Expert Systems With Applications 241 (2024) 122466

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Thunderstorm prediction during pre-tactical air-traffic-flow management using convolutional neural networks

Aniel Jardines [a,*], Hamidreza Eivazi [b], Elias Zea [b], Javier García-Heras [a], Juan Simarro [c], Evelyn Otero [b], Manuel Soler [a], Ricardo Vinuesa [b]

[a] Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Spain
[b] Department of Engineering Mechanics, KTH Royal Institute of Technology, SE-1044, Stockholm, Sweden
[c] Agencia Estatal de Meteorología (AEMET), Valencia, Spain

ARTICLE INFO

Keywords:
Thunderstorms
Air-traffic management
Weather data
Numerical weather prediction
Satellite images
Convolutional neural network
Machine learning

ABSTRACT

Thunderstorms can be a large source of disruption for European air-traffic management causing a chaotic state of operation within the airspace system. In current practice, air-traffic managers are provided with imprecise forecasts which limit their ability to plan strategically. As a result, weather mitigation is performed using tactical measures with a time horizon of three hours. Increasing the lead time of thunderstorm predictions to the day before operations could help air-traffic managers plan around weather and improve the efficiency of air-traffic-management operations. Emerging techniques based on machine learning have provided promising results, partly attributed to reduced human bias and improved capacity in predicting thunderstorms purely from numerical weather prediction data. In this paper, we expand on our previous work on thunderstorm forecasting, by applying convolutional neural networks (CNNs) to exploit the spatial characteristics embedded in the weather data. The learning task of predicting convection is formulated as a binary-classification problem based on satellite data. The performance of multiple CNN-based architectures, including a fully-convolutional neural network (FCN), a CNN-based encoder–decoder, a U-Net, and a pyramid-scene parsing network (PSPNet) are compared against a multi-layer-perceptron (MLP) network. Our work indicates that CNN-based architectures improve the performance of point-prediction models, with a fully-convolutional neural-network architecture having the best performance. Results show that CNN-based architectures can be used to increase the prediction lead time of thunderstorms. Lastly, a case study illustrating the applications of convection-prediction models in an air-traffic-management setting is presented.

1. Introduction

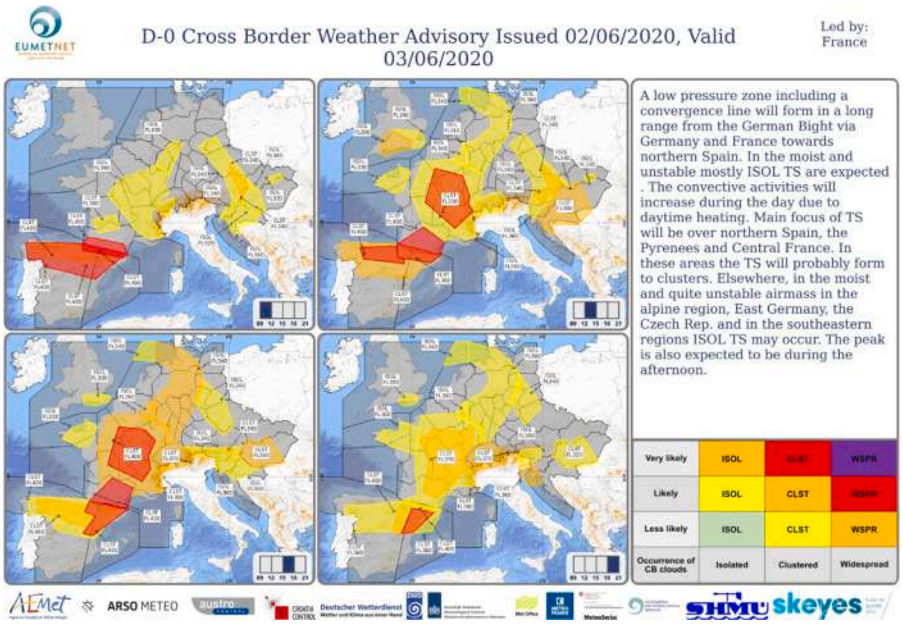
The European airspace can handle more than 30 thousand flights in a single day. Managing such a high volume of flights is a responsibility that falls on air navigation service providers (ANSPs). There are 37 ANSPs in Europe, each typically operating at the national level, for instance, ENAIRE in Spain or DSNA (Direction des Services de la navigation aérienne) in France. The coordination of flight plans and actual traffic among so many ANSPs is facilitated by the international organization EUROCONTROL, specifically the Network Manager Operations Center in Brussels. Most of the time this setup works well, and flight operations run smoothly, however, convective weather wreaks havoc on the air space system, with one-quarter of the total delay in the system being directly attributed to weather (EUROCONTROL, 2019).

Thunderstorms are frequent in the summer and coincide with a period of high air-traffic demand in the European airspace. This combination of bad weather and high demand causes significant disruption to air-traffic-management operations. Managing airborne traffic during thunderstorms can quickly become chaotic, with the sudden decreases in airport and airspace capacity, flights must divert from their intended flight paths, enter holding procedures in trying conditions, and manage fuel reserves as they scramble to find accommodations at alternate airports. On the ground, flights are subject to delays and cancellations that quickly propagate throughout the network.

In current European airspace operations, air-traffic managers are provided with weather information using a convection risk map, an example of which is shown in Fig. 1. This product is provided via pdf in an email on the morning of the day of operations and provides

[*] Corresponding author.
E-mail addresses: ajardine@ing.uc3m.es (A. Jardines), hamidre@kth.se (H. Eivazi), zea@kth.se (E. Zea), gcarrete@ing.uc3m.es (J. García-Heras), jsimarrog@aemet.es (J. Simarro), otero@kth.se (E. Otero), masolera@ing.uc3m.es (M. Soler), rvinuesa@mech.kth.se (R. Vinuesa).

https://doi.org/10.1016/j.eswa.2023.122466
Received 27 September 2022; Received in revised form 9 September 2023; Accepted 4 November 2023
Available online 10 November 2023



D-0 Cross Border Weather Advisory Issued 02/06/2020, Valid 03/06/2020

Led by: France

A low pressure zone including a convergence line will form in a long range from the German Bight via Germany and France towards northern Spain. In the moist and unstable mostly ISOL TS are expected . The convective activities will increase during the day due to daytime heating. Main focus of TS will be over northern Spain, the Pyrenees and Central France. In these areas the TS will probably form to clusters. Elsewhere, in the moist and quite unstable airmass in the alpine region, East Germany, the Czech Rep. and in the southeastern regions ISOL TS may occur. The peak is also expected to be during the afternoon.

Very likely	ISOL	CLST	WSPR
Likely	ISOL	CLST	WSPR
Less likely	ISOL	CLST	WSPR
Occurrence of CB clouds	Isolated	Clustered	Widespread

Fig. 1. Example of Cross-border Convection Advisory Product.
Source: Figure extracted from https://www.eumetnet.eu/.

convective weather advisories in 3-hour blocks. Due to the lack of spatial and temporal resolution in the product, ANSPs typically do not make strategic modifications to their operational plans; instead, choosing to make tactical adjustments in real-time according to the evolving weather situation. This reactive approach to handling disruptive weather events is not coordinated among the multiple ANSPs in Europe and leads to inefficiency in the system.

Forecasting the origin and evolution of convective weather remains a challenge. The general meteorological conditions necessary for convection are well understood, however, the exact timing and location of initiation triggers can be difficult to identify. As a result, nowcasting is the preferred technique for thunderstorm prediction. Nowcasting consists of short-term (1–3 h) predictions based on the extrapolation of observational data such as satellite images or radar (Wilson et al., 1998). In the United States, the Corridor Integrated Weather System (CIWS) is a nowcasting-based system offering meteorological information to the aviation community (Evans & Ducot, 2006). Nowcasting can offer precise weather predictions in the near term, but the extrapolation quickly breaks down for longer time horizons.

In this research, we propose increasing the lead time of convection prediction by exploiting the advances in numerical-weather-prediction (NWP) products. NWPs model the atmospheric processes on a computational grid using computer simulations and can estimate a large set of atmospheric parameters at each grid cell by using partial differential equations to capture the fluid flow and thermodynamic characteristics among neighboring grid cells.

With the aid of supercomputers, NWPs can provide fairly accurate forecasts of the state of the atmosphere multiple days into the future. Note that, while NWPs serve as the source for the majority of the weather forecast we encounter in our daily lives, they have not traditionally been used for thunderstorm prediction because the size and lifespan of thunderstorms are small compared with the spatial and temporal resolution of NWP models.

Advances in weather science and high-performance computing have greatly improved the prediction capabilities of NWPs in recent years. In our research, we set out to leverage these improvements, and the great potential of machine-learning techniques (Vinuesa et al., 2020), to predict thunderstorms using NWPs at the timescales (greater than 24 h) required for the pre-tactical phase of air-traffic-flow management (ATFM).

Convolutional neural networks (CNNs) have been used for the analysis of satellite images in a wide range of applications (Jean et al., 2016; Sirmacek & Vinuesa, 2022), including the development of interpretable models which enable obtaining insight into the model structure (Vinuesa & Sirmacek, 2021). For shorter time horizons, CNN-based methods have been used successfully to improve the nowcasting of weather phenomena, and, generally, machine-learning methods have been shown to effectively improve numerical simulations of fluid flows (Vinuesa & Brunton, 2022).

