute $|\langle S_{ij} \rangle|$, where the angle brackets denote averaging over the top percentile samples. We take the absolute value (absolute saliency), as we consider feature importance to be linked to the (sample-averaged) intensity of the ML model's linear response, irrespective of whether this response is positive or negative. In what follows, we refer to $|\langle S_{ij} \rangle|$ simply as saliency, i.e., with sample averaging and taking the absolute value being implicitly implied unless stated otherwise. The resulting saliency map for the top percentile is shown in Fig. 8. Note that average saliencies of the different fields are stacked on top of each other. As a consequence, the saliency envelope quantifies how much individual height levels affect the model outcome.

The saliency envelope displays two distinct peaks, at $z = 12$ km and $z = 5$ km, respectively. The upper-level peak receives the most contributions by horizontal wind speed. Indeed, the saliencies of U and V are maximal near the tropopause, where the average horizontal wind velocity difference between the top and bottom percentiles is greatest (Fig. 7). This suggests that the model relies to a considerable extent on the learned climatological propagation direction of thunderstorms. Conversely,

W saliency is approximately one order of magnitude smaller than the saliencies of U and V . QI saliency is maximal at the top of the troposphere, contributing significantly to the upper-level peak, as well. QI is present only at this height (Fig. 7),

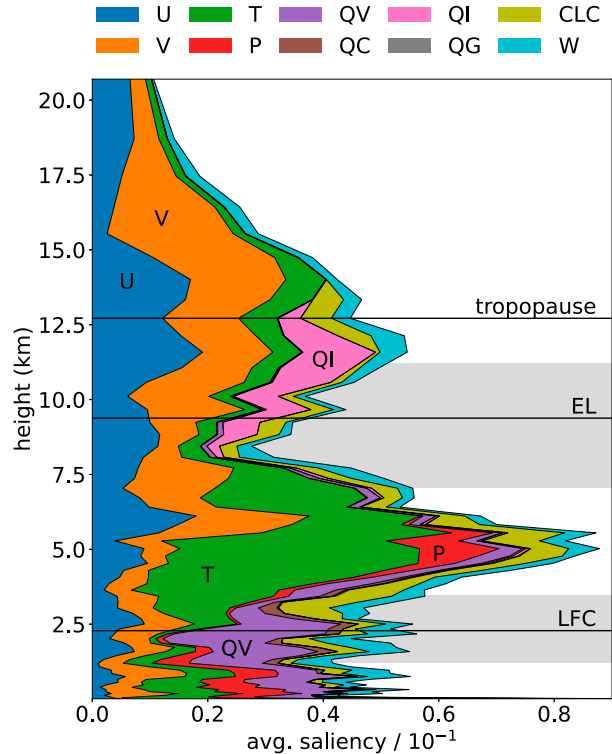


FIG. 8. Vertical profiles of average saliency for the top percentile of test set samples (based on model probability), with annotated levels of the tropopause, as well as the LFC and EL of a mixed-layer parcel. Saliencies for the different fields are stacked on top of each other in the order given in Table 1.

which suggests that the model actively takes ice particle content into consideration to infer thunderstorm occurrence. The only other hydrometeor field with significant nonvanishing saliency is specific humidity (QV). QV saliency peaks near the LFC, coinciding with the vicinity of maximal tropospheric moisture (Fig. 7). Similarly, CLC saliency is low but nonvanishing at all heights with nonzero CLC (Fig. 7).

Next, we turn to the midlevel peak. Apart from horizontal wind velocity, the midlevel peak receives most contributions from temperature. This may be partly due to convection feeding back to the temperature field. However, since pressure saliency is nonvanishing only at the midlevel peak and near the surface, we hypothesize that the ML model also reconstructs lapse rates. To understand this, note that our model is not informed about the height of individual vertical levels; rather, these levels are randomly shuffled (section 3a). Thus, our model can reliably infer height levels—and in particular level spacings—only from pressure, which monotonously decreases with height (Fig. 7). Therefore, we expect pressure saliency to be a proxy for how much the model relies on vertical gradients. Finally, as pressure saliency contributes to the midlevel peak and temperature saliency is high, we conjecture that our model reconstructs midlevel lapse rates. This is supported by the fact that the midlevel peak is bounded by the LFC and the equilibrium level (EL) of parcels lifted from the mixed layer, meaning that such parcels are buoyant for height levels in the vicinity of the midlevel peak. Positive buoyancy occurring in a conditionally unstable troposphere is known to be a crucial ingredient for thunderstorm development (Doswell et al. 1996) and constitutes the basis for several traditional thunderstorm predictors, such as CAPE.