In Han et al. (2019), CNNs were applied to weather radar data to improve the nowcasting of convective weather. In Lagerquist et al. (2020), CNN methods were also used on radar images to predict tornadoes in the next hour. However, predictions at these time scales are incompatible with pre-tactical ATFM operations. Note that machine learning has also been used on NWP data to predict thunderstorms for longer time horizons. In Šaur (2017), NWP and historical weather data were used to train a fully connected network with one hidden layer, to predict convective precipitation that may cause flash floods over the Zlin region of the Czech Republic up to 24 h in advance. In Collins and Tissot (2015), a deep-neural-network model was developed using cloud-to-ground lightning data to predict the occurrence of thunderstorms in certain regions of Texas (US), within two-hour time steps for time horizons of up to 15 h. Random forests have also been applied on NWP to predict the probability of lightning strikes over the Alaskan tundra (He & Loboda, 2020). In Simon et al. (2018), thunderstorm occurrence within a six-hour period was predicted over the European eastern Alps up to five days in advance using generalized additive models (GAMs) and gradient boosting with cloud-to-ground lightning data. Convolutional neural networks have also been applied on NWP tools to predict multiple types of convective weather within six hours up to 72 h in advance (Zhou et al., 2019). While these studies have been successful in using machine learning to predict convective weather, their specific applications did not require spatiotemporal resolution or the continental-scale geographic domain necessary for pre-tactical ATFM applications.

In this paper, we apply machine learning to predict thunderstorm occurrence over a large portion of western Europe, in hourly time steps for time horizons of up to 36 h. An ensemble NWP with 0.25-degree spatial resolution and satellite observations from the EUMETSAT NWC-SAF Rapid-Development Thunderstorm product are used to train a convolutional neural network to provide the likelihood of convective weather up to 36 h in advance. The goal of this research is to provide a pan-European convective weather forecast at time scales compatible with pre-tactical air traffic flow management operations.

Computer-vision methods, such as convolutional neural networks (CNNs) (LeCun et al., 2015), are able to exploit spatial information in the data to improve the predictions. In fact, and thanks to the successive application of convolution operations, the network can hierarchically build progressively more complex features relevant to the predictions at hand. The applicability of CNNs has been extensively demonstrated in the context of fluid mechanics (Guastoni et al., 2021), and even more complex architectures such as the generative adversarial networks (GANs) produce excellent performance in this type of benchmark (Güemes et al., 2021). CNNs and GANs have a great advantage over *e.g.* multilayer perceptrons (MLPs), in which the spatial information in the data cannot be effectively exploited, as also thoroughly assessed in previous work (Srinivasan et al., 2019). Consequently, and with the aim of exploiting the spatial information present in the current datasets, the learning task of predicting convective weather from NWP is formulated as an image-segmentation problem. Image segmentation has been widely applied on tasks related to medical-image analysis, robotic perception and video surveillance (Minaee et al., 2021).

This research aims to improve the forecasting of convective weather by applying an image segmentation approach that combines satellite observations with NWPs. While our model is not trained with the precise labeled image data often used in segmentation problems, here we show that the coupling of NWP grids with satellite images will allow to train the CNN models to correctly classify convective regions. To the authors' knowledge, this is the first time that an image-segmentation methodology has been applied to NWPs. Previous use of CNNs on NWPs have used architectures that conclude with fully connected layers for the task of classification tasks (Zhou et al., 2019). In Weyn et al. (2020) an encoder–decoder architecture is utilized but for the prediction of basic atmospheric variables rather than severe weather. This work presents a novel approach that combines different types of datasets and improves the forecast.

The rest of the paper is organized into the following sections: Section 2 presents an overview of the weather data utilized, Section 3 details the methodology and CNN architectures, with results presented in Section 4. Examples of model application within an ATFM context are presented in Section 5 and lastly, a summary and conclusions are provided in Section 6.

2. Description of the weather database

In developing the present convection-prediction model, data from ensemble NWP forecasts and satellite thunderstorm observations are used. Given the lead times required for pre-tactical ATFM, the model input is provided by ensemble NWP forecasts, as these are available 36 h in advance. Satellite-image data is used for training and evaluation of the model as it provides an accurate representation of convective events. The data used for training and validation is from June and July 2018, respectively, with a geographical domain covering vast portions of western Europe and northern Africa as seen in Fig. 2. Test data is used from dates in July 2018 as well as July 2019 (unseen by the networks), within the aforementioned geographical domain. It is important to note that the NWP grid size corresponding to the geographic domain has 128 × 128 grid points. Since we apply an image-segmentation methodology, this grid size is favorable for performing multiple instances of max-pooling and up-sampling calculations (see Section 3).

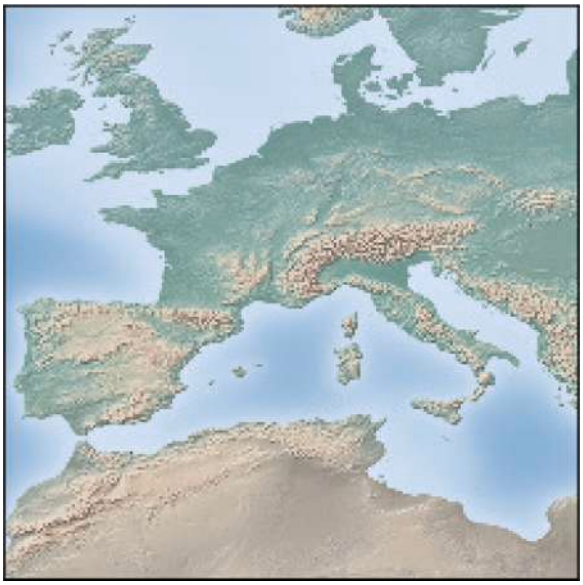


Fig. 2. Geographical domain of forecast and observational weather data.

2.1. Ensemble NWP

The model input is comprised of NWP data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS). Ensemble forecasting is a technique where rather than providing on possible future scenario of the atmosphere as with traditional NWPs, various scenarios are created by running multiple models, perturbing initial conditions, or using different combinations of physical parameterization schemes. Perturbations in parameters are in line with the observation errors in the current state of the atmosphere. By pooling the resulting multiple forecasts it is possible to provide an estimate of the uncertainty associated with predictions of the atmosphere. The ECMWF EPS product comprises 50 individual forecast "members", plus a control member based on the most accurate estimate of the initial conditions. The assumption is that the probability of occurrence of each of the 50 members is equally likely, and while one ensemble member may prove to be most accurate at a given geographical location, this need not be the case at another location (Palmer et al., 2006). The ECMWF EPS releases new forecasts four times per day at 00, 06, 12, and 18 UTC, and can predict the state of the atmosphere for up to 15 days.

The spatial resolution of the forecast is 0.25 degrees in longitude and latitude, equivalent to roughly 15 nautical miles.

The parameters selected to train the model were chosen based on their relevance to the following physical characteristics of convective weather and thunderstorms (CAE, 2015):

- Lifting force or trigger mechanism to produce early saturation of air. In convective storms, this trigger action is typically caused by heat from the Earth's surface causing moist air to rise.
- Sufficient moisture in the atmosphere to form and maintain the cloud.
- Atmospheric instability determined by the vertical temperature profile or lapse rate.

Based on these conditions 18 EPS parameters we selected as inputs to the model. These 18 parameters were chosen to capture the three essential elements of convection; lifting force, moisture, and instability. Lifting force was captured by parameters relating to temperature, pressure, and heat flux. Parameters relating to dewpoint, rain rate, and column water were meant to capture the moisture in the atmosphere.

Table 1
Total list of parameters used to train the models.

Parameter	Short name
2-m dewpoint	2d
2-m temperature	2t
Convective available potential energy	cape
Convective available potential energy 1 h before	cape-1
Convective available potential energy 2 h before	cape-2
Convective available potential energy 3 h before	cape-3
Convective inhibition	cin
Convective precipitation	cp
Convective rain rate	crr
Height of convective cloud top	hcct
Hour of day	hour
K index	kx
Large-scale precipitation	lsp
Large-scale rain rate	lsrr
Surface latent heat flux	slhf
Surface pressure	sp
Surface sensible heat flux	sshf
Range of forecast	range
Total cloud cover	tcc
Total column water	tcw
Total column water vapor	tcwv
Total totals index	totalx
Geopotential	z

Lastly, instability was obtained from parameters such as convective available potential, convective inhibition, K index and total totals index. Apart from the 18 EPS parameters, additional parameters *hour of the day* and *range of forecast* were also included to account for diurnal weather patterns and capture any range-dependent bias that could exist. Additionally, the convective-available-potential-energy (CAPE) parameter values from the three previous time steps were also included, given that large values of CAPE are common during periods leading up to the storm. A total of 23 input parameters (18 EPS parameters, one hour of day, one range of forecast, and three time-lagged CAPE) were chosen as input to the model, the complete list is provided in Table 1.

2.2. Satellite data

Target data for the model is provided using the Rapid-Development Thunderstorm (RDT) product developed by Météo-France within the EUMETSAT NWC-SAF framework. The RDT algorithm makes use of geostationary satellite data to monitor and track active convective cells in the atmosphere. The product outputs information on a 15-minute time interval related to convective clouds from the mesoscale (200–2000 km) down to hundreds of meters (Lee et al., 2020). The RDT algorithm is capable of capturing the shape, cloud top, movement, and severity of convective cells. Despite the rich characterization provided by the RDT product, our model was formulated as a binary classifier and only consider the shape and location of the convective cells.

2.3. Experimental data set

ECMWF EPS forecast data was blended with RDT observations to create the training, validation, and test data sets. The entire experimental data set consisted of forecasts and observations from summer 2018. A description of the exact days selected for each of the training, as well as the validation and testing data subsets, is provided in Table 2. Note that the test data is taken from dates unseen by the network during training.

In blending the data, hourly EPS data from the 00 and 12 UTC forecast releases up to 36 h from the 50 perturbed forecast members was considered. The RDT images were aggregated hourly to be consistent with the EPS time step. By overlaying the RDT convective cell polygons over the EPS grid, it was possible to provide a binary classification of the grid points where a convective cell was present during the hour.

Table 2
Data sets used for training, validation and testing.

Dataset	Date range	Number of days
Training	June 1–30, 2018	30
Validation	July 14–16, 2018	3
Testing	July 20–26, 2018	7

Fig. 3 provides an example of how four RDT images are processed to establish the target function. Given the forecast range of 36 h, and the forecasts release frequency of 12 h, different range forecasts valid for the same time were used to train, validate and test the model. Having data with varying forecast ranges allowed us to analyze how the forecast degrades with an increasing time horizon.

3. Methodology

The objective of this research is to explore how machine learning models with convolutional neural network architectures can be applied on weather data to improve the prediction of thunderstorms. The problem is formulated similarly to an image segmentation task, where the model must learn to classify specific regions within the image. However, rather than using an RGB image with 3 color channels, the input is provided using N channels, where N represents the number of NWP parameters. For the output, rather than training with labeled segmented images, the model is trained with convection observations. A visual comparison of the methodology used for convection prediction with the traditional image segmentation approach is provided in Fig. 4.

In this article we explore several CNN-based architectures including a fully-convolutional network, a CNN-based encoder–decoder, a U-Net, and a pyramid-scene parsing network. Within the methodology section, a basic overview of how layers and operations are done within a CNN-based model is provided the first subsection. Next, we provide a description of the four model architectures explored in this study along with architectural schematics in Figs. 4, 5, 6, and 7. Lastly, details on the computer software and hardware utilized during the training of the models is provided.

3.1. Convolutional neural networks

In this study, we employ CNN-based architectures for the classification of convective and non-convective regions. The objective is to exploit the spatial information in the input data through the convolutional layers and use it for classification. As the input fields exhibit coherent features and spatial correlations, learning such features should lead to improved accuracy and generalization. At a given two-dimensional convolution layer, the input is convolved with a filter of size $H \times W \times K$, where K is the number of kernels, which are the two-dimensional slices of the filter. This is mathematically expressed as:

$$z_{ijm}^{(l)} = \varphi\left(\sum_{k=1}^{K}\sum_{p=1}^{H}\sum_{q=1}^{W} z_{i+p,j+q,k}^{(l-1)} w_{pqkm}^{(l)} + b_{ijm}^{(l)}\right), \tag{1}$$

where $z^{(l-1)}$ and $z^{(l)}$ indicate the inputs and outputs of lth layer, respectively, and m denotes the number of output channels. Each pixel is represented by (i, j, k) and (i, j, m) for the input and output feature maps, respectively. Furthermore, $w^{(l)}$ and $b^{(l)}$ represent weights and biases of the lth layer, respectively, and φ denotes the non-linear activation function. Usually, the number of filter kernels K changes with the depth of the network. To obtain a feature map with M channels at the output, we need $H \times W \times K \times M$ trainable variables as $w^{(l)}$.

A schematic of the convolution operation denoted by $*$ is illustrated in Fig. 5(a). It is common to combine convolutional layers with pooling and upsampling operations. The pooling operation compresses the data by a factor of $(1/P)^2$ so that a region with the size of $P \times P$ is

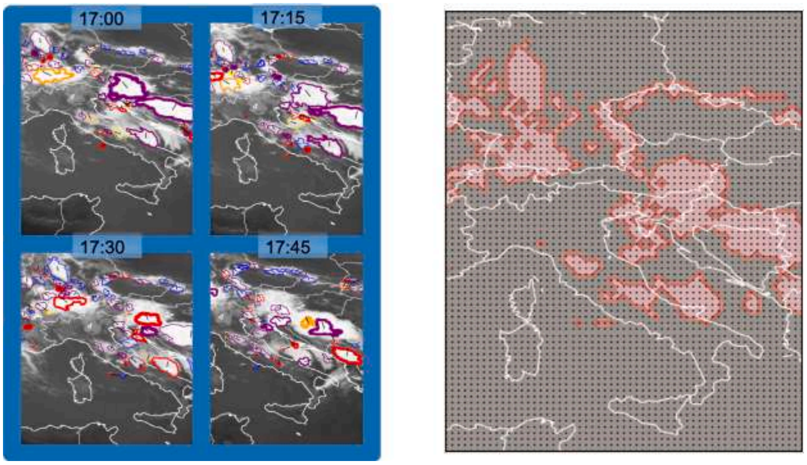


Fig. 3. (Left) RDT satellite observations and (right) resulting target function for thunderstorms occurring at 17:00 on June 8th, 2018.

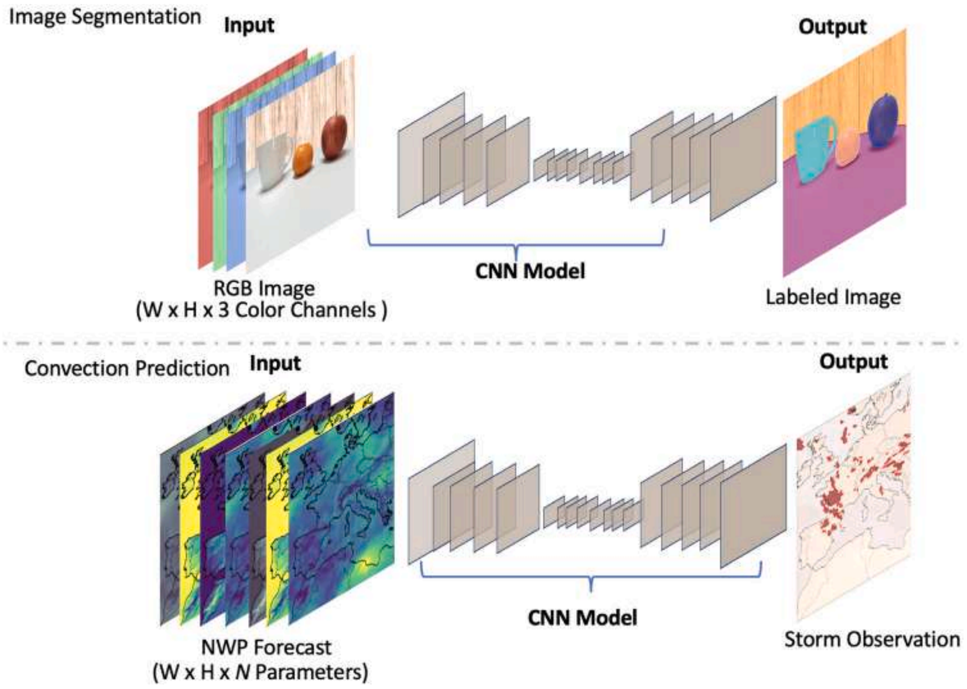


Fig. 4. Methodology for prediction convection using Convolutional Neural Network is based on an image segmentation approach.

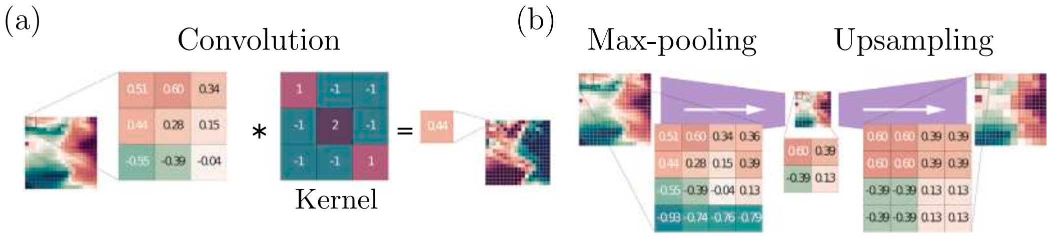


Fig. 5. A schematic view of (a) convolution operation and (b) max-pooling and upsampling operations.

represented by its maximum or mean value, which correspond to max- or average-pooling operations, respectively. Moreover, the upsampling operation increases the data size by a factor of P^2, e.g. through a nearest-neighbor or bilinear interpolation. Schematic representations of max-pooling and upsampling operations are depicted in Fig. 5(b). For all the CNN models in this paper, we consider two-dimensional snapshots of the 23 variables from the NWP forecast (see Table 1) as the inputs, and the binary images representing convective and non-convective classes as the outputs, as shown in Fig. 3(b).

3.1.1. *Depthwise separable convolutions*

A convolution layer aims at learning filters in a three-dimensional space; two spatial dimensions and a channel dimension. Therefore, a convolution kernel maps both spatial correlations and cross-channel

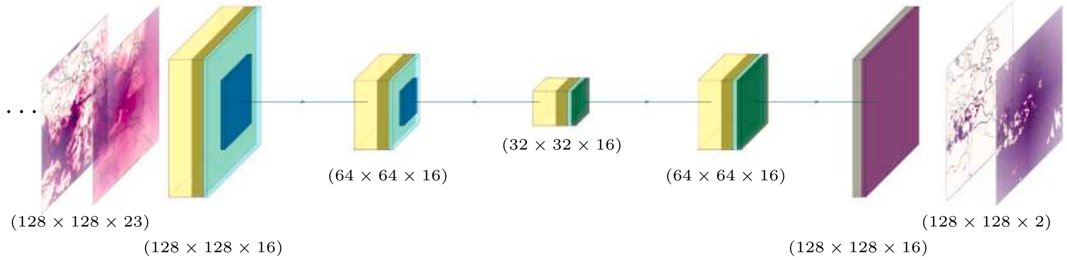


Fig. 6. A schematic view of the Enc-Dec model. The color coding for each layer is: 2D separable convolution (🟨), ReLU activation (🟫), spatial dropout (🟩), max pooling (🟦), upsampling (🟥), and Sigmoid activation (🟪). The numbers in brackets denote the size of the feature maps.