To test whether SALAMA 1D considers midlevel lapse rates, we show in Fig. 9a the distribution of 500–300-hPa lapse rates for the top and bottom probability percentiles. Indeed, essentially all high-probability samples are associated with conditionally unstable midlevel lapse rates, whereas the bottom percentile distribution extends further into absolutely stable midlevel lapse rates. On the other hand, the distributions of the 2 percentiles show a significant overlap, which implies that considering midlevel lapse rates alone is not sufficient to infer a high probability of thunderstorm occurrence.

Complementary to midlevel lapse rates, we show the distributions of CAPE for the 2 percentiles in Fig. 9c. Most samples in both percentiles have a low value of CAPE, with around 50% of the high-probability samples and 90% of the low-probability samples falling into the lowest bin. Nevertheless, the top percentile distribution has a longer tail, toward higher CAPE values, than the bottom percentile distribution does. The low CAPE values of the bottom percentile samples are consistent with conditional instability failing to develop. As for the top percentile samples, we expect CAPE to be considerably reduced inside the core of a convective cell, whereas some CAPE is expected to remain in lightning regions of less intense precipitation, producing the long tail of the top percentile distribution.

As pressure saliency is nonvanishing also near the surface, we show in Fig. 9b distributions of near-surface (10–1000 m) lapse rates. Indeed, the distributions differ for the 2 percentiles. In contrast to the midlevel peak, near-surface lapse rates

for the top percentile tend to be lower, and mostly absolutely stable, whereas lapse rates for the bottom percentile are mostly conditionally unstable. Figure 9d shows the corresponding distributions of convective inhibition (CIN). While most samples of both percentiles fall into the lowest bin, the top percentile distribution has a longer tail toward higher values of CIN. This might seem surprising as this result seems to suggest that more stable low-level lapse rates (or higher values of CIN) favor thunderstorm occurrence. However, the reduction in CIN is only conducive to thunderstorm development if sufficient CAPE has been able to build up beforehand. The bottom percentile samples are essentially from environments with low CAPE and low CIN, which suggests that in these cases, the constantly low CIN (due to, e.g., continuous mixing of low-level air) caused any instabilities aloft to be released prematurely, preventing CAPE from building up or thunderstorms from forming (Carlson et al. 1983; Tuckman et al. 2023). As for the top percentile samples, CIN is expected to be mostly removed within storm centers, while some CIN can be expected to remain in the less intense regions of convective precipitation or for samples with simulated convection occurring nearby or in the near future. This effect likely produces the longer tail of the top percentile distribution of CIN.

In summary, SALAMA 1D appears to rely on two categories of patterns. One category consists of patterns related to

- tropopause ice particle content, or
- cloud cover.

These patterns pertain to regions within ongoing convection, in which precipitation is most intense. In contrast, we identify a category of patterns related to

- horizontal wind direction,
- midlevel and near-surface lapse rates, or
- low-level moisture.

We refer to the latter patterns as mesoscale since the underlying fields, such as temperature and pressure, vary more slowly in the horizontal than the fine-grained hydrometeor variables do. Conversely, we refer to the former patterns as submesoscale. Mesoscale patterns in the above sense tend to be characteristic of regions with sufficient distance to the convective cores such that precipitation is less intense and some CAPE remains to be released. SALAMA 1D is sensitive to both categories to identify thunderstorm occurrence. While submesoscale patterns are useful for the identification of convective storm centers, the sensitivity to mesoscale patterns could explain the ML model's skill at increasing the areas of simulated convection observed in the case study (Fig. 4), namely, by accounting for modestly precipitating regions with low-level moisture and leftover CAPE.