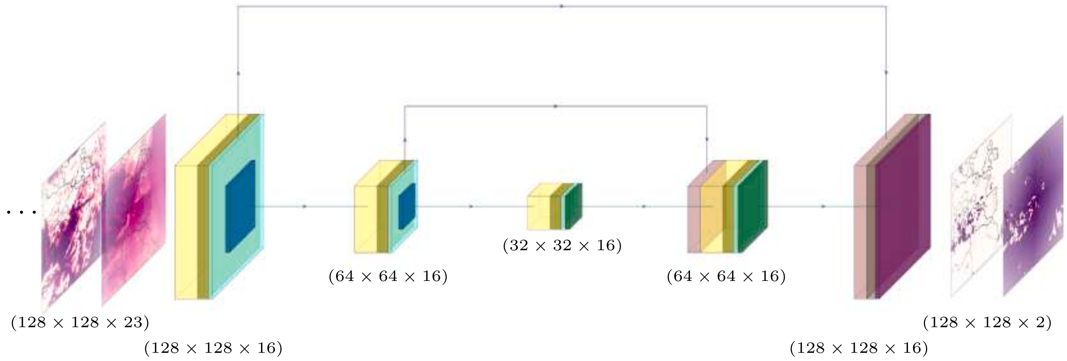


Fig. 7. A schematic view of the U-Net model. The color coding for each layer is: 2D separable convolution (🟨), ReLU activation (🟫), spatial dropout (🟩), max pooling (🟦), upsampling (🟥), concatenation (🟥), and Sigmoid activation (🟪). The connections on the top concatenate the present layer with features from previous layers.

correlations simultaneously. The main idea behind *depthwise separable convolutions* (Chollet, 2017) is to perform an efficient convolution operation by explicitly factoring it to a series of operations accounting for spatial correlations and cross-channel correlations independently. A depthwise separable convolution, also known as *separable convolution*, consists of a spatial convolution performed independently over each channel of the input followed by a 1 × 1 convolution (the so-called *pointwise convolution* (Lin et al., 2014)) which accounts for cross-channel correlations. This can be mathematically expressed as:

$$\begin{aligned} s^{(l)}_{ijk} &= \sum_{p=1}^{H}\sum_{q=1}^{W} z^{(l-1)}_{i+p,j+q,k}\,\mathcal{W}^{(l)}_{pqk}, \\ z^{(l)}_{ijm} &= \varphi\left(\sum_{k=1}^{K} s^{(l)}_{i,j,k}\,\omega^{(l)}_{km} + b^{(l)}_{ijm}\right), \end{aligned} \tag{2}$$

where $s^{(l)}$ indicates the output of spatial convolution. $\mathcal{W}^{(l)}$ and $\omega^{(l)}$ represent the trainable weights for the spatial and pointwise convolutions, respectively. To compute a feature map with M channels at the output of depthwise separable convolutions, we need $H \times W \times K + K \times M$ trainable weights; this leads to a lower number of parameters and a reduced computational cost in comparison with conventional (non-separable) convolution.

Here, we employ depthwise separable convolution in all the CNN-based models to reduce the model complexity and avoid overfitting. Hereafter, we refer to the depthwise separable convolution as separable convolution.

3.2. Models

In this paper, four models are considered: a fully-convolutional network, a CNN-based encoder–decoder, a U-Net, and a pyramid-scene parsing network. These four networks were chosen to explore the various possibilities of exploiting spatial information in the data with architectures of different but complementary capabilities. The fully-convolutional network is the basic model involving convolutions, which should serve as an adequate baseline to assess the prediction capabilities of these computer-vision-based strategies. As opposed to the standard CNN, in the CNN-based encoder–decoder (Eivazi et al., 2022) the data is first downsampled and then upsampled while the subsequent convolutions are being applied. This allows to first focus on the most essential features from the input data, and when the smallest dimension is reached these features are again increased and highlighted, adding additional detail in the generation of patterns. The U-net is characterized by skip connections, in which the feature map from the first layers is fed directly into the last ones. This allows to combine simpler and more complex features when reaching the last layers, and also exhibits benefits in the context of the back-propagation process. Finally, the pyramid-scene-parsing network (PSPNet) relies on the so-called pyramid-parsing module (PPM) to combine local information with features characteristic of the global features in the input data. This combination of features and scales has the potential to produce more nuanced and accurate predictions. The strengths and weaknesses of the various methods applied to our particular problem will be discussed in the next sections.

3.2.1. Fully-convolutional neural networks

We implement a fully-convolutional network (FCN) as the first CNN-based model. Here, we use three blocks of separable convolutions followed by spatial dropouts of fraction 0.2 (without down-sampling or upsampling) to process the input data, and at the final layer, a 1 × 1 convolution with a sigmoid activation function to map the feature maps to the pixel-level binary classes. We consider two channels at the output: in the first channel, a pixel value of one represents the convective region and zero indicates the non-convective region. For the second output channel, we consider the opposite, i.e., a pixel value of one represents a non-convective region and zero represents a convective region. Each separable convolution consists of 16 filters with a kernel size of $H \times W = 3 \times 3$ and rectified linear unit (ReLU) as the activation function. Same class representation and final layer architecture are used for other CNN-based models.

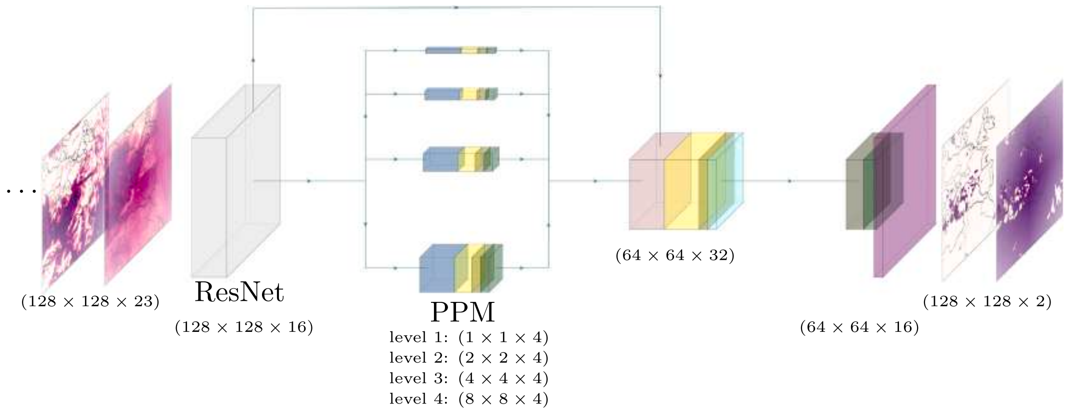


Fig. 8. A schematic view of the PSPNet model. The color coding for each layer is: ResNet model for feature extraction (⬜), 2D separable convolution (🟨), average pooling (🟦), upsampling through bilinear interpolation (🟥), ReLU activation (🟫), sigmoid activation (🟪), spatial dropout (🟩) and concatenation (▨).

3.2.2. CNN-based encoder–decoder architecture

We employ a CNN-based encoder–decoder (Enc-Dec) architecture as the second model. A schematic representation of this model is depicted in Fig. 6. As it can be seen, we use repeated application of a separable convolution layer consisting of 16 filters with a kernel size of $H \times W = 3 \times 3$ and rectified linear unit (ReLU) as the activation function, followed by a spatial dropout of fraction 0.2 and a max-pooling operation with $P = 2$ for down-sampling through the encoder. We keep the number of feature channels equal to 16 throughout the network. The upsampling process is performed through the decoder part, which comprises two consecutive steps of a 3 × 3 separable convolution with a ReLU activation function, followed by a spatial dropout of fraction 0.2 and an upsampling operation using nearest-neighbor interpolation. At the final layer, a 1 × 1 convolution with a sigmoid activation function is implemented to map the feature maps to the desired number of classes.

3.2.3. U-Net architecture

In an encoder–decoder architecture, the implementation of multiple down-sampling steps in the encoder allows the extraction of larger and more complex features that are useful for classification. However, this leads to a loss of spatial resolution of the feature maps through the pooling operations which can reduce the localization accuracy. Since deeper-layer feature maps contain more semantic meaning and less location information, combining multi-scale feature maps can improve performance. Therefore, as the third CNN-based model, we employ the so-called U-Net architecture for the classification of convective and non-convective regions. U-Net was first proposed by Ronneberger et al. (2015) for biomedical-image segmentation. The main idea is to combine the high-resolution features from each step of down-sampling with the corresponding upsampled feature maps to improve the localization. Fig. 7 shows a schematic representation of the U-Net architecture implemented in this study. The network architecture differs from the CNN-based encoder–decoder model only in the connections between the encoder and decoder. Through each upsampling step, the concatenated feature maps are processed by a separable convolution layer with a ReLU activation function, a spatial dropout of fraction 0.2, and an upsampling operation.

3.2.4. Pyramid-scene-parsing network

The next CNN-based architecture we employ is a pyramid-scene-parsing network (PSPNet) (Zhao et al., 2017), motivated by its state-of-the-art performance on semantic-segmentation tasks (Zhou et al., 2018). The main idea behind the PSPNet is to exploit global image-level context information by different-region-based context aggregation using the so-called pyramid-parsing module (PPM). This leads to the incorporation of global and local information together for prediction, which improves performance. Following the original architecture of the PSPNet (Zhao et al., 2017), we use a CNN-based residual-neural-network (ResNet) (He et al., 2016) architecture to extract the features. The extracted features are passed to a pyramid-parsing module to perform pixel-level classification. The ResNet model comprises of two consecutive 3 × 3 separable convolutions, followed by a residual summation, a ReLU activation, a spatial dropout of fraction 0.2, and a max-pooling operation with $P = 2$. A schematic of the PSPNet implemented in this study is depicted in Fig. 8. We employ a four-level pyramid-pooling module that fuses features at four different scales. The coarsest level is a global pooling which leads to a single bin output. For the finer levels, i.e. 2, 3 and 4, we consider kernel sizes of 2 × 2, 4 × 4, and 8 × 8, respectively. At each level, we perform an average-pooling operation followed by a 1 × 1 separable convolution that reduces the number of channels by $1/N$, where N is the level size of the pyramid. Then we perform a ReLU activation and an upsampling operation that directly upsamples the low-dimensional feature maps to the size of the ResNet feature maps using bilinear interpolation. Then, we concatenate the PPM feature maps as the global-context information with the ResNet feature maps followed by a 3 × 3 separable convolution, a ReLU activation, and a spatial dropout of fraction 0.2. In the final part, the feature maps are processed by a 1 × 1 convolution, an upsampling operation using bilinear interpolation, and a sigmoid activation to generate the final prediction maps.

3.3. Computer implementation

All the deep-learning algorithms are developed using Python 3.8.8 and TensorFlow 2.4.1, and all the training procedures for CNN-based models are executed on a server with four NVIDIA GeForce RTX 2080 Ti GPUs, each with 11 GB memory, and CUDA 11.6.

4. Results

In this section, we present a comparison of results for the CNN-based models over the seven days in our test data set. Additionally, results are also compared with a point-based MLP model architecture presented in previous research (Jardines et al., 2021). The MLP model in this paper is trained with the same data set used for the CNN-based methods. The point-based MLP model consists of one input layer of 23 nodes, three hidden layers of 16 nodes each, and one output layer of a single node. Each hidden layer is followed by a dropout of fraction 0.2. In our discussion, this MLP model will serve as a baseline to compare the CNN-based model performance. The aim of this research was to explore machine learning model architectures best suited to predict weather phenomena. In order to best capture the effect of the model architecture on the performance it was decided to compare results with previous research that utilized the exact same data set.

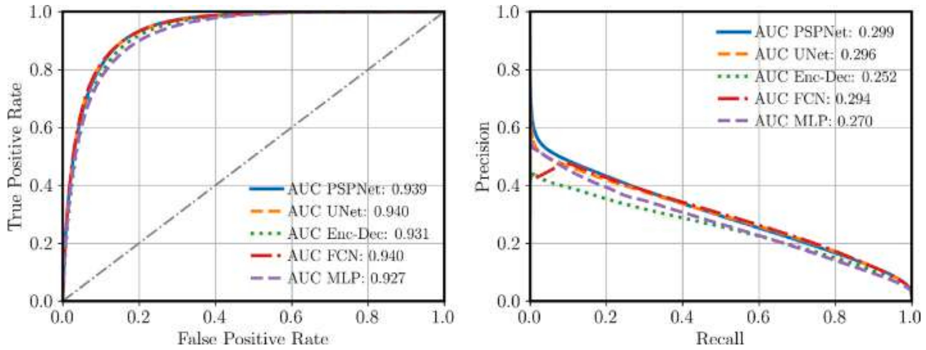


Fig. 9. Receiver operating characteristic (ROC) curve (left) and precision–recall curve (right) comparing multiple models for the entire test data set.

4.1. Model comparisons

4.1.1. Comparison based on probabilistic representations

The effectiveness of the CNN-based convection indicators is compared with the baseline MLP indicator using a receiver operating characteristic (ROC) and precision–recall curves. The ROC curve evaluates a binary classifier by plotting the sensitivity, or true positive rate (TPR), against (1-specificity), or the false positive rate (FPR), for various threshold settings (Mandrekar, 2010). The probability of detection is provided by the TPR, while the FPR provides the probability of false alarms.

However, the ROC curve does not provide information about false negative predictions corresponding here to the cases where a pixel belonging to the convective class is falsely predicted as non-convective, which can be a crucial factor for ATM applications. Another useful indicator for the comparison of the models is the precision–recall curve, which shows the tradeoff between precision and recall for different thresholds. A high AUC for the precision–recall curve represents both high precision and high recall, where high precision indicates a low FPR, and high recall shows a low false negative rate (FNR).

With both metrics, the ideal classifier will maximize the area under the curve (AUC).

In Fig. 9 we show a comparison between the results of the CNN-based models presented here and the baseline MLP indicator for the seven days in the test data set. Results are reported in the form of the ROC curves (left), the precision–recall curves (right), and their AUCs. From Fig. 9 (left), it is evident that all the CNN-based models outperform the baseline MLP indicator based on the ROC AUC. Moreover, because the CNN curves are always above the baseline MLP curve, they outperform the baseline independently of the chosen threshold value. In Fig. 9 (right), it can be observed that all the CNN-based models, except the Enc-Dec model, provide a higher value for the precision–recall AUC. The PSPNet curve is always above the MLP curve independently of the chosen threshold value and shows the highest AUC. It is important to note that the AUC value depends on the particular data set under analysis. The NN models are good at identifying areas without convection (true negatives), thus, analysis of days with few convective storms will yield greater AUC values.

4.1.2. Decision making

Probabilistic representation of the indicator allows the user to evaluate the risks in making a decision. However, in a real scenario, a general framework for discussion-making may be required. Here, we propose an approach for thresholding and binary representation of the model predictions. To this end, we employ F_β-score, the weighted harmonic mean of precision and recall, as:

$$F_\beta = (1+\beta^2)\frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}, \quad (3)$$

where $\beta \geq 0$ introduces a weighting for recall in the combined score. A value of $\beta < 1$ gives more weight to precision, whereas a $\beta > 1$ lends more weight to recall in the calculation of the score. Considering more weight for recall in the calculation of the F_β-score is beneficial for applications in which reducing the FNR is crucial. F_β-score reaches its optimal value at 1 and its wort value at 0. The thresholding value is computed based on the maximum F_β-score over the training data set.

We utilize the intersection over union (IoU) or the Jaccard index, which is one of the most commonly used metrics for semantic segmentation (Minaee et al., 2021), to evaluate the model predictions after thresholding the data. It is defined as the area of the intersection between the predicted segmentation map for storm class and the reference map, divided by the area of their union:

$$\text{IoU} = J(\text{ref.}, \text{pred.}) = \frac{|\text{ref.} \cap \text{pred.}|}{|\text{ref.} \cup \text{pred.}|}, \quad (4)$$

where ref. and pred. indicate the reference and the predicted segmentation maps for the storm class, respectively.

Fig. 10 shows the probability density function (PDF) of different score metrics over the testing data set obtained from different models and three values of β, i.e. 0.5, 1.0, and 2.0. The metrics are calculated per each snapshot of the data, and the PDFs show the distribution of the metrics over the whole testing data set. The vertical dashed lines show the mean values. Fig. 10(a) depicts the PDF of the ROC AUC over the testing data set; it can be seen that all the CNN-based models outperform the MLP model with respect to the ROC AUC, with higher mean and higher PDFs for larger ROC AUCs. It should be noted that the ROC AUC is computed using the raw indicator values and does not depend on the value of β. In Fig. 10(e~g) we illustrate the PDFs of the F_β-score for β of 2.0, 1.0 and 0.5, respectively. The superior performance of the PSPNet, UNet and FCN is evident in comparison with the MLP and Enc-Dec models for all the values of β. The same conclusion can be drawn from Fig. 10(b~d) where the PDFs of IoU are depicted. It can be seen that employing a β value of 2.0, which lends more weight to recall, leads to a better performance with respect to the IoU of the final thresholded predictions. For β of 2.0, superior performance of PSPNet, UNet, and FCN is evident in comparison with the MLP and Enc-Dec models, and FCN leads to the best performance with a mean IoU of 16.40 over the testing data set.

In Fig. 11 we illustrate the results obtained from different models and β of 2.0 for one of the samples in the testing data set. From the left, columns represent the normalized raw model predictions, the thresholded predictions, the reference data, and the pixel-level illustration of the true positive, true negative, false positive, and false negative predictions. Each row shows the results obtained from one of the models. It can be seen that for this sample, the FCN model provides the best ROC AUC and F_β-score equal to 0.930 and 0.58, respectively, while the MLP model leads to the ROC AUC of 0.903 and F_β-score of 0.51. The best IoU is obtained from the UNet model and it is equal to 25.14%. The FCN model leads to a slightly lower IoU of 25.03%. The lowest IoU is obtained from the MLP model and it is equal to 20.75%.

In Fig. 11(d) the results are reported as the pixel-level illustration of the true positive (white), true negative (light green), false positive

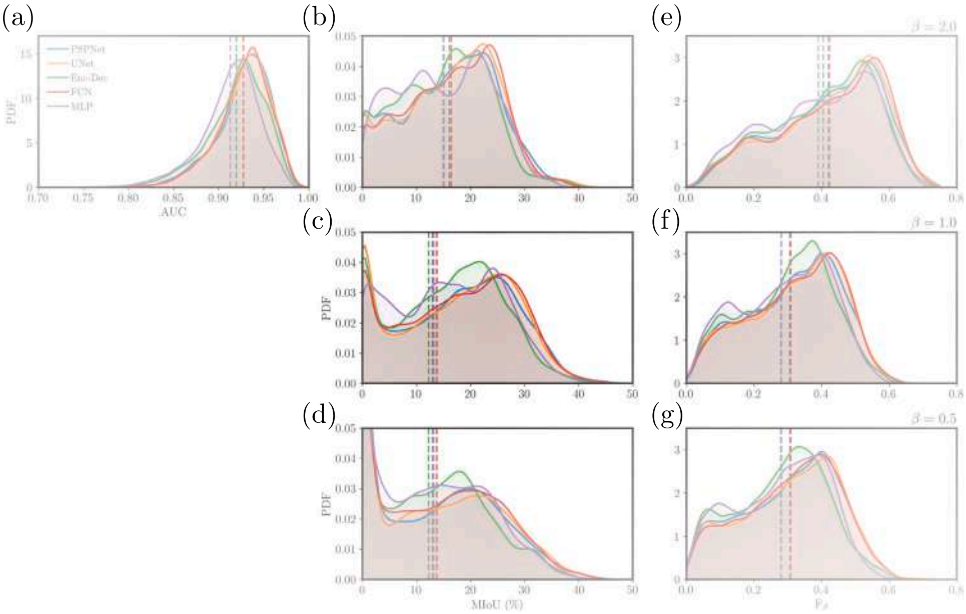


Fig. 10. Probability density function (PDF) of the ROC AUC, F_β-score and IoU metric over the testing data set obtained from different models and three values of β, i.e. 0.5, 1.0, and 2.0.