5. Discussion and conclusions

Bypassing the traditional use of derived single-level predictors from NWP data, we developed SALAMA 1D, an ML model for predicting the probability of thunderstorm

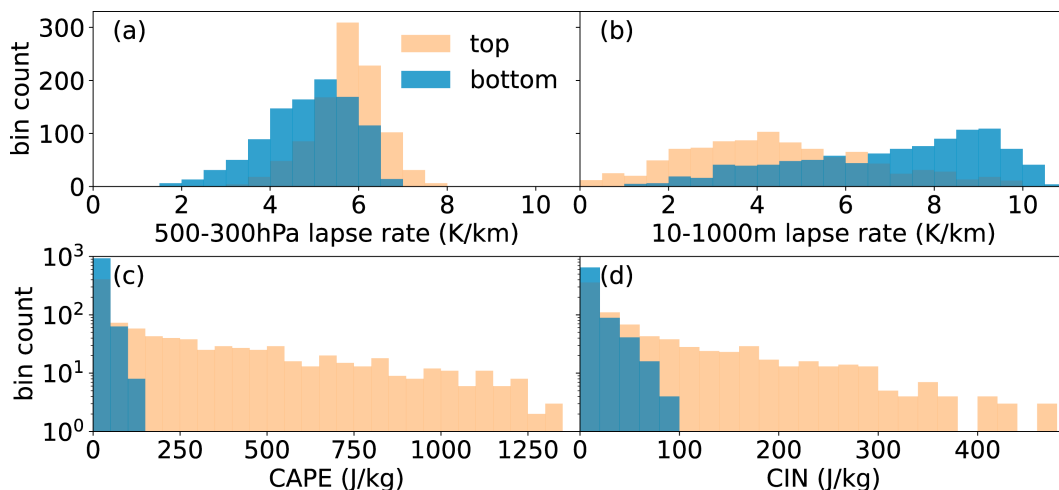


FIG. 9. (a) Distribution of midlevel lapse rates, (b) near-surface lapse rates, (c) mixed-layer CAPE, and (d) mixed-layer CIN for the top and bottom probability percentiles of test set samples.

occurrence on a pixelwise basis by processing vertical profiles of three-dimensional variables from convection-permitting NWP forecasts. The model's architecture was motivated by physical considerations. In particular, a sparse layer reduces parameter size by encouraging interactions at similar height levels, while a shuffling mechanism prevents the model from learning patterns tied to the vertical grid structure. The latter also adds a form of regularization that limits overfitting.

In comparison to SALAMA 0D, an ML baseline model that infers thunderstorm occurrence from derived single-level features, our model demonstrated higher skill across a wide range of metrics and for lead times up to at least 11 h. This result indicated that information relevant to thunderstorm occurrence, while intricately encoded in vertical profiles, can be successfully extracted by machine learning, resulting in an improved ability to recognize thunderstorm occurrence in NWP forecasts. Notably, our model remains lightweight in terms of computational complexity, making it just as suitable for real-time operational use as SALAMA 0D is. Case studies suggested that SALAMA 1D is capable of correcting the raw NWP output when convective areas are of incorrect size or when NWP fails to produce convection in the first place. Furthermore, doubling the number of days used to compile the training set (while keeping the training set size constant) also increased skill, underscoring the importance of a large and diverse database of NWP data. We anticipate further skill improvements with the collection of more NWP data.

A sensitivity analysis based on saliency maps revealed that many learned patterns are physically interpretable. For instance, our results suggested that SALAMA 1D has learned the climatological propagation direction of thunderstorms in the study region and relies on fine-grained (submesoscale) structures, such as ice particle content near the tropopause, or cloud cover, to identify high-precipitation regions of ongoing convection. Conversely, mesoscale patterns related to

atmospheric instability and moisture are used, possibly to account for regions with less intense precipitation and leftover CAPE. We hypothesized that mesoscale patterns are instrumental in correcting the areal size of simulated convection.

To improve the ML model's capability of correcting for NWP-related biases, it may be beneficial to adapt the model in such a way that horizontally extended input, or several forecast times around the target time, is processed. This would allow for improvement on correcting location and timing errors of convection. Furthermore, one could train the model in such a way that it can process all ensemble members simultaneously in one forward pass, which would enable the ML model to account for the NWP forecast uncertainty.