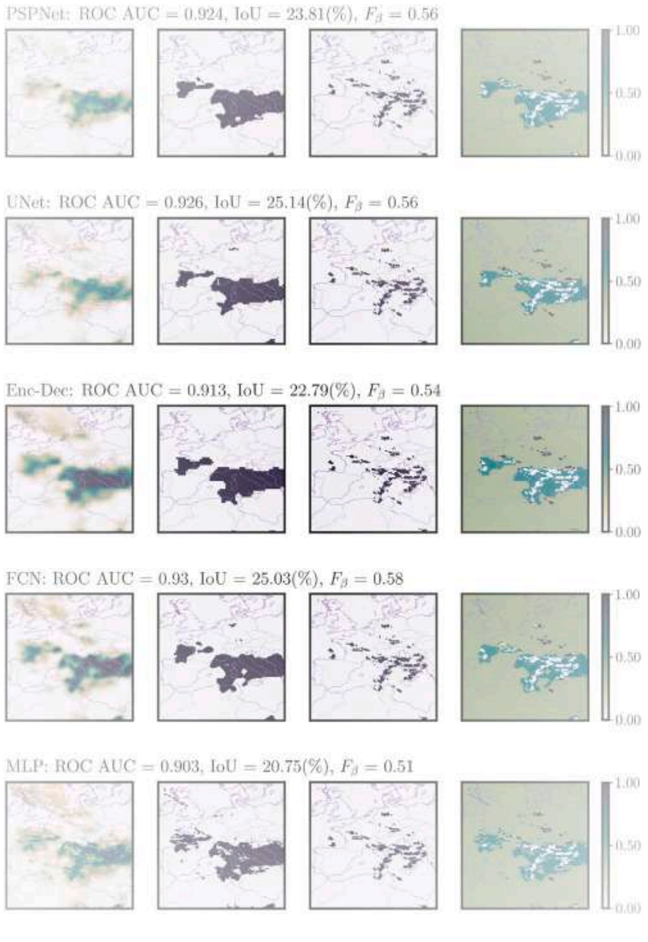

PSPNet: ROC AUC = 0.924, IoU = 23.81(%), F_β = 0.56
UNet: ROC AUC = 0.926, IoU = 25.14(%), F_β = 0.56
Enc-Dec: ROC AUC = 0.913, IoU = 22.79(%), F_β = 0.54
FCN: ROC AUC = 0.93, IoU = 25.03(%), F_β = 0.58
MLP: ROC AUC = 0.903, IoU = 20.75(%), F_β = 0.51
1.00
0.50
0.00

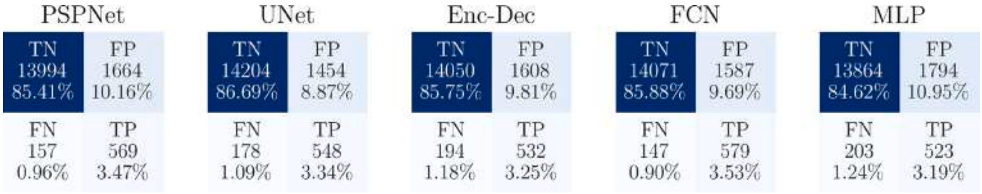
Fig. 11. Results obtained from different models and β = 2.0 for one of the samples in the testing data set. From the left, columns represent the normalized raw model predictions, the thresholded predictions, the reference data, and the pixel-level illustration of the true positive (white), true negative (light green), false positive (dark green), and false negative (black) predictions.

(dark green) and false negative (black) predictions. In general, very good performance of all the models in capturing the convective regions can be observed; however, for instance, the FCN model outperforms the MLP model by providing higher TPR and lower FPR and FNR. This can be observed in Fig. 12 where we reported the confusion matrices corresponding to the data represented in Fig. 11.

Fig. 13(a) illustrates the final thresholded predictions obtained from the FCN model using different values for β with the same color coding as Fig. 11(d). Fig. 13(b) shows the same results in a comparable fashion together with the reference convective region (colored in black). It can be seen that increasing the value of β, which gives recall more weight against precision in computing the F_β-score, leads to the reduction of false-negative predictions but an increase in the false-positive predictions. This can also be observed in Fig. 14 where we reported the confusion matrices obtained using different values of β corresponding to the data presented in Fig. 13. Results obtained using β = 0.5 show the excellent performance of the FCN model in predicting the operable region corresponding to true-negative predictions with TNR of 92.80% and FPR of 2.76%; however, utilizing β = 0.5 leads to 2.75% of FNR. By employing β of 2.0 the FNR drops to less than 1%, but TNR decreases to 85.88%, and FPR increases to 9.69%. This tradeoff is a significant factor for AFTM applications representing the tradeoff between efficiency and safety, and it can be optimized by selecting an appropriate value for β.

4.2. Sensitivity to the prevalence of convection

Our results in Fig. 10(b~d) show that for a considerable number of samples in the testing data set the model predictions lead to very small values for IoU. It because some of the samples in the testing data set do not contain any convective regions. Let us define p as the ratio of the number of pixels belonging to the convective class to the total number of pixels for each sample (128 × 128). We observed that for very small values of p the obtained IoU could be low. This can be observed in Fig. 15 where we depict the joint PDF of IoU and p for the samples in the testing data set obtained from the FCN and MLP predictions. The PDFs of p and IoU are presented on the top and the right axes of the plot, respectively. It can be seen that for the samples with higher values of p very good IoUs can be obtained from both models. Moreover, it can be observed that FCN outperforms MLP by providing higher density for higher IoUs.



PSPNet	
TN 13994 85.41%	FP 1664 10.16%
FN 157 0.96%	TP 569 3.47%

UNet	
TN 14204 86.69%	FP 1454 8.87%
FN 178 1.09%	TP 548 3.34%

Enc-Dec	
TN 14050 85.75%	FP 1608 9.81%
FN 194 1.18%	TP 532 3.25%

FCN	
TN 14071 85.88%	FP 1587 9.69%
FN 147 0.90%	TP 579 3.53%

MLP	
TN 13864 84.62%	FP 1794 10.95%
FN 203 1.24%	TP 523 3.19%

Fig. 12. Confusion matrices obtained from different models corresponding to the data presented in Fig. 11.

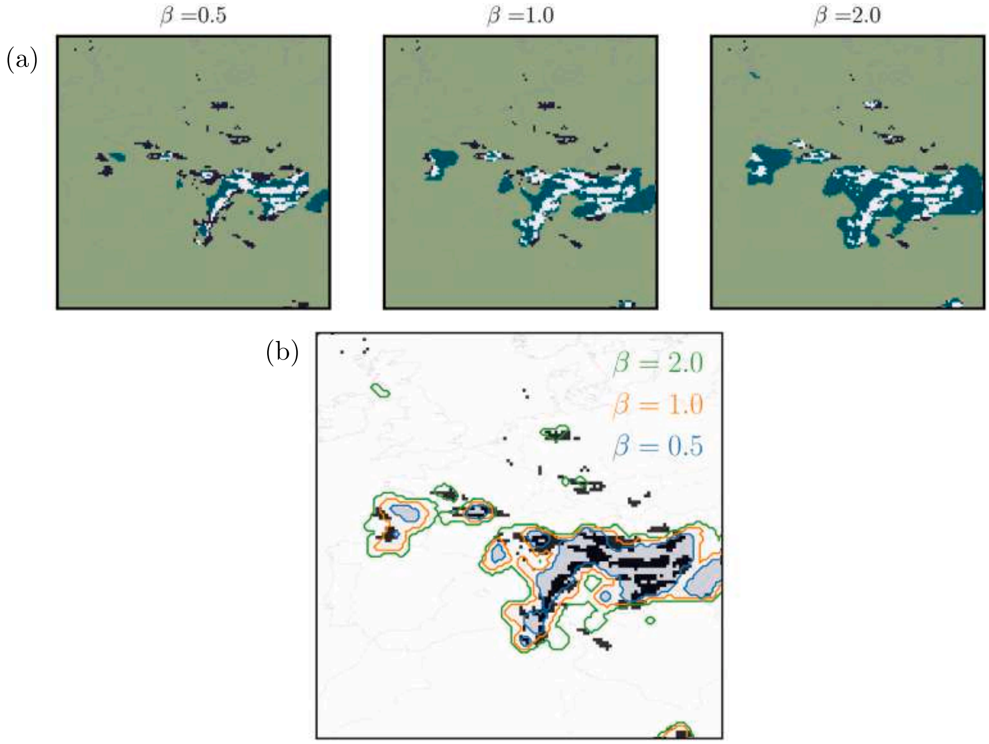
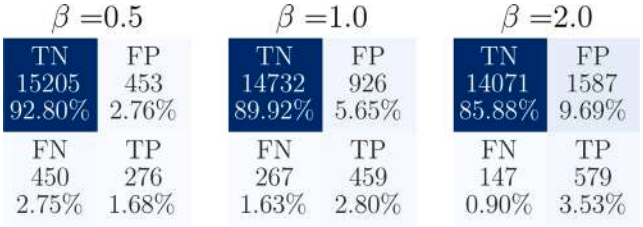


Fig. 13. Final thresholded results obtained from different values of β for one of the samples in the testing data set; (a) the pixel-level illustration of the true positive (white), true negative (light green), false positive (dark green) and false negative (black) predictions, and (b) the predicted convective class in a comparable fashion together with the reference convective region (colored in black).



β =0.5	
TN 15205 92.80%	FP 453 2.76%
FN 450 2.75%	TP 276 1.68%

β =1.0	
TN 14732 89.92%	FP 926 5.65%
FN 267 1.63%	TP 459 2.80%

β =2.0	
TN 14071 85.88%	FP 1587 9.69%
FN 147 0.90%	TP 579 3.53%

Fig. 14. Confusion matrices obtained from the FCN model predictions thresholded using different values of β and corresponding to the data presented in Fig. 13.