In closing, we stress that the methodology applied in this work may be useful for machine learning (ML) topics beyond the identification of thunderstorm occurrence in NWP data. Our approach demonstrates how incorporating physical constraints and symmetry principles can lead to more robust and computationally efficient ML models. Additionally, our use of saliency maps highlights a path toward more interpretable ML models, fostering greater trust as we gain insight into how models arrive at their predictions. As ML continues to play a growing role in severe weather forecasting, ensuring that these models are both accurate and transparent will be key to enhancing their operational utility in critical decision-making processes.

Acknowledgments. We gratefully acknowledge the computational and data resources provided through the joint high-performance data analytics (HPDA) project “terabyte” of the DLR and the Leibniz Supercomputing Center (LRZ). We thank Björn Brötz and Hessel Juliust for reviewing the ML code. C.M. carried out his contributions within the Italia-Deutschland Science-4-Services Network in Weather and Climate (IDEA-S4S; SESTO, 4823IDEAP4). This Italian-German research network of universities, research

institutes, and DWD is funded by the Federal Ministry of Digital and Transport (BMDV). The authors declare that there are no conflicts of interest to disclose.

Data availability statement. The training, validation, and testing datasets generated in this study are available on Zenodo (<https://doi.org/10.5281/zenodo.13981207>). The code for all ML models is released on GitHub (<https://github.com/kvahidyou/SALAMA>). The original simulation data from the ICON-D2-EPS model are accessible through the corresponding DWD database (<https://www.dwd.de/EN/ourservices/pamore/pamore>). Please note that the original data from the LINET are proprietary and cannot be shared in their raw form.