4.3. Sensitivity to temporal information

In this section, we are interested in assessing the impact on the temporal correlation of the training and test data sets. While the results presented above show the predictions for a week in July 2018, we also apply our models to data from July 2019. A second test data set was constructed based on NWP and satellite data from July 23–29, 2019. Results are summarized in Fig. 16 in the form of the ROC curves (left), the precision–recall curves (right), and their AUCs. It can be observed that although the value of AUCs slightly drops for this test, very good predictions can be obtained leading to a ROC AUC of above

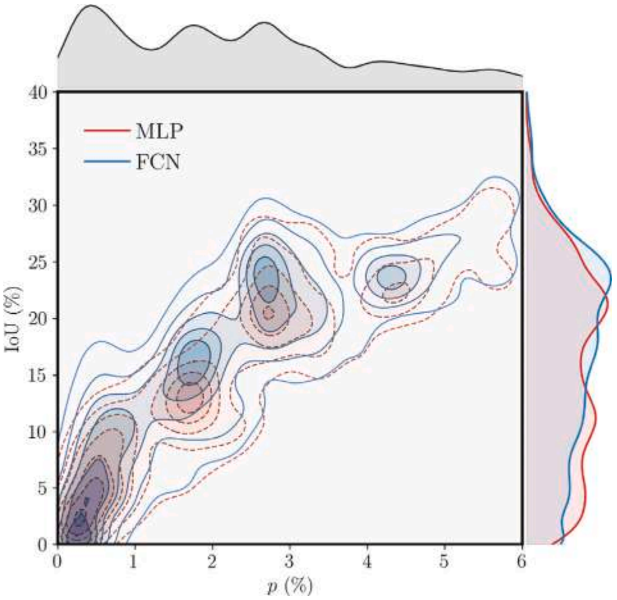


Fig. 15. Joint PDF of IoU and p for the samples in the testing data set obtained from the FCN and MLP predictions. The PDFs of p (black) and IoU are presented on the top and the right axes of the plot, respectively.

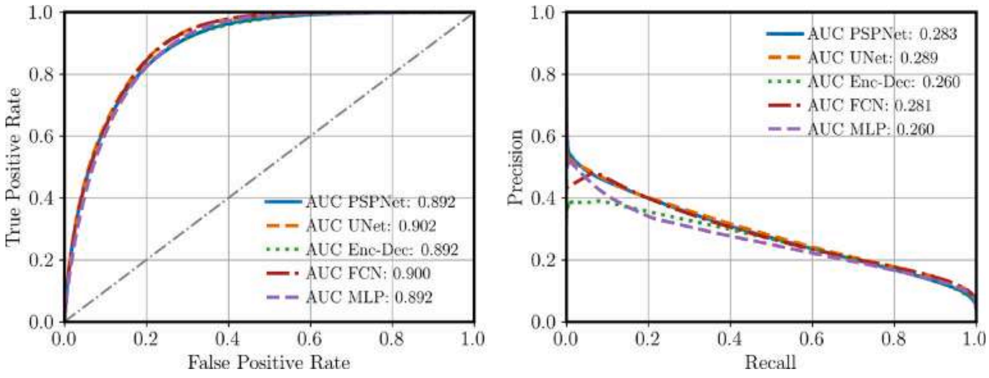


Fig. 16. ROC curve (left) and precision–recall curve (right) comparing multiple models for the entire second test data set from July 23–29, 2019.

0.89 for all the models. Moreover, this test shows the generalizability of the models on unseen data. The best ROC and precision–recall AUCs are obtained from the UNet model and they are equal to 0.902 and 0.289, respectively, while the MLP and Enc-Dec models lead to the lowest AUCs of 0.892 and 0.260 for ROC and precision–recall curves, respectively. While the results in Fig. 8 show a better performance of the CNN models than the MLP baseline more clearly, the results in Fig. 15 indicate that the performance across the CNN models, as well as the MLP baseline, is rather similar. This means that the lack of seen information dominates the performance of the models more than the choice of models. To further verify this claim, it is possible to test the models using data from another (unseen) week, e.g., July 2020.

5. ATFM application

The objective of this work is to provide air traffic managers with an awareness of where and when convective weather will develop. In this section, we present an example of a possible application of the machine-learning-based indicator in an ATFM operational setting. A sample case study is presented for July 26, 2019, a date within our test data set corresponding to strong convective activity within the European Civil Aviation Conference (ECAC) region. Post-operational data from Eurocontrol shows that on July 26th, 2019, 125 weather regulations were activated, resulting in over 120,000 min of ATFM delay.

Fig. 17 provides a sample case study of the timeline of weather information available leading up to the day of operations. In this study, we utilize results from the FCN model to provide an hourly prediction of convection activity during 10:00–13:00 UTC on July 26th, 2019. Figs. 17(a)–17(c) provide the FCN-based convection forecast that would have been made available at various release times leading up to the selected time period. A comparison with Fig. 17(d) shows that the model predictions correlate well with the RDT observations for the same period of time. While discrepancies do exist between the prediction and observation, the model is able to provide a fairly stable prediction of where and when storms will be present at lead times of up to 36 h. The stability of the prediction is important to air traffic managers, as it enables the possibility of taking decisive actions earlier in the planning process.

Lastly, in Fig. 17(e) sectors with active weather regulations during each hourly time period are illustrated.

In this qualitative example, one can notice that while there is some correlation between storms and regulated sectors, the problem is more complex than simply overlapping sectors and convective weather regions. Other factors regarding traffic demand, complexity, and sector capacity must also be considered. Furthermore, additional analysis is warranted to consider the applied capacity rates of each regulation. While this is merely a qualitative illustration, it is hypothesized that a convection indicator based on machine-learning techniques could provide traffic managers and flight dispatchers an early identification of problematic areas in the network as well as assistance in developing strategic mitigation measures to minimize the impact of weather on operations.

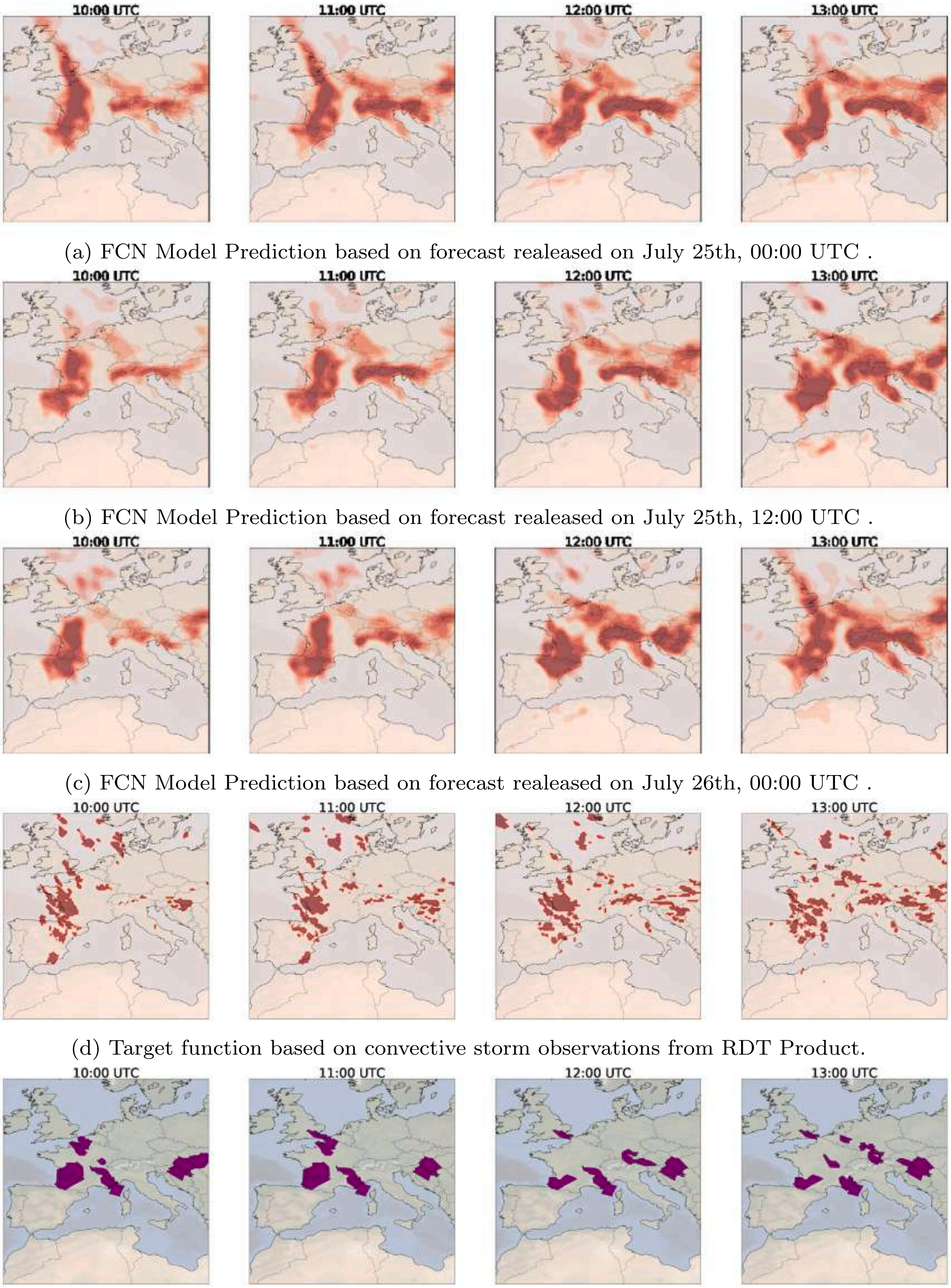
These results indicate to be an improvement over the current Cross-border Convection Advisory Product, (See Fig. 1), by providing predictions with hourly temporal resolutions and with longer lead times. Although additional analysis would be required to translate the model results into a format that is consistent with the standardized convection risk matrix (see Fig. 18) used in operations, the implementation of this work has great potential benefits. A machine learning-based approach allows for increased automation and digitization in the creation of aviation weather forecasts. Continuous training and evaluation of the ML algorithms using live feeds of NWP and observation data would equip decision-makers with the most current and precise weather information. Improved weather information earlier in the ATFM process would enable more coordinated and strategic mitigation strategies, allowing ANSPs to exercise finesse when applying weather regulations, more easily identify areas in the network with capacity for re-route opportunities, and minimize unnecessary delays.