REFERENCES

- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Betz, H. D., K. Schmidt, P. Laroche, P. Blanchet, W. P. Oettinger, E. Defer, Z. Dziewit, and J. Konarski, 2009: LINET—An international lightning detection network in Europe. *Atmos. Res.*, **91**, 564–573, <https://doi.org/10.1016/j.atmosres.2008.06.012>.
- Borsky, S., and C. Unterberger, 2019: Bad weather and flight delays: The impact of sudden and slow onset weather events. *Econ. Transp.*, **18**, 10–26, <https://doi.org/10.1016/j.ecotra.2019.02.002>.
- Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, <https://doi.org/10.1175/WAF993.1>.
- Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning-based probabilistic hail predictions for operational forecasting. *Wea. Forecasting*, **35**, 149–168, <https://doi.org/10.1175/WAF-D-19-0105.1>.
- Carlson, T. N., S. G. Benjamin, G. S. Forbes, and Y.-F. Li, 1983: Elevated mixed layers in the regional severe storm environment: Conceptual model and case studies. *Mon. Wea. Rev.*, **111**, 1453–1474, [https://doi.org/10.1175/1520-0493\(1983\)111<1453:EMLITR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1983)111<1453:EMLITR>2.0.CO;2).
- Diffenbaugh, N. S., M. Scherer, and R. J. Trapp, 2013: Robust increases in severe thunderstorm environments in response to greenhouse forcing. *Proc. Natl. Acad. Sci. USA*, **110**, 16361–16366, <https://doi.org/10.1073/pnas.1307758110>.
- Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, [https://doi.org/10.1175/1520-0434\(1996\)011<0560:FFFAIB>2.0.CO;2](https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2).
- Dramsch, J. S., and Coauthors, 2025: Explainability can foster trust in artificial intelligence in geoscience. *Nat. Geosci.*, **18**, 112–114, <https://doi.org/10.1038/s41561-025-01639-x>.
- Errico, R. M., 1997: What is an adjoint model? *Bull. Amer. Meteor. Soc.*, **78**, 2577–2591, [https://doi.org/10.1175/1520-0477\(1997\)078<2577:WIAAM>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2577:WIAAM>2.0.CO;2).
- Flora, M. L., C. K. Potvin, A. McGovern, and S. Handler, 2024: A machine learning explainability tutorial for atmospheric sciences. *Artif. Intell. Earth Syst.*, **3**, e230018, <https://doi.org/10.1175/AIES-D-23-0018.1>.
- Frankle, J., and M. Carbin, 2019: The lottery ticket hypothesis: Finding sparse, trainable neural networks. *Int. Conf. on Learning Representations 2019*, New Orleans, LA, ICLR, 1–42, <https://openreview.net/forum?id=rJl-b3RcF7>.
- Gagne, D. J., II, A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, <https://doi.org/10.1175/WAF-D-17-0010.1>.
- Geng, Y.-a., and Coauthors, 2021: A deep learning framework for lightning forecasting with multi-source spatiotemporal data. *Quart. J. Roy. Meteor. Soc.*, **147**, 4048–4062, <https://doi.org/10.1002/qj.4167>.
- Gerz, T., C. Forster, and A. Tafferner, 2012: Mitigating the impact of adverse weather on aviation. *Atmospheric Physics*, U. Schumann, Ed., Springer, 645–659, https://doi.org/10.1007/978-3-642-30183-4_39.
- Hagen, M., B. Bartenschlager, and U. Finke, 1999: Motion characteristics of thunderstorms in Southern Germany. *Meteor. Appl.*, **6**, 227–239, <https://doi.org/10.1017/S1350482799001164>.
- Hasanin, T., and T. Khoshgoftaar, 2018: The effects of random undersampling with simulated class imbalance for big data. *2018 IEEE Int. Conf. on Information Reuse and Integration (IRI)*, Salt Lake City, UT, Institute of Electrical and Electronics Engineers, 70–79, <https://doi.org/10.1109/IRI.2018.00018>.
- Herman, G. R., and R. S. Schumacher, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, <https://doi.org/10.1175/MWR-D-17-0250.1>.
- Holle, R. L., 2014: Some aspects of global lightning impacts. *2014 Int. Conf. on Lightning Protection (ICLP)*, Shanghai, China, Institute of Electrical and Electronics Engineers, 1390–1395, <https://doi.org/10.1109/ICLP.2014.6973348>.
- Hubbert, J. C., J. W. Wilson, T. M. Weckwerth, S. M. Ellis, M. Dixon, and E. Loew, 2018: S-Pol's polarimetric data reveal detailed storm features (and insect behavior). *Bull. Amer. Meteor. Soc.*, **99**, 2045–2060, <https://doi.org/10.1175/BAMS-D-17-0317.1>.
- Jardines, A., H. Eivazi, E. Zea, J. García-Heras, J. Simarro, E. Otero, M. Soler, and R. Vinuesa, 2024: Thunderstorm prediction during pre-tactical air-traffic-flow management using convolutional neural networks. *Expert Syst. Appl.*, **241**, 122466, <https://doi.org/10.1016/j.eswa.2023.122466>.
- Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, <https://doi.org/10.48550/ARXIV.1412.6980>.
- Kober, K., G. C. Craig, C. Keil, and A. Dörnbrack, 2012: Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 755–768, <https://doi.org/10.1002/qj.939>.
- LeCun, Y., J. Denker, and S. Solla, 1989: Optimal brain damage. *Advances in Neural Information Processing Systems*, Vol. 2, D. Touretzky, Ed., Morgan-Kaufmann, 598–605, <https://proceedings.neurips.cc/paperfiles/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf>.
- Leinonen, J., U. Hamann, U. Germann, and J. R. Mecikalski, 2022: Nowcasting thunderstorm hazards using machine learning: The impact of data sources on performance. *Nat. Hazards Earth Syst. Sci.*, **22**, 577–597, <https://doi.org/10.5194/nhess-22-577-2022>.
- , —, I. V. Sideris, and U. Germann, 2023: Thunderstorm nowcasting with deep learning: A multi-hazard data fusion model. *Geophys. Res. Lett.*, **50**, e2022GL101626, <https://doi.org/10.1029/2022GL101626>.
- Li, J., C. Forster, J. Wagner, and T. Gerz, 2021: Cb-fusion—Forecasting thunderstorm cells up to 6 hours. *Meteor. Z.*, **30**, 169–184, <https://doi.org/10.1127/metz/2020/1047>.

- Li, W., H. Chen, L. Han, and J. Xu, 2022: The interpretation of deep learning for convective storm nowcasting. *IGARSS 2022 - 2022 IEEE Int. Geoscience and Remote Sensing Symp.*, Kuala Lumpur, Malaysia, Institute of Electrical and Electronics Engineers, 7922–7925, <https://doi.org/10.1109/IGARSS46834.2022.9883967>.
- Lin, P.-F., P.-L. Chang, B. J.-D. Jou, J. W. Wilson, and R. D. Roberts, 2012: Objective prediction of warm season afternoon thunderstorms in Northern Taiwan using a fuzzy logic approach. *Wea. Forecasting*, **27**, 1178–1197, <https://doi.org/10.1175/WAF-D-11-00105.1>.
- Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1605–1631, <https://doi.org/10.1175/WAF-D-19-0258.1>.
- Markowski, P., and Y. Richardson, 2010: Basic equations and tools. *Mesoscale Meteorology in Midlatitudes*, John Wiley and Sons, 11–40, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470682104>.
- Marquis, J. N., Z. Feng, A. Varble, T. C. Nelson, A. Houston, J. M. Peters, J. P. Mulholland, and J. Hardin, 2023: Near-cloud atmospheric ingredients for deep convection initiation. *Mon. Wea. Rev.*, **151**, 1247–1267, <https://doi.org/10.1175/MWR-D-22-0243.1>.
- Mohammed, R., J. Rawashdeh, and M. Abdullah, 2020: Machine learning with oversampling and undersampling techniques: Overview study and experimental results. 2020 *11th Int. Conf. on Information and Communication Systems (ICICS)*, Online, Institute of Electrical and Electronics Engineers, 243–248, <https://doi.org/10.1109/ICICS49469.2020.239556>.
- Mueller, C., T. Saxen, R. Roberts, J. Wilson, T. Betancourt, S. Dettling, N. Oien, and J. Yee, 2003: NCAR auto-nowcast system. *Wea. Forecasting*, **18**, 545–561, [https://doi.org/10.1175/1520-0434\(2003\)018<0545:NAS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<0545:NAS>2.0.CO;2).
- Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2).
- Ntelekos, A. A., J. A. Smith, and W. F. Krajewski, 2007: Climatological analyses of thunderstorms and flash floods in the Baltimore metropolitan region. *J. Hydrometeor.*, **8**, 88–101, <https://doi.org/10.1175/JHM558.1>.
- Ortland, S. M., M. J. Pavolonis, and J. L. Cintineo, 2023: The development and initial capabilities of ThunderCast, a deep learning model for thunderstorm nowcasting in the United States. *Artif. Intell. Earth Syst.*, **2**, e230044, <https://doi.org/10.1175/AIES-D-23-0044.1>.
- Palmer, T., 2017: The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *J. Adv. Model. Earth Syst.*, **9**, 730–734, <https://doi.org/10.1002/2017MS000999>.
- Peters, J. M., D. R. Chavas, C.-Y. Su, H. Morrison, and B. E. Coffey, 2023: An analytic formula for entraining CAPE in midlatitude storm environments. *J. Atmos. Sci.*, **80**, 2165–2186, <https://doi.org/10.1175/JAS-D-23-0003.1>.
- Piper, D., M. Kunz, F. Ehmele, S. Mohr, B. Mühr, A. Kron, and J. Daniell, 2016: Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany—Part 1: Meteorological background. *Nat. Hazards Earth Syst. Sci.*, **16**, 2835–2850, <https://doi.org/10.5194/nhess-16-2835-2016>.
- Rädler, A. T., P. H. Groenemeijer, E. Faust, R. Sausen, and T. Púčik, 2019: Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. *npj Climate Atmos. Sci.*, **2**, 30, <https://doi.org/10.1038/s41612-019-0083-7>.
- Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597**, 672–677, <https://doi.org/10.1038/s41586-021-03854-z>.
- Reinert, D., and Coauthors, 2020: DWD database reference for the global and regional ICON and ICON-EPs forecasting system, version 2.1.7. Deutscher Wetterdienst Tech. Rep., 138 pp.
- Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, <https://doi.org/10.1175/2007MWR2123.1>.