6. Conclusions and outlook

In this study, convolutional neural networks have been applied to weather data for the creation of a machine-learning-based convection prediction methodology. Multiple models with convolutional architectures were trained on the same data set consisting of weather forecast and storm observations from satellite. A comparison of results showed that the best performance was obtained with the FCN model.

In this paper, we have demonstrated that convolutional methods provide an improved classification performance over a previously developed point-based model using an MLP neural network architecture. Furthermore, the ability of the model to generalize on unseen data is demonstrated by predicting on an entirely separate data set from 2019.

In the future research focus will be placed on making better use of the NWP data by formulating a model input that accounts for the distribution of the ensemble, rather than treating each member independently. Additionally, follow-on versions of this work must also consider the temporal relationships within the data by incorporating model architecture that can account for time series data, such as recurrent networks. Additionally, we plan to integrate physical knowledge into the loss functions by using physics-informed neural networks, which would leverage additional constraints on the optimization space explored by the networks during training and potentially improve their performance. There also exist limitations in the employed data sources. The use of additional parameters and higher-resolution NWPs could yield improvement in model performance. Additional storm-observation data, such as radar and lightning, could also be incorporated to create a target function that better describes the convective events. Future efforts should consider approaches capable of integrating multiple data sources to improve the measurement and prediction of



(a) FCN Model Prediction based on forecast realeased on July 25th, 00:00 UTC .

(b) FCN Model Prediction based on forecast realeased on July 25th, 12:00 UTC .

(c) FCN Model Prediction based on forecast realeased on July 26th, 00:00 UTC .

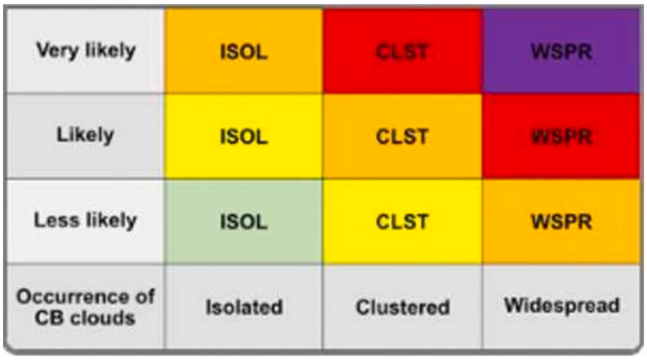
(d) Target function based on convective storm observations from RDT Product.

(e) Active weather en-route regulations in ECAC region.

Fig. 17. Case study example of showing timeline of model prediction starting at D-1, storm observations, and weather regulation for July 26th, 2019 10:00–13:00 UTC.

additional convective storm elements. For example, data from radar and lightning detection would help to better categorize the storm intensity, while satellite data could help to measure the cloud-top height. Rather than only providing the probability of convection, machine learning models can be designed to provide several outputs relating to storm elements that are most relevant to the task of air-traffic flow management, helping traffic managers take more informed decisions with greater lead times.

A possible application of this research could be to develop a convection advisory framework similar to what is in operation today. Machine learning models could be continuously trained and validated with incoming data in order to provide the best possible forecast. The foreseen product could provide automatic weather advisories in a digital format that could be distributed to various stakeholders and integrated with current operational ATFM platforms. Improved weather information at greater lead times would result in better situational



Very likely	ISOL	CLST	WSPR
Likely	ISOL	CLST	WSPR
Less likely	ISOL	CLST	WSPR
Occurrence of CB clouds	Isolated	Clustered	Widespread

Fig. 18. Standard convection risk matrix used in Cross-border Convection Advisory Product.

awareness, better resource allocation, and possible reduction of delays in the network. In this paper, we have provided a case study example based on a historic day with very high convective activity in the European airspace. Although the results show that the relevant convective information can be provided with higher time resolution and longer lead times, additional work is still required to transform the machine-learning model results into the convection risk matrix used by air-traffic managers.

Lastly, additional research is needed to better determine the impact of convective weather on traffic demand and sector capacities. This, in essence, is the real challenge in dealing with weather. Nevertheless, this study constitutes is a contribution toward providing air-traffic managers with tools to enable the pre-tactical planning of ATFM mitigation strategies during weather events.

CRediT authorship contribution statement

Aniel Jardines: Conceptualization, Methodology, Data curation, Writing – original draft. **Hamidreza Eivazi:** Methodology, Data curation, Writing – original draft. **Elias Zea:** Methodology, Data curation , Writing – original draft. **Javier García-Heras:** Supervision, Writing – review & editing. **Juan Simarro:** Conceptualization, Resources, Review. **Evelyn Otero:** Writing – review & editing. **Manuel Soler:** Supervision, Conceptualization, Funding acquisition, Writing – review & editing. **Ricardo Vinuesa:** Supervision, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We would like to acknowledge the financial support of the Spanish Ministry of Science, Innovation and Universities, Spain under grant RTI2018-098471-B-C32. The project has also received funding from the SESAR Joint Undertaking within the framework of the European Union's Horizon 2020 research and innovation programme under grant agreement No 891965. EZ, EO and RV acknowledge the financial support by the KTH, Sweden Digitalization Platform.

References

CAE (2015). *ATPL ground training, CAE Oxford aviation academy Meteorology*. CAE Oxford Aviation Academy (UK).

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1800–1807). http://dx.doi.org/10.1109/CVPR.2017.195.

Collins, W., & Tissot, P. (2015). An artificial neural network model to predict thunderstorms within 400 km^2 south Texas domains. *Meteorological Applications*, *22*(3), 650–665.

Eivazi, H., Le Clainche, S., Hoyas, S., & Vinuesa, R. (2022). Towards extraction of orthogonal and parsimonious non-linear modes from turbulent flows. *Expert Systems with Applications*, *202*, Article 117038.

EUROCONTROL (2019). 2018 Performance review report, "an assessment of air traffic management in Europe during the calendar year 2018". https://www.eurocontrol.int/air-navigation-services-performance-review.

Evans, J. E., & Ducot, E. R. (2006). Corridor integrated weather system. *Lincoln Laboratory Journal*, *16*(1), 59.

Guastoni, L., Güemes, A., Ianiro, A., Discetti, S., Schlatter, P., Azizpour, H., & Vinuesa, R. (2021). Convolutional-network models to predict wall-bounded turbulence from wall quantities. *Journal of Fluid Mechanics*, *928*, A27.

Güemes, A., Discetti, S., Ianiro, A., Sirmacek, B., Azizpour, H., & Vinuesa, R. (2021). From coarse wall measurements to turbulent velocity fields through deep learning. *Physics of Fluids*, *33*, Article 075121.

Han, L., Sun, J., & Zhang, W. (2019). Convolutional neural network for convective storm nowcasting using 3-D Doppler weather radar data. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(2), 1487–1495.

He, J., & Loboda, T. V. (2020). Modeling cloud-to-ground lightning probability in Alaskan tundra through the integration of weather research and forecast (WRF) model and machine learning method. *Environmental Research Letters*, *15*(11), Article 115009.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

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*, Article 100053.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794.

Lagerquist, R., McGovern, A., Homeyer, C. R., Gagne II, D. J., & Smith, T. (2020). Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Monthly Weather Review*, *148*(7), 2837–2861.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, J.-G., Min, K.-H., Park, H., Kim, Y., Chung, C.-Y., & Chang, E.-C. (2020). Improvement of the rapid-development thunderstorm (RDT) algorithm for use with the GK2a satellite. *Asia-Pacific Journal of Atmospheric Sciences*, *56*(2), 307–319,

Lin, M., Chen, Q., & Yan, S. (2014). Network in network. CoRR abs/1312.4400.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523–3542.

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., & Smith, L. (2006). Ensemble prediction: a pedagogical perspective. *ECMWF Newsletter*, *106*(106), 10–17.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing, ISBN: 978-3-319-24574-4.

Šaur, D. (2017). Forecasting of convective precipitation through NWP models and algorithm of storms prediction. In R. Silhavy, R. Senkerik, Z. Kominkova Oplatkova, Z. Prokopova, & P. Silhavy (Eds.), *Artificial intelligence trends in intelligent systems* (pp. 125–136). Cham: Springer International Publishing, ISBN: 978-3-319-57261-1.

Simon, T., Fabsic, P., Mayr, G. J., Umlauf, N., & Zeileis, A. (2018). Probabilistic forecasting of thunderstorms in the eastern Alps. *Monthly Weather Review*, *146*(9), 2999–3009.

Sirmacek, B., & Vinuesa, R. (2022). Remote sensing and AI for building climate adaptation applications. *Results in Engineering*, *15*, Article 100524.

Srinivasan, P. A., Guastoni, L., Azizpour, H., Schlatter, P., & Vinuesa, R. (2019). Predictions of turbulent shear flows using deep neural networks. *Physical Review Fluids*, *4*(5), Article 054603.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, *11*, 233.

Vinuesa, R., & Brunton, S. L. (2022). Enhancing computational fluid dynamics with machine learning. *Nature Computational Science*, *2*(6), 358–366.

Vinuesa, R., & Sirmacek, B. (2021). Interpretable deep-learning models to help achieve the sustainable development goals. *Nature Machine Intelligence*, *3*(11), 926.

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9), Article e2020MS002109.

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*(10), 2079–2100.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 6230–6239). http://dx.doi.org/10.1109/CVPR.2017.660.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2018). Semantic understanding of scenes through the ADE20K dataset. arXiv:1608.05442.

Zhou, K., Zheng, Y., Li, B., Dong, W., & Zhang, X. (2019). Forecasting different types of convective weather: A deep learning approach. *Journal of Meteorological Research*, *33*(5), 797–809.