- Saito, T., and M. Rehmsmeier, 2015: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, **10**, e0118432, <https://doi.org/10.1371/journal.pone.0118432>.
- Simon, T., P. Fabsic, G. J. Mayr, N. Umlauf, and A. Zeileis, 2018: Probabilistic forecasting of thunderstorms in the Eastern Alps. *Mon. Wea. Rev.*, **146**, 2999–3009, <https://doi.org/10.1175/MWR-D-17-0366.1>.
- Simonyan, K., A. Vedaldi, and A. Zisserman, 2014: Deep inside convolutional networks: Visualising image classification models and saliency maps. *Int. Conf. on Learning Representations*, Banff, AB, Canada, ICLR, <https://doi.org/10.48550/arXiv.1312.6034>.
- Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, <https://doi.org/10.1175/WAF-D-10-05046.1>.
- , G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, **35**, 1981–2000, <https://doi.org/10.1175/WAF-D-20-0036.1>.
- Sovrasov, V., 2018: ptflops: A flops counting tool for neural networks in pytorch framework. GitHub, <https://github.com/sovrasov/flops-counter.pytorch>.
- Taszarek, M., J. T. Allen, M. Marchio, and H. E. Brooks, 2021: Global climatology and trends in convective environments from ERA5 and rawinsonde data. *npj Climate Atmos. Sci.*, **4**, 35, <https://doi.org/10.1038/s41612-021-00190-x>.
- Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, <https://doi.org/10.1017/S1350482705001763>.
- Tuckman, P., V. Agard, and K. Emanuel, 2023: Evolution of convective energy and inhibition before instances of large CAPE. *Mon. Wea. Rev.*, **151**, 321–338, <https://doi.org/10.1175/MWR-D-21-0302.1>.
- Ukkonen, P., and A. Mäkelä, 2019: Evaluation of machine learning classifiers for predicting deep convection. *J. Adv. Model. Earth Syst.*, **11**, 1784–1802, <https://doi.org/10.1029/2018MS001561>.
- Vahid Yousefnia, K., T. Böhle, I. Zöbisch, and T. Gerz, 2024: A machine-learning approach to thunderstorm forecasting through post-processing of simulation data. *Quart. J. Roy. Meteor. Soc.*, **150**, 3495–3510, <https://doi.org/10.1002/qj.4777>.
- Vivekanandan, J., D. S. Zrnčić, S. M. Ellis, R. Oye, A. V. Ryzhkov, and J. Straka, 1999: Cloud microphysics retrieval using S-band dual-polarization radar measurements. *Bull. Amer. Meteor. Soc.*, **80**, 381–388, [https://doi.org/10.1175/1520-0477\(1999\)080<0381:CMRUSB>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<0381:CMRUSB>2.0.CO;2).
- Warder, S. C., K. J. Horsburgh, and M. D. Piggott, 2021: Adjoint-based sensitivity analysis for a numerical storm surge model. *Ocean Modell.*, **160**, 101766, <https://doi.org/10.1016/j.ocemod.2021.101766>.

- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.
- WMO, 1957: Definition of the tropopause. *Bull. WMO*, **6**, 136–137.
- Yang, R., and Coauthors, 2024: Interpretable machine learning for weather and climate prediction: A review. *Atmos. Environ.*, **338**, 120797, <https://doi.org/10.1016/j.atmosenv.2024.120797>.
- Yano, J.-I., and Coauthors, 2018: Scientific challenges of convective-scale numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **99**, 699–710, <https://doi.org/10.1175/BAMS-D-17-0125.1>.
- Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1185, [https://doi.org/10.1175/1520-0469\(2003\)060<1173:EOMCOM>2.0.CO;2](https://doi.org/10.1175/1520-0469(2003)060<1173:EOMCOM>2.0.CO;2).
- Zhou, K., J. Sun, Y. Zheng, and Y. Zhang, 2022: Quantitative precipitation forecast experiment based on basic NWP variables using deep learning. *Adv. Atmos. Sci.*, **39**, 1472–1486, <https://doi.org/10.1007/s00376-021-1207-7>.
- Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf, 2015: The ICON (ICOsahedral non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quart. J. Roy. Meteor. Soc.*, **141**, 563–579, <https://doi.org/10.1002/qj.2378>.